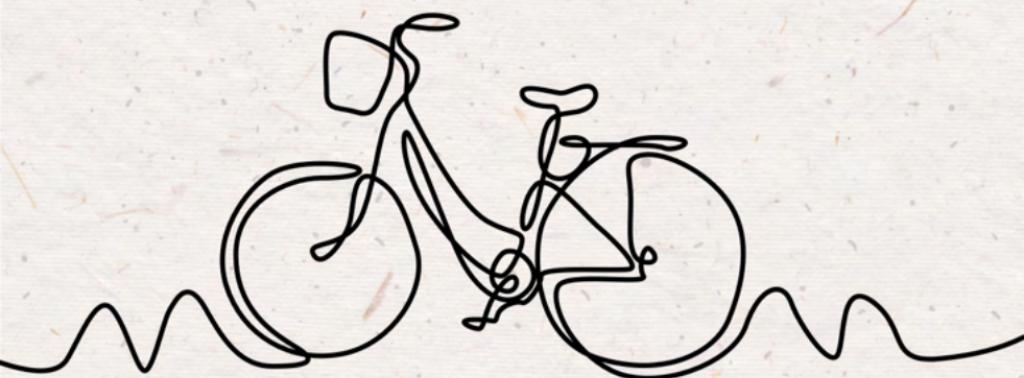


Title



Notes

(PCA) Dimensionality reduction

UN PCA = PRINCIPAL COMPONENT ANALYSIS, DATO UN DATASET CON MEDIA 0 HA IL COMITO RIDURRE LA DIMENSIONE DEL DATASET USANDO UNA TRANSFORMAZIONE LINEARE CHE PROIECTA SU UN SOTOSPAZIO, CON IL COMITO DI PRESERVARE LA MASSIMA PARTE DELLE INFORMAZIONI.

UN SOTOSPAZIO PUÒ ESSERE rappresentato da una matrice $P \in \mathbb{R}^{m \times n}$ LE CUI COLONNE SONO ORTHONORMALI (SONO ORTHONORMALI)

LA PROIEZIONE

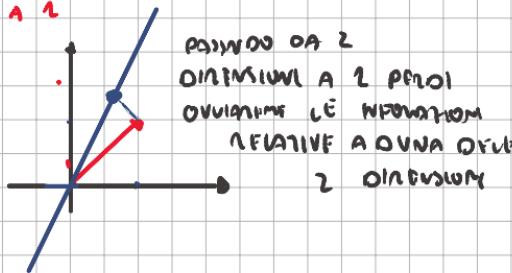
LINERAR MOLTO UNA \rightarrow LA LUNGHEZZA DEI VETTORI
ROTAZIONE DEGLI ASSI \rightarrow UNIFORME
E IL LUNGO PROIEZIONE SAREBBERE $\neq 0$

LA PROIEZIONE DI x (DATA DI x) SU P EQUIVALE AD UN PRODOTTO VETTORIALE $P \cdot x$ CHE IN TERMINI MATRICIALI EQUIVALE A $P^T x$

$$y = P^T x$$

IN 2 DIMENSIONI

$A \in \mathbb{R}^{2 \times 2}$



CONSIDERANDO y È QUNDO LA PROIEZIONE DI x SUL NUOVO SOTOSPAZIO (P^T) L'UNICO POSSIBILE x POSSIBILE INDIRIZZARE UNO SPAZIO ALLE STESE DIMENSIONI DI x INIZIALE E' MOLTIPLICARE PER $(P^T)^{-1}$

$$2 \begin{bmatrix} 3 \\ A \end{bmatrix} \begin{bmatrix} 3 \\ x \end{bmatrix} \rightarrow \text{POSSIBILE} x \text{ OGNI 3 A 2 DIMENSIONI}$$

CONSERVARE LE DIREZIONI

$$3 \begin{bmatrix} 2 \\ A \end{bmatrix} \begin{bmatrix} 2 \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ x \end{bmatrix}$$

PIÙ POSSIBILE ALLE 3 DIMENSIONI

CONSERVARE LE SAMI UN DIREZIONI

$A^T A^{-1}$ DÀ LO stesso COME $A^{-1} A^T$

$A^T A^{-1}$ NON È POSSIBILE MANI ALTRI AVVISTAMENTI

L'INFORMAZIONE PER UNA MISURAZIONE CONSIDERANDO L'AVVOLTA DI COSTRUZIONE ERRORE.



