1 Riassunto sulle Gaussiane

Nel caso unidimensionale, abbiamo definito la gaussiana canonica $N(0,1)$ e la gaussiana generica $N(\mu,\sigma^2)$ descrivendo esplicitamente le loro densità di probabilità. Si verifica con un po' di calcoli che μ e σ^2 sono media e varianza. Tra le varie proprietà c'è che

$$Z \sim N(0,1) \Rightarrow X := \sigma Z + \mu \sim N(\mu,\sigma^2)$$

e viceversa, data $X \sim N(\mu,\sigma^2)$, esiste $Z \sim N(0,1)$ (data da $Z := \frac{X-\mu}{\sigma}$) tale che $X = \sigma Z + \mu$. Quindi, una volta definita la gaussiana canonica tramite la sua densità, avremmo potuto definire le gaussiane generiche dicendo che erano le v.a. X della forma $X = \sigma Z + \mu$.

Nel caso multidimensionale, la definizione di gaussiana canonica è abbastanza naturale: un vettore $Z = (Z_1, ..., Z_n)$ si dice gaussiano canonico se le sue marginali $Z_1, ..., Z_n$ sono gaussiane canoniche indipendenti. Questo equivale a chiedere che la densità congiunta $f_Z(z_1, ..., z_n)$ di Z sia il prodotto delle densità marginali, quindi

$$f_Z(z_1, ..., z_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left(z_1^2 + ... + z_n^2\right)\right).$$

Invece la definizione di gaussiana generica è più laboriosa. Un modo veloce di risolvere questa difficoltà è di definire le gaussiane generiche X in più dimensioni come tutti i vettori aleatori $X = (X_1, ..., X_k)$ che si ottengono da gaussiane canoniche $Z = (Z_1, ..., Z_n)$ tramite operazioni affini:

$$X = AZ + \mu$$

dove A è una matrice $n \times k$ e $\mu = (\mu_1, ..., \mu_k)$ è un vettore dato. Abbiamo detto, senza dimostrazione, che se la matrice $k \times k$

$$Q = AA^T$$

è invertibile, allora X ha densità congiunta

$$f_X(x_1, ..., x_k) = \frac{1}{(2\pi)^{k/2}\sqrt{\det Q}} \exp\left(-\frac{1}{2}\left\langle Q^{-1}(x-\mu), x-\mu\right\rangle\right).$$

Si potrebbe partire da questa formula come definizione, ma col difetto che essa vale solo nel caso in cui Q sia invertibile. Si verifichi, a titolo di esempio,

che la densità unidimensionale di una $N(\mu, \sigma^2)$ è un caso particolare di questa espressione più complessa.

Sia sualla base della definizione $X = AZ + \mu$ sia sualla base della formula per la densità, si arriva a capire che gli insiemi di livello

$$f_X(x_1, ..., x_k) = \text{costante}$$

della densità sono ellissoidi (concentrici, aventi μ come centro). Inoltre, gli assi principali di questi ellissoidi sono dati dagli autovettori di Q. Gli autovalori di Q descrivono le ampiezze degli ellissoidi lungo i relativi assi. Quindi la matrice Q è essenziale per la comprensione della geometria delle gaussiane multidimensionali. Nle prossimo paragrafo interpreteremo Q come matrice di covarianza di X.

Osserviamo infine che se estraiamo un campione sperimentale da X, i punti sperimentali formeranno più o meno una nuvoletta ellissoidale simile agli ellissoidi di Q. Quindi Q è legato alla geometria dei dati sperimentali. Naturalmente questo permette anche di stimare Q a partire da tali dati.

2 Matrice di covarianza di una gaussiana

Partiamo dal fatto che una gaussiana generica $X = (X_1, ..., X_k)$ è definita come trasformazione affine di una gaussiana canonica $Z = (Z_1, ..., Z_n)$:

$$X = AZ + \mu$$
$$X_i = \sum_{j=1}^{n} A_{ij} Z_j + \mu_i$$

dove A è una matrice $n \times k$ e $\mu = (\mu_1, ..., \mu_k)$ è un vettore dato.

