L'analisi dei componenti principali e l'analisi dei fattori

Antonio Greco

Dipartimento di Matematica, via Ospedale 72,
09124 Cagliari. E-mail: greco@unica.it

20 agosto 2004

1. Introduzione.

L'analisi dei componenti principali *(principal components analysis)* e l'analisi dei fattori *(factor analysis)* sono tecniche matematiche per interpretare grandi quantità di dati provenienti dalle osservazioni sperimentali, tecniche che fanno parte della più generale *Analisi multivariata.*

Esistono vari programmi per calcolatore che consentono di eseguirle in modo automatico, fra i quali SPSS, StatSoft STATISTICA, MATLAB, ed altri ancora.

Queste tecniche trovano applicazione in discipline molto diverse tra loro, come ad esempio le Scienze della Terra, la Medicina, la Psicologia.

Furono degli psicologi ad introdurre l'analisi dei fattori, ed in particolare Spearman [5] nel 1904. Essi tentavano di calcolare, a partire dai punteggi ottenuti in vari test di intelligenza, un unico valore che fosse in grado di misurare l'intelligenza in assoluto (si veda [1, pag. 475], [3, pag. 127], [2, pag. 165], e la bibiliografia ivi citata).

Un esempio di applicazione dell'analisi dei fattori nel campo della medicina può essere il seguente: calcolare, a partire dai risultati di varie analisi effettuate su di un neonato, un unico valore che sia in grado di misurare la sua predisposizione a morire improvvisamente, onde effettuare con anticipo una terapia preventiva [4, pag. 2].

Sviluppiamo ora, più in dettaglio, un esempio legato alle scienze dalla terra. Supponiamo di avere esaminato N campioni rocciosi (N = numero dei campioni esaminati), e di avere misurato in ciascuno di essi la concentrazione di k elementi chimici (k = numero degli elementi considerati).

In altre parole, per ogni campione esaminato abbiamo k numeri, che indicano la concentrazione (espressa, poniamo, in parti per milione) di ciascuno dei k elementi considerati. Ad esempio, se siamo interessati solamente al contenuto di manganese e piombo, allora $k = 2$ ed in uno dei campioni potremmo trovare 150 ppm di manganese e 100 ppm di piombo.

La matrice (x_{nj}), che ha N righe e k colonne, viene detta matrice dei *dati grezzi (raw data).* Ogni sua riga riporta i valori delle diverse variabili osservate in un

singolo campione, mentre ogni sua colonna riporta i valori di una singola variabile osservata nei vari campioni.

Quando si considerano solo due variabili (quando, cioè, $k = 2$), si possono rappresentare i dati mediante un diagramma a dispersione *(scatterplot)*. In esso, ogni campione esaminato viene rappresentato in un piano cartesiano mediante un punto, le cui coordinate sono i valori delle due variabili osservate. Ad esempio, l'ipotetico campione menzionato sopra sarebbe rappresentato dal punto di coordinate $(150, 100)$.

La possibilità di servirsi di questa rappresentazione grafica, ancora possibile con $k = 3$, viene meno se $k > 3$, perché occorrono più di tre assi cartesiani. Tuttavia, ciascun campione può ancora essere rappresentato mediante k coordinate, e si continua a chiamarlo *punto* per analogia con il caso $k = 2$.

È qui che entrano in gioco le tecniche di analisi multivariata: esse possono servire per capire la distribuzione dei *punti* anche quando le variabili osservate sono molte, e non è possibile tracciare un grafico a dispersione.

2. Le direzioni principali.

Per riprendere l'esempio del paragrafo precedente, poniamo il caso che gli elementi chimici osservati siano presenti in tutti i campioni nelle stesse proporzioni, o almeno in proporzioni simili: noi ci aspettiamo che ciò avvenga, per le nostre conoscenze sullo specifico problema, e vogliamo mettere in evidenza questo fatto.

La situazione (ideale) sopra richiamata corrisponde, in un diagramma a dispersione, al caso in cui i *punti* che rappresentano i campioni sono disposti lungo una linea retta passante per l'origine.

È naturale, allora, considerare la direzione di questa retta come una *direzione principale*. Non solo: potremmo voler individuare ciascun campione tramite una nuova coordinata, corrente lungo l'asse principale, al posto delle due coordinate iniziali.

In realtà, però, i *punti* non sono distribuiti esattamente lungo una retta, e comunque, come abbiamo visto, non sempre è possibile ricorrere ad un grafico a dispersione. Come si può parlare, anche in tal caso, di direzioni principali?

