Alcune variabili casuali continue

La variabile casuale gamma

La v.c. Y è distribuita secondo una gamma di parametri $\alpha > 0$ e $\vartheta > 0$ se la funzione di densità è data da:

$$f(y) = \begin{cases} \dfrac{\vartheta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\vartheta y} & y > 0 \\ 0 & altrove \end{cases},$$

dove $\Gamma(\alpha) \equiv \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$.

Proprietà della funzione gamma:

- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$,
- $\Gamma(n) = (n-1)!$.

$$E(Y^r) = \int_0^{+\infty} y^r \frac{\vartheta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\vartheta y} dy =$$

$$= \frac{\vartheta^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(r+\alpha)}{\vartheta^{r+\alpha}} \int_0^{+\infty} \frac{\vartheta^{r+\alpha}}{\Gamma(r+\alpha)} y^{r+\alpha-1} e^{-\vartheta y} dy =$$

$$= \frac{1}{\vartheta^r} \frac{(r+\alpha-1)\cdot(r+\alpha-2)\cdot\ldots\cdot\alpha\cdot\Gamma(\alpha)}{\Gamma(\alpha)} =$$

$$= \frac{1}{\vartheta^r}(r+\alpha-1)\cdot(r+\alpha-2)\cdot\ldots\cdot\alpha.$$

Per $r = 1$, si ottiene $E(X)=\frac{\alpha}{\vartheta}$. Per $r = 2$, si ottiene $E(X^2)=\frac{\alpha\cdot(\alpha+1)}{\vartheta^2}$, da cui:

$$Var(X)=\frac{\alpha\cdot(\alpha+1)}{\vartheta^2}-\frac{\alpha^2}{\vartheta^2}=\frac{\alpha}{\vartheta^2}.$$

$$m_X(t)=\int_0^{+\infty} e^{ty}\,\frac{\vartheta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\vartheta y}dy=$$

$$=\frac{\vartheta^\alpha}{(\vartheta-t)^\alpha}\int_0^{+\infty}\frac{(\vartheta-t)^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-y(\vartheta-t)}dy=\left(\frac{\vartheta}{\vartheta-t}\right)^\alpha,$$

per $t<\vartheta$.

Casi particolari:

a) $\alpha=1$, v.c. esponenziale di parametro ϑ

b) $\alpha=\frac{k}{2},\vartheta=\frac{1}{2}$ (k intero), v.c. chi-quadrato con k gradi di libertà (g.d.l.).

(esplicitare i calcoli, e disegnare le densità; dare la definizione di quantile, e mostrare l'uso delle tavole.)

La v.c. normale

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad x\in R,\ \mu\in R,\ \sigma>0$$

Teorema

Sia X_1 una v.c. normale di aspettativa μ_1 e varianza σ_1^2. Sia X_2 una v.c. una v.c. normale di aspettativa μ_2 e varianza σ_2^2. Siano X_1 e X_2 indipendenti in probabilità. Allora la v.c. $X=X_1+X_2$ è una v.c. normale di aspettativa $\mu_1+\mu_2$ e varianza $\sigma_1^2+\sigma_2^2$.

Teorema

Sia Z una v.c. normale standardizzata. Allora la v.c. $Y=Z^2$ è una v.c. chi-quadrato con 1 g.d.l..

Dimostrazione

$$F_Y(y)=P\{Y\le y\}=P\left\{Z^2\le y\right\}=P\left\{-\sqrt{y}\le Z\le+\sqrt{y}\right\}=$$

$$=\int_{-\sqrt{y}}^{+\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=2\int_{0}^{+\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=$$

$$= 2\int_0^y \frac{1}{\sqrt{2}}\frac{1}{\sqrt{\pi}}\frac{1}{2\sqrt{v}}e^{-\frac{v}{2}}dv = \int_0^y \left(\frac{1}{2}\right)^{\frac{1}{2}}\frac{1}{\Gamma\left(\frac{1}{2}\right)}v^{\left(1-\frac{1}{2}\right)}e^{-\frac{v}{2}}dv,$$

che è la funzione cumulata della probabilità di una v.c. chi-quadrato con 1 g.d.l..

Teorema

Siano $(X_1, X_2, \cdots, X_n)$ n variabili casuali normali di aspettativa μ e varianza σ^2 e indipendenti in probabilità.

Allora:

a) La v.c. $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (media campionaria) si distribuisce come una normale di aspettativa μ e varianza $\frac{\sigma^2}{n}$.

b) Le variabili casuali $\overline{X}$ e $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ sono indipendenti

c) La v.c. $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2}$ (varianza campionaria corretta) si distribuisce come una chi-quadrato con $(n-1)$ g.d.l..

La distribuzione t di "Student"

La v.c. *X* ha distribuzione *t di Student*, con $k > 0$ g.d.l., se la sua densità è la seguente:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \cdot \frac{1}{\sqrt{k\pi}} \cdot \frac{1}{\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}}, \quad x \in R$$

(Disegnare il grafico)
Si può dimostrare che:

$$E(X) = 0 \quad (k > 1); \qquad Var(X) = \frac{k}{k-2} \quad (k > 2)$$

Inoltre μ_r esiste per $r < k$.

Teorema

Sia Z una v.c. distribuita come una normale standardizzata. Sia U una v.c. distribuita come una chi-quadrato con k g.d.l.. Siano Z e U indipendenti in probabilità. Allora la v.c. $X = \frac{Z}{\sqrt{U/k}}$ si distribuisce secondo una *t di Student* con k g.d.l..

Corollario

Sia $(X_1, X_2, \cdots, X_n)$ un campione casuale proveniente dalla v.c. normale di aspettativa μ e varianza σ^2. Allora la v.c. $T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}}$ ha distribuzione *t di Student* con $(n-1)$ g.d.l..

Dimostrazione

Si considerino le variabili casuali $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ e $U = \frac{(n-1)S^2}{\sigma^2}$. Poiché esse sono

indipendenti, si può applicare il teorema precedente, ottenendo la v.c.:

$$T = \frac{\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2} \Big/ (n-1)}} = \frac{\overline{X} - \mu}{\sqrt{S^2 / n}},$$

la cui distribuzione è *t di Student* con $(n-1)$ g.d.l..

La distribuzione F di Fisher

La v.c. *X* ha distribuzione *F di Fisher* con *m* e *n* gradi di libertà se ha la seguente funzione di densità:

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)} \left(\dfrac{m}{n}\right)^{\frac{m-2}{2}} \dfrac{x^{\frac{m-2}{2}}}{\left(1+\dfrac{m}{n}x\right)^{\frac{m+n}{2}}} & x > 0 \\ 0 & \textit{altrove} \end{cases}$$

m rappresenta i gradi di libertà del numeratore; *n* rappresenta i gradi di libertà del denominatore.

Si può dimostrare che:

- $E(X)=\dfrac{n}{n-2} \quad (n>2)$
- $Var(X)=\dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (n>4)$
- $\mu_r = E\left(X^r\right)$ esiste solo per $r<\dfrac{n}{2}$.

Teorema

Siano U e V due v.c. chi-quadrato indipendenti con m e n g.d.l. rispettivamente. Allora la v.c.

$$X=\frac{U/m}{V/n}$$

è distribuita come una v.c. F con m e n g.d.l..

Corollario 1

Sia $(X_1, X_2,\cdots,X_m)$ un campione casuale proveniente dalla v.c. normale di aspettativa μ_X e varianza σ^2. Sia $(Y_1, Y_2,\cdots,Y_n)$ un campione casuale proveniente dalla v.c. normale di aspettativa μ_Y e varianza σ^2.

Siano i due campioni indipendenti, cioè estratti da popolazioni differenti. Allora, la v.c.

$$X = \frac{\dfrac{\sum_{i=1}^{m}(X_i - \bar{X})^2}{\sigma^2} \Big/ (m-1)}{\dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2} \Big/ (n-1)} = \frac{S_X^2}{S_Y^2}$$

ha distribuzione F con $(m-1)$ e $(n-1)$ g.d.l.

Corollario 2

Se X ha distribuzione F con m e n g.d.l., allora $Y = \frac{1}{X}$ ha distribuzione F con n e m g.d.l..

Tale corollario è utile per la consultazione della tavole della distribuzione.

Si supponga di voler determinare il quantile y_α. Esso è definito come $P\{Y \le y_\alpha\} = \alpha$.

Vale la seguente relazione:

$P\{Y \le y_\alpha\} = P\left\{X = \frac{1}{Y} \ge \frac{1}{y_\alpha}\right\} = \alpha$, da cui

$$P\left\{X \le \frac{1}{y_\alpha}\right\} = 1-\alpha \;\Rightarrow \frac{1}{y_\alpha} = x_{(1-\alpha)} \;\Rightarrow y_\alpha = \frac{1}{x_{(1-\alpha)}}.$$

Le verifiche d'ipotesi

Si considerino n variabili casuali $(X_1,\cdots,X_n)$ indipendenti e aventi la medesima funzione di densità $f(x,\vartheta)$ - variabile casuale continua –, oppure la medesima funzione di probabilità $p(x,\vartheta)$ - variabile casuale discreta. Allora, si dirà che $(X_1,\cdots,X_n)$ è un campione casuale proveniente dalla funzione di densità $f(x,\vartheta)$, oppure dalla funzione di probabilità $p(x,\vartheta)$.

Per ipotesi statistica s'intende una congettura sulla forma della distribuzione dalla quale provengono i dati.

Esempio 1

1. l'aspettativa è pari a 3
2. la varianza è maggiore di 5
3. la distribuzione è simmetrica
4. la distribuzione è normale (gamma, Poisson, ecc.)

Nella teoria classica delle verifiche d'ipotesi, all'ipotesi statistica oggetto di verifica, detta ipotesi nulla (H_0), se ne deve affiancare una contraria, scelta del decisore, detta ipotesi alternativa (H_1).

Si dice test statistico una partizione dello spazio dei possibili risultati campionari in due sottoinsiemi disgiunti:

- la regione critica C, ovvero l'insieme dei risultati campionari per cui il test prescrive di rifiutare l'ipotesi nulla
- la regione di accettazione $\overline{C}$, ovvero l'insieme dei risultati campionari per cui il test prescrive di accettare l'ipotesi nulla (l'insieme complementare, o negazione, di C).

Si possono compiere due tipologie d'errore statistico:

- l'errore di prima specie, che consiste nel rifiutare l'ipotesi nulla quando essa è vera
- l'errore di seconda specie, che consiste nell'accettare l'ipotesi nulla quando essa è falsa.

$D \mid V$	H_0	H_1
H_0	-	II specie
H_1	I specie	-

D è l'ipotesi scelta, V è l'ipotesi vera.

Data la natura statistica dell'esperimento, tali errori si possono commettere con una certa probabilità:

- $\alpha \equiv \Pr(\textit{errore di I specie})$
- $\beta \equiv \Pr(\textit{errore di II specie})$.

