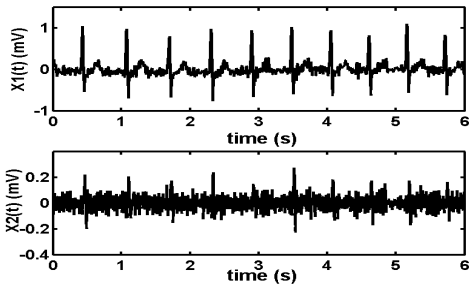
1 Tecniche di separazione delle sorgenti[1]

1.1 Introduzione

Nei metodi di **separazione delle sorgenti** si sfrutta la natura multicanale delle acquisizioni per estrarre caratteristiche comuni che viaggiano su canali diversi: ci si pone l'obiettivo di separare, da miscele di segnali ricevute simultaneamente da sensori distinti, i contributi di sorgenti diverse. In letteratura questo approccio è noto come **Blind Source Separation (BSS)**→ Separazione "cieca" delle sorgenti, in cui il termine "cieca" indica che sia le sorgenti, sia il processo di mescolamento a cui sono sottoposte, sono incogniti. Abbiamo disponibili soltanto le osservazioni, i segnali acquisiti dai sensori. Le tecniche BSS permettono di separare segnali sorgenti anche quando sono sovrapposti in frequenza. Ad esempio gli artefatti da movimento o da EMG nel segnale ECG sono i più difficili da rimuovere proprio perché presentano lo stesso contenuto frequenziale del segnale ECG.

Elenchiamo di seguito alcune delle applicazioni biomediche tra le più diffuse che impiegano l'approccio BSS:

- Rimozione di artefatti da segnali acquisiti simultaneamente ed aff etti da contaminazioni reciproche (vedi esempio in figura 1 relativo ad un segnale ECG ed un EMG)
- Estrazione di caratteristiche in un processo di classificazione.
- Separazione di attività atriale (onda P) e attività ventricolare (complesso QRS – onda T) da segnali ECG, per esempio per studi di fibrillazione atriale.
- Estrazione dell'ECG del feto da quello materno.
- Cancellazione o riduzione degli artefatti e del rumore nel segnale EEG.
- Miglioramento dei potenziali evocati (EP) nell'EEG e caratterizzazione dei segnali cerebrali (i potenziali cerebrali evocati da stimolazioni sensoriali di tipo visivo, acustico o somatosensoriale sono generalmente chiamati potenziali evocati).
- Applicazioni di Risonanza Magnetica funzionale: estrazione di serie temporali e mappe di connettività funzionale, che rappresentano l'andamento temporale e la distribuzione spaziale dell'attività cerebrale come misurata con immagini sensibili all'effetto BOLD (Fig. 2)
-



- Fig. 1

Un esempio di applicazione BSS consiste nel separare segnali audio emessi da diversi sorgenti (soggetti) e rivelati con un numero arbitrario di microfoni (Fig. 3). Le registrazioni consistono quindi in un mescolamento (*mixture*) di tutti i segnali audio. Tuttavia non sono noti né il modo in cui queste voci si siano combinate per dare origine ai segnali registrati al microfono, né le singole voci, che per questo motivo sono dette sorgenti latenti.

[1] A cura di Matteo Milanesi, rivisto (ma non necessariamente migliorato) da Nicola Vanello

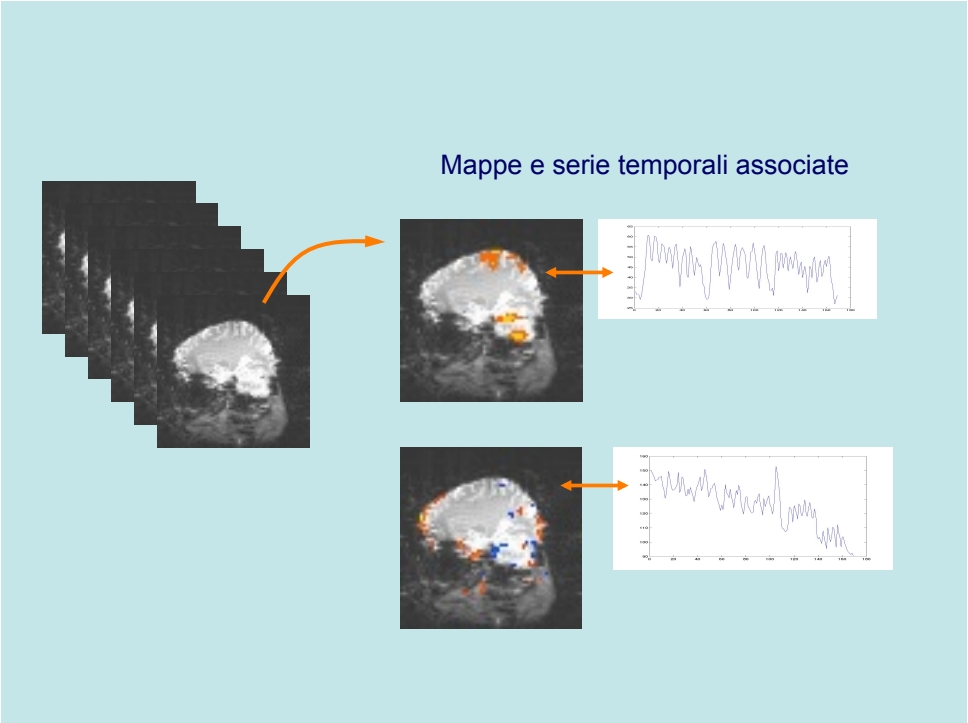


Fig. 2

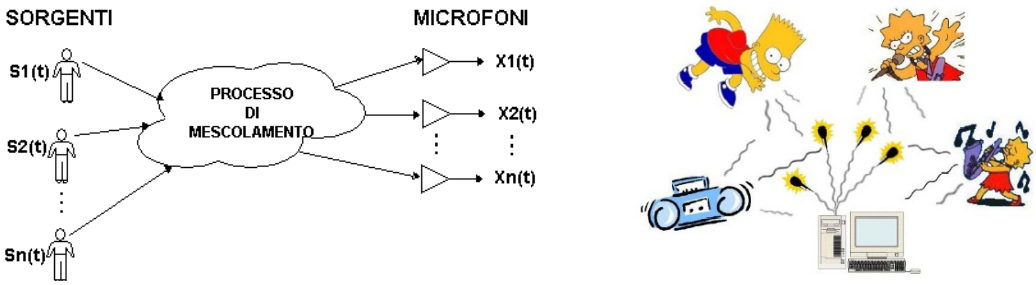


Fig. 3

I modelli di mixing possono essere suddivisi nelle seguenti classi:

- *Non lineare* → è il modello più generale: il fatto che i segnali non soddisfino la sovrapposizione degli effetti rende tali problemi di difficile trattazione.
- *Lineare convolutivo* → i segnali sorgente hanno ritardi differenti in ciascun segnale osservato a causa delle velocità finita di propagazione nel canale di comunicazione. Ciascun segnale osservato può anche contenere versioni ritardate dello stesso segnale. Da un punto di vista matematico si tratta di simulare un processo di filtraggio tra le sorgenti e gli elettrodi, rappresentato da una convoluzione con dei filtri a risposta impulsiva da determinare.
- *Lineare istantaneo* → i segnali ricevuti in un dato istante sono una combinazione lineare dei segnali sorgente nello stesso istante. Questo è il caso più semplice e più diffusamente studiato e anche quello che tratteremo nel seguito.

1.2 **Modello lineare istantaneo**

Supponiamo di disporre di n sorgenti indipendenti $(s_1, s_2, \cdots, s_n)$. In questa trattazione considereremo le variabili come funzione del tempo, ma più in generale possono essere funzione della posizione, come nel caso delle bioimmagini. *Il modello lineare istantaneo di mescolamento è dato da (Fig. 4):*

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + \cdots + a_{1n}s_n(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + \cdots + a_{2n}s_n(t)$$
$$\ldots$$
$$x_m(t) = a_{m1}s_1(t) + a_{m2}s_2(t) + \cdots + a_{mn}s_n(t)$$

dove i vettori $(x_1, x_2, \cdots, x_m)$ rappresentano le misure.