619

$$A = R_2 Z R_2 \quad A^{-1} = R_2^{-1} Z^{-1} R_2^{-1}$$

$$(A^{-1})^T = R_2 Z^{-1} R_2 \quad A^T = R_2^T Z^T R_2^T$$

$$R_2^{-1} Z^{-1} R_2$$

[$R_2 \in R_2$]

Sono ormai
Z è diagonale

$A^T + A^{-1}$ A NUO CIR

$Z = Z^{-1}$ C'È UN ERRORE DI

NUO CIR

A^T È L'UNICO POSSIBILE X

Trovare l'indiano utilizzando UNA

trasformazione simile AD A^{-1}

L'ERRORE DI NUO CIR E CALCOLATO COME DISTANZA EULLIDICA TRA I 2 PUNTI $x \in \mathbb{R}_n$ E $\hat{x} = x$ NUO CIR

$$\text{errore} = \frac{1}{n} \sum_{i=1}^n \|x_i - \hat{x}_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|x_i - p_i\|^2 =$$

OLI VETTORE

COLONNA VIFNE

TRASPOSTO SVILLO

SPAZIO PIÙ GRANDE

$$\frac{1}{n} \sum_{i=1}^n \|x_i - p_i\|^2$$

VOLIAMO QUINDI TROVARE I VALORI DI P PER CUI L'ERRORE È IL PIÙ DICCOLO POSSIBILE

$$p^* = \underset{P}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \|x_i - p_i\|^2$$

XH_P ADDIAMO QENDO CHE P È ORTHOGONALE QUINDI $P^T = P^{-1}$, $P^T P = I$

Norma la norma

$$\|v\|^2 = \langle v, v \rangle = v^T v$$

$$\begin{aligned}
 L(p) &= \frac{1}{h} \sum_{i=1}^h (x_i - p p^T x_i)^T (x_i - p p^T x_i) = (x_i^T - x_i^T p p^T) (x_i - p p^T x_i) \\
 &= \frac{1}{h} \sum_{i=1}^h (x_i^T x_i - x_i^T p p^T x_i - x_i^T p p^T x_i + x_i^T p p^T p p^T x_i) \\
 &= \frac{1}{h} \sum_{i=1}^h (x_i^T x_i - 2 x_i^T p p^T x_i + x_i^T p p^T x_i) \\
 &\quad \frac{1}{h} \sum_{i=1}^h (x_i^T x_i - x_i^T p p^T x_i)
 \end{aligned}$$

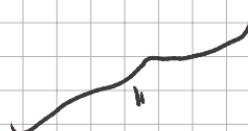
Consideriamo la traccia e poniamo fatto per lui

$p p^T = I$ → quindi alla FWF convergono tutti i valori su una
matrice diagonale ($x_i^T x_i$ non è invertibile) → inoltre

$$= \frac{1}{h} \sum_{i=1}^h \text{Tr}(x_i^T p p^T x_i)$$

+ ogni vettore
colonna
 $\vec{v} w = \text{Tr}(vw)$

Perche $\text{Tr}(AB) = \text{Tr}(BA)$



$$\begin{aligned}
 &= \frac{1}{h} \sum_{i=1}^h \text{Tr}(p^T x_i x_i^T p) = \text{Tr}\left(p^T \left[\sum_{i=1}^h x_i x_i^T\right] p\right) \\
 &\quad \text{matrice } h \times h \quad \text{e } x_i \in \mathbb{R}^n \\
 &\quad \text{vetture colonne} \quad \left[\begin{matrix} \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot \end{matrix} \right]
 \end{aligned}$$

Si ottengono le FWF $p^* = [v_1 \dots v_m]$ ovvero la soluzione che minimizza
l'errore di approssimazione e fornita dalli $m < m$ a tutte le FWF non
corrispondenti alle autovetture (ma solo) della matrice Σ

$$\frac{1}{h} \sum_{i=1}^h x_i x_i^T = U \Sigma U^T$$

Secondo la SVD

$$p^* = [v_1 \dots v_m] \quad \begin{array}{l} (\text{non sono}) \\ \text{autovetture di } A \\ \text{ma} \\ \text{sono} \\ \text{colonne di } P^* \\ \text{o } P^{-1} \end{array}$$

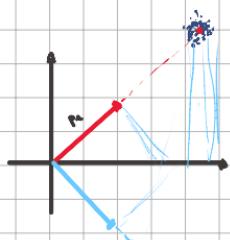
simile valore
di Σ di Σ

LA CARATTERISTICA DI V È CHE I SUOI AUTOVETTORI SONO SEMPRE ORTHONORMALI
E PERCIÙ SONO MATRICI SIMMETRICHE

COSA succede se il dataset non è zero mean??