Un test ideale sarebbe quello che renda minimi contemporaneamente le due probabilità d'errore. Sfortunatamente, le due probabilità d'errore hanno andamenti contrapposti, come mostra il seguente

Esempio 2
Si voglia verificare l'ipotesi nulla che una moneta sia regolare, sulla base di $n = 5$ lanci. Pertanto, si ha un campione casuale da una variabile casuale indicatore di parametro *p* (probabilità di ottenere testa).
Sappiamo che la variabile casuale "numero di teste in 5 lanci" *X* ha distribuzione $bin(n = 5, p)$.
Se la moneta è regolare, *p* sarà pari a ½.
Occorre, come si diceva, definire un'ipotesi alternativa, supponiamo, *p* pari a 0,8.
Questo problema di verifica d'ipotesi può formalizzarsi così:

$$H_0 : p = 0{,}5$$
$$H_1 : p = 0{,}8 \,.$$

Si consideri il test avente la seguente regione critica:

$$C = \{ X \geq 4 \},$$

in parole, il test prescrive di rifiutare l'ipotesi nulla se su 5 lanci almeno in 4 esce testa.

Per tale test, calcoliamo le probabilità d'errore statistico:

- $\alpha = P\{X \geq 4 | p = 0{,}5\} = \sum_{x=4}^{5} \binom{5}{x} \cdot 0{,}5^{x} \cdot 0{,}5^{5-x} =$

$$= \binom{5}{4} \cdot 0{,}5^{5} + \binom{5}{5} \cdot 0{,}5^{5} = 6 \cdot 0{,}5^{5} = 0{,}1875$$

- $\beta = P\{X < 4 | p = 0{,}8\} = 1 - P\{X \geq 4 | p = 0{,}8\} =$

$$= 1 - \sum_{x=4}^{5} \binom{5}{x} \cdot 0{,}8^{x} \cdot 0{,}2^{5-x} = 1 - \binom{5}{4} \cdot 0{,}8^{4} \cdot 0{,}2^{1} - \binom{5}{5} \cdot 0{,}8^{5} =$$
$$= 1 - 0{,}4096 - 0{,}32768 = 0{,}26272$$

Consideriamo ora un secondo test, avente per regione critica:

$$C_1 = \{ X \geq 5 \},$$

e calcoliamone i due tipi di errore:

- $\alpha_1 = P\{X \geq 5 | p = 0{,}5\} = \binom{5}{5} \cdot 0{,}5^5 \cdot 0{,}5^{5-5} = 0{,}03125$

- $\beta_1 = P\{X < 5 | p = 0{,}8\} = 1 - P\{X = 5 | p = 0{,}8\} =$

$= 1 - \binom{5}{5} \cdot 0{,}8^5 \cdot 0{,}2^{5-5} = 0{,}67232.$

Si nota che $\alpha > \alpha_1$ e $\beta < \beta_1$, cioè che al crescere di α, β diminuisce.

Non essendo possibile minimizzare i due tipi d'errore, la teoria classica delle verifiche d'ipotesi, fissa la probabilità dell'errore di prima specie α, e cerca il test che, a parità di α, rende minima la probabilità d'errore di seconda specie β (test più potente).

Un'ipotesi statistica si dice semplice, se specifica completamente la distribuzione da cui provengono i dati. Altrimenti, essa si dice composta.

Le ipotesi dell'*esempio 2* sono entrambe semplici. Nell'*esempio 1*:

- l'ipotesi 1. è semplice se si campiona dalla distribuzione normale con varianza nota, composta se la varianza non è nota
- l'ipotesi 2. è composta.

Il lemma di Neyman-Pearson fornisce la regione critica del test più potente, qualora si considerino due ipotesi semplici. Generalizzando opportunamente la teoria, è possibile ricavare regioni critiche di test che possiedono buone proprietà statistiche.

Per il campionamento da distribuzione normale, ecco le regioni critiche dei test sui parametri, per vari problemi di verifica d'ipotesi:

Verifiche d'ipotesi su μ

- Caso 1 σ^2 nota.

$H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$

Si rifiuta H_0 se $\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > z_{1-\alpha}$, o, equivalentemente, se $\bar{X} > \mu_0 + z_{1-\alpha} \dfrac{\sigma}{\sqrt{n}}$.

$H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$

Si rifiuta H_0 se $\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha}$, o, equivalentemente, se $\overline{X} < \mu_0 - z_{1-\alpha}\dfrac{\sigma}{\sqrt{n}}$.

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

Si rifiuta H_0 se $\dfrac{|\overline{X}-\mu_0|}{\sigma/\sqrt{n}} > z_{1-\frac{\alpha}{2}}$, o, equivalentemente, se

$$\overline{X} \notin \left[\mu_0 - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \mu_0 + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right].$$

• Caso 2 σ^2 non nota.

$H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$

Si rifiuta H_0 se

$$\frac{\overline{X}-\mu_0}{S/\sqrt{n}} > t_{1-\alpha}(n-1),$$

o, equivalentemente, se

$$\bar{X} > \mu_0 + t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}.$$

$t_{1-\alpha}(n-1)$ indica il quantile di ordine $(1-\alpha)$ della distribuzione t con $(n-1)$ g.d.l..

$H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$

Si rifiuta H_0 se $\frac{\bar{X} - \mu_0}{t/\sqrt{n}} < -t_{1-\alpha}(n-1)$, o, equivalentemente, se $\bar{X} < \mu_0 - t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}$.

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

Si rifiuta H_0 se

$$\frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} > t_{1-\frac{\alpha}{2}}(n-1),$$

o, equivalentemente, se

$$\bar{X} \notin \left[\mu_0 - t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}; \mu_0 + t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}} \right].$$

Verifiche d'ipotesi su σ^2.

- Caso 1 μ nota

$$H_0 : \sigma^2 \leq \sigma_0^2$$

$$H_1 : \sigma^2 > \sigma_0^2$$

Si rifiuta H_0 se vale

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma_0^2} > \chi^2_{1-\alpha}(n),$$

dove $\chi^2_{1-\alpha}(n)$ indica il quantile di ordine $(1-\alpha)$ della distribuzione chi-quadrato con *n* g.d.l..

$$H_0 : \sigma^2 \geq \sigma_0^2$$

$$H_1 : \sigma^2 < \sigma_0^2$$

Si rifiuta H_0 se vale

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma_0^2} < \chi^2_{\alpha}(n).$$

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

Si rifiuta H_0 se vale

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} \notin \left[\chi^2_{\frac{\alpha}{2}}(n);\chi^2_{1-\frac{\alpha}{2}}(n)\right].$$

- Caso 2 μ non nota

$$H_0 : \sigma^2 \le \sigma_0^2$$

$$H_1 : \sigma^2 > \sigma_0^2$$

Si rifiuta H_0 se vale

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma_0^2} > \chi^2_{1-\alpha}(n-1),$$

dove $\chi^2_{1-\alpha}(n-1)$ indica il quantile di ordine $(1-\alpha)$ della distribuzione chi-quadrato con $(n-1)$ g.d.l..

$$H_0 : \sigma^2 \ge \sigma_0^2$$

$$H_1 : \sigma^2 < \sigma_0^2$$

Si rifiuta H_0 se vale

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma_0^2} < \chi^2_{\alpha}(n-1).$$

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

Si rifiuta H_0 se vale

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2} \notin \left[\chi^2_{\frac{\alpha}{2}}(n-1); \chi^2_{1-\frac{\alpha}{2}}(n-1) \right].$$

Definizione

Dicesi *p-value* il valore minimo (massimo) della probabilità dell'errore di prima specie per il quale si rifiuta (si accetta) l'ipotesi nulla H_0.

(*Segue lezione su regressione lineare multipla, file in e-bacheca)*

L'analisi delle componenti principali

Si tratta di una tecnica di riduzione dei dati molto diffusa e applicata, affidata a trasformazioni lineari dei dati in un opportuno sottospazio delle variabili. Le variabili trasformate (le prime poche componenti principali) sono utili in connessione con l'investigazione preliminare di un gran numero di campioni. Per un'idea orientativa, sarebbe sufficiente la sola prima componente principale (di quattro) perché solitamente rappresenta circa i ¾ (75%) della varianza. Potrebbe sembrare ai non matematici che la tecnica proposta sia altamente arbitraria, ma tale dubbio scompare allorchè si consideri l'interpretazione geometrica: l'analisi delle componenti principali semplicemente porta a nuovi angoli di visione dei dati, angoli meglio adatti a mostrare l' ordine di grandezza e la variazione di forma. Come vedremo, ai fini di estrarre informazioni da una considerevole mole di dati, dal punto di vista statistico, è fondamentale saper interpretare i risultati, in termini di interazioni e legami concettuali fra le variabili.

Sia $\underset{n\times p}{X} = \{x_{hj}\}$ la matrice dei dati associata a *n* unità statistiche su ognuna delle quali siano state osservate *p* variabili (senza perdere di generalità esse si possono assumere a media nulla).

Sia $\underset{p\times p}{S} = \{s_{ij}\}$ la matrice di varianze e covarianze. Se le variabili sono a media nulla essa avrà forma $S = X'X/(n-1)$.

L'obiettivo è quello di trovare *r* trasformazioni lineari delle variabili osservate

$$f_i = \sum_{j=1}^{p} w_{ji} x_j \qquad i = 1, \cdots, r$$

ove *r* è il rango della matrice.

E' bene che fra le variabili osservate in *X* sussista una relazione lineare. Qualora esse siano legate da relazioni non lineari, sarà bene linearizzare.

Se $f = Xw$, si tratta di massimizzare la funzione $f'f/(n-1) = w'Sw$,

soggetta al vincolo di normalizzazione

$w'w = 1$

A tal fine si predispone la funzione lagrangiana:

$l = w'Sw - \Omega(w'w - 1)$

Derivando rispetto a *w'* e ad Ω si ottiene il sistema

$$2Sw - 2\Omega w = 2(S - \Omega I)w = 0 \qquad (1)$$

$$w'w - 1 = 0 \qquad (2)$$

La (1) soggetta al vincolo (2) ha soluzione non nulla se il suo determinante è nullo.

$$|S - \Omega I| = 0$$

Tale equazione equivale ad eguagliare a zero un polinomio ordinato di grado *p* nella variabile Ω.

Poiché la matrice $S - \Omega I$ è simmetrica, si otterranno *p* radici reali dette autovalori associati a *S*:

$$\Omega_1 \geq \Omega_2 \geq \cdots \geq \Omega_p \geq 0 \qquad (3)$$

Sostituendo la radice più grande nella (1), si ottiene:

$(S - \Omega_1 I)w_1 = 0,$

vettore che porge i pesi della prima componente principale, detto anche autovettore associato a all'autovalore Ω_1.

Per trovare la seconda componente, si calcola la matrice di varianze e covarianze residua, cioè privata della variabilità della prima componente:

$S^* = S - \Omega_1 w_1 w_1'$.

Si massimizza $w_2' S^* w_2,$

soggetta al vincolo di normalizzazione

$w_2' w_2 = 1,$

che pertanto porgerà la soluzione:

$(S - \Omega_2 I)w_2 = 0,$

secondo valore ordinato della (3) e l'autovettore associato, e via di seguito sino alla *p*-esima componente.

