Dispense di Statistica II

Franco Flandoli

2013-14

Indice

Capitolo 1

Correlazione e regressione

1.1 Covarianza e coefficiente di correlazione

Definizione 1 *Date due v.a. X, Y, si chiama covarianza tra X ed Y il numero*

$$Cov\left(X,Y\right)=E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]$$

dove μ_X e μ_Y sono le medie di X ed Y. La definizione ha senso se μ_X e μ_Y sono finiti ed il valor medio complessivo è finito, cosa che accade ad esempio se si suppone che sia $E\left[X^2\right]<\infty$ e $E\left[Y^2\right]<\infty$.

La definizione è quindi analoga, algebricamente, a quella di varianza, e risulta infatti

$$Var\left[X\right]=Cov\left(X,X\right)$$

e

$$Cov\left(X,Y\right)=E\left[XY\right]-\mu_X\mu_Y$$

come per la varianza. Però il numero $Cov\left(X,Y\right)$ può avere segno qualsiasi. Ad esempio, se $\mu_X=0$ e prendiamo $Y=-X$, vale $Cov\left(X,Y\right)=-E\left[X^2\right]$.

Anche la covarianza soffre dei problemi di scala noti per la varianza. Qui, non potendo prendere la radice quadrata ($Cov\left(X,Y\right)$ non è sempre positiva), si normalizza in quest'altro modo, dividendo per le deviazioni standard.

Definizione 2 *Chiamiamo coefficiente di correlazione tra X ed Y il numero definito da*

$$\rho\left(X,Y\right)=\frac{Cov\left(X,Y\right)}{\sigma_X\sigma_Y}.$$

Scriveremo anche ρ_{XY} al posto di $\rho\left(X,Y\right)$.

Si noti che, per la disuguaglianza di Hölder,

$$|Cov(X,Y)| \leq \sqrt{E\left[(X-\mu_X)^2\right] E\left[(Y-\mu_Y)^2\right]}$$

e quindi $|\rho(X,Y)| \leq 1$. Questo dimostra la prima delle seguenti proprietà, che tutte insieme chiariscono l'aspetto di universalità, o invarianza per cambio di unità di misura (invarianza di scala), di ρ, a differenza della covarianza.

Proposizione 1 *Vale*

$$-1 \leq \rho(X,Y) \leq 1.$$

Vale inoltre

$$Cov(aX,bY) = abCov(X,Y)$$

per ogni $a,b \in \mathbb{R}$, *e*

$$\rho(aX,bY) = \rho(X,Y)$$

per ogni $a,b>0$.

Proof. Abbiamo già visto come mai $-1 \leq \rho(X,Y) \leq 1$. Dimostriamo la seconda proprietà. Vale

$$\begin{aligned} Cov(aX,bY) &= E[(aX-\mu_{aX})(bY-\mu_{bY})] = E[(aX-a\mu_X)(bY-b\mu_Y)] \\ &= abE[(X-\mu_X)(Y-\mu_Y)] = abCov(X,Y). \end{aligned}$$

Vale poi

$$\rho(aX,bY) = \frac{Cov(aX,bY)}{\sqrt{Var[aX]\,Var[bY]}} = \frac{abCov(X,Y)}{\sqrt{a^2b^2}\sqrt{Var[X]\,Var[Y]}} = \frac{ab}{|ab|}\rho(X,Y)$$

e quindi la formula desiderata, se $a,b>0$. ■

Nello stesso modo si dimostra la seguente proprietà, che in un certo senso è la linearità della convarianza nei suoi argomenti. Si noti che le costanti additive spariscono, come per la varianza.

Proposizione 2

$$Cov(aX+bY+c,Z) = aCov(X,Z) + bCov(Y,Z)$$

$$Cov(X,\alpha Y+\beta Z+\gamma) = \alpha Cov(X,Y) + \beta Cov(X,Z).$$

Proof. Basta dimostrare la prima in quanto la covarianza è simmetrica. Vale

$$\begin{aligned} Cov(aX+bY+c,Z) &= E\left[\left(aX+bY+c-\mu_{aX+bY+c}\right)(Z-\mu_Z)\right] \\ &= E[(a(X-\mu_X)+b(Y-\mu_Y))(Z-\mu_Z)] \\ &= aCov(X,Z) + bCov(Y,Z). \end{aligned}$$

■

Ricordiamo che se X ed Y sono v.a. indipendenti, allora $E[XY] = \mu_X\mu_Y$ (mentre il viceversa non è vero in generale). Ne discende subito il seguente risultato.

Teorema 1 *Se X ed Y sono v.a. indipendenti, allora*

$$Cov(X,Y)=0, \quad \rho(X,Y)=0.$$

Viceversa, se $Cov(X,Y)=0$, non è detto che X ed Y siano indipendenti. Se però (X,Y) è gaussiano (definizione che daremo nel seguito) e $Cov(X,Y)=0$, allora X e Y sono indipendenti. Anche questo fatto contribuisce alla tendenza pratica a ritenere che la condizione $Cov(X,Y)=0$ sia un notevole sintomo di indipendenza.

Definizione 3 *Diciamo che X e Y sono scorrelate se hanno correlazione nulla, $\rho(X,Y)=0$, o equivalentemente se $Cov(X,Y)=0$.*

Quindi l'indipendenza implica la scorrelazione.

A livello numerico su dati sperimentali, se la correlazione è molto vicino a zero, questo è un buon indicatore di indipendenza, o più precisamente di scorrelazione (invece, dipendendo il numero $Cov(X,Y)$ dalla scala scelta, la sua vicinanza a zero è meno assoluta, quindi può trarre in inganno). Precisiamo cosa intendiamo con correlazione di dati sperimentali. Stiamo pensando di avere n coppie di valori sperimentali (x_1,y_1), ... , (x_n,y_n), o più espressivamente una tabella

	X	Y
1	x_1	y_1
...	...	...
...	...	...
n	x_n	y_n

in cui le colonne corrispondono alle variabili e le righe agli "individui" (unità sperimentali, unità osservate). Di questi dati sperimentali possiamo calcolare la *varianza empirica* ed il *coefficiente di correlazione empirico* definiti da

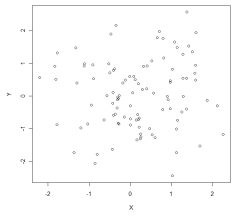
$$\widehat{Cov}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}), \quad \widehat{\rho}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

Questi indicatori sono buone stime di quelli teorici, ad esempio per via della legge dei grandi numeri. Fatte queste premesse, la vicinanza a zero di $\widehat{\rho}$ si interpreta come sintomo di indipendenza o comunque bassa dipendenza, la vicinanza ad 1 come elevato legame positivo, a -1 come elevato legame negativo.

Esaminiamo questi fatti per mezzo del software R. Innanzi tutto generiamo due campioni di cardinalità 100, con distribuzione gaussiana standard, mostriamo le coppie (x_i,y_i) nel piano cartesiano e calcoliamo la correlazione empirica:

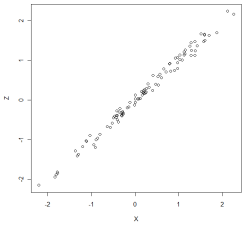
```
X=rnorm(100); Y=rnorm(100)
cor(X,Y)
```

```
[1] 0.06068838
plot(X,Y)
```



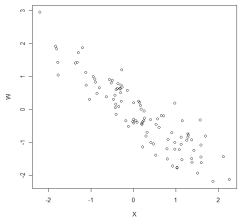
Questa è una situazione a correlazione sostanzialmente nulla. Costruiamo invece un campione Z simile a X ma un po' perturbato in modo aleatorio:

```
Z=X+0.1*rnorm(100)
cor(X,Z)
[1] 0.9949628
plot(X,Z)
```



Questa è una situazione ad elevatissima correlazione positiva. Proviamo poi

```
W=-X+0.5*rnorm(100)
cor(X,W)
[1] -0.8987381
plot(X,W)
```



Quest'ultima è una situazione a moderata/elevata correlazione negativa. Si noti che il coefficiente di correlazione non è esattamente uguale al coefficiente angolare, come si potrebbe pensare dai nomi. Si veda sotto il legame.

Il segno di $\widehat{Cov}$ corrisponde all'inclinazione positiva o negativa della nuvola di punti, come abbiamo visto negli esempi numerici. Gli esempi però sono solo alcuni dei possibili esempi, quindi può essere utile un ragionamento di carattere generale. La formula $\widehat{Cov} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ permette di effettuarlo, anche se in modo un po' vago.

Il punto $(\overline{x}, \overline{y})$ sta, diciamo, al centro della nuvola, comunque essa sia orientata. Analizziamo il caso in cui sia $\widehat{Cov} > 0$. Questo significa che nella somma $\sum_{i=1}^{n}$ predominano addendi positivi. Per semplicità di ragionamento, supponiamo che siano tutti positivi. Se $(x_i - \overline{x})(y_i - \overline{y}) > 0$, significa che i fattori $(x_i - \overline{x})$ e $(y_i - \overline{y})$ hanno lo stesso segno. Se sono positivi, significa che il punto (x_i, y_i) sta a nord-est di $(\overline{x}, \overline{y})$; se sono negativi, significa che sta a sud-ovest. In ogni caso, i punti (x_i, y_i) si trovano, rispetto a $(\overline{x}, \overline{y})$, come nel secondo disegno visto sopra. Per semplicità abbiamo ipotizzato che tutti i termini $(x_i - \overline{x})(y_i - \overline{y})$ fossero positivi: in genere non è così, ce ne saranno anche di negativi, ma predominano quelli positivi, quindi predomina la struttura grafica del tipo detto sopra, anche se alcuni punti si troveranno a nord-ovest o sud-est di $(\overline{x}, \overline{y})$. Il ragionamento nel caso $\widehat{Cov} < 0$ è identico.

Osservazione 1 *Il nome "covarianza", inteso come "covariazione", corrisponde a quanto appena descritto: si calcolano le variazioni congiunte dei dati x_i e y_i rispetto ai valori medi $\overline{x}$ ed $\overline{y}$.*

Menzioniamo infine il fatto che il numero $\rho(X, Y)$ è una *misura del legame lineare* tra X ed Y. Questo verrà chiarito sotto con lo studio della regressione lineare.

1.1.1 Regressione lineare semplice

Ipotizziamo che tre v.a. X, Y ed ε siano legate dalla relazione lineare

$$Y = aX + b + \varepsilon$$

dove a e b sono numeri reali. Interpretiamo questa scrittura pensando che X ed Y siano legate da una relazione lineare (graficamente una retta di equazione $y = ax + b$, per cui a si dirà coefficiente angolare e b intercetta), perturbata però da un *errore* casuale ε. La v.a. X verrà detta *input*, o *predittore*, o fattore, la Y *output*, o quantità da predire.

Supporremo sempre che ε sia standardizzato:

$$E[\varepsilon] = 0.$$

L'ventuale media di ε è inglobata in b. Supporremo inoltre che ε ed X siano indipendenti o almeno incorrelate:

$$Cov(X, \varepsilon) = 0.$$

Chiameremo *modello lineare* (semplice) la relazione precedente. Diremo anche *modello di regressione lineare* (semplice), e chiameremo *retta di regressione* la retta $y = ax + b$.

Ci poniamo due scopi:

1. trovare formule che permettano di calcolare approssimativamente a, b e la deviazione standard σ_ε dell'errore ε a partire da dati sperimentali, quando si ipotizza il modello lineare ma non si conoscono i coefficienti;

2. interpretare rigorosamente il concetto di coefficiente di correlazione nell'ambito del modello lineare.

Raggiungeremo entrambi gli scopi calcolando valori medi, varianze e covarianze tra le diverse grandezze in gioco.

Proposizione 3 *Se tre v.a. X, Y ed ε sono legate dalla relazione lineare $Y = aX + b + \varepsilon$, X ed ε sono scorrelate e $\mu_\varepsilon = 0$, e se $\sigma_X > 0$, allora i coefficienti a e b sono univocamente determinati:*

$$a = \frac{Cov(Y, X)}{\sigma_X^2}$$
$$b = \mu_Y - a\mu_X.$$

Inoltre

$$\sigma_\varepsilon^2 = \sigma_Y^2 - a^2\sigma_X^2.$$

Proof. Vale, per linearità e per la proprietà $E[\varepsilon] = 0$,

$$E[Y] = aE[X] + b.$$

Vale inotre, per le regole sulla varianza (qui usiano la scorrelazione tra X ed ε),

$$Var[Y] = a^2 Var[X] + Var[\varepsilon].$$

Infine, per analoghe ragioni vale

$$\begin{aligned} Cov(Y, X) &= Cov(aX + b + \varepsilon, X) \\ &= aCov(X, X) + Cov(\varepsilon, X) \end{aligned}$$

da cui

$$Cov(Y, X) = aVar[X].$$

Da queste relazioni si ricavano le tre formule. ■

Supponiamo di avere n dati sperimentali, che in questo contesto significa avere n coppie (x_1, y_1), ... , (x_n, y_n) (n individui sono stati esaminati e per ciascuno sono stati trovati i valori di due grandezze X ed Y). Possiamo calcolare i numeri

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2, \quad \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2$$
$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

e considerarli come approssiamzioni (stime) rispettivamente di

$$\begin{gathered} E[X], \quad E[Y] \\ Var[X], \quad Var[Y] \\ Cov(X,Y). \end{gathered}$$

Tramite queste approssimazioni possiamo stimare a, b e σ. Indichiamo con $\widehat{a}$ e $\widehat{b}$ le stime empiriche di a, b così trovate.

Osservazione 2 (retta di regressione) *La retta di equazione*

$$y = \widehat{a}x + \widehat{b}$$

è detta retta di regressione associata alla tabella di dati da cui sono stati calcolati $\widehat{a}$ e $\widehat{b}$. Chiameremo quindi $\widehat{a}$ coefficiente angolare e $\widehat{b}$ intercetta, rella regressione lineare. A volte si usa lo stesso linguaggio per i parametri teorici a e b del modello teorico $Y = aX + b + \varepsilon$. Molti software permettono di tracciare la retta di regressione in sovrapposizione al plot dei dati sperimentali, per apprezzare la pendenza.

1.1.2 Interpretazione di ρ_{XY}

La Proposizione 3 stabilice una relazione tra coefficiente di correlazione al coefficiente angolare a della retta di regressione:

Corollario 1

$$\rho_{XY} = \frac{\sigma_X}{\sigma_Y} a.$$

Proof. Dalla Proposizione 3 e dalla definizione di coefficiente di correlazione, abbiamo

$$a = \frac{Cov(Y,X)}{\sigma_X^2} = \frac{Cov(Y,X)}{\sigma_X \sigma_Y} \frac{\sigma_Y}{\sigma_X} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

quindi $a = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$. ■

Innanzi tutto questo chiarisce che a *non è* il coefficiente di correlazione, come invece per una sorta di gioco di parole si è spesso portati a credere. Del resto, ρ_{XY} può variare solo tra -1 e 1, mentre la pendenza di una retta può essere maggiore di quella delle bisettrici.

Vale però la regola: $a > 0$ se e solo se $\rho_{XY} > 0$ (ed analogamente per valori negativi). Quindi $\rho_{XY} > 0$ è indice di legame lineare diretto, cioè con coefficiente angolare positivo, mentre $\rho_{XY} < 0$ è indice di legame lineare inverso (nel senso: una variabile cresce se l'altra cala), cioè con coefficiente angolare negativo. Almeno il segno di ρ_{XY} è facilmente interpretabile.

Supponiamo di *standardizzare sia X sia Y*. In realtà non importa che sottraiamo la media, ma è essenziale che dividiamo per la deviazione standard, in modo da ricondurci ad avere $\sigma_X = 1$ e $\sigma_Y = 1$. In questo caso

$$\rho_{XY} = a$$

Questo può offrire un'interpretazione più stretta.

L'interpretazione più precisa viene invece dallo studio dell'errore. Abbiamo visto sopra che

$$\sigma_\varepsilon^2 = \sigma_Y^2 - a^2\sigma_X^2.$$

Sostituendo $a = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$ si trova

$$\sigma_\varepsilon^2 = \sigma_Y^2\left(1 - \rho_{XY}^2\right).$$

Questo dice che la varianza dell'errore, cioè la grandezza che misura quanto preciso sia il legame lineare tra X ed Y, è tanto maggiore quanto più vicino a zero è ρ_{XY}: valori vicini a zero di ρ_{XY} implicano un cattivo legame lineare (errore elevato). Viceversa, valori di ρ_{XY} vicini a ± 1 (non importa il segno!), implicano σ_ε piccolo e quindi un legame lineare stretto:

$$\begin{array}{rcl} \rho_{XY} \sim 0 & \text{corrisponde a} & \sigma_\varepsilon \text{ elevato } (\sim \sigma_Y) \\ \rho_{XY} \sim \pm 1 & \text{corrisponde a} & \sigma_\varepsilon \sim 0. \end{array}$$

Quindi, salvo che si esegua una standardizzazione di entrambe le variabili, ρ_{XY} non è legato tanto all'inclinazione della retta di regressione quanto piuttosto alla *precisione con cui essa descrive il legame tra le variabili.*

Nel ragionamento precedente bisogna osservare che la grandezza o piccolezza di σ_ε è relativa anche alla grandezza o piccolezza di σ_Y. Questa è solo una questione di unità di misura delle quantità aleatorie che stiamo esaminando. Il discorso diventa indipendente dall'unità di misura e dall'ordine di grandezza dei valori tipici di Y se introduciamo la *varianza standardizzata* dell'errore:

$$\frac{\sigma_\varepsilon^2}{\sigma_Y^2}.$$

Per essa vale

$$\frac{\sigma_\varepsilon^2}{\sigma_Y^2} = 1 - \rho_{XY}^2$$

portando ad un ragionamento più universale circa il legame tra entità dell'errore e valore di ρ_{XY}^2.

Infine, introduciamo alcuni nuovi nomi. Essi si ispirano all'idea che con un modello lineare stiamo cercando di dare una *spiegazione della variabilità* della grandezza Y. Abbiamo una grandezza Y, essa varia in modo imprevedibile, aleatorio, e noi vorremmo

capire se queste variazioni sono almeno in parte spiegabili tramite un legame lineare con un predittore X: quando osserviamo ad es. valori di Y pià grandi della media, questo non è dovuto semplicemente al caso, ma al fatto che il predittore ha assunto valori ad es. più grandi del solito (se $a > 0$). Tutto però è pur sempre corrotto dall'errore, per cui la spiegazione della variabilità di Y offerta dalla retta di regressione non è mai una spiegazione completa.

In quest'ottica, Y ha una sua varianza, una sua variabilità. L'espressione $aX + b$ riesce a spiegarne una parte, l'altra resta non spiegata. La parte non spiegata di Y è la differenza tra Y e la parte spiegata, cioè $aX + b$. Quindi la parte non spiegata di Y è proprio l'errore ε (non c'è niente di nuovo, è solo una questione di linguaggio).

Con questo nuovo linguaggio:

Definizione 4 *Chiamiamo varianza spiegata la percentuale della varianza che è stata spiegata da $aX + b$ e varianza non spiegata la percentuale complementare. Siccome la parte di Y non spiegata è ε, in termini matematici la varianza non spiegata è*

$$\frac{\sigma^2_\varepsilon}{\sigma^2_Y}$$

mentre la varianza spiegata è

$$1 - \frac{\sigma^2_\varepsilon}{\sigma^2_Y}.$$

Ma questa è pari a ρ^2_{XY}! Siamo arrivati al seguente risultato:

Proposizione 4 *Il coefficiente di correlazione al quadrato, ρ^2_{XY}, è la varianza spiegata $1 - \frac{\sigma^2_\varepsilon}{\sigma^2_Y}$ dalla relazione lineare.*

Più ρ^2_{XY} è alto (vicino a 1) più la relazione lineare riesce a spiegare la variabilità di Y.

1.1.3 Domande

1. Definizione teorica di covarianza, di coefficiente di correlazione e loro formule empiriche (stessa cosa naturalmente per la varianza e la deviazione standard)
2. Invarianza di scala (o meno) di covarianza e correlazione: enunciato e dimostrazione
3. Legami tra indipendenza e condizioni sulla covarianza (risp. correlazione)
4. Vicinanza a zero di $\widehat{\rho}$ rispetto a vicinanza a zero di $\widehat{Cov}$
5. Com'è fatta una tabella di dati sperimentali relativi ad una coppia di variabili (X, Y)

6. Interpretazione grafica del segno di $\widehat{Cov}$, sulla base della formula empirica
7. Significato della vicinanza di ρ a zero o a ± 1, in termini di legame lineare tra le variabili
8. Definizione di varianza spiegata nella regressione
9. Legami matematici tra varianza spiegata, coefficiente di correlazione e coefficiente angolare nella regressione lineare semplice
10. Quando coefficiente di correlazione e coefficiente angolare coincidono
11. Scrivere i comandi per generare un insieme di 30 numeri casuali con $\widehat{\rho} > 0$.

1.2 Matrice di covarianza

Definizione 5 *La matrice di covarianza Q (risp. di correlazione ρ) di un vettore $X = (X_1, ..., X_n)$ è la matrice $n \times n$ definita da $Q_{ij} = Cov(X_i, X_j)$ (risp. $\rho_{ij} = \rho(X_i, X_j)$), per $i, j = 1, ..., n$.*

Esempio 1 *Nella Lezione 1 abbiamo esaminato il legame tra due variabili aleatorie X ed Y. Possiamo costruire la matrice di covarianza del vettore (X, Y), data semplicemente da*

$$Q = \begin{pmatrix} \sigma_X^2 & Cov(X,Y) \\ Cov(X,Y) & \sigma_Y^2 \end{pmatrix}.$$

La matrice di correlazione è

$$\rho = \begin{pmatrix} 1 & \rho(X,Y) \\ \rho(X,Y) & 1 \end{pmatrix}.$$

Le seguenti proprietà vengono descritte per Q ma valgono anche per ρ.

Proposizione 5 *La matrice di covarianza Q di un vettore aleatorio $X = (X_1, ..., X_n)$ è simmetrica*

$$Q_{ij} = Q_{ji}$$

e definita non-negativa

$$x^T Q x \geq 0$$

per ogni $x \in \mathbb{R}^n$.

Proof. La simmetria di Q discende dalla simmetria dell'operazione di covarianza:

$$Q_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i) = Q_{ji}.$$

Per la proprietà di definita non-negatività, preso $x = (x_1, ..., x_n) \in \mathbb{R}^n$, vale

$$x^T Q x = \sum_{i,j=1}^{n} Q_{ij} x_i x_j = \sum_{i,j=1}^{n} Cov\left(X_i, X_j\right) x_i x_j = \sum_{i,j=1}^{n} Cov\left(x_i X_i, x_j X_j\right)$$
$$= Cov\left(\sum_{i=1}^{n} x_i X_i, \sum_{j=1}^{n} x_j X_j\right) = Var\left[W\right] \geq 0$$

dove $W = \sum_{i=1}^{n} x_i X_i$. Abbiamo usato la linearità della covarianza in entrambi gli argomenti. ■

Ricordiamo il

Teorema 2 (spettrale) *Se Σ è una matrice simmetrica $n \times n$, allora esiste una base ortonormale $\{e_1, ..., e_n\}$ di $\mathbb{R}^n$ i cui elementi sono autovettori di Σ: per opportuni numeri $\lambda_1, ..., \lambda_n$ vale*

$$\Sigma e_i = \lambda_i e_i, \qquad i = 1, ..., n.$$

Supporremo sempre di aver ordinato gli autovalori in modo che sia $\lambda_1 \geq ... \geq \lambda_n$.

Corollario 2 *Detta U la matrice avente i vettori $e_1, ..., e_n$ come colonne, detta Σ_e la matrice*

$$\Sigma_e = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

vale

$$\Sigma = U \Sigma_e U^T.$$

Diremo che Σ può essere diagonalizzata.

Per il teorema spettrale, la matrice di covarianza Q di un vettore aleatorio piò essere diagonalizzata:

$$Q = U Q_e U^T. \tag{1.1}$$

Inoltre, essendo Q definita non-negativa, vale

$$\lambda_i \geq 0, \qquad i = 1, ..., n.$$

Infatti $\lambda_i = \lambda_i e_i^T e_i = e_i^T Q e_i \geq 0$. Abbiamo usato il fatto che $e_i^T e_i = 1$ (vettori di norma uno) e che $\lambda_i e_i = Q e_i$, più la non-negatività di Q.

Usando entrambi questi fatti si può definire la radice quadrata di Q. La dimostrazione della proposizione fornisce anche la sua costruzione, che richiede la conoscenza di autovettori e autovalori di Q.

Proposizione 6 *Esiste una matrice simmetrica definita non-negativa $\sqrt{Q}$, detta radice quadrata di Q, avente la proprietà $\left(\sqrt{Q}\right)^2 = Q$.*

Proof. In primo luogo possiamo definire facilmente la radice quadrata di Q_e, ponendo

$$\sqrt{Q_e} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sqrt{\lambda_n} \end{pmatrix}.$$

Si vede subito che questa è simmetrica ed il suo quadrato è Q_e. Tramite questa matrice, "tornando indietro", possiamo definire

$$\sqrt{Q} := U\sqrt{Q_e}U^T.$$

Si verifica facilmente che la matrice $\sqrt{Q}$ è simmetrica ed il suo quadrato è uguale a Q. Infatti Abbiamo

$$\left(\sqrt{Q}\right)^T = U\left(\sqrt{Q_e}\right)^T U^T = U\sqrt{Q_e}U^T = \sqrt{Q}$$

e

$$\left(\sqrt{Q}\right)^2 = U\sqrt{Q_e}U^T U\sqrt{Q_e}U^T = U\sqrt{Q_e}\sqrt{Q_e}U^T = UQ_eU^T = Q$$

in quanto $U^TU = Id$. ■

1.2.1 Commenti di algebra lineare

Le seguenti note, esterne al corso di statistica, possono essere utili per rammentare il significato preciso e l'interpetazione geometrica di alcune espressioni ed alcuni fatti ricordati sopra.

Osservazione 3 *Lo spazio $\mathbb{R}^n$ delle n-ple $(x_1, ..., x_n)$ è uno spazio vettoriale. Su $\mathbb{R}^n$ consideriamo il prodotto scalare euclideo $\langle ., . \rangle$ e la norma euclidea $|\cdot|$. Una base ortonormale $\{e_1, ..., e_n\}$ è un insieme di n vettori di norma uno e perpendicolari, $\langle e_i, e_j \rangle = \delta_{ij}$ per $i, j = 1, ..., n$. Data una base ortonormale $\{e_1, ..., e_n\}$, ogni elemento (vettore) $x \in \mathbb{R}^n$ si scrive univocamente nella forma*

$$x = \alpha_1 e_1 + ... + \alpha_n e_n$$

e vale

$$\alpha_i = \langle x, e_i \rangle, \qquad i = 1, ..., n$$

(dette coordinate di x nella base $\{e_1, ..., e_n\}$).

Osservazione 4 *Indichiamo con $u_1, ..., u_n$ la base canonica di $\mathbb{R}^n$, cioè quella data da*

$$u_1 = \begin{pmatrix} 1 \\ 0 \\ ... \\ 0 \end{pmatrix}, ..., u_n = \begin{pmatrix} 0 \\ 0 \\ ... \\ 1 \end{pmatrix}.$$

Le coordinate di un punto $x = (x_1, ..., x_n)$ in questa base sono i numeri x_i stessi. Presa un'altra base ortonormale $\{e_1, ..., e_n\}$, indichiamo con U la matrice la cui prima colonna è il vettore e_1, la seconda e_2 e così via. Potremmo scrivere $U = (e_1, ..., e_n)$. Vale

$$Uu_i = e_i, \qquad i = 1, ..., n.$$

La matrice U è ortogonale:

$$U^{-1} = U^T.$$

Infatti si verifica subito che $U^T U$ è la matrice identica, quindi anche UU^T. Ricordiamo che le trasformazioni ortogonali sono isometrie, come le rotazioni o le riflessioni.

Osservazione 5 *Se $v \in \mathbb{R}^n$ è un vettore di componenti $(x_1, ..., x_n)$ nella base canonica $u_1, ..., u_n$, se $e_1, ..., e_n$ è una nuova base ortonormale, se U è matrice suddetta avente come colonne i vettori $e_1, ..., e_n$, posto*

$$\begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} = U^T \begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix}$$

allora $y_1, ..., y_n$ sono le componenti di v rispetto alla base $e_1, ..., e_n$. Infatti, $y_1 = \langle e_1, v \rangle$ e così via.

Osservazione 6 *Dimostriamo che il teorema spettrale implica il corollario, ovvero che vale $\Sigma = U\Sigma_e U^T$. Basta osservare che*

$$U^T \Sigma U = U^T (\Sigma e_1, ..., \Sigma e_n) = (\langle e_i, \Sigma e_j \rangle) = (\lambda_j \langle e_i, e_j \rangle) = (\lambda_j \delta_{ij}) = \Sigma_e$$

da cui

$$U\Sigma_e U^T = U \left(U^T \Sigma U \right) U^T = \Sigma$$

essendo $U^T U = UU^T = I$.

1.2.2 Matrici di covarianza e correlazione emprica

Supponiamo di avere una tabella di dati del tipo

	X_1	...	X_p
1	$x_{1,1}$	...	$x_{1,p}$
2	$x_{2,1}$	...	$x_{2,p}$
...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$

dove le colonne rappresentano diverse variabili (ad esempio X_1 = PIL, ... , X_{p-1} = spese per istruzione, X_p = spese per sanità), le righe rappresentano diversi "individui" (ad esempio le nazioni europee) ed i valori numerici sono noti, sono stati misurati.

Dal punto di vista teorico, abbiamo visto il concetto di matrice di covarianza Q delle v.a. $(X_1, ..., X_p)$. Dal punto di vista empirico, possiamo calcolare la matrice di covarianza empirica $\widehat{Q}$ associata alla precedente tabella. La componente $\widehat{Q}_{ij}$ si ottiene calcolando la covarianza empirica tra le colonne i e j:

$$\widehat{Q}_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{k,i} - \overline{x}_i)(x_{k,j} - \overline{x}_j)$$

dove $\overline{x}_i = \frac{1}{n} \sum_{k=1}^{n} x_{k,i}$. Analogamente si può definire la matrice di correlazione empirica $\widehat{\rho}$ di componenti

$$\widehat{\rho}_{ij} = \frac{\sum_{k=1}^{n} (x_{k,i} - \overline{x}_i)(x_{k,j} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{k,i} - \overline{x}_i)^2 \sum_{k=1}^{n} (x_{k,j} - \overline{x}_j)^2}}.$$

Il significato applicativo di queste matrici è la conseguenza del significato di covarianza e correlazione tra stringhe di dati; nel caso della correlazione, il numero $\widehat{\rho}_{ij}$ misura il legame lineare tra le colonne i e j, con le interpretazioni del segno descritte al paragrafo corrispondente.

I legami così catturati sono però a due a due, non globali tra tutte le variabili nel loro complesso. Per ogni coppia di variabili X_i e X_j, abbiamo il corrispondente coefficiente $\widehat{\rho}_{ij}$.

Con software `R`, data una matrice `A`, basta calcolare `cor(A)` e viene restituita la tabella delle correlazioni (la matrice di correlazione). Chiedendo `plot(A)` si ottiene una figura piuttosto ricca che indica i plot a due a due delle diverse variabili, da cui si può intuire l'eventuale legame lineare ed il suo segno. Confrontando la tabella `cor(A)` col grafico `plot(A)`, si ottiene un primo insieme di informazioni sui legami tra le variabili oggetto di studio. Esemplifichiamo con la seguente tabella, il cui significato verrà descritto altrove:

	PLIC	SC	SA.SC	TD	TMI
Piem	0.088	0.471	-0.707	-0.607	-0.395
Vaos	-1.545	0.348	-0.642	-0.813	1.578
Lomb	0.202	1.397	-0.836	-0.790	-0.538
TrAA	0.677	0.435	-1.269	-0.966	-0.075
Vene	0.088	1.334	-1.210	-0.848	-0.497
FrVG	0.639	-0.005	-1.028	-0.804	-1.301
Ligu	1.190	-0.247	0.470	-0.429	-0.354
EmRo	0.658	1.177	-1.315	-0.863	-0.347
Tosc	0.126	1.092	-0.795	-0.644	-1.355
Umbr	-1.431	0.675	-0.140	-0.524	-1.287
Marc	0.278	1.090	-0.265	-0.702	-0.0006
Lazi	2.329	0.546	-0.080	-0.113	-0.014
Abru	0.335	-0.373	0.402	-0.456	0.040
Moli	0.658	-1.289	0.065	0.451	-1.151
Camp	-1.811	-1.314	2.031	1.664	0.414
Pugl	-0.766	-0.926	1.038	0.648	1.109
Basi	-0.747	-1.154	0.661	0.844	2.001
Cala	-0.500	-1.727	1.571	2.153	0.632
Sici	-0.918	-1.130	1.332	1.517	1.783
Sard	0.449	-0.403	0.717	1.285	-0.238

(per inciso, si può provare a caricarla in R con un semplice copia-incolla tramite il comando

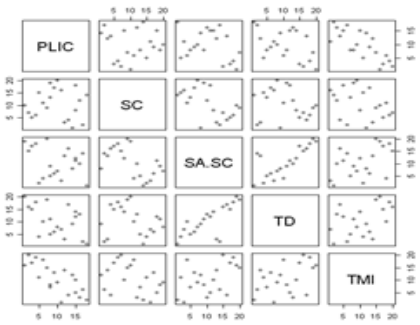
```
A<-read.table(clipboard,dec=',',header=T,row.names=1)
```

che funziona così: si copia da questo file la tabella, si scrive il comando su `R` e si dà l'invio)

```
cor(A)
```

	PLIC	SC	SA.SC	TD	TMI
PLIC	1	0.32	-0.41	-0.36	-0.44
SC	0.32	1	-0.84	-0.85	-0.48
SA.SC	-0.41	-0.84	1	0.90	0.51
TD	-0.36	-0.85	0.90	1	0.48
TMI	-0.44	-0.48	0.51	0.48	1

```
plot(A)
```



1.2.3 Trasformazione lineare (affine) dei momenti

La soluzione dei seguenti esercizi è basata sulla linearità del valore atteso (e quindi della covarianza, rispetto a ciascuno dei suoi argomenti)

Esercizio 1 *Sia $X = (X_1, ..., X_n)$ un vettore casuale, A una matrice $n \times d$, cioè $A : \mathbb{R}^n \to \mathbb{R}^d$, $b \in \mathbb{R}^d$, ed $Y = (Y_1, ..., Y_d)$ un vettore casuale definito da*

$$Y = AX + b.$$

Sia $\mu^X = \left(\mu_1^X, ..., \mu_n^X\right)$ il vettore dei valori medi di X, ovvero $\mu_i^X = E[X_i]$ e sia $\mu^Y = \left(\mu_1^Y, ..., \mu_d^Y\right)$ il vettore dei valori medi di Y. Allora

$$\mu^Y = A\mu^X + b.$$

Soluzione. L'identità $Y = AX + b$, per componenti significa

$$Y_i = \sum_{j=1}^{n} A_{ij} X_j + b_i.$$

Pertanto, per la linearità del valor medio,

$$E[Y_i] = E\left[\sum_{j=1}^{n} A_{ij} X_j + b_i\right] = \sum_{j=1}^{n} A_{ij} E[X_j] + b_i$$

che è la versione per componenti dell'identità da dimostrare.

Esercizio 2 *Sotto le stesse ipotesi, se Q^X e Q^Y sono le matrici di covarianza di X ed Y, allora*

$$Q^Y = AQ^X A^T.$$

Soluzione. Sempre usando l'identità per componenti scritta sopra,

$$\begin{aligned} Q^Y_{ij} &= Cov\left(Y_i, Y_j\right) = Cov\left(\sum_{i'=1}^{n} A_{ii'}X_{i'} + b_i, \sum_{j'=1}^{n} A_{jj'}X_{j'} + b_j\right) \\ &= \sum_{i'=1}^{n} A_{ii'} Cov\left(X_{i'}, \sum_{j'=1}^{n} A_{jj'}X_{j'} + b_j\right) \\ &= \sum_{i'=1}^{n} A_{ii'} \sum_{j'=1}^{n} A_{jj'} Cov\left(X_{i'}, X_{j'}\right) = \sum_{i'=1}^{n}\sum_{j'=1}^{n} A_{ii'} Q^X_{i'j'} A_{jj'} \end{aligned}$$

avendo usato la linearità della covarianza nelle due componenti. Ricordiamo che, date due matrici A e B, vale $(AB)_{ij} = \sum_k A_{ik}B_{kj}$. Allora

$$\sum_{i'=1}^{n}\sum_{j'=1}^{n} A_{ii'} Q^X_{i'j'} A_{jj'} = \sum_{j'=1}^{n} \left(AQ^X\right)_{ij'} A_{jj'}.$$

Per interpretare anche quest'ultimo come prodotto tra matrici bisogna trasporre A:

$$= \sum_{j'=1}^{n} \left(AQ^X\right)_{ij'} A^T_{j'j} = \left(AQ^X A^T\right)_{ij}.$$

L'esercizio è risolto.

1.2.4 Domande

1. Definire matrice di covarianza e correlazione, teorica ed empirica
2. Dimostrazione della simmetria
3. Enunciato e dimostrazione della definita non-negatività
4. Se le variabili $X_1, ..., X_p$ sono standardizzate, che differenza c'è tra la la matrice di covarianza e quella di correlazione?
5. Diagonalizzazione della matrice di covarianza (cosa si può dire su autovalori e autovettori per una matrice di covarianza)
6. Cosa produce il plot di una tabella
7. Enunciare e dimostrare la formula di trasformazione della covarianza sotto trasformazioni affini.

1.3 Regressione lineare multipla

Supponiamo di avere una tabella di numeri del tipo

	X_1	...	X_p	Y
1	$x_{1,1}$	...	$x_{1,p}$	y_1
2	$x_{2,1}$	...	$x_{2,p}$	y_2
...	...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$	y_n

dove le colonne rappresentano diverse variabili (ad esempio X_1 = reddito, ... , X_p = numero anni istruzione, Y = spese per mostre e musei), le righe rappresentano diverse unità sperimentali o individui (persone, città ecc.). Si noti che, a differenza della tabella su cui abbiamo calcolato la matrice di correlazione empirica, qui c'è una colonna privilegiata, quella della variabile Y.

Ci chiediamo se le variabili Y ed $X_1, ..., X_p$ *siano legate* da una relazione funzionale, a meno di errore, della forma:

$$Y = f(X_1, ..., X_p, \varepsilon).$$

Osservazione 7 (causa-effetto) *In quasi tutte le applicazioni concrete di questi modelli si vorrebbe, con essi, quantificare o catturare una relazione causa-effetto: le variabili X_i corrispondono alle cause, Y all'effetto. Si vorrebbe capire se le variabili $X_1, ..., X_p$* ***influiscano*** *su Y; se Y* ***dipenda*** *da $X_1, ..., X_p$. Un maggior reddito induce a maggiori spese per mostre e musei? Ed un maggior grado di istruzione (misurato ad esempio come numero di anni si studio)? Purtroppo bisogna convincersi che la teoria che stiamo sviluppando permette solo di stabilire se ci sia un legame tra queste grandezze, se Y ed $X_1, ..., X_p$* ***siano legate*** *da una relazione funzionale (come abbiamo detto sopra). Una cosa è un legame, un'altra è una relazione causa-effetto. Il numero di mezzi pubblici di trasporto ed il numero di insegnanti di una nazione occidentale sono legati, correlati positivamente (entrambi sono più elevati nelle nazioni più grandi), ma non hanno alcuna relazione causa-effetto l'uno con l'altro. Essi hanno una concausa: la dimensione della nazione. Questo è un fatto abbastanza generale. La presenza di una concausa provoca spesso dei legami, delle correlazioni, ma spesso le grandezze influenzate dalla comune causa non hanno relazione cuasale le une con le altre, non si influenzano a vicenda.*

Osservazione 8 *Nonostante le precedenti precisazioni, è così intutivo usare un linguaggio del tipo "Y dipende da $X_1, ..., X_p$" che lo faremo anche noi in modo informale. Va sempre ricordato che non può essere preso alla lettera. In alcuni casi corrisponderà alla realtà, in altri no e gli strumenti matematici che sviluppiamo non sono in grado di dirimere questo aspetto interpretativo.*

Le variabili (grandezze) $X_1, ..., X_p$ verranno dette *input, fattori, predittori*, mentre la grandezza Y è l'*output*.

