

Si hanno dei vettori in uno spazio D-dimensionale (D in genere indicato in maiuscolo per dire che la dimensionalità è elevata).

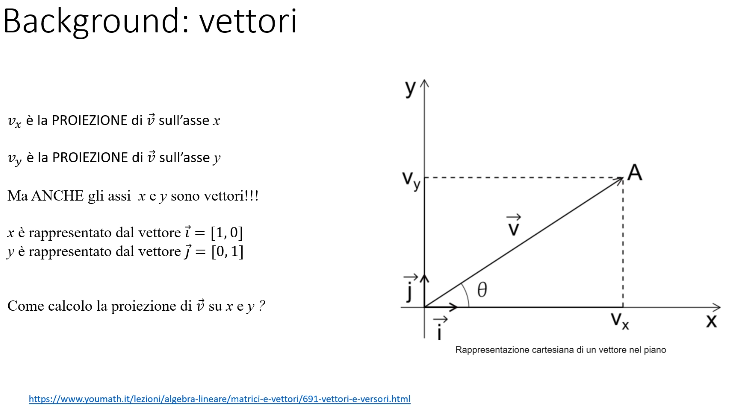
Come facciamo a visualizzare questi vettori?

Con le visualizzazioni abbiamo a disposizione spazi 2D o 3D.

Se avessimo 1000 dimensioni, dovremmo potenzialmente prendere le triplette di valori ma non ce la facciamo, è difficile, allora si cerca di prendere questi vettori ***xi*** e mapparli in vettori ***yi*** che sono vettori in uno spazio vettoriale di dimensionalità molto minore di D e nella visualizzazione scientifica è tipicamente 2 o 3.

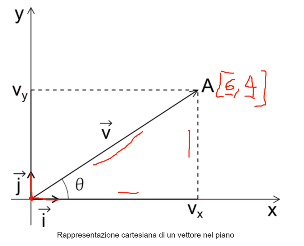
Principal Component Analysis (hopefully) simply explained

Uno dei metodi a questo fine è l’analisi delle compenenti principali (**PCA**).

Cos’è un vettore?

Un vettore è visualizzato come una freccia o segmento orientato.

Ha un **punto di applicazione** , ha un **punto di fine** e generalmente il vettore è indicato con il numero di componenti lungo un piano composto da **assi ortogonali**.

Se disegniamo il punto A di coordinate (6,4) nel piano cartesiano, allora possiamo tracciare il vettore che parte dall’origine e arriva al punto A.

Si tratta di un vettore orientato che è indicato con i due elementi 6 e 4.

Dove 6 è ***vx*** (proiezione di A sull’asse x) e 4 è ***vy*** (proiezione di A sull’asse y).

Anche x ed y, gli assi, sono due considerati due vettori (infatti hanno la freccia).

Per rappresentare x ed y usiamo i vettori normalizzati.

Tutti i vettori hanno una norma, ossia una lunghezza di un vettore, che è pari alla radice del: quadrato del primo elemento + il quadrato del secondo elemento.

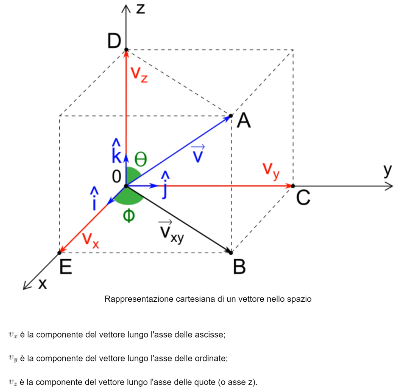
È a tutti gli effetti il teorema di Pitagora per calcolare la diagonale di un triangolo che ha come due cateti le proiezioni del vettore sugli assi.

vx alla 2 + vy alla 2 dà la norma, ossia la lunghezza del vettore A (che è un vettore bidimensionale).