IL DATASET VA CENTRATO IN QUANTO, LA PRIMA DIREZIONE DELLA PCA, NON CHÉ UN PIANO AUTOVETTORE SARA' VICINO ALLO SPAN CHE PASSA DALLA MEDIA, IL CUI NON E' NIENTE INTERESSANTE (NON HA PESO NELLA COMBINAZIONE)

Inoltre l'errore di ricostruzione sarà più alto, perché gli autovettori della matrice P che rappresentano il suo sottospazio poggiano tutte direzionali più distante dall' dataset



ERRORE DI RICOSTRUZIONE
SICURAMENTE PIÙ ALTO
(VEDILO X PIÙ DISTANZI)

DI CONSEGUENZA dato che quando l'origine dello spazio è allontanata, (non conoscendo nulla di informazioni), i dati vengono spostati in modo tale che l'origine sia nella loro media (cosa dei binari?)

IN PRATICHA TRAVERSO

UNO SPazio DI PROIEZIONE CHE

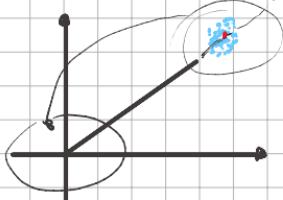
CI PERMETTE DI SCRIVERE I DATI IN
PONTELLI MOLTI OLTRE

(SEMPLIFICA RACCOLTA Dati UNIVARI)

notiziarie x

una partite

POTRAVEMENTE VOLIAMO TRAVERSARE UNO SHIFT DEI DATI PER CUI QUANDO APPLICHIAMO LA PROIEZIONE P SUI DATI SCRITI, LIMITIAMO L'ERRORE DI RICOSTRUZIONE.



SI PUÒ DIMOSTRARE CHE
LA SOLUZIONE OTTIMALE SI
HA SHIFTONDO I DATI
ALLEGGERITO GLI DATI
NELL' X

AQISSO CONSIDERANO O SEANZA LA FORMULA PRECEDENTE
 X CUI IL SOGGERITO P CHE MINIMIZZA L'INVERSA DI COVARIANZA È DATO DA
 UU AUTOVETTORI DELLA EALD DECOMPOSITION DI $\frac{1}{n} \sum_{i=1}^n x_i x_i^T$

Allora dato che i dati sono SHIFTED BASTA CONSIDERARE

$$C = \frac{1}{n} \sum_i (x_i - \bar{x})(x_i - \bar{x})^T = V \Sigma V^T$$

OVRNO LIU UN AUTOVETTORE DI V CORRISPONDENTI ALLI AUTOVETTORI
 PIÙ GRANDI

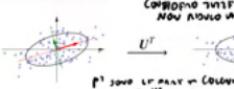
C'È CHIARATA ESPRESSA COVARIANCE MATRIX OF THE DATASET

MALESIMA QUANTO SPRAZI SONO I PUNTI LUNGO LE

FINE

Nota $V^T(x_i - \bar{x})$ È UNA TRANSFORMAZIONE LINEARE, DATA DA UNA
 MATE. OMOCARTE V^T (A QUILI SE È FULL RANK È
 PROVATA UNA ROTAZIONE DEL DATASET, E UNA PROIEZIONE
 DGSF)

Projection over V^T transforms our data so that the different directions
 are uncorrelated



To keep only the m directions with highest variance we keep only
 the first m transformed directions.

MATRICI DI
 COVARIANZA
 DOPO LA
 TRANSFORMAZIONE:

$$C' = \frac{1}{n} \sum_i [V^T(x_i - \bar{x})] [V^T(x_i - \bar{x})]^T$$

$$\begin{aligned} & \frac{1}{n} \sum_i V^T(x_i - \bar{x})(x_i - \bar{x})^T V = \\ & = V^T (V \Sigma V^T) V = \Sigma \end{aligned}$$

QUESTO SIGNIFICA CHE LA MATE. DI COVARIANZA CHE OTTERMO
 PER LE PRINCIPALI TRANSFORMATE È DIAGONALE, OVVERO LE PRINCIPALI
 DIREZIONI DELLA MATE. DI COVARIANZA SONO ALLINEATI CON
 GLI ASSI CARTESIANI

Non, la prima oinfitione, è la oinfitione in cui i dati hanno varianza massim, sono più distanti dal centro, la seconda oinfitione ci da la oinfitione con la seconda varianza massim e così via...