Osservazione 9 (fattori e predittori) *Il termine "fattori" (che allude a "cause") è legato ai remark precedenti; non va quindi inteso in senso letterario. E' invece interessante capire il perché del termine "predittori". Supponiamo di aver stabilito che certe grandezze Y e $X_1, ..., X_p$ sono legate da una relazione funzionale di quelle che stiamo studiando. Non importa se questa relazione stia effettivamente descrivendo una relazione causa-effetto oppure se, dal punto di vita interpretativo, sia del tutto artificiosa (come potrebbe essere una relazione tra numero di mezzi pubblici ed il numero di insegnanti). Tale relazione può essere utilizzata per eseguire predizioni: calcolata in nuovi valori delle $X_1, ..., X_p$ (non già speriementati) essa produce dei corrispondenti valori di Y. Permette di predire il valore di Y in corrispondenza di input mai sperimentati. Se ad esempio Y sono volumi di vendita di un prodotto e le X_i sono grandezze socio-economiche del territorio, una volta costruito il modello sulla base di dati speriementati in una regione, possiamo esportare il modello in regioni meno note prevedendo le vendite, sulla base dei dati socio-economici di quella nuova regione.*

Nel seguito ci limiteremo allo studio di relazioni funzionali di tipo lineare (affine) ovvero della forma:

$$Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$$

(b è detta *intercetta* e gli a_i, così come b, sono i *parametri* del modello).

In realtà, la relazione appena scritta è astratta, riguarda le variabili aleatorie (è un po' come una legge fisica tipo $F = ma$). Noi invece siamo partiti da una tabella di dati, e quindi possiamo formulare il problema in modo più concreto, più attinente ai dati. A partire dai dati specifici, esaminiamo una relazione lineare tra essi della forma

$$y_i = a_1 x_{i,1} + ... + a_p x_{i,p} + b + \varepsilon_i \qquad i = 1, ..., n \tag{1.2}$$

dove abbiamo introdotto i numeri ε_i, detti *residui*, definiti dalla relazione stessa, cioè da

$$\varepsilon_i = y_i - (a_1 x_{i,1} + ... + a_p x_{i,p} + b)\,.$$

Così facendo, non stiamo ipotizzando una relazione lineare tra le variabili: l'identità (1.2) vale sempre, con la definizione data dei residui. Il problema è: quanto sono grandi i residui, affinché valga la (1.2)? I residui sono una misura dell'errore insito nella relazione lineare.

Osservazione 10 *I residui, così definiti, sono funzione dei parametri del modello. Più tardi, i parametri verranno fissati in modo ottimale e chiameremo "residui" i corrispondenti residui (con questa precisazione intendiamo dire che i residui, a seconda del momento in cui se ne parla, solo funzione dei parametri, oppure sono numeri univocamente individuati da una condizione di ottimalità; in questo momento, sono funzioni dei parametri).*

Possiamo poi calcolare lo *scarto quadratico medio* dei residui, ancora funzione dei parametri $a_1, ..., a_p, b$,

$$SQM\left(a_1, ..., a_p, b\right) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \left(a_1 x_{i,1} + ... + a_p x_{i,p} + b\right)\right)^2.$$

La grandezza $SQM\left(a_1, ..., a_p, b\right)$ misura la bontà del modello lineare $Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$: se piccola, il modello è buono.

La strategia, che illustreremo in dettaglio nel paragrafo 1.3.1, sarà di cercare i parametri $a_1, ..., a_p, b$ che rendono minima la grandezza $SQM\left(a_1, ..., a_p, b\right)$. Tali parametri forniscono il migliore tra i modelli lineari. Indichiamo con $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$ i parametri ottimali, che troveremo nel paragrafo 1.3.1. Chiamiamo

$$Y = \widehat{a}_1 X_1 + ... + \widehat{a}_p X_p + \widehat{b} + \varepsilon$$

il *modello di regressione lineare multipla* (*regressione lineare semplice* nel caso $n = 1$) associato alla tabella di dati precedente. La varianza dell'errore, o dei residui, è ora

$$\sigma_\varepsilon^2 = SQM\left(\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}\right) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2$$

dove i residui (ora numeri univocamente determinati dai dati) sono dati da

$$\widehat{\varepsilon}_i = y_i - \left(\widehat{a}_1 x_{1,i} + ... + \widehat{a}_p x_{p,i} + \widehat{b}\right).$$

La *varianza spiegata*, o indice R^2, è definita da

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}$$

dove σ_Y^2 è la varianza dei dati $y_1, ..., y_n$. L'idea è che i dati $y_1, ..., y_n$ hanno una loro variabilità, descritta da σ_Y^2, in situazione di completa ignoranza; ma quando abbiamo a disposizione un modello, esso spiega i dati $y_1, ..., y_n$ in una certa misura, cioè a meno degli errori $\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n$. Quindi la variabilità di questi errori è la variabilità inspiegata, residua. Da qui il nome di (percentuale di) varianza spiegata per il numero $1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}$.

1.3.1 Calcolo dei coefficienti

Il calcolo dei coefficienti si può svolgere in vari modi diversi. Iniziamo con quello più teorico, concettualmente non molto soddisfacente, ma molto veloce e chiaro.

Calcolo a partire da un modello teorico

Consideriamo, come abbiamo fatto nel caso della regressione semplice ($p = 1$), il modello

$$Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$$

dove le X_i, la Y ed ε sono v.a. aleatorie (non dati sperimentali o residui numerici).

Proposizione 7 *Supponiamo che le v.a. $X_1, ..., X_p$, Y ed ε siano legate dalla relazione precedente, ε sia scorrelato con ciascuna X_i ed abbia media nulla, e la matrice di covarianza Q del vettore $(X_1, ..., X_p)$ sia invertibile. Indichiamo con c il vettore di coordinate $c_j = Cov(Y, X_j)$ e con a_0 il vettore dei parametri $(a_1, ..., a_p)$. Allora i parametri $a_1, ..., a_p, b$ sono univocamente determinati dalle relazioni:*

$$\begin{aligned} a_0 &= Q^{-1}c \\ b &= E[Y] - (a_1 E[X_1] + ... + a_p E[X_p]). \end{aligned}$$

Proof. Calcoliamo la covarianza tra Y ed X_j:

$$\begin{aligned} Cov(Y, X_j) &= Cov(a_1 X_1 + ... + a_p X_p + b + \varepsilon, X_j) \\ &= a_1 Cov(X_1, X_j) + ... + a_p Cov(X_p, X_j) + Cov(\varepsilon, X_j). \end{aligned}$$

Usando la matrice $Q = (Cov(X_i, X_j))$ possiamo riscrivere questa indentità come

$$c_j = \sum_{i=1}^{p} Q_{ij} a_i + Cov(\varepsilon, X_j).$$

Avendo ipotizzato che ε sia scorrelato da ciascuna X_j, troviamo

$$c_j = \sum_{i=1}^{p} Q_{ji} a_i$$

(avendo usato che Q è simmetrica) ovvero $Qa_0 = c$. Quindi

$$a_0 = Q^{-1}c.$$

Poi, per calcolare b, basta calcolare il valor medio a destra e sinistra dell'equazione che definisce il modello:

$$E[Y] = a_1 E[X_1] + ... + a_p E[X_p] + b.$$

Allora vale

$$b = E[Y] - (a_1 E[X_1] + ... + a_p E[X_p]).$$

■

Questo risultato fornisce anche un modo per calcolare i parametri a partire da una matrice di dati. Si calcola la matrice di covarianza *empirica* $\widehat{Q}$ della matrice di dati riguardante le variabili X_i, si calcola il vettore $\widehat{c}$ delle covarianze empirche tra le variabili X_j ed Y, e si calcolano i valori

$$\widehat{a}_0 = \widehat{Q}^{-1}\widehat{c}.$$

In modo simile si calcola il valore empirico $\widehat{b}$ a partire dai dati.

Metodo dei minimi quadrati

Questo metodo ha il pregio di partire dai dati, piuttosto che dal modello teorico, quindi la sua giustificazione è più chiara. Ha inoltre il pregio di essere un metodo universale per la ricerca di parametri ottimali (nel senso che con opportune modifiche lo si applica ai più svariati contesti, ad esempio allo studio delle serie storiche che vedremo in un'altra lezione). Richiede però calcoli un po' meno immediati per arrivare alle formule finali.

Come già illustrato in una sezione precedente, il metodo consiste nella ricerca dei valori $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$ che rendono minimo lo scarto quadratico medio

$$SQM\left(a_1, ..., a_p, b\right) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(a_1 x_{i,1} + ... + a_p x_{i,p} + b\right)\right)^2.$$

Per trovare i parametri ottimali si può procedere in due modi. il più naturale ma più laborioso consiste nello studio delle condizioni necessarie di ottimalità, che rimandiamo al paragrafo 1.3.1. Vediamo qui invece un metodo compatto di tipo più geometrico.

Riscriviamo l'equazione (1.2) in forma vettoriale. Introduciamo la matrice degli input

$$X = \begin{pmatrix} x_{1,1} & ... & x_{1,p} & 1 \\ ... & ... & ... & ... \\ x_{n,1} & ... & x_{n,p} & 1 \end{pmatrix}$$

(la colonna di "uni" serve fittiziamente per manipolare l'intercetta in modo unificato agli altri parametri), ed i vettori dei parametri, degli output e dei residui:

$$a = \begin{pmatrix} a_1 \\ ... \\ a_p \\ b \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{pmatrix}$$

ottenendo la riscrittura vettoriale di (1.2):

$$y = Xa + \varepsilon$$

L'errore quadratico medio (come funzione dei parametri $a_1, ..., a_p, b$) si può scrivere

$$SQM\,(a) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\left|\varepsilon\right|^2 = \frac{1}{n}\left|y - Xa\right|^2.$$

Vogliamo minimizzarlo.

In questa forma il problema ha un'interessante interpretazione geometrica. Nello spazio $\mathbb{R}^n$ abbiamo un punto dato, y, ed il sottospazio $\operatorname{Im} X$, l'immagine di $\mathbb{R}^{p+1}$ attraverso X:

$$\operatorname{Im} X = \left\{z : z = Xa, a \in \mathbb{R}^{p+1}\right\}.$$

Cercare il punto $\widehat{a} \in \mathbb{R}^{p+1}$ che minimizza $|y - Xa|$ significa cercare il punto $\widehat{z} \in \operatorname{Im} X$ che minimizza $|y - z|$, dovendo poi come passo finale trovare $\widehat{a} \in \mathbb{R}^{p+1}$ tale che $X\widehat{a} = \widehat{z}$. Operativamente, quindi, si devono risolvere due problemi: una minimizzazione

$$\min_{z \in \operatorname{Im} X} |y - z|$$

e, detto $\widehat{z}$ il punto di minimo, la risoluzione dell'equazione (risolubile perché $\widehat{z} \in \operatorname{Im} X$)

$$Xa = \widehat{z}.$$

Abbiamo sempre detto "il punto..." ma l'unicità degli oggetti che stiamo cercando va anch'essa dimostrata.

Lemma 1 *Dati un punto $y \in \mathbb{R}^n$ ed un sottospazio $V \subset \mathbb{R}^n$, esiste uno ed un solo punto $\widehat{z} \in V$ tale che $|y - \widehat{z}| \leq |y - z|$ per ogni $z \in V$. Il vettore $y - \widehat{z}$ è perpendicolare a V.*

Proof. Sia d la dimensione di V, $d \leq n$. Possiamo costruire una base ortonormale $e_1, ..., e_n$ di $\mathbb{R}^n$ tale che $e_1, ..., e_d$ sia una base di V. I vettori $e_{d+1}, ..., e_n$ sono ortogonali a V. Il punto y si rappresenta univocamente come

$$\begin{aligned} y &= \alpha_1 e_1 + ... + \alpha_n e_n \\ &= \alpha_1 e_1 + ... + \alpha_d e_d + \alpha_{d+1} e_{d+1} + ... + \alpha_n e_n \\ &= \widehat{z} + w \end{aligned}$$

dove $\widehat{z} = \alpha_1 e_1 + ... + \alpha_d e_d \in V$, $w = y - \widehat{z} = \alpha_{d+1} e_{d+1} + ... + \alpha_n e_n$ è perpendicolare a V. Con un po' di calcoli che omettiamo si verifica che $\widehat{z}$ è il punto dato; l'interpretazione grafica è evidente. ■

Lemma 2 *Supponiamo che valga* $\ker X = \{0\}$. *Allora, dato $\widehat{z} \in \operatorname{Im} X$, esiste uno ed un solo punto $\widehat{a} \in \mathbb{R}^{p+1}$ tale che $X\widehat{a} = \widehat{z}$.*

Proof. L'esistenza è insita nell'ipotesi $\widehat{z} \in \operatorname{Im} X$. L'unicità discende dall'ipotesi $\ker X = \{0\}$: se $\widehat{a}, \widetilde{a}$ fossero due soluzioni, allora $X(\widehat{a} - \widetilde{a}) = 0$, quindi $\widehat{a} - \widetilde{a} \in \ker X$, ovvero $\widehat{a} - \widetilde{a} = 0$. ■

Lemma 3 *Se* $\ker X = \{0\}$*, allora* $\det\left(X^T X\right) \neq 0$.

Proof. Se $\ker X = \{0\}$ allora, per ogni vettore $v \neq 0$ vale $Xv \neq 0$, quindi $|Xv|^2 \neq 0$, quindi $\left\langle X^T X v, v \right\rangle \neq 0$. Ma allora $X^T X v \neq 0$ (altrimenti, se fosse $X^T X v = 0$, allora sarebbe anche $\left\langle X^T X v, v \right\rangle = 0$). Abbiamo dimostrato che $v \neq 0$ implica $X^T X v \neq 0$ quindi $\ker\left(X^T X\right) = \{0\}$. Ma $X^T X$ è una matrice quadrata, quindi è non singolare, ovvero $\det\left(X^T X\right) \neq 0$. ■

Teorema 3 *Supponiamo che valga* $\ker X = \{0\}$*. Allora, esiste uno ed un solo vettore* $\widehat{a}$ *che minimizza la funzione* $f(a) = |y - Xa|^2$*. Esso è dato da*

$$\widehat{a} = \left(X^T X\right)^{-1} X^T y.$$

Se inoltre $p + 1 = n$ *allora il minimo è nullo e vale* $\widehat{a} = X^{-1} y$.

Proof. Dato $y \in \mathbb{R}^n$, siano $\widehat{z} \in \operatorname{Im} X$ e $\widehat{a} \in \mathbb{R}^{p+1}$ tali che $\widehat{z}$ minimizza $|y - z|$, $\widehat{a}$ risolve $X\widehat{a} = \widehat{z}$. Preso $a \neq \widehat{a}$, posto $z = Xa$, vale $z \neq \widehat{z}$ e $z \in \operatorname{Im} X$, quindi

$$|y - \widehat{z}| < |y - z|$$

ovvero

$$|y - X\widehat{a}| < |y - Xa|.$$

Abbiamo dimostrato che $f(\widehat{a}) < f(a)$ per ogni $a \neq \widehat{a}$, quindi $\widehat{a}$ è punto di minimo ed è l'unico.

Siccome soddisfa $X\widehat{a} = \widehat{z}$, allora (moltiplicando a sinistra ambo i membri per X^T) soddisfa anche $X^T X\widehat{a} = X^T\widehat{z}$. Per ipotesi e per uno dei lemmi, $X^T X$ è invertibile, quindi $\widehat{a} = \left(X^T X\right)^{-1} X^T\widehat{z}$. Ma

$$\left\langle X^T(\widehat{z} - y), a \right\rangle = \langle \widehat{z} - y, Xa \rangle = 0$$

per ogni a, perché $\widehat{z} - y$ è perpendicolare a $\operatorname{Im} X$. Quindi $X^T(\widehat{z} - y) = 0$, cioè $X^T\widehat{z} = X^T y$. Abbiamo dimostrato che $\widehat{a} = \left(X^T X\right)^{-1} X^T y$.

Infine, se vale anche la condizione $p+1 = n$ (quindi $\det X \neq 0$ perché $\ker X = \{0\}$), le matrici X ed X^T sono invertibili, quindi

$$\widehat{a} = \left(X^T X\right)^{-1} X^T y = X^{-1}\left(X^T\right)^{-1} X^T y = X^{-1} y.$$

La dimostrazione è completa. ■

Osservazione 11 *L'errore relativo ai parametri ottimali*

$$\widehat{\varepsilon} = y - X\widehat{a}$$

è ortogonale a $\operatorname{Im} X$ *(è la proprietà di* $y - \widehat{z} = y - X\widehat{a} = \widehat{\varepsilon}$ *descritta nella dimostrazione). Questo è il riflesso empirico delle proprietà* $Cov(X_i, \varepsilon) = 0$, $i = 1, ..., p$, *e* $E[\varepsilon] = 0$ *alla base del metodo del paragrafo 1.3.1. Vediamo in dettaglio. L'ortogonalità di* $\widehat{\varepsilon}$ *rispetto a* $\operatorname{Im} X$ *si può esprimere con la condizione*

$$X^T\widehat{\varepsilon} = 0$$

(come nella dimostrazione appena svolta, vale $\langle \widehat{\varepsilon}, Xa \rangle = \left\langle X^T\widehat{\varepsilon}, a \right\rangle$ *per ogni* a*, da cui si vede l'equivalenza tra ortogonalità ad* $\operatorname{Im} X$ *e condizione* $X^T\widehat{\varepsilon} = 0$*). La matrice* X^T *è data da*

$$X^T = \begin{pmatrix} x_{1,1} & ... & x_{n,1} \\ ... & ... & ... \\ x_{1,p} & ... & x_{n,p} \\ 1 & ... & 1 \end{pmatrix}$$

per cui, ad esempio, l'ultima componente del vettore $X^T\widehat{\varepsilon}$ *è*

$$\left(X^T\widehat{\varepsilon}\right)_{p+1} = \widehat{\varepsilon}_1 + ... + \widehat{\varepsilon}_n.$$

Quindi $\left(X^T\widehat{\varepsilon}\right)_{p+1} = 0$ *equivale a* $\frac{\widehat{\varepsilon}_1 + ... + \widehat{\varepsilon}_n}{n} = 0$*, che è la versione empirica di* $E[\varepsilon] = 0$*. Analogamente si vede che* $\left(X^T\widehat{\varepsilon}\right)_i = 0$ *è la versione empirica di* $Cov(X_i, \varepsilon) = 0$*, per* $i = 1, ..., p$*.*

Osservazione 12 *In particolare, abbiamo verificato che* $\frac{\widehat{\varepsilon}_1 + ... + \widehat{\varepsilon}_n}{n} = 0$*. Pertanto la varianza empirica* σ_ε^2 *che compare, ad esempio, nella definizione di* R^2 *è data da* SQM*, cioè* $\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2$*, senza bisogno di sottrarre la media empirica, essendo nulla.*

Condizioni di ottimalità

Posto

$$\begin{aligned} f(a_1, ..., a_p, b) &= SQM(a_1, ..., a_p, b) \\ &= \frac{1}{n}\sum_{i=1}^n (y_i - (a_1 x_{i,1} + ... + a_p x_{i,p} + b)) \end{aligned}$$

le condizioni necessarie di ottimalità sono

$$\begin{aligned} &\frac{\partial f}{\partial a_j}\left(\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}\right) = 0, \qquad j = 1, ..., p \\ &\frac{\partial f}{\partial b}\left(\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}\right) = 0 \end{aligned}$$

(con un ragionamento sulla convessità di SQM si può mostrare che sono anche sufficienti, sotto l'ipotesi $\ker X = \{0\}$). Se si ha dimestichezza col calcolo vettoriale si arriva subito al risultato finale. Infatti, con le notazioni vettoriali del paragrafo precedente,

$$f(a) = \frac{1}{n} |y - Xa|^2$$

da cui, per ogni direzione $h \to \mathbb{R}^{p+1}$,

$$\begin{aligned} \langle \nabla f(a), h \rangle &= \frac{2}{n} \langle y - Xa, Xh \rangle \\ &= \frac{2}{n} \left\langle X^T (y - Xa), h \right\rangle \end{aligned}$$

ovvero $\nabla f(a) = \frac{2}{n} X^T (y - Xa)$. Le condizioni di ottimalità diventano quindi

$$X^T (y - X\widehat{a}) = 0$$

ovvero

$$X^T X \widehat{a} = X^T y$$

come nel paragrafo precedente. Bisognerebbe poi giustificare che il minimo è unico.

Per eventuale maggior chiarezza, svolgiamo i conti, però parecchio laboriosi, senza notazione vettoriale. Vale

$$\begin{aligned} \frac{\partial f}{\partial a_j} &= -\frac{2}{n} \sum_{i=1}^{n} (y_i - (a_1 x_{i,1} + ... + a_p x_{i,p} + b)) x_{i,j} \\ &= -\frac{2}{n} \langle y, x_j \rangle + \frac{2}{n} a_1 \langle x_{\cdot,1}, x_{\cdot,j} \rangle + ... + \frac{2}{n} a_p \langle x_{\cdot,p}, x_{\cdot,j} \rangle + 2b\overline{x}_j \end{aligned}$$

$$\begin{aligned} \frac{\partial f}{\partial b} &= -\frac{2}{n} \sum_{i=1}^{n} (y_i - (a_1 x_{1,i} + ... + a_p x_{p,i} + b)) \\ &= -2\overline{y} + 2a_1 \overline{x}_1 + ... + 2a_p \overline{x}_p + 2b \end{aligned}$$

dove $\overline{y}$ è la media degli $y_1, ..., y_n$ e, per ciascun $j = 1, ..., p$, $\overline{x}_j$ è la media degli $x_{1,j}, ..., x_{n,j}$; ed inoltre abbiamo posto

$$\langle y, x_{\cdot,j} \rangle = \sum_{i=1}^{n} y_i x_{i,j}, \qquad \langle x_{\cdot,k}, x_{\cdot,j} \rangle = \sum_{i=1}^{n} x_{i,k} x_{i,j}.$$

Quindi deve valere

$$\begin{gathered} a_1 \langle x_{\cdot,1}, x_{\cdot,j} \rangle + ... + a_p \langle x_{\cdot,p}, x_{\cdot,j} \rangle + nb\overline{x}_j = \langle y, x_{\cdot,j} \rangle, \qquad i = 1, ..., p \\ a_1 \overline{x}_1 + ... + a_p \overline{x}_p + b = \overline{y} \end{gathered}$$

Questo è un sistema di $p+1$ equazioni lineari in $p+1$ incognite.

Si può a questo punto introdurre la matrice A quadrata a $p+1$ righe e colonne ed il vettore w

$$A = \begin{pmatrix} \langle x_{\cdot,1}, x_{\cdot,1}\rangle & \ldots & \langle x_{\cdot,p}, x_{\cdot,1}\rangle & \langle x_{\cdot,1}, 1\rangle \\ \ldots & \ldots & \ldots & \ldots \\ \langle x_{\cdot,1}, x_{\cdot,p}\rangle & \ldots & \langle x_{\cdot,p}, x_{\cdot,p}\rangle & \langle x_{\cdot,p}, 1\rangle \\ \langle x_{\cdot,1}, 1\rangle & \ldots & \langle x_{\cdot,p}, 1\rangle & 1 \end{pmatrix}, \qquad w = \begin{pmatrix} \langle y, x_{\cdot,1}\rangle \\ \ldots \\ \langle y, x_{\cdot,p}\rangle \\ \langle y, 1\rangle \end{pmatrix}$$

dove 1 indica il vettore con tutti "1", ed y è il vettore delle y_i. Allora il calcolo del vettore

$$\widehat{a} = \begin{pmatrix} \widehat{a}_1 \\ \ldots \\ \widehat{a}_p \\ \widehat{b} \end{pmatrix}$$

si può vedere come la risoluzione di

$$A\widehat{a} = w.$$

Infine, possiamo riconoscere che la matrice A si ottiene col prodotto

$$A = X^T X$$

dove

$$X = \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} & 1 \\ \ldots & \ldots & \ldots & \ldots \\ x_{n,1} & \ldots & x_{n,p} & 1 \end{pmatrix}$$

è la matrice già introdotta nella sezione precedente. Inoltre,

$$w = X^T y.$$

Quindi il problema $Av = w$ ha la forma

$$X^T X\widehat{a} = X^T y.$$

La sua soluzione è quella trovata nella sezione precedente.

1.3.2 Domande

1. Com'è fatta una tabella di dati in un problema di regressione multipla
2. La regressione stabilisce un legame causa-effetto? Può essere usata per fare previsioni?

3. Scrivere un modello di regressione lineare multipla, sia nella versione teorica di legame tra variabili aleatorie, sia in quella più pratica di legame tra dati sperimentali a meno di errore

4. Definizione di residui e formulazione del metodo dei minimi quadrati

5. Formule per i coefficienti della regressione multipla (enunciato e dimostrazione del Teorema 3; possibilmente anche un altro approccio)

6. Descrivere graficamente (geometricamente) il procedimendo di ricerca dei parametri ottimali adottato dal Teorema 3.

7. I residui, dipendono dai parametri oppure no?

1.4 Osservazioni sulla regressione lineare multipla

Overfitting

Quando $p+1=n$ e $\det X \neq 0$, i residui relativi ai parametri $\widehat{a}$ sono nulli (il minimo di SQM è zero). Da ciò si deduce anche che la varianza spiegata R^2 è 1. Sembrerebbe quindi che un tale modello sia il migliore possibile. Invece per le usuali applicazioni non è così. Il suo potere previsivo può essere scarsissimo, così come la sua capacità di aver capito il problema, i dati. Infatti, un tale modello si adatta perfettamente ai dati sperimentali seguendo al millimetro i suoi accidenti casuali, senza filtrarli, senza riconoscere che alcune variazioni sono rumore, non struttura da catturare. Si può capire bene questo pensando al caso di due variabili X, Y, di cui si conoscano solo due dati sperimentali $(1, y_1)$, $(2, y_2)$, tali che y_1 e y_2 sono numeri casuali $N(0,1)$. Il modello corretto sarebbe

$$Y = \varepsilon$$

(cioè $a = b = 0$), mentre se obblighiamo il software a cercare un modello del tipo $Y = aX + b + \varepsilon$ ($p = 1$, $p+1 = 2$), esso fitterà perfettamente i due dati trovando la retta che passa per i due punti $(1, y_1)$, $(2, y_2)$: residui nulli, ma modello dato da una retta senza senso, completamente sbagliata rispetto a $Y = \varepsilon$, ed oltretutto fortemente dipendente dai valori casuali y_1 e y_2. Quando p è troppo grande rispetto al numeri di dati a disposizione, si parla di *overfitting*.

Variabilità statistica dei parametri ottimali

Per capire l'enunciato (o lo spirito) del seguente teorema bisogna immaginare la seguente situazione sperimentale (in realtà questa immaginazione poteva essere descritta sin dall'inizio del capitolo, come la base logica del nostro problema di stima dei parametri). Svolgiamo degli esperimenti che coinvolgono le grandezze $X_1, ..., X_p$ ed Y. Le grandezze

$X_1, ..., X_p$ sono sotto il nostro preciso controllo: possiamo fissarne i valori $x_{i,j}$ (ovvero la matrice X), quando effettuiamo gli esperimenti. Invece non possiamo fissare i valori della grandezza Y, ma li possiamo osservare e registrare. Supponiamo infine che le grandezze siano legate dalla relazione $Y = \overline{a}_1 X_1 + ... + \overline{a}_p X_p + \overline{b} + \varepsilon$, secondo ben precisi valori $\overline{a}_1, ..., \overline{a}_p, \overline{b}$ che però non conosciamo; e dove ε è un disturbo casuale di media nulla e deviazione standard σ_ε.

Effettuiamo n esperimenti, ottenendo dei valori $y_1, ..., y_n$, che registriamo. Questi valori sono dati dalla formula

$$y_i = \overline{a}_1 x_{i,1} + ... + \overline{a}_p x_{i,p} + \overline{b} + \varepsilon_i$$

dove $\varepsilon_1, ..., \varepsilon_n$ sono le realizzazioni casuali dell'errore ε accadute in tali esperimenti; ma ricordiamo che noi non conosciamo i parametri di questa formula. Si noti che l'unico aspetto causale, imprevedibile, in ciascun esperimento è proprio il valore che assume ε.

A questo punto, partendo dai dati noti $x_{i,j}$ più i dati misurati y_i, calcoliamo il vettore dei parametri ottimali $\widehat{a}$, secondo il metodo dei minimi quadrati descritto sopra. Il valore ottenuto per $\widehat{a}$ non coincide col valore esatto $\overline{a} = \left(\overline{a}_1, ..., \overline{a}_p, \overline{b}\right)$ (sconosciuto). Ne è una *stima*. Per inciso, questa filosofia (di avere un modello con parametri esatti ma incogniti ed un procedimento di stima approssimata di tali parametri) è quella generale di tutti i problemi di stima dei parametri, come la stima della media e della varianza di una grandezza aleatoria gaussiana partendo da un campione sperimentale.

La mancata coincidenza di $\widehat{a}$ con $\overline{a}$, o se si preferisce la variabilità del risultato $\widehat{a}$, dipendono dalla casualità di ε.

La domanda è: che proprietà statistiche ha $\widehat{a}$? E' uno *stimatore corretto* (*non distorto*), cioè tale che $E\left[\widehat{a}\right] = \overline{a}$? La variabilità di $\widehat{a}$ (cioè $Var\left[\widehat{a}\right]$) è grande? Qui, coi simboli $E\left[\widehat{a}\right]$ e $Var\left[\widehat{a}\right]$, si sta pensando a medie rispetto alla casualità di ε: siccome ε è aleatorio, anche gli output y misurati sono aleatori, quindi anche la stima ottimale $\widehat{a}$ è aletoria.

Non comportando maggiori difficoltà, impostiamo la casualità dell'errore in modo persino un po' più generale: supponiamo che gli errori, casuali, degli n esperimenti, cioè le v.a. $\varepsilon_1, ..., \varepsilon_n$, abbiano matrice di covarianza Q_ε. E' poi più che sufficiente pensare al caso base in cui essi sono indipendenti ed ugualmente distribuiti, caso in cui Q_ε è diagonale, multiplo dell'identità: $Q_\varepsilon = \sigma_\varepsilon^2 I$.

Teorema 4 *Lo stimatore $\widehat{a}$ è non distorto:*

$$E\left[\widehat{a}\right] = \overline{a}$$

e la matrice di covarianza di $\widehat{a}$ è

$$Q_{\widehat{a}} = \left[\left(X^T X\right)^{-1} X^T\right] Q_\varepsilon \left[X \left(X^T X\right)^{-1}\right].$$

In particolare, se $Q_\varepsilon = \sigma_\varepsilon^2 Id$, cioè se gli errori ε_i sono indipendenti tra loro e tutti di varianza σ_ε^2, allora

$$Q_{\widehat{a}} = \sigma_\varepsilon^2 \left(X^T X\right)^{-1}.$$

Proof. Con notazioni vettoriali, abbiamo $y = X\overline{a} + \varepsilon$ dove X ed $\overline{a}$ sono dati mentre ε è un vettore aleatorio, centrato di covarianza Q_ε. Siccome $\widehat{a} = \left(X^T X\right)^{-1} X^T y$, vale

$$\begin{aligned} \widehat{a} &= \left(X^T X\right)^{-1} X^T X\overline{a} + \left(X^T X\right)^{-1} X^T \varepsilon \\ &= \overline{a} + \left(X^T X\right)^{-1} X^T \varepsilon. \end{aligned}$$

Conosciamo le regole di trasformazione di media e covarianza per trasformazioni lineari: $E\left[\widehat{a}\right] = \overline{a}$ e

$$Q_{\widehat{a}} = \left[\left(X^T X\right)^{-1} X^T\right] Q_\varepsilon \left[\left(X^T X\right)^{-1} X^T\right]^T = \left[\left(X^T X\right)^{-1} X^T\right] Q_\varepsilon \left[X\left(X^T X\right)^{-1}\right].$$

Se $Q_\varepsilon = \sigma_\varepsilon^2 Id$, allora

$$Q_{\widehat{a}} = \sigma_\varepsilon^2 \left[\left(X^T X\right)^{-1} X^T\right] Id \left[X\left(X^T X\right)^{-1}\right] = \sigma_\varepsilon^2 \left(X^T X\right)^{-1}.$$

La dimostrazione è completa. ■

Allineamento tra fattori

Due fattori X_i, X_j si dicono allineati se sono fortemente correlati. E' una definizione vaga, di natura applicativa o pratica, perché non si stabilisce con precisione quanto debbano essere correlati. Il caso limite sarebbe quello in cui $\rho\left(X_i, X_j\right) = \pm 1$, che fornirebbe una definizione rigorosa, inutile però perché priva di applicazioni. Accontentiamoci quindi di parlare di allineamento ogni volta che ci sembra elevata la correlazione.

Siccome si tratta di un concetto pratico, lo si riferisce ai dati sperimentali, quindi si sta allundendo alla correlazione $\widehat{\rho}$ tra le colonne i e j della matrice di dati. Il caso $|\widehat{\rho}| = 1$ significa che le colonne coincidono, a meno del segno o di un fattore moltiplicativo, cioè sono proporzionali. Elevato $|\widehat{\rho}|$ significa che le colonne sono quasi proporzionali, pur non esattamente proporzionali. Se due colonne della matrice X fossero esattamente proporzionali, $X^T X$ sarebbe degenere ed il suo determinante nullo. Se, più realisticamente, due colonne della matrice X sono quasi proporzionali, $X^T X$ è quasi degenere. Cosa significa questo in termini di coefficienti della matrice $\left(X^T X\right)^{-1}$ che compare nelle due formule

$$\begin{aligned} \widehat{a} &= \left(X^T X\right)^{-1} X^T y \\ Q_{\widehat{a}} &= \sigma_\varepsilon^2 \left(X^T X\right)^{-1} \end{aligned}$$

dimostrate ai paragrafi precedenti?

Distinguiamo due casi. Il primo riguarda il caso in cui i numeri di partenza (quelli della tabella dei dati) sono grosso modo unitari, magari addirittura standardizzati. In

questo caso, X^TX quasi degenere significa che il suo determinante è quasi nullo (rispetto al numero 1). La matrice inversa, $\left(X^TX\right)^{-1}$ pur ben definita, avrà determinante molto elevato ed anche alcuni valori molto elevati. Questo provoca una serie di probemi che discuteremo tra un momento.

Il secondo caso è quello in cui, magari a causa di una specifica scelta della scala, dell'unità di misura, la tabella dei dati di partenza contiene elementi di grandezza non unitaria. Se c'è un allienamento, ancora in un certo senso X^TX è quasi degenere ma la quantificazione numerica di questa frase dipenderà dalla grandezza dei numeri in gioco. Questo secondo caso è un po' più complicato da capire, per cui lo rimandiamo alla fine del paragrafo. Supponiamo quindi, nella prima parte del paragrafo, che i dati di partenza siano standardizzati o comunque di grandezza grosso modo unitaria.

In questo caso base, come abbiamo già detto, alcuni valori della matrice $\left(X^TX\right)^{-1}$ saranno molto elevati (in caso contrario, il determinante di $\left(X^TX\right)^{-1}$ avrebbe un valore moderato e quindi $\det\left(X^TX\right)$ non sarebbe quasi nullo).

Dalla formula $\widehat{a} = \left(X^TX\right)^{-1}X^Ty$ si vede allora che, in caso di allineamenti tra fattori, i coefficienti stimati $\widehat{a}$ possono avere valori molto elevati. E dalla formula $Q_{\widehat{a}} = \sigma_\varepsilon^2\left(X^TX\right)^{-1}$ si vede che, in caso di allineamenti tra fattori, i coefficienti stimati $\widehat{a}$ possono essere molto instabili, possono variare molto in conseguenza di piccoli errori.

La grandezza di certi valori di $\widehat{a}$ può essere quindi del tutto fittizia, non corrispondere al fatto che l'influenza dei corrispondenti fattori è molto intensa. In caso di allineamento, alcuni parametri stimati possono essere molto grandi ed esserlo in modo causale, fortemente variabile.

Osservazione 13 *Si può essere un po' più precisi circa* ***quali*** *elementi di* $\left(X^TX\right)^{-1}$ *saranno plausibilmente molto grandi. Per far questo bisogna ricordare che, data una matrice quadrata* B *invertibile, l'elemento generico della matrice inversa* B^{-1} *è dato da*

$$\left(B^{-1}\right)_{ij} = \frac{(-1)^{i+j}}{\det(B)}\det\left(B_{ji}\right)$$

dove B_{ji} *è la sotto-matrice di* B *ottenuta rimuovendo la riga* j*-esima e la colonna* i*-esima. Se (sempre nell'ipotesi di dati di ordine di grandezza 1) la riga* j *o la colonna* i *sono gemelle di un'altra nel senso dell'allineamento detto sopra, la loro rimozione elimina la singolarità della matrice* X^TX*, o per lo meno l'attenua (potrebbero esserci altri fattori allineati oltre a quello che stiamo rimuovendo). Quindi* $\det\left(\left(X^TX\right)_{ji}\right)$ *non è così piccolo come* $\det\left(X^TX\right)$*, ovvero il rapporto* $\frac{\det\left(\left(X^TX\right)_{ji}\right)}{\det(X^TX)}$ *è (o potrebbe essere) grande. In altre parole,* $\left(\left(X^TX\right)^{-1}\right)_{ij}$ *è grande, più del normale (cioè più degli altri termini della matrice inversa). In conclusione, possono essere grandi gli elementi di* $\left(X^TX\right)^{-1}$ *che hanno indici in comune con i fattori allineati.*

In caso di allineamento si originano fenomeni strani come quello di coefficienti di segno opposto alla correlazione tra fattore e output. La ragione è sempre la stessa descritta sopra, numeri alti e fortemente variabili, ma lo si può anche intuire dal seguente ragionamento più pratico. Supponiamo di avere una grandezza Y che dipende da X_1 e X_2, con buona correlazione con entrambe: $\rho_{Y,X_1} \sim 0.8$, $\rho_{Y,X_2} \sim 0.8$ (per esempio). Ma supponiamo anche che X_1 e X_2 siano molto correlate, ad es. $\rho_{X_1,X_2} \sim 0.9$. Se usassimo solo X_1 come predittore, troveremmo una buona relazione lineare, ad esempio

$$Y = 3.4 \cdot X_1 - 6.2 + \varepsilon.$$

Se usassimo solo X_2 come predittore, troveremmo una relazione lineare, molto simile, perché X_1 e X_2 sono quasi la copia una dell'altra: ad es.

$$Y = 3.2 \cdot X_2 - 6.1 + \varepsilon.$$

Volendo usare entrambe le variabili X_1 e X_2 nel modello, il metodo dei minimi quadrati applicato a dati sperimentali potrebbe ad esempio produrre

$$Y = 9.8 \cdot X_1 - 6.3 \cdot X_2 - 6.3 + \varepsilon. \tag{1.3}$$

Infatti, essendo $X_2 \sim X_1$, grossolanamente parlando abbiamo

$$9.8 \cdot X_1 - 6.3 \cdot X_2 \sim 9.8 \cdot X_1 - 6.3 \cdot X_1 = 3.5 \cdot X_1.$$

Ma il modello (1.3) è insensato: 9.8 non è la misura giusta dell'influenza di X_1 su Y, e tantomeno -6.3 che è addirittura negativo, contrariamente a $\rho_{Y,X_2} > 0$.