Molte volte però del vettore A non ci interessa, ci interessano di più invece gli assi normalizzati.

Siccome a noi degli assi non interessa la lunghezza ma solo la loro direzione e nel caso di spazio ortogonale ci interessa che siano assi ortogonali; in genere i vettori che rappresentano gli assi nello spazio euclideo cartesiano sono i vettori ***i*** e ***j*** che sono vettori con due semplici elementi [1,0] e [0,1].

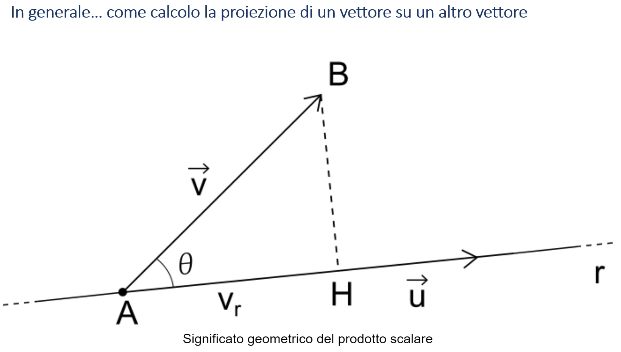
Cosa si intende per proiezione e come si fa a calcolare la proiezione di un vettore sugli assi e quindi sui vettori i e j?



A → Si tratta di un vettore tridimensionale.

In questo caso il vettore A corrisponde al punto [5,5,5], ha tre elementi perché ha tre proiezioni: una sul vettore che identifica l’asse x, una sul vettore che identifica l’asse y e una sul vettore che identifica l’asse z.

Anche in questo caso si parla di proiezioni.

(Siamo in un caso bidimensionale.)

Prendiamo i vettori ***u*** e ***v***.

Si vuole proiettare il vettore v sul vettore u: significa che si prende il vertice estremo di v, traslo v e piazzo la sua origine in u (ovunque, tendenzialmente si piazza in modo intelligente), poi si traccia all’altro vertice di v la perpendicolare a u.

Generalmente se u parte, dove si piazza l’origine del vettore v? Sull’origine del vettore u.

Partendo dalle stesse origini, si prende l’altro vertice di v e lo si proietta perpendicolarmente al vettore u.

La lunghezza del segmento ***vr*** è la lunghezza della proiezione di v su u.

Questo è il modo in cui si calcola il prodotto scalare di v per u:



Si tratta della somma di prodotti degli elementi corrispondenti di v ed u.

Se v è un vettore nello spazio D-dimensionale, avrà elementi v1, v2,v3, v4 fino a vD.

Qundi ad esempio in uno spazio 4-dimensionale, v dovrebbere essere il vettore [1,2,3,4].

Per cui la proiezione sul primo asse è 1, la proiezione sul secondo asse è 2, la proiezione sul terzo asse è 3 e la proiezione sul quarto asse è 4.

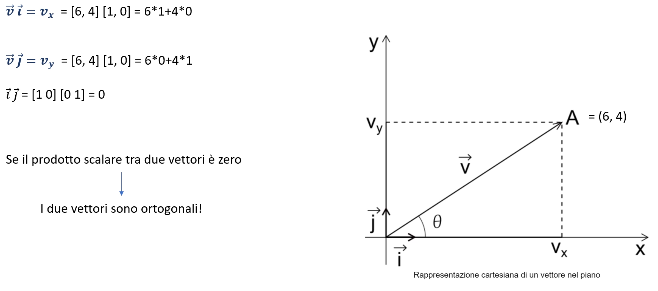
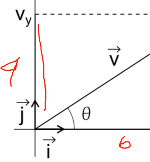
[ Non riusciamo ad immaginarlo perché noi umani vediamo solo 3 dimensioni. ]

E u supponiamo sia [7,7,7,7].