↓

Si evince quindi che la PIA ci permette di preferire le caratteristiche con varianza maggiore.

Quindi i dati più
distanti rispetto - quindi quelli
alla 1^ volta che si distinguono
dalle altre

ALESSIO COME SCEGLIO m ? $\ln^n \rightarrow \ln^m$, $m < n$

il valore di m può essere scelto a tentativi, usando la cross validation e verificando quale che ha l'errore minore, ovviamente va esclusa l'identity matrix quindi evitando $m=n$

$$\arg \min_m \min_{\theta} \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2 \neq 0$$

$$\text{se } x_i = x_i$$

Dopo che l'errore non si può usare potrei usare l'accuracy e vedere quanto ci attesta

Ottima si usa un criterio basato su un t (threshold) cui misura quando le deviazioni stanno mantenendo in percentuale varianza / varianze di Σ

$$\min_m m \text{ s.t. } \frac{\sum_{i=1}^m b_i}{\sum_{i=1}^n b_i} \geq t$$

Solitamente si sceglie m in base quanto a quanta informazione (varianze voluto mantenere) dei soli valori tra 90% 95% 99%.

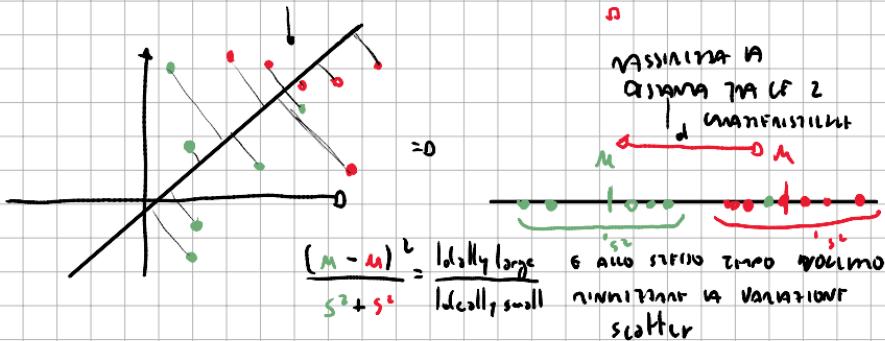
comunque si sceglono valori casuali

Non uno dei svantaggi della PIA è la difficoltà di calcolare la matrice di covarianza per valori di n (molte mani)

LDA (Linear Discriminant Analysis)

LA LDA SI CONCENTRA SUL TRUVARE LA DIREZIONE O SUBSPAZIO CHE PERMETTE DI PRESERVARE LA MASSIMA VARIANZA TRA I DATI, MA NI INTENDE EFFETTIVAMENTE QUESTA INFORMAZIONE?! INOLTRE ESSA NON È TOLERANTE ALLE TRANSFORMAZIONI LINEARI (SE MASCHERO LINEARMENTE LA DIREZIONE DELLA PIA (CLASSA))

SE VOLERSI PRESERVARE ALTRI INFORMAZIONI USO LDA, LDA NI PERMETTE DI TRUVARE UNO SPAZIO CHE MI VA A MASSIMIZZARE LA DISTANZA TRA LE MEDEIE DEI CLASSE, E MINIMIZZARE LA DISTANZA TRA I PUNTI NELLA CLASSE



CAPITO COSA VOLGO FAR. DEVO TRUVARE APROVATO UN NUOVO
CHE MI PERMETTA DI PRESERVARE LA DISTANZA TRA LE CLASSE E LA DISTANZA TRA I
PUNTI DELLA CLASSE PROLUNGANDO NEL NUOVO SOTTOSPAZIO.