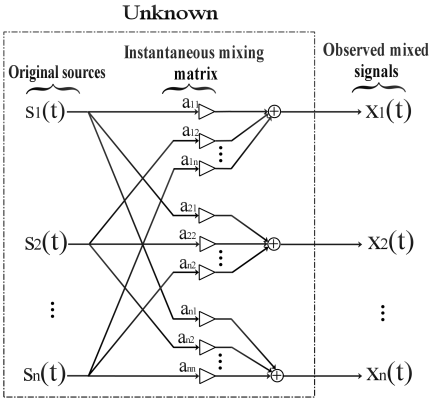


Fig. 4

In notazione matriciale:

$\boldsymbol{x(t)=As(t)}$

In generale il numero delle misure m è maggiore o uguale al numero delle sorgenti n, $m \geq n$, ma per semplicità di trattazione nel seguito assumiamo $m=n$.

Dunque l'obiettivo delle tecniche di separazione delle sorgenti (Fig. 5) è trovare una trasformazione lineare $\boldsymbol{W}$ tale che $\boldsymbol{y(t)=Wx(t)}$ sia una buona approssimazione delle sorgenti $\boldsymbol{s(t)}$ (nel caso m=n, matrice A quadrata, $\boldsymbol{W \approx A^{-1}}$).

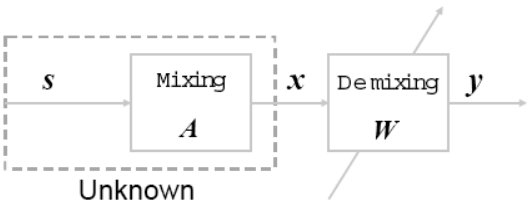


Fig. 5

Per separare le sorgenti useremo ***un approccio statistico multivariato*** in cui ogni campione temporale del segnale i-esimo $x_i(t)$ è considerato come una osservazione di una variabile aleatoria. Considerata questa assunzione, l'indice temporale t può essere trascurato perché ciò che interessa è la distribuzione della variabile, non l'ordine in cui si presentano i campioni. Quindi, x_{ik} rappresenta l'*i-esimo* segnale biomedico monodimensionale campionato (o osservato) all'istante *k-esimo.* Pertanto, il dato multivariato misurato può essere messo sotto forma di matrice X:

$$X = \begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{m1} & \cdots & \cdots & x_{mp} \end{pmatrix}$$

con m misure, ciascuna composta da p campionamenti o osservazioni. Ciascun misura x_{ik} è la combinazione lineare di n sorgenti istantanee e indipendenti al tempo k-esimo:

$$\mathbf{X = A\ S}$$

con:

$$S = \begin{pmatrix} s_{11} & \cdots & \cdots & s_{1p} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ s_{n1} & \cdots & \cdots & s_{np} \end{pmatrix}$$

Nel seguito tratteremo l'analisi delle componenti indipendenti (Independent Component Analysis o ICA) che si occupa di risolvere il modello appena visto.

Tale metodo fa parte dei metodi ***esplorativi*** o ***data-driven*** perché forniscono dei risultati basati su assunzioni generali riguardo alla formazione dei segnali come ad esempio la legge di distribuzione statistica delle variabili. Per contro i metodi ***hypothesis driven*** si basano su ipotesi più forti che richiedono di conoscere il modello probabilistico, non permettendo di trovare fenomeni inaspettati, ovvero non modellati a priori.

Un metodo analogo è l'analisi delle componenti principali (Principal Component Analysis o PCA) che ha l'obiettivo di ricavare le componenti incorrelate che possiedono la massima ***energia*** contenuta nelle miscele mentre l'***ICA*** permette di ricavare quelle che possiedono la massima ***informazione.***

La **PCA** fa uso di **statistiche del secondo ordine** per stimare le sorgenti originarie in quanto ricerca componenti tra loro incorrelate. D'altra parte l'**ICA,** che ricerca sorgenti statisticamente indipendenti, fa uso di **statistiche di ordine superiore,** poiché come vedremo tra breve l'indipendenza statistica è una condizione più forte dell'incorrelazione**.**

1.3 **Richiami di teoria della probabilità**

1.3.1 Vettore aleatorio e statistiche di ordine superiore

Per descrivere il comportamento statistico di una singola variabile aleatoria X_1 è necessario considerare la sua funzione distribuzione di probabilità:

$$F_{X1}(x_1) = \Pr\{X_1 < x_1\}$$

e la funzione densità di probabilità:

$$fx_1(x_1) = \frac{\partial F_{X1}(x_1)}{\partial x_1}$$

Dalla conoscenza della $f_{X1}(x_1)$ della singola variabile aleatoria X_1 è possibile ricavare le statistiche del primo ordine.

Ad esempio ricordiamo i momenti di ordine n della singola variabile aleatoria:

$$\alpha_i = E\{X_1^j\} = \int_{-\infty}^{+\infty} x_1^j f_{X1}(x_1) dx_1$$

dove la lettera *E* sta per expectation (aspettazione). I momenti centrali di ordine n sono descritti dalla seguente formulazione:

$$E\{(X_1 - \alpha_1)^j\} = \int_{-\infty}^{+\infty} (x_1 - m_x)^j f_{X1}(x_1) dx_1$$

I momenti centrali sono pertanto calcolati attorno al valor medio m_x che è uguale al momento di ordine 1.

Il momento di ordine due è $E\{X_1^2\}$, cioè la potenza media di X_1 mentre il momento centrale di ordine due è la varianza σ^2.

Altre statistiche del primo ordine molto usate sono le cumulanti:

$$\begin{aligned}
k1 &= E\{X_1\} \\
k2 &= E\{X_1^2\} - [E\{X_1\}]^2 \\
k3 &= E\{X_1^3\} - 3E\{X_1^2\}E\{X_1\} + 2[E\{X_1\}]^3 \\
k4 &= E\{X_1^4\} - 3[E\{X_1^2\}]^2 - 4E\{X_1^3\}E\{X_1\} + 12E\{X_1^2\}[E\{X_1\}]^2 - 6[E\{X_1\}]^4
\end{aligned}$$

e per variabili aleatorie a media nulla si semplificano in:

$$\begin{aligned}
k1 &= 0 \\
k2 &= E\{X_1^2\} \\
k3 &= E\{X_1^3\} \\
k4 &= E\{X_1^4\} - 3[E\{X_1^2\}]^2
\end{aligned}$$

e quindi le prime tre cumulanti coincidono con i primi tre momenti.

Nel caso multivariato, cioè nello studio di un sistema di n variabili aleatorie, si considera una notazione più compatta. Le n variabili aleatorie vengono disposte in un vettore *n-dimensionale* $\boldsymbol{X}$ detto vettore aleatorio:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = (X_1 \quad X_2 \quad \cdots \quad X_n)^T$$

La caratterizzazione statistica completa del vettore aleatorio richiede la conoscenza completa della classe di distribuzioni di probabilità congiunta di ordine n:

$$F_X(X) = F_{X1,X2,..,Xn}(x_1, x_2, ..., x_n) = \Pr\{X_1 < x_1, X_2 < x_2, ..., X_n < x_n\}$$
$$f_{x_1,x_2,...x_n}(x_1, x_2, ..., x_n) = \frac{\partial^n F_{X_1X_2,..,Xn}(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}$$

Se vogliamo quindi studiare il comportamento congiunto di due variabili aleatorie è necessario conoscere la funzione distribuzione di probabilità congiunta di ordine 2, da cui si ricava la rispettiva densità di probabilità:

$$F_{X_1,X_2}(x_1, x_2) = \Pr\{X_1 < x_1, X_2 < x_2\} \qquad f_{X_1,X_2}(x_1, x_2) = \frac{\partial^2 F_X(x)}{\partial x_1 \partial x_2}$$

Dalla funzione densità di probabilità congiunta di due variabili aleatorie, si possono ricavare le statistiche del secondo ordine come ad esempio la correlazione:

$$r_{X1,X2} = E\{(X_1X_2)\} = \int_{x1=-\infty}^{+\infty} \int_{x2=-\infty}^{+\infty} x_1 x_2 f_{X1X2}(x_1, x_2) dx_1 dx_2 .$$

Essa rappresenta la cumulante di ordine due per un vettore aleatorio a media nulla. Nel caso di n > 2 si parla di **statistiche di ordine superiore (Higher Order Statistics o HOS)**

Esempi in questo senso sono i momenti e le cumulanti del terzo e quarto ordine (nel caso multivariato si usa spesso il termine cross-cumulanti) dette appunto statistiche del terzo e quarto ordine:

$$K(X_{i,}X_j, X_k) = E\{X_i X_j X_k\}$$
$$K(X_i, X_j, X_k, X_l) = E\{X_i X_j X_k X_k\} - E\{X_i X_j\}E\{X_k X_l\} - E\{X_j X_k\}E\{X_j X_l\} - E\{X_i X_l\}E\{X_j X_k\}$$

Uno svantaggio nell'uso di statistiche di ordine superiore è che esse richiedono molti più campioni per essere stimate rispetto alle statistiche del secondo ordine.