E' possibile dimostrare che:

1. Le componenti principali sono a due a due incorrelate, cioè $w_i' w_j = 0$. Ciò si può anche esprimere dicendo che $W_{p \times r}$ la matrice dei primi *r* autovettori è ortogonale:

$W'W = I$ e in modo equivalente $W' = W^{-1}$.

2. L'autovalore Ω_i rappresenta la varianza dell'i-esima componente principale, infatti:
   $w_i' S w_i = w_i'(X'X) w_i /(n-1) = f_i' f_i /(n-1) =$
   $= \Omega_i = Var(f_i)$
   Per la 1. la matrice di varianze e covarianze delle componenti principali è diagonale
   $\Omega = F'F /(n-1) = diag(\Omega_i)$

3. $\sum_{i=1}^{p} \Omega_i = tr(S)$, essendo *tr*(.), la somma degli elementi della diagonale principale, detta traccia della matrice. Se la matrice fattorizzata è una matrice di correlazione, allora *tr*(*R*)=*p*.

4. Il prodotto degli autovalori è il determinante della matrice di partenza: $\prod_{i=1}^{p} \Omega_i = |S|$. Ciò è conseguenza del fatto che $|AB| = |A||B|$ e della proprietà di *W* al punto 1.

5. Scomposizione di Eckart e Young (1936), o scomposizione spettrale:

$$S = W\Omega W' = \sum_{i=1}^{p} \Omega_i w_i w_i'$$

(seguono commenti)

Osservazioni

- Le componenti principali non sono indipendenti dalla scala di misura delle variabili. Se si moltiplica una variabile osservata per un valore costante, la matrice di varianze e covarianze cambia, come anche le componenti generate.
- Le componenti non variano se le variabili hanno uguale varianza, oppure sono standardizzate in senso statistico.
- Non è necessario standardizzare se tutte le variabili sono dello stesso tipo, o sono tutte dicotomiche, oppure punteggi di una stessa scala, oppure percentuali, oppure rapporti tra due grandezze che variano entro intervalli limitati, oppure misure ripetute sulle stesse unità che presentano varianze tendenzialmente costanti (misure con varianze dello stesso ordine di

grandezza si ottengono anche con trasformazioni logaritmiche delle osservazioni)

Scelta del numero di componenti

- Percentuale di varianza cumulata
- Con variabili standardizzate, si considerano solo autovalori maggiori di 1
- Con variabili centrate si considerano solo autovalori superiori alla loro media aritmetica
- Scree plot.

Esempio 1

$$S = \begin{bmatrix} 7 & -5 & 0 \\ -5 & 10 & -\sqrt{6} \\ 0 & -\sqrt{6} & 4 \end{bmatrix}$$

$$|S - \omega I| = -\omega^3 + 21\omega^2 - 107\omega + 138 =$$

$$= (\omega - 2)(-\omega^2 + 19\omega - 69) = 0$$

$$\omega_3 = 2 \quad \omega_{1,2} = \frac{19 \pm \sqrt{85}}{2} = \frac{19 \pm 9{,}21}{2}$$

Verifico che $\sum_{i=1}^{p} \Omega_i = tr(S)$

$$2 + \frac{19 + \sqrt{85}}{2} + \frac{19 - \sqrt{85}}{2} = 10 + 7 + 4 = 21$$

$$(S-\omega_1 I)w_1 = \begin{bmatrix} -7{,}105 & -5 & 0 \\ -5 & -4{,}105 & -\sqrt{6} \\ 0 & -\sqrt{6} & -10.105 \end{bmatrix} \cdot w_1 = 0$$

$w_1' = [0{,}565 \quad -0{,}802 \quad 1{,}237]$, prima componente principale.

ω_i	$\omega_i / tr(S)$	Cum.
14,105	0,6716	0,6716
4,8950	0,2331	0,9047
2,0000	0,0953	1,0000
21	1	

Esempio 2

Facciamo una simulazione. Poniamo di disporre di un'indagine che ci riporta per 10 soggetti: voto medio (da 0 a 33), intelligenza (da 0 a 10), media ore studiate in un giorno e zona d'origine (che varia da 1 a 3). Standardizziamo i valori con la formula:

$z=(Xi-E(X))/SD$

(con "E(x)" che è X medio).

Dopo di che calcoliamo la matrice dei coefficienti di correlazione che sarà:

		Zscore (VotoMedio)	Zscore (Intelligenza)	Zscore (Provenienza)	Zscore (OreMed Studio)
Correlation	Zscore(VotoMedio)	1,000	,600	-,838	,788
	Zscore(Intelligenza)	,600	1,000	-,222	,022
	Zscore(Provenienza)	-,838	-,222	1,000	-,918
	Zscore(OreMedStudio)	,788	,022	-,918	1,000

Chiaramente la diagonale principale è composta da valori uguali ad 1 (il coefficiente di correlazione di una variabile con se stessa deve dare necessariamente questo valore). È pure una matrice simmetrica (il coefficiente di correlazione tra la variabile "x" e la variabile "y" sarà uguale a quello tra "y" e "x"). Vediamo come ci sia un forte legame tra voto, media ore studio e intelligenza.

Studiamo allora gli autovalori (eigenvalues) e quanto spiegano:

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,828	70,708	70,708	2,828	70,708	70,708
2	1,070	26,755	97,463	1,070	26,755	97,463
3	,084	2,088	99,551			
4	,018	,449	100,000			

Abbiamo posto gli autovalori più alti per primi, e, come detto, il loro rapporto con la somma degli autovalori ci dà la varianza che spiegano. Abbiamo selezionato (arbitrariamente) solo quelli che hanno valore maggiore di 1: i più significativi, che ci spiegano il 70,708% e il 26,755% rispettivamente.

Guardiamo ora alla **matrice delle componenti principali**:

	Component	
	1	2
Zscore(VotoMedio)	,966	,204
Zscore(Intelligenza)	,442	,894
Zscore(Provenienza)	-,947	,228
Zscore(OreMedStudio)	,897	-,420

Il fattore 1 (che, facendo una congettura, si potrebbe chiamare bravura) pesa dunque fortemente sul voto medio. Sembrerebbe pure che pesi in maniera negativa sulla variabile della zona di origine (chiaramente questa affermazione non avrebbe senso perché invertiremmo il nesso di causalità, spetta infatti allo statistico saper dare una spiegazione e una lettura sensate).

Dopo di che ottengo la **matrice di punteggio fattoriale**:

	Component	
	1	2
Zscore(VotoMedio)	,341	,191
Zscore(Intelligenza)	,156	,836
Zscore(Provenienza)	-,335	,213
Zscore(OreMedStudio)	,317	-,392

Come vediamo la variabile provenienza continua ad avere un influsso di segno negativo sull'autovalore principale. Le altre variabili invece hanno peso positivo.

Esempio 3 (e-bacheca)

(Segue lezione sul modello lineare generale, regressione multipla e model selection)

L'analisi dei gruppi (cluster analysis)

(i) L'obiettivo delle tecniche raccolte sotto questo nome è l'assegnazione di unità statistiche, sulle quali siano stati rilevati più caratteri, a pochi gruppi (clusters), non definiti a priori. Il principio è che le unità appartenenti al medesimo gruppo devono essere "simili" o "somiglianti", secondo qualche criterio formale (di vicinanza).

(ii) Per svolgere una cluster analysis si devono prendere varie decisioni:

(iii) Identificare le variabili di classificazione. Si consideri la matrice $\underset{n\times p}{X} = \{x_{hj}\}$ la matrice dei dati associata a *n* unità statistiche su ognuna delle quali siano state osservate *p* variabili, misurate su qualsiasi scala.

(iv) Selezione della misura di prossimità tra unità

(v) Selezione della tecnica di raggruppamento delle entità

(vi) Identificazione del numero dei gruppi entro i quali ripartire le entità

(vii) Uso integrato di altri metodi di analisi multivariata.

Tecniche di analisi dei gruppi

Con la definizione di criteri specifici, volti a calcolare una misura del grado di similarità-dissimilarità fra due qualsiasi degli elementi di un dataset, sulla base dei valori delle diverse variabili prescelte, si realizza una sintesi dell'informazione originariamente disponibile, che rende più agevole il processo di classificazione.

Le più importanti misure di somiglianza-diversità sono gli indici di similarità (e di dissimilarità) e di distanza.
Nella letteratura dell'analisi dei dati sono stati stabiliti alcuni requisiti formali secondo cui una funzione, definita su tutte le coppie di oggetti appartenenti ad un insieme dato, può essere definita rispettivamente un indice di similarità, di dissimilarità o di distanza.
Affinché si possa parlare di indice di similarità, è necessario che la funzione:
a) sia non negativa;
b) sia simmetrica: la similarità di un oggetto A rispetto ad un oggetto B deve essere uguale alla similarità dell'oggetto B rispetto ad A;
c) assuma il valore massimo quando si riferisce al rapporto di un oggetto con se stesso.
Analogamente, perché si possa parlare di indice di dissimilarità, è necessario che la funzione:
a) sia non negativa;
b) sia simmetrica: la dissimilarità di un oggetto A rispetto ad un oggetto B deve essere uguale alla dissimilarità dell'oggetto B rispetto ad A;
c) assuma il valore nullo quando si riferisce al rapporto di un oggetto con se stesso.
Perché si possa parlare di indice di distanza, infine, si richiedono proprietà più restrittive: le proprietà formali che vengono attribuite abitualmente ad una "metrica".
a) la distanza di un oggetto da se stesso è nulla;

b) simmetria: la distanza di un oggetto A rispetto ad un oggetto B è uguale alla distanza dell'oggetto B rispetto ad A;
c) disuguaglianza triangolare: la distanza fra A e C è minore, o al più uguale, alla somma delle distanze fra A e B e fra B e C (per ogni A,B,C).
In particolare, la disuguaglianza triangolare esprime l'idea che sia più breve andare da A a C "direttamente", che non passando per un altro punto B non posto sulla linea retta congiungente i primi due.
Infine, va notato che un indice di distanza può ovviamente essere considerato un indice di dissimilarità, ma in generale non vale l'inverso.
Nel caso di caratteri quantitativi, possono essere utilizzati vari tipi di indici di distanza.
Tra questi, i più utilizzati sono:
a) la distanza euclidea: corrisponde al concetto geometrico di distanza nello spazio multidimensionale. Per misurare la distanza tra x_{hv} e x_{kv} si utilizza il teorema di Pitagora:

$$d_{hk} = \left\{ \sum_{v=1}^{p} w_v (x_{hv} - x_{kv})^2 \right\}^{1/2}$$

dove x_{hv} e x_{kv} sono le coordinate dei due punti P_h e P_h nello spazio cartesiano sulla variabile x_v e w_v è il peso attribuito alla variabile.
La distanza euclidea può creare dei problemi, poiché considera sullo stesso piano variabili di natura diversa e espresse in unità di

misura differenti; per ovviare a tale problema, spesso si ricorre alla preventiva standardizzazione della matrice dei dati.
Un ulteriore problema è dato dal fatto che la distanza euclidea non considera eventuali correlazioni tra le variabili quindi, se due variabili sono fortemente correlate e fanno riferimento allo stesso fattore, quest'ultimo è come se pesasse due volte nel calcolo della distanza. Pertanto è utile ponderare le variabili, dando un peso maggiore a quelle meno correlate con le altre.
b) il quadrato della distanza euclidea: questo tipo di distanza viene impiegato qualora si voglia dare un peso progressivamente maggiore agli oggetti che stanno oltre una certa distanza.

