# Ci si può fidare dei risultati dell'analisi delle componenti principali?

**Ricci C, Milani S**

**Istituto di Statistica Medica e Biometria G.A. Maccacaro – Facoltà di Medicina e Chirurgia, Università degli Studi di Milano**

## Scopo dello studio

L'analisi delle componenti principali (PCA) ha lo scopo di trasformare un set di *m* variabili y in un set di *m* nuove variabili x (dette *componenti*) tra di loro ortogonali e combinazioni lineari delle precedenti, in modo tale che gran parte della varianza possa essere spiegata dalle prime $p<m$ componenti, dette appunto *componenti principali*. Si tratta quindi di un metodo per rappresentare sinteticamente, con un minor numero di variabili, una struttura complessa, come quelle emergenti dagli studi antropometrici, psicologici o sociologici. Il significato delle nuove variabili viene in genere attribuito, *a posteriori*, attraverso l'esame delle correlazioni tra le nuove variabili e quelle originali. Alternativamente, le variabili originali possono essere pensate, *a priori*, come *proxy* di uno o più *fattori* che costituiscono le caratteristiche individuali non direttamente osservabili (*domini*). In questo senso, la PCA è un caso particolare dell'analisi dei fattori. Benché la PCA sia una tecnica essenzialmente descrittiva, il ricercatore non rinuncia a scegliere, sulla scorta dei risultati emersi dall'analisi, il numero di componenti da considerare e a dare un'interpretazione a tali componenti, spesso senza domandarsi se la componente cui si sforza di dare un significato biologico, psicologico o comportamentale rifletta un aspetto della realtà o sia solo frutto del caso.

Scopo di questo studio di simulazione è descrivere la distribuzione degli autovalori nel caso di *m* deviate gaussiane indipendenti o correlate a fattori non osservabili, e valutare quanto la distribuzione delle variabili originali ha effetto sui risultati. Quest'ultimo punto riveste particolare importanza quando le variabili sono rilevate, come spesso accade, con scale Likert.

## Metodi

In una prima fase, si è studiata la relazione tra distribuzione degli autovalori e numero *m* di variabili e dimensione *n* del campione, nell'ipotesi di *m* deviate gaussiane standard indipendenti: si è posto *m* pari a 4, 9, 16, 25, 36, 49 ed *n* pari a 80, 400, 2000, 10000, e si sono considerate tutte le combinazioni possibili di *m* ed *n*. Per ognuna di tali combinazioni si sono generati 4000 campioni. Questa stessa procedura è stata ripetuta dopo aver trasformato i valori delle deviate gaussiane (z) in valori di scale Likert simmetriche o asimmetriche (*sim, asim*), senza e con trasformazione *normit* (*simN, asimN*) secondo lo schema:

| | distribuzione simmetrica (*sim* / *simN*) | | | | | distribuzione asimmetrica (*asim* / *asimN*) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *norm* (z) | <–1.20 | –1.20, –0.40 | –0.40, +0.40 | +0.40, +1.20 | >+1.20 | <–0.45 | –0.45, +0.30 | 0.30, +1.05 | +1.05, +1.80 | >+1.80 |
| p(z) | 0.115 | 0.230 | 0.310 | 0.230 | 0.115 | 0.326 | 0.292 | 0.235 | 0.111 | 0.036 |
| Likert-score | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| *normit*-score | –1.685 | –0.758 | 0.000 | +0.758 | +1.685 | –1.104 | –0.072 | +0.644 | +1.359 | +2.188 |

In una seconda fase, si sono valutati gli effetti del numero di variabili e della dimensione campionaria sulla probabilità di considerare significative le prime *p* componenti principali, sulla correlazione tra fattori latenti e componenti principali, e sulla capacità di riottenere per mezzo degli score delle componenti principali le classificazioni definite per mezzo dei fattori latenti (soggetti con valore negativo *versus* soggetti con valore positivo del fattore latente). A tal fine si sono ripetute integralmente le procedure della prima fase, ma nell'ipotesi di deviate gaussiane standard correlate a due fattori latenti gaussiani e non correlati. Si è inoltre ipotizzato che metà delle variabili fosse correlata con il primo fattore latente (con $\rho$=0.6), un quarto con il secondo (con $\rho$=0.6) e le restanti con nessuno dei due. Si è assunta nulla la correlazione parziale delle variabili condizionata ai fattori.