Quanto descritto in questo esempio immaginario è la norma, non l'eccezione, se ci sono fattori allineati. Inoltre, se si ripete la regressione su nuovi dati (dello stesso problema applicativo), essendo questi dati affetti da nuovi errori, producono stime diverse $\widehat{a}$, magari molto differenti tra loro. Potremmo trovare il modello:

$$Y = -12.1 \cdot X_1 + 15.4 \cdot X_2 - 6.2 + \varepsilon$$

di nuovo plausibile perché

$$-12.1 \cdot X_1 + 15.4 \cdot X_2 \sim -12.1 \cdot X_1 + 15.4 \cdot X_1 = 3.3 \cdot X_1.$$

In altre parole, i coefficienti stimati non hanno la naturale interpretazione che ci si aspetterebbe, sono molto casuali; e probabilmente molto più grossi di quelli giusti.

Quando si individuano due fattori allineati, può convenire eliminarne uno dei due. Il vantaggio è ottenere un modello (più economico e) con coefficenti sensati. Per certi versi si butta via una ricchezza di dati, quindi tale eliminazione va valutata con cura. Però, se i coefficienti, prodotti dal metodo di minimi quadrati, sono assurdi, poi queste assurdità si ripercuoteranno sull'utilizzo del modello ad esempio in previsioni future. Un coefficiente assurdamente grosso amplificherà le imprecisioni dei valori delle X quando faremo previsioni col modello.

Ovviamente la difficoltà nella pratica è decidere se sia maggiore il danno prodotto da tali coefficienti strani o il danno dell'eliminazione di fattori e dati importanti.

Osservazione 14 *Discutiamo infine il caso in cui la tabella dei dati di partenza contenga elementi di grandezza non unitaria. Se questo è dovuto ad una disomogeneità dell'ordine di grandezza dei dati (alcuni unitari, altri no), si può concludere poco, perché ogni ragionamento è viziato da tale disomogeneità. Ad esempio, possono esserci elementi di ampiezza anomala in* $\left(X^T X\right)^{-1}$ *anche senza allineamenti, solo per effetto dei valori anomali di* X*.*

Osservazione 15 *Se invece il problema è di scala, cioè* X *ha elementi che differiscono moltiplicativamente di un fattore di scala* λ *da valori circa unitari, siamo nella situazione in cui*

$$X = \lambda X^{(u)}$$

dove $X^{(u)}$ *è una matrice di dati di scala unitaria. Quindi*

$$X^T X = \lambda^2 \left(X^{(u)T} X^{(u)}\right), \qquad \left(X^T X\right)^{-1} = \lambda^{-2} \left(X^{(u)T} X^{(u)}\right)^{-1}.$$

Alla matrice $X^{(u)}$ *possiamo applicare i ragionamenti detti sopra: l'allineamento tra due fattori, cioè l'elevato coefficiente di correlazione, provoca che* $\det\left(X^{(u)T} X^{(u)}\right)$ *sia quasi zero ed alcuni elementi di* $\left(X^{(u)T} X^{(u)}\right)^{-1}$ *(in particolare quelli con indici corrispondenti ai fattori allineati) possano essere molto più grandi del normale. Dall'identità*

$$\left(X^T X\right)^{-1}_{ij} = \lambda^{-2} \left(X^{(u)T} X^{(u)}\right)^{-1}_{ij}$$

deduciamo che i corrispondenti elementi $\left(X^T X\right)^{-1}_{ij}$ *possono essere molto più grandi degli altri. Si ha cioè un'anomalia di* ***grandezza relativa****. In senso assoluto, a causa del fattore* λ^{-2}*, gli elementi di* $\left(X^T X\right)^{-1}$ *possono essere tutti enormi o tutti piccoli, rispetto ad 1. Ma in senso relativo, gli uni rispetto agli altri, si ripresenta il fenomeno di anomalia descritto sopra.*

Osservazione 16 *Quanto detto riguarda però gli elementi della matrice* $\left(X^T X\right)^{-1}$*. Il discorso cambia (in positivo) se si considerano le due formule*

$$\begin{aligned} \widehat{a} &= \left(X^T X\right)^{-1} X^T y \\ Q_{\widehat{a}} &= \sigma^2_\varepsilon \left(X^T X\right)^{-1}. \end{aligned}$$

Se siamo nel caso detto poco fa in cui $X = \lambda X^{(u)}$*, plausibilmente avremo anche* $y = \lambda y^{(u)}$ *e* $\varepsilon = \lambda \varepsilon^{(u)}$ *(la scelta che avremo fatto dell'unità di misura sarà stata adottata anche per le altre variabili del modello). Ma allora*

$$\begin{aligned} \left(X^T X\right)^{-1} X^T y &= \left(X^{(u)T} X^{(u)}\right)^{-1} X^{(u)T} y^{(u)} \\ \sigma^2_\varepsilon \left(X^T X\right)^{-1} &= \sigma^2_{\varepsilon^{(u)}} \left(X^{(u)T} X^{(u)}\right)^{-1} \end{aligned}$$

ovvero $\widehat{a}$ e $Q_{\widehat{a}}$ sono invarianti per cambiamento di scala. Quindi i ragionamenti circa il fatto che alcuni elementi della matrice $\left(X^{(u)T}X^{(u)}\right)^{-1}$ possono essere molto grandi si traducono in analoghi ragionamenti per $\widehat{a}$ e $Q_{\widehat{a}}$, senza bisogno di dire che per grandezza si intende la grandezza relativa.

Osservazione 17 *Purtroppo quello che abbiamo appena descritto è giusto solo parzialmente per il fatto che la colonna di "uni" della matrice X non si conserva per cambio di scala: nell'identità $X = \lambda X^{(u)}$ non può esserci tale colonna in entrambe le matrici X ed $X^{(u)}$. Il problema può essere superato introducendo la matrice $\widetilde{X}$ uguale ad X salvo che per la colonna di "uni" e sviluppando tutta la teoria in termini di $\widetilde{X}$. Questo però è lungo e meno elegante della teoria che abbiamo descritto, e servirebbe solo per correggere l'errore ora segnalato. Trascuriamo quindi questo particolare e prendiamo le osservazioni precedenti come solo approssimativamente corrette.*

Uso del software e considerazioni pratiche

Descriviamo un esempio eseguito col software `R`. Riprendiamo la tabella della Sezione 1.2.2, in cui le variabili erano PLIC, SC, SA.SC, TD, TMI. In quella sezione, erano tutte sullo stesso piano. Qui, immaginiamo di cercare una relazione del tipo

$$\texttt{TD} = \mathtt{a}_1 \cdot \texttt{PLIC} + \mathtt{a}_2 \cdot \texttt{SC} + \mathtt{a}_3 \cdot \texttt{SA.SC} + \mathtt{a}_4 \cdot \texttt{TMI} + \mathtt{b} + \varepsilon$$

(si cerca di spiegare il tasso di disoccupazione tramite alcune variabili economiche e sociali). Costruiamo i vettori con le singole variabili:

```
PLIC<-A[,1]; SC<-A[,2]; SA.SC<-A[,3]; TD<-A[,4]; TMI<-A[,5]; Nord<-A[,6]
```

Poi eseguiamo la regressione:

```
reg<- lm(TD ~PLIC+SC+SA.SC+TMI)
```

Chiediamo infine informazioni sull'esito con comando `summary(reg)`. Esso fornisce una tabella di cui riportiamo qui una sintesi:

	Estimate	Pr(>|t|)	
(Intercept)	1.065e-01	1.00000	
PLIC	6.308e-04	0.99576	
SC	-3.006e-01	0.13320	
SA.SC	6.481e-01	0.00496	**
TMI	8.899e-03	0.94400	
Multiple R-squared:	0.8465	Adjusted R-squared:	0.8055
p-value:	5.793e-06		

La prima colonna riporta semplicemente i nomi delle variabili in input, inclusa l'intercetta. La seconda colonna fornisce la stima $\widehat{a}$ dei parametri; nella tabella completa qui non riportata, a fianco ci sono informazioni sulla precisione di tale stima. Nella

terza colonna c'è il p-value dei singoli fattori, di cui parleremo in un prossimo paragrafo. Da un punto di vista "fenomenologico", se il p-value di un fattore è alto (ad es. 0.99576 è sicuramente molto alto), quel fattore è poco rilevante nel modello; se invece il p-value è basso (ad es. 0.00496 è molto basso, mentre 0.13320 è tendente al basso ma non così tanto), quel fattore è molto rilevante nel modello. Nel nostro esempio, SA.SC è la grandezza che più spiega TD, con un po' di aiuto da parte di SC; invece PLIC e TMI possono essere tolte dal modello, sono del tutto irrilevanti. La colonna successiva enfatizza questi fatti tramite asterischi.

Più sotto, viene dato il valore di $R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}$, la varianza spiegata, che già sappiamo essere una grandezza importante. A fianco, c'è la *varianza spiegata corretta* (adjusted) che è una modifica di R^2 che tiene conto del numero di fattori, tramite la formula

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right).$$

L'idea di questa grandezza è che, pur essendo simile ad R^2, diminuisce se il numero di fattori è troppo grande (rispetto al numero di unità sperimentali). Dietro i modelli c'è sempre il desiderio che siano economici, sia per varie ragioni di semplicità e praticità, sia perché potrebbero essere migliori: un modello dipendente da tantissime grandezze, in cui tale dipendenza sia stata stabilita sulla base di un certo set di dati sperimentali, si espone al rischio che molte di queste dipendenze siano fittizie, casualità di quei particolari dati sperimentali. Ne abbiamo anche già parlato sopra a proposito dell'overfitting. Si vorrebbe catturare la "fisica" del problema, l'essenza, non rincorrere particolarità dei dati che non corrispondono a fatti reali, ma sono solo dovute al caso. Con poche variabili, questo rischio è più contenuto. Detto questo, R^2_{adj} premia contemporaneamente una buona varianza spiegata ed un buon grado di economia di variabili.

Infine, più sotto, viene dato il p-value globale dell'intero modello (anche di esso diremo qualcosa in un prossimo paragrafo).

Il significato del valore numerico dei parametri

Supponiamo di aver eseguito una regressione ed aver trovato i valori numerici dei parametri ottimali $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$. Supponiamo ad esempio che valga $\widehat{a}_1 = 3.5$. Questo valore ha un'interpretazione? Ricordiamo che nel caso $p = 1$ (regressione semplice) esso è il coefficiente angolare della retta di regressione, quindi ha un'interpretazione geometrica. In teoria lo stesso vale per la regressione multipla: la retta andrebbe sostituita con un iperpiano; però la visualizzazione non è più efficace e l'interpretazione geometrica si perde.

Numericamente, $\widehat{a}_1 = 3.5$ significa che, se incrementiamo il valore di X_1 di un'unità, il valore di Y aumenta di 3.5 unità: se

$$y = \widehat{a}_1 x_1 + ... + \widehat{a}_p x_p + \widehat{b} + \varepsilon$$

allora

$$y + \widehat{a}_1 = \widehat{a}_1 (x_1 + 1) + ... + \widehat{a}_p x_p + \widehat{b} + \varepsilon.$$

Quindi il valore numerico dei parametri fornisce la misura della variazione di Y in corrispondenza di variazioni unitarie dei fattori.

Oltre a questo, la cosa basilare è il segno dei parametri: se $\widehat{a}_1 > 0$, allora Y cresce al crescere di X_1; se $\widehat{a}_1 > 0$, allora Y decresce, al crescere di X_1. Questo permette ragionamenti interpretativi, sul fatto che certe grandezze varino in modo concorde o discorde. Lo stesso risultato però si ottiene guardando la matrice di correlazione. Anzi, come discusso nel paragrafo degli allineamenti, i parametri della regressione possono diventare ingannevoli, sia come segno sia come entità, in caso di allineamento.

I valori così piccoli dei parametri nell'esempio del paragrafo precedente sono dovuti al fatto che i dati sono standardizzati.

Non si deve pensare che ci sia una relazione semplice tra il parametro $\widehat{a}_1$ ed il coefficiente di correlazione $\widehat{\rho}(X_1, Y)$, come avviene invece per la regressione lineare semplice. La formula $\widehat{a} = \left(X^T X\right)^{-1} X^T y$ indica che tutti i coefficienti di correlazione $\widehat{\rho}(X_i, Y)$, $i = 1, 2, ..., p$, dati dal vettore $X^T y$, entrano in gioco nel calcolo di $\widehat{a}_1$. Se la matrice $\left(X^T X\right)^{-1}$ fosse diagonale, della forma

$$\left(X^T X\right)^{-1} = \begin{pmatrix} \sigma_1^{-2} & 0 & ... & 0 \\ 0 & \sigma_2^{-2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & \sigma_{p+1}^{-2} \end{pmatrix}$$

allora avremmo

$$\widehat{a}_1 = \frac{\widehat{\rho}(X_1, Y)}{\sigma_1^2}$$

e così via per gli altri parametri ottimali. Ma questo è un caso molto particolare. Questo dimostra che la regressione lineare multipla mescola in modo complesso l'interazione tra tutti i fattori e l'output, non a due a due come la matrice di correlazione o il plot di una tabella.

Esaminiamo però un caso ideale. Supponiamo che le v.a. $X_1, ..., X_p$ siano indipendenti. Supponiamo di avere una tabella di dati sperimentali relativi a queste variabili e standarizziamo la tabella (così che le matrici di covarianza e correlazione empirica coincidano). Supponiamo che la tabella sia così tipica che tali matrici siano diagonali:

$$\widehat{Q} = I$$

(gli zeri fuori dalla diagonale sono il riflesso dell'indipendenza, e su questo stiamo idealizzando la realtà perché le correlazioni empriche non sono mai nulle; gli uni sulla diagonale sono conseguenza esatta della standardizzazione della tabella). Allora $X^T X = I$, cioè

$$\widehat{a} = X^T y.$$

In questo caso, $\widehat{a}_i = \widehat{\rho}(X_i, Y)$ per ogni $i = 1, ..., p$.

p-value dei singoli fattori

Per capire pienamente questa sezione si suggerisce di rivedere il concetto di p-value nell'ambito della statistica di base, ad esempio nel test per la media di una gaussiana. Qui ricordiamo solo che è un concetto della teoria dei test statistici, può essere definito in almeno due modi, uno dei quali è il punto di demarcazione tra i valori della significatività α (o livello del test) per cui il test risulta significativo oppure no. La definizione intuitivamente più comoda è la seguente: il p-value è la probabilità che una determinata grandezza statistica (intesa in senso teorico, come variabile aleatoria, e presa con la distribuzione valida sotto l'ipotesi nulla) sia più estrema del valore sperimentale osservato. Ad esempio, nel test per la media μ di una gaussiana X di varianza nota σ^2, dove l'ipotesi nulla è che la media sia pari ad un certo valore μ_0, e dove si usa come grandezza statistica con cui eseguire il test la v.a.

$$Z = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

che, sotto l'ipotesi nulla, è una gaussiana standard, il p-value è la probabilità che Z sia più estrema del valore sperimentale osservato, diciamo $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$. Poi, la frase "più estrema" può a sua volta essere intesa in vari modi, a seconda che si preferisca eseguire il test bilaterale, o quello unilatrale destro, o unilatrale sinistro. Nei tre casi, il p-value è

$$P\left(|Z| > |z|\right), \qquad P\left(Z > z\right), \qquad P\left(Z < z\right).$$

Se un p-value è piccolo, significa che, sotto l'ipotesi nulla, il valore z era improbabile; siccome z è il valore osservato, siamo spinti a credere che non valesse l'ipotesi nulla. Da qui nasce la regola "fenomenologica" che se il p-value è piccolo, riteniamo falsa l'ipotesi nulla. Se invece non è abbastanza piccolo, non possiamo ritenerla falsa. In pratica, per darsi un'unità di misura (ma questo è del tutto soggettivo), si pensa che i p-value piccoli siano quelli inferiori a 0.05. Quando un p-value è maggiore di 0.05, l'ipotesi nulla potrebbe essere valida.

Fatte queste premesse, veniamo al concetto di p-value relativo al singolo fattore X_i. L'ipotesi nulla è $a_i = 0$. La grandezza statistica scelta per mettere in discussione questa ipotesi è sostanzialmente $\widehat{a}_i$. Ma $\widehat{a}_i$, analoga ad $\overline{X}$ nel test per la media ricordato sopra, risentirebbe ad esempio dell'unità di misura, per cui conviene standardizzarla (anche se oggi giorno si potrebbe adottare un metodo Monte Carlo per il calcolo del p-value). Conviene quindi considerare $\frac{\widehat{a}_i - E[\widehat{a}_i]}{\sigma_{\widehat{a}_i}}$, dove però $E\left[\widehat{a}_i\right] = a_i$ e $a_i = 0$ sotto l'ipotesi nulla, quindi $\frac{\widehat{a}_i - E[\widehat{a}_i]}{\sigma_{\widehat{a}_i}} = \frac{\widehat{a}_i}{\sigma_{\widehat{a}_i}}$. Usando poi il teorema precedente, nell'ipotesi più specifica su ε, vale

$$\frac{\widehat{a}_i}{\sigma_{\widehat{a}_i}} = \frac{\widehat{a}_i}{\sqrt{(Q_{\widehat{a}})_{ii}}} = \frac{\widehat{a}_i}{\sigma_\varepsilon\sqrt{(X^TX)^{-1}_{ii}}}.$$

Questa è la grandezza statistica scelta, di cui si calcola la probabilità che superi il suo valore sperimentale.

L'ipotesi nulla, $a_i = 0$, significa che il modello migliore non ha bisogno di includere il fattore X_i. Detto altrimenti, i modelli con e senza X_i hanno prestazioni uguali; "X_i non influisce su Y". Se il p-value di un fattore X_i è piccolo, diciamo minore di 0.05 per esemplificare, riteniamo falsa l'ipotesi nulla, quindi falso che X_i non influisca su Y; detto altrimenti, riteniamo che X_i influisca su Y. Se il p-value, invece, non è abbastanza piccolo, dubitiamo dell'importanza di X_i per spiegare Y.

Il p-value di due (o più) fattori allineati è in genere elevato, pessimo. La ragione è che il modello senza uno dei due fattori si comporta altrettanto bene, perché c'è l'altro fattore che compensa, che fa le veci di quello eliminato. Il discorso vale per entrambi, simmetricamente, quindi paradossalmente sembra dai p-value che entrambi i fattori siano irrilevanti. Ciascuno in effetti è irrilevante, perché c'è l'altro che ne fa le veci.

Infine, senza dettagli, segnaliamo che il p-value globale, che compare nell'ultima riga del summary dato dal software quando si esegue una regressione, è associato al modello nel suo insieme. La sua ipotesi nulla, detta a parole, è che il modello costante, con solo l'intercetta, senza alcun fattore, spieghi Y tanto bene quanto il modello sottoposto a giudizio. Tale ipotesi nulla viene rifiutata da un p-value globale piccolo. Siccome è un'ipotesi alquanto estrema, normalmente si trovano valori molto piccoli di questo p-value globale. Solo quando tutti i fattori fossero davvero fallimentari, anche questo p-value risulterebbe non piccolo. Nell'esempio visto sopra, in cui il modello proposto non è certo molto buono (un solo fattore è decente), il p-value globale è piccolissimo.

Come varia R^2 col numero di fattori

Supponiamo di esaminare due modelli

$$Y = a_1X_1 + ... + a_pX_p + b + \varepsilon$$

$$Y = a_1X_1 + ... + a_pX_p + a_{p+1}X_{p+1} + b + \varepsilon$$

relativamente alla stessa tabella di dati. Ci chiediamo quale abbia il valore più elevato di R^2.

Proposizione 8 *Il modello con più fattori ha R^2 maggiore.*

Proof. Ricordiamo che $R^2 = 1 - \frac{\sigma^2_\varepsilon}{\sigma^2_Y}$. Il numero σ^2_Y è lo stesso per i due modelli. Cambia invece σ^2_ε. Indichiamo con R^2_p ed R^2_{p+1} i valori di R^2 relativi al modello con p e con $p+1$ fattori rispettivamente. In modo simile definiamo $\sigma^2_{\varepsilon,p}$ e $\sigma^2_{\varepsilon,p+1}$. Se dimostriamo che

$$\sigma^2_{\varepsilon,p} \geq \sigma^2_{\varepsilon,p+1}$$

allora $\frac{\sigma^2_{\varepsilon,p}}{\sigma^2_Y} \geq \frac{\sigma^2_{\varepsilon,p+1}}{\sigma^2_Y}$, $1 - \frac{\sigma^2_{\varepsilon,p}}{\sigma^2_Y} \leq 1 - \frac{\sigma^2_{\varepsilon,p+1}}{\sigma^2_Y}$ ovvero $R^2_p \leq R^2_{p+1}$.

Ricordiamo poi che σ^2_ε è la varianza empirica dei residui ottimali, cioè quelli relativi ai parametri ottimali trovati col metodo dei minimi quadrati:

$$\begin{aligned}\sigma^2_{\varepsilon,p} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}^2_{i,p} = SQM_p\left(\widehat{a}^p_1,...,\widehat{a}^p_p,\widehat{b}^p\right)\\ \sigma^2_{\varepsilon,p+1} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}^2_{i,p+1} = SQM_{p+1}\left(\widehat{a}^{p+1}_1,...,\widehat{a}^{p+1}_p,\widehat{a}^{p+1}_{p+1},\widehat{b}^{p+1}\right)\end{aligned}$$

con ovvia notazione per $\widehat{\varepsilon}^2_{i,p},\widehat{\varepsilon}^2_{i,p+1}$ e dove $\widehat{a}^p_1,...,\widehat{a}^p_p,\widehat{b}^p$ sono i parametri ottimali del problema con p variabili, mentre $\widehat{a}^{p+1}_1,...,\widehat{a}^{p+1}_p,\widehat{a}^{p+1}_{p+1},\widehat{b}^{p+1}$ sono i parametri ottimali del problema con $p+1$ variabili. Esplicitamente,

$$SQM_p\left(\widehat{a}^p_1,...,\widehat{a}^p_p,\widehat{b}^p\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\widehat{a}^p_1x_{i,1}+...+\widehat{a}^p_px_{i,p}+\widehat{b}^p\right)\right)^2$$

$$\begin{aligned}&SQM_{p+1}\left(\widehat{a}^{p+1}_1,...,\widehat{a}^{p+1}_p,\widehat{a}^{p+1}_{p+1},\widehat{b}^{p+1}\right)\\ =\ &\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\widehat{a}^{p+1}_1x_{i,1}+...+\widehat{a}^{p+1}_px_{i,p}+\widehat{a}^{p+1}_{p+1}x_{i,p+1}+\widehat{b}^{p+1}\right)\right)^2\end{aligned}$$

da cui vediamo che

$$SQM_{p+1}\left(\widehat{a}^p_1,...,\widehat{a}^p_p,0,\widehat{b}^p\right) = SQM_p\left(\widehat{a}^p_1,...,\widehat{a}^p_p,\widehat{b}^p\right).$$

Non c'è però una relazione semplice tra i due vettori di parametri ottimali.

Ora, per la condizione di ottimalità,

$$\begin{aligned}SQM_{p+1}\left(\widehat{a}^{p+1}_1,...,\widehat{a}^{p+1}_p,\widehat{a}^{p+1}_{p+1},\widehat{b}^{p+1}\right) &\le SQM_{p+1}\left(\widehat{a}^p_1,...,\widehat{a}^p_p,0,\widehat{b}^p\right)\\ &= SQM_p\left(\widehat{a}^p_1,...,\widehat{a}^p_p,\widehat{b}^p\right)\end{aligned}$$

quindi $\sigma^2_{\varepsilon,p+1}\le\sigma^2_{\varepsilon,p}$, come volevasi dimostrare. ■

Riduzione di un modello

Come già osservato, in genere si tende a preferire un modello economico, con pochi fattori. Nei casi applicativi fortunati, però, si posseggono molti dati e molti fattori potenzialmente utili. Quali utilizzare? Si immagini una situazione con 20 fattori, che potrebbero essere utili a spiegare una variabile Y. Per esempio, potrebbe essere uno studio massiccio sul problema della disoccupazione, in cui si cerca di capirla esplorando una ventina di variabile economiche, sociali ecc. Gli individui potrebbero essere le

regioni italiane, le nazioni europee o mondiali. Quale gruppo di variabili e tramite quale modello regressivo, può spiegare al meglio la disoccupazione? Un simile modello potrebbe permettere di eseguire simulazioni, variando il valore dei fattori, per capire quali condizioni socio-economiche migliorerebbero il gradi di occupazione.

Nel paragrafo precedente abbiamo visto che l'indice R^2 aumenta all'aumentare del numero di fattori. Quindi R^2 non può essere usato per decidere il modello migliore, perché preferirebbe sempre quello con tutti i fattori. Però può essere usato per apprezzare il grado di peggioramento che si ottiene eliminando dei fattori, come spiegheremo tra un attimo.

Un metodo (non l'unico) ragionevole consiste nel partire dal modello completo, con tutti i fattori, ed eliminarne uno alla volta sulla base di varie considerazioni. Una delle considerazioni possibili è basata sui p-value. La strategia più semplice è quella di eliminare il fattore col peggior p-value (il più grande). Qui si capisce perché i fattori vanno eliminati uno per volta: per evitare il rischio di eliminare una coppia di fattori allineati ma entrambi rilevanti. Eliminandone uno solo, l'altro si fa carico della capacità di spiegazione che entrambi hanno ed il suo p-value diventa subito molto migliore.

Quando fermarsi? Qui può essere utile R^2. Eliminado fattori, esso peggiora, decresce. Ma accadrà che eliminando certi fattori esso non cambi molto, ed in questo caso l'eliminazione non peggiora granché la spiegazione che il modello dà della variabilità di Y, guadagnando invece in economia; mentre ad un certo punto, eliminando un ulteriore fattore, R^2 cala di più, segno che con tale eliminazione il modello è peggiorato. Così si può decidere di terminare il procedimento di eliminazione, prima di eliminare tale fattore determinante.

Ovviamente tutte queste indicazioni sono vaghe, le variazioni numeriche possono essere sfumate, quindi molte decisioni restano soggettive. Si consiglia di avere sotto controllo altri parametri, per non prendere decisioni superficiali. Un parametro è R^2_{adj} (a cui eventualmente si può delegare l'intero procedimento selettivo, senza nemmeno usare i p-value; però è più saggio affiancarlo all'uso dei p-value, avere cioè due metri paralleli di giudizio, che ci rafforzano nelle decisioni quando danno la stessa indicazione operativa, mentre ci allertano di una situazione difficile quando danno indicazioni operative contrastanti o troppo vaghe). Un altro oggetto è la coppia matrice di correlazione - plot della tabella, intendendo la tabella completa con anche Y. Queste correlazioni a due a due col relativo grafico di dispersione, offrono un controllo ulteriore dei legami tra le variabili, permettono di accorgersi dei fattori allineati, fanno capire in altro modo quali siano le variabili più legate ad Y, quindi sono uno strumento utile per prendere decisioni circa la riduzione del modello.

Infine, il grafico di PCA, che studieremo più avanti, può essere affiancato come ulteriore strumento grafico per visualizzare i legami tra le variabili.

1.4.1 Domande

1. Chiarire in che senso $\widehat{a}$ è un vettore aleatorio e calcolarne media e matrice di covarianza
2. Problemi causati da allineamenti di fattori
3. Importanza del p-value dei singoli fattori e cenni della sua formulazione matematica.
4. Che si può dire dei p-value dei singoli fattori in caso di allineamento
5. Dipendenza di R^2 dal numero di fattori
6. Overfitting e difetti di un modello con troppi fattori
7. Significato numerico dei parametri ottimali; legame coi coefficienti di correlazione
8. Come ridurre la numerosità di un modello.
9. La formula $a_i = \rho(X_i, Y)$ è vera in generale?

1.4.2 Domande riassuntive su correlazione e regressione

1. Formulare e dimostrare l'invarianza di scala del coefficiente di correlazione.
2. Perché la matrice di covarianza di un vettore aleatorio è simmetrica.
3. Data una matrice di covarianza, indicare come si costruisce la sua radice quadrata e descrivere un suo utilizzo (la seconda parte può essere rinviata dopo lo studio dei capitoli successivi).
4. Definizione della matrice di covarianza di un vettore aleatorio.
5. A cosa corrisponde algebricamente (*non* graficamente), per un campione di dati (x_i, y_i), $i = 1, ..., n$, un segno positivo di $\widehat{Cov}$?
6. Definire i residui di un modello di regressione lineare. Essi sono numeri univocamente determinati dai dati oppure sono funzioni dei parametri?
7. Definire con simboli matematici precisi il problema della regressione lineare multipla, fornire il risultato principale riguardante la stima dei parametri (enunciato e dimostrazione), discutere il ruolo di eventuali allieneamenti dal punto di vista matematico.
8. Dimostrare la formula $Q_Y = AQ_XA^T$, ricordando chi sono gli oggetti coinvolti.

9. Spiegare il p-value dei singoli fattori nella regressione multipla, fornendo anche alcuni elementi matematici sia relativi al concetto di p-value in generale, sia relativi allo specifico p-value dei singoli fattori.

10. Cosa si intende per overfitting nella regressione multipla, che apparenti pregi e che presumibili difetti ha?

11. Dimosrare che la matrice di covarianza di un vettore aleatorio è semi-definita positiva.

12. Nella regressione semplice, il coefficiente angolare non coincide sempre col coefficiente di correlazione. Discutere sinteticamente questo problema fornendo alcune dimostrazioni delle formule rilevanti per la discussione, estendendo poi, se possibile, il discorso al caso della regressione multipla.

13. Enunciare e dimostrare la formula per i coefficienti della regressione multipla ottenuta tramite il metodo minimi quadrati. Dedurne media e covarianza degli stimatori dei parametri del modello.

14. Dimostrare che R^2 aumenta all'aumentare dei fattori.

15. Definizione di varianza spiegata nella regressione.

16. Descrivere dal punto di vista matematico il p-value dei singoli fattori in una regressione multipla. Precisamente: ricordare il concetto di p-value in generale, descrivere il caso particolare richiesto, facendo riferimento alle formule ed i teoremi sulla regressione multipla.

17. Quando, per i coefficienti della regressione multipla, vale $a_i = \rho(X_i, Y)$? Spiegare anche i preliminari necessari.

18. L'allineamento tra fattori visto dal punto di vista strettamente matematico (cosa significa matematicamente e quali conseguenze matematiche ha).

Capitolo 2

Vettori gaussiani e PCA

2.1 Vettori gaussiani

Ricordiamo che una v.a. gaussiana o normale $N(\mu, \sigma^2)$ è una v.a. con densità di probabilità

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|x-\mu|^2}{2\sigma^2}\right).$$

Si dimostra che μ è la media e σ^2 la varianza. La normale standard è il caso $\mu = 0$, $\sigma^2 = 1$. un fatto importante è il seguente: se Z è una normale standard allora $\mu + \sigma Z$ è $N(\mu, \sigma^2)$, ed ogni gaussiana $N(\mu, \sigma^2)$ si può scrivere nella forma $\mu + \sigma Z$ con $Z \sim N(0, 1)$.

Si può dare la definizione di *vettore* gaussiano, o gaussiana multidimensionale, in più modi. Seguiremo la seguente strada: prendiamo la proprietà che $\mu + \sigma Z$ è una $N(\mu, \sigma^2)$ come definizione; poi calcoleremo la densità, sotto opportune ipotesi.

Definizione 6 *i) Chiamiamo vettore normale (o gaussiano) standard in dimensione d un vettore aleatorio $Z = (Z_1, ..., Z_d)$ con densità congiunta*

$$f(z_1, ..., z_d) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} = \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{z_1^2 + ... + z_d^2}{2}}.$$

ii) Un vettore $X = (X_1, ..., X_n)$ si dice gaussiano se può essere rappresentato nella forma

$$X = AZ + b$$

dove $Z = (Z_1, ..., Z_d)$ è un vettore normale standard, $A : \mathbb{R}^d \to \mathbb{R}^n$ è una matrice e b è un vettore di $\mathbb{R}^n$.

In base alla Proposizione 15, la condizione del punto (i) equivale al fatto che le componenti $Z_1, ..., Z_d$ sono gaussiane standard indipendenti. Questa è quindi una formulazione equivalente: un vettore $Z = (Z_1, ..., Z_d)$ è normale standard se le componenti $Z_1, ..., Z_d$ sono normali standard indipendenti.

Dai risultati sulla trasformazione dei momenti sotto applicazioni affini, Sezione 1.2.3, abbiamo le seguenti fondamentali proprietà:

Proposizione 9 *Se X è un vettore gaussiano della forma $X = AZ+b$ (con le notazioni della Definizione 6), il vettore dei valori medi μ e la matrice di covarianza Q di X sono dati da*

$$\mu = b$$
$$Q = AA^T.$$

Proposizione 10 *Sia $X = (X_1, ..., X_n)$ un vettore gaussiano secondo la Definizione 6, B una matrice $n \times m$, c un vettore di $\mathbb{R}^m$. Allora*

$$Y = BX + c$$

è un vettore gaussiano di dimensione m (sempre secondo la Definizione 6). La relazione tra medie e covarianze è

$$\mu_Y = B\mu_X + c$$
$$Q_Y = BQ_XB^T.$$

Spesso usaremo la frase "sia $X = (X_1, ..., X_n)$ un vettore gaussiano avente matrice di covarianza Q"; scriveremo inoltre che X è $N(\mu, Q)$, se ha media μ e matrice di covarianza Q. Secondo la definizione, questo significa che X è della forma $X = AZ + \mu$ (con Z, A, μ come sopra) e vale $AA^T = Q$.

Una domanda naturale però è: esiste? Data una matrice Q, simmetrica e definita non-negativa, c'è un vettore gaussiano con tale matrice di covarianza? La frase evidenziata sopra, a seconda del punto di vista, allude implicitamente a questo. Data Q, simmetrica e definita non-negativa, sappiamo che esiste la radice quadrata $\sqrt{Q}$, costruita nella Sezione 1.2. Posto $A = \sqrt{Q}$, vale $AA^T = Q$. Prendiamo allora un qualsiasi vettore normale standard $Z = (Z_1, ..., Z_n)$ ed un qualsiasi $b \in \mathbb{R}^n$. Il vettore

$$X = \sqrt{Q}Z + b$$

è gaussiano ed ha covarianza Q. Abbiamo risposto affermativamente alla domanda di esistenza posta sopra.

Discutiamo ora della densità di probabilità congiunta di un vettore gaussiano. Sia Q la matrice di covarianza di X. Ricordiamo alcuni fatti visti nella Sezione 1.2. La matrice Q è diagonalizzabile (in quanto simmetrica), ovvero vale $Q = UQ_eU^T$ dove U è ortogonale e Q_e diagonale, con gli autovalori di Q sulla diagonale. Inoltre Q è definita

non-negativa; i suoi autovalori sono non-negativi. Dalla formula $Q = UQ_eU^T$ discende che Q è invertibile se e solo gli autovalori sono strattamente positivi, ovvero se e solo se Q è definita positiva. In tal caso anche Q^{-1} è definita positiva, quindi l'espressione $(x-\mu)^T Q^{-1}(x-\mu)$ che compare tra poco è non-negativa. Inoltre $\det(Q) > 0$.

Proposizione 11 *Se Q è invertibile, allora $X = (X_1, ..., X_n)$ possiede densità congiunta data da*

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(Q)}} \exp\left(-\frac{(x-\mu)^T Q^{-1}(x-\mu)}{2}\right)$$

dove $x = (x_1, ..., x_n) \in \mathbb{R}^n$.

Proof. Ci limitiamo, per brevità, al caso in cui $d = n$ nella Definizione 6. Supponiamo quindi che sia $X = AZ + \mu$ con Z di dimensione n. Siccome Q è invertibile e $Q = AA^T$ allora anche A è invertibile (si verifica per assurdo). La densità congiunta di Z è $f_Z(z) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{z^T z}{2}\right)$. Per il Corollario 3, X ha densità congiunta, data da

$$f_X(x) = \frac{f_Z(A^{-1}(x-\mu))}{|\det A|}.$$

Sostituendo la formula di f_Z troviamo

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{(2\pi)^n |\det A| |\det A|}} \exp\left(-\frac{(A^{-1}(x-\mu))^T (A^{-1}(x-\mu))}{2}\right) \\ &= \frac{1}{\sqrt{(2\pi)^n |\det(AA^T)|}} \exp\left(-\frac{(x-\mu)^T (AA^T)^{-1}(x-\mu)}{2}\right). \end{aligned}$$

La dimostrazione è completa. ■

La formula precedente per la densità può essere utilizzata come definizione alternativa del concetto di vettore gaussiano, limitatamente al caso in cui Q sia invertibile; si può poi verificare che le due definizioni sono equivalenti.

Osservazione 18 *La densità di un vettore gaussiano (quando Q è invertibile) è determinata dal vettore dei valori medi e dalla matrice di covarianza. Questo fatto fondamentale verrà usato più tardi nello studio dei processi stocastici. Usando il concetto di legge di un vettore aleatorio, questo fatto vale anche nel caso degenere, senza densità, in cui si deve usare la definizione 6, però i dettagli sono meno elementari.*

Infine, abbiamo gli elementi per dimostrare il seguente fatto.

Proposizione 12 *Un vettore gaussiano $X = (X_1, ..., X_n)$, centrato ($\mu_X = 0$), è standard se e solo se ha matrice di covarianza Q uguale all'identità.*

Proof. Se X è standard, $Q_{ii} = Var[X_i] = 1$, $Q_{ij} = Cov(X_i, X_j) = 0$ per $j \neq j$ (perché le componenti di X sono gaussiane standard indipendenti), quindi Q è l'identità.

Se Q è l'identità, allora $X = (X_1, ..., X_n)$ possiede densità congiunta data da

$$f(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{x^T x}{2}\right)$$

e quindi è una gaussiana standard, per la Definizione 6. ■

Analogo e rilevante è il seguente fatto:

Proposizione 13 *Se il vettore $X = (X_1, X_2)$ è gaussiano, la proprietà $Cov(X_1, X_2) = 0$ implica che X_1 e X_2 siano indipendenti.*

Proof. La matrice di covarianza è diagonale, della forma

$$Q = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

quindi $\det(Q) = \sigma_1^2 \sigma_2^2$,

$$Q^{-1} = \begin{pmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{pmatrix}$$

da cui si deduce facilmente

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2}\right) \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right).$$

Per la Proposizione 15, X_1 e X_2 sono indipendenti. ■

Un'altra definizione

Esistono altre definizioni di vettore gaussiano. Per curiosità enunciamo la seguente, che è forse la più veloce ma può apparire più oscura di altre.

Definizione 7 *Un vettore aleatorio $X = (X_1, ..., X_n)$ si dice gaussiano se accade che per ogni vettore di numeri reali $u = (u_1, ..., u_n)$ la v.a.*

$$\langle u, X \rangle = \sum_{i=1}^{n} u_i X_i$$

sia gaussiana.

Questa definizione generalizza la nota proprietà che le combinazioni lineari di gaussiane indipedenti sono gaussiane. Con questa definizione è immediato verificare che le trasformazioni lineari di vettori gaussiani sono vettori gaussiani.

La definizione data ha anche una certa interpretazione geometrica. Se u ha lunghezza unitaria, l'epressione $\langle u, X \rangle$ è la proiezione di X su u. La definizione afferma quindi che tutte le proiezioni uni-dimensionali sono gaussiane.

Vettori gaussiani degeneri

Diciamo che un vettore gaussiano $X = (X_1, ..., X_n)$ è degenere se la sua matrice di covarianza Q è singolare, non invertibile. In questo caso si potrebbe dimostrare quanto segue. Per semplicità espositiva, supponiamo $\mu_X = 0$, altrimenti le proprietà seguenti varranno per $X - \mu_X$. C'è un sottospazio vettoriale V, di dimensione strettamente minore di n, tale che i valori del vettore X appartengono a V. Non c'è densità congiunta su $\mathbb{R}^n$. Prendendo come V l'insieme dei valori possibili di X, c'è una densità su V.