Da qui calcoliamo media e varianza di ogni componente X_i e covarianza tra le componenti, usando il fatto che $E[Z_j] = 0$, $Var[Z_j] = 1$, $Cov(Z_j, Z_{j'}) = 1$ per $i = i'$, $Cov(Z_j, Z_{j'}) = 0$ per $i \neq i'$:

$$E[X_i] = E\left[\sum_{j=1}^{n} A_{ij} Z_j + \mu_i\right] = \sum_{j=1}^{n} A_{ij} E[Z_j] + \mu_i = \mu_i$$

$$Var\left[X_i\right] = Var\left[\sum_{j=1}^{n} A_{ij}Z_j + \mu_i\right] = Var\left[\sum_{j=1}^{n} A_{ij}Z_j\right]$$

$$\sum_{j=1}^{n} Var\left[A_{ij}Z_j\right] = \sum_{j=1}^{n} A_{ij}^2 Var\left[Z_j\right] = \sum_{j=1}^{n} A_{ij}^2 = \left(AA^T\right)_{i,i}$$

$$\begin{aligned} Cov\left(X_i, X_{i'}\right) &= Cov\left(\sum_{j=1}^{n} A_{ij}Z_j, \sum_{j'=1}^{n} A_{i'j'}Z_{j'}\right) \\ &= \sum_{j=1}^{n}\sum_{j'=1}^{n} A_{ij}A_{i'j'}Cov\left(Z_j, Z_{j'}\right) \\ &= \sum_{j=1}^{n} A_{ij}A_{i'j} = \left(AA^T\right)_{i,i'} \end{aligned}$$

Emerge quindi la matrice $Q = AA^T$, le cui componenti

$$Q_{ij} = Cov\left(X_i, X_j\right)$$

sono le covarianze tra le componenti. Q è detta quindi *matrice di covarianza* della v.a. X.

La matrice di correlazione Θ è invece definita

$$\Theta_{ij} = \frac{Cov\left(X_i, X_j\right)}{\sqrt{Var\left[X_i\right]Var\left[X_j\right]}}.$$

3 Caso 2-dimensionale

La matrice di covarianza in questo caso ha la forma

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

quindi

$$\begin{aligned} Q^{-1} &= \frac{1}{\sigma_X^2\sigma_Y^2 - \rho^2\sigma_X^2\sigma_Y^2}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} \\ &= \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_X^2} & -\frac{\rho}{\sigma_X\sigma_Y} \\ -\frac{\rho}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y^2} \end{pmatrix}. \end{aligned}$$

Pertanto

$$\begin{pmatrix} x & y \end{pmatrix} Q^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{1-\rho^2} \left[\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} - 2\frac{\rho xy}{\sigma_X \sigma_Y} \right]$$

ed essendo $\det Q = (1-\rho^2)\,\sigma_X^2 \sigma_Y^2$, troviamo infine

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} - 2\frac{\rho xy}{\sigma_X\sigma_Y} \right] \right).$$

4 Ricostruzione da Q della rappresentazione $Q = AA^T$

Data una gaussiana $X = (X_1, ..., X_k)$ oppure un suo campione sperimentale, come possiamo risalire alla rappresentazione $X = AZ + \mu$? Il vettore μ è dato semplicemente dai valori medi. Poi possiamo calcolare (o stimare) Q. Come ricostruiamo A?

Intanto osserviamo che A non è unica: se U è una matrice di rotazione $n \times n$, avente la proprietà UU^T= identità (matrice ortogonale), allora $A' = AU$ soddisfa $A'(A')^T = Q$. Pertanto ci poniamo il problema di trovare almeno una A.

Passo 1: bisogna diagonalizzare Q, nel senso che bisogna trovare autovettori $v_1, ..., v_k$ ed autovalori $\lambda_1, ..., \lambda_k$ di Q. Questa parte è implicita nel metodo PCA. Con R le *radici quadrate* degli autovalori si leggono col comando *princomp*, gli autovettori col comando *princomp(...)$loadings*.

Passo 2. Se U indica la matrice avente come colonne i vettori $v_1, ..., v_k$, essa è ortogonale, e si verifica subito che

$$\Lambda := U^T Q U$$

è una matrice diagonale di componenti λ_i. E' la matrice Q letta nella base $v_1, ..., v_k$. Di conseguenza vale la rappresentazione

$$Q = U \Lambda U^T.$$

Allora basta prendere

$$A := U\sqrt{\Lambda}$$

dove $\sqrt{\Lambda}$ indica la matrice diagonale di componenti $\sqrt{\lambda_i}$.

Vale quindi la rappresentazione

$$X = U\sqrt{\Lambda} Z + \mu.$$

Eseguito il metodo PCA, abbiamo quindi anche una rappresentazione di X.