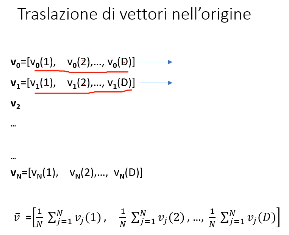
Per fare la proiezione del primo vettore v sul seondo vettore u: si prende il primo elemento di v, ossia 1, lo si proietta per 7 (primo eleemto di u), questo prodotto si somma per il secondo elemento di v per il secondo elemento di u e così via.

AH = 1\*7 + 2\*7 + 3\*7 + 4\*7

La somma di tutto questo dà la lunghezza del segmento **AH** che è la proiezione di v su u.



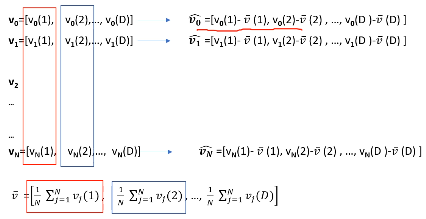
Vogliamo vedere qual è la proiezione del vettore A sugli assi, quindi sul vettore i e sul vettore j che rappresentano gli assi x ed y rispettivamente.

Cosa significa traslare un vettore nell’origine?

Se si hanno i vettori in uno spazio D-dimensionale, quindi:

il vettore v0 è composto dal primo elemento, dal secondo elemento e dal D-esimo elemento e così tutti gli altri vettori.

Per traslare i punti dei vari vettori nell’origine, bisogna calcolare il vettore medio: si prende la media di tutti i primi elementi, secondi elementi e D elementi.

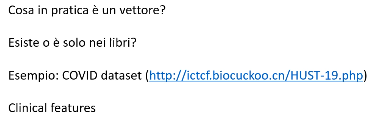
Questo è ciò che bisogna fare effettivamente per far sì che il punto medio di tutti i punti che si vedevano prima diventino l’origine degli assi: semplicemente bisogna sottrarre ad ogni vettore il vettore medio.

La sottrazione tra vettori avviene sottraendo elemento per elemento.

Quindi la sottrazione permette di avere un nuovo vettore della stessa dimensione dei due vettori.

Il vettore medio è il vettore composto dalle medie di elementi corrispondenti.

[ Il vettore si può indicare con la freccia sopra altrimenti col carattere in bold, il carattere non bold indica un elemento del vettore. ]

Ogni riga del dataset è un vettore che rappresenta il paziente.

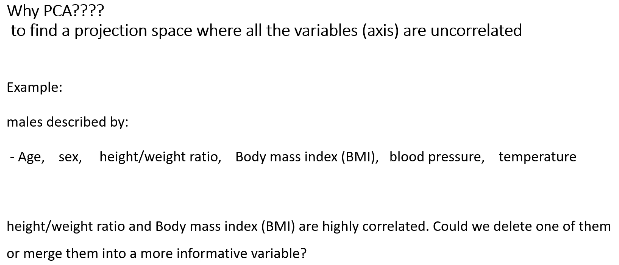
[ Il vettore in questo caso è 300-dimensionale.

Ogni elemento del vettore è una caratteristica del vettore. ]

Quindi i vettori esistono anche nella realtà.

La PCA si applica tendenzialmente a variabili continue cioè a vettori che sono rappresentati da valori interi o valori continui.

Per valori categorici si applicano tecniche diverse, anche se applicare la PCA non è sbagliato.

La PCA è per proiettare i punti in un nuovo spazio dove tutti gli assi siano scorrelati.

Supponiamo di avere un dataset composto da pazienti o persone di cui si vuole fare uno studio.

Ogni persona è descritta da: età , sesso, rapporto altezza/peso, BMI, pressione del sangue e temperatura.

Ogni singola variabile è un asse.

Quindi c’è l’asse delle età e tutti i primi elmenti di tutti i pazienti saranno le proiezioni del vettore che rappresenta quei pazienti sull’asse età.

La proiezione che rappresenta un paziente sull’asse sesso.