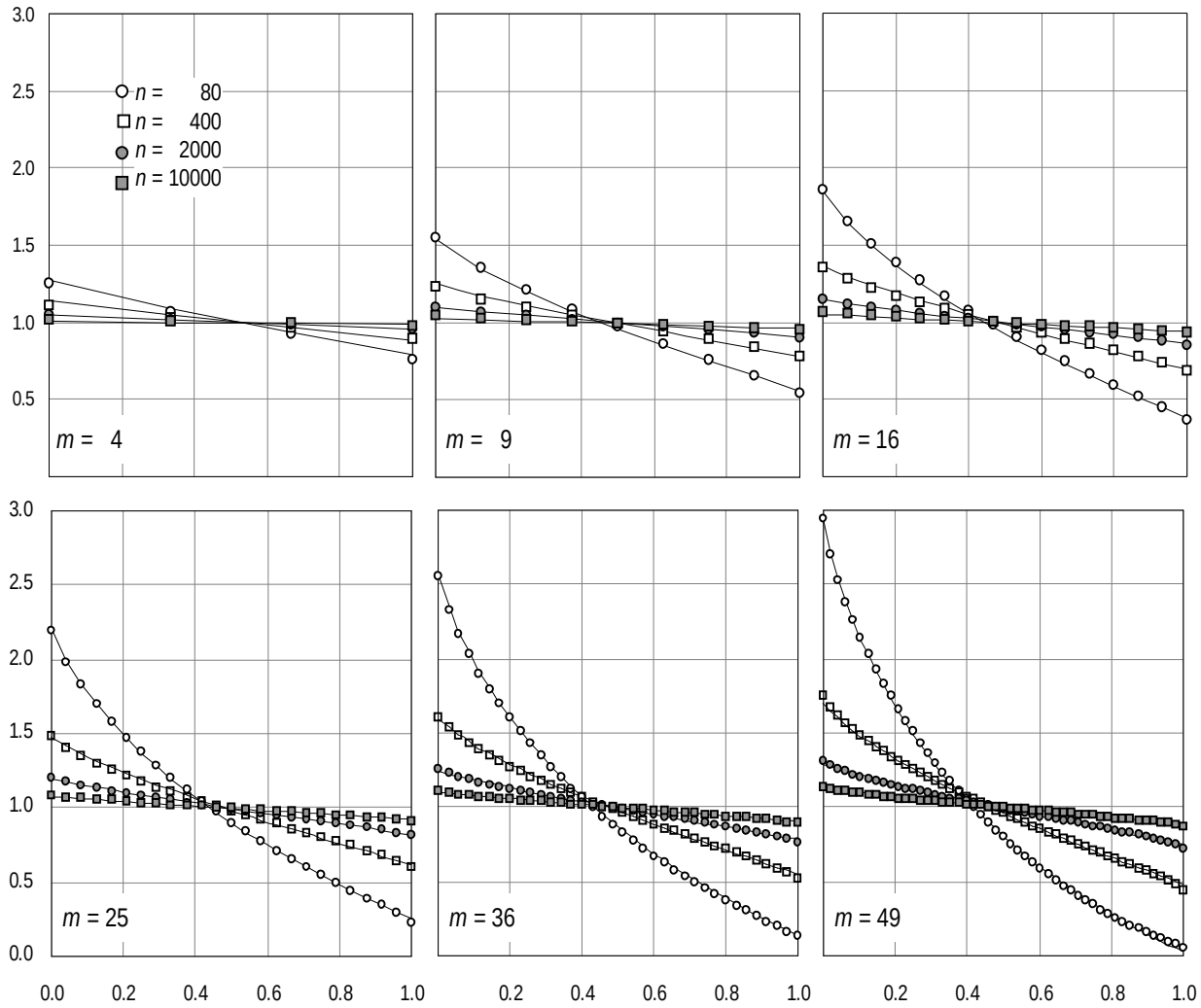
**Risultati**

La figura 1 rappresenta i valori attesi degli autovalori, nell'ipotesi di *m* deviate gaussiane standard indipendenti, in funzione della loro posizione relativa [*r*=(*i*-1)/*m*] nell'insieme ordinato in ordine decrescente, del numero di variabili analizzate, e della dimensione del campione (*n*). I valori attesi del primo autovalore aumentano all'aumentare del numero di variabili e diminuiscono all'aumentare della dimensione del campione. La forma della relazione tra valore atteso dell'autovalore e posizione relativa (*r*) è monotona decrescente,

ma varia al variare di *m* e di *n*, potendo presentare un punto di flesso in prossimità di *i*=0.5, per valori elevati di *n* e *m*.

Per ognuno dei setting descritti nella figura 1, si è calcolato il 95° centile della distribuzione degli autovalori. La probabilità che un autovalore ha di superare tale soglia è pressoché invariante al variare della distribuzione delle variabili originali, almeno nell'ambito delle distribuzioni qui esaminate: il rischio d'errore di tipo I è in ogni caso compreso tra 0.04 e 0.06.

**Figura 1.** Valori attesi degli autovalori in funzione della loro posizione relativa ($0 \le r \le 1$) nell'insieme ordinato in ordine decrescente del numero di variabili analizzate (*m* = 4, 9, 16, 25, 36, 49) e della dimensione del campione (*n*=80, 400, 2000, 10000). Risultati ottenuti dall'analisi di 4000 campioni generati per simulazione, nell'ipotesi di *m* deviate gaussiane standard indipendenti.




(ρ=0.6), un quarto con il secondo (ρ=0.6) e il restante quarto con nessuno dei

due, che la correlazione parziale delle variabili condizionata ai fattori sia nulla, e che i due fattori latenti siano gaussiani e non correlati.

| Assi | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| **0** | 41.28 | 49.75 | 49.60 | 52.80 | 52.30 | 0.52 | 1.47 | 1.53 | 2.18 | 2.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 53.42 | 45.72 | 45.75 | 43.30 | 43.82 | 46.45 | 53.08 | 53.25 | 54.73 | 54.90 | 1.60 | 3.20 | 3.33 | 4.65 | 4.30 |