2.1.1 Raffigurazioni e simulazioni

Il grafico della densità normale standard in 2 dimensioni è invariante per rotazioni:

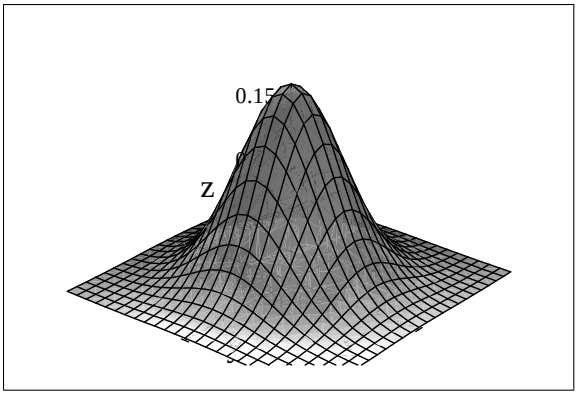
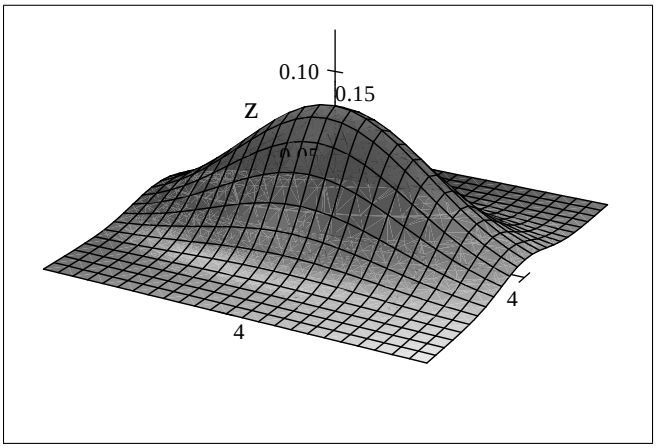


Grafico della normale standard in due dimensioni

Il grafico delle altre densità gaussiane può essere immaginato eseguendo trasformazioni lineari del piano base xy (deformazioni definite da A) e traslazioni (di b). Per esempio, se A è la matrice

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

che amplifica l'asse x di un fattore 2, otteniamo il seguente grafico:



Tuttavia questo tipo di grafico è poco interpretabile, salvo casi facili; grazie alle ombreggiature riusciamo ad intuire qualcosa, ma con poca precisione. Vengono allora in aiuto due tipi di raffigurazione: tramite curve di livello e tramite campione aleatorio generato con un software.

Curve (o insiemi) di livello

Un modo di visualizzare un grafico in due dimensioni è quello di tracciare le sue curve di livello. Sviluppiamo l'argomento in dimensione qualsiasi. Data $f : \mathbb{R}^n \to \mathbb{R}$, l'insieme di livello a è il luogo dei punti $x \in \mathbb{R}^n$ tali che $f(x) = a$. Nel caso di una densità f, essendo positiva, ha senso esaminare solo il caso $a > 0$. Nel caso della gaussiana $N(\mu, Q)$, dobbiamo capire l'equazione

$$\frac{1}{\sqrt{(2\pi)^n \det(Q)}} \exp\left(-\frac{(x-\mu)^T Q^{-1} (x-\mu)}{2}\right) = a.$$

Per semplicità algebrica, restringiamoci al caso $\mu = 0$, altrimenti basterò traslare di μ il risultato finale. Inoltre, poniamo $C = \sqrt{(2\pi)^n \det(Q)}$. Allora vale

$$\exp\left(-\frac{1}{2} x^T Q^{-1} x\right) = aC$$

e posto $a' := -2\log(aC)$, l'equazione diventa

$$x^T Q^{-1} x = a'.$$

Questa, per $a' \geq 0$, è l'equazione di un'ellissoide[1]. Infatti, usando la scomposizione $Q = U Q_e U^T$, si veda la (1.1), e ricordando che

$$Q^{-1} = \left(U^T\right)^{-1} Q_e^{-1} U^{-1} = U Q_e^{-1} U^T$$

[1]Può sembrare che ci sia un assurdo nel fatto che partiamo da insiemi di livello $\{f(x) = a\}$ definiti per ogni $a > 0$ e poi troviamo la restrizione $a' \geq 0$ quando invece $a' = -2\log(aC)$ può assumere qualsiasi segno. La contraddizione non esiste perché se $a > \max f$ l'insieme di livello $\{f(x) = a\}$ è vuoto, e svolgendo bene i conti si può riconoscere che questo corrisponde ai valori $a' < 0$.

l'equazione diventa

$$x^T U Q_e^{-1} U^T x = a'.$$

Posto $y = U^T x$ troviamo

$$y^T Q_e^{-1} y = a'$$

che in coordinate si legge

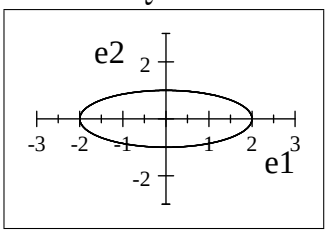
$$\frac{y_1^2}{\lambda_1} + ... + \frac{y_n^2}{\lambda_n} = a'$$

ovvero un'ellissoide.

Più precisamente, le coordinate y sono quelle nella base $\{e_1, ..., e_n\}$ di autovettori di Q (si veda l'osservazione (5), essendo $y = U^T x$), quindi si tratta di un ellissoide rispetto a tale base, cioè avente $e_1, ..., e_n$ come assi. Inoltre, la lunghezza dei semiassi è $\sqrt{\lambda_1}, ..., \sqrt{\lambda_n}$. In conclusione (torniamo a μ qualsiasi):

Proposizione 14 *Le ipersuperfici (curve per $n = 2$) di livello di un vettore gaussiano $N(\mu, Q)$ in $\mathbb{R}^n$ sono ellissoidi di centro μ ed assi dati dagli autovettori $e_1, ..., e_n$ di Q, con lunghezze degli assi pari alle radici degli autovalori $\sqrt{\lambda_1}, ..., \sqrt{\lambda_n}$.*

Il seguente disegno raffigura il caso $\frac{y_1^2}{4} + \frac{y_2^2}{1} = 1$. Questa è l'ellisse rispetto agli assi e_1, e_2, base di autovettori di Q.

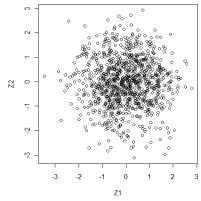


Se invece vogliamo vedere l'ellisse nella base canonica originaria, quella delle variabili x_i, bisogna eseguire la rotazione U e la traslazione di μ. Non c'è bisogno di sapere con esattezza di che rotazione si tratta, basta sapere come appaiono i vettori e_1, e_2 nella base canonica (cioè avere le loro coordinate), e tracciare l'ellisse con tali assi.

Simulazione

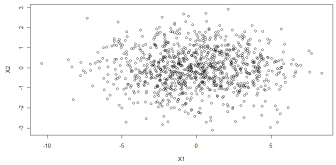
Un altro modo di visualizzare un vettore aleatorio X è di raffigurare un suo campione sperimentale, possibilmente numeroso: è (come per le curve di livello) una raffigurazione nello spazio $\mathbb{R}^k$, lo spazio dei *valori* possibili di X; se i punti sono molti, si riesce ad intuire la struttura delle curve di livello. Tornando alla Definizione 6, un modo di avere un campione di numerosità N da X è quello di averlo da Z e trasformarlo. Iniziamo allora osservando che un campione di numerosità $N = 1000$ (per fare un esempio) estratto da Z in dimensione $n = 2$ si ottiene e raffigura coi comandi

```
Z1<-rnorm(1000); Z2<-rnorm(1000); plot(Z1,Z2)
```



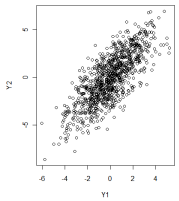
Possiamo poi osservare l'effetto di una matrice del tipo $A_1 = \begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix}$ con $\lambda \neq 1$, ad esempio $\lambda = 3$:

```
X1<-3*Z1; X2<-Z2; plot(X1,X2)
```



Infine, possiamo vedere l'effetto di una successiva rotazione (in senso antiorario) di θ radianti, ottenuta componendo ulteriormente con la matrice $A_2 = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ (componiamo le due, cioè applichiamo prima la matrice A_1 poi la A_2); vediamo ad esempio $\theta = 1$:

```
A11 = cos(1); A12 = -sin(1); A21 = sin(1); A22 = cos(1)
Y1 <- A11*X1+A12*X2; Y2 <- A21*X1+A22*X2; plot(Y1,Y2)
```



Infine, poniamoci questa domanda: data una matrice di covarianza Q ed un vettore medio μ, come possiamo simulare un vettore X di tipo $N(\mu, Q)$? Vogliamo generare un punto casuale $(x_1, ..., x_n)$ dalla $N(\mu, Q)$. Per far questo possiamo generare n numeri casuali indipendenti $z_1, ..., z_n$ dalla normale standard 1-dimensionale e calcolare

$$\sqrt{Q}z + \mu$$

dove $z = (z_1, ..., z_n)$. Per trovare le componenti della matrice $\sqrt{Q}$ possiamo usare la formula$\sqrt{Q} = U\sqrt{Q_e}U^T$. La matrice $\sqrt{Q_e}$ è ovvia. Per ottenere la matrice U si ricordi che le sue colonne sono gli autovettori $e_1, ..., e_n$ scritti nella base di partenza. Basta quindi che il software sia in grado di effettuare la decomposizione spettale di Q. Ecco i comandi di `R`.

Innanzi tutto,

```
e <- eigen(Q)
```

```
U <- e$vectors
U %*% diag(e$values) %*% t(U)
```

restituisce Q. Corrisponde alla formula $Q = UQ_eU^T$. Per avere la radice $\sqrt{Q} = U\sqrt{Q_e}U^T$ basta ora fare

```
B <- U %*% diag(sqrt(e$values)) %*% t(U)
```

e la matrice B è $\sqrt{Q}$.

I comandi ora scritti sono molto utili: descrivono come trasporre matrici, moltiplicare matrici, calcolare autovettori ed autovalori, costruire una matrice diagonale.

Appendice: richiamo su densità congiunta e marginali

In questa sezione abbiamo usato il concetto di densità congiunta di un vettore aleatorio, che ora richiamiamo.

Una densità di probabilità f su $\mathbb{R}^n$ è una funzione non negativa (misurabile) e tale che

$$\int_{\mathbb{R}^n} f(x_1, ..., x_n)\, dx_1 \cdots dx_n = 1.$$

Un vettore aleatorio $X = (X_1, ..., X_n)$ ha densità di probabilità *congiunta* $f_X(x_1, ..., x_n)$ se

$$P((X_1, ..., X_n) \in A) = \int_A f_X(x_1, ..., x_n)\, dx_1 \cdots dx_n$$

per ogni insieme $A \subset \mathbb{R}^n$ per cui queste operazioni abbiano senso (es. unioni di plurirettangoli; in generale i cosidetti boreliani di $\mathbb{R}^n$).

Parallelamente, presa una generica componente X_i del vettore aleatorio X, sopravvive il vecchio concetto di densità di probabilità di X_i, funzione $f_{X_i}(x_i)$ di una sola variabile; è detta densità *marginale* di X_i.

Nasce allora la domanda circa il legame tra congiunta e marginali.

Proposizione 15 *In generale (quando le densità esistono), vale*

$$f_{X_1}(x_1) = \int_{\mathbb{R}^{n-1}} f(x_1, ..., x_n)\, dx_2 \cdots dx_n$$

e così per le altre. Quando $X_1, ..., X_n$ sono v.a. indipendenti, vale inoltre

$$f(x_1, ..., x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

e vale anche il viceversa (se la densità congiunta è il prodotto delle marginali, allora le v.a. sono indipendenti).

Omettiamo la dimostrazione, non troppo difficile peraltro. Osserviamo come interpretazione che, mentre dalla congiunta è sempre possibile calcolare le marginali, viceversa dalle marginali è in genere molto difficile risalire alla congiunta, salvo nel caso di

indipendenza. Questo non deve stupire: è come il problema di calcolare la probabilità di una intersezione $P(A \cap B)$. In generale, abiamo bisogno di conoscere ad esempio $P(A|B)$, che è un'informazione ben più complessa delle probabilità "marginali" $P(A)$ e $P(B)$.

Appendice: trasformazione di una densità

Esercizio 3 *Se X ha cdf $F_X(x)$ e g è strettamente crescente e continua, allora $Y = g(X)$ ha cdf*

$$F_Y(y) = F_X\left(g^{-1}(y)\right)$$

per tutte le y nell'immagine di g. Se g è strettamente decrescente e continua, la formula è

$$F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$$

Soluzione:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right).$$

La seconda è identica.

Esercizio 4 *Se X ha una pdf continua $f_X(x)$ e g è strettamente crescente e differenziabile, allora $Y = g(X)$ ha pdf*

$$f_Y(y) = \frac{f_X\left(g^{-1}(y)\right)}{g'\left(g^{-1}(y)\right)} = \left.\frac{f_X(x)}{g'(x)}\right|_{y=g(x)}$$

per tutte le y nell'immagine di g. Se g è decrescente e differenziabile, la formula è

$$f_Y(y) = -\left.\frac{f_X(x)}{g'(x)}\right|_{y=g(x)}.$$

Soluzione: Vale

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X\left(g^{-1}(y)\right) = F_X'\left(g^{-1}(y)\right)\frac{d}{dy}g^{-1}(y) \\ &= f_X\left(g^{-1}(y)\right)\frac{1}{g'\left(g^{-1}(y)\right)} = \left.\frac{f_X(x)}{g'(x)}\right|_{y=g(x)}. \end{aligned}$$

La seconda è identica.

Quindi, in generale, abbiamo:

Proposizione 16 *Se g è monotona e differenziabile, la trasformazione di densità è data da*

$$f_Y(y) = \left.\frac{f_X(x)}{|g'(x)|}\right|_{y=g(x)}$$

Osservazione 19 *Se g non è monotona, sotto ipotesi opportune la formula si generalizza a*

$$f_Y(y) = \sum_{x:y=g(x)} \frac{f_X(x)}{|g'(x)|}.$$

Esercizio 5 *Se X è una v.a. esponenziale di parametro λ, trovare la densità di $Y = X^2$ seguendo il metodo di risoluzione degli esercizi precedenti e confrontare il risultato con la formula generale.*

Osservazione 20 *Una seconda dimostrazione della formula precedente proviene dalla seguente caratterizzazione delle densità: f è la densità di X se e solo se*

$$E[h(X)] = \int_{\mathbb{R}} h(x) f(x) dx$$

per tutte le funzioni continue e limitate h. Usiamo questo fatto per dimostrare che $f_Y(y) = \frac{f_X(x)}{|g'(x)|}\Big|_{y=g(x)}$ è la densità di $Y = g(X)$. Calcoliamo $E[h(Y)]$ per una generica funzione continua e limitata h. Dalla definizione di Y e dalla caratterizzazione precedente applicata a X, abbiamo

$$E[h(Y)] = E[h(g(X))] = \int_{\mathbb{R}} h(g(x)) f(x) dx.$$

Usiamo il teorema di cambio di variabile negli integrali, con $y = g(x)$, se g è monotona, biunivoca e differenziabile. Abbiamo $x = g^{-1}(y)$, $dx = \frac{1}{|g'(g^{-1}(y))|} dy$ (abbiamo scritto il valore assoluto per non cambiare gli estremi di integrazione) così che

$$\int_{\mathbb{R}} h(g(x)) f(x) dx = \int_{\mathbb{R}} h(y) f\left(g^{-1}(y)\right) \frac{1}{|g'(g^{-1}(y))|} dy.$$

Se poniamo $f_Y(y) := \frac{f_X(x)}{|g'(x)|}\Big|_{y=g(x)}$ abbiamo dimostrato che

$$E[h(Y)] = \int_{\mathbb{R}} h(y) f_Y(y) dy$$

per ogni funzione continua e limitata h. Usando di nuovo la caratterizzazione, deduciamo che $f_Y(y)$ è la densità di Y. Questa dimostrazione è basata sul cambio di variabile negli integrali.

Osservazione 21 *La stessa dimostrazione funziona nel caso multidimensionale, in cui non riusciamo più a lavorare con le cdf. Bisogna usare il teorema di cambio di variabile negli integrali multipli. Ricordiamo che in esso al posto di $dy = g'(x)dx$ si deve usare $dy = |\det Dg(x)| dx$ dove Dg è la matrice jacobiana (la matrice delle derivate prime)*

della trasformazione $g : \mathbb{R}^n \to \mathbb{R}^n$. In realtà abbiamo bisogno della trasformazione inversa, quindi usiamo la formula

$$dx = \left|\det Dg^{-1}(y)\right| dy = \frac{1}{\left|\det Dg\left(g^{-1}(y)\right)\right|} dy.$$

Con gli stessi passaggi visti sopra nel caso 1-dimensionale, otteniamo il seguente risultato.

Proposizione 17 *Se g è biunivoca e differenziabile con matrice jacobiana invertibile e $Y = g(X)$, allora*

$$f_Y(y) = \left.\frac{f_X(x)}{\left|\det Dg(x)\right|}\right|_{y=g(x)}.$$

Corollario 3 *Sia $X = (X_1, ..., X_n)$ un vettore casuale, A una matrice $n \times n$ invertibile, $b \in \mathbb{R}^n$, ed $Y = (Y_1, ..., Y_n)$ un vettore casuale definito da*

$$Y = AX + b.$$

Se X ha densità congiunta $f_X(x)$ allora anche Y ha densità congiunta, data da

$$f_Y(y) = \frac{f_X\left(A^{-1}(y-b)\right)}{|\det A|}.$$

Proof. La trasformazione $g(x) = Ax+b$ è invertibile, con inversa $g^{-1}(y) = A^{-1}(y-b)$. La matrice jacobiana di $g(x)$ è A, costante. Basta allora sostituire questi fatti nella formula precedente. ■

Esercizio 6 *Se X (in $\mathbb{R}^n$) ha densità $f_X(x)$ e $Y = UX$, dove U è una trasformazione ortogonale di $\mathbb{R}^n$ (ovvero $U^{-1} = U^T$), allora Y ha densità*

$$f_Y(y) = f_X\left(U^T y\right).$$

Soluzione. Le trasformazioni ortogonali sono invertibili ed hanno determinante pari a ± 1, in quanto

$$1 = \det I_d = \det\left(UU^T\right) = \det(U)\det\left(U^T\right) = \det(U)^2.$$

Basta quindi sostituire nella formula precedente.

2.1.2 Domande

1. Densità congiunta e marginali, legami (senza dimostrazione)
2. Definizione di vettore gaussiano standard

3. Definizione di vettore gaussiano qualsiasi

4. Vettore medio e matrice di covarianza di un vettore gaussiano standard e di uno qualsiasi

5. Un vettore gaussiano è standard se e solo se ha media nulla e matrice di covarianza identità

6. Diverse raffigurazioni di un vettore gaussiano standard o meno: comprensione delle curve di livello nel caso generale; come generare punti gaussiani relativi a diverse situazioni (rotazioni, dilatazioni)

7. Fornire esempi di vettori gaussiani particolari, nei seguenti due casi: i) le cui componenti sono $N(0,1)$ ma, come vettori, non sono standard; ii) gaussiana degenere

8. Le trasformazioni affini di gaussiane sono gaussiane

9. Costruzione della radice quadrata di una matrice simmetrica semi-definita positiva. Conseguenze sulla generazione relativa ad un vettore aleatorio gaussiano con matrice data

2.2 Il metodo delle componenti principali

Una tabella della forma

	X_1	...	X_p
1	$x_{1,1}$	...	$x_{1,p}$
2	$x_{2,1}$	...	$x_{2,p}$
...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$

può essere pensata come un insieme di n punti nello spazio $\mathbb{R}^p$: i punti $(x_{i,1}, ..., x_{i,p})$ al variare di $i = 1, ..., n$. Le visualizzazioni aiutano moltissimo l'esplorazione dei dati, quindi questa associazione geometrica può essere molto utile. Il problema però è duplice: da un lato, se $p > 2$, la raffigurazione è difficile ($p = 3$) o impossibile ($p > 3$). Quindi dobbiamo effettuare fotografie (proiezioni) 2-dimensionali dell'insieme dei punti. Dall'altro, nel momento in cui eseguiamo proiezioni 2-dimensionali, esse potrebbero essere poco leggibili se i punti sono troppo sovrapposti. Per questo bisogna trovare la visuale 2-dimensionale più conveniente. Questo è il problema di PCA: *rappresentare un insieme di punti di $\mathbb{R}^p$ nel modo bidimensionale più sparpagliato possibile.*

La soluzione è immediata se si utilizzano le idee della teoria dei vettori gaussiani. Si immagini che i punti suddetti in $\mathbb{R}^p$ siano realizzazioni sperimentali di un vettore gaussiano $X_1, ..., X_p$ (questa ipotesi è molto difficile da verificare, ma non viene realmente

utilizzata nel senso rigoroso del termine, serve solo per suggerire l'idea). Sotto tale ipotesi, essi si dispongono in modo pressapoco ellissoidale. Per conoscere gli ellissoidi, insiemi di livello della densità, serve la matrice di covariaza Q. A partire dai punti $(x_{i,1}, ..., x_{i,p})$, $i = 1, ..., n$, calcoliamoci allora la matrice di covarianza empirica $\widehat{Q}$. Essa è semplicemente la matrice di covarianza della tabella di dati.

Possiamo ora trovare autovettori ed autovalori, che indicheremo con $\widehat{e}_1, ..., \widehat{e}_n$ e $\widehat{\lambda}_1, ..., \widehat{\lambda}_n$ per sottolineare che sono empirici, cioè relativi a dati empirici.

Gli ellissoidi di cui abbiamo parlato hanno come assi i vettori $\widehat{e}_1, ..., \widehat{e}_n$ e come lunghezze le radici $\sqrt{\widehat{\lambda}_1}, ..., \sqrt{\widehat{\lambda}_n}$. Gli assi con le lunghezze maggiori sono quelli da preferire nell'ottica detta sopra di avere una visione il più sparpagliata possibile dei punti sperimentali. Volendo una visione 2-dimensionale, prendiamo gli assi $\widehat{e}_1, \widehat{e}_2$ (adottiamo sempre la convenzione che sia $\widehat{\lambda}_1 \geq ... \geq \widehat{\lambda}_n$; si noti inoltre che nell'ambito di oggetti empirici le coincidenze non avvengono praticamente mai, per cui di solito vale $\widehat{\lambda}_1 > ... > \widehat{\lambda}_n$ e non c'è ambiguità circa la scelta dei due autovalori più grandi).

Il piano individuato da $\widehat{e}_1, \widehat{e}_2$ è quello rispetto a cui è maggiore lo sparpagliamento dei dati. E' detto *piano principale*. Dobbiamo proiettare i dati su tale piano, ovvero calcolare i punti di coordinate

$$\langle x, \widehat{e}_1 \rangle , \langle x, \widehat{e}_2 \rangle$$

al variare di x nell'insieme di punti sperimentali considerato.

Osservazione 22 *In sintesi queste sono le operazioni compiute dal software: prende in input la tabella A scritta sopra, calcola* $\widehat{Q}$*, calcola* $\widehat{e}_1, ..., \widehat{e}_n$ *e* $\widehat{\lambda}_1, ..., \widehat{\lambda}_n$*, calcola* $\langle x, \widehat{e}_1 \rangle , \langle x, \widehat{e}_2 \rangle$ *al variare dei punti* x*. Il risultato è una nuvola di punti del piano, che rappresenterà graficamente. I comando di R sono*

B=princomp(A); biplot(B).

Il risultato del comando biplot, come vedremo negli esempi, è una buona visione della nuvola di punti iniziale. A che può servire? Lo scopo è di solito diverso da quanto abbiamo fatto fino ad ora nelle altre lezioni. Quando abbiamo calcolato la matrice di correlazione di una tabella, lo scopo era capire le relazioni tra le variabili $X_1, ..., X_p$. Quando abbiamo eseguito regressioni, lo scopo era di nuovo l'esame di relazioni tra le variabili (non più solo a due a due) e la creazione di modelli magari utili per la previsione; pur sempre questioni riguardanti le variabili. Ora, col metodo PCA e la sua visualizzazione finale, lo scopo è *vedere gli individui* (nazioni, regioni ecc.), come si pongono gli uni rispetto agli altri, se sono raggruppati, chi sta da una parte chi dall'altra lungo gli assi principali.

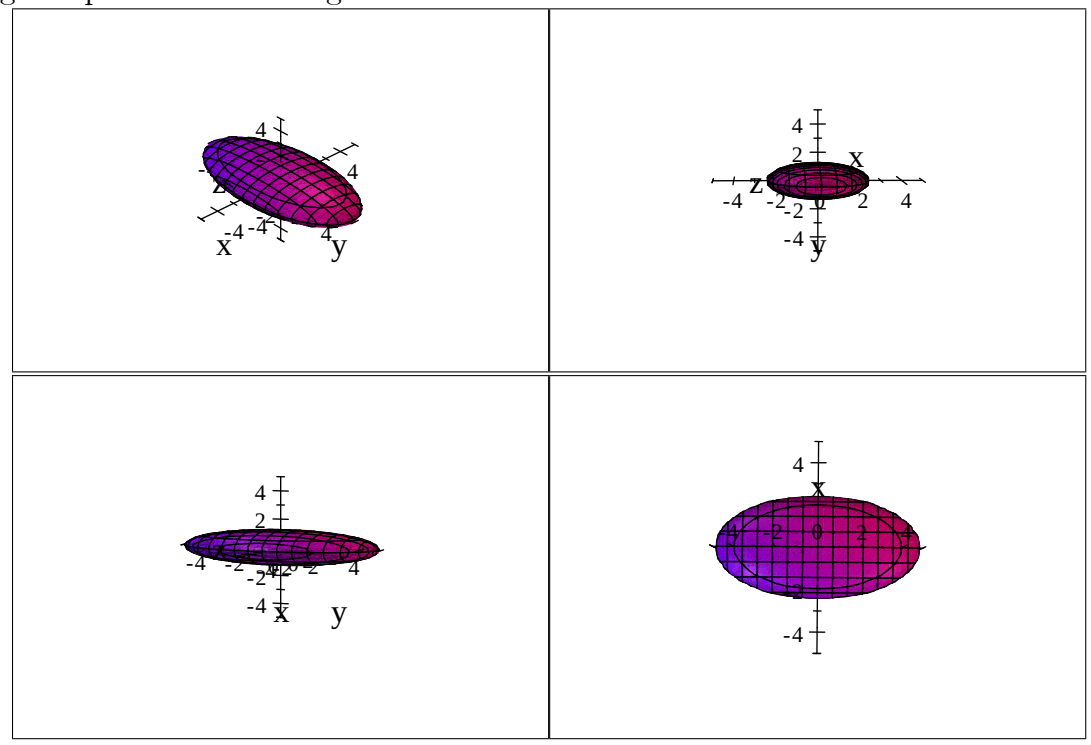
L'altra cosa da osservare è che si tratta di una tecnica *esplorativa*: si propone solo di esplorare la struttura dei dati, non di creare modelli come la regressione.

Esempio 2 *Giusto per rendere l'idea, immaginiamo un esempio tridimensionale. Supponiamo di avere* 3 *variabili aleatorie e 100 punti sperimentali, e supponiamo che il*

corrispondente centinaio di punti si disponga nello spazio tridimensionale secondo un ellissoide con i tre assi lunghi rispettivamente 4, 1 e 0.2 (in qualche unità di misura). Se guardiamo la figura dal punto di vista sbagliato, vediamo una sottilissima striscetta di punti tutti accalcati, e non saremo in grado di vedere eventuali relazioni tra loro. Se invece guardiamo la figura di piatto vediamo un'ellisse di assi 4 ed 1, che ci dà un'idea molto più realistica della disposizione tridimensionale dei dati, e ci permette di cogliere meglio eventuali relazioni tra di essi.

Osservazione 23 *Il fatto che uno dei tre assi della figura ellissoide sia molto piccolo (rispetto agli altri) ci dice che le variabili aleatorie 'variano di poco' in quella direzione: 'scartando' tale direzione (che è quello che facciamo proiettando l'ellissoide tridimensionale sul cerchio bidimensionale 'visto di piatto') abbiamo cioè la minima perdita di informazioni. Questo commento si riallaccia al probema della dimensione dei dati, che discuteremo sotto.*

Raffiguriamo un ellissoide in 3 dimensioni visto inclinato (visuale migliore per la comprensione 3-dimensionale) e poi visto secondo i tre piani possibili. Il terzo è il migliore per vedere i dettagli.



Punteggi e classifiche

Possiamo anche limitarci a calcolare i numeri

$$\langle x, \widehat{e}_1 \rangle$$

ovvero le proiezioni sul cosidetto *asse principale*. Essi forniscono i *punteggi* ottenuti dagli individui (nazioni, regioni ecc.) rispetto al *nuovo indicatore* che è $\widehat{e}_1$. Tramite tali punteggi possiamo poi stilare una *classifica* dei nostri individui. La classifica sarà interessante se l'asse principale ha un significato.

Esempio 3 *Supponiamo di voler esaminare la brillantezza economica di una nazione. Non c'è un'unica grandezza osservabile che corrisponde a tale brillantezza, ma varie che colgono vari aspetti: il tasso di occupazione, la produzione industriale, il numero di brevetti, il PIL, ecc. Immaginiamo di raccogliere i dati di una tabella in cui* $X_1, ..., X_p$ *sono le variabili ora elencate e varie altre, mentre* $1, ..., n$ *sono le nazioni europee. Se usiamo solo la variabile* X_1*, es. tasso di occupazione, possiamo fare una classifica tra le nazioni, ma tale classifica terrà conto solo di* X_1*. E così per ciascuna* X_i*. Invece, facciamo il seguente ragionamento. Tutte le variabili* X_i *sono abbsatanza correlate tra loro e la nuvola dei punti sperimentali (le nazioni) sarà quindi abbastanza allungata lungo una direzione principale* $\widehat{e}_1$*. Possiamo immaginare che tale direzione sia la brillantezza economica. Se facciamo questo assunto, la classifica dei valori* $\langle x, \widehat{e}_1 \rangle$ *è quella che cerchiamo. Si tenga presente che la matematica si ferma nel fornirci gli strumenti con cui trovare* $\widehat{e}_1$*,* $\langle x, \widehat{e}_1 \rangle$ *ecc., mentre l'interpretazione della direzione principale* $\widehat{e}_1$ *come brillantezza economica è un'estrapolazione non matematica, puramente soggettiva; istruita però dalla matematica. Non dobbiamo incolpare la matematica se tiriamo delle conclusioni sbagliate a livello interpretativo!*

Per calcolare effettivamente tali punteggi serve il vettore e_1. Esso si trova o tramite PCA leggendo i loading, oppure più basilarmente cercando il primo autovettore della matrice `Q`. Se `A` è la nostra tabella, basta fare:

```
Q=cov(A)
e1 = eigen(Q)$vector[,1]
```

Ora bisogna moltiplicare le righe della tavola A (cioè gli individui, le regioni) per e1. Il problema è che A è una tavola, non una matrice e gli usuali comandi di moltiplicazione vettoriale non funzionano. Bisogna tradurla in matrice. Basta introdurre:

AA <- as.matrix(A).

Visualizzando AA si vede che non è cambiato nulla, apparentemente, ma ora le moltiplicazioni vettoriali funzionano. Se calcoliamo

AA%*%e1

Stiamo moltiplicando ogni riga di AA per e1, cioè che desideriamo. Otterremo i punteggi desiderati. Il segno non è necessariamente in accordo con l'interpretazione che volgiamo dare, per cui va visto a posteriori. In questo esempio va cambiato. Ecco il risultato relativo all'esempio che descriveremo in dettaglio nel seguito di questa sezione:

> -AA%*%e1

Piem	1.077	EmRo	2.024	Camp	-3.246
Vaos	-0.166	Tosc	1.822	Pugl	-1.972
Lomb	1.781	Umbr	0.713	Basi	-2.323
TrAA	1.591	Marc	1.112	Cala	-3.137
Vene	1.920	Lazi	1.093	Sici	-2.966
FrVG	1.622	Abru	-0.069	Sard	-0.985
Ligu	0.358	Moli	-0.253		

Vediamo che EmRo 2.024, Vene 1.920, Tosc 1.822, Lomb 1.781, FrVG 1.622, TrAA 1.591 spiccano in positivo, mentre Camp -3.246, Cala -3.137, Sici -2.966, Basi -2.323, Pugl -1.972 in negativo.

2.2.1 Varianza lungo le componenti principali

Abbiamo detto sopra che i punti proiettati lungo l'asse principale sono dati dai numeri $\langle x, \widehat{e}_1 \rangle$. Chiamiamo $\widehat{e}_1$ componente principale, o meglio *prima componente principale*, ma usiamo anche il linguaggio vago di componenti principali per $\widehat{e}_2, \widehat{e}_3$ ecc. (es. $\widehat{e}_2$ è la seconda componente principale). In modo simile al caso della prima, può essere interessante considerare i punti proiettati lungo la generica i-esima componente principale, ovvero i numeri

$$\langle x, \widehat{e}_i \rangle$$

al variare di x.

Possiamo chiederci quale sia la varianza di questi numeri (da qui il titolo del paragrafo). Ci aspettiamo che sia maggiore per $\widehat{e}_1$ e via via più bassa per $\widehat{e}_2, \widehat{e}_3$ ecc., sulla base delle intuizioni geometriche sviluppate sopra. Vediamo un risultato rigoroso.

Impostiamo il risultato in modo astratto piuttosto che empirico, per maggior chiarezza di notazioni, ma si potrebbe dare un analogo empirico. Supponiamo di avere il vettore aleatorio $X = (X_1, ..., X_p)$ (non serve che sia gaussiano), aver calcolato la sua matrice di covarianza Q ed averla diagonalizzata; usiamo le solite notazioni $e_1, \ldots, e_p$, $\lambda_1, \ldots, \lambda_p$, U, Q_e. Proiettiamo il vettore aleatorio sulla componente principale e_i, ed indichiamo con V_i tale proiezione:

$$V_i = \langle X, e_i \rangle .$$

Teorema 5 *$Var[V_i] = \lambda_i$. Inoltre, $Cov(V_i, V_j) = 0$, quindi $V_1, ..., V_p$ sono indipendenti se X è gaussiano.*

Proof. Poniamo

$$V = U^T X.$$

Si riconosce subito che $V = (V_1, ..., V_p)$.

Sappiamo che

$$Q_V = U^T Q U$$

(Q indica Q_X). Ricordiamo che vale $Q = U Q_e U^T$, ovvero $Q_e = U^T Q U$. Quindi

$$Q_V = Q_e = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_p \end{pmatrix}.$$

Ma $V = (V_1, ..., V_p)$, quindi abbiamo dimostrato le identità $Var[V_i] = \lambda_i$, $Cov(V_i, V_j) = 0$. Infine, se X è gaussiano allora lo è anche V, e sappiamo che per un vettore gaussiano la scorrelazione delle componenti corrisponde all'indipendenza, quindi anche l'ultima affermazione del teorema è dimostrata. ■

Osservazione 24 *Il teorema offre un'ulteriore interpretazione dei numeri λ_i.*

Osservazione 25 *Le proiezioni V_i sono le coordinate di X rispetto alla base $e_1, \ldots, e_p$:*

$$X = \langle X, e_1 \rangle e_1 + ... + \langle X, e_p \rangle e_p = V_1 e_1 + ... + V_p e_p.$$

Osservazione 26 *Anche per via dell'identità scritta nella precedente osservazione, risulta abbastanza naturale dire che la variabilità (varianza) totale del vettore X è*

$$\lambda_1 + ... + \lambda_p$$

(pari a $Var[V_1] + ... + Var[V_p]$). Comunque, prendiamola come definizione di varianza complessiva di X. Allora è naturale vedere il numero

$$\frac{\lambda_1}{\lambda_1 + ... + \lambda_p}$$

come la proporzione di varianza catturata (spiegata) dall'asse principale,

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + ... + \lambda_p}$$

come la proporzione di varianza spiegata dal piano principale e così via. Viene anche chiamata varianza spiegata cumulativa.

Osservazione 27 *La traccia di una matrice è invariante per cambi di base. Quindi*

$$Traccia(Q) = Traccia(Q_e).$$

Ne discende

$$Var[X_1] + ... + Var[X_p] = \lambda_1 + ... + \lambda_p.$$

Anche questo consolida la visione di $\lambda_1 + ... + \lambda_p$ come indicatore della variabilità complessiva di X.

2.2.2 Un esempio

Esemplifichiamo quanto detto. Esaminiamo cinque potenziali indicatori di benessere nelle diverse regioni italiane:

X_1 = PLIC (posti letto in istituti di cura)

X_2 = SC (spese complessive per famiglia)

X_3 = SA.SC (proporzione di SC dedicata agli alimentari)

X_4 = TD (tasso di disoccupazione)

X_5 = TMI (tasso di mortalità infantile)

Ad ogni regione italiana R possiamo associare un vettore con cinque coordinate:

$$R \leftrightarrow \begin{pmatrix} X_1(R) & X_2(R) & X_3(R) & X_4(R) & X_5(R) \end{pmatrix}$$

Per non falsare l'indagine è conveniente standardizzare i dati: calcoliamo per ogni indicatore X_n la sua media μ_n e la sua deviazione standard σ_n e costruiamo una nuova tabella di dati dove sostituiamo ad ogni valore x di ogni indicatore X_n il valore standardizzato $\frac{x-\mu_n}{\sigma_n}$. In questo modo ora ogni indicatore ha la stessa media 0 e la stessa deviazione standard 1, e la matrice di covarianza Q coincide quindi con la matrice di correlazione.

Ci troviamo quindi con 20 punti disposti su una nuvoletta simile a un'ellissoide in 5 dimensioni. Nessuno riesce a visualizzare una tale figura, e l'idea di base del metodo PCA è quella di operare un cambio di variabili, cioè un cambio di base nel nostro spazio vettoriale di dimensione 5, che grosso modo ruoti la nuvola elissoidale in modo da poterla vedere dall'angolazione migliore, cioè in modo da averne una proiezione bidimensionale dove i dati sono il più distinti possibile tra loro.

Detta `A` la tabella dei dati regionali relativi ai cinque indicatori suddetti, calcoliamo

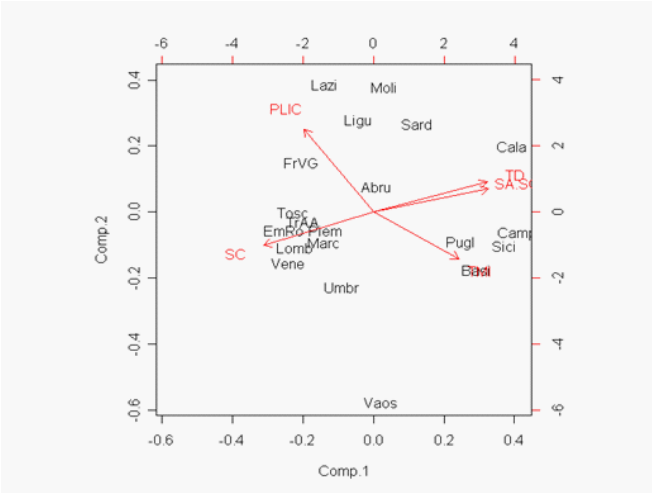
```
pca<-princomp(A)
```

Il nome `pca` vale come qualsiasi altro, nonostante l'allusione al metodo; il comando di R è `princomp(A)`.

Con il comando:

```
biplot(pca)
```

si ottiene un'immagine del piano principale, molto ricca di informazioni:



Essa contiene tre elementi: i due nuovi assi, la proiezione dei punti sul piano principale ed alcune freccie rosse corrispondenti alle variabili originarie. Gli assi orizzontale e verticale sono rispettivamente la prima e la seconda componente principale.