Le altre proiezioni per altezza/peso, BMI …

[ BMI è una misura di quanto una persona è nomopeso, magra o fatty. ]

BMI è inversamente proporzionale alla ratio altezza/peso: se una persona è bassa e magra il BMI è basso.

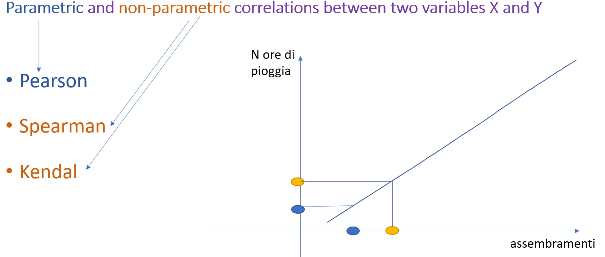
Il BMI è più alto per le persone obese.

Queste due variabili sono correlate da una correlazione inversa: al crescere di una decresce l’altra.

Queste informazioni sono in qualche modo ridondanti è necessario perciò averle entrambe? NO.

Si potrebbe avere un solo rappresetnante delle due variabili, si ridurrebbe la dimensionalità e si avrebbero tutte variabili tra di loro scorrelate, ovvero varibaili che portano tra di loro informazioni completamente diverse l’una dall’altra → è inutile avere variabili che indicano la stessa informazione.

Nel campo delle visualizzazoni si sta usando un asse in più che rende meno chiara la visualizzazione stessa e in più comunica un’informazione già ripetuta.

[ N° ore di pioggia – asse Y Assembramenti – asse X ]

Il prodotto scalare di due vettori ha una correlazione molto stretta con la correlazione tra due variabili.

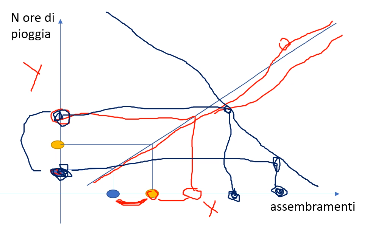
Una variabile può essere vista come un vettore.

Quindi si prendono tutti i pazienti, tutte le età, quella è una variabile.

L’insieme di tutte le età dei pazienti è la variabile X.

Se metto insieme alle età la pressione del sangue, questa è la variabile y.

Di due variabili si vuole vedere spesso se sono correlate perché se lo sono altamente, direttamente o inversamente, non interssano perché portano la stessa informazione: perché si sa che al crescere di una cresce o decresce quell’altra.

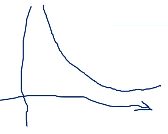


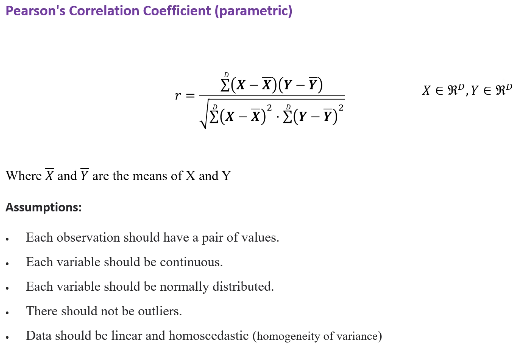
Correlazione diretta: le due variabili crescono insieme → se confronto due pazienti se il paziente giallo ha un valore più alto per la prima variabile, si sa che questa variabile x è direttamente correlata con una vraibile y, si saprà che tendenzialmente anche la seconda variabile del paziente giallo avrà un valore più alto di quella blu.

Se per il paziente blu la variabile x è bassa allora si sa che anche la variabile y è più bassa.

Correlazione inversa: al crescere della variabile x decresce la variabile y.

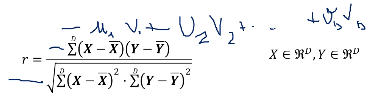
Queste correlazioni sono dette lineari, perché se le si plottano si ottiene una retta, quindi c’è una *dipendenza lineare*.