Per rispondere a questa domanda cominciamo con l'osservare che, se $k = 2$ e se i punti del grafico a dispersione sono disposti lungo una retta *parallela ad uno degli assi coordinati*, allora le due variabili osservate (i due elementi chimici, nell'esempio) sono *incorrelate*, cioè la loro *covarianza* è nulla.

Se la covarianza non è nulla, come capita in generale, possiamo *ruotare* gli assi cartesiani e ricalcolare le coordinate dei punti rispetto ai nuovi assi: questo comporta vantaggi e svantaggi.

Uno svantaggio è che le nuove coordinate dei punti hanno un significato diverso dalle vecchie coordinate, un significato la cui interpretazione resta a carico dell'analista. Ad esempio, se inizialmente le coordinate dei punti rappresentavano la

concentrazione di certi elementi chimici, le nuove coordinate dopo la rotazione degli assi non rappresentano affatto la concentrazione di un qualche elemento.

Un vantaggio è che, se i nuovi assi cartesiani sono scelti in modo opportuno, le nuove variabili (il cui significato resta da stabilire) sono *incorrelate*, e perciò le direzioni dei nuovi assi si possono considerare *direzioni principali*, per analogia con il caso in cui i punti sono disposti lungo una retta.

3. Diagonalizzazione.

Ricordiamo che i k valori associati ad un campione individuano un punto $X_n = (x_{n1}, \ldots, x_{nk})$ in uno spazio a k dimensioni (lo spazio $\mathbb{R}^k$). Il numero k delle variabili osservate viene stabilito dallo sperimentatore, come pure il numero N dei campioni esaminati. La variabile n va da 1 ad N.

Il punto di partenza matematico sia per l'analisi dei componenti principali che per l'analisi dei fattori è costituito dal seguente teorema, per la cui dimostrazione si rimanda a [6, Prop. 9.31], o ad un qualunque testo di algebra lineare:

Teorema 1. *Qualunque matrice simmetrica è diagonalizzabile mediante una matrice di passaggio ortogonale.*

Nell'analisi dei componenti principali, e nell'analisi dei fattori, una *matrice simmetrica* alla quale può essere applicato il teorema è la *matrice di correlazione* (ρ_{ij}) delle variabili considerate, e cioè la matrice le cui componenti sono:

$$\rho_{ij} = \frac{1}{N\,\sigma_i\,\sigma_j} \sum_{n=1}^{N} (x_{ni} - \overline{x}_i)\,(x_{nj} - \overline{x}_j)\,, \tag{1}$$

dove $\overline{x}_i$ e σ_i denotano la media e la deviazione standard dei valori $(x_{1i}, \ldots, x_{ni})$ (relativi, cioè, all'i-esima variabile osservata). Poiché le deviazioni standard figurano al denominatore (e non si può dividere per zero), affinché la matrice di correlazione sia ben definita è necessario che la varianza di ciascuna delle k variabili in gioco sia maggiore di zero.

Si osservi che la matrice (ρ_{ij}) è quadrata e di ordine k, cioè, per riprendere l'esempio precedente, ha tante righe e tante colonne quanti sono i componenti misurati in ciascun campione. Inoltre, scambiando i e j tra loro, i prodotti $(x_{ni} - \overline{x}_i)\,(x_{nj} - \overline{x}_j)$ non cambiano, né cambia la loro somma al variare dell'indice n, quindi $\rho_{ij} = \rho_{ji}$. Quest'ultima proprietà si esprime dicendo che la matrice di correlazione è *simmetrica*, come vuole il Teorema 1.

La *matrice ortogonale* di cui si parla nel teorema rappresenta, invece, la *rotazione* degli assi cartesiani.

Il Teorema 1 dice dunque che *ruotando opportunamente gli assi cartesiani* di $\mathbb{R}^k$ si può far sì che la matrice di correlazione diventi diagonale, cioè tale che tutti gli elementi fuori della diagonale principale (quelli con $i \neq j$) siano nulli. Quindi le nuove variabili, correnti lungo gli assi ruotati, sono incorrelate.

Gli assi cartesiani rispetto ai quali la matrice di correlazione è diagonale vengono considerati *direzioni principali* della *nuvola* dei dati.