Osservazione 28 *Mentre i punti del biplot sono le vere proiezioni dei punti di partenza sul piano principale, quindi i legami tra gli individui mostrati dal biplot si possono prendere alla lettera (modulo il fatto che sono una proiezione) le frecce rosse non sono le proiezioni degli assi canonici sul piano principale, come si potrebbe pensare (se così fosse, sarebbe comune trovarne alcune molto più corte di altre). Tralasciamo l'algoritmo con cui vengono calcolate, osservando solo sperimentalmente che essere danno informazioni tra i loro legami e quelli con gli assi principali, attraverso allineamenti e perpendicolarità. Per quantificare la vera correlazione tra gli indicatori di partenza, bisogna leggere la matrice di correlazione delle variabili di partenza, tramite il comando* `cor(A)`*, non prendere alla lettera il grado di perpendicolarità delle frecce rosse. Analogamente, per conoscere il legame matematico preciso tra gli assi di partenza e quelli principali, bisogna leggere i loadings, di cui parleremo nella prossima sezione.*

Una prima analisi qualitativa può essere svolta in base ai rapporti tra i vettori che rappresentano i nostri indicatori (ortogonalità, parallelismo con versi concordi o discordi, ecc.), e ai raggruppamenti e alle posizioni dei dati. Nel nostro esempio, guardando la figura, alcune delle considerazioni che possiamo fare (per quanto naturali e più o meno note, visto che conosciamo abbastanza bene la situazione nazionale del benessere) sono:

- SC, TD e SA.SC sono tutti essenzialmente paralleli, a indicare una forte correlazione tra di loro: potremmo ad esempio leggere la loro direzione comune come un indicatore complessivo di benessere economico.

- Il verso di SC e opposto a quelli di TD e SA.SC, segno che questi indicatri sono correlati negativamente: come ci si aspetta, una maggior disoccupazione media si riflette su una minore spesa complessiva media (a TD alto corrisponde SC basso, e viceversa), mentre se la spesa complessiva media e molto bassa questa sarà, come e naturale, in gran parte dedicata agli alimentari (a SC basso corrsiponde SA.SC alto, e viceversa). Allo stesso modo, la correlazione positiva tra TD e SA.SC indica che nelle zone di più alta disoccupazione le (poche) spese sono destinate per lo più ai generi alimentari.

- PLIC e TM sono abbastanza paralleli tra loro (in analogia a quanto visto sopra potremmo leggere la loro direzione comune come un indicatore complessivo di salute), ma correlati negativamente, come è naturale.

- PLIC e TM sono abbastanza perpendicolari agli altri indicatori, segno che i due gruppi, e quindi le due direzioni indicatore complessivo di benessere economico e indicatore complessivo di salute, sono abbastanza scorrelati tra loro. Tuttavia notiamo le lievi correlazioni positive nelle direzioni che ci aspettiamo: maggiori posti letto dove ci sono maggiori spese complessive, e maggior mortalità infantile dove c'è più disoccupazione e le spese sono in prevalenza alimentari.

- L'area di maggior benessere e quella nella direzione positiva di SC, con un po' di spostamento verso PLIC. In tale zona si trovano fortemente raggruppate varie regioni (Veneto, Trentino Alto Adige, Lombardia, Piemonte, Emilia Romagna, Marche e Toscana), che pertanto risultano molto simili rispetto agli indicatori considerati.

- Le altre regioni del centro-nord (Liguria, Friuli, Lazio) non eccellono in SC ma eccellono in PLIC, a indicare una buona cura sanitaria nonostante un tenore di vita medio più modesto rispetto al gruppo precedente.

- Particolarmente negativo, sia rispetto all'asse del benessere economico che a quello della salute, risulta il raggruppamento composto da Campania, Sicilia, Basilicata e Puglia, in maniera molto più accentuata rispetto ad altre regioni meridionali o insulari (come Calabria e Sardegna) che nell'immaginario colettivo potremmo invece credere ad esse simili. Questo potrebbe indicare uno sforzo di miglioramento di alcune regioni, e potrebbe ad esempio suggerire l'analisi di altri dati più mirati per averne verica o smentita.

L'orientazione delle variabili di partenza rispetto alle componenti principali può inoltre suggerire delle potenziali interpretazioni delle due componenti principali. E' ragionevole associare e_1 alle tre variabili SC, SA.SC e TD, in quanto ha componenti maggiori in tali direzioni. Allo stesso modo, ha senso associare e_1 a PLIC e TMI. Una possibile interpretazione delle prime due componenti principali, cioè delle nuove

variabili aleatorie, potrebbe quindi essere quella dove la prima descrive il benessere di topo economico e la seconda quello sanitario.

Nell'esempio sugli indicatori di benessere, possiamo così verificare quanto avevamo già stabilito: la forte correlazione (con il giusto segno) tra SC, SA.SC e TD, l'assenza di legame tra PLIC e SC e TD, la correlazione negativa ma non troppo marcata tra PLIC e TMI, e via dicendo. Notiamo, rispetto a quanto già detto basandoci sulla figura, la correlazione (anche se non forte) di TMI non solo con PLIC, ma quasi allo stesso modo anche con le tre variabili economiche, negativa o positiva nel modo che ci aspettiamo.

Per quanto appena detto, può essere una buona idea affiancare l'esplorazione visiva offerta da PCA agli altri metodi quando si voglia creare un modello regressivo, in particolare quando si voglia ridurne la dimensione: oltre a vedere i valori della matrice di correlazione ed i p-value, conviene avere un ulteriore controllo offerto dal biplot di PCA. Si intende che PCA può essere applicata alla tabella contenente anche la variabile di output.

2.2.3 Il metodo PCA per capire la dimensione di un problema

Con "dimensione" di un problema intendiamo, intuitivamente parlando, un'indicatore della sua complessità. I cinque indicatori socio-economici del solito esempio, parlano di due cose diverse o di più di due? A stretto rigore di cinque, ma in realtà descrivono una, due o più caratteristiche? Quanto complesso è il fenomeno da essi descritto?

Il comando `plot(pca)` illustra la varianza lungo le diverse componenti principali, cioè i numeri $\lambda_1, ..., \lambda_p$, da cui e possibile farsi un'idea della dimensione dei dati, cioè di quante componenti sono necessarie o utili per analizzare i dati. Tornando al nostro solito esempio, è chiaro come Comp.4 e Comp.5 siano inutili, e quindi la dimensione dei dati sia 2 o 3. Questo significa che l'ellissoide 5-dimensionale dei dati ha in realtà solo 2 o 3 dimensioni effettive, e quindi che una rappresentazione ottimale dei dati si ottiene con una opportuna proiezione in dimensione 2 o 3. Detto altrimenti, per rappresentare le 5 variabili iniziali in realtà bastano solo 2 o 3 variabili aleatorie (cioè Comp.1, Comp.2 e, eventualmente, Comp.3).

Si possono avere i dati numerici precisi con il comando

`summary(pca)`

La prima riga riporta la deviazione standard di ogni componente principale. Essendo le componenti principali tra loro scorrelate (si ricordi il Teorema 5), la varianza della somma delle nuove variabili aleatorie è la somma delle rispettive varianze: possiamo quindi calcolare per ogni componente principale la parte di varianza totale da essa spiegata, valore che viene riportato nella seconda riga. Ad esempio, per gli indicatori di benessere in esame, Comp.1 spiega circa il 67% della varianza totale, mentre Comp.2 e Comp.3 rispettivamente il 17% e l'11%. La terza riga riporta la varianza cumulativa, che e semplicemente la somma delle percentuali di varianza spiegata da quella com-

ponente principale e da tutte le precedenti (per cui e ovvio che l'ultima componente abbia varianza cumulativa 1).

La varianza spiegata cumulativa è il pricipale parametro dell'efficacia del metodo PCA, dato che quantifica quanto accurata e la visualizzazione dei dati data dal piano principale. Nel nostro esempio, le prime due componenti principali (cioè il piano principale) spiegano complessivamente l'84% della varianza totale, e quindi la rappresentazione e decisamente soddisfacente. Una rappresentazione tridimensionale, contando quindi anche Comp.3, sarebbe praticamente perfetta (95%!). In genere, si considera il metodo PCA efficace quando il piano principale rappresenta l'80 - 90% della varianza totale dei dati, cioè quando la parte di informazione persa (rappresentata dalla varianza delle altre componenti principali: Comp.3, Comp.4, eccetera) si aggira sul 10 - 20% del totale. Tuttavia, anche quando la rappresentazione bidimensionale data dal piano principale e insufficiente, il metodo PCA contribuisce comunque a comprendere meglio i dati analizzati, in particolare indicandone l'effettiva dimensione, cioè quante variabili al minimo bastano per rappresentarli efficacemente.

2.2.4 Domande

1. Spiegare il legame tra la teoria delle v.a. gaussiane e quella del metodo PCA (descrivete l'idea alla base di PCA, facendo riferimento alla teoria dei vettori gaussiani)

2. In PCA, definire le variabili V_i e dimostrare il Teorema $Var[V_i] = \lambda_i$

3. Cos'è il vettore principale calcolato dal metodo PCA

4. Quali operazioni svolge il software per trovare il piano principale e rappresentare i punti sperimentali?

5. Come si calcola la classifica delle unità sperimentali secondo l'asse principale?

6. Cosa rappresentano i numeri λ_i in PCA? Come si leggono col software?

7. Cosa rappresentano i numeri

$$\lambda_1 + ... + \lambda_p, \frac{\lambda_1}{\lambda_1 + ... + \lambda_p}, \frac{\lambda_1 + \lambda_2}{\lambda_1 + ... + \lambda_p}?$$

8. Mostrare che le componenti principali sono scorrelate (formulare precisamente quest'affermazione e dimostrarla).

2.3 Loadings, modelli lineari, analisi fattoriale

2.3.1 Loadings in PCA

In PCA sono in gioco due basi ortonormali: la base canonica $\mathbf{u}_1, ..., \mathbf{u}_p$ e la base dei vettori principali $\mathbf{e}_1, \ldots, \mathbf{e}_p$ (usiamo il neretto solo qui, per evidenziare maggiormente la differenza tra vettori e loro componenti; inoltre, per non appesantire la notazione, non indichiamo esplicitamente che $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sono empirici), autovettori di Q con autovalori $\lambda_1, \ldots, \lambda_p$. Scritti i vettori $\mathbf{e}_i$ in componenti

$$\mathbf{e}_i = \left(e_i^1, ..., e_i^p\right)$$

valgono le ovvie identità

$$\begin{aligned} \mathbf{e}_1 &= e_1^1\mathbf{u}_1 + ... + e_1^p\mathbf{u}_p \\ &... \\ \mathbf{e}_p &= e_p^1\mathbf{u}_1 + ... + e_p^p\mathbf{u}_p \end{aligned}$$

che esprimono il fatto che i vettori $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sono combinazioni lineari dei vettori $\mathbf{u}_1, ..., \mathbf{u}_p$. La tabella di numeri

$$U^T = \begin{pmatrix} e_1^1 & ... & e_1^p \\ ... & ... & ... \\ e_p^1 & ... & e_p^p \end{pmatrix}$$

è detta tabella dei *loadings*. Ricordando che $U^T = U^{-1}$, valgono anche le identità

$$\begin{aligned} \mathbf{u}_1 &= e_1^1\mathbf{e}_1 + ... + e_p^1\mathbf{e}_p \\ &... \\ \mathbf{u}_p &= e_1^p\mathbf{e}_1 + ... + e_p^p\mathbf{e}_p \end{aligned}$$

che esprimono i vettori $\mathbf{u}_1, ..., \mathbf{u}_p$ come combinazioni lineari dei vettori $\mathbf{e}_1, \ldots, \mathbf{e}_p$.

Nel seguito di questo paragrafo confondiamo un po' $\mathbf{u}_1, ..., \mathbf{u}_p$ con le variabili $X_1, ..., X_p$, e $\mathbf{e}_1, \ldots, \mathbf{e}_p$ con $V_1, ..., V_p$. La ragione è che, oltre all'aspetto intuitivo, valgono esattamente le stesse relazioni:

$$\begin{aligned} V_1 &= e_1^1X_1 + ... + e_1^pX_p \\ &... \\ V_p &= e_p^1X_1 + ... + e_p^pX_p \end{aligned}$$

$$\begin{aligned} X_1 &= e_1^1V_1 + ... + e_p^1V_p \\ &... \\ X_p &= e_1^pV_1 + ... + e_p^pV_p \end{aligned}$$

e quindi tutti i ragionamenti su $\mathbf{u}_1, ..., \mathbf{u}_p$ e $\mathbf{e}_1, \ldots, \mathbf{e}_p$ si applicano inalterati a $X_1, ..., X_p$ e $V_1, ..., V_p$. [Si riconosca che valgono queste identità.]

Vediamo quindi che, ad esempio, il numero (loading) e_1^1 è (in entrambe le tabelle) il coefficiente che lega $\mathbf{e}_1$ con $\mathbf{u}_1$; analogamente, e_1^p è quello che lega $\mathbf{e}_1$ con $\mathbf{u}_p$; e così via: in generale, il loading e_i^j è il numero che lega $\mathbf{e}_i$ con $\mathbf{u}_j$. Esso esprime l'intensità del legame tra queste due variabili (V_i e X_j). Se entrambe le variabili hanno un significato, es. $\mathbf{u}_j$ (ovvero X_j)è il tasso di occupazione ed $\mathbf{e}_i$ (ovvero V_i) è la brillantezza economica, il loading e_i^j quantifica il legame tra esse; e lo quantifica in rapporto agli altri legami: perché la brillantezza economica $\mathbf{e}_i$ è legata a tutte le variabili $\mathbf{u}_1, ..., \mathbf{u}_p$, con diversi coefficienti $e_i^1, ..., e_i^p$. Quindi i loadings quantificano il legame tra variabili.

In un certo senso, abbiamo costruito dei modelli di tipo regressivo

$$\mathbf{e}_i = e_i^1 \mathbf{u}_1 + ... + e_i^p \mathbf{u}_p$$

che esprimono l'influenza (non necessariamente nel senso di causa-effetto) delle variabili $\mathbf{u}_1, ..., \mathbf{u}_p$ sulle $\mathbf{e}_1, \ldots, \mathbf{e}_p$ o viceversa

$$\mathbf{u}_i = e_1^i \mathbf{e}_1 + ... + e_p^i \mathbf{e}_p$$

Tuttavia, questo tipo di modelli non è un sottocaso della regressione lineare multipla, per la seguente fondamentale ragione: mentre le variabili $\mathbf{u}_1, ..., \mathbf{u}_p$ sono quelle osservate sperimentalmente, quelle della tabella di dati, le variabili $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sono create dalla matematica, dal metodo PCA, non erano state osservate e misurate. Invece nella regressione possiamo misurare sia gli input sia l'output ed usiamo entrambe le misurazioni per costruire il modello. Qui invece è usccessa una cosa per certi versi sorprendente: abbiamo creato un modello sulla base delle sole misurazioni degli input, o degli output!

Le utilità dei loadings sono almeno due. La prima è quella di quantificare il legame tra le $\mathbf{u}_i$ e le $\mathbf{e}_j$ permettendo di creare associazioni (es. PLIC e TMI più associate a $\mathbf{e}_2$, mentre TD, SC, SA.SC più associate a $\mathbf{e}_1$, nel nostro solito esempio), dalle quali possono nascere delle interpretazioni delle $\mathbf{e}_i$ ($\mathbf{e}_1$ più legata all'aspetto prettamente economico, $\mathbf{e}_2$ a quello sanitario). Naturalmente il biplot di PCA già forniva questo tipo di informazioni attraverso parallelismi e perpendicolarità delle frecce tra loro e rispetto agli assi principali, ma i loadings quantificano questo numericamente ed oltretutto lo esprimono anche rispetto alle altre componenti $\mathbf{e}_3$ ecc.

La seconda è quella di creare modelli, simili a quelli regressivi, di legame tra variabili, che esploreremo più a fondo nei prossimi paragrafi.

I loadings, con R, si trovano col semplice comando:

```
> loadings(B)
Loadings:
```

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
PLIC	-0.310	0.769	-0.553		
SC	-0.491	-0.309		-0.813	
SA.SC	0.512	0.216	0.120	-0.433	-0.699
TD	0.506	0.279	0.115	-0.381	0.713
TMI	0.380	-0.435	-0.816		

Abbiamo riportato il risultato relativo ai soliti indicatori PLIC ecc. Si noti che il software non mette alcun numero quando il valore è molto piccolo. Esso non è esattamente nullo, come si può verificare col comando

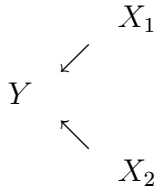
```
> loadings(B)[1,5]
[1] -0.03449247
>
```

La ragione di questa omissione del software è che così diventa più facile per l'utilizzatore riconoscere quali variabili incidono maggiormente su altre. Nel nostro esempio però questo è solo moderatamente utile perché si tratta delle componenti principali meno significative. Intendiamo dire che, per lo scopo di dare un nome, interpretare $\mathbf{e}_1$ ed $\mathbf{e}_2$, questi vuoti non aiutano. Invece, possono aiutare per fare rgionamenti di secondo livello: le tre variabili TD, SC, SA.SC, più associate a $\mathbf{e}_1$, contengono però più informazione della singola variabile $\mathbf{e}_1$, non sono completamente uni-dimensionali. Hanno un carattere multidimensionale catturato da $\mathbf{e}_4$ ed $\mathbf{e}_5$ (mentre $\mathbf{e}_2$ ed $\mathbf{e}_3$ sono quasi inattive, rispetto a TD, SC, SA.SC). Il vettore $\mathbf{e}_4$ cattura quella parte di SC non spiegata da $\mathbf{e}_1$, mentre $\mathbf{e}_5$ cattura quella parte di TD e SA.SC non spiegata da $\mathbf{e}_1$.

2.3.2 Analisi fattoriale

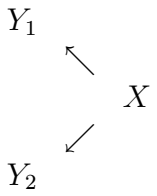
Un fattore, più risposte

La regressione lineare mette in relazione più fattori (predittori, input) con una risposta (output); tutti misurabili per ciascun individuo del "training set", l'insieme di individui o dati che serve per costruire il modello. Successivamente il modello può essere utilizzato su un "test set", in cui siano state misurate solo le variabili di input per i nuovi individui; il modello fornirà il valore dell'output. Oppure, se degli individui del "test set" è misurato anche l'output, questi nuovi dati possono essere utilizzati per valutare la bontà del modello (proprio nela filosofia di un test). Il diagramma



descrive la situazione con più input ed un output.

Una situazione diametralmente opposta alla precedente è quella in cui c'è un solo fattore X e diverse risposte Y_i, ad esempio tre grandezze X, Y_1, Y_2 tali che Y_1 sia influenzata da X ed anche Y_2 sia influenzata dallo stesso X.



Se disponiamo di dati sperimentali per tutte queste grandezze, basta esaminare un modello lineare per la coppia (X, Y_1) e poi un'altro, separatamente, per la coppia (X, Y_2). In questo caso non ci sarebbe nulla di nuovo rispetto alla regressione già studiata. Il discorso si generalizza senza problemi al caso di più input e più output.

Completamente diverso invece è il caso in cui le due grandezze Y_1, Y_2 sono misurabili, disponiamo di loro dati sperimentali, mentre X non è misurabile, anzi forse non è nemmeno completamente ben definita. Immaginiamo ad esempio, sempre con riferimento all'esempio della Sezione 2.2, di sospettare che ci sia un fattore X che influenza Y_1 =SA.SC e Y_2 =SC. Abbiamo le misurazionii delle grandezze Y_1 e Y_2 ma non di X, di cui anzi non ci è chiaro nemmeno il significato. la domanda è: c'è un fattore, che ancora dobbiamo scoprire, mettere allo scoperto, che influenza entrambe SA.SC e SC? Un *fattore nascosto*? Che *spiega* (si suol dire) la variabilità di SA.SC e SC? Cosa c'è dietro il fatto che certe regioni hanno una minore spesa complessiva familiare ed una maggior proporzione di spesa per alimentari, rispetto ad altre regioni in cui queste grandezze sono invertite?

A modo suo, *il metodo PCA serve anche a questo scopo*: esso ha individuato una nuova grandezza aleatoria, Comp1, a cui abbiamo attribuito un significato del tipo "benessere economico", legata ad entrambe SC e SA.SC. Discuteremo sotto il metodo dell'analisi fattoriale, alternativo a PCA, tornando però anche su PCA in relazione al problema ora posto.

Esempi di Analisi Fattoriale

Consideriamo alcuni esempi semplicissimi di Analisi Fattoriale (FA, Factorial Analysis), col solo scopo di far capire alcune idee strutturali del problema.

Consideriamo il modello

$$Y_1 = a_1 X + b_1 + \varepsilon_1$$
$$Y_2 = a_2 X + b_2 + \varepsilon_2$$

cioè un fattore e due output. Ma immaginiamo di avere dati solo delle variabili (Y_1, Y_2). Anzi, X non sappiamo nemmeno a priori cosa sia, che variabile sia, se ci sia. E' possibile risalire alla X ed ai coefficienti del modello?

Le incognite sono $a_1, a_2, b_1, b_2, \varepsilon_1, \varepsilon_2$. Se supponiamo di conoscere le medie di Y_1 e Y_2 e supponiamo convenzionalmente che X, ε_1, ε_2, abbiano media nulla, allora $b_1 = E[Y_1]$, $b_2 = E[Y_2]$, e quindi le incognite davvero difficili da determinare sono $a_1, a_2, \varepsilon_1, \varepsilon_2$.

Il problema è quello descritto sopra: misuriamo due variabili Y_1, Y_2, magari ben correlate, ma che la logica ci dice non essere in relazione causa-effetto. Ci chiediamo invece se ci sia, alle loro spalle, a monte di esse, una variabile X che sia loro causa, che le "spieghi", nel senso che spieghi come mai Y_1 ed Y_2 variano in modo coordinato (sono correlate). Si tratta di *spiegare le variazioni (coordinate) degli output*. In termini matematici, *spiegare la matrice di covarianza* Q_Y di Y.

Abbiamo enfatizzato il problema illustrando il caso in cui X sia *causa* di Y_1 e Y_2, ma non è necessario che sia proprio una relazione di causa-effetto. Magari si tratta solo di rintracciare una variabile riassuntiva X, di cui Y_1 e Y_2 siano manifestazioni misurabili. Ad esempio X può essere il grado di benessere economico, e le Y_i essere varie misurazioni di indicatori di benessere (spese per cultura, per vacanze ecc.).

Si noterà che in tutti i nostri esempi prendiamo sempre meno fattori che output, altrimenti varrebbe la risposta banale: un fattore per ogni output. Se accettassimo di cercare un numero di fattori pari (o addirittura superiore) agli output, fattori che spieghino gli output, la risposta banale sarebbe prendere come fattori gli output stessi. Essi spiegherebbero perfettamente tutta la variabilità degli output. Solo imponendo il vincolo che i fattori sono di meno, sopravvive un problema non ovvio di spiegare delle variazioni coordinate degli output.

Se abbiamo una tabella di dati per le variabili $Y = (Y_1, Y_2)$ calcoliamo dai dati la matrice di correlazione

$$Q_Y = \begin{pmatrix} \sigma^2_{Y_1} & Cov(Y_1, Y_2) \\ Cov(Y_1, Y_2) & \sigma^2_{Y_2} \end{pmatrix}.$$

Non abbiamo altro (eventualmente i valori medi delle Y_1, Y_2) per tentare di risalire al modello.

A livello teorico, se vale un modello di questo genere, con $\varepsilon_1, \varepsilon_2, X$ indipendenti (ricordiamo che questa richiesta, nella regressione, rimpiazzava la minimizzazione dei quadrati), vale

$$\begin{aligned} Cov(Y_1, Y_2) &= a_1 a_2 \sigma^2_X \\ \sigma^2_{Y_1} &= a_1^2 \sigma^2_X + \sigma^2_{\varepsilon_1} \\ \sigma^2_{Y_2} &= a_2^2 \sigma^2_X + \sigma^2_{\varepsilon_2}. \end{aligned}$$

Supponiamo $\sigma^2_X = 1$ altrimenti questa grandezza la si fa rientrare nei coefficenti incogniti a_1 e a_2. Quindi

$$\begin{aligned} a_1 a_2 &= Cov(Y_1, Y_2) \\ a_1^2 + \sigma^2_{\varepsilon_1} &= \sigma^2_{Y_1} \\ a_2^2 + \sigma^2_{\varepsilon_2} &= \sigma^2_{Y_2}. \end{aligned}$$

Sono tre equazioni nelle quattro incognite $(a_1, a_2, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2})$. Il problema è sottodeterminato. Ci sono quindi (almeno in linea di principio, visto che è un problema nonlineare, quindi non del tutto banale) infinite soluzioni. Il software cerca quella che rende minima la somma dei residui $\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2}$.

Se però erano tre output ed un solo fattore, cioè il modello

$$\begin{aligned} Y_1 &= a_1 X + b_1 + \varepsilon_1 \\ Y_2 &= a_2 X + b_2 + \varepsilon_2 \\ Y_3 &= a_3 X + b_3 + \varepsilon_3 \end{aligned}$$

avevamo

$$\begin{aligned} Cov\left(Y_1, Y_2\right) &= a_1 a_2 \\ Cov\left(Y_1, Y_3\right) &= a_1 a_3 \\ Cov\left(Y_2, Y_3\right) &= a_2 a_3 \\ \sigma^2_{Y_1} &= a_1^2 + \sigma^2_{\varepsilon_1} \\ \sigma^2_{Y_2} &= a_2^2 + \sigma^2_{\varepsilon_2} \\ \sigma^2_{Y_3} &= a_3^2 + \sigma^2_{\varepsilon_3}. \end{aligned}$$

Sono 6 equazioni nelle 6 incognite $(a_1, a_2, a_3, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}, \sigma_{\varepsilon_3})$ per cui in linea di principio c'è una sola soluzione. Può accadere in alcuni casi che ci siano più soluzioni o che non ce ne sia alcuna; se ce ne sono più d'una, il software determina quella che rende minima la somma dei residui $\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_3}$.

Con 4 output certamente è sovradeterminato: ci sono 4 equazioni del tipo $\sigma^2_{Y_i} = a_i^2 + \sigma^2_{\varepsilon_i}$ e 3+2+1=6 equazioni del tipo $Cov\left(Y_i, Y_j\right) = a_i a_j$. Quindi 10 equazioni nelle 4+4=8 incognite $a_1, ..., a_4, \sigma_{\varepsilon_1}, ..., \sigma_{\varepsilon_4}$. In questi casi, di non risolubilità, il criterio è costruire con i parametri $(a_1, a_2, a_3, a_4, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}, \sigma_{\varepsilon_3}, \sigma_{\varepsilon_4})$ una matrice di covarianza più vicina possibile (in una certa metrica) alla matrice Q_Y. Si costruisce cioè la matrice di covarianza $Q_{\widetilde{Y}}$ del vettore $\widetilde{Y} = \left(\widetilde{Y}_1, ..., \widetilde{Y}_4\right)$ dato da $\widetilde{Y}_i = a_i X + b_i + \varepsilon_i$; tael matrice è funzione dei coefficienti $(a_1, a_2, a_3, a_4, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}, \sigma_{\varepsilon_3}, \sigma_{\varepsilon_4})$; e, scelta un'opportuna *distanza tra matrici*, indicata con $d\left(\cdot, \cdot\right)$, cerchiamo i coefficienti che minimizzano $d\left(Q_{\widetilde{Y}}, Q_Y\right)$.

Vediamo anche il caso di due fattori e tre output:

$$\begin{aligned} Y_1 &= a_{11} X_1 + a_{12} X_2 + b_1 + \varepsilon_1 \\ Y_2 &= a_{21} X_1 + a_{22} X_2 + b_2 + \varepsilon_2 \\ Y_3 &= a_{31} X_1 + a_{32} X_2 + b_3 + \varepsilon_3 \end{aligned}$$

Qui vale, sempre prendendo i fattori standardizzati, e supponendoli indipendenti tra

loro e dagli errori,

$$Cov\left(Y_1, Y_2\right) = a_{11}a_{21} + a_{12}a_{22}$$
$$Cov\left(Y_1, Y_3\right) = a_{11}a_{31} + a_{12}a_{32}$$
$$ecc.$$

cioè 6 equazioni in 9 incognite. E' un sistema sottodeterminato, quindi si cerca la soluzione che minimizza $\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_3}$.

Si noti che il numero di equazioni si può contare nel seguente modo. La matrice di covarianza Q_Y è quadrata, ed ogni suo elemento corrisponde ad un'equazione del tipo $Cov\left(Y_i, Y_j\right) = ...$, con l'accortezza però che la parte triangolare inferiore è una copia di quella superiore (la matrice è simmetrica). Pertanto, se Q_Y è 3x3, ha 3 elementi sulla diagonale più 3 nel triangolo superiore, quindi 6: il sistema avrà 6 equazioni. Se Q_Y è 4x4, ha 4 elementi sulla diagonale più 6 nel triangolo superiore, quindi 10: il sistema avrà 10 equazioni. D'altra parte, il conteggio dei coefficienti incogniti è piuttosto ovvio dopo aver scritto il modello. Quindi si riesce a calcolare facilmente se siamo in un caso sovra o sotto determinato, o con soluzione unica (fermo restando che, come abbiamo detto, a causa della natural nolineare delle equazioni, possono aversi varie anomalie).

Forma matriciale del problema

Si può sintetizzare tutto con le matrici. Immaginiamo le variabili Y_i raccolte nel vettore aleatorio $Y = (Y_1, ..., Y_n)$, le X_i nel vettore $X = (X_1, ..., X_d)$, gli errori ε_i nel vettore $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$, quindi

$$Y = AX + b + \varepsilon.$$

I parametri incogniti sono i coefficienti della matrice A e le varianze $\sigma^2_{\varepsilon_i}$, che raggruppiamo nella matrice diagonale

$$Q_\varepsilon = \begin{pmatrix} \sigma^2_{\varepsilon_1} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sigma^2_{\varepsilon_n} \end{pmatrix}.$$

Nota è la matrice di covarianza dell'output, Q_Y, e le medie $E\left[Y_i\right]$ da cui si calcoleranno immediatamente i b_i. Le ipotesi sono che le variabili $X_1, ..., X_d, \varepsilon_1, ..., \varepsilon_n$ sono tutte indipendenti tra loro ed a media nulla; e che $Var\left[X_i\right] = 1$ per ciascuna X_i.

La forma matriciale, anche se a prima vista più oscura degli esempi particolari, sintetizza l'idea base della FA: abbiamo un vettore aleatorio Y, in pratica un set di variabili aleatorie $Y_1, ..., Y_n$, di cui conosciamo la variabilità (la matrice di covarianza) e vogliamo capire se questa variabilità, se i valori congiunti delle v.a. Y_i, possono derivare dalla variazione di meno grandezze, poche grandezze riassuntive $X_1, ..., X_d$ con $d < n$; in modo lineare, cioè nella forma $Y = AX + b$; tuttavia, come sempre, accettiamo che la relazione valga a meno di un errore, quindi nella forma $Y = AX + b + \varepsilon$.

Dobbiamo, data Q_Y, determinare A e Q_ε. Se non ci fosse il vettore ε, conosceremmo la regola $Q_Y = AQ_XA^T$. Il vettore ε è indipedente da X; usando questo fatto si può dimostrare che vale $Q_Y = AQ_XA^T + Q_\varepsilon$. Inoltre, Q_X è l'identità in quanto X ha componenti indipedenti e standard (è come un vettore gaussiano standard). Quindi, dal modello e dalle ipotesi otteniamo la relazione

$$Q_Y = AA^T + Q_\varepsilon.$$

Questa è un'equazione matriciale nelle incognite A e Q_ε.

Se $n = d$, possiamo prendere come soluzione $A = \sqrt{Q_Y}$, $Q_\varepsilon = 0$. Il problema della relazione $Y = AX+b+\varepsilon$ è risolto come nel capitolo sulle gaussiane, pensando a Y come ad una gaussiana generica, avente covarianza nota Q_Y, che vogliamo rappresentare nella forma $Y = AX$ con X gaussiano standard. La soluzione è $A = \sqrt{Q_Y}$, e non c'è bisogno di introdurre errori ε.

La vera FA nasce quando $n > d$ e quindi la soluzione di $AA^T + Q_\varepsilon = Q_Y$ non è ovvia. Il software calcola A e Q_ε con le strategie illustrate sopra nei casi particolari. A seconda dei valori di d ed n, il problema è risolubile univocamente, oppure per infinite matrici A, Q_ε (ed in tal caso si minimizza la varianza globale dell'errore), oppure non è risolubile esattamente, nel qual caso si cercano A e Q_ε tali che

$$d\left(Q_Y, AA^T + Q_\varepsilon\right)$$

sia minima, avendo indicato con $d(\cdot,\cdot)$ un'opportuna distanza tra matrici.

Loadings e comandi del software. Rotazioni e interpretazioni

La matrice A è detta matrice dei *loadings*, esattamente come per PCA. I suoi coefficienti sono i numeri che legano le variabili ausiliarie X_i alle variabili originarie misurate Y_j. Col comando `factanal(A,d)` del software R si ottengono i loadings di una FA con d fattori nascosti $X_1, ..., X_d$.

I numeri $\sigma^2_{\varepsilon_1}, ..., \sigma^2_{\varepsilon_n}$ sono detti *unicità* (uniqueness) delle variabili Y_j: $\sigma^2_{\varepsilon_1}$ è l'unicità della variabile Y_1, e così via. Anch'essi compaiono con `factanal(A,d)`. Per inciso,

$$\sigma^2_{Y_1} - \sigma^2_{\varepsilon_1} = a^2_{11} + ... + a^2_{1d}$$

è detta *communality* della v.a. Y_1, e così via.

Analogamente a PCA, si può parlare di varianza spiegata. Anch'esse compaiono con quel comando. [La varianza spiegata cumulativa di tutto il modello è $1-\frac{\sigma^2_{\varepsilon_1}+...+\sigma^2_{\varepsilon_n}}{\sigma^2_{Y_1}+...+\sigma^2_{Y_n}}$.]

Confrontando su un singolo esempio l'esito di PCA e di FA si nota una forte somiglianza. Si deve paragonare ad esempio `factanal(A,2)` con le prime due componenti principali di PCA. Sia i loadings sia le varianze spiegate, sono abbastanza simili, pur non uguali (lo discuteremo teoricamente tra un momento).

La soluzione dell'equazione $Q_Y = AA^T + Q_\varepsilon$ non è unica. Supponiamo di aver trovato una soluzione (A, Q_ε). Sia U una matrice ortogonale, un cambio di base, una *rotazione*, tale che UU^T è l'identità. Allora (AU, Q_ε) è un'altra soluzione:

$$(AU)(AU)^T + Q_\varepsilon = AUU^TA^T + Q_\varepsilon = AA^T + Q_\varepsilon = Q_Y.$$

In termini di modello $Y = AX + b + \varepsilon$ si tratta di averlo scritto nella forma

$$\begin{aligned} Y &= (AU)X' + b + \varepsilon \\ X' &= U^TX. \end{aligned}$$

In altre parole, "ruotando" i fattori e modificando A, si risolve il problema nello stesso modo. Che vantaggio può avere una soluzione rispetto ad un'altra, che differiscano per una rotazione? Se si riesce a trovare una rotazione in cui A sia particolarmente ricca di zeri (o valori molto piccoli), questo può venire a vantaggio di una buona *interpretazione* dei fattori, dell'attribuire un significato ai fattori. Ragioniamo su un esempio.

Si pensi al solito esempio degli indicatori PLIC, SC ecc. Suggestionati dall'analisi svolta con PCA, che suggerisce la presenza di due fattori, immaginiamo ci siano appunto due fattori X_1, X_2 che influenzano le variabili PLIC, SC ecc., che spiegano la particolare struttura di variabilità di queste grandezze tra le regioni italiane:

$$\begin{aligned} PLIC &= a_{11}X_1 + a_{12}X_2 + b_1 + \varepsilon_1 \\ SC &= a_{21}X_1 + a_{22}X_2 + b_2 + \varepsilon_2 \\ &ecc. \end{aligned}$$

Immaginiamo di eseguire una FA e di trovare una soluzione (A, Q_ε). Il software, nel calcolare (A, Q_ε), ovviamente ignora ogni possibile interpretazione applicativa (per il SW, che si parli di PLIC o di risultati calcistici è la stessa cosa). Quindi, a priori, il SW non aiuta lo studioso a dare un'interpretazione dei risultati. Ma supponiamo che tramite una rotazione si ottenga una matrice A con numeri nettamente distinti, in numeri grandi e numeri piccoli. Un loading piccolo significa che c'è poca relazione tra il fattore e la variabile che esso lega. Ad esempio, se venisse che a_{11} è piccolo, vorrebbe dire che il fattore X_1 non è legato a PLIC, ma serve per spiegare le altre variabili. Questo minore o maggiore grado di associazione di un fattore a certe variabili può contribuire a dare un nome, un significato, a quel fattore.

Le rotazioni col software R si operano col comando `varimax`.

FA e PCA

Sperimentalmente si vede una somiglianza tra PCA e FA. Concettualmente, entrambe le metodologie identificano fattori nascosti a partire da un set di variabili osservabili. C'è un legame matematico preciso? Per rispondere a questa domanda dobbiamo ricordare il sistema

$$\begin{aligned} X_1 &= e_1^1 V_1 + ... + e_p^1 V_p \\ &... \\ X_p &= e_1^p V_1 + ... + e_p^p V_p \end{aligned}$$

visto in un paragrafo precedente su PCA. Esso dice che le v.a. originarie $X_1, ..., X_p$ a cui applichiamo PCA vengono, da PCA, rappresentate come combinazione lineare di nuove variabili $V_1, ..., V_p$, che corrispondono alle proiezioni sugli assi principali. Cambiamo notazioni, per capire il legame con FA. Supponiamo di avere le variabili osservate $Y_1, ..., Y_n$ e di applicare PCA. Otterremo

$$\begin{aligned} Y_1 &= e_1^1 V_1 + ... + e_n^1 V_n \\ &... \\ Y_n &= e_1^n V_1 + ... + e_n^n V_n \end{aligned}$$

Supponiamo per semplicità espositiva di voler eseguire una FA a due fattori. Evidenziamo nel sistema precedente i primi due fattori, quelli con maggior variabilità:

$$\begin{aligned} Y_1 &= e_1^1 V_1 + e_2^1 V_2 + ... + ... \\ &... \\ Y_n &= e_1^n V_1 + e_2^n V_2 + ... + ... \end{aligned}$$

Ora cambiamo ulteriormente notazione: indichiamo V_1 e V_2 con X_1 e X_2 e indichiamo tutta la somma sopra indicata con $+... + ...$ con $\varepsilon_1, ..., \varepsilon_n$:

$$\begin{aligned} Y_1 &= e_1^1 X_1 + e_2^1 X_2 + \varepsilon_1 \\ &... \\ Y_n &= e_1^n X_1 + e_2^n X_2 + \varepsilon_n. \end{aligned}$$

Questo è precisamente un modello FA a due fattori! Quindi PCA produce modelli di tipo FA.

In realtà questo non è completamente vero, c'è una piccola differenza. In FA richiediamo che le v.a. $X_1, X_2, \varepsilon_1, ..., \varepsilon_n$ siano tutte indipendenti tra loro. Vediamo se questa condizione è soddisfatta dalla soluzione trovata con PCA. Supponiamo che tutti i vettori in gioco siano gaussiani. Allora, per un noto teorema di PCA, V_i e V_j sono indipendenti per $i \neq j$. Quindi $X_1, X_2, V_3, ..., V_n$ sono indipendenti tra loro; in particolare X_1 e X_2 sono indipendenti tra loro ed indipendenti dalle $\varepsilon_1, ..., \varepsilon_n$, perché quest'ultime sono composte da $V_3, ..., V_n$. Ma $\varepsilon_1, ..., \varepsilon_n$ non sono indipendenti tra loro! Infatti dipendono tutti dalle stesse variabili $V_3, ..., V_n$ (ciascun ε_j dipende da tutte le $V_3, ..., V_n$). Questa è l'unica differenza tra PCA e FA: gli errori di PCA (interpretato

nel modo appena visto come modello con due o pochi fattori) non sono indipendenti tra loro. Detto altrimenti, la matrice Q_ε non è diagonale.