① LA VARIANZA DEI PUNTI RISPETTO ALLA MEDIA DELLA CLASSE A CUI APPARTENGONO È UNA MISURA DELLO SPAREGGIO NELLA CLASSE

$$S_w = \sum_{i=1}^k N_i \text{Var}_i = N_1 \text{Var}_1 + N_2 \text{Var}_2 + \dots$$

② LO SPAREGGIO TRA LA MEDIA DI OGNI CLASSE E LA MEDIA TOTALE

$$S_B = N_1 \| \mu_1 - \bar{\mu} \|^2 + N_2 \| \mu_2 - \bar{\mu} \|^2$$

VOLUO IN PARTICOLARE ASSICURARE CHE (2) È MINIMIZZANTE LA (2)

Più il numero delle classi

$$S_B \triangleq \frac{1}{N} \sum_{c=1}^n m_c \underbrace{\left(\mu_c - \mu \right) \left(\mu_c - \mu \right)^T}_{\text{INTER CLASS}}$$

m_c n. di
classi
di classe
m. classi

$$S_W \triangleq \frac{1}{N} \sum_{c=1}^n m_c \left[\frac{1}{m_c} \sum_{i=1}^{m_c} (x_{ci} - \mu_c) (x_{ci} - \mu_c)^T \right] =$$

$$= \frac{1}{N} \sum_{c=1}^n \sum_{i=1}^{m_c} (x_{ci} - \mu_c) (x_{ci} - \mu_c)^T \quad \text{MINORE DI DISPERSIONE
NELL'ELABORAZIONE}$$

Dove

$$\mu = \frac{1}{N} \sum_{c=1}^n \sum_i x_{ci} \quad \text{MEAN}$$

$$\mu_c = \frac{1}{m_c} \sum_{i=1}^{m_c} x_{ci} \quad \text{MEAN OF THE CLASS}$$

ALESSIO PUNTO OVE CONSIDERARE I PUNTI PROFETATI, PER SEGUIMENTO CONSIDERO
LA PROIEZIONE RISPETTO A UN VETTORE UNIDIMENSIONALE w .

GOLF NUOVO W
X SEDILE SPARE
I NEL
OBBLIGATORI?

INIZIATIVA CONSIDERAR LA GLUOM MEAN E IN CLASS MEAN NELL'
DIREZIONE w

$$m = w^T \mu \quad m_c = w^T \mu_c$$

E POSSO CALCOLARE NUOVA BETWEEN E WITHIN CLASS COVARIANCE

$$S_B = \frac{1}{N} \sum_{c=1}^n m_c (w^T \mu_c - w^T \mu) (w^T \mu_c - w^T \mu)^T =$$

$$\frac{1}{N} \sum_{c=1}^n m_c w^T (\mu_c - \mu) (\mu_c - \mu)^T w =$$

$$\frac{1}{N} \sum_{c=1}^n w^T m_c (\mu_c - \mu) (\mu_c - \mu)^T w \quad \begin{matrix} w \text{ NUOVO VETTORE} \\ \text{DALLA SUMMA} \end{matrix}$$

$$= w^T \frac{1}{N} \left[\sum_{c=1}^C n_c (\mu_c - \mu) (\mu_c - \mu)^T \right] w = w^T S_{\text{D}} w$$

||
 S_{D}

$$S_{\text{W}} = \frac{1}{N} \sum_{c=1}^C \sum_{i=1}^{n_c} (w^T x_{ci} - w^T \mu_c) (w^T x_{ci} - w^T \mu_c)^T =$$

$$= \frac{1}{N} \sum_{c=1}^C \sum_{i=1}^{n_c} w^T (x_{ci} - \mu_c) [w^T (x_{ci} - \mu_c)]^T =$$

$$= \frac{1}{N} \sum_{c=1}^C \sum_{i=1}^{n_c} w^T (x_{ci} - \mu_c) (x_{ci} - \mu_c)^T w =$$

$$w^T \left[\frac{1}{N} \sum_{c=1}^C \sum_{i=1}^{n_c} (x_{ci} - \mu_c) (x_{ci} - \mu_c)^T \right] w = w^T S_{\text{W}} w$$

||
 S_{W}

W NON HA INDETERMINATA PRO AVERE: $n_m \neq 2$ (E QUINDI NON INDICARE UNA DIREZIONE
MA NON E' PIENO, X MIGLIORARE IL NASTRO DIO CHE W E' UNIDIMENSIONALE E
 $w^T S_{\text{D}} w$ E' UNO SCALARE MI SEMBRA CONSIDERABILE)

$$L(w) = \frac{w^T S_{\text{D}} w}{w^T S_{\text{W}} w} \times \text{MIGLIORARE QUESTA PARTE}$$

DISOLVA CONSIDERARE IL MASSIMO DI L(w) MASSIMIZZARE S(w) SINTESI QUINDI
TROVARE IL MASSIMO E POPOLO = AL VETTORE NUVO.