4. La standardizzazione.

Se abbiamo dei valori numerici $y_1, \ldots, y_N$, il valor medio $\overline{y}$ e la varianza σ^2 sono dati rispettivamente da

$$\overline{y} = \frac{1}{N} \sum_{n=1}^{N} y_n, \qquad \sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \overline{y})^2. \tag{2}$$

Si dice *deviazione standard* il valore σ dato dalla radice quadrata della varianza. A partire da una distribuzione $(y_1, \ldots, y_N)$, attraverso la trasformazione

$$x_n = \frac{y_n - \overline{y}}{\sigma}, \qquad n = 1, \ldots, N, \tag{3}$$

si ottiene una distribuzione $(x_1, \ldots, x_N)$ il cui valor medio e la cui varianza sono uguali a 0 e 1, rispettivamente. Per verificarlo basta svolgere i calcoli usando, in luogo di x_n, la sua espressione data dalla (3). Si dice che si è effettuata una *standardizzazione*, ed i valori $x_1, \ldots, x_N$ si dicono *standardizzati*.

A titolo puramente esemplificativo, è facile *standardizzare* la distribuzione di due valori $y_1 = 1$, $y_2 = 2$. Applicando la (3) si trova $x_1 = -1$, $x_2 = 1$. Un esercizio molto istruttivo è quello di standardizzare un campione di due valori numerici scelti a piacere.

Affinché sia possibile effettuare la standardizzazione è necessario che la varianza σ^2 sia diversa da zero, altrimenti si annulla il denominatore nella (3) e tale espressione non ha significato. La varianza è uguale a zero soltanto nel caso, piuttosto inverosimile, in cui tutti i valori $y_1, \ldots, y_N$ sono uguali tra loro. Questo discende dal fatto che, come si vede dalla (2), la varianza si ottiene come una somma di quadrati, e quindi è nulla se e solo se ciascun addendo è uguale a zero.

Tornando a considerare la matrice di correlazione (1), supponiamo che i dati siano stati preventivamente *standardizzati* per portare ai valori 0 e 1, rispettivamente, la media e la varianza di ognuna delle k variabili osservate. Allora gli elementi ρ_{ij} di tale matrice assumono la seguente semplice espressione:

$$\rho_{ij} = \frac{1}{N} \sum_{n=1}^{N} x_{ni}\, x_{nj}. \tag{4}$$

5. Direzioni *più principali* di altre.

Oltre ad individuare gli assi principali (che, lo ricordiamo, sono tanti quante le variabili osservate, e quindi possono essere decine) è anche possibile stabilire matematicamente se alcuni di essi sono più importanti degli altri, cioè seguono meglio, in

qualche senso, la nuvola dei dati. Questa possibilità è legata al concetto di *autovalore* (eigenvalue).

Un numero λ si dice *autovalore* per una matrice A se vi sono vettori X (diversi dal vettore nullo) le cui coordinate soddisfano l'uguaglianza

$$AX = \lambda X, \tag{5}$$

dove AX rappresenta il prodotto della matrice A per il vettore X, mentre λX è il prodotto del vettore X per lo scalare λ. Il vettore $X = (0, \ldots, 0)$, detto *vettore nullo*, non viene preso in considerazione nella definizione degli autovalori perché soddisfa l'uguaglianza (5) qualunque sia il valore di λ.

I vettori X non nulli che soddisfano l'uguaglianza (5) si dicono *autovettori* della matrice A. Si veda il paragrafo 11.3 per un esempio numerico.

Osservazioni.

1. Una matrice può avere vari autovalori, e cambiando autovalore cambiano anche gli autovettori da sostituire nella (5) affinché l'uguaglianza sia soddisfatta.
2. Gli autovettori associati ad un dato autovalore sono infiniti. Infatti se X è un autovettore allora lo sono anche tutti i suoi multipli per un coefficiente reale (diverso da zero).
3. Se X e Y sono due autovettori *associati al medesimo autovalore*, allora anche la somma $X + Y$ lo è.

Avendo in mente le osservazioni precedenti possiamo meglio interpretare il seguente teorema, strettamente legato al Teorema 1. Per la dimostrazione si rimanda a [6], o ad un qualunque testo di algebra lineare.

Teorema 2.

a. Gli assi rispetto ai quali una matrice simmetrica A diventa diagonale sono formati da autovettori di A.
b. Gli elementi della diagonale principale della matrice diagonalizzata sono gli autovalori di A.
c. La somma degli elementi della diagonale principale non viene modificata dalla diagonalizzazione.