$$d_{hk} = \sum_{v=1}^{p} w_v (x_{hv} - x_{kv})^2$$

dove x_{hv} e x_{kv} sono le coordinate dei due punti P_h e P_h nello spazio cartesiano sulla variabile x_v e w_v è il peso attribuito alla variabile.
c) la distanza di Manhattan (o city-block o distanza assoluta): è fornita dalla differenza media fra le dimensioni. La distanza tra due punti, con questa tecnica, viene calcolata come somma delle differenze, in valore assoluto, tra le loro coordinate. Questa tecnica è consigliata nel caso in cui si abbia a trattare con classificazioni su scala ordinale.

$$d_{hk} = \sum_{v=1}^{p} w_v |x_{hv} - x_{kv}|$$

d) la distanza di Chebyshev: può essere appropriata nei casi in cui si vogliano definire due oggetti come "differenti" se essi sono diversi in ciascuna delle dimensioni:

$$d_{hk} = \max|x_{hv} - x_{kv}|$$

e) la distanza di Minkowski: è una generalizzazione delle precedenti, infatti è data

dalla formula:

$$d_{hk} = \left\{ \sum_{v=1}^{p} w_v |x_{hv} - x_{kv}|^s \right\}^{1/s}$$

nel casi particolari di s=1 e s=2 si ottengono rispettivamente la distanza di Manhattan e la distanza euclidea; mentre, nel caso limite in cui s tenda a infinito si ottiene la distanza di Chebyshev.

I metodi di aggregazione

Esistono varie classificazioni delle tecniche di clustering comunemente utilizzate.

Una prima categorizzazione dipende dalla possibilità che un elemento possa o meno essere assegnato a più clusters:

a) clustering esclusivo: ogni elemento può essere assegnato ad uno e ad un solo gruppo. I clusters risultanti, quindi, non possono avere elementi in comune.

Questo approccio è detto anche Hard Clustering.

b) clustering non-esclusivo: in questo tipo di approccio un elemento può appartenere a più clusters con gradi di appartenenza diversi. Questo approccio è

noto anche con il nome di Soft Clustering o Fuzzy Clustering.

Un'altra suddivisione delle tecniche di clustering tiene conto della tipologia dell'algoritmo utilizzato per dividere lo spazio:

a) clustering partitivo (detto anche k-clustering): per definire l'appartenenza ad un gruppo viene utilizzata una distanza da un punto rappresentativo del cluster (centroide, medioide), avendo prefissato il numero di gruppi della partizione risultato.

b) clustering gerarchico: viene costruita una gerarchia di partizioni caratterizzate da un numero (de)crescente di gruppi, visualizzabile mediante una rappresentazione ad albero (dendrogramma), in cui sono rappresentati i passi di accorpamento/divisione dei gruppi.

Queste due suddivisioni sono del tutto trasversali; molti algoritmi, nati come "esclusivi", sono stati, in seguito, adattati nel caso "non-esclusivo" e viceversa.

Il clustering gerarchico

Quando si procede per partizioni successive da un solo cluster iniziale, contenente tutti i dati osservati, oppure da un insieme di cluster pari al numero degli elementi osservati, uno per cluster, allora si parla di clustering gerarchico.

Le tecniche di clustering gerarchico si fondano essenzialmente su due "filosofie":

1) dal basso verso l'alto (metodi aggregativi o Bottom-Up): questa filosofia prevede che, inizialmente, tutti gli elementi siano considerati cluster a sé e, successivamente, l'algoritmo provveda ad unire i cluster più vicini. L'algoritmo continua a unire elementi al cluster fino a ottenere un numero prefissato di cluster, oppure fino a che la distanza minima tra i cluster non superi un certo

valore, o, ancora, in relazione ad un determinato criterio statistico prefissato.

2) dall'alto verso il basso (metodi divisivi o Top-Down): all'inizio tutti gli elementi formano un unico cluster, in seguito l'algoritmo inizia a dividere lo stesso in tanti clusters di dimensioni inferiori. Il criterio che guida la divisione è, naturalmente, quello di ottenere gruppi sempre più omogenei. L'algoritmo procede fino a che non viene soddisfatta una regola di arresto, generalmente legata al raggiungimento di un numero prefissato di clusters. Queste tecniche, pur essendo più ricche di "proprietà matematiche, hanno un carattere meno empirico dei metodi aggregativi, basandosi su note proprietà statistiche della suddivisione della matrice delle devianze e codevianze".

Tuttavia esse, a causa della complessità dei calcoli richiesti, trovano scarso impiego all'interno della ricerca.

Le tecniche basate sugli algoritmi scissori si distinguono in due categorie:

1) I metodi monotetici, che realizzano la suddivisione dei gruppi basandosi sui valori assunti da una sola variabile.

2) I metodi politetici, che prendono in considerazione i valori assunti da tutte le variabili prescelte per la classificazione.

Il metodo divisivo più noto è quello di Edwards - Cavalli Sforza. Il criterio seguito da questo metodo è di esaminare ad ogni stadio tutte le possibili suddivisioni in due parti di tutti i gruppi. Sarà operata la divisione che fa diminuire maggiormente la varianza entro i gruppi.

La procedura operativa può essere schematizzata nei seguenti passi:

1) inizializzazione: date n unità statistiche o osservazioni, ogni elemento rappresenta un gruppo di un elemento (si hanno n cluster iniziali) e gli stessi vengono numerati da 1 a n;
2) selezione: vengono calcolate le distanze e selezionati i clusters più vicini rispetto ad una misura di prossimità fissata;
3) aggiornamento: si aggiorna il numero di clusters (n-1) attraverso l'unione di due gruppi a minima distanza tra loro; in corrispondenza si aggiorna la matrice delle distanze, sostituendo alle due righe che riferiscono la minima distanza una colonna con le distanze aggiornate rispetto ai nuovi clusters, per tenere conto del nuovo gruppo;
4) ripetizione: si eseguono nuovamente i passi 2 e 3 per n-1 volte;
5) arresto: la procedura viene fermata quando tutti gli elementi vengono incorporati in un unico cluster.

I vari tipi di clustering gerarchico

In base al modo in cui vengono calcolate le distanze tra i dati di input si distinguono diversi metodi gerarchici di clustering.

a) metodo del legame singolo (single linkage)

Questo metodo è anche noto come tecnica "del confinante più vicino" (nearest neighbour technique). La distanza tra i gruppi è posta pari alla più piccola delle distanze istituibili a due a due tra tutti gli elementi dei due gruppi. Se C e D sono due gruppi, la loro distanza è definita come la più piccola tra tutte le $n_1 \cdot n_2$ distanze che si possono calcolare tra ciascuna unità i di C e ciascuna unità j di D:

$$d_{hk} = \min(d_{ij}) \quad \forall i \in C, \quad \forall j \in D$$

L'adozione di questo algoritmo per la composizione dei gruppi evidenzia, in maniera netta, tutte le similitudini e somiglianze tra gli elementi, privilegia la differenza tra i gruppi piuttosto che l'omogeneità degli elementi di ogni gruppo. Il dendrogramma costruito su questa matrice ha i rami molto corti ed è più compatto: ciò proprio perché vengono valorizzate le somiglianze. Tale tecnica gode della particolare proprietà, delle sue partizioni, di essere invarianti rispetto a trasformazioni monotone delle variabili.

b) metodo del legame completo (complete linkage)

Il metodo è anche noto come quello “del confinante più lontano” (farthest neighbour technique) e rappresenta l’opposto della tecnica del legame singolo. In questo procedimento, si considera la maggiore delle distanze istituibili, a due a due, tra tutti gli elementi dei due gruppi:

$$d_{hk} = \max(d_{ij}) \quad \forall i \in C, \quad \forall j \in D$$

Questa tecnica tende a favorire l'omogeneità tra gli elementi del gruppo a scapito della differenziazione netta tra gruppi. Il dendrogramma costruito su questa matrice ha i rami molto più lunghi, i gruppi (e soprattutto i rami) si formano a distanze maggiori. In uno stesso range di valori, rispetto al legame singolo, gli elementi sono molto meno compatti e più diluiti.

c) metodo del legame medio (average linkage)

Per determinare la distanza tra due gruppi C e D utilizzando questa tecnica, si prendono in considerazione tutte le distanze fra

gli n_1 oggetti membri del gruppo C e gli n_2 oggetti membri del gruppo D. Con questa tecnica, la distanza fra i due gruppi è calcolata in base alla media aritmetica fra le distanze:

$$d(C,D) = \frac{1}{n_1 \cdot n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{ij} \quad \forall i \in C, \quad \forall j \in D$$

d) metodo del centroide

La tecnica del centroide fa riferimento ad una rappresentazione spaziale degli oggetti da classificare, infatti identifica, per ogni gruppo, un centroide definito come un vettore, che ha per coordinate i valori medi delle p variabili all'interno del gruppo. La distanza tra i gruppi coincide con la distanza tra i rispettivi centroidi. Se $\overline{X}_C$ e $\overline{X}_D$ sono i centroidi

dei due gruppi C e D, avremo:

$$d(C,D) = d(\overline{X}_C, \overline{X}_D)$$

Il metodo del centroide e quello del legame medio presentano delle analogie: il metodo del legame medio considera la media delle distanze tra le unità di ciascuno dei suoi gruppi, mentre il metodo del centroide calcola le medie di ciascun gruppo e, in seguito, misura le distanze tra esse.

e) metodo di Ward

Questa tecnica si propone di realizzare una classificazione gerarchica tramite la minimizzazione della varianza delle variabili entro ciascun gruppo. La tecnica è iterativa, ad ogni passo sono fusi i cluster che presentano la minima variazione della varianza entro i gruppi.

Questo metodo permette di generare dei gruppi composti da un numero di elementi comparabile. Il metodo si fonda sulla minimizzazione di una funzione obiettivo che vuole realizzare la massima coesione interna a ciascun gruppo e la massima separazione esterna tra gruppi diversi. La devianza totale delle p variabili analizzate viene scomposta in devianza nei gruppi e devianza fra i gruppi e, ad ogni passo della procedura gerarchica, si aggregano tra loro i cluster che comportano il minore incremento della devianza nei gruppi e il maggiore incremento della devianza tra gruppi, in modo da ottenere la maggiore coesione interna possibile e la maggiore separazione esterna tra gruppi.

La scelta del numero di gruppi

Non esiste un metodo unico per determinare il numero di cluster ottimale. Vi sono, piuttosto, dei possibili criteri da utilizzare.