Sia i momenti che le cumulanti portano lo stesso contenuto informativo da un punto di vista statistico, poiché le cumulanti possono essere espresse come somme di prodotti di momenti. E' tuttavia preferibile lavorare con le cumulanti perché presentano in modo più chiaro le informazioni addizionali fornite dalle statistiche di ordine superiore. Le cumulanti mostrano anche delle proprietà che non condividono con i momenti. Infatti:

- la cumulante della somma di vettori aleatori statisticamente indipendenti, è uguale alla somma delle rispettivi cumulanti
- se la distribuzione di un vettore aleatorio è gaussiana multivariata, tutti le sue cumulanti del terzo ordine e superiore sono nulle

Le stesse proprietà valgono per le cumulanti calcolate sulla singola variabile aleatoria, anziché sul vettore. Quindi le cumulanti di ordine superiore misurano quanto un vettore aleatorio (o una singola variabile) sia distante da un vettore aleatorio con densità di probabilità gaussiana.

1.3.2 Incorrelazione e indipendenza

Sappiamo che due variabili aleatorie x e y (o due variabili appartenti allo stesso vettore aleatorio **x**) si dicono *incorrelate* se la loro covarianza è nulla.

$$c_{xy} = E\{(x - m_x)(y - m_y)\} = 0$$

dove m_x e m_y indicano i rispettivi valori medi.

Poiché la correlazione è definita come $r_{xy} = E\{xy\}$, la covarianza può essere scritta come: $c_{xy}=r_{xy}-m_x m_y$. Quando le due variabili aleatorie sono incorrelate si ha, $r_{xy}=m_x m_y$.

Per vettori di variabili aleatorie $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)^T$ si ricava la matrice di covarianza: quando il valor medio è nullo, correlazione e covarianza coincidono e si è soliti scrivere la matrice di correlazione come $C_x = E\{\mathrm{x}\mathrm{x}^T\}$.

Se prendiamo in considerazione un singolo vettore $\boldsymbol{x}$ di variabili aleatorie tra loro mutuamente incorrelate, la matrice di covarianza C_x diventa diagonale; sulla diagonale compariranno i valori della varianza delle singole variabili. Un particolare vettore aleatorio a media nulla e matrice di covarianza unitaria è detto *bianco*.

L'indipendenza tra due variabili è invece definita nel modo seguente:

$$f_{x,y}(x,y) = f_x(x) f_y(y)$$

cioè la distribuzione congiunta $f_{xy}(x,y)$ di x e y si scompone nel prodotto tra le densità marginali $f_x(x)$ e $f_y(y)$. Questo implica che la conoscenza sul valore di una delle variabili non ci dice nulla sui possibili valori assunti dall'altra. L'indipendenza statistica implica anche la seguente proprietà:

$$E\{g(x)h(y)\} = E\{g(x)\}E\{h(y)\}$$

dove $g(x)$ e $h(y)$ sono due qualsiasi funzioni assolutamente integrabili di x e di y. Da qui si vede che l'indipendenza statistica è più forte dell'indipendenza lineare che si ha per e . Quindi variabili indipendenti sono incorrelate ma non è necessariamente vero il contrario.

Per esempio, la precedente fornisce il mezzo per calcolare il momento del secondo ordine come prodotto dei rispettivi valori medi.

$$\mathrm{E[XY] = E[X]E[Y] = \int_{-\infty}^{\infty} x f_X(x)dx \int_{-\infty}^{\infty} y f_Y(y)dy = \eta_X \eta_Y}$$

Se le variabili aleatorie hanno una distribuzione gaussiana, l'incorrelazione tra di esse assicura anche l'indipendenza.

(a) (b)

Fig. 6

In Fig. 6a sono riportati gli scatter plots di due variabili distribuite in modo uniforme e indipendenti. Si ricorda che due sorgenti sono indipendenti quando il valore assunto da una non fornisce alcuna informazione su quello assunto dall'altra e ciò è riscontrabile anche dalla forma della loro densità di probabilità congiunta. Intuitivamente si può affermare che, fissato un valore di una delle due variabili, ogni valore dell'altra variabile è possibile. In (b) i campioni sono incorrelati, ma non indipendenti e quindi, fissato un valore di una delle due variabili, solo un numero limitato di valori dell'altra variabile è possibile; in particolare un punto sul vertice del quadrilatero determina univocamente i valori che devono assumere le due variabili.

Si richiama inoltre il **teorema del limite centrale** secondo il quale la distribuzione della somma di variabili aleatorie indipendenti e identicamamente distribuite tende ad una gaussiana.

Fig. 7

In Fig. 7 è riportato lo scatter plot di due variabili distribuite in modo uniforme ed indipendente (sinistra) e l'istogramma dei valori assunti da una delle due variabili con confronto con pdf gaussiana (linea tratteggiata) **N**(0, 1) (destra).

Fig. 8

In Fig. 8 si vede come la distribuzione sotto forma di istogramma della combinazione lineare dei campioni relativi alle due variabili distribuite in modo uniforme ed indipendenti tenda maggiormente ad una pdf gaussiana, se confrontato con l'istogramma (uniforme) di partenza.

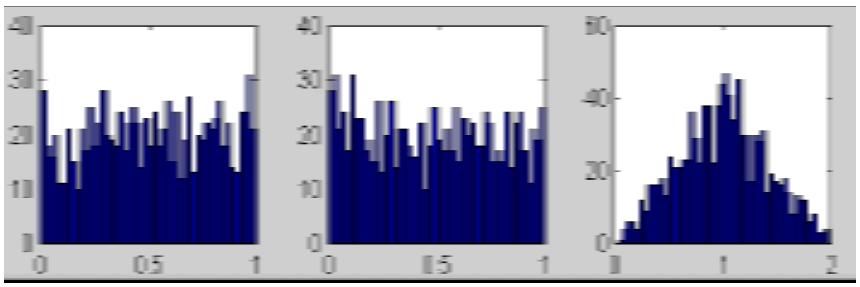


Fig. 9

In Fig. 9 è riportata una simulazione al computer per conferma di quanto sopra esposto.

1.4 **Analisi delle Componenti Principali**

L'analisi delle componenti principali (PCA) è un metodo di analisi multivariata che permette di determinare nuove variabili come combinazione lineare di quelle di partenza, ordinate secondo valori decrescenti di varianza e incorrelate tra di loro.

Tale analisi viene anche indicata con trasformata di Karhunen-Loeve o di Hotelling.

Lo scopo dell'analisi delle componenti principali è quello di ridurre la dimensionalità dei dati, studiare la correlazione tra le variabili e classificare i dati.

Se consideriamo il vettore di variabili di partenza $\boldsymbol{x}=[x_1, x_2, \ldots, x_n]$ la PCA trova un nuovo set di variabili $\boldsymbol{y}=[y_1, y_2, \ldots, y_n]$ legate alle precedenti secondo un modello lineare.

Quindi è possibile esprimere le singole variabili come

- $y_i = a_{1i}x_1 + a_{2i}s_2 + \ldots + a_{ni}x_n$
- $\mathrm{Cov}(y_i, y_j)=\delta_{ij}$
- $\mathrm{var}(y_1) > \mathrm{var}(y_2) > \mathrm{var}(y_3) > \ldots > \mathrm{var}(y_n)$

La stima dei coefficienti a_{ij} è ottenuta dall'analisi della matrice di covarianza dei dati

$$\Sigma = E\left[\left(x-\mu\right)\left(x-\mu\right)^T\right]$$

dove $\boldsymbol{\mu}$ è il vettore delle medie di $\boldsymbol{x}$.