| **1-2** | **5.30** | **4.53** | **4.65** | **3.90** | **3.88** | **48.70** | **41.50** | **41.30** | **39.25** | **39.35** | **96.72** | **94.12** | **93.97** | **92.60** | **93.00** |
| **1-3** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.28 | 3.85 | 3.85 | 3.77 | 3.52 | 1.68 | 2.65 | 2.67 | 2.67 | 2.62 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.10 | 0.07 | 0.07 | 0.01 | 0.00 | 0.03 | 0.03 | 0.08 | 0.08 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| **0** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 0.00 | 0.10 | 0.00 | 0.27 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-2** | **99.15** | **98.75** | **98.78** | **98.03** | **98.10** | **99.87** | **99.73** | **99.67** | **99.30** | **99.30** | **99.90** | **99.85** | **99.83** | **99.73** | **99.67** |
| **1-3** | 0.85 | 1.15 | 1.12 | 1.67 | 1.55 | 0.13 | 0.27 | 0.33 | 0.70 | 0.70 | 0.10 | 0.15 | 0.17 | 0.27 | 0.33 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Dimensione campionaria = 80

| Assi | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| **0** | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 87.78 | 88.05 | 88.17 | 87.42 | 87.95 | 0.00 | 0.13 | 0.13 | 0.20 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-2** | **12.22** | **11.95** | **11.83** | **12.55** | **12.05** | **92.57** | **91.52** | **91.57** | **91.60** | **91.40** | **99.75** | **99.67** | **99.75** | **99.43** | **99.45** |
| **1-3** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 7.38 | 8.33 | 8.28 | 8.20 | 8.42 | 0.25 | 0.33 | 0.25 | 0.57 | 0.55 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.02 | 0.02 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| **0** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-2** | **99.98** | **99.85** | **99.90** | **99.85** | **99.80** | **100.00** | **100.0** | **100.00** | **99.97** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **1-3** | 0.02 | 0.15 | 0.10 | 0.15 | 0.20 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Dimensione campionaria = 400