L'idea di fondo è quella di scegliere il numero di clusters in base ai "salti" nelle distanze alle quali questi si formano, per identificare i gruppi più evidenti. Per esempio, se molti clusters si formano in sequenza, con piccole differenze di similarità, allora è difficile stabilire un numero di gruppi, ma se, a un certo punto, c'è un salto molto forte, ciò significa che, per ottenere la configurazione successiva, occorre abbassare molto la similarità (cioè fondere due clusters distanti). Questo salto indica una possibile scelta per il numero di gruppi.

Il criterio è, in ultima analisi, soggettivo, ma esistono degli indicatori che possono essere utilizzati per identificare il momento in cui fermare il procedimento.

Nel caso di una cluster gerarchica, la scelta del numero di gruppi può essere effettuata utilizzando, in primo luogo, la distanza di fusione. Per valutare l'entità dell'incremento della distanza di fusione si può ricorrere all'incremento relativo della distanza di fusione definito come:

$$\delta_k = \frac{d_k - d_{k+1}}{d_{k+1}}$$

verrà scelto il numero di gruppi k per cui δ_k è massimo.

La distanza di fusione, in termini di distanza riscalata, può essere facilmente evinta dall'osservazione del dendrogramma: se nel passaggio da K gruppi a K+1 si registra un forte incremento della distanza di fusione è consigliabile "tagliare" a K gruppi.

Attraverso l'osservazione del dendrogramma, un criterio di scelta può consistere nell'arresto della procedura di fusione prima di uno dei "salti" che vengono generati da aggregazione di gruppi molto distanti tra loro e, quindi, disomogenei.

Il clustering non gerarchico

Con le tecniche di classificazione "a partizioni ripetute" si cerca di determinare una partizione degli N oggetti in K gruppi che ottimizzi un criterio prefissato, fornendo come prodotto finale una sola partizione delle N osservazioni.

Le diverse partizioni sono determinate, partendo da quella iniziale, spostando in successione i singoli oggetti, secondo criteri prefissati, fino a raggiunge una situazione in cui lo spostamento di singoli elementi non migliorerebbe più il valore della funzione obiettivo (es. massimizzazione dell'omogeneità all'interno dei gruppi).

L’inizializzazione dell’algoritmo avviene indicando G centri di partenza intorno ai quali aggregare le unità. A differenza dei metodi gerarchici, l’assegnazione di un oggetto ad un cluster non è irrevocabile. Ovvero, le unità vengono riassegnate ad un diverso cluster se l’allocazione iniziale risulta inappropriata.
Nei metodi non gerarchici, l’inizializzazione è definita a partire da centroidi scelti a caso tra i punti del dataset o all’esterno del dataset. Supposto che, a priori, sia stato fissato il numero dei gruppi in cui si vuole ripartire il collettivo di partenza, le procedure non gerarchiche si articolano sostanzialmente nelle seguenti fasi:
1) inizializzazione: si scelgono G centri provvisori, i quali inducono una prima partizione temporanea. Tale raggruppamento avviene sulla base della minima distanza di un’unità da uno di questi centri. L’elemento A apparterrà al primo gruppo se la distanza tra A e il centro del gruppo è inferiore a quella tra A e qualunque altro centro dei restanti gruppi;
2) determinazione dei nuovi centri: si calcolano i baricentri dei gruppi ottenuti attraverso il processo di aggregazione del passo precedente. Si assumono i baricentri appena calcolati come nuovi centri provvisori;
3) Individuazione della nuova partizione: si ripete il procedimento di allocazione delle unità ai centri sulla base della minima distanza. Si itera la partizione tornando al passo 1;
4) Arresto della procedura: se tra un passo e il successivo non vi sono riallocazioni dei punti tra un gruppo e un altro, la procedura si arresta, giacché la partizione ottenuta può ritenersi soddisfacente.

Il metodo delle k-medie

In letteratura sono stati proposti diversi algoritmi che differiscono tra loro nei seguenti aspetti:

a) come sono inizializzati i centri di partenza;

b) come gli elementi vengono assegnati ai diversi centri;

c) come alcune o tutte le unità vengono eventualmente riassegnate ad un diverso gruppo.

Tra questi metodi, il più popolare è il cosiddetto metodo delle k-medie (k-means method) introdotto da MacQueen nel 1963. Questo presuppone che sia definito a priori il numero di gruppi ai quali le unità debbono essere assegnate. Il suo obiettivo, quindi, è quello di minimizzare la distanza tra ogni unità e il centroide del cluster, attraverso un'appropriata procedura iterativa.

L'algoritmo segue i passi generali di una procedura non gerarchica, con alcune varianti, che consentono di raggiungere più velocemente la soluzione ottimale:

1) inizializzazione: si scelgono G punti iniziali che fungono da centri provvisori e si costruisce la partizione iniziale assegnando ogni punto al gruppo il cui centro risulta più vicino;

2) determinazione dei nuovi centri: si calcolano i baricentri dei gruppi ottenuti attraverso il processo di aggregazione del passo precedente. Si assumono i baricentri appena calcolati come nuovi centri provvisori;

3) individuazione della nuova partizione: si ripete il procedimento di allocazione delle unità ai centri sulla base della minima distanza. Ad ogni assegnazione di un nuovo punto a un nuovo gruppo, si procederà alla rideterminazione del baricentro del nuovo e del vecchio gruppo. Si itera la partizione tornando al

passo 1;

4) arresto della procedura: se tra un passo e il successivo non vi sono riallocazioni dei punti tra un gruppo e un altro, la procedura si arresta in quanto la partizione ottenuta può ritenersi soddisfacente.

La maggiore velocità con cui si raggiunge la soluzione ottimale dipende, essenzialmente, dal fatto che, nel momento di individuazione della nuova partizione, ogni volta che un punto viene assegnato a un nuovo gruppo, viene ricalcolato il baricentro del nuovo e del vecchio gruppo e questi vengono utilizzati nel proseguo dell'analisi come centri provvisori.

Osservazioni sul metodo delle k-medie

L'algoritmo delle k-medie si distingue dalla procedura generica di raggruppamento non gerarchico in un elemento chiave.

Nella procedura classica, al momento di individuazione della nuova partizione, i nuovi centri sono ricalcolati solo al termine della fase di aggregazione di tutti i punti ai diversi gruppi. Nelle k-medie, invece, ogni volta che un punto è assegnato ad un gruppo, il baricentro di questo gruppo viene immediatamente ricalcolato (così come quello del gruppo a cui il punto apparteneva in precedenza).

Ciò implica un assestamento dei centri ad ogni singola assegnazione e non al termine di ogni passo della procedura.

In questo modo si garantisce un raggiungimento della convergenza dell'algoritmo più rapido (cioè con un minor numero di iterazioni) rispetto alle soluzioni classiche.

La procedura iterativa, su cui si fonda l'algoritmo delle k-medie, presenta alcuni potenziali problemi. In particolare, i risultati dell'analisi possono dipendere da:

a) la scelta dei centri iniziali: la scelta dei centri iniziali rappresenta il punto di partenza da cui prende avvio la ricerca di una partizione finale, che possa essere considerata una soluzione soddisfacente. L'approccio più banale si basa sulla scelta di G punti in maniera completamente casuale. Ovviamente, tale modo di procedere rende fortemente instabile la procedura (che se eseguita più volte fornirà risultati spesso differenti).

Un modo alternativo consiste nel generare, per ognuno dei G gruppi, L punti casuali il cui baricentro rappresenterà il centro provvisorio iniziale del gruppo, oppure, in alternativa, quello di eseguire più volte l'analisi e considerare come partizione finale dei dati la partizione media, ottenuta aggregando le singole soluzioni;

b) la scelta del numero dei gruppi: in merito al numero di gruppi G, se questa informazione non è nota a priori sulla base di informazioni possedute sulla popolazione, si potrà eseguire più volte la procedura, facendo variare G e scegliendo, poi, il valore di G che consente di avere la migliore partizione (in termini di massima eterogeneità tra i gruppi o minima interna ai gruppi);

c) l'ordine in cui le osservazioni sono riportate nella matrice dei dati: infine, per ridurre il problema legato all'ordine delle osservazioni, si può condurre una analisi con l'approccio classico in cui i centri vengono ricalcolati solo alla fine del processo di aggregazione e non punto dopo punto.

Confronto tra cluster gerarchica e non gerarchica

La scelta dell'algoritmo delle K-medie, piuttosto che di uno di classificazione gerarchica risiede, oltre che nella possibilità di scegliere a priori il numero di cluster da ottenere, principalmente nel risparmio computazionale che si ottiene utilizzando il primo metodo, piuttosto che il secondo.

Il metodo delle k-medie, d'altro canto, potrebbe dipendere eccessivamente dalla scelta iniziale dei centri e dal numero di gruppi; scelte che potrebbero rendere eccessivamente instabile la procedura e portare a una soluzione ben peggiore dell'ottimo globale.

I metodi gerarchici, d'altra parte, sono poco flessibili: se due unità sono state aggregate, non possono essere divise in una iterazione successiva, come invece è possibile fare attraverso un algoritmo non gerarchico; infatti, nel secondo caso, una partizione iniziale insoddisfacente può essere modificata successivamente, perché è possibile spostare ogni singola unità da un gruppo all'altro.

In sintesi, generalmente è consigliato applicare una prima analisi gerarchica, che consenta di individuare il numero ottimale di gruppi, da assegnare, in seguito, come input in una posteriore analisi non gerarchica, che consenta di ottenere la configurazione finale dei gruppi.

Ambiti di applicazione

Per una varietà di obiettivi di ricerca, gli studiosi hanno la necessità di scoprire quali oggetti in un set sono simili o dissimili tra loro. Una delle ragioni per cui la cluster analysis è così utile risiede nel fatto che i ricercatori, in ogni campo, hanno la continua necessità di fare classificazioni.

Nell’ambito economico, si cerca spesso di procedere a una classificazione dei concorrenti prima di attuare una strategia di ingresso in un mercato.
Un’altra fondamentale attività di marketing consiste nel dividere in gruppi clienti e prodotti: poiché le aziende non possono raggiungere la totalità dei clienti potenziali, esse devono dividere gli stessi in segmenti caratterizzati da bisogni e desideri simili.
Solo dopo aver adoperato una segmentazione, le imprese potranno scegliere quale gruppo adottare come target, posizionandosi in modo univoco sul mercato.
Al contrario delle ricerche di mercato che, troppo spesso, indirizzano verso la formazione di segmenti eccessivamente basati su considerazioni personali, sull’esperienza dei ricercatori e sulla pratica aziendale, la cluster analysis offre una metodologia grazie alla quale i segmenti formati dipendono in maniera ragguardevolmente limitata da elementi soggettivi.