Nell'impossibilità di conoscere la matrice di covarianza delle variabili nel vettore $\boldsymbol{x}$, questa viene stimata dalla matrice di covarianza campionaria $\hat{\Sigma} = \mathbf{X}\mathbf{X}^T$

dove la matrice $\mathbf{X}$ è la matrice che contiene tutte le p osservazioni delle n-variabili

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

dove con x_{hk} si indica la k-esima osservazionne della h-esima variabile.

1.4.1 Interpretazione geometrica delle PCA

L'analisi delle componenti principali si presta ad un'interpretazione geometrica.

Infatti le variabili che costituiscono il vettore $\boldsymbol{x}$ possono essere viste come coordinate di uno spazio n-dimensionale.

La k-esima osservazione dei valori delle variabili può essere vista come un punto in questo spazio dimensionale.

La determinazione delle PCA, può essere vista come la determinazione di una nuova base rispetto alla quale è possibile descrivere i dati osservati.

I vettori della nuova base, descritti nello spazio di partenza saranno forniti dagli autovettori della matrice di covarianza

$$\mathbf{u}_j = \left\{u_{1j}, u_{2j}, \ldots, u_{nj}\right\}$$

tali per cui la matrice di covarianza può essere scritta come

$$\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^T$$

dove le colonne di $\boldsymbol{U}$ sono date dati vettori $\boldsymbol{u}_j$, che rappresentano una base ortonormale dello spazio delle variabili. Infatti $\mathbf{U}^T\mathbf{U} = I$.

La matrice S^2 rappresenta una matrice quadrata nxn diagonale. I valori sulla diagonale rappresentano gli autovalori λ_i della matrice di covarianza ovvero le varianze delle componenti principali.

La trasformazione in oggetto è la seguente

$Y = \mathbf{U}^T X$

con $\mathbf{Y}$ matrice delle componenti principali.

Visto che $\mathbf{U}^T\mathbf{U} = I$ possiamo scrivere che $\mathbf{U}Y = \mathbf{U}\mathbf{U}^T X = X$, ottenendo quindi la trasformazione inversa, per la quale la *i*-esima variabile di partenza può scriversi come.

$$x_i = \sum_{k=1}^{r} u_{ik} y_k$$

1.4.2 Spettro degli autovalori

É possibile dimostrare che la somma delle varianze delle variabili di partenza è uguale alla somma delle varianze delle component principali. Quindi la grandezza $\frac{\lambda_i}{\sum_{i=1}^{r}\lambda_i}$ rappresenta la percentuale di varianza totale spiegata dalla *i*-esima variabile.

In generale gli autovettori associati ad autovalori non nulli, saranno $r \leq n$, dove con *r* si indica il rango della matrice di covarianza. Quindi *n-r* autovettori della matrice di covarianza saranno autovettori dello spazio nullo. In questo caso, alcune variabili di partenza saranno linearmente dipendenti.

Il caso di un autovalore piccolo, si può interpretare come il caso di una componente principale che spiega una piccolo porzione della varianza dei dati. La decisione riguardo all'interpretazione di tale componente varia caso per caso, così come l'opportunità di eliminare tale componente dall'analisi successiva. Infatti non sempre minore varianza comporta una minore importanza di una componente.

1.4.3 Riduzione dei dati

Spesso, l'analisi delle component principali viene utilizzata per ridurre la dimensionalità del problema. Si utilizzano infatti non più le *n* variabili di partenza, ma le prime componenti principali. Solitamente si possono scegliere quelle che rappresentano la maggior parte della varianza nei dati. La scelta del numero di componenti comunque non è un passaggio ovvio, e per rispondere a tale quesito è possibile fare ricorso a criteri della teoria dell'informazione come il criterio di Akaike.

1.4.4 Rappresentazione dei dati

La determinazione della nuova base, permette di descrivere le osservazioni in un nuovo spazio. In questo spazio potrebbe essere più semplice caratterizzare e classificare le osservazioni. Tale risultato potrebbe essere ottenuto andando a evidenziare eventuali raggruppamenti in questo spazio o in generale studiare la distribuzione delle osservazioni. Nel caso di studi su una popolazione, dove le osservazioni sono gli individui, questa rappresentazione potrebbe facilitare la classificazione degli individui. Nel caso in cui le variabili siano serie temporali, e quindi le osservazioni siano campioni nel tempo di diversi segnali, questa analisi potrebbe permettere la classificazione di diversi istanti temporali.

1.4.5 Cerchio delle correlazioni.

Una tipica applicazione della PCA è lo studio delle correlazioni tra le variabili di partenza e le nuove componenti. In particolare si può creare una rappresentazione sul piano delle componenti principali delle variabili di partenza in termini di coefficiente di correlazione.

La variabile di partenza j-esima, sarà quindi rappresentata nel piano delle componenti h e k utilizzando come coordinate il coefficiente di correlazione tra x_j e y_h e tra x_j e y_k rispettivamente.

Il coefficiente di correlazione tra la j-esima variabile di partenza x_j e la h-esima componente principale può essere calcolato come

$$\rho_{xj,yh} = \frac{\operatorname{cov}\left(x_j, y_k\right)}{\sigma_{x_j}\sigma_{y_h}} = \frac{\operatorname{cov}\left(\sum_{k=1}^{r} u_{jk} yk, y_h\right)}{\sigma_{x_j}\sigma_{y_h}}$$

vista l'incorrelazione tra le PC, si ha

$$\frac{\operatorname{cov}\left(\sum_{k=1}^{r} u_{jk} y_k, y_h\right)}{\sigma_{x_j}\sigma_{y_h}} = \frac{u_{jh}\sigma^2_{y_h}}{\sigma_{x_j}\sigma_{y_h}} = \frac{u_{jh}\sigma_{y_h}}{\sigma_{x_j}}$$

In questo modo è possibile:

- verificare la relazione tra le componenti principali e le variabili di partenza. Questo può essere utile per capire il significato delle componenti principali
- classificare le variabili di partenza. Ad esempio le variabili di partenza che correlano maggiormente con le prime componenti principali, saranno le più importanti, in termini di varianza spiegata. Le variabili che correlano di meno con le prime e maggiormente con le ultime, saranno invece le meno importanti, e in alcune applicazioni potrebbero essere eliminate.
- comprendere eventuali ridondanze delle variabili di partenza, ad esempio nel caso due o più variabili di partenza si vadano a collocare in regioni molto vicine del cerchio delle correlazioni.

1.5 **Analisi delle componenti indipendenti**

1.5.1 Ambiguità del modello

Si prenda in considerazione il seguente modello di mescolamento delle componenti indipendenti: ***x=As.*** Nel modello, sono presenti alcune ambiguità che non possono essere eliminate.

- La prima ambiguità riguarda l'impossibilità di determinare le varianze, e quindi le energie, delle singole componenti indipendenti. Essendo sia il vettore delle sorgenti ***s*** che A incogniti, non siamo in grado di stabilire in quale misura l'ampiezza di **x** sia da attribuire alla sorgente **s** oppure al canale di trasmissione rappresentato in **A**. Infatti, dato uno scalare α_i e considerata la sorgente s si può scrivere:

$$x_{ik} = \sum_{j=1}^{m} a_{ij} s_{jk} = \sum_{j=1}^{m} \alpha a_{ij} \frac{s_{jk}}{\alpha}$$

  Quindi, durante il processo di stima delle componenti indipendenti si assume che le componenti indipendenti abbiano tutte varianza unitaria: $E\{s_i^2\} = 1$. I metodi utilizzati per risolvere il modello dovranno tener conto di ciò riadattando la stima della matrice A. Resta comunque irrisolta l'ambiguità sul segno poiché una componente indipendente può essere moltiplicata per (-1) senza che il modello venga in qualche modo modificato.

- Il secondo problema riguarda l'impossibilità di determinare l'ordine delle componenti indipendenti. Matematicamente si può scrivere che: $\mathrm{x} = AP^{-1}P\mathrm{s}$, dove P è una matrice di permutazione. Gli elementi di $P\mathrm{s}$ sono le variabili originali s_j, ma in un ordine diverso. La matrice AP^{-1} è la nuova matrice di mescolamento che deve essere ricavata.