 [ Ci sono anche correlazioni anche non lineari per cui si ottiene una parabola magari. ]



Pearson misura la correlazione lineare tra due variabili.

Somma le moltiplicazioni tra elementi corrispndenti delle variabili sotratti dalla loro media.

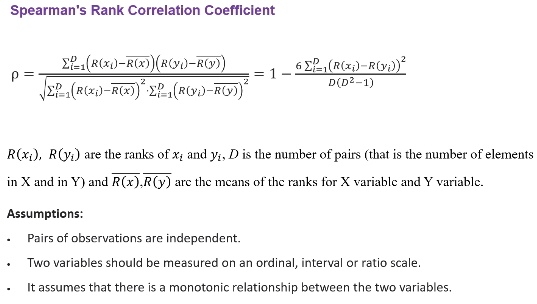
Prendiamo un vettore e lo normalizziamo:

• Al numeratore si sta facendo un prodotto scalare tra due vettori dopo la normalizzazione.

• Al denominatore si ha il fattore di normalizzazione che fa sì che la correlazione di Pearson vada sempre tra -1 (quando c’è una correlazione inversa), +1 (correlazione diretta) e 0 (quando x ed y sono ortogonali tra di loro ossia tra di loro non c’è alcuna correlazione).

Con PCA si vogliono prendere questi vettori (età, sesso, ..) e si vuole ottenere un nuovo spazio vettoriale in cui tutte queste variabili sono massimamente scorrelate.

PCA non “butta” variabili ma le combina per ottenere variabili al massimo scorrelate.



Mentre Pearson si applica a variabili continue, Spearman e Kendal si applicano anche a variabili discrete o comunque a variabili categoriche anche, che devono essere **ordinabili** (ossia per cui ci può essere un ordinamento).

Non confrontano il valore delle variabili, ma il *rank* degli elementi.

Quindi se si ha un vettore X ed un vettore Y e ad ogni elemento del vettore X corrisponde un elemento del vettore Y, quindi si hanno diversi pazienti rappresentati da Età e Blood Pressure.

Per il primo paziente la sua età corrisponde al blood pressure del primo paziente.

Per il secondo paziente la sua età corrisponde al blood pressure del secondo paziente.

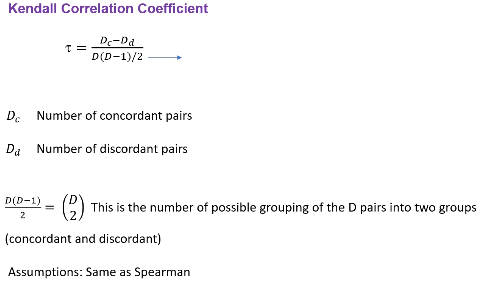
Quindi si hanno coppiette xy associate tra di loro.

Si fa il sort di tutti i valori che si ha per X (età), quindi per tutti i pazienti, si fa il sort di tutti i valori che si hanno per y (blood pressure).

Ad ogni elemento non si associa il suo valore ma il suo rank, cioè se era il primo dopo l’ordinamento…

Quindi si confronta → il rank degli elementi x meno il rank medio delle x + rank delle y meno rank medio delle y.

Spearman assume una correlazione monotona tra le due variabili X ed Y, cioè che siano rappresentate da una correlazione lineare sempre crescente o sempre decrescente.