La segmentazione di clienti, prodotti e mercati è, quindi, una delle principali applicazioni della cluster analysis, ma la stessa può anche essere applicata a contesti più particolari.
L’analisi dei gruppi trova, inoltre, numerose applicazioni in diverse discipline, nel momento in cui si vogliano fare delle classificazioni: per esempio, gli astronomi potrebbero essere interessati a comprendere quante classi di stelle esistono in base ad alcuni criteri statistici.
Anche in psichiatria si è via via sviluppato un notevole interesse nell’uso di questa tecnica finalizzato a perfezionare o ridefinire le correnti categorie di diagnosi, poiché i disagi della mente sono più sfuggenti di quelli del corpo. Gli psichiatri, spesso, classificano

i loro pazienti in base ai risultati di test, con lo scopo di migliorare la comprensione dei problemi mentali che affliggono gli stessi e pianificare un trattamento specifico.
Un ulteriore campo di applicazione si può riscontrare in ambito meteorologico: un'enorme quantità di dati inerenti il clima viene raccolta quotidianamente in tutto il mondo. L'esplorazione di essi, attraverso una analisi dei gruppi, potrebbe portare a nuove intuizioni riguardanti le nuove tendenze climatiche e ambientali, che potrebbero avere una significatività sia scientifica che pratica.

- La cluster analysis, inoltre, ha rilevanza per:
- Ridurre i dati in forma grafica
- Generare ipotesi di ricerca (con la dovuta circospezione, "E' la teoria che deve prescrivere cosa misurare", Einstein)
- Identificare tipi
- Costruire sistemi di classificazione automatica (tassonomia)
- Stratificare popolazioni da sottoporre a campionamento.

Analisi discriminante

L’analisi statistica multivariata comprende un corpo di metodologie statistiche che permettono di analizzare simultaneamente misurazioni riguardanti diverse caratteristiche (variabili qualitative o quantitative) di un insieme di individui in esame. Gli obiettivi principali delle metodologie di analisi multivariata sono riassumibili nella sintesi delle osservazioni ovvero nella semplificazione della loro struttura (riduzione del numero delle variabili), nell’ordinamento e nel raggruppamento (classificazione) di osservazioni, nello studio delle interdipendenze tra le variabili, nella formulazione e verifica di ipotesi operative.
Le diverse tecniche di analisi multivariata possono essere distinte a seconda che facciano o meno riferimento ad un modello distributivo assunto per le osservazioni e alla base degli sviluppi inferenziali. In questo senso le tecniche collegate allo studio della dipendenza (modello lineare generale, modelli lineari generalizzati) si contrappongono ad un insieme di metodologie giustificate prevalentemente da argomenti logico-intuitivi note sotto il nome di metodi di analisi dei dati. Sono questi dei metodi esplorativi (L. Fabbris,1991) ovvero di statistica descrittiva multidimensionale (L. Lebart, A. Morineau, J.P. Fénelon, 1982) che partono dal presupposto espresso chiaramente nella frase di uno dei maggiori esponenti della scuola francese di analisi dei dati: “il modello deve seguire i dati, non viceversa” (J.P. Benzécri, 1980). Un tale approccio porta a procedure di analisi euristiche, ovvero di carattere intuitivo-analogico, i cui risultati devono essere controllati e convalidati in un secondo tempo (logica del trovare), e si contrappone all’approccio confermativo per il quale la verifica della sussistenza di assunzioni effettuate prima ancora della rilevazione dei dati, viene condotta sulla base

di metodi statistico inferenziali (logica del giustificare). La scelta di uno dei due approcci dipende sia dagli obiettivi del ricercatore che dalle informazioni disponibili riguardo alla distribuzione delle variabili in esame, ovvero dalla possibilità di controllare sperimentalmente l'osservazione dei fenomeni.
Per questo motivo l'analisi dei dati è tradizionalmente collegata alle applicazioni in ambito socio-economico, mentre i metodi modellistico-inferenziali vengono maggiormente utilizzati nelle scienze sperimentali.

6.1 Funzione discriminante lineare di Fisher

Per analisi discriminante si intende un corpo di metodologie che, considerando un universo campionario *k*-dimensionale X suddiviso in *p* sottopopolazioni $X_1, X_2, \cdots, X_p$, permettono di assegnare una generica osservazione x ad una delle p sottopopolazioni.
Uno tra i primi a parlare di analisi discriminante multivariata fu R. A. Fisher (1936) con riferimento all'attribuzione di alcuni reperti fossili alla categoria dei primati o a quella degli umanoidi in base a diverse misurazioni effettuate sugli stessi. Nell'approccio di Fisher l'obiettivo dell'analisi discriminante è quello di individuare la sottopopolazione di appartenenza di un'osservazione multidimensionale in base alla conoscenza campionaria del comportamento delle diverse sottopopolazioni.
Non facendo alcuna assunzione sulla forma distributiva delle p sottopopolazioni da cui vengono estratti i campioni $X_1, X_2, \cdots, X_p$, l'assegnazione dell'osservazione *x* viene effettuata tramite una combinazione lineare $W = a'X$
delle *k* componenti della variabile X rilevata, tale da rendere massima la separazione (o discriminazione) tra i *p*

campioni. Il criterio che viene utilizzato per definire la trasformazione, ovvero il vettore *k*-dimensionale di costanti a, consiste pertanto nel pretendere che sia massima la differenza tra le medie di W nei *p* campioni, in modo da rendere meno ambigua la classificazione dell'osservazione $w = a'x$.
L'informazione parziale di partenza sia dunque costituita da p campioni $X_1, X_2, \cdots, X_p$ di numerosità n_j da ciascuna sottopopolazione X_j con $j = 1,2,\cdots, p$:

$$X_j = \begin{pmatrix} x_{11j} & \cdots & x_{1kj} \\ \vdots & & \vdots \\ x_{n_j 1 j} & \cdots & x_{n_j kj} \end{pmatrix} = \left[x_{ihj} \right]$$

$i = 1,\cdots, n_j,\ h = 1,\cdots, k$ e $j = 1,2,\cdots, p$.

Siano inoltre

$$\overline{X}_j = \frac{1}{n_j} X'_j u_{n_j} = (\overline{X}_{1j}, \cdots, \overline{X}_{kj})'$$

la media campionaria del *j*-esimo campione ed

$$S_j = \frac{1}{n_j} (X_j - u_{n_j} \overline{X}'_j)'(X_j - u_{n_j} \overline{X}'_j) = \left[S_{hlj} \right]$$

la matrice *k*×*k* delle varianze e covarianze campionarie del j-esimo campione (nelle espressioni precedenti *h, l* = 1, . . . , k e *j* = 1, . . . , *p*).
Trasformando tramite il vettore a la matrice *nj*×*k* del generico campione *j*-esimo, si ottiene per *j* = 1, . . . , *p* il vettore *nj*-dimensionale

$$W_j = X_j a$$

con media e varianza campionarie date da

$$\overline{W_j} = \frac{1}{n_j} W_j' u_{n_j} = \frac{1}{n_j} a' X_j' u_{n_j} = a' \overline{X}_j$$

$$S^2_{W_j} = \frac{1}{n_j} (X_j a - u_{n_j} a' \overline{X}_j)' (X_j a - u_{n_j} a' \overline{X}_j)$$

$$= \frac{1}{n_j} a' (X_j - u_{n_j} \overline{X}_j')' (X_j - u_{n_j} \overline{X}_j') a = a' S_j a$$

Complessivamente, posto $n = \sum_{j=1}^{p} n_j$ sia

$$X = (X_1', X_2', \cdots, X_p')'$$

la matrice
n × *k* di tutte le osservazioni disponibili ed inoltre sia

$$\overline{X} = \frac{1}{n} X' u_n = (\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_k)'$$

il vettore delle medie campionarie complessive ed

$$S = \frac{1}{n}(X - u_n \overline{X}')'(X - u_n \overline{X}') = [S_{hl}]$$

la matrice *k* × *k* delle varianze e covarianze campionarie calcolate in base a tutti i p campioni.
Considerando il generico elemento *(h,l)*-esimo della matrice *S*

$$S_{hl} = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ihj} - \overline{X}_h)(x_{ilj} - \overline{X}_l) =$$

$$= \sum_{j=1}^{p} \frac{n_j}{n} S_{hlj} + \frac{1}{n} \sum_{j=1}^{p} n_j (\overline{X}_{hj} - \overline{X}_h)(\overline{X}_{lj} - \overline{X}_l)$$

la matrice di varianze e covarianze S può dunque essere scomposta nel modo seguente

$$S = S_{(w)} + S_{(b)}$$

dove S(w) indica la matrice delle varianze e covarianze all'interno dei *p* campioni (within) data da

$$S_{(w)} = \sum_{j=1}^{p} \frac{n_j}{n} S_j$$

mentre S(b) è la matrice di varianze e covarianze tra i *p* campioni (between).

Analogamente trasformando tramite a la matrice $\underset{n\times k}{X}$ di tutte le osservazioni disponibili si ottiene il vettore *n*-dimensionale

$$W = Xa$$

con media e varianza date dalle espressioni seguenti

$$\overline{W} = a'\overline{X}$$

$$S_W^2 = a'Sa = a'S_{(w)}a + a'S_{(b)}a$$

Si voglia ora definire W (ovvero determinare a) in modo tale da massimizzare le differenze tra le medie campionarie $\overline{W_1},\cdots,\overline{W_p}$. Ciò implica la massimizzazione della varianza between di W ovvero di $a'S_{(b)}a$. Ovviamente quanto maggiori in valore assoluto sono gli elementi del vettore a, tanto più elevato `e il valore della forma quadratica. Quindi affinchè il problema della determinazione del massimo assoluto di $a'S_{(b)}a$ rispetto ad a sia ben definito, si considera un vincolo sulla dimensione di a dato dall'espressione $a'Sa = 1.$
Tale vincolo corrisponde a pretendere che W abbia varianza unitaria. Pertanto per la determinazione di *a* bisogna risolvere il seguente problema di massimo vincolato:

$$\begin{cases} \max a'S_{(b)}a \\ a'Sa = 1 \end{cases}$$

La funzione lagrangiana prende la forma seguente, dove lambda è il moltiplicatore di Lagrange

$$L = a'S_{(b)}a - \lambda(a'Sa - 1)$$

Il problema di massimo vincolato si traduce nella soluzione del sistema

$$\begin{cases} \dfrac{\partial L}{\partial a} = 2S_{(b)}a - 2\lambda Sa = 0 \\ \dfrac{\partial L}{\partial \lambda} = a'Sa - 1 = 0 \end{cases} = \begin{cases} \lambda = a'S_{(b)}a \\ a'Sa = 1 \end{cases}$$

si noti che la prima equazione del sistema può essere espressa nella forma di equazione caratteristica (o equazione agli autovalori)

$$S^{-1}S_{(b)}a = \lambda a$$

Dalla quale risulta come λ sia uno degli autovalori di $S^{-1}S_{(b)}$ ed a l'autovettore ad esso associato. Inoltre, affinchè si verifichi $\lambda = a'S_{(b)}a$, bisogna scegliere tra gli autovalori di $S^{-1}S_{(b)}$ quello che assume valore massimo. La variabile

$$W_{(1)} = a'_{(1)}X$$

definita tramite l'autovettore a(1) associato al maggiore degli autovalori λ_1 corrisponde dunque alla combinazione lineare

delle componenti della variabile k-dimensionale di partenza che separa maggiormente i p campioni ed è detta prima funzione discriminante lineare. L'autovalore λ_1 equivalente alla varianza between della variable W(1) `e detto potere discriminante di W(1) e misura la capacità di W(1) di separare le medie dei p campioni.
La definizione della seconda funzione discriminante lineare W(2) prevede che questa soddisfi la condizione di massimo e il vincolo precedenti, ed inoltre che sia incorrelata con W(1). In tal caso

$$W_{(2)} = a'_{(2)}X$$

dove il vettore a(2) `e dato dalla soluzione del sistema

$$\begin{cases} \max a'_{(2)}S_{(b)}a_{(2)} \\ a'_{(2)}Sa_{(2)} = 1 \\ a'_{(1)}Sa_{(2)} = 0 \end{cases}$$

Dopo manipolazioni algebriche basate sul metodo dei moltiplicatori di Lagrange, che esula dallo scopo del corso presentare, si ottiene:

$$S^{-1}S_{(b)}a_{(2)} = \mu a_{(2)}$$

L'equazione definisce pertanto in $\mu = \lambda_2$ il secondo autovalore più grande di $S^{-1}S_{(b)}$.