| Assi | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| **0** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 87.55 | 87.10 | 87.15 | 87.05 | 87.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-2** | **12.45** | **12.90** | **12.85** | **12.95** | **12.80** | **93.85** | **94.30** | **94.33** | **94.27** | **94.00** | **99.95** | **99.88** | **99.88** | **99.82** | **99.82** |
| **1-3** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.15 | 5.70 | 5.67 | 5.73 | 6.00 | 0.05 | 0.12 | 0.12 | 0.18 | 0.18 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| **0** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-2** | **99.98** | **99.97** | **99.98** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **1-3** | 0.02 | 0.03 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Dimensione campionaria = 2000

| Assi | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* | *norm* | *sim* | *simN* | *asim* | *asimN* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| **0** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 88.22 | 87.98 | 87.78 | 87.50 | 87.52 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1-2 | **11.78** | **12.02** | **12.22** | **12.50** | **12.48** | **93.30** | **93.38** | **93.55** | **92.95** | **92.95** | **99.88** | **99.92** | **99.90** | **99.85** | **99.80** |
| 1-3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.70 | 6.62 | 6.45 | 7.05 | 7.05 | 0.12 | 0.08 | 0.10 | 0.15 | 0.20 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| **0** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-2** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **1-3** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **1-4** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Dimensione campionaria = 10000

I risultati riportati nelle tabelle 1-3, sono stati ottenuti, nell'ipotesi che metà delle variabili sia correlata con il primo fattore latente ($\rho$=0.6), un quarto con il secondo ($\rho$=0.6) e le restanti con nessuno dei due, e nell'ulteriore ipotesi che la

correlazione parziale delle variabili condizionata ai fattori sia nulla, e che i due fattori latenti siano gaussiani e non correlati.

La tabella 1 riporta la probabilità di identificare come significativi i primi *p* assi fattoriali. Al crescere del numero di variabili (*m*) cresce la probabilità di identificare correttamente i due fattori latenti. Indipendentemente dalla distribuzione delle variabili originali, tale probabilità è maggiore del 98% quando *m* è almeno 25, con appena 80 soggetti, e quando *m* è almeno 16, con 400 soggetti o più. Un numero limitato di variabili rende difficoltosa l'identificazione del 2° asse fattoriale anche per elevate dimensioni campionarie: quando *m*=4 il 2° autovalore è significativo nel 5% circa dei casi per *n*=80 e nel 12% circa dei casi per *n* pari a 400 o più; quando *m*=9 il 2° autovalore è significativo in meno del 50% circa dei casi per *n*=80 e in oltre il 90% circa dei casi per *n* pari a 400 o più. La probabilità di identificare i due fattori latenti si riduce di poco quando le variabili sono espresse in scala Likert, la trasformazione in *normit* non produce effetti rilevanti.

La tabella 2 riporta la correlazione tra le prime due componenti principali e i fattori latenti (*F.1, F.2*). La correlazione tra componente e fattore cresce al crescere del numero di variabili correlate con il fattore e, in misura minore, con la dimensione campionaria. La correlazione si riduce di poco quando le variabili sono espresse in scala Likert: tale riduzione è più evidente quando la scala è asimmetrica. La trasformazione *normit* è irrilevante quando applicata alla scala Likert simmetrica e produce un piccolo aumento dei coefficienti di correlazione se applicata alla scala Likert asimmetrica.