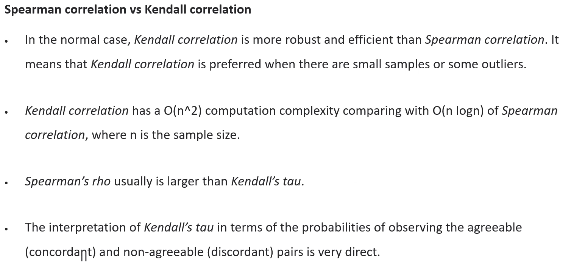


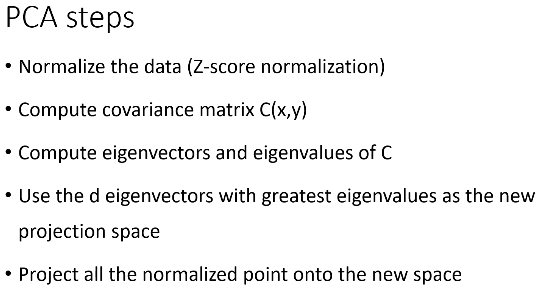
Kendal è simile.

Va a vedere i rank e misura il numero di ***pair*** che sono concordanti (stesso rank) e sconcordanti (diverso rank).

Identifica dei gruppi tra le variabili, dà un rank, e misura la correlazione tra variabili continue o discrete (come Spearman).

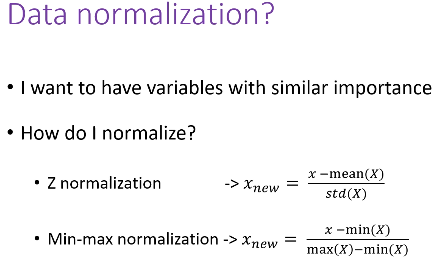
Quando si deve valutrare la correlazione tra variabili continue, viene utilizzato Pearson (se si vuole valutare una correlazione semplice, lineare).





La matrice di covarianza utilizza principalmente una *Pearson correlation*.

Tutto questo per diminuire la dimensionalità dello spazio vettoriale su cui proiettare i vettori.

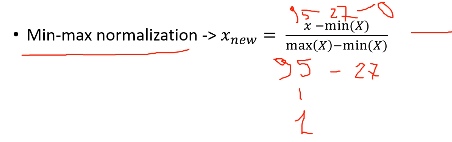


Ci sono due tipi di normalizzazione dei dati.

Quella più semplice è la min-max che porta tutti i valori ad essere tra 0 ed 1.

Supponiamo di avere la avriabile dell’età, il paziente con età minimo è 27, con età massimo è 95.

Si prende per ogni paziente la sua età si sottrae 27, si divide ..:



Il paziente con età massima diventa 1.

Tutti gli altri valori nel mezzo vanno a stare nel mezzo, quindi si sono normalizzati i dati per fare in modo che siano tra 0 ed 1.

La z-normalization calcola le medie di tutti i valori delle variabili età, ad ogni paziente sottrae il valore della media e divide per la deviazione standard di tutte le età dei pazienti.

PCA applica questa normalizzazione perché vuole che tutte le variabili abbiano tutte la stessa importanza.

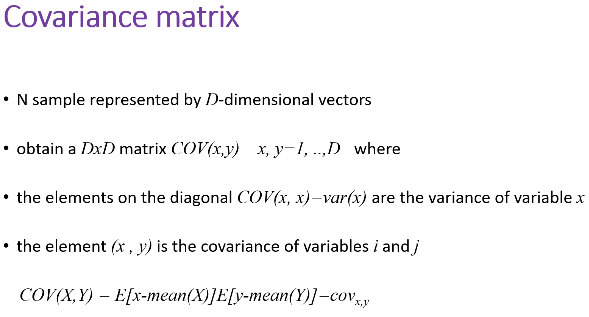
Se si ha età e blood pressure, blood pressure raggiunge valori 150, l’età massimo 100.

Se si andasse a pesare le variabili, se si volessero calcolare delle differenze euclidee tra pazienti.

Se si hanno variabili a range completamenrte diversi, non normalizzati, si ha il rischio che quelle variabili per cui per convenzione si ha un’unità di misura che è più alta, valgano di più (blood pressure in questo caso).

E non si vuole, si vuole dare la stessa potenza a tutte le variabili.