Si possono individuare tante funzioni discriminanti lineari quanti sono gli autovalori non nulli della matrice $S^{-1}S_{(b)}$, ossia un numero pari al rango della matrice stessa $g = r(S^{-1}S_{(b)})$. In genere si considera un numero $t < g$ di funzioni discriminanti, interrompendo l'analisi quando il potere discriminante della *(t + 1)*-esima funzione discriminante lineare, ossia il valore del *(t + 1)*-esimo autovalore di $S^{-1}S_{(b)}$, diviene trascurabile. Una misura del potere discriminante complessivo delle prime *t* funzioni discriminanti è data dal rapporto

$$\frac{\sum_{q=1}^{t} \lambda_q}{\sum_{q=1}^{g} \lambda_q} = \frac{\sum_{q=1}^{t} \lambda_q}{tr(S^{-1}S_{(b)})}$$

Nel caso in cui si considerino *t* funzioni discriminanti lineari, l'osservazione *x* `e assegnata individuando il valore j* tale che, calcolato $w_{(q)} = a'_{(q)}x$ per $q = 1, \cdots, t$ si abbia:

$$\sum_{q=1}^{t} \left| w_{(q)} - \overline{W}_{(q),j^*} \right| = \min_j \sum_{q=1}^{t} \left| w_{(q)} - \overline{W}_{(q),j} \right|$$

essendo $\overline{W}_{(q),j}$ la media di $W_{(q)}$ nel j-esimo campione, per q = 1, . . . , t.

Da un punto di vista geometrico l'analisi discriminante consiste nel rappresentare le *p* nuvole *k* dimensionali di *nj* punti (i p campioni) in uno spazio euclideo di dimensione $t < k$ tale da evidenziare opportunamente le distanze tra i campioni. L'output dell'analisi discriminante deve perciò includere il rango *t* del nuovo riferimento (ovvero del modello discriminante), la posizione di ciascuna dimensione del modello discriminante rispetto al riferimento originario (i vettori a(q)), la posizione dei *p* campioni di osservazioni nel sottospazio delle variabili discriminanti (le medie $\overline{W}_{(q),j}$).

L'analisi delle corrispondenze

A termine lezione, verranno discussi gli esempi sul libro (Fabbris pp.286-296), pertanto consigliamo agli studenti di munirsi delle pagine del testo.

Sia $\underset{H\times M}{N}$ una matrice di frequenze e $P = N/n$

La frequenza relativa di un'unità si dice massa. La massa dell'unità *h* è $p_{h.}$ e quella dell'unità *m* è $p_{.m}$.
Le distribuzioni condizionate sulle righe e sulle colonne si si dicono profili. Il profilo dell'unità *h* è la riga *h*-esima, il profilo della modalità *m* è la colonna *m*-esima.
I profili marginali, medie aritmetiche ponderate delle distribuzioni condizionate, si dicono baricentri o centroidi.
Il punto in cui s'incontrano i baricentri delle entità si dice centroide o baricentro della configurazione.
Dalla matrice *P* si derivano le matrici diagonali ${}_uD$ e ${}_vD$ aventi come elementi non nulli le frequenze relative marginali rispettivamente di colonna e di riga.

Pertanto, ${}_uD^{-1}P$ rappresenta le distribuzioni condizionate di riga, mentre ${}_vD^{-1}P'$ rappresenta le distribuzioni condizionate di colonna e $Tr({}_uD) = Tr({}_vD) = 1$.
La soluzione fattoriale è l'insieme degli autovalori e autovettori della matrice quadrata ma non simmetrica

$$S = P'{}_uD^{-1}P{}_vD^{-1}$$

La ricerca degli autovalori e degli autovettori equivale a trovare il vettore β di M elementi che massimizza la media aritmetica ponderata dei quadrati dei valori del vettore

$$ {}_u f = {}_u D^{-1} P_v D^{-1} \beta = \{ {}_u f_k \} = \left\{ \sum_{m=1}^{M} \frac{p_{m|k}}{p_{\cdot m}} \beta_m \right\} = \left\{ \sum_{m=1}^{M} \frac{n_{km}}{n^*_{km}} \beta_m \right\} $$

$n^*_{km} = \frac{n_{k.} \times n_{.m}}{n}$ essendo le frequenze teoriche in ipotesi di indipendenza.

$$ \begin{cases} \max_u f' {}_v D {}_u f = \left\{ \sum_{h=1}^{H} {}_u f_h^2 p_{h.} \right\} \\ \beta' {}_v D^{-1} \beta = 1 \end{cases} $$

Nello spazio duale delle unità

$$ \begin{cases} \max_v f' {}_u D {}_v f = \left\{ \sum_{m=1}^{M} {}_v f_m^2 p_{.m} \right\}, \\ \beta' {}_v D^{-1} \beta = 1 \end{cases} $$

ove

$$ {}_u f = {}_v D^{-1} P' {}_u D^{-1} \alpha $$
$$ \alpha' {}_u D^{-1} \alpha = 1 $$

La soluzione porge gli autovettori associati agli stessi autovalori della soluzione duale.
Anche se sono in corrispondenza duale, gli autovettori $\alpha \; e \; \beta$ si ricavano indipendentemente.

Per ciascun autovalore Ω_k gli autovettori sono legati dalla relazione di "mutua transizione", dalle due "formule di transizione":

$$\alpha_k = {}_u D^{-1} P \beta_k \Big/ \sqrt{\Omega_k} = \{\alpha_{kh}\} = \left\{ \sum_{m=1}^{M} \frac{p_{hm}\beta_{km}}{p_{h.}\sqrt{\Omega_k}} \right\}$$

$$\beta_k = {}_v D^{-1} P' \alpha_k \Big/ \sqrt{\Omega_k} = \{\beta_{km}\} = \left\{ \sum_{h=1}^{H} \frac{p_{hm}\alpha_{kh}}{p_{.m}\sqrt{\Omega_k}} \right\}$$

Lebart *et al.* (1977: II.2) chiamano i vettori $\alpha_k \; e \; \beta_k$ "assi fattoriali" e "fattori" i vettori aventi gli stessi elementi divisi per la massa delle entità corrispondenti.
Il numero massimo di autovalori non banali non nulli è pari a

$$r = rango(X) \leq \min\{H, M\} - 1$$

La tabella delle frequenze può ricostruirsi tramite la formula:

$$n_{km} = n^*_{km} \left\{ 1 + \sum_{k=1}^{r} \alpha_{km}\beta_{km} \right\}$$

Proprietà degli autovalori

L'autovalore Ω_k è una misura di variabilità (denominata anche "inerzia", termine in uso nella fisica) del *k*-esimo asse fattoriale sul quale si proiettano i punti.

1. Il primo autovalore è sempre uguale a 1 e per questo è detto banale. La somma degli altri autovalori eguaglia la somma delle varianze delle nuvole di punti (modalità-riga e modalità-colonna) osservati. Per la dualità della soluzione fattoriale delle corrispondenze, un autovalore rappresenta la lunghezza del vettore della soluzione basata sulle entità-riga così come delle entità-colonna. L'autovalore Ω_k è, dunque, la media aritmetica ponderata dei punteggi fattoriali sia delle entità-riga che delle entità-colonna:

$$\sum_{h=1}^{H} p_{h.} f_{kh}^2 = \sum_{m=1}^{M} p_{.m} f_{km}^2 = \Omega_k \qquad k = 1, \cdots, r$$

2. La somma degli autovalori si può esprimere come traccia della matrice *S*:

$$1 + \Omega_1 + \Omega_2 + \cdots = tr(S) = \sum_{h=1}^{H} \sum_{m=1}^{M} \frac{n_{km}^2}{n_{h.} n_{.m}},$$

da cui si ricava che la somma degli autovalori è funzione del coefficiente $\Phi^2 = \chi^2 / n$ calcolato sulla tabella:

$$\Omega_1 + \Omega_2 + \cdots = \frac{1}{n} \sum_{h=1}^{H} \sum_{m=1}^{M} \frac{(n_{km} - n_{km}^*)^2}{n_{km}^*} = \frac{\chi^2}{n} = \Phi^2$$

Gli autovalori assumono valori al più pari a 1

- Il primo uguale a 1, è l'unico non nullo nel caso di indipendenza statistica (nel senso delle righe e delle colonne)
- I rimanenti $r - 1$ sono minori o uguali a 1
- Se ce n'è un altro uguale a 1, la matrice di frequenze può essere scomposta in due sottomatrici a blocchi diagonali, pertanto esistono due gruppi di modalità riga-colonna fra loro indipendenti (e così via se ci sono altri *s* autovalori pari a 1)
- Se la matrice è diagonale, allora si hanno tutti autovalori pari a 1 (caso limite di perfetta dipendenza)

3. Il generico autovalore Ω_k si può interpretare come il quadrato del coefficiente di correlazione canonica, ossia il coefficiente di correlazione tra la *k*-esima combinazione lineare delle modalità di riga e di quelle sulle colonne:

$$corr({}_u f_k, {}_v f_k) = \sum_{h=1}^{H} \sum_{m=1}^{M} f_{kh} f_{km} p_{hm} = \sqrt{\Omega_k}$$

Coordinate fattoriali

Le coordinate fattoriali delle unità *h* sul fattore *k*, ${}_u f_{kh}$, ottenute proiettando il valori della *h*-esima riga della tabella sull'asse fattoriale *k*-esimo, sono medie aritmetiche ponderate degli *M* valori delle frequenze n_{hm} della riga *h* della tabella con pesi dati dagli *M* elementi del vettore β a loro volta ponderati. Il peso di ciascun coefficiente β_m è il rapporto n_{hm} / n^*_{hm}. Dunque:

- Il valor medio dei pesi dei coefficienti è 1

- In corrispondenza della modalità la cui frequenza teorica è superiore a quella empirica il peso della frequenza è minore

di 1; se, invece, la frequenza nell'ipotesi di indipendenza è inferiore a quella empirica, il peso è maggiore di 1

La relazione tra coordinate e coefficienti degli assi fattoriali è analoga se si considerano le modalità-colonna invece di quelle riga. I coefficienti saranno gli *H* elementi del vettore α.