Senza perdere in generalità possiamo assumere che le variabili osservate e le componenti indipendenti siano a media nulla. Se nella realtà tale ipotesi non è verificata possiamo *centrare* i dati di cui disponiamo sottraendovi i rispettivi valori medi prima di effettuare la ICA, $\mathrm{x} = \mathrm{x} - E\{\mathrm{x}\}$. Allora anche le componenti indipendenti stimate saranno a media nulla poiché $E\{\mathbf{s}\} = E\{\mathbf{A^{-1}x}\} = \mathbf{A^{-1}}E\{\mathbf{x}\}$. Una volta ricavata la matrice di separazione (*unmixing*) W, come stima di A^{-1}, i valori medi sottratti possono essere ripristinati sommando alle componenti indipendenti il termine $A^{-1}E\{\mathrm{x}\}$: $\mathbf{s' = s + A^{-1}}E\{\mathbf{x'}\}$

1.5.2 ICA e sbiancamento

Abbiamo già avuto modo di definire, a proposito della PCA, l'operazione di sbiancamento di un vettore di n variabili aleatorie x, tramite l'applicazione di una matrice ortogonale ai dati stessi. L'operazione di sbiancamento è rappresentata da una trasformazione lineare $\mathbf{z = Vx}$, dove $V = \Sigma^{-1/2}U^T$, U è una matrice ortogonale e Σ è una matrice di normalizzazione.

E' possibile giungere allo stesso risultato partendo dalla decomposizione a valori singolari della matrice di covarianza C_x dei dati:

$$C_x = E\{\mathbf{x}\mathbf{x}^T\} = U\Sigma U^T$$

dove $U = (u_1, \ldots, u_n)$ è la matrice ortogonale di autovettori che soddisfa la proprietà $U^T U = UU^T = I$, e $\Sigma = diag(d_1, \ldots, d_n)$ è la matrice diagonale dei suoi autovalori. La matrice di sbiancamento, normalizzata , risulta essere:

$$V = U\Sigma^{-1/2}U^T$$

Infatti: $E\{\mathrm{z}\mathrm{z}^T\} = VE\{\mathrm{x}\mathrm{x}^T\}V^T = VC_xV^T = VU\Sigma U^T V^T$. Mediante sostituzione del termine V, si ottiene:

$$E\{\mathrm{z}\mathrm{z}^T\} = U\Sigma^{-1/2}U^T U\Sigma U^T U\Sigma^{-1/2}U^T = U\Sigma^{-1/2}\Sigma\Sigma^{-1/2}U^T = I$$

perciò *z* è bianco.

L'operazione di sbiancamento fornisce una nuova matrice di mixing $\tilde{A}$ tale che:

$$\mathrm{z} = VA\mathrm{s} = \tilde{A}\mathrm{s}$$

Questa operazione ovviamente non è sufficiente da sola a risolvere il problema della separazione delle sorgenti indipendenti perché l'incorrelazione ottenuta con lo sbiancamento è una proprietà meno forte dell'indipendenza.

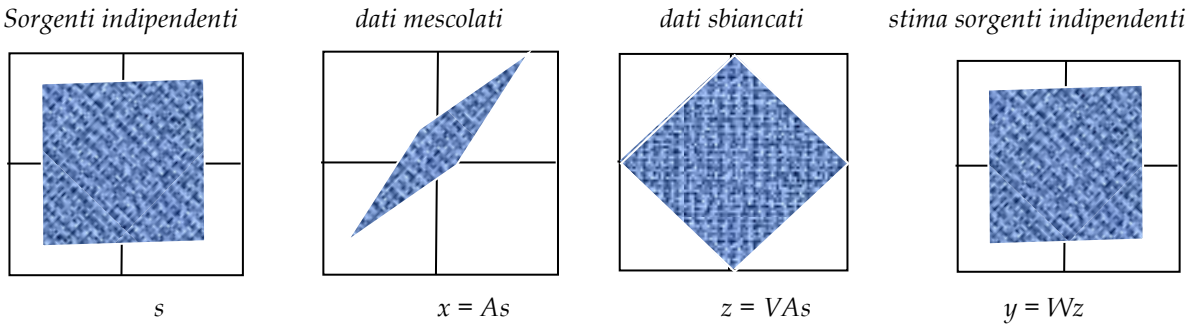


Fig. 10

Com'è descritto in Fg. 10, l'operazione di sbiancamento è il primo passo nella ricerca delle componenti indipendenti: dopo l'operazione di sbiancamento non c'è correlazione, ma dipendenza; l'ICA si riduce all'operazione di rotazione tramite una trasformazione ortogonale dalla quale è possibile stimare le sorgenti indipendenti.

Pertanto, il processo di sbiancamento è semplicemente una variazione lineare di coordinate dei dati mescolati. Dopo aver trovato la soluzione ICA nel sistema di coordinate sbiancate, è facile riproiettare la soluzione ICA sul sistema originale.

L'operazione di sbiancamento risulta comunque molto utile dal momento che la nuova matrice di mescolamento $\widetilde{A}$ è ortogonale. Infatti, dal momento che s è il vettore aleatorio delle sorgenti statisticamente indipendenti, e quindi anche incorrelate, si può scrivere che:

$$E\left\{ \mathrm{zz}^T \right\} = \widetilde{A} E\left\{ \mathrm{ss}^T \right\} \widetilde{A}^T = \widetilde{A} I \widetilde{A}^T = I$$

Quindi l'operazione di sbiancamento si ottiene la riduzione della dimensionalità del problema. Infatti una matrice ortogonale possiede $n(n-1)/2$ gradi di libertà anziché n^2 e quindi $n(n+1)/2$ equazioni indipendenti.

Una volta effettuata l'operazione di sbiancamento è necessario utilizzare ulteriori metodi per stimare le componenti indipendenti.

1.5.3 ICA e PCA

Ambedue i metodi cercano una nuova base per i dati, ma mentre la PCA massimizza la varianza e quindi le proiezioni sulla nuova base non sono altro che le osservazioni mescolate, la ICA cerca la base sulla quale le proiezioni sono indipendenti.

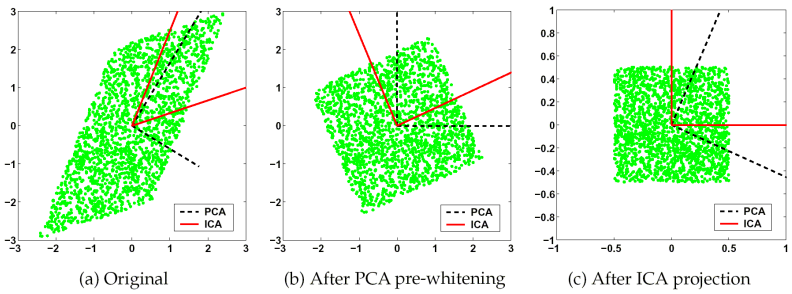


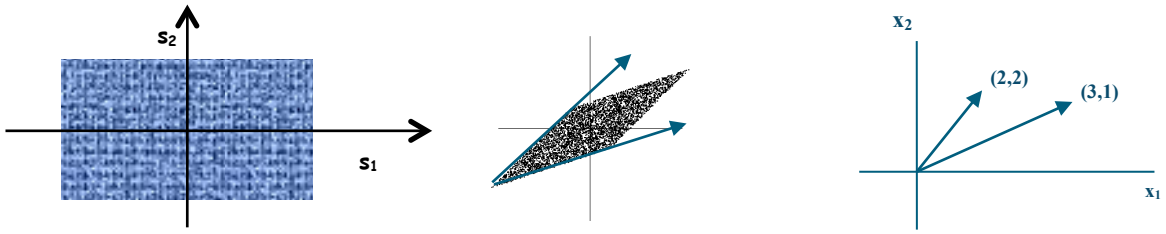
Fig. 11

In Fig. 11 sono riportati esempi di estrazione di PCA e ICA in diverse condizioni di distribuzioni di dati. Come si può osservare, la prima componente principale è situata nella direzione di massima varianza e la seconda componente è perpendicolare alla prima.

D'altra parte le componenti indipendenti non necessariamente sono ortogonali, in quanto sono situate sulle direzioni di massima indipendenza di dati non gaussiani.

1.5.4 ICA e variabili gaussiane

Consideriamo due sorgenti statisticamente indipendenti s_1 e s_2 con funzioni densità di probabilità marginali uniformi (Fig. 12a) e due variabili x_1 e x_2 derivanti dal loro mixing attraverso la matrice A (Fig. 12b). Le variabili originali si trasformano nelle variabili x_1 e x_2, perdendo quindi l'indipendenza statistica. Supponiamo che la matrice A assuma la seguente espressione: $A = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}$.