Si normalizzano perciò i dati per avere dati che sono comparabili, ossia nello stesso range.



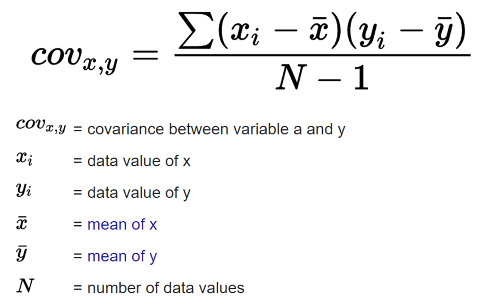
A questo punto la PCA applica *z-normalization*.

Dopo *Z score* (sottratto la media diviso per la deviazione standard), prende tutte le variabili e calcola la matrice di covarianza.

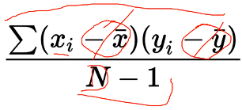
La matrice di covarianza tra un vettore e sé stesso è la varianza del vettore.

La matrice di covarianza tra due vettori, due variabili, è:

il valore atteso della prima variabile – la sua media \* la seconda variabile – la sua media.



Questo è il calcolo preciso per la covarianza.

Le medie si rimuovono per via della PCA, varrebbero 0.

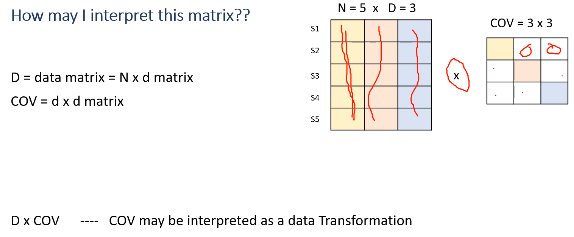
Al numeratore alla fine si ha il prodotto scalare tra due vettori, ma è anche molto simile alla correlazione di Pearson.

Quindi la covarianza misura quanto correlano le due variabili: se una variabile tende a crescere l’altra tende a crescere se la covarianza è positiva, se la covarianza è negativa allora si ha una relazione inversa ossia una cresce e l’altra deresce; se la covarianza è zero le due varibaili sono socrrelate e quindi ortogonali tra di loro.

Al denominatore N-1 è il fattore di normalizzazione.

Ogni paziente è un vettore D-dimensionale, ma quando si parla della varibaili, si parla di vettori N-dimensionali.

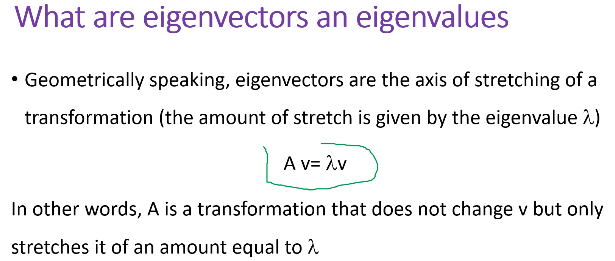
SI può calcolcare anche la covarianza di un paziente rispetto un altro.



La moltiplicazione di matrici è il prodotto righe per colonne tra due matrici che dà luogo ad un'altra matrice.

Alla prima riga della matrice risultante (dal prodotto tra le due matrici) si ha la proiezione del vettore sugli assi che sono individuati nella matrice di covarianza (quindi questa matrice è una matrice di assi che non sono ortogonali tra di loro ma raccontano la varianza tra le variabili).

Dopo aver calcolato la matrice di covarianza, la si diagonalizza, ossia si trovano degli assi ortogonali di questa matrice.



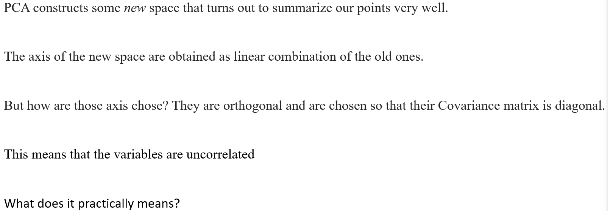
Gli autovettori sono quegli assi di una matrice di trasformazione che non vengono trasformati dalla matrice di trasformazione, ma sono solo allungati (stretchati) dall’autovalore corrispondente all’autovettore.