Grazie a questo teorema possiamo misurare l'*importanza* di una direzione principale andando a guardare quanto è grande l'*autovalore* corrispondente ad essa. Questo è legato al significato statistico degli autovalori della matrice di correlazione, che sono le *varianze* delle *k nuove variabili*, cioè delle coordinate *riferite ad un sistema di assi principali.*

Quando un autovalore è grande, grande è l'allontanamento dei dati dal piano ortogonale all'asse corrispondente, e dunque la distribuzione dei punti X_n segue la direzione di questo asse.

6. Analogie e differenze.

Accenniamo ora alla differenza tra l'analisi dei componenti principali e l'analisi dei fattori. L'analisi dei componenti principali consiste essenzialmente nella determinazione degli autovettori della matrice di correlazione. Tali autovettori, che, lo ricordiamo, sono infiniti, per comodità vengono individuati da k particolari autovettori, ortogonali a due a due e di modulo 1, che indicheremo con $e'_1, \ldots, e'_k$.

Nell'analisi dei fattori, invece, soltanto alcuni dei k particolari autovettori sopra menzionati vengono presi in considerazione, in quanto una delle finalità di questa tecnica è quella di *ridurre* il numero delle variabili considerate, sfruttando l'esistenza di *relazioni* tra di esse.

Inoltre, ciascuno degli autovettori che si è scelto di considerare viene moltiplicato per la *radice quadrata* dell'autovalore corrispondente, e cioè per la deviazione standard della *nuova variabile* che gli corrisponde.

Il vettore così ottenuto, che ha per modulo tale deviazione standard, viene detto *fattore*, e la sua lunghezza viene usata come unità di misura lungo l'asse individuato dal fattore stesso. Ne segue che le variabili correnti lungo tali assi hanno varianza uguale a 1.

Con lo stesso termine di *fattori*, o meglio, di *punteggi fattoriali*, ci si riferisce anche alle *coordinate* dei punti dati, espresse rispetto a tali vettori. Su questo argomento ritorneremo nel prossimo paragrafo.

Concludiamo con l'osservare che, se da un lato è chiaro che i primi autovettori ad essere presi in considerazione sono quelli i cui *autovalori* risultano più grandi, d'altro lato la scelta del *numero* dei fattori da considerare resta a carico dell'analista, come pure il problema di *dare un significato* ai fattori estratti.

7. Punteggi *(scores)*.

Sappiamo che ciascuna *unità statistica* (cioè ciascun campione, nell'esempio) è rappresentata da un punto nel diagramma a dispersione. Le coordinate di tali punti si trovano nella matrice dei dati grezzi, ed hanno un significato legato alla natura del problema sotto osservazione.

Una volta scelti i *fattori*, è possibile esprimere rispetto ad essi le coordinate dei punti. Ad ogni punto restano associate tante coordinate quanti sono i fattori estratti. Tali coordinate vengono dette *punteggi fattoriali (factor scores)*.

Poiché, di solito, il numero m dei fattori è inferiore al numero k delle variabili, occorrono altre $k - m$ coordinate per individuare ciascun punto. Le coordinate mancanti vengono trascurate. In effetti, uno degli scopi dell'analisi dei fattori è proprio quello di *ridurre* il numero delle variabili considerate.

8. Coefficienti fattoriali *(factor score coefficients).*

Ciascuno degli assi cartesiani di un diagramma a dispersione individua un *versore*, cioè un vettore di modulo 1, orientato e diretto come l'asse stesso. Il versore del primo asse si indica con e_1 ed ha coordinate $(1, 0, \ldots, 0)$, quello dell'ultimo asse si indica con e_k, se k è il numero degli assi, ed ha coordinate $(0, \ldots, 0, 1)$.

Solitamente, le coordinate di un punto sono le componenti del suo vettore di posizione rispetto ai versori degli assi. Tuttavia, dopo avere scelto i *fattori* di una distribuzione empirica multivariata, è possibile, con una inversione dei ruoli, trovare le componenti dei versori $e_1, \ldots, e_k$ rispetto ai fattori.

Le componenti dei versori $e_1, \ldots, e_k$ rispetto ai fattori vengono dette *coefficienti fattoriali (factor score coefficients).*

9. Carichi *(loadings).*

Le coordinate dei fattori rispetto alle variabili iniziali vengono dette *pesi fattoriali*, o anche *carichi fattoriali (factor loadings).* Supponendo che le variabili iniziali siano *standardizzate*, i carichi hanno un'importante significato: ciascun *carico* di un dato fattore è il *coefficiente di correlazione* tra quel fattore e la variabile iniziale corrispondente a quel carico.