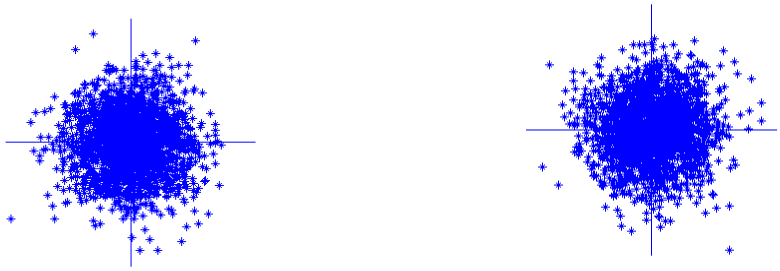


(a) (b) (c)

Fig. 9

Le direzioni dei segmenti che delimitano i due lati del parallelogramma contengono informazioni sulle colonne della matrice di mixing incognita A: i lati del parallelogramma sono proprio le direzioni delle colonne dei vettori di A. Il procedimento non è generalizzabile a variabili con densità di probabilità non uniformi e dal punto di vista computazionale è difficilmente realizzabile.

Le componenti indipendenti nel modello ICA devono essere non gaussiane, dato che dopo una trasformazione ortogonale le variabili gaussiane (Fig. 13a) rimangono tali (Fig. 13b). Infatti, dopo l'applicazione di una trasformazione ortogonale si ottiene una distribuzione con simmetria rotazionale in modo che non si possono ricavare informazioni sulle direzioni delle colonne della matrice A.



(a) (b)

Fig. 13

Prendiamo due componenti indipendenti s_1 e s_2 la cui densità di probabilità congiunta sia gaussiana:

$$f(s_1, s_2) = \frac{1}{2\pi} \exp\left(-\frac{s_1^2 + s_2^2}{2} \right) = \frac{1}{2\pi} \exp\left(-\frac{\|s\|^2}{2} \right)$$

Applichiamo un mescolamento tramite una matrice $\widetilde{A}$ ortogonale e otteniamo $\mathbf{x} = \widetilde{A}\,\mathbf{s}$. Ricordiamo che data una trasformazione lineare $x = As$, la densità di probabilità è: $f_x(x) = \frac{1}{|\det A|} f_s\left(A^{-1}x\right)$ e grazie alla proprietà $\widetilde{A}^{-1} = \widetilde{A}^T$, si ricavala densità congiunta dei dati x_1 e x_2:

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{\left\|\widetilde{A}^T x\right\|^2}{2} \right) \frac{1}{\left|\det \widetilde{A}\right|}$$

Grazie all'ortogonalità di $\widetilde{A}$, abbiamo che $\left\|\widetilde{A}^T x\right\|^2 = \|x\|^2$ e che $\left|\det \widetilde{A}\right| = 1$. Perciò la densità congiunta diventa:

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{\|x\|^2}{2}\right)$$

La trasformazione ortogonale non ha cambiato la densità di probabilità, la quale è ancora gaussiana.

Il motivo per cui non si possa stimare una matrice di unmixing per variabili gaussiane è collegato al fatto che variabili aleatorie incorrelate a densità di probabilità congiunta gaussiana sono necessariamente indipendenti. Quindi, in questo caso, l'informazione sull'indipendenza non porta oltre lo sbiancamento dei dati.

Se alcune delle componenti che vogliamo stimare sono gaussiane ed altre non gaussiane, possiamo stimare soltanto le componenti non gaussiane mentre quelle gaussiane non possono essere separate le une dalle altre. In altre parole, alcune componenti dopo l'applicazione di algoritmi di separazione resteranno combinazioni lineare arbitrarie di componenti gaussiane.

1.5.5 Massimizzazione della nongaussianità

Nel paragrafo precedente è stato mostrato come non sia possibile stimare le componenti indipendenti se queste presentano una densità di probabilità gaussiana. Si può quindi pensare a tecniche di ricerca di variabili meno gaussiane o, in altre parole, che massimizzino la non-gaussianità.

Assumiamo per semplicità che le sorgenti $s_{jk}, j = (1,....n)$ siano uniformemente distribuite. Si denoti con y_k una combinazione lineare delle misure x_{ik} all'istante k-esimo, $k = (1,....p)$

$$y_k = \sum_{i=1}^{m} b_i x_{ik} = b^T \mathrm{x} = b^T A s_k = q^T s_k = \sum_{i=1}^{m} q_i s_{ik}$$

Il termine $y_k = b^T A s_k$ è una combinazione lineare delle misure, ciascuna delle quali è combinazione lineare delle sorgenti indipendenti. E' stato già accennato che, sulla base del ***teorema del limite centrale,*** tale somma è più gaussiana delle componenti originarie.

Tornando quindi al modello ICA, è chiaro che un mescolamento di sorgenti indipendenti del tipo: $x_{ik} = \sum_{j=1}^{m} a_{ij} s_{jk}$ di componenti indipendenti, sarà più prossimo ad una gaussiana di quanto non lo siano le componenti segnali di partenza. Esso diventa meno gaussiano quando y_k uguaglia una delle componenti indipendenti, ossia quando $b^T A$ ha un solo elemento non nullo. Possiamo quindi variare b e osservare la distribuzione di y_j muovendoci nella direzione che massimizza la nongaussianità di $b^T \mathrm{x}$.

Più in generale, la non gaussianità delle y è massima quando la matrice di unmixing è l'inversa di $b^T A$.

1.5.5.1 Misura della nongaussianità tramite Kurtosis

Per poter utilizzare i concetti precedentemente discussi è necessario uno stimatore della nongaussianità. Utilizziamo la cumulante del quarto ordine, introdotta nel paragrafo relativo alle statistiche di ordine superiore, che, per una variabile aleatoria x, è detta *kurtosis* e definita come:

$$kurt(x) = E\{x^4\} - 3\left(E\{x^2\}\right)^2 - 4E\{x^3\}E\{x\} + 12E\{x^3\}\left(E\{x\}\right)^2 - 6\left(E\{x\}\right)^4$$

e nel caso in cui x sia a media nulla, cioè per $E\{x\} = 0$ diviene

$$kurt(x) = E\{x^4\} - 3\left(E\{x^2\}\right)^2$$

Se lavoriamo con dati a varianza unitaria $E\{x^2\} = 1$, l'espressione si semplifica ulteriormente e diventa una versione normalizzata del momento del quarto ordine.

E' interessante notare che per variabili gaussiane il momento del quarto ordine uguaglia proprio $3\left(E\left\{ x^2 \right\}\right)^2$ e quindi la kurtosis è nulla; mentre per quasi tutte le altre variabili la kurtosis è diversa da zero. In particolare, variabili che hanno una kurtosis positiva sono dette supergaussiane e quelle a kurtosis negativa (es. distribuzione uniforme) subgaussiane (Fig. 14).

Massimizzare la nongaussianità significa quindi massimizzare il valore della kurtosis: $\max\left|kurt(x)\right|$.

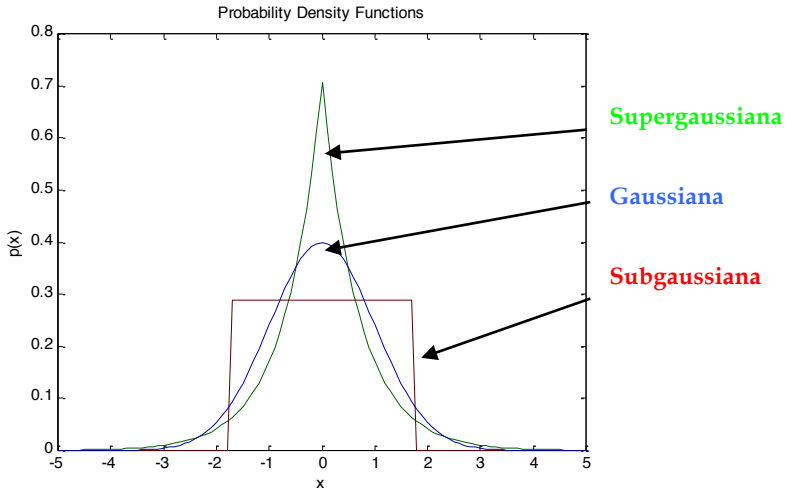


Fig. 14

In generale si può dimostrare che il valore della kurtosis è massimo quando $y = b^T x$ uguaglia una delle componenti indipendenti.