Analisi multipla delle corrispondenze

Si consideri la matrice ottenuta osservando su *n* unità statistiche *p* variabili qualitative. Il numero complessivo di modalità delle *p* variabili è

$$M = \sum_{i=1}^{p} M_i$$

I dati sono rilevati in una supermatrice $\underset{n\times M}{X}$ detta matrice disgiuntiva, avente le unità statistiche nel senso delle righe e le *M* modalità di tutte le variabili dicotomizzate nel senso delle colonne:

$$X = \lfloor X_1, \cdots, X_i, \cdots, X_p \rfloor,$$

dove $\underset{n\times M_i}{X_i}$ contenente $M_i - 1$ zeri e un solo 1 per ogni riga, quest'ultimo sulla colonna corrispondente alla modalità del carattere x_i posseduta dall'unità sulla riga *h*. Chiaramente il numero di 1 sarà *p* per ogni unità statistica e $n_{.m}$ per la colonna *m*-esima.

Le frequenze marginali delle *M* modalità si pongono sulla diagonale principale di una matrice diagonale

$$\underset{M\times M}{D} = diag\lfloor D_1, \cdots, D_i, \cdots, D_p \rfloor = \{d_{mm}\} = \{n_{.m}\}$$

Per la ricerca degli assi fattoriali su cui proiettare i punti unità si costruisce la matrice

$$\underset{n\times n}{A_i} = X_i D_i^{-1} X_i',$$

dove

$$\underset{n\times n}{A_i} = \{_i a_{hh'}\} = \sum_{m=1}^{M} \frac{n_{hm} n_{h'm}}{n_{\cdot m}}$$

Il primo fattore $_u f$ è l'autovettore *n*-dimensionale di media nulla associato al più grande autovalore non banale Ω della matrice

$$\sum_{i=1}^{p} A_i / p,$$

dove

$$_u f = \frac{1}{p\Omega} \sum_{i=1}^{p} X_{i\,v}\, f_i$$

e $_v f_i$ è il vettore di M_i coordinate della variabile x_i sul primo asse fattoriale. Il vettore rende massima la media aritmetica ponderata dei quadrati delle M_i coordinate delle *p* variabili nella forma disgiuntiva (che equivale a massimizzare la media di *p* varianze):

$$\frac{1}{p} \sum_{i=1}^{p} {}_v f_i' D_{i\,v}\, f_i$$

Gli altri autovettori si trovano in corrispondenza degli autovalori ottenuti decurtando progressivamente la matrice $\sum_{i=1}^{p} A_i / p$ della

variabilità e covariabilità estratta dai passi precedenti. Se $M \leq n$, il numero massimo di autovalori non banali è

$$r = rango(X) \leq M - p$$

La matrice $\underset{M \times M}{B} = X'X$, detta matrice di Burt, è formata da p^2 blocchi:

- Quelli sulla diagonale principale sono matrici diagonali $M_i \times M_i$ con le frequenze delle modalità della variabile x_i

- Quelli esterni alla diagonale principale sono matrici rettangolari $M_i \times M_j$ che rappresentano ognuna le frequenze incrociate delle variabili x_i e x_j

Gli autovettori della matrice $D^{-1}B$, dove $\underset{M \times M}{D} = \{d_{mm}\}$ è la matrice diagonale il cui elemento generico è $d_{mm} = b_{mm}$ sono vettori di *M* elementi $_v f$ con i quali si possono calcolare i punteggi fattoriali delle *n* unità statistiche.

La somma degli autovalori di $D^{-1}B$ (che sono Ω^2) è uguale alla media dei p^2 valori dei coefficienti Φ^2 tra tutte le possibili coppie delle *p* variabili analizzate:

$$\sum_{k=1}^{r} \Omega_k^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \Phi_{ij}^2$$

(Ricordiamo che $\Omega_1 + \Omega_2 + \cdots = \frac{1}{n} \sum_{h=1}^{H} \sum_{m=1}^{M} \frac{(n_{km} - n_{km}^*)^2}{n_{km}^*} = \frac{\chi^2}{n} = \Phi^2$)

Da cui la considerazione che l'analisi delle corrispondenze multiple è un metodo per studiare le relazioni multiple fra *p* variabili utilizzando solo la dipendenza fra coppie di variabili.

L'analisi delle corrispondenze mediante il modello analitico di Burt è plausibile per insiemi di variabili di dimensioni contenute, perché il numero di modalità che si analizzano incide notevolmente sul calcolo.
Se *p* = 2, si può applicare qualsiasi tipo di analisi, quella disgiuntiva, l'approccio di Burt e l'analisi della tabella di frequenze. I tre tipi di modelli danno la stessa soluzione fattoriale, tuttavia gli autovalori non sono gli stessi, con quelli dell'analisi disgiuntiva inferiori a tutti.

Proprietà degli autovalori nell'analisi multipla

1. La somma degli autovalori non banali è data da:

$$\Omega_1 + \Omega_2 + \cdots = tr\left(\sum_{i=1}^{p} A_i / p\right) - 1$$

che si dimostra essere uguale a

$$\sum_{k=1}^{r} \Omega_k = \frac{1}{p} \sum_{i=1}^{p} (M_i - 1) = \frac{M}{p} - 1$$

La dipendenza esistente tra il massimo valore che gli autovalori possono assumere e le modalità della costruzione della matrice disgiuntiva indica che la frazione di varianza non ha significato statistico. Pertanto l'interpretazione della soluzione va basata sull'analisi dei contributi delle unità e delle modalità, come si vedrà in seguito.

2. Con l'eccezione del caso p=2, non si può trovare un'equazione caratteristica per ogni variabile x_l, ma solo per l'intero insieme delle p variabili, ossia per il supervettore ${}_v f = ({}_v f_1', \cdots, {}_v f_p')$ di M elementi.

3. Gli assi dei primi fattori della soluzione analitica delle corrispondenze sono gli stessi, a meno di un parametro di scala, sia che si analizzino i punti-riga, sia i punti-colonna. L'analisi delle corrispondenze semplice della matrice disgiuntiva X, considerata come una tabella di frequenze, è pertanto equivalente a un'analisi multipla delle corrispondenze.

Proiezione di punti supplementari

$$\underset{n\times M}{X} = \begin{bmatrix} \underset{n\times M_a}{X_a} & \underset{n\times M_s}{X_s} \end{bmatrix}$$

Le coordinate delle modalità supplementari si calcolano proiettando gli M_s punti sull'asse β_k

$${}_s f_k = {}_s D^{-1} Y \beta_k' \Big/ \sqrt{\Omega_k} = \{ {}_s f_{km} \} = \left\{ \sum_{h=1}^{n} \frac{{}_s x_{km}}{{}_s x_{\cdot m}} u \; f_{km} \right\},$$

dove ${}_sD$ è la matrice diagonale delle frequenze marginali delle M_s modalità supplementari e *Y* la matrice $M_s \times M_a$ di frequenze incrociate dei due insiemi di modalità attive e illustrative messe l'una nel senso delle righe e l'altra in quello delle colonne:

$$ {}_s f_{km} = \sum_{h=1}^{n} \frac{{}_s x_{km}}{{}_s x_{\cdot m}} {}_u f_{km} \qquad m = 1, \cdots, M_s, $$

dove ${}_u f_{km}$ è dato dalla precedente:

$$ {}_u f = {}_u D^{-1} P_v D^{-1} \beta = \{ {}_u f_k \} = \left\{ \sum_{m=1}^{M} \frac{p_{m|k}}{p_{\cdot m}} \beta_m \right\} = \left\{ \sum_{m=1}^{M} \frac{n_{km}}{n_{km}^*} \beta_m \right\} $$

${}_s x_{km}$ è il generico elemento della matrice X_s e ${}_s x_{\cdot m} = \sum_{h=1}^{H} {}_s x_{hm}$.

Le M_s coordinate associate all'autovettore *k*-esimo si possono interpretare come covarianze tra il fattore ${}_u f_k$ e la *m*-esima modalità dicotomica ($m = 1, \cdots, M_s$)

Criteri per l'interpretazione dei fattori

L'interpretazione dell'esito di un'analisi delle corrispondenze si basa sull'analisi del grafico che risulta alla fine dell'analisi e sulla valutazione dei contributi delle modalità alla determinazione della soluzione.

La denominazione degli assi ottenuti si basa quasi esclusivamente sull'analisi del grafico sul quale sono rappresentati i punti corrispondenti alle modalità analizzate. E' bene fidarsi soprattutto di criteri "obiettivi", meglio se quantitativi, quali quelli basati sulle misure del prossimo paragrafo.

Scomposizione della variabilità di modalità e fattori: contributi

Il contributo della modalità *h*-esima alla determinazione del fattore *k*-esimo è dato da:

$$C_{h|k} = \frac{p_{h\cdot} f_{kh}^2}{\sum_{h=1}^{H} p_{h\cdot} f_{kh}^2} = \frac{p_{h\cdot} f_{kh}^2}{\Omega_k}$$

Similmente per la modalità *m*-esima dello stesso fattore:

$$C_{m|k} = \frac{p_{\cdot m} f_{km}^2}{\sum_{m=1}^{M} p_{\cdot m} f_{km}^2} = \frac{p_{\cdot m} f_{km}^2}{\Omega_k}$$

Il contributo del generico asse *k*-esimo alla variabilità della modalità *h*-esima è dato da:

$$C_{k|h} = \frac{p_{h\cdot} f_{kh}^2}{\sum_{k=1}^{r} p_{h\cdot} f_{kh}^2} = \frac{f_{kh}^2}{\sum_{k=1}^{r} f_{kh}^2}$$

$$C_{k|m} = \frac{p_{\cdot m} f_{kh}^2}{\sum_{k=1}^{r} p_{\cdot m} f_{kh}^2} = \frac{f_{km}^2}{\sum_{k=1}^{r} f_{km}^2}$$

La somma dei contributi $C_{k|h}$ estesa ai *q* assi è detta comunanza:

$$\sum_{k=1}^{q} C_{k|h} = \sum_{k=1}^{q} \frac{f_{kh}^2}{\sum_{k=1}^{r} f_{kh}^2}$$

Assume valore 0 quando la modalità non è correlata ai fattori trovati, valore massimo quando gli assi della soluzione spiegano tutta la variabilità della modalità. Se la comunanza è bassa, non è molto illuminante proiettare la modalità sugli assi per studiarne la posizione (la modalità tenderà a collocarsi vicino al baricentro della posizione, insieme alle altre a comunanza bassa). Se il valore della comunanza è elevato, allora è da considerarsi una modalità che discrimina le entità poste nel senso delle righe.

Forme tipiche e configurazioni

Le più comuni:

- Ellissoide
- Nuvole separate
- Ferro di cavallo
- Triangolo o tetraedro
-

(*Illustrare le configurazioni p.279 testo, e commentarle opportunamente)*

A termine lezione, verranno discussi gli esempi sul libro (Fabbris pp.286-296), pertanto consigliamo agli studenti di munirsi delle pagine del testo.