Gli autovettori sono gli assi ortogonali di quello spazio.

L’autovalore associato ad ogni autovettore spiega come quell’asse tiene i dati contenuti in quello spazio.

(La matrice di trasformazione si tiene spesso da dei dati.)

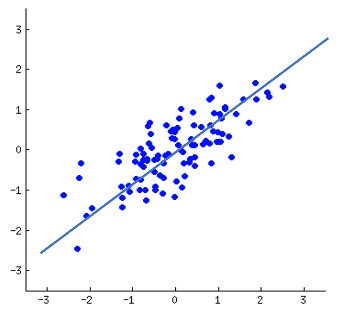
Si rimuovono gli autovalori meno importanti perché spiegano meno le cose.



Quindi la PCA:

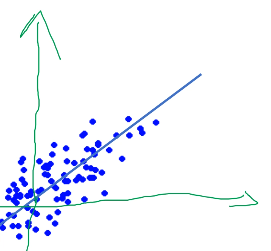
prende dei punti, li normalizza con la *z-normalization*, calcola la matrice di covarianza, considera la matrice di covarianza come una trasformazione in uno spazio che esprime la covarianza e cerca di abbattere la covarianza tra le variabili, trovando un nuovo spazio in cui tutte le vairbaili siano scorrelate.

Gli autovettori fanno da base dello spazio su cui si proiettano i punti.



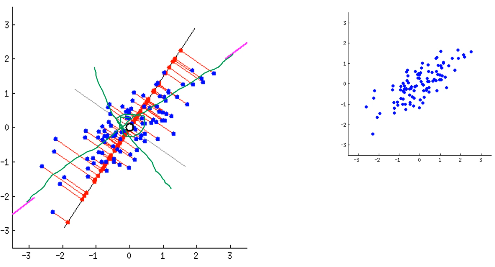
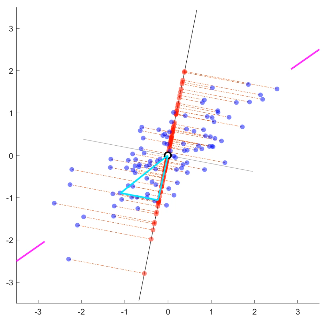
In pratica.

Si suppone di partire da tutto questo insieme di punti.

 Ciò che fa la PCA è normalizzare, quindi sposta gli assi nella media dei punti, alternativamente trasla i punti.

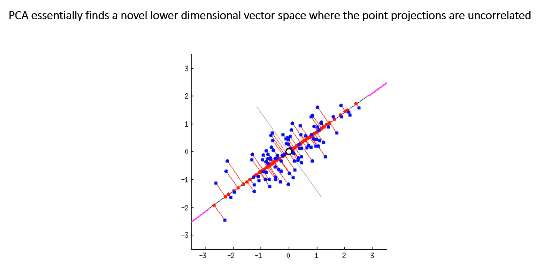
Poi calcola la correlazione e dalla matrice di correlazione (covarianza) trova gli assi di massima variazione.

La matrice di correlazione si può vedere come un nuovo spazio in cui si trovano gli assi dello spazio che esprimono meglio la correlazione, che sono ortogonali tra di loro.



In questo spazio si hanno punti 2D, si potrebbe diminuire la dimensionalità per avere punti 1D.

Quindi si può considerare un unico asse su cui si rappresentano i punti.

Quindi la PCA:

trova l’asse su cui le proiezioni dei punti sono scorrelate, massimamente separate.

Aiuta a trovare un nuovo spazio in cui i punti sono molto più distanziati tra di loro e ben visibili.