La kurtosis è semplice da calcolare ma ha il difetto di essere sensibile agli outlier, cioè a valori delle osservazioni, probabilmente dovuti a rumore, che si discostano molto dal valore degli altri dati. Questi possono essere identificato come valori che si posizionano sulle code della distribuzione di probabilità dei dati e possono influenzare pesantemente il valore della kurtosis.

1.5.5.2 *Esempio di applicazione al caso bidimensionale*

Facciamo un esempio di come utilizzare la kurtosis per stimare il modello ICA. Supponiamo di essere nel caso bidimensionale, noi osserviamo due grandezze x_1 e x_2 che modelleremo come due variabili aleatorie ottenute tramite un processo di mixing da altre due variabili aleatorie s_1 e s_2 che rappresentano le componenti indipendenti. Quindi $\boldsymbol{x = As}$ dove $x = \{x_1, x_2\}$ e $s = \{s_1, s_2\}$.

Si fanno le ipotesi che $\boldsymbol{kurt}(s_1), \boldsymbol{kurt}(s_2) \neq 0$ e le varianze di s_1 e s_2 siano uguali a 1.

Il problema consiste nel trovare una $\boldsymbol{y = b^T x}$ che rappresenti una stima di una delle componenti indipendenti. Dobbiamo trovare il vettore che massimizza il valore assoluto della kurtosis di y.

Dal modello abbiamo che $\boldsymbol{y = b^T x = b^T As = q^T s = q_1 s + q_2 s_2}$

La kurtosis, essendo una cumulante, possiede le seguenti proprietà, date due variabili aleatorie x_1 e x_2 e uno scalare a,

kurt(x_1+x_2) = kurt(x_1)+kurt(x_2)
kurt(a x_1) = a^4 kurt(x_1)

Quindi, la kurtosis di y è data da $\boldsymbol{kurt}(y) = \boldsymbol{kurt}(q_1 s_1) + \boldsymbol{kurt}(q_2 s_2) = q_1^4 \boldsymbol{kurt}(s_1) + q_2^4 \boldsymbol{kurt}(s_2)$

Visto che le componenti indipendenti sono state assunte a varianza unitaria, possiamo fare la stessa ipotesi per y, che ne è una stima quindi avremo di conseguenza un vicolo su q, infatti:

$$E\left\{y^2\right\} = q_1^2 + q_2^2 = 1$$

Questo implica che il vettore q è vincolato al cerchio di raggio unitario in un piano bidimensionale. Se per semplicità si assumono le kurtosis delle componenti indipendenti uguali a uno, il problema si riduce a

trovare il massimo della funzione $|kurt(y)| = |q_1^4 + q_2^4|$ al variare del vettore q sul cerchio di raggio unitario. È facile dimostrare che il massimo coincide con i punti per cui uno solo degli elementi di q è diverso da zero e in particolare è uguale a +1 o –1, e quindi la y coincide, a meno del segno, con una delle componenti indipendenti. Nel caso in cui le kurtosis abbiano valori arbitrari si può dimostrare che il risultato è sempre valido.

Se lavoriamo con dei dati sbiancati $z = Vx$ cercheremo un vettore w tale che la non gaussianità di $w^T z$ sia massima. Il vincolo posto su q si traduce nel medesimo vincolo su w infatti, visto che $y = w^T VAs$ per cui $q = (VA)^T w$, abbiamo

$$\|q\|^2 = (w^T VA)(A^T V^T w) = \|w\|^2$$

per cui dobbiamo massimizzare la kurtosis con il vincolo $\|w\| = 1$.

Il ragionamento può essere steso nel caso di più componenti indipendenti per cui il vincolo $\|w\| = 1$ significa che il vettore w deve giacere sulla sfera n-dimensionale di raggio unitario. La componente indipendente stimata $y = w^T z$ rappresenta la proiezione dei dati nella direzione di w. Noi possiamo quindi stimare la kurtosis di y al variare di w, che nel caso bidimensionale equivale a variarne l'angolo che forma con l'asse orizzontale.

Metodo di discesa al gradiente - Un metodo utilizzato per risolvere il problema è quello della discesa al gradiente: si parte da un valore di w qualsiasi e si varia w andando a stimare, sulla base dei campioni disponibili, la direzione lungo la quale il valore della kurtosis cresce maggiormente. L'algoritmo si ferma quando è stato trovato un minimo o un massimo della kurtosis.

Il gradiente del valore assoluto della kurtosis può essere scritto come

$$\frac{\partial |kurt(w^T z)|}{\partial w} = 4 sign(kurt(w^T z)\left[E\{z(w^T z)^3\} - 3w\|w\|^2 \right]$$

Visto che w deve giacere sulla sfera di raggio unitario, dopo aver variato w, bisogna dividerlo per la sua norma. Inoltre visto che a noi interessa solo la variazione della direzione di w, non interessano le variazioni della kurtosis al variare della norma di w e quindi possiamo trascurare il termine nella parentesi quadra a destra nell'espressione del gradiente. In questo modo l'algoritmo di discesa al gradiente si può scrivere come

$$\Delta w \propto sign(kurt(w^T z) E\{z(w^T z)^3\}$$
$$w \leftarrow w / \|w\|$$

Una volta stimata una componente indipendente, per stimare le rimanenti, si può sfruttare il fatto che i vettori w_i nello spazio sbiancato, sono tra di loro ortogonali (si ricorda che la matrice di mixing dopo lo sbiancamento è una matrice ortogonale). Infatti ricordiamo che l'indipendenza statistica delle componenti implica la loro incorrelazione quindi $E\{(w_i^T z)(w_j^T z)\} = w_i^T w_j$: questo significa che l'incorrelazione è equivalente alla ortogonalità dei vettori w_i.

Per stimare più componenti bisogna fare in modo che i vettori w_i siano tra di loro ortogonali. Questo può essere ottenuto con il metodo di ortogonalizzazione di Gram-Schmidt: supponiamo di avere stimato p vettori w_1, ..., w_p e si voglia stimare il vettore w_{p+1}. Dopo aver eseguito un passo dell'algoritmo di discesa al gradiente bisogna sottrarre dalla stima di w_{p+1} le proiezioni $(w_{p+1}^T w_j) w_j$ con $j=1,...,p$ e rinormalizzare w_{p+1}.

1.5.5.3 Misura della nongaussianità tramite Negentropia

Una misura più robusta della non gaussianità è data dalla Negentropia.

Il concetto di Entropia nasce nel campo della teoria dell'informazione come misura della lunghezza della codifica necessaria per descrivere un certo dato; in riferimento alla misura di una variabile aleatoria essa rappresenta il grado di informazione che porta l'osservazione di tale variabile. Più una variabile è casuale più la sua entropia è elevata poiché necessita di una codifica più lunga; se invece la variabile aleatoria ha una distribuzione racchiusa in uno stretto *range* di valori allora essa potrà essere codificata con un numero inferiore di "bit" e la sua entropia risulterà più bassa.

L'entropia differenziale H di un vettore aleatorio y con densità di probabilità $f(\mathrm{y})$ è definita

$$H(\mathrm{y}) = -\int f_{\mathrm{y}}(\mathrm{y}) \log f_{\mathrm{y}}(\mathrm{y}) d\,\mathrm{y}$$

Un risultato fondamentale della teoria dell'informazione è che variabili gaussiane hanno l'entropia maggiore tra le variabili di uguale varianza. Quindi tale grandezza può essere sfruttata per avere una misura della nongaussianità.

Per ottenere una misura della nongaussianità che sia positiva e zero per variabili gaussiane, possiamo introdurre una versione normalizzata della entropia differenziale, detta neg-entropia:

$$J(\mathrm{y}) = H(\mathrm{y}_{gauss}) - H(\mathrm{y})$$

dove y_{gauss} è un vettore aleatorio gaussiano che presenta la stessa matrice di correlazione di $\boldsymbol{y}$. La negentropia è nulla per variabili gaussiane (quelle con la massima entropia H).