La tabella 3 riporta le probabilità di classificare correttamente un soggetto rispetto al 1° (*F.1*) o al 2° (*F.2*) asse fattoriale, calcolate nell'ipotesi che si sia stabilito *a priori* di considerare i primi due assi, e solo i primi due assi, indipendentemente dai risultati della PCA. Emerge che la probabilità che un soggetto sia classificato in modo corretto in base al 1° (o al 2°) asse fattoriale dipende essenzialmente dal numero di variabili analizzate, e in misura trascurabile dalla dimensione del campione e dalla distribuzione delle variabili. Tale probabilità va da poco più del 70%, quando le variabili analizzate sono solo 4, ad oltre il 90% quando le variabili sono 49. L'effetto della distribuzione delle variabili originali è trascurabile.

**Tabella 2.** Correlazione tra le prime due componenti principali e i fattori latenti (*F.1, F.2*) nell'ipotesi che metà delle variabili sia correlata con il primo fattore latente (ρ=0.6), un quarto con il secondo (ρ=0.6) e le restanti con nessuno dei due, che la correlazione parziale delle variabili condizionata ai fattori sia nulla, e che i due fattori latenti siano gaussiani e non correlati.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dimensione campionaria = 80 | | | | | | | | | | | | | | |
| | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 0.685 | 0.656 | 0.656 | 0.641 | 0.643 | 0.808 | 0.786 | 0.786 | 0.769 | 0.773 | 0.892 | 0.878 | 0.878 | 0.863 | 0.867 |
| ***F.2*** | 0.361 | 0.333 | 0.334 | 0.319 | 0.321 | 0.620 | 0.581 | 0.581 | 0.556 | 0.559 | 0.783 | 0.759 | 0.759 | 0.741 | 0.744 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 0.923 | 0.912 | 0.912 | 0.899 | 0.903 | 0.946 | 0.938 | 0.938 | 0.926 | 0.931 | 0.958 | 0.951 | 0.952 | 0.940 | 0.944 |
| ***F.2*** | 0.839 | 0.820 | 0.820 | 0.803 | 0.806 | 0.883 | 0.868 | 0.869 | 0.854 | 0.858 | 0.906 | 0.895 | 0.895 | 0.881 | 0.885 |
| | Dimensione campionaria = 400 | | | | | | | | | | | | | | |
| | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 0.718 | 0.693 | 0.692 | 0.676 | 0.678 | 0.827 | 0.808 | 0.807 | 0.793 | 0.795 | 0.902 | 0.889 | 0.888 | 0.876 | 0.878 |
| ***F.2*** | 0.409 | 0.391 | 0.389 | 0.380 | 0.379 | 0.702 | 0.675 | 0.674 | 0.659 | 0.661 | 0.822 | 0.803 | 0.802 | 0.788 | 0.790 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 0.931 | 0.921 | 0.921 | 0.909 | 0.912 | 0.952 | 0.945 | 0.944 | 0.934 | 0.936 | 0.963 | 0.957 | 0.956 | 0.946 | 0.949 |
| ***F.2*** | 0.870 | 0.854 | 0.854 | 0.840 | 0.843 | 0.907 | 0.895 | 0.894 | 0.882 | 0.885 | 0.927 | 0.917 | 0.917 | 0.905 | 0.908 |
| | Dimensione campionaria = 2000 | | | | | | | | | | | | | | |
| | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 0.725 | 0.702 | 0.702 | 0.685 | 0.688 | 0.831 | 0.813 | 0.813 | 0.798 | 0.800 | 0.904 | 0.891 | 0.892 | 0.878 | 0.881 |
| ***F.2*** | 0.419 | 0.401 | 0.401 | 0.389 | 0.391 | 0.722 | 0.698 | 0.698 | 0.682 | 0.684 | 0.830 | 0.812 | 0.812 | 0.796 | 0.799 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 0.933 | 0.923 | 0.923 | 0.911 | 0.914 | 0.954 | 0.946 | 0.946 | 0.935 | 0.938 | 0.965 | 0.958 | 0.958 | 0.948 | 0.951 |
| ***F.2*** | 0.877 | 0.862 | 0.862 | 0.848 | 0.851 | 0.912 | 0.901 | 0.901 | 0.889 | 0.891 | 0.932 | 0.922 | 0.922 | 0.910 | 0.913 |
| | Dimensione campionaria = 10000 | | | | | | | | | | | | | | |
| | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 0.727 | 0.704 | 0.704 | 0.688 | 0.690 | 0.832 | 0.814 | 0.814 | 0.799 | 0.802 | 0.904 | 0.892 | 0.892 | 0.879 | 0.882 |
| ***F.2*** | 0.422 | 0.405 | 0.405 | 0.393 | 0.395 | 0.726 | 0.703 | 0.704 | 0.687 | 0.690 | 0.832 | 0.814 | 0.814 | 0.799 | 0.802 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 0.933 | 0.923 | 0.924 | 0.911 | 0.915 | 0.954 | 0.947 | 0.947 | 0.935 | 0.939 | 0.965 | 0.959 | 0.959 | 0.948 | 0.952 |
| ***F.2*** | 0.878 | 0.863 | 0.864 | 0.849 | 0.853 | 0.914 | 0.902 | 0.902 | 0.889 | 0.893 | 0.933 | 0.923 | 0.924 | 0.911 | 0.915 |