Di conseguenza, il *quadrato* di ciascun carico, moltiplicato per 100, esprime *la percentuale della varianza della variabile iniziale* che viene *spiegata* da quel fattore in un modello di regressione lineare a due variabili.

Le tecniche di *rotazione* dei fattori (denominate varimax, quartimax, equamax, ecc.) sono spesso tese ad avvicinare ulteriormente ad 1 i *carichi* già abbastanza grandi, avvicinare ulteriormente a -1 i carichi già abbastanza vicini a -1, e ad avvicinare ulteriormente a zero i carichi piccoli in valore assoluto. Tutto ciò per *caratterizzare* meglio i fattori stessi.

10. Comunanze *(communalities).*

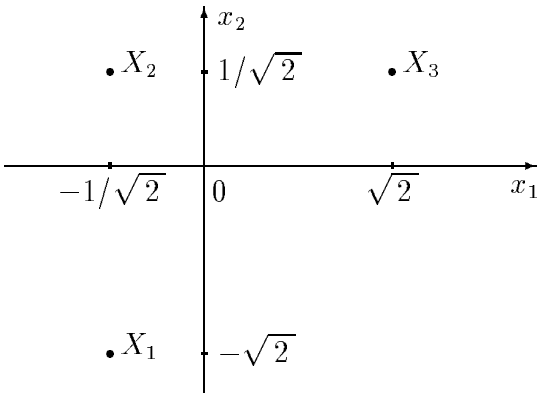
La somma dei quadrati dei carichi di tutti i fattori rispetto ad una data variabile si chiama *comunanza* di quella variabile. Essa si suole interpretare come un indice della attitudine dei fattori a *spiegare* quella variabile.

Per capire meglio questo concetto, consideriamo il caso in cui si utilizzino *tutti* i fattori, che sono tanti quante sono le variabili. In questo caso, la *comunanza* di ciascuna variabile uguaglia la *varianza* della variabile stessa, che a sua volta vale 1 se la variabile in questione è standardizzata.

Quando, come accade in pratica, il numero dei fattori prescelti è *inferiore* al numero delle variabili, ciascuna variabile si considera tanto meglio *spiegata* quanto maggiore è la sua *comunanza*, cioè quanto più la comunanza si avvicina alla *varianza* della variabile stessa.

11. Un esempio numerico.

I programmi per calcolatore reperibili in commercio possono effettuare in pochi istanti il calcolo di una matrice di correlazione anche quando i dati da elaborare sono molto numerosi. Il seguente esempio, invece, è basato su pochi dati, privi di significato concreto e scelti in modo tale da semplificare i risultati. Esso è rivolto a coloro che volessero seguire più da vicino le fasi del calcolo.



Consideriamo tre punti X_1, X_2, X_3, ciascuno dei quali individuato da due coordinate. Abbiamo dunque $N = 3$ e $k = 2$. Le coordinate di tali punti siano le seguenti: $X_1 = (-1/\sqrt{2}, -\sqrt{2})$, $X_2 = (-1/\sqrt{2}, 1/\sqrt{2})$, $X_3 = (\sqrt{2}, 1/\sqrt{2})$. In altri termini, la matrice dei *dati grezzi* è:

$$(x_{nj}) = \begin{pmatrix} -1/\sqrt{2} & -\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \\ \sqrt{2} & 1/\sqrt{2} \end{pmatrix}.$$

Si trova che il valor medio e la varianza di ciascuna variabile valgono, rispettivamente, 0 e 1, cioè i dati sono già *standardizzati*.

11.1. Calcolo della matrice di correlazione.

Con riferimento ai dati precedenti, ed applicando la definizione (1) oppure la formula (4), calcoliamo ρ_{12}. Abbiamo: $x_{11}\,x_{12} = 1$, $x_{21}\,x_{22} = -1/2$, $x_{31}\,x_{32} = 1$, quindi $\rho_{12} = 1/2$. Volendo, si può calcolare analogamente ρ_{21} per verificare che, come osservato nel paragrafo 3, si ha $\rho_{12} = \rho_{21}$. Quanto a ρ_{11} e ρ_{22} si trova che sono entrambi uguali ad 1. Perciò abbiamo:

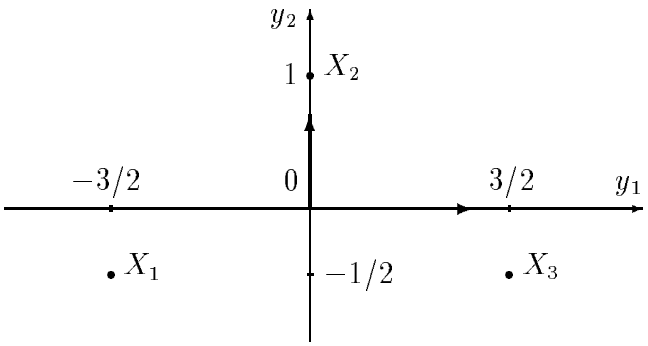
$$(\rho_{ij}) = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \tag{6}$$

La circostanza che $\rho_{11} = \rho_{22} = 1$ è un fatto del tutto generale: gli elementi della diagonale principale di una matrice di correlazione sono uguali ad 1, come si verifica facilmente considerando $i = j$ nella (1).

11.2. Riduzione agli assi principali. Punteggi.

Nel paragrafo precedente abbiamo calcolato la matrice di correlazione di una semplicissima distribuzione bivariata. Come applicazione della teoria sin qui svolta, è naturale chiedersi quali siano le direzioni principali di tale distribuzione. Ebbene, in quel particolare caso, esse sono le bisettrici dei quattro quadranti.

Per verificarlo, dobbiamo vedere se, esprimendo le coordinate dei punti rispetto a tali assi, la nuova matrice di correlazione è diagonale. La figura seguente riporta la posizione dei punti X_1, X_2 e X_3 riferiti ai nuovi assi. Rispetto a tali assi si ha: $X_1 = (-3/2,\, -1/2)$, $X_2 = (0,1)$, $X_3 = (3/2,\, -1/2)$.



Attenzione: ruotando gli assi la standardizzazione si perde. I valori medi restano uguali a zero, ma le varianze possono risultare diverse da 1. Infatti nel nostro esempio si ha $\sigma_1^2 = (9/4 + 9/4)/3 = 3/2$, $\sigma_2^2 = (1/4 + 1 + 1/4)/3 = 1/2$.

Perciò, per calcolare la matrice di correlazione, bisogna usare la formula (1). Occorre anche tener presente che i nuovi assi sono stati chiamati y_1 e y_2 per distinguerli dai vecchi. Si ha $\overline{y}_1 = \overline{y}_2 = 0$, $y_{11}\, y_{12} = 3/4$, $y_{21}\, y_{22} = 0$, $y_{31}\, y_{32} = -3/4$, quindi si trova $\rho_{12} = 0$, e quindi la nuova matrice di correlazione è diagonale, come volevasi dimostrare.

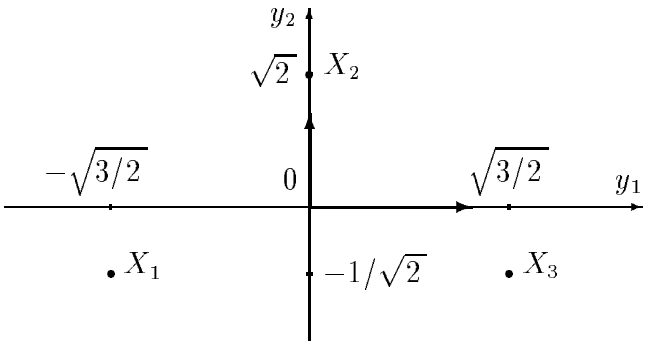
Avendo verificato che $\rho_{12} = 0$ possiamo concludere che i nuovi assi sono *direzioni principali* della distribuzione considerata. Ma qual è il primo fattore? Qual è, cioè, l'asse *più importante*? Intuitivamente è l'asse y_1, ma come si giustifica matematicamente questa intuizione?

Come già detto in precedenza (paragrafo 5), per misurare l'*importanza* di un asse si va a vedere *la varianza* ad esso associata. Nel nostro caso σ_1^2 è maggiore di σ_2^2, per questo motivo il *primo fattore* ha la direzione dell'asse y_1.

In base alla definizione data nel paragrafo 6, costruiamo i fattori moltiplicando i versori dei nuovi assi per le deviazioni standard delle nuove variabili. Così facendo, troviamo che il primo fattore ha componenti $(\sqrt{3/2}\,,\, 0)$ ed il secondo $(0,\, 1/\sqrt{2}\,)$. Essi sono rappresentati come frecce nere nella figura sopra.