[Osserviamo un'altro dettaglio, però secondario, più convenzionale che altro: in FA richiediamo convenzionalmente che $Var[X_i] = 1$; invece le v.a. X_1, X_2 prodotte da PCA hanno varianze pari a λ_1, λ_2. Se vogliamo rispettare la convenzione di FA dobbiamo inglobare un fattore $\sqrt{\lambda}$ nei coefficienti, cioè riscrivere il sistema precedente nella forma

$$\begin{aligned} Y_1 &= e_1^1\sqrt{\lambda_1}\left(\frac{X_1}{\sqrt{\lambda_1}}\right) + e_2^1\sqrt{\lambda_2}\left(\frac{X_2}{\sqrt{\lambda_2}}\right) + \varepsilon_1 \\ &\dots \\ Y_n &= e_1^n\sqrt{\lambda_1}\left(\frac{X_1}{\sqrt{\lambda_1}}\right) + e_2^n\sqrt{\lambda_2}\left(\frac{X_2}{\sqrt{\lambda_2}}\right) + \varepsilon_n. \end{aligned}$$

Le nuove v.a. $\frac{X_1}{\sqrt{\lambda_1}}$ e $\left(\frac{X_2}{\sqrt{\lambda_2}}\right)$ hanno varianza unitaria. I loadings di FA sono quindi

$$A = \begin{pmatrix} e_1^1\sqrt{\lambda_1} & e_2^1\sqrt{\lambda_2} \\ \dots & \dots \\ e_1^n\sqrt{\lambda_1} & e_2^n\sqrt{\lambda_2} \end{pmatrix}$$

invece che $\begin{pmatrix} e_1^1 & e_2^1 \\ \dots & \dots \\ e_1^n & e_2^n \end{pmatrix}$ come in PCA.]

2.3.3 Domande

1. In che senso PCA può essere un modello a fattori non misurabili o nascosti? Scrivere anche le formule che legano i fattori alle variabili in uscita
2. Definire i loadings
3. A cosa possono servire i loadings?
4. Come scorrelare fattori allineati
5. Discutere la differenza tra FA e PCA come modelli a fattori nascosti
6. Tracciare il calcolo dei coefficienti di FA in un esempio particolare (teorico, non numerico).
7. Mostrare che una rotazione fornisce ancora una soluzione di FA.
8. Formulare il problema di FA in forma vettoriale e matriciale.

Capitolo 3

Classificazione e clustering

3.1 Regressione logistica

3.1.1 Premessa sul problema della classificazione in generale

Nei problemi di classificazione si hanno due classi C_1 e C_2 (o più di due) ed un individuo (o più individui) da classificare, cioè da assegnare ad una delle classi.

Per far questo, bisogna decidere una regola che, osservando alcune caratteristiche dell'individuo, lo assegni ad una delle classi.

Se le caratteristiche osservabili per l'individuo includono anche la classe stessa, non c'è nessun problema da risolvere (ad esempio, se dobbiamo decidere se un individuo ha i capelli biondi o neri e possiamo osservare i suoi capelli, l'assegnazione alla classe è ovvia).

Ci occuperemo quindi del caso in cui le caratteristiche osservabili siano in qualche modo indirette, rispetto alle classi. Si pensi ad esempio ad un medico che deve stabilire se un paziente è affetto da una certa malattia che sia direttamente osservabile solo tramite operazione chirurgica. In prima battuta, il medico vorrà scegliere tra le classi $C_1 =$ "sano", $C_2 =$ "malato" tramite esami non invasivi, come analisi del sangue o visite esterne (che però non daranno la certezza assoluta di aver eseguito la classificazione corretta).

Supporremo che le caratteristiche osservabili siano descritte da variabili $X_1, ..., X_p$, l'analogo dei fattori della regrassione. In un certo senso, parafrasando la regressione, l'output Y sarebbe la classe; questa è la visione adottata dalla regressione logistica che vedremo in questa prima sezione.

Quindi, per riassumere, dobbiamo inventare delle regole che permettano di assegnare un individuo ad una delle due classi C_1 o C_2 sulla base dei valori $x_1, ..., x_p$ che misuriamo per quell'individuo. Come costruire una regola?

L'idea è di usare n altri individui di cui si conoscano sia i valori $x_{i,1}, ..., x_{i,p}$, $i = 1, ..., n$ sia la classe, C_1 o C_2. Il gruppo di questi individui di classe nota viene chiamato

training set. Ogni metodo di classificazione descritto nel seguito (regressione logistica, analisi discriminante o altri) inventa una regola particolare basata sul training set, mediante la quale classifica ogni individuo successivo di cui si conoscano solo i valori $x_1, ..., x_p$. Il dato di partenza (training set) è quindi una tabella del tipo

	X_1	...	X_p	Classe
1	$x_{1,1}$	...	$x_{1,p}$	C_1
...	...	...	...	...
...	...	...	...	...
n_1	...	...	...	C_1
$n_1 + 1$	...	...	...	C_2
...	...	...	...	...
...	...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$	C_2

dove, degli individui, si conoscono i valori e la classe.

3.1.2 La soluzione offerta della regressione logistica

Definizione 8 *Un modello di regressione logistica tra p fattori $X_1, ..., X_p$ ed un output Y è una relazione del tipo*

$$Y \sim B(1, p)$$
$$g(p) = a_1 X_1 + ... + a_p X_p + b.$$

Abbiamo sintetizzato il modello in una definizione concisa perché questo concetto risulta in genere particolarmente oscuro e impreciso. Ora però cerchiamo di capirlo in modo più progressivo.

Come in tutti i modelli regressivi, anche nella regressione logistica ci sono dei fattori $X_1, ..., X_p$ misurabili, ed un output Y anch'esso misurabile, tutti relativamente ad un insieme di unità sperimentali. Tuttavia, nella regressione logistica l'output Y è dicotomico: 0 o 1, mentre i predittori assumono valori reali generici, come nella regressione lineare multipla tradizionale.

Si pensa che, dati i valori dei predittori, l'output $Y \in \{0, 1\}$ sia casuale ma con legge univocamente determinata dai predittori. Y è una v.a. di Bernoulli, quindi la sua legge è identificata dal parametro $p = P(Y = 1)$. Questo numero è univocamente determinato dai predittori, è funzione deterministica dei valori assunti dai predittori.

Inoltre, il modo di dipendere dai predittori, nel modello di regressione logistica, non è qualsiasi ma avviene solo attraverso una loro combinazione affine, detta *predittore lineare*

$$\eta = a_1 X_1 + ... + a_p X_p + b.$$

Non stiamo affermando che $p = \eta$, ma che p dipende da $X_1, ..., X_p$ solo attraverso una combinazione affine η di questo tipo, e non tramite espressioni magari quadratiche o altro.

Mettiamo a confronto regressione logistica (RLog) e regressione lineare multipla (RLM) tradizionale, per spiegare meglio il modello RLog. Nella RLM, dati i valori $x_1, ..., x_p$ dei predittori, noti i coefficienti $a_1, ..., a_p, b$, l'output è una v.a. gaussiana Y di media $\mu = \eta$ (media uguale al predittore lineare) e varianza σ^2, quindi rappresentabile nella forma

$$Y = a_1 x_1 + ... + a_p x_p + b + \varepsilon$$

con $\varepsilon \sim N(0, \sigma^2)$. Invece, nella RLog, dati i valori $x_1, ..., x_p$ dei predittori, noti i coefficienti $a_1, ..., a_p, b$, l'output è una v.a. di Bernoulli Y, di parametro p che dipende da η attraverso una certa funzione.

Pensiamo ad un esempio: supponiamo che gli individui siano le nazioni europee e che, per una certa nazione, sia $Y = 1$ se la nazione migliora la propria condizione economica ($Y = 0$ altrimenti) durante l'anno 2011. I predittori potrebbero essere gli investimenti in ricerca, e così via del 2010. Noti i valori dei predittori, la casualità non è certo esaurita, quindi Y resta aleatorio, ma la sua legge (cioè p) è ora determinata, nota. Nel modello RLog si suppone che la probabilità p di miglioramento sia nota quando sono noti i predittori. Inoltre si suppone che p dipenda dai predittori solo attraverso la loro combinazione affine η.

Un altro esempio: gli individui sono esemplari di complessi sistemi meccanici o elettronici, $Y = 1$ se il sistema funziona per un anno, i predittori possono essere valori misurati di caratteristiche meccaniche ecc. di sottoparti, del materiale ecc.

Essendo p una probabilità, non possiamo pensare che la relazione tra p ed η sia del tipo $p = \eta$, cioè

$$p = a_1 x_1 + ... + a_p x_p + b$$

altrimenti otterremmo per p valori anche esterni a $[0, 1]$. Si deve adottare un modello del tipo

$$g(p) = a_1 x_1 + ... + a_p x_p + b$$

dove g è una funzione definita in $[0, 1]$ a valori reali, invertibile. In modo che sia

$$p = g^{-1}(a_1 x_1 + ... + a_p x_p + b).$$

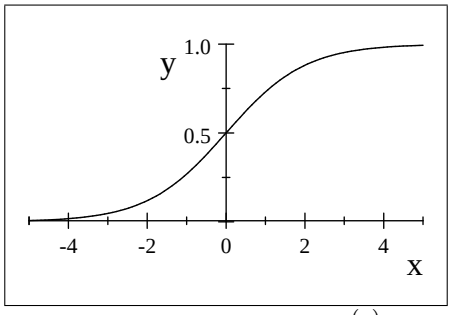
Una scelta molto comune è la funzione detta *logit*

$$g(p) = \log\left(\frac{p}{1-p}\right).$$

Per $p \to 0$ essa tende a $-\infty$, mentre per $p \to 1$ tende a $+\infty$; ed è strettamente crescente, oltre che regolare. La sua funzione inversa è

$$p = g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

rappresentata in figura:



Funzione logistica $\frac{\exp(\eta)}{1+\exp(\eta)}$.

Osservazione 29 *Verifichiamo che le due funzioni scritte sopra sono inversa una dell'altra:* $\log\left(\frac{p}{1-p}\right) = \eta$, $\frac{p}{1-p} = \exp(\eta)$, $p = (1-p)\exp(\eta)$, $p(1+\exp(\eta)) = \exp(\eta)$, $p = \frac{\exp(\eta)}{1+\exp(\eta)}$.

In definitiva, il modello è

$$Y \sim B(1,p) \text{ con } p = \frac{\exp(\eta)}{1+\exp(\eta)} \text{ dove } \eta = a_1x_1 + ... + a_px_p + b.$$

Quando i coefficienti $a_1, ..., a_p, b$ sono divenuti noti, preso un nuovo individuo, calcolati i valori dei suoi predittori $x_1, ..., x_p$, si calcola la probabilità p relativa a quell'individuo. Se p è molto elevata, siamo abbastanza sicuri che per quell'individuo sarà $Y = 1$, mentre se è molto bassa, conteremo che sia $Y = 0$; nel mezzo ovviamente c'è molta indecisione sul valore di Y di quell'individuo, pur valendo comunque che se $p > 1/2$ è più probabile $Y = 1$ e viceversa.

Nella teoria generale dei modelli lineari generalizzati, il numero $\eta = a_1x_1 + ... + a_px_p + b$ viene detto *predittore lineare*, la funzione g^{-1} viene detta *link function* e la funzione g viene detta *mean function*. Nella regressione logistica, la link function è la funzione logistica, rappresentata nella figura precedente.

Resta il problema di trovare i coefficienti. Si devono avere n individui di cui si conoscano i valori dei predittori X_i e di Y. Si usa il metodo della *massima verosimiglianza*. Noti i valori $x_1, ..., x_p$ dei predittori di un individuo, abbiamo detto che Y è $B(1,p)$, con $p = g^{-1}(\eta)$, $\eta = a_1x_1 + ... + a_px_p + b$. Quindi $P(Y = 1) = p$, $P(Y = 0) = 1 - p$. Se indichiamo uno dei due numeri 0 o 1 con y, si può scrivere in una sola formula

$$P(Y = y) = p^y(1-p)^{1-y}.$$

Supponiamo come abbiamo detto che, per un individuo noto, sia noto anche il valore di Y, che chiamiamo con y. Il numero $p^y(1-p)^{1-y}$ è la verosimiglianza relativa a quell'individuo. In astratto, la verosimiglianza è funzione di molte grandezze:

$x_1, ..., x_p,\ y,\ a_1, ..., a_p, b$. Trattandosi di un individuo con $x_1, ..., x_p,\ y$ noti, ben precisi, la verosimiglianza è funzione di $a_1, ..., a_p, b$. Se poi consideriamo gli n individui indipendenti, ed indichiamo con $x_1^{(i)}, ..., x_p^{(i)},\ y^{(i)}$ i loro valori noti, vale

$$P\left(Y^{(1)} = y^{(1)}, ..., Y^{(n)} = y^{(n)}\right) = \prod_{i=1}^{n} \left(p^{(i)}\right)^{y^{(i)}} \left(1 - p^{(i)}\right)^{1-y^{(i)}}$$

dove

$$p^{(i)} = g^{-1}\left(\eta^{(i)}\right), \qquad \eta^{(i)} = a_1 x_1^{(i)} + ... + a_p x_p^{(i)} + b.$$

Questa è la verosimiglianza del campione sperimentale, funzione di $a_1, ..., a_p, b$. Il metodo di massima verosimiglianza consiste nel cercare i valori di $a_1, ..., a_p, b$ che rendono massima la verosimiglianza, cioè $\prod_{i=1}^{n} \left(p^{(i)}\right)^{y^{(i)}} \left(1 - p^{(i)}\right)^{1-y^{(i)}}$. Tecnicamente, conviene massimizzare il logaritmo della verosimiglianza (è equivalente), cioè

$$\sum_{i=1}^{n} \left(y^{(i)} \log p^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - p^{(i)}\right)\right).$$

Il software esegue la massimizzazione con un procedimento iterativo.

3.1.3 Classificazione tramite regressione logistica

Il metodo della regressione logistica serve ad esempio per effettuare una classificazione *non perentoria*. L'output Y può assumere due valori, che per comodità espositiva chiamiamo A e B. Assume A con probabilità p.

A partire da un set di dati, cioè di individui di cui si conoscano sia i predittori sia la classe, A o B, a cui appartengono, si calcolano i coefficienti del modello. Poi, esaminando nuovi individui che vogliamo classificare sulla base della sola conoscenza dei predittori, calcoliamo il numero p di un individuo, ed assegnamolo alla classe che ha probabilità maggiore (A se $p > 1/2$). Eseguita così è una classificazione perentoria, ma è corredata dal numero p stesso, che fornisce un'indicazione del grado di sicurezza che abbiamo, nella classificazione appena eseguita.

E' la stessa logica della predizione tramite modello regressivo. Invece che desiderare una predizione numerica di una grandezza Y associata a certe unità sperimentali, desideriamo sapere se quelle unità appartengono ad una categoria o ad un'altra.

Esempio

Partiamo dalla solita tabella degli indicatori di benessere (per lo scopo che abbiamo, la loro standardizzazione non era necessaria). Assegnamo punteggio 1 alle nazioni del

Nord Italia: Piem, Vaos, Lomb, TrAA, Vene, FrVG, Ligu. EmRo.

	PLIC	SC	SA.SC	TD	TMI	Geo
Piem	0.088	0.471	-0.707	-0.607	-0.395	1
Vaos	-1.545	0.348	-0.642	-0.813	1.578	1
Lomb	0.202	1.397	-0.836	-0.790	-0.538	1
TrAA	0.677	0.435	-1.269	-0.966	-0.075	1
Vene	0.088	1.334	-1.210	-0.848	-0.497	1
FrVG	0.639	-0.005	-1.028	-0.804	-1.301	1
Ligu	1.190	-0.247	0.470	-0.429	-0.354	1
EmRo	0.658	1.177	-1.315	-0.863	-0.347	1
Tosc	0.126	1.092	-0.795	-0.644	-1.355	0
Umbr	-1.431	0.675	-0.140	-0.524	-1.287	0
Marc	0.278	1.090	-0.265	-0.702	-0.0006	0
Lazi	2.329	0.546	-0.080	-0.113	-0.014	0
Abru	0.335	-0.373	0.402	-0.456	0.040	0
Moli	0.658	-1.289	0.065	0.451	-1.151	0
Camp	-1.811	-1.314	2.031	1.664	0.414	0
Pugl	-0.766	-0.926	1.038	0.648	1.109	0
Basi	-0.747	-1.154	0.661	0.844	2.001	0
Cala	-0.500	-1.727	1.571	2.153	0.632	0
Sici	-0.918	-1.130	1.332	1.517	1.783	0
Sard	0.449	-0.403	0.717	1.285	-0.238	0

Copiamo la tabella in R, sotto il nome `IB.`

Costruiamo i vettori con le singole variabili:

```
PLIC<-IB[,1]; SC<-IB[,2]; SA.SC<-IB[,3]; TD<-IB[,4]; TMI<-IB[,5]; Nord<-IB[,6]
```

Eseguiamo la regressione logistica: si usano i Generalized Linear Models con distribuzione in uscita binomiale

```
Nordism<-glm(Nord ~SC+SA.SC+TD,family=binomial)
```

```
predict(Nordism,type = response)
```

	Nordism	Geo
Piem	0.88	1
Vaos	0.99	1
Lomb	0.67	1
TrAA	0.99	1
Vene	0.97	1
FrVG	0.99	1
Ligu	0.24	1
EmRo	0.99	1
Tosc	0.27	0
Umbr	0.040	0
Marc	0.24	0
Lazi	2.7 e-06	0
Abru	0.65	0
Moli	3.3 e-07	0
Camp	2.2 e-16	0
Pugl	1.0 e-11	0
Basi	1.1 e-12	0
Cala	2.2 e-16	0
Sici	2.2 e-16	0
Sard	2.2 e-16	0

La tabella riporta la colonna di uni decisa da noi inizialmente ed a fianco la colonna delle probabilità p di essere "regione di tipo nordico". Si tenga presente che tali probabilità non vanno viste come punteggi in modo analogo a PCA, perché non seguono una scala lineare, ma rappresentano la probabilità di essere classificati Regione del Nord Italia.

Se vogliamo usare il risultato precedente per una classificazione perentoria, mettendoin classe "nordica" le regioni con $p > 1/2$, vediamo che la Liguria non verrebbe classificata come tale (così come tutte le regioni

Esercizio 7 *eseguire la regressione* `Nord~SC + SA.SC + TD` *e fare predict.*

Esercizio 8 *(progetto complesso) Pare che certe agenzie di rating usino (o abbiano utilizzato in passato) i seguenti indicatori, tra altri, per assegnare le nazioni in questa o quella categoria:*

1) PIL (Prodotto Interno Lordo – valore complessivo dei beni e servizi prodotti)

2) debito pubblico (debito dello Stato nei confronti di chi ha sottoscritto obbligazioni – quali, in Italia, BOT e CCT – destinate a coprire il fabbisogno finanziario) sullo stesso

3) deficit del bilancio pubblico.

Più di recente, pare vengano utilizzati anche parametri come la differenza tra attività e passività finanziarie e l'entità dei debiti delle famiglie e delle imprese in relazione al PIL.

Si trovino i dati di alcuni di questi indicatori in rete, relativamente al periodo che si ritiene più opportuno, o usando dati incrementali, o medie o altro secondo la propria intuizione.

Si costruisca una tabella con le nazioni scelte come righe, gli indicatori scelti come prime p colonne, e come colonna (p+1)-esima una colonna di 0 e 1 così pensata.

Si possono eseguire vari esercizi. Si può prendere come ultima colonna una classificazione binaria proposta, relativamente ad un certo anno, da una agenzia di rating. Oppure si può scegliere la classificazione in nazioni che hanno già subito bancarotta rispetto a quelle che non l'hanno subita. Oppure una classificazione ideale in cui ad esempio Grecia e Irlanda hanno 1, in quanto sono le nazioni verso cui l'Europa sta già effettuando operazioni massicce di aiuto economico.

Si utilizzino poi i comandi della regressione logistica per assegnare una probabilità di "fallimento" alle varie nazioni esaminate.

3.1.4 Modelli lineari generalizzati

Sono una generalizzazione del modello base, gaussiano,

$$Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$$

e di quello bernoulliano (detto modello binomiale, in questo contensto) appena visto

$$Y \sim B(1, p) \text{ con } p = \frac{\exp(\eta)}{1 + \exp(\eta)} \text{ dove } \eta = a_1 x_1 + ... + a_p x_p + b.$$

In generale, si ipotizza che l'output Y abbia distribuzione di una certa classe, ad esempio appunto gaussiana, Bernoulli, Poisson, ecc., e si ipotizza che un suo parametro fondamentale θ, di solito la media (μ per la gaussiana, p per la Bernoulli, λ per la Poisson) sia legato ai fattori attraverso una formula del tipo

$$\theta = g^{-1}(\eta)$$

dove

$$\eta = a_1 x_1 + ... + a_p x_p + b$$

è chiamiato *predittore lineare*. Chiamiamo *link function* la funzione g. Nella regressione logistica si prende come g la funzione logit. Nella regressione tradizionale, g è l'identità.

Il comando di R che esegue la regressione per i modelli lineari generalizzati è `glm`.

3.2 Classificazione

3.2.1 Premessa: teoria delle decisioni e regola di Bayes

Per capire il prossimo metodo di classificazione, è utile qualche premessa di teoria delle decisioni.

L'idea base della teoria delle decisioni si può descrivere tramite le nozioni fondamentali del calcolo delle probabilità: l'universo degli eventi, le partizioni, la formula di fattorizzazione e quella di Bayes.

Supponiamo di avere un universo Ω, una partizione (C_k) (ad esempio la suddivisione di Ω in un insieme C_1 ed il suo complementare $C_2 = C_1^c$), e dobbiamo prendere una decisione: quale degli eventi C_k si è verificato (o si verificherà)? Abbiamo usato la lettera C come "classe", immaginando di voler effettuare una classificazione.

Supponiamo di conoscere le (cosidette) *probabilità a priori* dei C_k, i numeri $P(C_k)$. A volte sono note da statistiche precedenti (come nell'esempio 1 che vedremo tra poco), altre volte, più che conoscerle, le si ipotizza. Ad esempio, a volte si suppongono tutte uguali (C_k equiprobabili a priori) per sottolineare il nostro grado di ignoranza iniziale circa quale dei C_k si quello giusto.

Ipotizziamo che gli eventi C_k influiscano su (o comunque siano collegati ad) un evento A che possiamo osservare e che vediamo che si è verificato. Supponiamo di conoscere le probabilità condizionali

$$P(A|C_k)$$

per tutti i k. Tramite il teorema di Bayes, allora, possiamo calcolare le *probabilità a posteriori* dei C_k, i numeri

$$P(C_k|A) = \frac{P(A|C_k)\,P(C_k)}{\sum_i P(A|C_i)\,P(C_i)}.$$

Queste sono le probabilità dei C_k nel momento in cui sappiamo che l'evento A si è verificato.

La *regola decisionale di Bayes* è: scegliere tra i C_k quello con la *massima probabilità a posteriori*. In simboli: $C^{opt} := \arg\max_{C_k} P(C_k|A)$, ovvero

$$C^{opt} := \arg\max_{C_k} P(A|C_k)\,P(C_k)$$

in quanto il denominatore è uguale per tutti i $P(C_i|A)$. Va notato che, se pur in casi plausibilmente rari, potrebbero esistere due diversi C_i che massimizzano questa espressione. In questo caso il metodo non è in grado di prendere una decisione e si può ad esempio dire (anche se in un senso lievemente improprio) che il metodo ha commesso un errore, per cui includeremo questa eventualità negli eventi di errore studiati sotto.

Osservazione 30 *Nel caso equiprobabile, essendo $P(C_k)$ uguale per tutti, il criterio diventa semplicemente*

$$C^{opt} := \arg\max_{C_k} P(A|C_k).$$

Si adotta questo criterio semplificato non solo quando è noto che le probabilità a priori sono uguali, ma anche nelle situazioni di ignoranza, quando le probabilità a priori non sono note.

Esempio 1. Si sa a priori che lo 0.2% della popolazione soffre di una certa malattia dopo i 50 anni. Quella malattia non è ovvia da diagnosticare. Se la malattia è presente, una certa analisi la evidenzia nel 90% dei casi. Se non è presente, l'analisi produce un falso positivo nel 15% dei casi. Un medico esegue l'analisi a un paziente, che risulta positivo. Il medico che decisione prende? (intendiamo: è più propenso a credere che il paziente abbia o non abbia la malattia?). Soluzione: indichiamo con C_1 l'evento: ha la malattia, con A l'evento: risulta positivo all'analisi; conosciamo: $P(C_1) = 0.002$, $P(C_2) = 0.998$, $P(A|C_1) = 0.9$, $P(A|C_2) = 0.15$, quindi calcoliamo

$$P(A|C_1)\,P(C_1) = 0.9 \cdot 0.002 = 0.001\,8$$
$$P(A|C_2)\,P(C_2) = 0.15 \cdot 0.998 = 0.149\,7.$$

la conclusione è che il medico è ancora più propenso a credere che il paziente sia sano. Quell'analisi è poco discriminante. Non si deve però pensare che l'analisi non sia servita a niente. Ora, per la prossima analisi, si parte da una probabilità a priori diversa: il paziente cade in una categoria di persone che ha probabilità $\frac{0.001\,8}{0.001\,8+0.149\,7} = 0.01$ di essere ammalata, $\frac{0.149\,7}{0.001\,8+0.149\,7} = 0.99$ di essere sana (proporzioni ben diverse da quelle iniziali).

Esempio 2. Una rete di trasmissione invia messaggi codificati con 0 e 1. Sulla rete c'è un disturbo, che con probabilità 0.1 modifica 1 in 0 e con probabilità 0.1 modifica 0 in 1. Se riceviamo un 1, cosa decidiamo che sia stato spedito? Soluzione. Per ignoranza, supponiamo che siano equiprobabili l'invio di 0 o di 1. Indichiamo con C_1 l'evento: è stato inviato 1, con A l'evento: abbiamo ricevuto 1; conosciamo: $P(C_1) = P(C_2) = 0.5$, $P(A|C_1) = 0.9$, $P(A|C_2) = 0.1$. Siccome le alternative C_1 e C_2 sono equiprobabili, basta confrontare $P(A|C_1)$ con $P(A|C_2)$ e scegliere il più grande. Quindi ovviamente decidiamo che è stato spedito 1. Questo esempio, così formulato, appare ovvio e poco istruttivo; interessante sarebbe proseguirne l'analisi in un'altra direzione: la probabilità di errore, data da

$$P_{err} = P(A|C_2)\,P(C_2) + P(A^c|C_1)\,P(C_1)$$

è piuttosto alta (vale $P_{err} = 0.1$) e renderebbe troppo incerta la trasmissione di messaggi, quindi bisogna inventare procedimenti per limitare la possibilità di sbagliare. Da qui nascono i codici di correzione d'errore.

3.2.2 Punto di vista geometrico della teoria della classificazione

Riprendiamo il punto di vista descritto nella Sezione 3.1.1. Abbiamo la tabella

	X_1	...	X_p	Classe
1	$x_{1,1}$	...	$x_{1,p}$	C_1
...	...	...	...	...
...	...	...	...	...
n_1	...	...	...	C_1
n_1+1	...	...	...	C_2
...	...	...	...	...
...	...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$	C_2

Ogni individuo del training set è un punto nello spazio $\mathbb{R}^p$, come in PCA, solo che ora esso ha un ulteriore attributo, la classe. Possiamo immaginare gli individui come punti colorati (per distinguere la classe; si vedano i disegni nel seguito). Abbiamo quindi, in $\mathbb{R}^p$, due nuvole di punti, sperabilmente un po' separate (non completamente sovrapposte). Basandoci su queste due nuvole di punti, colorate in modo diverso, vogliamo dividere lo spazio $\mathbb{R}^p$ in due regioni A_1 e A_2 (circa uguale ad A_1^c); per usarle poi per classificare nuovi individui di cui si conoscano solo i valori $x_1, ..., x_p$. Tutti i nuovi individui la cui stringa $(x_1, ..., x_p)$ cadrà in A_1, veranno classificati C_1, gli altri C_2:

$$(x_1, ..., x_p) \in A_1 \longrightarrow \text{ classe } C_1$$
$$(x_1, ..., x_p) \in A_2 \longrightarrow \text{ classe } C_2.$$

Formalizziamo con le notazioni della tabella: abbiamo in $\mathbb{R}^p$ due insiemi di punti, relativi a individui del training set: i punti $P_1, ..., P_{n_1}$ degli individui di classe C_1 e quelli $P_{n_1+1}, ..., P_n$ di quelli di classe C_2.

Come effettuare la suddivisione di $\mathbb{R}^p$ in due regioni? Conviene fare in modo che A_1 contenga tutti i punti $P_1, ..., P_{n_1}$ ed A_2 tutti i punti $P_{n_1+1}, ..., P_n$? Non necessariamente. Una tale strategia ha vari difetti:

- non è univoca (infinite regioni hanno questa proprietà ed è proprio del tutto arbitrario sceglierne una);
- non tiene conto del fatto che le sole variabili $X_1, ..., X_p$ non dovrebbero permettere una classificazione sicura (salvo problemi molto particolari e privi di aleatorietà), quindi deve essere possibile che un individuo di classe A_1 stia nella regione A_2 e viceversa;
- è facile immaginare disposizioni dei punti P_i tali che, per dividerli come detto sopra, siamo costretti a immaginare regioni A molto contorte; se immaginiamo

che dietro il nostro tentativo di classificazione ci sia una realtà "fisica", una *struttura*, un legame reale tra le variabili $X_1, ..., X_p$ e la classe (a meno di errore ed altre variabili non identificate o considerate), è molto strano che questo legame passi attreverso complicate formule matematiche (quelle necessarie a descrivere una regione molto contorta); di solito i legami fisici tra grandezze hanno natura polinomiale o comunque abbastanza semplice.

Quindi si rinuncia al requisito che A_1 contenga tutti i punti $P_1, ..., P_{n_1}$ e A_2 tutti gli altri. Si vuole che ciò avvenga per la maggior parte dei punti, salvaguardando contemporaneamente qualche criterio di *struttura* e *semplicità geometrica* delle due regioni. Una scelta molto comune, che vedremo, è che le due regioni siano dei semispazi, cioè la divisione in due di $\mathbb{R}^p$ sia realizzata da un iperpiano.

Il discorso astratto si estende al caso di più classi, senza modifiche particolarmente rilevanti, se non notazionali. Le suddivisioni però saranno più complicate.

3.2.3 Esempio: suddivisione tramite regressione lineare multipla

Un approccio meno noto di altri ma concettualmente molto naturale è il seguente. Si riscrive la tabella del training set nella forma

	X_1	...	X_p	Y
1	$x_{1,1}$	...	$x_{1,p}$	-1
...	...	...	...	...
...	...	...	...	...
n_1	...	...	...	-1
n_1+1	...	...	...	1
...	...	...	...	...
...	...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$	1

e si applica la regressione lineare classica, cioè si cerca di spiegare il valore ± 1 (che è semplicemente una codifica, peraltro arbitraria, della classe) tramite i predittori $X_1, ..., X_p$. Supponiamo di aver trovato i coefficienti $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$ del modello regressivo. Se ora esaminiamo un nuovo individuo, di cui si conoscano i valori $x_1, ..., x_p$, possiamo calcolare il suo valore y:

$$y = \widehat{a}_1 x_1 + ... + \widehat{a}_p x_p + \widehat{b}.$$

Ovviamente questo valore non sarà uguale a ± 1, però adottiamo la regola: se $y < 0$ allora l'individuo è di classe C_1; se invece $y > 0$ allora è di classe C_2.

Questo metodo è quasi uguale alla regressione logistica: anche in essa si calcola, per un nuovo individuo, il predittore lineare

$$\eta = \widehat{a}_1 x_1 + ... + \widehat{a}_p x_p + \widehat{b}$$

con cui poi si calcola $p = g(\eta)$ e si classifica l'individuo di classe C_1 se $p < \frac{1}{2}$; ma questo corrisponde esattamente a dire che lo si calssifica C_1 se $\eta < 0$ (se g è la funzione logit). Quindi gli algoritmi sembrano identici. L'unica differenza è nel metodo di calcolo dei coefficienti $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$, eseguito massimizzando la verosimiglianza (nella regressione logistica) invece che col metodo dei minimi quadrati.

Come si interpreta questo metodo in termini di regioni A_1 ed A_2? Semplicemente,

$$\begin{aligned} A_1 &= \left\{ x \in \mathbb{R}^p : \widehat{a}_1 x_1 + ... + \widehat{a}_p x_p + \widehat{b} < 0 \right\} \\ A_2 &= \left\{ x \in \mathbb{R}^p : \widehat{a}_1 x_1 + ... + \widehat{a}_p x_p + \widehat{b} > 0 \right\}. \end{aligned}$$

Sono quindi dei semispazi e la separazione è fatta tramite l'iperpiano $\widehat{a}_1 x_1 + ... + \widehat{a}_p x_p + \widehat{b} = 0$. Lo stesso vale per la regressione logistica.

3.2.4 Unione delle idee bayesiana e geometrica: Discriminant Analysis

Descriviamo ora la teoria dell'Analisi Discriminante che unisce i punti di vista bayesiano e geometrico. Per determinare le regioni A_1 ed A_2 dell'approccio geometrico essa utilizza densità di probabilità gaussiane, seguendo l'impostazione bayesiana.

Ricordiamo che, dietro i valori sperimentali del training set, immaginiamo che ci siano le v.a. $X_1, ..., X_p$. Supponiamo che, per gli individui della classe C_1, il vettore aleatorio $X = (X_1, ..., X_p)$ sia gaussiano, di densità congiunta $f_X(x|C_1)$; analogamente, supponiamo che per gli individui della classe C_2, X sia gaussiano di densità congiunta $f_X(x|C_2)$. Dai dati sperimentali del training set possiamo calcolare media e matrice di covarianza empirica di $f_X(x|C_1)$ e di $f_X(x|C_2)$, che indichiamo con μ_i e Q_i, $i = 1, 2$.

Per Bayes (immaginiamo di usare un analogo della formula di Bayes nel caso di densità), vale

$$P(C_1|x) = \frac{f_X(x|C_1) P(C_1)}{f_X(x|C_1) P(C_1) + f_X(x|C_2) P(C_2)}$$

ed analogamente per $P(C_2|x)$. Ecco la regola di classificazione: se abbiamo un nuovo individuo di cui abbiamo misurato i valori $x = (x_1, ..., x_p)$, calcoliamo $f_X(x|C_1) P(C_1)$ e $f_X(x|C_2) P(C_2)$ e lo classifichiamo C_1 o C_2 a seconda della classe col valore maggiore.

Restringendoci per semplicità al caso $P(C_1) = P(C_2)$, la regola è: se $f_X(x|C_1) > f_X(x|C_2)$ mettiamo l'individuo x nella classe C_1; se $f_X(x|C_1) < f_X(x|C_2)$ lo mettiamo

nella classe C_2. Posto

$$A_1 = \{x \in \mathbb{R}^p : f_X(x|C_1) > f_X(x|C_2)\}$$
$$A_2 = \{x \in \mathbb{R}^p : f_X(x|C_1) < f_X(x|C_2)\}$$

se $x \in A_1$ classifichiamo C_1, mentre se $x \in A_2$ classifichiamo C_2. Il software, data la tabella iniziale di dati del training set, calcola le densità di probabilità empiriche $f_X(x|C_1), f_X(x|C_2)$ e, per ogni nuovo individuo, verifica se si trovi in A_1 oppure A_2.

Cerchiamo ora di capire l'aspetto geometrico di questa suddivisione.

3.2.5 Linear e Quadratic Discriminant Analysis

Ricordando che

$$f_X(x|C_i) = \frac{1}{\sqrt{(2\pi)^n \det Q_i}} \exp\left(-\frac{1}{2}(x-\mu_i)^T Q_i^{-1}(x-\mu_i)\right)$$

la disuguaglianza $f_X(x|C_1)P(C_1) > f_X(x|C_2)P(C_2)$ diventa, passando ai logaritmi,

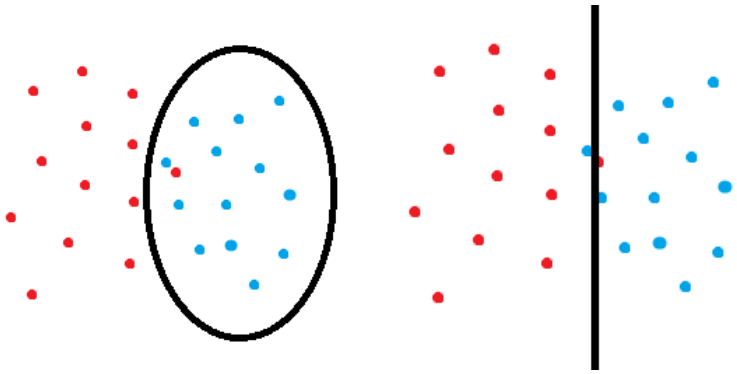
$$(x-\mu_2)^T Q_2^{-1}(x-\mu_2) - (x-\mu_1)^T Q_1^{-1}(x-\mu_1) > \log(\det Q_1) - \log(\det Q_2).$$

Si trova una condizione della forma

$$x^T\left(Q_2^{-1} - Q_1^{-1}\right)x + ...$$

con termini lineari e costanti, quindi, a seconda delle proprietà della matrice $Q_2^{-1} - Q_1^{-1}$, si trovano regioni curvilinee di vario tipo. Questa è la *Quadratic Discriminant Analysis*, su cui non entriamo in dettaglio. Preferiamo approfondire il caso successivo per due ragioni: la maggior complessità del caso quadratico ed un suo potenziale difetto di tipo geometrico.

Osservazione 31 *Si immagini la situazione descritta dal primo dei due disegni riportati sotto: in essa la quadratic discriminant analysis, per adattarsi al meglio alla forma ellissoidale del cluster azzurro, crea una suddivisione (la curva nera) che gira attorno al cluster. Ne discenderebbe che un nuovo individuo che si trovi alla destra dei punti azzurri, o sopra o sotto di essi, un po' staccato, verrebbe classificato rosso. Questa classificazione corrisponde alla nostra intuizione? Magari in alcuni problemi sì, ma in altri riterremmo più naturale una suddivisione lineare del tipo indicato dalla seconda figura.*



Ipotizziamo ora invece che le due gaussiane abbiano la stessa covarianza (corrisponde a dire che il legame di correlazione tra i predittori sia lo stesso per le due classi). Determiniamo la covarianza empirica Q usando tutti i dati del training set, indifferenziati. Invece non supponiamo che le medie coincidano, anzi è proprio la differenza tra i centri dei due cluster che permette di effettuare una classificazione (se i centri coincidessero le due nuvole sarebbero sovrapposte e non potremmo trovare una buona suddivisione). Riprendendo i calcoli precedenti ma nell'ipotesi che sia $Q_1 = Q_2 = Q$, i termini quadratici si semplificano e troviamo

$$2x^T Q^{-1} (\mu_1 - \mu_2) > \mu_1^T Q^{-1} \mu_1 - \mu_2^T Q^{-1} \mu_2.$$

E' l'equazione di un semispazio. In altre parole, $\mathbb{R}^p$ viene suddiviso dall'iperpiano

$$x \cdot v = \alpha$$
$$v = Q^{-1} (\mu_1 - \mu_2), \qquad \alpha = \frac{1}{2} \left(\mu_1^T Q^{-1} \mu_1 - \mu_2^T Q^{-1} \mu_2 \right).$$

Quindi si realizza ciò che è raffigurato dal secondo dei disegni precedenti, analogo inoltre a ciò che abbiamo visto nei paragrafi precedenti con la regressione lineare oppure logistica.

La procedura è:

1. prima si usano i punti $P_1, ..., P_{n_1}$ per stimare il punto medio μ_1 e si usano i punti $P_{n_1+1}, ..., P_n$ per stimare μ_2;

2. poi si centrano i punti, cioè si calcolano i punti $P'_i = P_i - \mu_1$ per $i = 1, ..., n_1$, $P'_i = P_i - \mu_2$ per $i = n_1 + 1, ..., n$;

3. infine si calcola la matrice di covarianza empirica usando tutti i punti P'_i.