**Tabella 3.** Probabilità di classificare correttamente un soggetto rispetto al 1° (*F.1*) o al 2° (*F.2*) asse fattoriale, nell'ipotesi che metà delle variabili sia correlata con il primo fattore latente (ρ=0.6), un quarto con il secondo (ρ=0.6) e le restanti con nessuno dei due, che la correlazione parziale delle variabili condizionata ai fattori sia nulla, e che i due fattori latenti siano gaussiani e non correlati. Le probabilità sono calcolate nell'ipotesi che si sia stabilito *a priori* di considerare i primi due assi, e soloi primi due assi, indipendentemente dai risultati della PCA.

| Dimensione campionaria = 80 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classi | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 73.75 | 72.50 | 72.50 | 72.50 | 72.50 | 80.00 | 78.75 | 78.75 | 78.75 | 78.75 | 85.00 | 83.75 | 83.75 | 83.75 | 83.75 |
| ***F.2*** | 61.25 | 60.00 | 60.00 | 60.00 | 60.00 | 70.00 | 70.00 | 68.75 | 68.75 | 68.75 | 78.75 | 77.50 | 77.50 | 77.50 | 77.50 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 87.50 | 86.25 | 86.25 | 86.25 | 86.25 | 88.75 | 88.75 | 88.75 | 88.75 | 88.75 | 90.00 | 90.00 | 90.00 | 88.75 | 90.00 |
| ***F.2*** | 81.25 | 81.25 | 81.25 | 80.00 | 80.00 | 83.75 | 83.75 | 83.75 | 83.75 | 83.75 | 85.00 | 85.00 | 85.00 | 85.00 | 85.00 |
| Dimensione campionaria = 400 | | | | | | | | | | | | | | | |
| Classi | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 75.25 | 74.50 | 74.50 | 74.50 | 74.50 | 81.00 | 80.25 | 80.25 | 79.75 | 80.00 | 85.75 | 85.00 | 85.00 | 84.75 | 84.75 |
| ***F.2*** | 63.00 | 62.50 | 62.50 | 62.50 | 62.50 | 74.75 | 74.00 | 74.00 | 73.75 | 73.75 | 80.75 | 80.00 | 80.00 | 79.50 | 79.75 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 88.00 | 87.50 | 87.50 | 87.00 | 87.25 | 90.00 | 89.50 | 89.50 | 89.00 | 89.25 | 91.25 | 90.75 | 90.75 | 90.25 | 90.50 |
| ***F.2*** | 83.50 | 83.00 | 82.75 | 82.50 | 82.50 | 86.00 | 85.50 | 85.50 | 85.00 | 85.25 | 87.75 | 87.25 | 87.25 | 86.75 | 86.75 |
| Dimensione campionaria = 2000 | | | | | | | | | | | | | | | |
| Classi | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 75.80 | 74.65 | 74.65 | 75.20 | 75.10 | 81.20 | 80.35 | 80.35 | 79.95 | 80.25 | 85.90 | 85.30 | 85.30 | 84.85 | 85.00 |
| ***F.2*** | 63.60 | 62.75 | 62.80 | 63.05 | 63.00 | 75.65 | 74.65 | 74.65 | 75.00 | 74.90 | 81.15 | 80.30 | 80.32 | 79.95 | 80.15 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 88.25 | 87.70 | 87.70 | 87.20 | 87.38 | 90.25 | 89.80 | 89.75 | 89.25 | 89.45 | 91.50 | 91.10 | 91.05 | 90.45 | 90.70 |
| ***F.2*** | 84.00 | 83.30 | 83.30 | 83.00 | 83.05 | 86.55 | 86.00 | 85.95 | 85.50 | 85.65 | 88.15 | 87.70 | 87.65 | 87.15 | 87.30 |
| Dimensione campionaria = 10000 | | | | | | | | | | | | | | | |
| Classi | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** | ***norm*** | ***sim*** | ***simN*** | ***asim*** | ***asimN*** |
| | *4 variabili* | | | | | *9 variabili* | | | | | *16 variabili* | | | | |
| ***F.1*** | 75.90 | 74.65 | 74.66 | 75.25 | 75.23 | 81.28 | 80.35 | 80.37 | 80.11 | 80.38 | 85.97 | 85.29 | 85.32 | 85.02 | 85.10 |
| ***F.2*** | 63.82 | 62.63 | 62.68 | 63.34 | 63.30 | 75.88 | 74.65 | 74.66 | 75.26 | 75.20 | 81.26 | 80.35 | 80.39 | 80.06 | 80.34 |
| | *25 variabili* | | | | | *36 variabili* | | | | | *49 variabili* | | | | |
| ***F.1*** | 88.29 | 87.73 | 87.75 | 87.38 | 87.45 | 90.29 | 89.84 | 89.84 | 89.22 | 89.52 | 91.53 | 91.13 | 91.12 | 90.58 | 90.77 |
| ***F.2*** | 84.11 | 83.36 | 83.38 | 83.08 | 83.16 | 86.67 | 86.05 | 86.06 | 85.44 | 85.78 | 88.28 | 87.75 | 87.75 | 87.32 | 87.46 |