Avendo effettuato, in questo caso particolare, una rotazione di 45 gradi, si capisce che le componenti dei versori e_1, e_2 degli assi x_1, x_2 (non rappresentati in figura) sono: $e_1 = (1/\sqrt{2}\,, -1/\sqrt{2})$, $e_2 = (1/\sqrt{2}\,, 1/\sqrt{2})$.

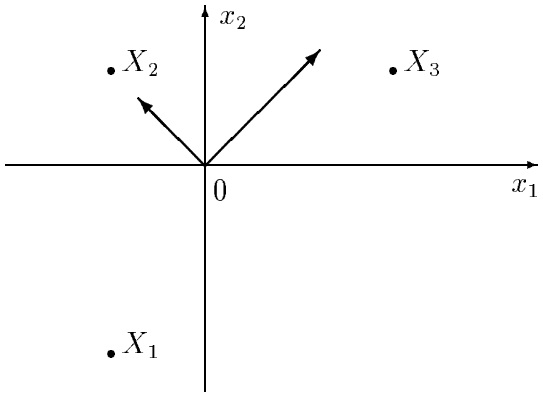
Se, infine, cambiamo scala lungo i nuovi assi, ed utilizziamo come unità di misura il modulo dei fattori, come descritto nel paragrafo 6, allora anche le nuove variabili y_1, y_2 risultano standardizzate e le coordinate dei tre punti X_1, X_2, X_3 diventano: $X_1 = (-\sqrt{3/2}\,, -1/\sqrt{2})$, $X_2 = (0, \sqrt{2})$, $X_3 = (\sqrt{3/2}\,, -1/\sqrt{2})$ (vedi figura sotto). Tali coordinate sono i *punteggi fattoriali* di cui si parla nel paragrafo 7.



Dopo il suddetto cambiamento di scala, le componenti dei versori degli assi x_1, x_2 diventano: $e_1 = (1/\sqrt{3}, -1)$, $e_2 = (1/\sqrt{3}, 1)$. Tali componenti sono i *coefficienti fattoriali* di cui al paragrafo 8.

11.3. Determinazione dei carichi fattoriali.

All'inizio del paragrafo 11 abbiamo introdotto, a titolo di esempio, una semplice distribuzione bivariata e ne abbiamo dato una rappresentazione grafica. Nel paragrafo 11.2, invece, abbiamo rappresentato la stessa distribuzione *rispetto agli assi principali*, evidenziando i due *fattori*. Ora che conosciamo i fattori della distribuzione, vogliamo vederli rappresentati rispetto al sistema di riferimento iniziale $x_1\, x_2$, che avevamo prima della rotazione degli assi.



A tal fine, basta effettuare la rotazione inversa. In tal modo troviamo che il primo fattore ha componenti $(\sqrt{3}/2, \sqrt{3}/2)$, ed il secondo $(-1/2, 1/2)$. Tali componenti sono i *carichi fattoriali*. Possiamo anche verificare che i fattori sono *autovettori* della matrice di correlazione (6), e che gli *autovalori* corrispondenti sono 3/2 e 1/2. Ad

esempio, sostituendo nella (5) i valori $X = (\sqrt{3}/2, \sqrt{3}/2)$, $\lambda = 3/2$, ed al posto di A la matrice (6), si trova che l'uguaglianza è soddisfatta. Alla stessa conclusione si perviene sostituendo $X = (-1/2, 1/2)$ e $\lambda = 1/2$.

Bibliografia.

[1] J. C. Davis, Statistics and data analysis in Geology, Wiley, New York 1973.

[2] L. Fabbris, Statistica multivariata, McGraw-Hill, Milano 1997.

[3] P. Machek, B. Testa, Dispense del corso di Geostatistica, CNR, Centro di studio per la stratigrafia e petrografia delle alpi centrali, Milano 1991.

[4] M. J. Norušis/SPSS Inc., Manuale di istruzioni di *SPSS for Windows: Professional Statistics*, Release 6.0, Chicago 1993.

[5] C. Spearman, *General intelligence, objectively determined and measured,* American Journal of Psychology **15**, 201–293.

[6] M. I. Stoka, Lezioni di algebra lineare e geometria proiettiva, Edizioni C.E.L.U.P., Palermo 1982.