La Negentropia è uno stimatore ottimo della non gaussianità, tuttavia la sua valutazione secondo la definizione è difficoltosa da un punto di vista computazionale, in quanto richiede la stima della densità di probabilità della variabile. Nella pratica si utilizzano delle approssimazioni di cui quella classica, ma non la migliore a causa dell'utilizzo della Kurtosis, è:

$$\mathrm{J}(\mathrm{y}) \approx \frac{1}{12}\mathrm{E}\left\{y^3\right\}^2 + \frac{1}{48} kurt(y)^2$$

dove y è una variabile aleatoria a valor medio nullo e varianza unitaria. Questa approssimazione porta tuttavia all'uso della Kurtosis visto nel paragrafo precedente in quanto il primo termine della relazione precedente, detto skewness, vale zero nel caso di variabili con distribuzione (approssimativamente) simmetrica, cosa che accade piuttosto comunemente.

Si preferisce perciò usare delle *funzioni non-polinomiali* con cui generalizzare le cumulanti di ordine superiore. Possiamo sostituire le funzioni y^3 e y^4 con altre funzioni G^i (dove i è un indice e non una potenza) e avere un modo semplice di approssimare la Negentropia basato sulle aspettazioni di $E\left\{G^i(y)\right\}$. Questo metodo deriva direttamente dal principio della massima entropia con cui si va a stimare il valore massimo di entropia compatibile con i nostri dati; le funzioni non polinomiali crescono più lentamente di una parabola (y^2) e questo riduce la sensibilità dell'approssimazione agli outliers.

Per esempio possiamo usare due funzioni non quadratiche G^1 e G^2, con G^1 dispari e G^2 pari, ottenendo:

$$J(y) \approx k_1\left(E\left\{G^1(y)\right\}\right) + k_2\left(E\left\{G^2(y)\right\} - E\left\{G^2(v)\right\}\right)^2$$

con k_1 e k_2 costanti positive, e variabile gaussiana a valor medio nullo e varianza unitaria.

Se vogliamo usare una sola funzione non quadratica, l'approssimazione diventa:

$$J(y) \propto \left[E\left\{G(y)\right\} - E\left\{G(v)\right\}\right]^2$$

Scegliendo opportunamente la funzione G si possono ottenere ottime approssimazioni della neg-entropia. Tra le più usate troviamo:

$$G(y) = \frac{1}{a_1}\log(\cosh a_1 y)$$

$$G(y) = -\exp(-y^2/2)$$

con $1 \le a_1 \le 2$ costante che decide il passo di convergenza. Notiamo che per $G(y)=y^4$ riotteniamo l'espressione che utilizza la Kurtosis.

In Fig. 16 è riportato un esempio di simulazione del processo di mixing e unmixing per quattro segnali sorgente.

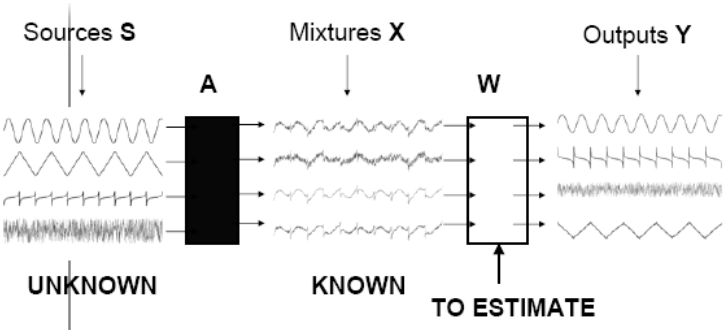
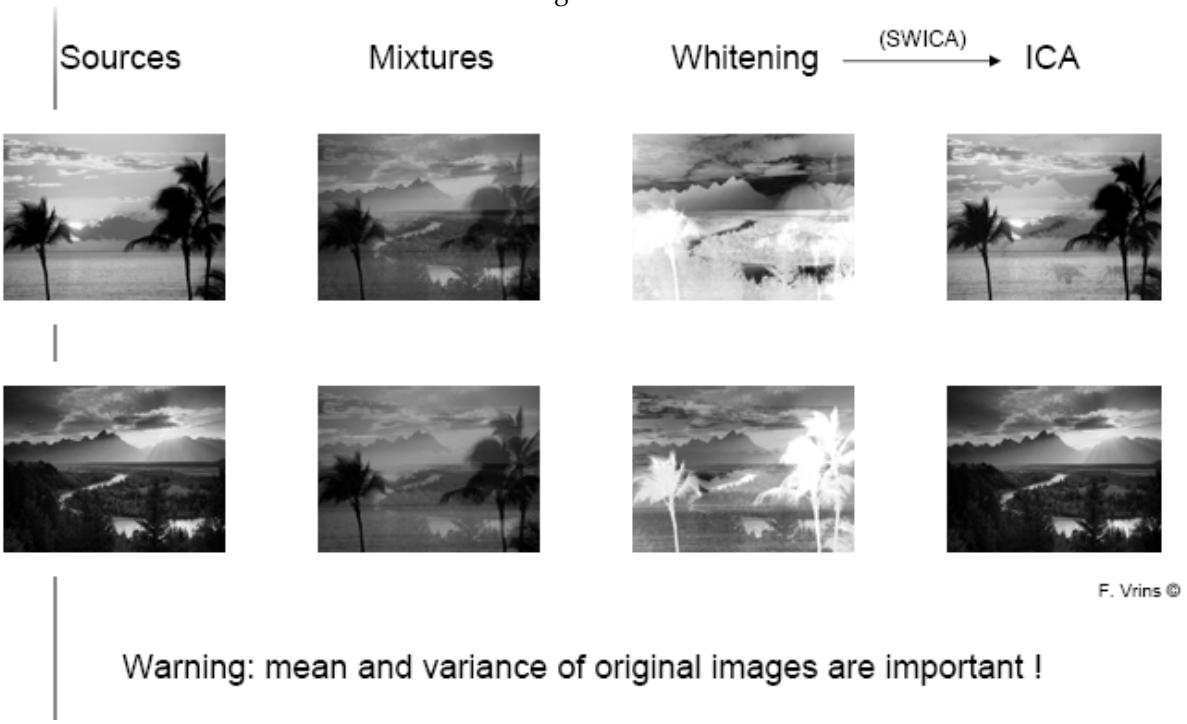


Fig. 16



1.5.6 Minimizzazione della muta informazione

Un altro approccio utilizzato per la stima delle componenti indipendenti è quello che fa uso della mutua informazione. Se consideriamo m variabili aleatorie, $y_1,\ldots, y_m$ si definisce mutua informazione tra le m variabili come

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\boldsymbol{y})$$

dove con $\boldsymbol{y}$ si indica il vettore formato dalle m variabili aleatorie. La mutua informazione è una misura naturale della dipendenza tra le variabili causali: è sempre maggiore di zero e nulla se le variabili sono statisticamente indipendenti.

Le $H(y_i)$ indicano la lunghezza della codifica per le y_i, quando vengono considerate separatamente, mentre H(y) quando sono considerate insieme. La mutua informazione di dice che riduzione di codifica si ottiene se le variabili sono considerate insieme piuttosto che separatamente: se sono indipendenti i due diversi modi di considerare le variabili sono equivalenti.

Vediamo come può essere utilizzata la mutua informazione per la stima delle componenti indipendenti.

Se prendiamo una trasformazione lineare invertibile $\mathbf{y = Wx}$ avremo che $\mathbf{H(y) = H(x) + \log|det(W)|}$ per cui la mutua informazione delle variabili trasformate, che sono la stima delle componenti indipendenti, è data da

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\mathbf{x}) - \log|det(W)|$$

se le y_i sono incorrelate e hanno varianza unitaria $\boldsymbol{E\{yy^T\} = WE\{xx^T\}W^T = I}$ quindi

$$det\,\boldsymbol{I} = 1 = det(\boldsymbol{WE\{xx^T\}W^T}) = (det\,\boldsymbol{W})det(\boldsymbol{E\{xx^T\}})det(\boldsymbol{W^T})$$

questo implica che, visto che $\det \boldsymbol{E\{xx^T\}}$ non dipende da W allora il det(W) è costante.

Visto poi che entropia e neg-entropia differiscono solo per il segno e una costante la mutua informazione può essere scritta come

$$I(y_1, y_2, \ldots, y_m) = C - \sum_i J(y_i)$$

La relazione precedente mostra che trovare una trasformazione W che minimizza la mutua informazione tra le componenti stimate è equivalente a trovare le direzioni lungo le quali la neg-entropia è massimizzata. Si deve notare che il metodo della mutua informazione *non necessita del vincolo di incorrelazione tra le componenti indipendenti*.