L'immagine dell'iperpiano serve solo idealmente per capire il risultato. Infatti, per eseguire la classificazione di un nuovo individuo, rappresentato da una nuova stringa $x = (x_1, ..., x_p)$, basta:

1. calcolare $v = Q^{-1}(\mu_1 - \mu_2)$ e $\alpha = \frac{1}{2}\left(\mu_1^T Q^{-1}\mu_1 - \mu_2^T Q^{-1}\mu_2\right)$ (usando gli oggetti stimati)

2. assegnare il nuovo individuo alla classe C_1 se $x \cdot v > \alpha$, mentre alla classe C_2 se $x \cdot v < \alpha$.

Naturalmente la prima delle due operazioni si può svolgere una volta per tutte, essendo la stessa per tutti i nuovi individui. Osserviamo che la situazione di mancata classificazione, cioè il caso $x \cdot v = \alpha$, in pratica non può mai avvenire.

Quella appena descritta è la *Linear Discriminant Analysis*. Per utilizzarla col software `R` bisogna prima caricare il package MASS, col comando `require(MASS)`, poi usare il comando `lda` (si consiglia di leggerlo con `?lda`).

3.3 Clustering

Le tecniche di classificazione appena descritte partono dall'esistenza di classi prestabilite e si pongono il problema di assegnare nuovi individui alle classi (*classificare* nuovi individui).

Le tecniche di clustering invece si pongono l'obiettivo di trovare delle buone suddivisioni di un gruppo di individui, avendo a priori solo il numeri di suddivisioni desiderate, ma nulla riguardo alla loro natura. In questo senso, somiglia un po' all'idea di FA, in cui i fattori non sono noti a priori, solo il loro numero.

Si pensi ad un insieme W di punti Q del piano ($Q \in W$), sparpagliati, ciascuno rappresentante un individuo (descritto quindi da due variabili, due predittori). Ci saranno casi in cui i punti sono un po' separati in due gruppi, o più di due gruppi, pur essendo vaga la separazione. Si pensi alle case di due città limitrofe in zone molto abitate: si va da una città all'altra quasi senza soluzione di continuità, però il grado di addensamento è diverso nelle due zone proprie delle città rispetto alla parte intermedia, dove c'è ancora un po' di campagna qua e là. Abbiamo quindi questo insieme di punti. Ipotizziamo che esso sia suddividibile in due classi (il caso con tre o più classi è simile, ma torneremo su questo punto). Vediamo alcune idde generali per trovare una buona suddivisione.

Alcune idee dei paragrafi precedenti sarebbero perfettamente adatte: cercare una retta, o una parabola (linear o quadratic discriminant analysis) che separa bene l'insieme dei punti. Sviluppiamo altre idee.

Immaginiamo che le due classi siano come due nuvole un po' ellittiche, pur con vaghezza (magari senza una vera soluzione di continuità tra le nuvole). Iniziamo col cercare i *centri* delle nuvole. Avendo deciso che sono due, si cercano due centri, M_1 e M_2 (qui entra in gioco il numero di classi deciso a priori: se avessimo deciso di dividere in tre classi, avremmo cercato tre centri). Si inizi mettendo a caso due punti M_1 e M_2 nel piano, in assenza di suggerimenti migliori (se invece c'è un'idea migliore la si

usi). Poi si trovino gli *insiemi di Voronoi* di questi due punti, che chiamiamo V_1 e V_2: V_i è l'insieme dei punti del piano che distano da M_i meno che dall'altro centro. Sono due semipiani. Se partivamo da tre centri M_1, M_2, M_3 trovavamo una divisione in tre "angoli", e così via. Poi, chiamiamo W_1 e W_2 gli insiemi dei punti originari che cadono in V_1 e V_2 rispettivamente: W_i è l'insieme dei punti $Q \in W$ che appartengono a V_i, quindi che distano da M_i meno che dall'altro centro. Questa è già una suddivisione possibile, però relativa ad una scelta iniziale dei centri, fatta a caso o comunque non ancora ottimizzata in alcun modo.

Diamo un punteggio alla suddivisione trovata: calcoliamo la somma delle distanze al quadrato di tutti i punti di W_1 da M_1

$$d_1^2 = \sum_{Q \in W_1} d^2 (Q, M_1)$$

ed analogamente per W_2: $d_2^2 = \sum_{Q \in W_2} d^2 (Q, M_2)$. Questa suddivisione è caratterizzata dal numero $d_1^2 + d_2^2$; se tale numero è alto, la suddivisione viene considerata poco buona (i punti di ciascun gruppo distano troppo dal loro centro). In generale, per k gruppi, il numero da calcolare è

$$\sum_{i=1}^{k} \sum_{Q \in W_i} d^2 (Q, M_i) .$$

Si vorrebbero trovare i punti M_i che rendono minima questa espressione. Si possono inventare vari algoritmi che cercano di trovare dei buoni centri M_i. L'algoritmo k-means lavora su centri M_i che vengono presi, ad ogni passo dell'algoritmo iterativo, pari alla media aritmetica dei punti di W_i (poi vengono ricalcolati i W_i, poi i loro punti medi M_i e così via). L'algoritmo k-medoids utilizza invece come centri alcuni dei punti di W stesso, aggiornando iterativamente i medoidi (alla ricerca dei migliori) attravero scambi causali tra i medoidi e gli altri punti di W. Gli algoritmi differiscono poi, tra altre cose, per la distanza $d (Q, M_i)$ che viene utilizzata (rimandiamo alla letteratura specializzata per questi ed altri dettagli).

Questi algoritmi hanno un difetto: raggruppano secondo la minima distanza dai centri, quindi tendono a costruire dei raggruppamenti equilibrati, della stessa grandezza. Questa simmetria può essere poco adatta a certe applicazioni, in cui si capisce ad occhio che i punti $Q \in W$ sono divisi in gruppi di ampiezza differente, per esempio una grossa nuvola con una piccola nuvola satellite. Gli algoritmi descritti fino ad ora forzerebbero la suddivisione ad essere abbastanza simmetrica, attribuendo una parte di punti della grossa nuvola alla parte W_i relativa al piccolo satellite. C'è allora una variante, detta algoritmo EM (Expectation-Maximization) basata sulle misture di gaussiane e la massima verosimiglianza, che permette di trovare partizioni diseguali, più aderenti a certe dituazioni pratiche.

In genere il software, come input di un particolare metodo di clustering (k-means ecc.), chiede i punti $Q \in W$ (una tabella di dati come quella di PCA) ed il numero

di classi k in cui vogliamo suddividerli. Come output fornisce le classi trovate, in genere elencando gli elementi delle classi, e fornendo una raffigurazione grafica dei punti separati in gruppi, raffigurazione spesso legata a PCA. Infatti, se i punti $Q \in W$ stanno in uno spazio a dimensione maggiore di 2, il modo più naturale è innanzi tutto mostrare questi punti attraverso una visione che li distingua il più possibile (e questo è svolto da PCA), sovrapponendo poi ad essa la suddivisione in gruppi. Esistono anche visualizzazioni tridimensionali a colori.

Oltre a questo, il software fornisce in output dei parametri numerici che servono a giudicare la suddivisione ottenuta, il più comune dei quali è la *silhoutte*. Tramite questi numeri abbiamo una quantificazione della bontà o vaghezza dei cluster ottenuti che, oltre ad essere un metro di giudizio di tipo assoluto, può essere utilizzato in modo comparativo per decidere il numero k. Esso era stato scelto a priori, ma con quale criterio? Ci saranno casi in cui, o per ragioni di evidenza grafica o per motivi applicativi, sapremo come decidere k a priori; altri in cui si va per tentativi e si sceglie k a posteriori: quello che massimizza la silhouette.

Descriviamo la silhouette secondo una delle sue possibili definizioni. La silhouette di un singolo individuo $Q \in W$, relativa alla partizione $W_1, ..., W_k$ trovata con un qualsiasi metodo tipo k-means ecc., è data dall'espressione

$$s(Q) = \frac{b(Q) - a(Q)}{\max(a(Q), b(Q))}.$$

Indicando con $W(Q)$ il cluster, tra i vari $W_1, ..., W_k$, che contiene il punto Q, il numero $a(Q)$ è la distanza media quadratica di Q dagli altri punti del proprio cluster $W(Q)$:

$$a(Q) = \sum_{Q' \in W(Q)} d(Q, Q')^2.$$

Il numero $b(Q)$ invece è la distanza media quadratica di Q dai punti del cluster "successivo", così definito: si calcolano i numeri

$$\sum_{Q' \in W_i} d(Q, Q')^2$$

per ogni $W_i \neq W(Q)$ e si prende il minimo; questo è $b(Q)$. Si verifica che il numero $s(Q)$ soddisfa

$$-1 \leq s(Q) \leq 1.$$

Più $s(Q)$ è vicino a 1, più si ritiene che la clusterizzazione di Q sia buona. Infatti, supponiamo che $s(Q)$ sia vicino a 1. Innanzi tutto questo implica che $b(Q) - a(Q)$ è positivo, quindi $\max(a(Q), b(Q)) = b(Q)$ e vale

$$s(Q) = \frac{b(Q) - a(Q)}{b(Q)} = 1 - \frac{a(Q)}{b(Q)}.$$

Ora, se questo rapporto vale quasi 1, significa che $a(Q)$ è molto piccolo rispetto a $b(Q)$, cioè che la distanza media di Q dai suoi compagni di gruppo è decisamente minore di quella dai membri del gruppo "successivo". Questo è sintomo di buona clusterizzazione di Q.

La silhouette di un singolo individuo Q serve a giudicare quali individui sono stati raggruppati bene e quali no. Poi, mediando sugli individui di un gruppo W_i si ottiene la silhouette media di W_i, che descrive quanto preciso o vago sia il gruppo W_i. Infine, mediando sui gruppi si ottiene una silhouette media complessiva della clusterizzazione $W_1, ..., W_k$, che può essere utilizzata per confrontare vari k tra loro (oltre che vari metodi anche di natura diversa).

Si suggerisce, col software R, l'uso del comando `pam` (partition around medoids), che svolge la cluster analysis con metodo dei medoidi.

3.4 Domande

1. Descrivere la struttura dei dati di un training set
2. Cosa si determina a partire da un training set?
3. In termini geometrici, quali metodi producono suddivisioni lineari?
4. Spiegare perché i metodi regressivi producono suddivisioni lineari
5. Determinare le suddivisioni prodotte dall'analisi discriminante
6. Spiegare l'origine bayesiana e geometrica dell'analisi discriminante
7. definire il modello della regressione logistica e spiegarne i dettagli
8. come vengono determinati i coefficienti nella regressione logistica?
9. A cosa servono le silhouette.
10. Che differenza c'è tra metodi di classificazione e metodi di clustering.

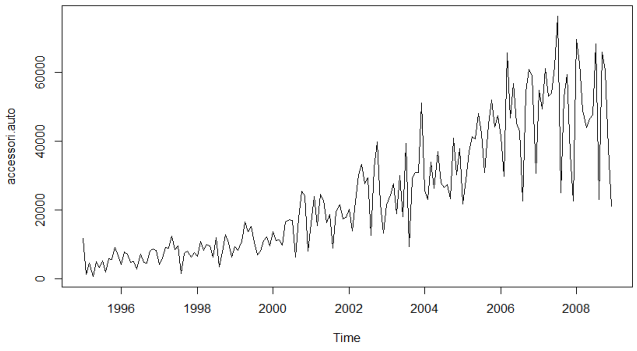
Capitolo 4

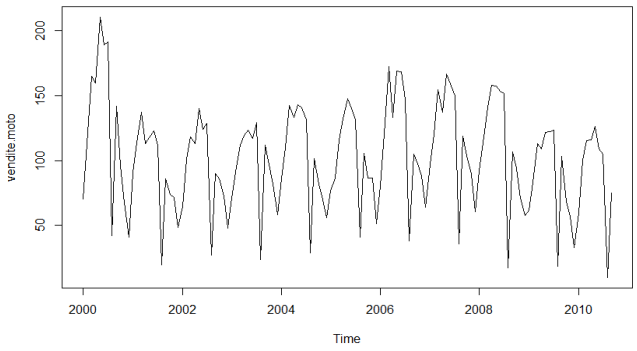
Serie storiche e processi stocastici

4.1 Introduzione alle serie storiche

4.1.1 Struttura, trend e stagionalità

Una serie storica è una sequenza finita di numeri $x_1, ..., x_n$ in cui, a livello interpretativo, l'indice corrisponde al tempo. Il tempo, pari ad $1, 2, ..., n$ può ad esempio essere misurato in mesi. Ecco due esempi grafici di serie storiche:





Esse rappresentano le vendite di accessori per auto e le vendite di moto relative al periodo indicato, da parte di certe aziende.

Lo studio delle serie storiche si può dividere in *analisi* e *previsione*. L'analisi a volte è utile di per sé, per capire un fenomeno già avvenuto, ma più spesso è funzionale alla previsione. L'interesse per la previsione non ha invece bisogno di giustificazione.

Nel tentativo di eseguire una previsione dei valori futuri di una serie storica, in genere si usano solamente i valori noti (passato e presente) della serie stessa; solo con metodi più complessi e di rado si cerca di usare anche qualche informazione esterna. Volendo usare solo i dati noti della serie stessa, ci si convince che l'unica regola è ricopiare il passato. Però non in senso banale. Si vorrebbe capire la *struttura* della serie nel passato, al netto delle variazioni accidentali; in modo da ricopiare nel futuro solamente la struttura e non le variazioni (che non si ripeterebbero identiche). Il problema è quindi distinguere la struttura dalle fluttuazioni casuali.

La struttura non è però un concetto matematicamente ben definito, è solo un'idea. Due elementi strutturali sono il *trend* e la *periodicità* (o *stagionalità*). Anche questi due concetti sfuggono ad una definizione matematica univoca e precisa ma l'intuito umano li cattura ugualmente, grazie alle capacità di confronto e sintesi che abbiamo.

La serie delle vendite di accessori auto ha un trend crescente piuttosto evidente, meno intenso all'inizio e più intenso negli anni 2001-2007 (poi subentra la crisi economica). La serie delle vendite di moto non ha un trend marcato, tutt'al più delle lievi variazioni di trend (la crisi del 2007 è visible anche lì, con un lieve trend negativo).

La serie delle vendite di accessori auto non mostra alcuna periodicità chiaramente visibile, anche se l'occhio umano qui potrebbe confondersi a causa delle fluttuazioni. Invece la serie delle vendite di moto ha un'evidente stagionalità, una periodicità annuale. Si noti che in questo ambito, per periodicità, non intendiamo che valga esattamente la proprietà $f(x+T) = f(x)$ dell'analisi matematica; varrà qualcosa di simile ma perturbato dalle fluttuazioni e magari da un più o meno lieve trend; questo rende impossibile una definizione precisa. Ma l'intuito è chiaro.

4.1.2 Funzione di autocorrelazione empirica

Data una serie storica $x_1, ..., x_n$, preso un intero non negativo k minore di n che chiameremo ritardo, immaginiamo graficamente di affiancare la serie originaria alla sua traslata di k; il seguente disegno mostra il caso $k = 2$:

x_1	x_2	x_3	...	...	...	...	x_{n-2}	x_{n-1}	x_n		
		x_1	x_2	x_3	...	...	...	...	x_{n-2}	x_{n-1}	x_n

Ora si prenda la parte comune:

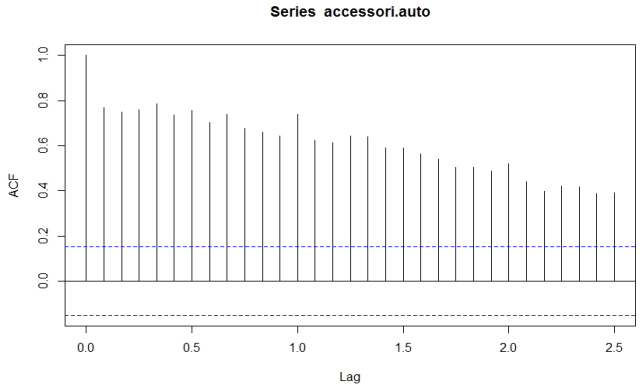
x_3	...	...	...	...	x_{n-2}	x_{n-1}	x_n
x_1	x_2	x_3	...	...	...	...	x_{n-2}

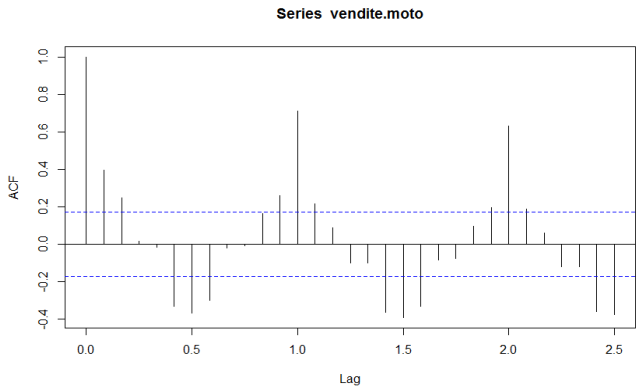
Queste sono due stringhe di $n-2$ numeri ciascuna. Se ne calcoli la correlazione $\widehat{\rho}$. Ricoridamo la formula per la correlazione, scrivendola in generale con k qualsiasi:

$$\widehat{\rho}(k) = \frac{\sum_{i=1}^{n-k}(x_i - \overline{x}_0)(x_{i+k} - \overline{x}_k)}{\sqrt{\sum_{i=1}^{n-k}(x_i - \overline{x}_0)^2 \sum_{i=1}^{n-k}(x_{i+k} - \overline{x}_k)^2}} \tag{4.1}$$

dove $\overline{x}_0 = \frac{1}{n-k}\sum_{i=1}^{n-k} x_i$ $\overline{x}_k = \frac{1}{n-k}\sum_{i=k+1}^{n} x_i$.

Ripetendo il procedimento per $k = 0, 1, ..., k_0$, dove k_0 è un intero positivo minore di n, troviamo una nuova serie storica $\widehat{\rho}(k)$, detta *funzione di autocorrelazione empirica*, che possiamo raffigurare. Ad esempio, le figure seguenti mostrano il calcolo svolto dal software R relativamente agli esempi di serie storiche del paragrafo precedente:





Se il valore di k_0 è troppo vicino ad n, le parti comuni delle due stringhe sono troppo brevi ed il risultato perde di validità statistica; per questo bisogna stare un po' bassi con k_0.

Osservazione 32 *Vale sempre* $\widehat{\rho}(0) = 1$.

Osservazione 33 *Per essere precisi, il software R ha scelto la politica di rappresentare, tramite il comando acf usato per i due disegni precedenti, una lieve variante*

della funzione $\widehat{\rho}(k)$*, data dalla seguente formula:*

$$\widehat{\rho}_k^R = \frac{\widehat{cov}^R(0,k)}{\widehat{cov}^R(0,0)}$$

dove

$$\widehat{cov}^R(0,k) = \frac{1}{n}\sum_{i=1}^{n-k}(x_i - \overline{x}_0)(x_{i+k} - \overline{x}_k).$$

Vale

$$\widehat{\rho}_k^R = \frac{\sum_{i=1}^{n-k}(x_i - \overline{x}_0)(x_{i+k} - \overline{x}_k)}{\sum_{i=1}^{n}(x_i - \overline{x}_0)(x_{i+k} - \overline{x}_k)}$$

quindi la somma a denominatore ha n *addendi invece che* $n-k$ *come il numeratore. Pertanto, al crescere di* k*, il rapporto avrà una certa tendenza a diminuire. Questo varrà per motivi puramente algebrici, indipendenti dalle proprietà statistiche della serie storica. Invece la funzione* $\widehat{\rho}(k)$ *definita da (4.1) non ha questa tendenza alla diminuzione; o meglio, su dati sperimentali veri tende a diminuire ma non per motivi algebrici, bensì per la naturale perdita di correlazione (memoria) man mano che il tempo avanza.*

Per com'è definita, la funzione di autocorrelazione cattura le somiglianze interne alla serie storica, vede cioè se la serie traslata opportunamente somiglia a se stessa; e tali somiglianze corrispondono ad aspetti strutturali.

Più precisamente, se una serie ha un'accentuata periodicità ma non ha un accentuato trend, allora $\widehat{\rho}(k)$ avrà un valore elevato in corrispondenza del periodo (e, anche se un po' meno, dei suoi multipli). Se invece c'è un trend accentuato, $\widehat{\rho}(k)$ è abbastanza elevato dappertutto (calando lentamente al crescere di k, per le ragioni spiegate nell'osservazione 33). Se oltre al trend accentuato c'è anche una periodicità, il valore di $\widehat{\rho}(k_{per})$ relativo al periodo k_{per} si staglierà un pochino al di sopra di tutti gli altri $\widehat{\rho}(k)$, però ne sarà mascherato, in maniera maggiore o minore. Nel primo esempio visto sopra, $\widehat{\rho}(12)$ è un po' più alto, segno che c'è una lieve periodicità, ma il trend la maschera quasi del tutto. Perciò la funzione $\widehat{\rho}(k)$ è davvero utile solo quando il trend è lieve, altrimenti è di dubbia interpretazione.

Vediamo di capire algebricamente come mai un trend, positivo o negativo, dovrebbe generare valori grandi di $\widehat{\rho}(k)$. Supponiamo che la serie storica $x_1, ..., x_n$ abbia la forma

$$x_i = a \cdot i + \varepsilon_i$$

cioè una retta di coefficiente angolare a ed intercetta nulla (per semplicità di esposizione), perturbata dagli errori ε_i che supponiamo molto piccoli. Trasliamo di k, ad esempio $k = 2$, e vediamo la parte comune:

$3a + \varepsilon_3$	$4a + \varepsilon_4$	$5a + \varepsilon_5$	...	...
$a + \varepsilon_1$	$2a + \varepsilon_2$	$3a + \varepsilon_3$	...	...

che riscriviamo

$a+2a+\varepsilon_3$	$2a+2a+\varepsilon_4$	$3a+2a+\varepsilon_5$	...	...
$a+\varepsilon_1$	$2a+\varepsilon_2$	$3a+\varepsilon_3$	...	...

La prima stringa è circa pari alla seconda più $2a$ e quindi, in base alla definizione di correlazione empirica, $\widehat{\rho}(2)$ è elevato. Il ragionamento vale per ogni k. Il segno di a non modifica il ragionamento (in particolare non si deve pensare che un trend negativo renda negativi i numeri $\widehat{\rho}(k)$).

Analogamente, vediamo che una serie storica molto prossima ad essere periodica di periodo k_{per} avrà un picco di $\widehat{\rho}(k)$ per $k = k_{per}$ mentre gli altri valori di $\widehat{\rho}(k)$ saranno generici, non necessariamente grandi. Supponiamo che la serie storica $x_1, ..., x_n$ abbia la forma

$$x_i = f(i) + \varepsilon_i$$

dove f è una funzione di periodo k_{per}, cioè soddisfa $f(t + k_{per}) = f(t)$ per ogni t; e di nuovo supponiamo che gli errori ε_i siano molto piccoli. Trasliamo di $k = 2$ e vediamo la parte comune:

$f(3)+\varepsilon_3$	$f(4)+\varepsilon_4$	$f(5)+\varepsilon_5$	...	...
$f(1)+\varepsilon_1$	$f(2)+\varepsilon_2$	$f(3)+\varepsilon_3$	...	...

Se $k_{per} = 2$, vale $f(3) = f(1)$, $f(4) = f(2)$ e così via, quindi le due stringhe sono quasi uguali e la correlazione è elevata. Se invece $k_{per} \neq 2$, non c'è a priori alcun legame tra $f(3)$ ed $f(1)$, tra $f(4)$ ed $f(2)$, e così via, quindi la correlazione non sarà, in genere, particolarmente elevata. Spiccherà (in genere) il valore $\widehat{\rho}(k_{per})$; inoltre spiccheranno, per le stesse ragioni, i valori $\widehat{\rho}(2k_{per})$, $\widehat{\rho}(3k_{per})$ ecc., mentre gli altri resteranno sempre causalmente alti o bassi.

4.1.3 Metodi di decomposizione

Il problema che affrontiamo qui è quello di trovare trend e stagionalità di una serie storica $x_1, ..., x_n$. Ovvero di scriverla nella forma

$$x_i = t_i + s_i + \varepsilon_i$$

dove la serie $t_1, ..., t_n$ è il *trend*, la serie $s_1, ..., s_n$ è la *stagionalità*, la serie $\varepsilon_1, ..., \varepsilon_n$ è quella dei *residui*.

La precedente è la cosidetta forma *additiva* di decomposizione. In alcuni casi, specialmente in ambito economico, è più naturale una decomposizione *moltiplicativa*:

$$x_i = t_i (s_i + \varepsilon_i).$$

In questo caso le fluttuazioni, rappresentate da $s_i + \varepsilon_i$ (più precisamente dal loro effetto su x_i) aumentano quando aumenta il trend. I balzelli o strutturali (es. estate-inverno) o casuali hanno ampiezza maggiore se si applicano a volumi maggiori di attività economica.

L'idea forse più naturale è quella di:

- mediare la serie, tramite medie locali, in modo da identificare il trend
- sottrarlo, riconducendosi ad una serie senza trend
- mediare la nuova serie sul periodo (nel senso che spiegheremo nel paragrafo 4.1.3), trovando così la componente stagionale
- sottrarla, trovando così i residui.

Medie locali e media mobile

Per media locale di una serie intendiamo un'operazione del tipo

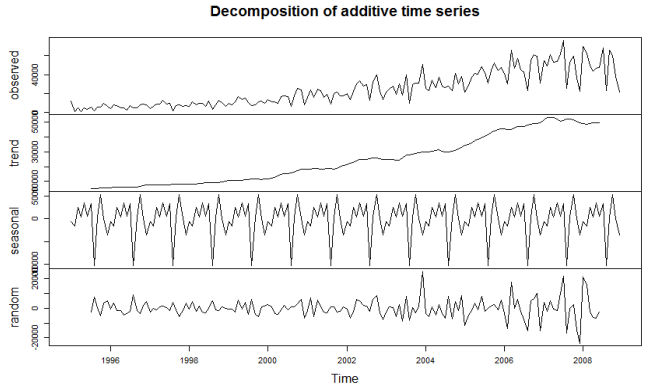
$$\frac{x_i + ... + x_{i+k}}{k}$$

dove k è (anche molto) minore di n. Si può eseguire una simile operazione con due scopi, uno di ricerca del trend ed uno di previsione. La ricerca del trend significa ricerca di una componente della serie nota, mentre la previsione riguarda i valori futuri. Le formule sono quasi uguali ma con piccole differenze sugli indici.

Se si vuole trovare, in corrispondenza di un valore $i = 1, ..., n$ (la parte nota della serie), un valore medio della serie, che al variare di i ne rappresenti il trend, la cosa più naturale è fare una media simmetrica attorno ad i, del tipo

$$t_i = \frac{x_{i-k_1} + ... + x_{i+k_1}}{2k_1 + 1}. \tag{4.2}$$

Questa è l'operazione effettuata dal comando `decompose` di R. A causa di tale simmetria, il risultato di `decompose` è un trend definito solo a partire da un certo tempo (maggiore di 1) e solo fino ad un certo istante (minore di n). Si veda un tipico risultato grafico:



Il trend non è lungo quanto la serie originaria.

Se invece si vuole usare la media locale dei valori per eseguire una previsione, si fa una media unilaterale, basata solo sui valori passati. Questo è il *metodo di media mobile*, il più semplice fra tutti i metodi di previsione di serie storiche (tuttavia molto usato, soprattutto in situazioni di enorme incertezza e poca struttura, come nelle situazioni finanziarie). Fissato un intero positivo k, si pone

$$p_{i+1} = \frac{x_{i-k+1} + ... + x_i}{k}.$$

Ha senso eseguire questo calcolo per $i = k, k+1, ...$ ecc.; per i primi di questi valori, l'indice $i+1$ appartiene ancora al passato, alla parte nota della serie, quindi il metodo sta eseguendo una previsione come se ci mettessimo nel passato (es. cosa avrebbe previsto il metodo circa il valore di ottobre 2009 sapendo i valori fino a settembre 2009). Solo per $i = n$ il metodo esegue una vera previsione sul futuro. Eseguire però le previsioni anche sul passato è un'operazione utilissima per vedere la bontà del metodo, quanto esso avrebbe sbagliato nel passato, quindi viene sempre eseguita.

La scelta di k_1 (per la media simmetrica) o k (per il metodo di media mobile) è arbitraria, soggettiva. Un valore basso di k_1 produrrà un trend abbastanza variabile ed oscillante, perché esso risentirà parecchio delle fluttuazioni e variazioni della serie storica (anche della stagionalità, ad esempio). Un valore alto di k_1 media di più, produce un trend più regolare, poco variabile, ma i valori di i per cui si può calcolare t_i (con la formula (4.2)) sono pochi. A noi la scelta di quanto stare in mezzo. Possiamo anche vedere visivamente il risultato ed operare la scelta a posteriori. Si può anche immaginare di imbastire un procedimento di minimo quadrati e scelgliere il k che li minimizza; non sviluppiamo questa idea, simile a quanto diremo per il metodo SE più avanti, strategia di scelta che ha gli stessi difetti di quelli che evidenzieremo per SE.

Detrendizzazione

Il secondo passo nella lista del paragrafo 4.1.3 è la semplice sottrazione del trend, la detrendizzazione. In realtà con questo nome si denota anche la fase di ricerca del trend,

discussa nel paragrafo precedente tramite operazioni di media locale, e ripresa in futuro mediante altri metodi (es. SET).

Trovato un trend, lo sottraiamo, trovando la serie detrendizzata:

$$y_i = x_i - t_i.$$

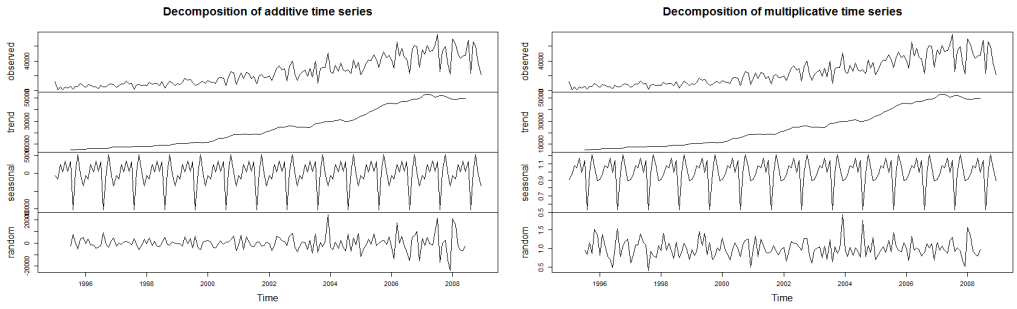
Unica osservazione: se si usa un modello moltiplicativo, del tipo $x_i = t_i\,(s_i + \varepsilon_i)$, si deve porre

$$y_i = \frac{x_i}{t_i}.$$

In entrambi i casi, la nuova serie y_i avrà trend pressapoco nullo e si cercherà di scomporla nella forma

$$y_i = s_i + \varepsilon_i.$$

I seguenti due disegni mostrano la decomposizione effettuata dal comando `decompose` nella modalità additiva (già vista sopra) e moltiplicativa.



La serie originaria ha una struttura chiaramente moltiplicativa e questo si vede bene nella maggior regolarità dei residui (residui meno strutturati, = miglior modello) ottenuti con la decomposizione moltipicativa. Si noti che il trend è identico, perché così è il procedimento che lo determina. Poi, eseguita le detrendizzazione, le serie y_i nei due casi sono diverse, da cui segue la diversità delle componenti stagionali e dei residui.

Componente stagionale

Per trovarla col metodo che illustriamo ora si deve aver deciso in anticipo il periodo. E questo si può cercare di fare mescolando l'intuito (sia visivo della serie, sia basato sulla natura della serie, es. dati mensili di vendite che risentono delle stagioni) con le informazioni provenienti dalla funzione di autocorrelazione.

Deciso il periodo k_{per}, si isolino i vari blocchi di un periodo:

$$x_1, ..., x_{k_{per}}$$
$$x_{k_{per}+1}, ..., x_{2k_{per}}$$
$$x_{2k_{per}+1}, ..., x_{3k_{per}}$$
$$...$$

che, se k_{per} è un buon periodo, dovrebbero somigliare. Attenzione: se la serie è stata detrendizzata, allora possono davvero somigliare in senso stretto, mentre se non è stata detrendizzata, essi possono differire per una costante, cioè avremmo

$$\begin{aligned} x_{k_{per}+1} &\sim x_1 + C \\ x_{2k_{per}+1} &\sim x_{k_{per}+1} + C \sim x_1 + 2C \end{aligned}$$

ecc. Quindi è importante aver eseguito prima la detrendizzazione.

Eseguiamo ora una media sul periodo:

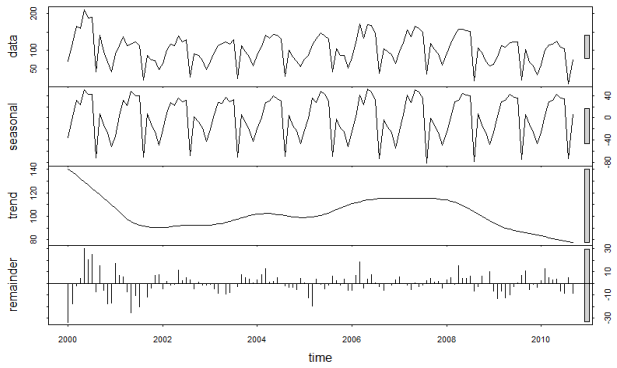
$$\begin{aligned} &\frac{x_1 + x_{k_{per}+1} + x_{2k_{per}+1} + ...}{k} \\ &\frac{x_2 + x_{k_{per}+2} + x_{2k_{per}+2}}{k} \\ &... \end{aligned}$$

se decidiamo di mediare k valori. Per esemplificare, se la serie è fatta di dati mensili, stiamo mediando i valori di gennaio, poi mediamo i valori di febbraio, ecc. in modo da trovare il valor medio di gennaio, quello di febbraio e così via. Il risultato è un profilo annuale medio.

Però possiamo prendere k piccolo (es. 2, 3) oppure pari al numero n_0 di periodi (ipotizziamo che sia $n_0 \cdot k_{per} = n$). Se lo prendiamo piccolo, stiamo mediando solo i primi mesi di gennaio della serie, mentre se lo prendiamo pari a n_0 stiamo mediando tutti i valori di gennaio della serie. Sono quindi due strategie diverse, tra cui possiamo scegliere. Se riteniamo, da un'analisi visiva o del significato applicativo della serie, che la periodicità sia abbastanza uniforme nel tempo lungo la serie, allora forse conviene $k = n_0$. Se invece ci sembra che la struttura periodica cambi nel tempo (nei primi anni sia di un tipo, poi via via di un'altro), allora conviene prendere k piccolo (in questo caso, il segmento di mesi di gennaio mediati va spostato nel tempo, con la stessa logica dei metodi di media locale).

Questa descrizione concettuale è resa operativa dal software R attraverso due comandi. Il comando `decompose` esegue una media globale ($k = n_0$) dei periodi. Come si può vedere nei disegno precedenti, la componente periodica è uniforme su tutto l'arco temporale.

Invece, il comando `stl` esegue una media locale. L'algoritmo di media locale è più complicato (ad esempi bisogna far avanzare la finestra su cui si media) ed il comando `stl` esegue un algoritmo piuttosto complesso che non descriviamo in dettaglio; l'idea comunque è quella detta: esso esegue una media locale sia per la ricerca del trend sia per la stagionalità. Ecco il risultato sulla serie delle moto:



Il comando `stl` richiede l'assegnazione di un parametro k; se è piccolo, la media dei periodi è molto locale, su pochi periodi; se è grande, la media è eseguita su molti periodi ed il risultato somiglia a quello di decompose.

Previsione

Il calcolo di trend e stagionalità (in generale il calcolo di ogni elemento strutturale) può permettere di eseguire previsioni. Bisogna essere in grado innanzi tutto di prolungare il trend nel futuro. Se il trend trovato è abbastanza rettilineo, basterà trovare la retta di regressione del trend, che quasi conicide con esso, e proungarla. Più in generale, salvo usare polinomi o altre funzioni al posto di rette, si possono usare dei metodi che spiegheremo in seguito, es. SET.

Prolungato il trend nel futuro, gli si aggiunge (o moltipica) la componente stagionale. Se si usa quella fornita dal comando `decompose`, non ci sono scelte da operare. Se invece si preferisce quella fornita da `stl`, essendo essa non uniforme, va scelto quale usare, e la cosa più naturale è usare l'ultima.

Infine, se dei residui si riesce a cogliere qualche elemento strutturale, questo può essere utilizzato in alcuni casi (rari) per migliorare ancora le previsioni, oppure (più spesso) per quantificare l'errore di previsione.

Calcolo ed analisi dei residui

Trovata anche la stagionalità s_i della serie, si calcolano i residui

$$\varepsilon_i = y_i - s_i.$$

Si vedano ad esempio i residui dati dai metodi `decompose` e `stl`, nelle figure precedenti.

Essi potrebbero avere ugualmente qualche elemento di struttura, cioè non essere del tutto casuali. Per residui del tutto casuali si intenderebbe una successione di numeri simile a quelle generate in modo indipendente da una stessa distribuzione, ad esempio numeri gaussiani di media nulla e deviazione standard σ_ε. Come capire se sono casuali o meno? A parte l'uso di test statistici un po' complessi, alcune cose elementari che

si possono fare sono ovviamente la loro visualizzazione, ma anche un loro istogramma, un q-q-plot, la loro funzione di autocorrelazione.

Esercizio 9 *Si generi una stringa di numeri casuali gaussiani e si tracci l'acf. Ripetendo più volte i comandi si vede la struttura di una acf emprica di numeri casuali indipendenti. Essa assume sostanzialmente valori irrilevanti a partire già da* $k = 1$.

Più difficile è sfruttare queste informazioni. Se ad esempio i residui sono abbastanza causali ma crescono in σ nel tempo, che possiamo fare per arricchire il modello? Sapere che i residui aumentano via via nel tempo non permette di migliorare la previsione ma può darci una valutazione più realistica dell'errore di previsione.

4.2 Il metodo di Holt-Winters

Lo scopo di questa sezione è quello di illustrare il metodo di Holt-Winters (HW), partendo dalle versioni particolari dette dello smorzamento esponenziale (SE) e dello smorzamento esponenziale con trend (SET).

4.2.1 Metodo di Smorzamento Esponenziale (SE)

Sia $x_1, ..., x_n$ la serie storica in esame. Nel metodo SE si introduce una serie storica ausiliaria $p_1, ..., p_{n+1}$ che rappresenta la serie delle *previsioni*. Siccome è facile equivocare, spieghiamo in dettaglio cosa si intenda per p_i: essa è *la previsione del valore relativo al tempo* i*, eseguita al tempo* $i - 1$. In altre parole, p_2 è ciò che, al tempo 1, siamo in grado di prevedere del valore relativo al tempo 2 (il tempo 2 è il futuro, quando siamo al tempo 1); p_3 è ciò che, al tempo 2, siamo in grado di prevedere del valore relativo al tempo 3 (che è il futuro, quando siamo al tempo 2); e così via. Siccome l'ultimo tempo presente possibile è n, esiste anche p_{n+1} nella serie delle previsioni, che è il valore del vero futuro (gli altri erano dei futuri condizionati a dei finti presenti).

Il Metodo di Smorzamento Esponenziale sceglie come previsione p_{i+1} del futuro una media pesata tra la previsione p_i del presente, fatta in precedenza, ed il valore attuale x_i della serie storica:

$$p_{i+1} = \alpha x_i + (1 - \alpha)\, p_i.$$

Il parametro α è (di solito) compreso tra 0 ed 1.

Se $\alpha = 1$, significa che decidiamo come previsione futura il valore odierno: $p_{i+1} = x_i$, cioè la pura ripetizione del presente. Se ad esempio la serie x_i ha delle oscillazioni, queste vengono riprodotte esattamente, in ritardo di un'unità temporale.

Se $\alpha = 0$, vale $p_{i+1} = p_i$, quindi $p_{i+1} = p_i = ... = p_1$, cioè la previsione è costante. Scegliamo come previsione una costante (nella migliore delle ipotesi la media dei dati). Ignoriamo ogni struttura ulteriore, ogni variazione.