5 Previsione con le componenti principali?

Nel caso 2D, trovare le componenti principali equivale ad effettuare la regressione lineare. Si pensi ad una nuvola di punti nel piano. Se la nuvola è abbastanza ellissoidale, l'asse principale dell'ellissoide - che si trova col metodo PCA - è anche la retta di regressione, cioè la retta che passa più vicina ai dati. Quando si possiede la retta di regressione tra dati X ed Y, la si può usare per prevedere i valori di Y in corrispondenza a valori di X ancora non sperimentati.

In realtà questa corrispondenza tra PCA e regressione non è completamenta vera. Infatti l'usuale regressione lineare utilizza come criterio di vicinanza tra dati e retta la distanza in verticale tra questi. Invece PCA è più geometrico, legato quindi alla distanza euclidea (perpendicolare alla retta, per così dire).

In più di due dimensioni, l'intuizione geometrica è meno ovvia e non l'approfondiamo. Però il metodo Factor analysis dewscritto nel seguito è in un certo senso una risposta (se pur teorica) a questa domanda.

6 Ricerca di fattori comuni nascosti (Factor analysis)

Nella regressione lineare, che affronteremo altrove, si immagina di avere a che fare con un modello della forma

$$X_i = \mu_i + \sum_{j=1}^{m} \Lambda_{ij} f_j + \sigma_i \varepsilon_i$$

dove le f_j sono variabili d'ingresso, dette *fattori* o *predittori* o variabili indipendenti, le X_i sono variabili d'uscita dette anche variabili dipendenti, μ_i sono i valori medi delle X_i e $\sigma_i \varepsilon_i$ sono i loro errori casuali. In tale ambito, si conoscono sia le variabili indipendenti sia quelle dipendenti, mentre è sconosciuta la relazione che le lega (ovvero i coefficienti Λ_{ij} e l'ampiezza degli errori σ_i).

Nell'*Analisi Fattoriale* si hanno a disposizione solamente le variabili d'uscita X_i ed un loro campione sperimentale. Si immagina però che esistano dei fattori (o predittori) nascosti f_j, che il nostro intuito magari riesce ad immaginare grossolanamente ma che non siamo in grado di esplcitare a priori.

Si immagina queindi che debba valere un modello del tipo

$$X_i = \mu_i + \sum_{j=1}^{m} \Lambda_{ij} f_j + \sigma_i \varepsilon_i$$

dove le v.a. f_j e ε_i sono tutte indipendenti gaussiane canoniche. Un tale modello sarebbe interessante in quanto *spiegherebbe* il risultato delle variabili x_i sulla base dei *fattori comuni* f_j, a meno degli errori ε_i (che essendo indipendenti, non contengono ulteriore possibilità di essere *spiegati*). I coefficienti Λ_{ij} sono detti *loadings*.

Lo scopo della Factor analysis è di identificare questi fattori nascosti. I risultati quasi sempre sono un po' deludenti o impliciti o non univoci, insomma pieni di difetti. Non per questo deve essere sottostimata la metodologia, in quanto si deve riconoscere che cerca di affrontare un problema di enorme difficoltà concettuale. Limitiamoci ad alcuni cenni.

Se X è un vettore gaussiano, vale (si veda il paragrafo precedente sulla ricostruzione di A)

$$X_i = \mu_i + \sum_{j=1}^{n} (v_j)_i \sqrt{\lambda_j} Z_j$$

con le Z_j indipendenti gaussiane canoniche. Le Z_j fungono quindi da fattori comuni che spiegano le X_i. Ma hanno il difetto di essere troppi per essere utili a livello interpretativo.

Torniamo ora all'ipotesi iniziale che valga il modello

$$X_i = \mu_i + \sum_{j=1}^{m} \Lambda_{ij} f_j + \sigma_i \varepsilon_i$$

e vediamo come, fissato m, si possono calcolare i loadings Λ_{ij} e le ampiezze di errore σ_i (dette *uniquenesses* in quanto descrivono una componente di fluttuazione casuale *unica* delle singole X_i, cioè non comune a più X_i con-

giuntamente). Allora

$$\begin{aligned}
Cov\left(x_i, x_{i'}\right) &= Cov\left(\sum_{j=1}^{m} \Lambda_{ij} f_j + \sigma_i \varepsilon_i, \sum_{j'=1}^{m} \Lambda_{i'j'} f_{j'} + \sigma_{i'} \varepsilon_{i'}\right) \\
&= \sum_{j=1}^{m} \sum_{j'=1}^{m} \Lambda_{ij} \Lambda_{i'j'} Cov\left(f_j, f_{j'}\right) + \sum_{j=1}^{m} \Lambda_{ij} \sigma_{i'} Cov\left(f_j, \varepsilon_{i'}\right) \\
&+ \sum_{j'=1}^{m} \Lambda_{i'j'} \sigma_i Cov\left(\varepsilon_i, f_{j'}\right) + \sigma_i \sigma_{i'} Cov\left(\varepsilon_i, \varepsilon_{i'}\right) \\
&= \sum_{j=1}^{n} \Lambda_{ij} \Lambda_{i'j} + \sigma_i \sigma_{i'} \delta_{ii'}
\end{aligned}$$