La tabella 3 riporta le probabilità di classificare correttamente un soggetto rispetto al 1° (*F.1*) o al 2° (*F.2*) asse fattoriale, calcolate nell'ipotesi che si sia stabilito *a priori* di considerare i primi due assi, e solo i primi due assi, indipendentemente dai risultati della PCA. Emerge che la probabilità che un soggetto sia classificato in modo corretto in base al 1° (o al 2°) asse fattoriale dipende essenzialmente dal numero di variabili analizzate, e in misura trascurabile dalla dimensione del campione e dalla distribuzione delle variabili.

Tale probabilità va da poco meno dell'80%, quando le variabili analizzate sono solo 4, ad oltre il 90% quando le variabili sono 49.

## Conclusioni

Questo studio di simulazione mostra che l'analisi delle componenti principali costituisce una tecnica affidabile per identificare le strutture latenti anche in campioni di dimensione limitata, purché vi sia un numero sufficiente di variabili e gran parte di queste (almeno 3/4) sia almeno moderatamente correlata con uno dei fattori latenti. La tecnica appare alquanto robusta nei confronti della distribuzione delle variabili rilevate, anche con scale ordinali a pochi livelli e in presenza di forti asimmetrie. Pertanto, almeno nell'ambito dei setting qui esplorati, la trasformazione *normit* appare del tutto ininfluente. La PCA appare meno efficace nel classificare i soggetti in base agli score delle componenti principali. Nel caso di una classificazione dicotomica (valori positivi *versus* valori negativi), la percentuale di soggetti classificati correttamente secondo il primo fattore supera di poco il 90%, anche con elevato numero di variabili e grandi dimensioni campionarie.

Quali siano le performance della PCA nella descrizione di strutture più complesse (ad esempio 3 o più fattori non necessariamente indipendenti, minori livelli di correlazione variabile-fattore, distribuzione differente delle variabili che concorrono alla struttura) sarà oggetto di future indagini.