IL GRADIENTE (INOLTRE LA PENDOLA DELLA SPERONE MULTIDIMENSIONALE OTENNEVA GLOCCATO UNA
DIMENSIONE E CALCOLARO AVVENTURE LE ALTRE, SE E' > 0 LA FUNZIONE SI E' TESENDO, SE E' < 0
SI E' DIVERTITO GE = 0 A TROVARE IN UN NUVO / MASSIMO LOCALI (STAVO IN CIMA)

$$\text{Non: } \nabla_w \left(\frac{f(w)}{g(w)} \right) = \frac{g(w) \nabla_w f(w) - f(w) \nabla_w g(w)}{(g(w))^2}$$

$$\nabla_w (w^T S_0 w) = (S_0 + S_0^T) w = \text{dimo che } S_0 \text{ è simmetrico} \Rightarrow 2S_0 w$$

$$\begin{aligned} \nabla_w L(w) &= \frac{w^T S_w w \nabla_w (w^T S_w w) - w^T S_w w \nabla_w (w^T S_w w)}{(w^T S_w w)^2} = \\ &= \frac{2w^T S_w w S_{ow} - 2w^T S_{ow} S_w w}{(w^T S_w w)^2} = \frac{2S_{ow}}{(w^T S_w w)^2} - \frac{2w^T S_{ow} S_w w}{(w^T S_w w)^2} \end{aligned}$$

S_{ow} > 0

$$= 0 \Rightarrow S_{ow} = \frac{w^T S_{ow} S_w w}{w^T S_w w} = S_{ow} w = \lambda(w) S_w w =$$

$$\Rightarrow S_w^{-1} S_{ow} w = \lambda(w) w$$

QUESTO CI DICE CHE w È PIANO ALI AUTOVETTORI DI $S_w^{-1} S_{ow}$, POICHÉ SE È COSTRUITA L'EQUAZIONE VENNE SOLUZIONATA.

DATO CHE $S_w^{-1} S_{ow}$ PUÒ POSSERE AVERE + AUTOVETTORI QUALI È QUESTO CHE CI PERMETTE DI ESEGUIRE UNA MASSIMIZZAZIONE IN DIREZIONE DI CLUSTER E MINIMIZZARE IN DIREZIONE DI PUNTI NEL CLUSTER?

DATO CHE $\lambda(w) = \frac{w^T S_{ow}}{w^T S_w w}$ E NOI VOLIAMO MASSIMIZZARE QUESTO.

SCELGO CON w , L'ANALOGHEZZE CORRISPONDENTI ALL'ANNUALITÀ A PIÙ GRANDE.

Dopo ciò LDA POTREBBE ANCHEESSERE USATA PER
la dimensionality reduction e non solo come criterio di
separazione dei dati

CHE FACCIO? Dopo trovare il somspazio W sul quale
proiettare i dati

$$I punti proiettati sono \hat{x} = W^T x$$

LA PROIEZIONE BETWEEN & WITHIN CLASS COVARIANCE PER I PUNTI
PROIETTATI È PARI A

$$\hat{S}_B = W^T S_B W$$

$$\hat{S}_W = W^T S_W W$$

PER MASSIMIZZARE IL TUTTO POSSO USARE IL RATIO, DATO CHE NON HO PIÙ
DATI, SI USA

$$\lambda = \text{Tr}(S_W^{-1} \hat{S}_B)$$

QUINDI CONFRONTA IL RATIO, E DISOMA
TROVARE IL MASSIMO DI STA FUNZIONE

SI DISOMA CHE LA SOLUZIONE È DATA DAGLI m (right)
AUTOVETTORI CORRISPONDENTI ALLI ANNUALI MIGLIORI DI $S_W^{-1} S_B$

Dato che x costituisce S_B ha un numero di annui $\neq 0$
punti a $C-1$, dato che uno dei vettori di cui è composta può essere
escluso dalle combinazioni lineari degli altri.

DI CONSEGUENZA UN LDA SI PUÒ DIRE STANTE AL MENO C-1
OPZIONI. $\text{MIN}(S_B) = C-2$

E quindi avremo S_W^{-1}

Dato che S_W deve essere ben definita, con autovettori $\neq 0$, DEVE FARSI
CHE LE FEATURE SONO LINEARMENTE INDEPENDENTI, E QUINDI COMBINAZIONI DELLE STesse SONO "ELIMINATE"
E PER CIÒ PRIMA DELLA LDA DI SOLONO SI APPLICA UNA PIAZZA

SE NO LOA FALLISCE (PERCHÉ NON

Posso calcolare S_{W^2})