cioè in termini matriciali

$$Q = \Lambda \Lambda^T + \Psi$$

dove Ψ è la matrice diagonale di componenti σ_i^2. Viceversa, si può dimostrare che se Q si può riscrivere in questa forma, allora vale il modello della factor analysis. Il problema quindi è di:

- prescrivere il numero m di fattori comuni con cui si desidera costruire il modello,
- cercare una matrice $n \times n$ diagonale positiva Ψ ed una matrice $m \times n$ di coefficienti Λ, tali che valga $Q = \Lambda \Lambda^T + \Psi$.

Osserviamo che se si accetta $m = n$, cioè si cerca A, $n \times n$, tale che $Q = AA^T$, basta prendere $A = \sqrt{Q}$. Nelle coordinate della base principale, essendo Q diagonale di componenti $\sigma_{X_i}^2$, basterà prendere A diagonale di componenti σ_{X_i}. Poi però va scritta A nella base canonica originaria, quindi servono le formule di cambio di base, che non riportiamo e che un SW eseguirebbe senza difficoltà. Nel caso invece, più utile, in cui m è scelto molto minore di n, non descriviamo in dettaglio la procedura iterativa adottata dal SW per il calcolo di loadings ed uniquenesses.

Osserviamo infine che la matrice Λ è definita a meno di *rotazioni* U nello spazio m-dimensionale dei fattori: se Λ risolve il problema, anche ΛU lo risolve, in quanto

$$\left(\Lambda U\right)\left(\Lambda U\right)^T = \Lambda U U^T \Lambda^T = \Lambda \Lambda^T$$

essendo UU^T pari all'identità. Quindi non c'è unicità dei fattori (per m fissato) ma essi possono essere cambiati con rotazioni. Queste rotazioni a volte sono abbastanza utili a livello interpretativo in qunato i fattori opportunamente ruotati possono apparire collegati in modo più naturale a certe delle variabili X_i. In altre parole, mentre i fattori trovati inizialmente dal SW hanno magari componenti significativamente non nulle con tutte le X_i e per questo hanno scarsa interpretabilità, opportunamente ruotati possono avere componenti significative solo con alcuni degli X_i e questo aiuta a darne un'interpretazione applicativa.

7 Confronto

I risultati di PCA e FA non coincidono. Se ad esempio si prendono le prime due componenti di PCA e si esegue una FA con due fattori, non vengono gli stessi vettori. Questo a prima vista sembra strano, se si pensa che PCA cerca i vettori che catturano maggiormente la variabilità dei dati. Però si deve ricordare che FA vuole resti indipendenti tra le diverse componenti, mentre PCA non ha questo requisito. Vediamo un esempio astratto, con media nulla. PCA produce la rappresentazione

$$X = \sum_{j=1}^{n} \sqrt{\lambda_j} Z_j v_j.$$

Isoliamo la prima componente principale:

$$X = Z_1 \sqrt{\lambda_1} v_1 + \sum_{j=2}^{n} \sqrt{\lambda_j} Z_j v_j.$$

Una FA con un solo fattore avrebbe prodotto un modello del tipo

$$X = f_1 \Lambda^{(1)} + \sigma\varepsilon$$

dove $\Lambda^{(1)}$ è il vettore della prima colonna di Λ, $\sigma\varepsilon$ è un vettore di gaussiane indipendenti. Oltre alla maggior semplicità del modello FA, si vede che l'altro può esserne ben lontano: pensiamo al caso $n = 2$. Il resto nel modello PCA è

$$r^{PCA} = \sqrt{\lambda_2} Z_2 v_2$$

mentre nel modello FA è

$$r^{FA} = \sigma\varepsilon.$$

$$Cov\left(r_1^{PCA}, r_2^{PCA}\right) = (v_2)_1 (v_2)_2 \lambda_2$$

può essere sensibilmente diversa da zero. Il metodo FA identifica vettori molto più allineati con la base canonica di quanto non faccia PCA.