Se invece prendiamo $\alpha \in (0,1)$, mediamo tra questi due estremi: otterremo una previsione meno oscillante dei dati reali, ma non del tutto costante, lievemente concorde con l'ultimo dato.

La previsione p_{i+1} è una media pesata tra un valore *conservativo*, p_i, ed uno *innovativo*, x_i.

Un'ulteriore interpretazione viene dalla formula che si ottiene applicando la ricorsione:

$$\begin{aligned} p_{i+1} &= \alpha x_i + (1-\alpha)\, p_i \\ &= \alpha x_i + (1-\alpha)\,(\alpha x_{i-1} + (1-\alpha)\, p_{i-1}) \\ &= \alpha x_i + (1-\alpha)\,(\alpha x_{i-1} + (1-\alpha)\,(\alpha x_{i-2} + (1-\alpha)\, p_{i-2})) \end{aligned}$$

ecc. che si riscrive

$$p_{i+1} = \alpha x_i + \alpha\,(1-\alpha)\, x_{i-1} + \alpha\,(1-\alpha)^2\, x_{i-2} + ...$$

Vediamo che la previsione futura è una media pesata di tutti i valori passati, con pesi che descrescono esponenzialmente (da qui il nome SE). Si sceglie come previsione una media pesata che dà più importanza agli ultimi valori rispetto ai precedenti. Quanta più importanza, lo decide α (α vicino ad 1 vuol dire molta più importanza ai valori più recenti). E' un po' come un metodo di media mobile ma pesata in modo non uniforme.

Un difetto del metodo è che, se la serie storica x_i ha un trend, per $\alpha = 0$ il metodo darebbe una costante, pessima previsione; per $\alpha = 1$, viceversa fa il meglio che può, ma può solo inseguire il trend in ritardo di un'unità temporale (si pensi a due rette parallele).

Ricerca automatica di α. SE come metodo per la ricerca di un trend

Il software R di default ricerca il miglior α secondo il seguente criterio (che vedremo essere opinabile). Dato α, calcoliamo la serie ausiliaria delle previsioni $p_1, ..., p_n$ (anche solo fino al tempo n) che, per evidenziare che dipendono dalla scelta fatta di α, scriviamo come $p_i(\alpha)$. Poi, introduciamo i numeri

$$\varepsilon_i(\alpha) = p_i(\alpha) - x_i, \qquad i = 1, ..., n.$$

Essi somo gli errori che il metodo ha commesso nelle previsioni dei dati noti; possiamo chiamarli *residui*, a patto di non confonderli con i residui dei modelli di decomposizione visti sopra. Cerchiamo allora il valore di α che minimizza la somma dei quadrati dei residui:

$$\min_{\alpha\in[0,1]} \sum_{i=1}^{n} \varepsilon_i(\alpha)^2 .$$

Il software trova così il valole di α.

L'automatismo potrebbe far credere che questa sia una regola obbligata o giusta. Tuttavia, questa regola è ispirata al principio che le previsioni p_i debbano essere più vicine possibili ai valori reali x_i, principio ovvio se si sta veramente parlando di previsioni. Se invece si utilizza SE per trovare un trend (vista la sua somiglianza col metodo di media mobile), non conviene inseguire i dati, le loro oscillazioni, ma mediare di più, quindi usare un valore più conservativo di α. Questo ragionamento, pir giusto, viene comunque annullato dalla sezione successiva, dove vedremo il metodo SET comunque vincente se si cerca di individuare un trend. Resta invece utile il principio con cui si cerca il parametro α, perché è lo stesso per SET ed Holt-Winters.

4.2.2 Metodo di Smorzamento Esponenziale con Trend (SET)

Due caratteristiche concettuali del metodo SE sono l'introduzione di una grandezza ausiliaria (il vettore delle previsioni) e la formula iterativa che ne lega i valori futuri a quelli presenti ed alla serie storica. Il metodo SET che ora vedremo conserva entrambi questi elementi potenziandoli però tramite due ulteriori grandezze ausiliarie. Geometricamente parlando, esso introduce una retta ausiliaria per ogni istante di tempo, retta che descrive la pendenza (il trend) a quell'istante. Una retta è identificata da due numeri, intercetta e coefficiente angolare; quindi, algebricamente, il metodo introduce due grandezze ausiliarie.

Indichiamo sempre con $x_1, ..., x_n$ la serie storica di partenza e con $p_1, ..., p_{n+1}, p_{n+2}, ...$ la serie ausiliaria delle previsioni. Le ulteriori due serie storiche ausiliarie di SET, che corrispondono a intercetta e coefficiente angolare, sono:

$$s_1, ..., s_n$$
$$m_1, ..., m_n.$$

Se siamo al tempo i, il presente, la previsione dei valori futuri p_{i+1}, p_{i+2}, ecc., è data dalla formula

$$p_{i+k} = m_i \cdot k + s_i$$

(equazione della retta di coefficiente angolare m_i ed intercetta s_i. E' utile pensare che l'asse delle ordinate sia collocato al tempo i, per farsi un'idea grafica.

L'idea di far dipendere m e s dal tempo è basilare: vogliamo sì una previsione con trend lineare, ma dobbiamo poterla modificare nel tempo, se il trend cambia.

Mentre nel metodo SE legavamo i valori futuri di p a quelli passati, ora leghiamo i valori futuri delle grandezze ausiliarie m ed s a quelli passati. Continuiamo ad usare una logica di media pesata tra un valore conservativo ed uno innovativo. Per il coefficiente angolare m la prima idea che viene in mente è

$$m_i = \beta\,(x_i - x_{i-1}) + (1 - \beta)\,m_{i-1}$$

media pesata tra il valore conservativo m_i e quello innovativo $x_i - x_{i-1}$, che è la pendenza osservata sui dati dell'ultimo periodo. Ma così ci si espone troppo alle fluttuazioni

casuali dei dati: la pendenza $x_i - x_{i-1}$ può deviare marcatamente dall "pendenza media" dell'ultimo periodo. Serve una grandezza simile a $x_i - x_{i-1}$ ma più stabile, meno esposta alle fluttuazioni causali. Essa è $s_i - s_{i-1}$, come capiremo tra un momento.

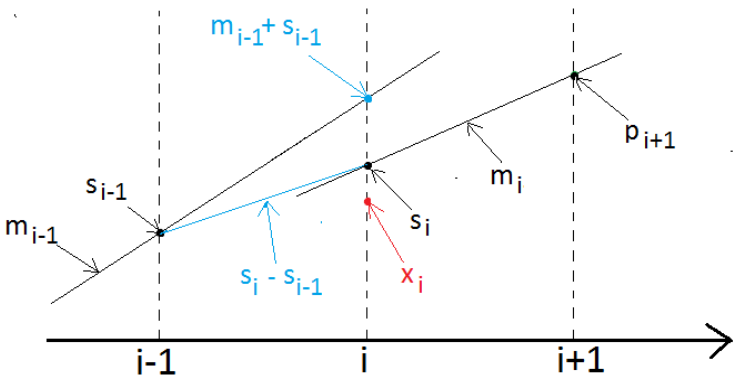
Veniamo alla ricorsione per s_i. Se disegnamo un grafico con due assi verticali delle ordinate, uno al tempo $i-1$ ed uno al tempo i, vediamo che l'intercetta al tempo i non deve essere simile all'intercetta al tempo $i-1$ ma al valore sulla retta $k \longmapsto m_{i-1} \cdot k + s_{i-1}$. I due istanti $i-1$ e i distano di un'unità, quindi s_i è giusto che sia legata a $m_{i-1} + s_{i-1}$ (cioè $k = 1$). Questa è la parte conservativa della ricorsione. La parte innovativa è dire che s_i deve essere legato al valore vero x_i. Quindi s_i sarà una media pesata tra x_i e $m_{i-1} + s_{i-1}$. In conclusione, le due equazioni ricorsive sono:

$$\begin{aligned} s_i &= \alpha x_i + (1-\alpha)(m_{i-1} + s_{i-1}) \\ m_i &= \beta(s_i - s_{i-1}) + (1-\beta) m_{i-1}. \end{aligned}$$

In pratica è necessario calcolare prima s_i dalla prima equazione, per poterlo sostituire nella seconda. Infine, si calcola la previsione al tempo successivo tramite

$$p_{i+1} = m_i + s_i.$$

Tutti questi ragionamenti sono illustrati nel disegno seguente:



In esso si vedono tracciate le due rette ausiliarie ai tempi $i-1$ ed i (in nero, con le frecce che indicano intercetta e coefficiente angolare di entrambe). La retta del tempo $i-1$ fornisce il valore (indicato in azzurro) $m_{i-1} + s_{i-1}$ al tempo i, che viene preso come valore conservativo di s_i; mentre il valore innovativo x_i è raffigurato in rosso. Il valore s_i è una combinazione convessa di $m_{i-1} + s_{i-1}$ ed x_i. Poi viene calcolato il coefficiente angolare m_i come combinazione convessa tra $s_i - s_{i-1}$ (pendenza innovativa, raffigurata in azzurro) ed m_{i-1} (pendenza conservativa); si noti che abbiamo rispettato, nel disegno, il fatto che m_i è intermedio tra $s_i - s_{i-1}$ e m_i, così come il fatto che s_i è intermedio tra $m_{i-1} + s_{i-1}$ ed x_i. Infine, viene raffigurata la previsione p_{i+1} relativa al tempo $i+1$, eseguita al tempo i tramite la retta ausiliaria appena calcolata.

Il metodo è ottimo per catturare i trend. Ma se c'è anche un'evidente periodicità, il metodo potrebbe non riuscire a riconoscerla come tale, potrebbe inseguirla come se

fosse una modifica continua del trend. Il problema si presenta in fase di previsione: siccome la previsione successiva è data dalla retta $p_{n+k} = m_n \cdot k + s_n$, si ottiene una previsione che ha come pendenza quella dell'ultimo tratto del profilo periodico, cosa senza senso per k grandi (ragionevole invece per k molto piccoli).

Il calcolo dei parametri ottimali effettuato dal software avviene secondo il criterio

$$\min_{\alpha,\beta\in[0,1]} \sum_{i=1}^{n} \varepsilon_i (\alpha, \beta)^2$$

dove

$$\varepsilon_i (\alpha, \beta) = p_i (\alpha, \beta) - x_i$$

$p_i (\alpha, \beta)$ essendo la previsione calcolata relativamente ai parametri fissati (α, β).

Inizializzazione di SE ed SET

Le equazioni per ricorrenza vanno inizializzate. Supponiamo che la serie temporale x_i parta da $i = 1$. Per SE, si tratta di stabilire p_1, la previsione al primo istante temporale. In questo modo, se ci troviamo al tempo $i = 1$ (nostro presente) e conosciao quindi x_1, potremo poi calcolare la previsione futura p_2:

$$p_2 = \alpha x_1 + (1 - \alpha) p_1.$$

Quando poi ci troveremo al tempo $i = 2$, che sarà diventato il nostro presente, conosceremo x_2 e potremo calcolare la previsione futura p_3:

$$p_3 = \alpha x_2 + (1 - \alpha) p_2.$$

E così via. Il valore di inizializzazione p_1 è abbastanza casuale da scegliere. Per semplicità si può prendere ad esempio $p_1 = x_1$.

Per SET, se ci troviamo al tempo $i = 1$ e vogliamo calcolare la previsione p_2, servono i valori m_1 e s_1. Vista la natura di s descritta sopra (si pensi al grafico con l'asse verticale per $i = 1$), è naturale prendere $s_1 = x_1$. La pendenza iniziale però è indecidibile senza vedere il futuro, quindi scegliamo $m_1 = 0$, salvo abbiamo informazioni diverse di tipo previsivo. Fatte queste scelte, possiamo calcolare le previsioni

$$p_2 = m_1 + s_1.$$

Poi il tempo $i = 2$ diventerà il nostro presente. Calcoleremo m_2 e s_2 con le formule (ora x_2 è noto)

$$\begin{aligned} s_2 &= \alpha x_2 + (1 - \alpha)(m_1 + s_1) \\ m_2 &= \beta (s_2 - s_1) + (1 - \beta) m_1 \end{aligned}$$

da usarsi nell'ordine scritto. Avendo calcolato m_2 e s_2, la previsione del futuro è

$$p_3 = m_2 + s_2.$$

Si può inizializzare SET in un secondo modo: attendere alcuni istanti in più prima di iniziare la previsione, ed utilizzarli per calcolare una retta di regressione, da usarsi come stima iniziale della pendenza. Chiaramente, più è lungo il periodo iniziale che attendiamo, più è precisa la stima della pendenza, quindi le previsioni inizieranno molto megli che con la semplice posizione $m_1 = 0$. Ma è anche vero che, se iniziamo all'istante iniziale con $m_1 = 0$, dopo alcuni istanti questa anomalia sarà stata automaticamente aggiustata dal metodo, tramite le iterazioni che correggono di volta in volta m tramite i valori degli incrementi $s_i - s_{i-1}$. Quindi alla fine le cose si equivalgono abbastanza: i primi istanti o non vengono proprio previsti oppure sono previsti un po' male; i successivi vengono previsti piuttosto bene.

4.2.3 Smorzamento esponenziale con trend e stagionalità (Holt-Winters)

Di questo metodo esiste una versione per stagionalità additiva ed una per quella moltiplicativa; descriviamo solo quest'ultima, essendo l'altra del tutto simile e forse più elementare. Si ipotizza il modello

$$x_i = (a \cdot i + b)\, f(i) + \varepsilon_i$$

con f funzione periodica di periodo k_{per}. Per capire nel modo più semplice possibile come sono state idete le equazioni ricorsive del modello, fingiamo di non avere il rumore ε_i, quindi di lavorare sull'equazione

$$x_i = (a \cdot i + b)\, f(i).$$

Idealmente, si introduce la grandezza ausiliaria $y_i = \frac{x_i}{f(i)}$ che soddisfa

$$y_i = a \cdot i + b.$$

A questa possiamo applicare quindi lo smorzamento con trend. Detta p^y la previsione di y e detti s^y, m^y i valori calcolati dal metodo SET relativamente ad y, si trova

$$p^y_{i+1} = m^y_i + s^y_i$$

dove (si noti che c'è y_i e non x_i)

$$\begin{aligned} s^y_i &= \alpha y\,(t) + (1-\alpha)\left(m^y_{i-1} + s^y_{i-1}\right) \\ m^y_i &= \beta\left(s^y_i - s^y_{i-1}\right) + (1-\beta)\, m^y_{i-1}. \end{aligned}$$

Il problema è che per innescare questo sistema bisogna conoscere y_i e per questo bisognerebbe conoscere $\frac{x_i}{f(i)}$, mentre $f(i)$ per ora è incognita. L'idea è di stimare anche la funzione periodica f in modo iterativo, così da aggiustarla se è il caso. Allora al posto di y_i si mette $\frac{x_i}{f(i-k_{per})}$, immaginando che nella struttura iterativa che troveremo alla fine il valore $f(i-k_{per})$ sia noto (e riteniamo sia una buona approssimazione di $f(i)$ in quanto cerchiamo una f periodica).

Poi bisogna creare un'equazione iterativa per f. Un'idea ispirata alla filosofia dello smorzamento esponenziale è

$$f(i) = \gamma \frac{x_i}{y_i} + (1-\gamma)\, f(i-k_{per})$$

(si ricordi la definizione $y_i = \frac{x_i}{f(i)}$; se non si mettesse alcun termine legato al passato useremmo l'equazione $f(i) = \frac{x_i}{y_i}$). Però non conosciamo y_i. Noti però s_i^y ed m_i^y, s_i^y è una stima di y_i. In definitiva, si arriva al sistema:

$$\begin{aligned} s_i &= \alpha \frac{x_i}{f(i-k_{per})} + (1-\alpha)\,(m_{i-1} + s_{i-1}) \\ m_i &= \beta\,(s_i - s_{i-1}) + (1-\beta)\, m_{i-1} \\ f(i) &= \gamma \frac{x_i}{s_i} + (1-\gamma)\, f(i-k_{per}) \end{aligned}$$

dove abbiamo smesso di indicare y a pedice in quanto ormai usiamo queste equazioni come equazioni finali per stimare x.

La previsione futura si svolge con la formula

$$p_{n+h} = (s_n + m_n \cdot h) \cdot f\,(n - k_{per} + h)\,, \qquad h = 1, 2, ..., k_{per}.$$

Si noti che questa previsione futura è più articolata, non semplicemente rettilinea, come per i metodi SE ed SET. Si può facilmente prolungare a valori più grandi di h.

Il software determina i parametri col criterio

$$\min_{\alpha,\beta,\gamma\in[0,1]} \sum_{i=1}^{n} \varepsilon_i\,(\alpha,\beta,\gamma)^2$$

con le solite convenzioni per i simboli.

Inizializzazione di HW

L'inizializzazione qui è più complessa. Serve f su un intero periodo per innescare l'iterazione. Allora si sacrifica il primo periodo (a volte più di uno), su quello si trova una retta di regressione

$$z\,(i) = a \cdot i + b$$

e la si usa come se fosse una stima di y_i. Quindi dalla definizione $y_i = \frac{x_i}{f(i)}$ si stima

$$f(i) = \frac{x_i}{a \cdot i + b}.$$

In definitiva, per $i = 1, 2, ..., k_{per}$ si prendono questi valori di f e poi si pone $s_{k_{per}} = a \cdot k_{per} + b$, $m_{k_{per}} = a$. Si comincia quindi a prevedere il valore al tempo $k_{per} + 1$.

4.3 Regressione lineare multipla applicata alle serie storiche

Essa parte dall'idea che la serie storica $x_1, ..., x_n$ possegga una struttura di legame interno tra i valori, del tipo

$$x_i = a_1 x_{i-1} + a_2 x_{i-2} + ... + a_p x_{i-p} + b + \varepsilon_i.$$

Ovviamente, a patto di definire i residui per differenza, una tale identità è sempre vera, qualsiasi siano i parametri. Il punto è sperare che, per un'opportuna scelta dei parametri, i residui siano molto piccoli, così che la formula ricorsiva rispecchi fedelmente la struttura dei dati. Se, per certi parametri $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$ ciò avviene (cioè i residui sono molto piccoli), diventa sensato prevedere il valore al tempo $n + 1$ tramite lo stesso modello coi residui posti a zero:

$$p_{n+1} = \widehat{a}_1 x_n + \widehat{a}_2 x_{n-1} + ... + \widehat{a}_p x_{n+1-p} + \widehat{b}. \tag{4.3}$$

Naturalmente non siamo obbligati ad usare tutti i ritardi in input, ma possiamo restringerci a quelli che per qualche ragione riteniamo più significativi (magari trovati con un procedimento di eliminazione).

Un esempio tipico è il seguente. Si immagini che, per sua natura (ad esempio dati mensili di tipo economico), i valori della serie storica distanziati di 12 unità (= mesi) siano simili, o comuque legati; e che però anche l'ultimo valore (mese) precedente sia rilevante, perché contiene informazioni relative all'andamento attuale del mercato. Allora si può ipotizzare un modello del tipo

$$x_i = a_1 x_{i-1} + a_{12} x_{i-12} + b + \varepsilon_i. \tag{4.4}$$

L'implementazione col software della regressione alle serie storiche non è completamente automatica. Bisogna crearsi le strighe di dati che corrispondono alle variabili di input $X_1, ..., X_p$ ed alla variabile di output Y della regressione. Per facilitare la comprensione, illustriamo il procedimento per una serie storica del tipo

$$x_1, ..., x_{100}$$

e per il modello semplice (4.4). Il vettore di dati che corrisponde all'output Y è

$$x_{13}, ..., x_{100}$$

mentre i vettori di dati che corrispondono alle variabili di input X_1 e X_{12} sono rispettivamente

$$x_{12}, ..., x_{99}$$
$$x_1, ..., x_{88}$$

(per convincersi basta pensare che $x_{13} = a_1x_{12}+a_{12}x_1+b+\varepsilon_{13}$, $x_{100} = a_1x_{99}+a_{12}x_{88}+b+\varepsilon_{100}$). Così è come se avessimo costruito la tabella di dati a cui applicare la regressione:

X_1	X_{12}	Y
x_{12}	x_1	x_{13}
...	...	...
x_{99}	x_{88}	x_{100}

I comandi col software R sono i seguenti. Supponiamo di aver caricato in R la serie storica nel vettore `X`. Poniamo

```
Y=X[13:100]; X1=X[12:99]; X12=X[1:88]
```

e poi eseguiamo la regressione

```
Reg = lm(Y~X1+X12)
```

Vediamo infine come si effettua la previsione. Detti `a1`, `a12`, `b` i coefficienti usciti dalla regressione appena eseguita, possiamo calcolare p_{n+1} con la formula (4.3). Cerchiamo però di andare oltre: supponiamo di volere la previsione di più istanti successivi. Per capire il problema, si pensi al solito esempio semplice (4.4). Se vogliamo p_{n+2}, il modello avrebbe bisogno di x_{n+1}:

$$p_{n+2} = \widehat{a}_1 x_{n+1} + \widehat{a}_{12} x_{n-10} + \widehat{b}$$

($n+2-12 = n-10$). Ma non possediamo x_{n+1}. Allora usiamo la sua previsione p_{n+1}:

$$p_{n+2} = \widehat{a}_1 p_{n+1} + \widehat{a}_{12} x_{n-10} + \widehat{b}.$$

E così via, osservando che dopo 12 mesi inizieremo a dover sostituire anche il termine che moltiplica $\widehat{a}_{12}$. Allora un trucco per realizzare questo col software è di introdurre un vettore ausiliario `P` che contenga all'inizio la serie storica ed alla fine delle caselle vuote da riempire con le formule precedenti. Supponiamo di voler prevedere 12 mesi:

```
P=1:112; P[1:100]=X
P[101]=a1*P[100]+a12*P[89]+b
P[102]=a1*P[101]+a12*P[90]+b
```

ecc., sintetizzabile con un ciclo di for. L'aver sostituito `X` alle prime componenti di `P` ha permesso di evitare di distinguere i vari casi (cioè quando, in input, si può usare il vero `X` e quando invece si è costretti ad usare `P` stesso).

4.3.1 Variabili esogene, cross-correlazione

Tra le caratteristiche uniche dell'uso della regressione lineare multipla nell'ambito delle serie storiche c'è la possibilità di inserire, tra i predittori, serie storiche diverse da quella data, serie che riteniamo possano essere utili per la previsione della serie data; esse corrispondono alle cosidette *variabili esogene*. Possiamo quindi creare modelli del tipo (a volte detti anche detti ARX)

$$\begin{aligned} x_i &= a_1 x_{i-1} + a_2 x_{i-2} ... \\ &\quad + c_1 z_{i-1} + c_2 z_{i-2} ... \\ &\quad + b + \varepsilon_i \end{aligned}$$

dove $z_1, ..., z_n$ è un'altra serie storica (ovviamente si possono usare, come predittori, diverse serie storiche). Ad esempio, si può immaginare che il costo di certi beni di consumo siano influenzati dal prezzo del petrolio nel mese precedente. Allora x_i è il costo del bene in oggetto, z_{i-1} è il prezzo del petrolio nel mese precedente, x_{i-1} è il costo del bene considerato, relativo al mese precedente, ecc. E' chiara la flessibilità di questo metodo. Purtroppo la sua applicazione pratica richiede pazienza ed arte. E' esperienza comune che, a dispetto delle aspettative, il legame tra diverse serie storiche potenzialmente collegate sia più debole del previsto e spesso insignificante rispetto al legame interno (autocorrelazione) della serie $x_1, ..., x_n$ stessa; in altri termini, il vantaggio come varianza spiegata dall'aggiunta dei fattori esogeni è spesso molto basso, anche se a priori ci si aspetterebbe maggior successo visto il significato intuitivo delle serie in gioco. La metodologia è comunque significativa per le sue potenzialità e fa parte della cosidetta *econometria*.

Prima di buttarsi nell'uso di tali modelli conviene assicurarsi che ci sia un legame tra le serie storiche che si vogliono mettere insieme, qui chiamate $z_1, ..., z_n$ e $x_1, ..., x_n$. Più precisamente, serve scoprire che ritardi sono più significativi, er usare il minor numero di parametri possibile. Se ad esempio la variabile esogena Z influisce su X, ma il suo influsso tarda tre mesi (mediamente) a manifestarsi, conviene usare un modello del tipo

$$\begin{aligned} x_i &= a_1 x_{i-1} + a_2 x_{i-2} ... \\ &\quad + c_3 z_{i-3} + b + \varepsilon_i \end{aligned}$$

piuttosto che un modello più completo con anche tutti i termini $c_1 z_{i-1}$ e così via.

La *cross correlation function* (ccf) serve a questo scopo. Esso calcola la correlazione tra le due serie (opportunamente troncate), rispetto a tutte le traslazioni possibili. Il procedimento è identico a quello descritto per la acf. Preso un intero k, qui però positivo o negativo, in valore assoluto minore di n, che chiameremo genericamente ritardo, ad es. $k = 2$, affianchiamo le due serie

x_1	x_2	x_3	...	...	...	...	x_{n-2}	x_{n-1}	x_n		
		z_1	z_2	z_3	...	...	...	...	z_{n-2}	z_{n-1}	z_n

prendiamo la parte comune

x_3	...	...	...	...	x_{n-2}	x_{n-1}	x_n
z_1	z_2	z_3	...	...	...	...	z_{n-2}

e ne calcoliamo la correlazione

$$\widehat{\rho}_{XZ}(k) = \frac{\sum_{i=1}^{n-k}(x_i - \overline{x}_0)(z_{i+k} - \overline{z}_k)}{\sqrt{\sum_{i=1}^{n-k}(x_i - \overline{x}_0)^2 \sum_{i=1}^{n-k}(z_{i+k} - \overline{z}_k)^2}}$$

dove $\overline{x}_0 = \frac{1}{n-k}\sum_{i=1}^{n-k} x_i$ $\overline{z}_k = \frac{1}{n-k}\sum_{i=k+1}^{n} z_i$. Ripetendo il procedimento per $k = -k_0, ..., -1, 0, 1, ..., k_0$, dove k_0 è un intero positivo minore di n, troviamo una nuova serie storica $\widehat{\rho}_{XZ}(k)$, detta *funzione di correlazione mutua empirica* (*cross correlation function*). I valori di k per cui essa è più grande sono i ritardi desiderati. Naturalmente il risultato è statisticamente sigificativo solo per k_0 basso.

Il software R esegue questo calcolo tramite il comando `ccf`. Attenzione quando lo si usa: l'ordine con cui si danno le due serie al comando `ccf` è essenziale; per k positivi si vedrà la correlazione tra l'una e le traslazioni in avanti dell'altra, e viceversa, ma come essere sicuri dell'ordine? Rileggendo l'help del comando. Esso recita: The lag k value returned by ccf(x,y) estimates the correlation between x[t+k] and y[t]. Quindi, supponiamo che si voglia scoprire se una certa serie storica x influisce su una y qualche mese dopo (i valori di x di gennaio si ripercuotono sui valori di y di marzo, e così via, ad esempio). Allora ccf(x,y) con k=-2 ci dice come x[t-2] è collegata a y[t].

Se la correlazione tra $z_1, ..., z_{n-k}$ e $x_{k+1}, ..., x_n$ è elevata (vicina ad 1), allora queste serie hanno un legame. Se lo riteniamo sensato dal punto di vista applicativo, possiamo estrapolare che $z_1, ..., z_{n-k}$ influisca su $x_{k+1}, ..., x_n$ e che quindi abbia senso impostare un modello regressivo con fattore esogeno, che spiega la serie x non solo attraverso la propria struttura ricorsiva, ma anche facendo uso del predittore z. Si noti che dev'essere k strettamente positivo; già $k = 0$ (cioè la correlazione tra le due serie storiche, senza traslazioni) non serve a nulla, perché dobbiamo cercare di prevedere il fururo della X usando passato e presente della Z, non il futuro (incognito) della Z. Peggio ancora ovviamente se il picco della ccf è per k negativi.

4.4 Residui

Il concetto dei residui compare più volte nel corso, sempre in modo complementare. Per sottolineare alcuni elementi di unità ed alcune differenze tra i vari ambiti, discutiamo nuovamente alcuni tratti della teoria dei residui.

4.4.1 Le definizioni

La *definizione* di residuo cambia un poco da un ambito all'altro, pur conservando la stessa idea di fondo.

Nella regressione lineare multipla (RLM) si parla del residuo dell'individuo i-esimo:

$$\varepsilon_i = y_i - (a_1 x_{i,1} + ... + a_p x_{i,p} + b)$$

dove $i = 1, ..., n$ è l'indice che distingue un individuo da un altro. Rammentiamo che si presenta la doppia visione: se i parametri $a_1, ..., a_p, b$ non sono ancora stati determinati, i residui sono funzioni dei parametri, $\varepsilon_i = \varepsilon_i(a_1, ..., a_p, b)$; se invece sono stati trovati (tipicamente col metodo dei minimi quadrati) allora i residui sono numeri fissati.

Nei metodi di decomposizione di serie storiche, se si usa un modello additivo del tipo $x_i = t_i + s_i + \varepsilon_i$, dove ora $i = 1, ..., n$ indica il tempo, t_i il trend e s_i la stagionalità, allora la definizione è

$$\varepsilon_i = x_i - (t_i + s_i).$$

Va però ricordato che trend e stagionalità non sono concetti univocamente definiti, per cui questa "definizione" di residui va intesa in senso più vago, oppure riferito ad uno specifico algoritmo di decomposizione. Ad esempio, è univoco parlare dei residui forniti dal comando `decompose`, o di quelli forniti da `stl` relativamente ad una ben precisa scelta di k.

Se si adotta invece un modello moltiplicativo della forma $x_i = t_i(s_i + \varepsilon_i)$, allora la definizione è

$$\varepsilon_i = \frac{x_i}{t_i} - s_i.$$

Nell'ambito dei tre metodi della categoria Holt-Winters, che indichiamo con SE, SET ed HW, detta sempre p_i la variabile ausiliaria della previsione (previsione del valore relativo all'istante i-esimo, effettuata all'istante $i-1$), la definizione di residuo al tempo i è

$$\varepsilon_i = p_i - x_i.$$

Ricordiamo che invece la definizione di p_i varia da un metodo all'altro:

$$p_{i+1} = \alpha x_i + (1-\alpha) p_i \qquad \text{per SE}$$

$$p_{i+1} = s_i + m_i \qquad \text{per SET}$$

$$p_{i+1} = (s_i + m_i) f(i - k_{per} + 1) \qquad \text{per HW.}$$

Infine, quando si usa la regressione nell'ambito delle serie storiche, si calcola la previsione usando la formula

$$p_{i+1} = \widehat{a}_1 x_i + \widehat{a}_2 x_{i-1} + ... + \widehat{a}_p x_{i+1-p} + \widehat{b}.$$

Anche in questo caso il residuo al tempo i è dato da

$$\varepsilon_i = p_i - x_i$$

Questo è comunque un caso particolare della definizione di residuo della RLM.

4.4.2 Uso ed utilità

Indichiamo almeno quattro utilizzi:

1. determinare i parametri ottimali di un modello
2. confrontare modelli diversi
3. stimare l'incertezza delle previsioni
4. rintracciare eventuale struttura residua.

Il punto 4 è troppo specialistico rispetto alla nostra trattazione, per cui non lo discutiamo.

Il punto 1 è già stato discusso ampiamente: sia per i modelli regressivi, sia per quelli di tipo Holt-Winters, i parametri vengono determinati col metodo dei minimi quadrati, cioè minimizzando la somma dei quadrati dei residui.

Confronto tra metodi

Discutiamo brevemente il punto 2. E' molto simile al punto 1: dati due metodi diversi che concorrono a studiare lo stesso problema, ad esempio SET e HW, oppure HW e RLM per serie storiche, calcoliamo la somma dei quadrati dei residui ed usiamo tale indicatore per dare un voto al metodo. Fermo restando che nulla garantisce che un voto migliore corrisponda a migliori prestazioni in fase di previsione.

Questa strategia di confronto tra metodi diversi può essere svolta in due modi diversi. Il primo modo, non avendo un nome, verrà chiamato scherzosamente del *conflitto di interessi*. Esso è del tutto simile al punto 1: si prendono tutti i residui (o solo quelli più recenti, se quelli troppo vecchi non sono più rappresentativi), di fa la somma dei quadrati e si calcola così il voto. Qui, per residui, si intendono quelli ottenuti applicando il metodo alla serie storica, dopo aver determinato i parametri ottimali. Il conflitto di interessi sta proprio in questo: si usano i dati storici $x_1, ..., x_n$ per determinare i parametri ottimali, ad es. di HW e di SET; poi si confrontano le prestazioni di HW ed SET sugli stessi dati che sono stati usati per ottimizzarli. Nella realtà invece, dovremmo confrontare HW ed SET sui dati futuri, per capire quale metodo sia davvero migliore. Ma siccome i dati futuri non li abbiamo, ci accontentiamo di ri-applicare HW e SET ai dati passati.

Di fatto, questo confronto dice solo quale dei due (o più) metodi fitta meglio i dati storici noti. Se poi questo produrrà migliori previsioni future, nessuna teoria può garantirlo.

Il secondo modo è quello della *cross-validation*. Il principio della cross-validation è abbastanza universale ed applicabile a tantissime teorie matematiche diverse; vediamolo all'opera sulle serie storiche. Si divide la serie storica in due parti: la prima, normalmente quella relativa ai dati più vecchi, è detta *training set*; la seconda, quella

dei dati più recenti, è detta *test set*. Si applicano i due metodi che vogliamo confrontare, es. SET ed HW, al training set, trovando così i parametri ottimali di entrambi i metodi. Poi, con essi, si effettua la previsione del periodo successivo, che è il test set. Si calcoleranno quindi i residui relativi al test set, poi la somma dei loro quadrati. In questo modo, si sta un po' simulando la realtà, in cui uno costruisce il modello sui dati a disposizione e poi deve esaminarne le prestazioni relativamente a come riesce a prevedere i dati futuri, non gli stessi con cui l'ha generato.

Questo metodo ha degli ovvi punti a suo favore, ma anche alcuni difetti. Uno è di tipo implementativo: richiede sicuramente più lavoro. L'altro è di tipo concettuale: ciò che noi possediamo davvero e vorremmo davvero confrontare sono (ad, es.) SET ed HW calibrati su *tutti* i nostri dati a disposizione; invece con questo metodo confrontiamo SET ed HW calibrati su dati vecchi, del passato, magari non più così rappresentativi del presente.

In definitiva, non ci sono metodi sicuri per decidere quale previsione sarà migliore. E' giusto così, nessuno può prevedere il futuro.

Valutazione dell'incertezza delle previsioni

Quando applichiamo un qualsiasi metodo di previsione di serie storiche, a partire dalla nostra serie $x_1, ..., x_n$ otteniamo un valore p_{n+1} (e magari altri successivi). Questa può essere detta *stima puntuale*, del valore al tempo $n+1$. A questa può essere onesto associare una *stima intervallare*, una valutazione dell'*incertezza* della stima puntuale. Essa consisterà in una frase del tipo: al 90% il valore al tempo n cadrà nell'intervallo

$$[p_{n+1} - \delta, p_{n+1} + \delta] .$$

Naturalmente ogni percentuale al posto del 90% ha ugualmente senso; il livello di confidenza lo dobbiamo decidere noi, in base alla gravità di un'eventuale stima sbagliata.

Il problema è quindi, fissato il livello di confidenza, es. 90%, come calcolare δ. Aggiungiamo un po' di flessibilità alla formulazione, accettando l'idea che sia più interessante o necessario un intervallo non simmetrico:

$$[p_{n+1} - \delta_-, p_{n+1} + \delta_+] .$$

Dovremo trovare δ_- e δ_+. Perché essi non siano definiti in modo ambiguo, stabiliamo che dividiamo 10% (il complementare di 90%) in parti uguali, 5% e 5%, e richiediamo che 5% sia la probabilità di avere valori inferiori a $p_{n+1} - \delta_-$, 5% quella di avere valori superiori a $p_{n+1} + \delta_+$, 90% in mezzo.

Si aprono due possibilità: calcolare δ_- e δ_+ in modo non parametrico, completamente empirico (nel senso di utilizzare solo i dati, in modo diretto); oppure in modo parametrico, (ovvero interponendo l'identificazione di una densità di probabilità).

Innanzi tutto però, si devono calcolare i residui ε_i. Essi vengono poi visti come un campione sperimentale (non si considera più il loro ordine temporale), come risultati

sperimentali di un'ipotetica v.a. residuo. Se si immagina che tale v.a. sia ad esempio gaussiana, useremo le gaussiane per calcolare δ_- e δ_+ (metodo parametrico), altrimenti useremo solo i residui, come dati grezzi, in modo non parametrico.

Il principio, come sempre, è quello che l'incertezza (l'errore) trovata sulla parte nota della serie storica, sia rappresentativa di quella futura. Quindi la distribuzione statistica dei residui, che rigurdano la parte nota della serie, ci darà informazioni sull'incertezza futura.

Nel metodo non parametrico dobbiamo trovare un numero δ_- tale che il 5% dei residui sia minore di δ_-; poi un numero δ_+ tale che il 5% dei residui sia maggiore di δ_+. Lo possiamo fare a mano: basta ordinare i residui in modo crescente (con un `sort`), poi dividerli in tre gruppi in modo proporzionare ai numeri 5%, 90%, 5%. Oppure, col software R, si può usare il comando `quantile`, precisamente `quantile(e,0.05)` e `quantile(e,0.95)`, dove `e` indica qui la serie storica dei residui.

Nel metodo parametrico, bisogna innanzi tutto decidere il tipo di densità che si preferisce o ritiene giusto usare. Questa fase può essere complessa ed esula da questo corso, per cui ci limitiamo a descrivere il seguito nel caso, sicuramente molto frequente, in cui si scelgano le gaussiane. Bisogna innanzi tutto prendere la gaussiana più ragionevole, che è quella con media e deviazione standard date da quelle empiriche dei residui (`mean(e)`, `sd(e)`). Poi si calcolano, per tale gaussiana, i quantili di ordine 0.05 e 0.95, ovvero

`qnorm(0.05,mean(e),sd(e))`

`qnorm(0.95,mean(e),sd(e)).`

Per inciso, essi sono simmetrici ($\delta_- = \delta_+$).

Questi sono, con l'uno o l'altro metodo, i valori $-\delta_-$ e δ_+, da aggiungere a p_{n+1} per trovare l'intervallo di confidenza al 90%.

4.5 Domande

1. Scrivere la formula per la funzione di autocorrelazione empirica
2. Quali elementi strutturali sono visibili tramite acf?
3. Come appare acf se la serie ha un trend accentuato? E per una serie ha accentuata stagionalità? Spiegare le affermazioni fatte tramite argomenti matematici.
4. Scrivere le formule generali di decomposizione di una serie storica. Descrivere una strategia generale per eseguire la decomposizione.
5. Scrivere le formule per la media simmetrica e la media mobile. A quali due diversi scopi sono indirizzate?
6. Che effetto ha la scelta dell'ampiezza k della finestra nel metodo della media mobile?

7. Scrivere le formule per la serie detrendizzata.

8. Come calcola la componente stagionale il comando decompose di R? In cosa differisce, da questo punto di vista, il comando stl?

9. Come si calcolano i residui di un metodo di decomposizione? E del metodo SE?

10. Descrivere le formule e le idee alla base del metodo SE (idem per SET, idem per Holt-Winters).

11. Cosa rappresenta p_i in uno di questi metodi?

12. Che effetto ha la scelta del parametro α nel metodo SE?

13. Come eseguono la ricerca automatica dei parametri, questi metodi?

14. Descrivere l'inizializzazione di SE (idem per SET, idem per Holt-Winters).

15. Descrivere, teoricamente e con i comandi del software, la regressione lineare multipla applicata alle serie storiche, evidenziando le difficoltà da superare in fase di previsione.

16. Descrivere i modelli a fattori esogeni ed il metodo ccf per la ricerca dei ritardi più opportuni.