UNIVERSITÀ DEGLI STUDI DI PALERMO

dipartimento di ingegneria unipa

Preparazione dei dati

CORSO DI BIG DATA

a.a. 2020/2021

Prof. Roberto Pirrone

LABORATORIO DI INTERAZIONE UOMO-MACCHINA
CHILAB

Sommario

- Estrazione delle feature rilevanti
- Portabilità tra diverse tipologie di dati
- Data cleaning
- Riduzione della dimensionalità

Processo di data preparation

- Estrazione delle feature e portabilità
  - Si ricercano le feature più significative in relazione al problema da risolvere e, in genere, tali feature vanno *convertite nelle tipologie di dati* più adatte per l'utilizzo con gli algoritmi di analisi
- Data cleaning
  - Eliminazione di dati erronei e/o inconsistenti
  - *Imputazione* dei dati mancanti attraverso un processo di stima
- Data reduction & transofrmation
  - Riduzione del volume dei dati tramite *campionamento* di un sotto insieme, *riduzione della dimensionalità* dei campioni ovvero *trasformazione* degli stessi secondo una diversa rappresentazione

Portabilità tra tipologie di dati

Source data type	**Destination data type**	**Methods**
Numeric	Categorical	Discretization
Categorical	Numeric	Binarization
Text	Numeric	Latent semantic analysis (LSA)
Time series	Discrete sequence	SAX
Time series	Numeric multidimensional	DWT, DFT
Discrete sequence	Numeric multidimensional	DWT, DFT
Spatial	Numeric multidimensional	2-d DWT
Graphs	Numeric multidimensional	MDS, spectral
Any type	Graphs	Similarity graph (Restricted applicability)

Portabilità tra tipologie di dati

- Discretizzazione
  - Si creano degli intervalli discreti per rappresentare la variazione del dato numerico
    - Intervalli ad ampiezza costante
    - Intervalli logaritmici: ogni intervallo $[a,b] \rightarrow \log(b) - \log(a) = \text{costante}$
    - Più in generale $[a,b] \rightarrow f(b) - f(a) = \text{costante}$ per una certa $f(.)$ che rappresenta la distribuzione dei dati
    - Intervalli a «profondità» costante: ogni intervallo contiene lo stesso numero di elementi

Portabilità tra tipologie di dati

- Binarizzazione
  - Ogni categoria possibile induce la creazione di un *one hot vector* i cui componenti sono tutti nulli tranne quello in posizione corrispondente alla categoria

$$a \in \{c_1, c_2, \dots, c_\phi\} \to \mathbf{x} \in \mathbb{R}^\phi,$$

$$a = c_k \to \mathbf{x} = [0_1, 0_2, \dots, 1_k, \dots, 0_\phi]$$

Portabilità tra tipologie di dati

- Latent Semantic Analysis (LSA)
  - Un corpus documentale può essere visto come una matrice binaria sparsa $\mathcal{D}$ che ha come righe i termini e come colonne i documenti che li contengono.
  - La LSA è una forma di *Singular Value Decomposition* (SVD) di $\mathcal{D}$ che mira a creare uno spazio euclideo in cui un corpus documentale viene trasformato in un insieme di vettori in $\mathbb{R}^d$ dove è possibile giudicare la similarità tramite una misura di distanza.
  - Di norma gli elementi di $\mathcal{D}$ sono pesati con una misura di tipo frequentista chiamata *Term Frequency – Inverse Document Frequency* TF-IDF

Portabilità tra tipologie di dati

- TF-IDF
  - Prodotto della frequenza $f_{t,d}$ del termine t nel documento d appartenente a $\mathcal{D}$, avente dimensione N, per l'inverso della frequenza dei documenti che contengono il termine t e cioè N/n_t
  - Diversi schemi di pesatura:

$$\mathrm{tf}(t,d) = f_{t,d} \Bigg/ \sum_{t' \in d} f_{t',d}$$

$$\mathrm{idf}(t,\mathcal{D}) = \log\left(\frac{N}{n_t}\right)$$

$$\text{tf-idf}(t,d,\mathcal{D}) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,\mathcal{D}) = MI(\mathcal{T},\mathcal{D})$$

Portabilità tra tipologie di dati

- Symbolic Aggregate Approximation (SAX)
  - Si effettua la media dei valori della serie temporale su finestre di osservazione contigue di data ampiezza *w*
  - Si discretizza la serie risultante con intervalli a profondità costante
  - Il risultato è una serie discreta di simboli

  - Si assume una distribuzione gaussiana dei valori delle serie temporali mediate che viene stimata in termini di media e varianza ed i cui *quantili* forniscono gli estremi degli intervalli

Portabilità tra tipologie di dati

- Discrete Wavelet Transform (DWT) e Discrete Fourier Transform (DFT)
  - Usate per trasformare serie temporali in vettori multidimensionali di coefficienti che sono meno interdipendenti dei dati originali
  - Riduzione di dimensionalità
  - La DWT si applica anche alla trasformazione di dati spaziali in dati numerici multidimensionali

Portabilità tra tipologie di dati

- Discrete Wavelet Transform (DWT) e Discrete Fourier Transform (DFT)
  - Una sequenza discreta di simboli può essere trasformata in un insieme di sequenze binarie che descrivono la presenza di un solo simbolo per volta e poi questi ultimi sono trasformati con la DWT

```
ACACACTGTGACTG
10101000001000
01010100000100
00000010100010
00000001010001
```

Portabilità tra tipologie di dati

- DWT (Haar Wavelets)

  - Si assuma una sequenza temporale (t_i; x_i) avente lunghezza q che è potenza di 2 e suddivisa ricorsivamente in due metà S^i_k (i-esima metà a profondità di suddivisione k) fino ad arrivare a segmenti di lunghezza unitaria

  - $k = 0, 1, \ldots, \log_2(q)$
  - $i = 1, 2, \ldots, q/2^k$, per ogni k

Portabilità tra tipologie di dati

- DWT (Haar Wavelets)
  - Il coefficiente *i*-esimo della DWT è, per ogni livello di decomposizione *k*, la semi-differenza tra i valori medi di due sotto-sequenze adiacenti S^{2i}_{k+1} e S^{2i-1}_{k+1} al livello di suddivisione *k*+1

$$\psi_k^i = \left(\Phi_{k+1}^{2\cdot i-1} - \Phi_{k+1}^{2\cdot i}\right)/2$$
$$\Phi_k^i = \left(\Phi_{k+1}^{2\cdot i-1} + \Phi_{k+1}^{2\cdot i}\right)/2$$
$$\Phi_{\log_2 q}^i \equiv x^i$$

Portabilità tra tipologie di dati

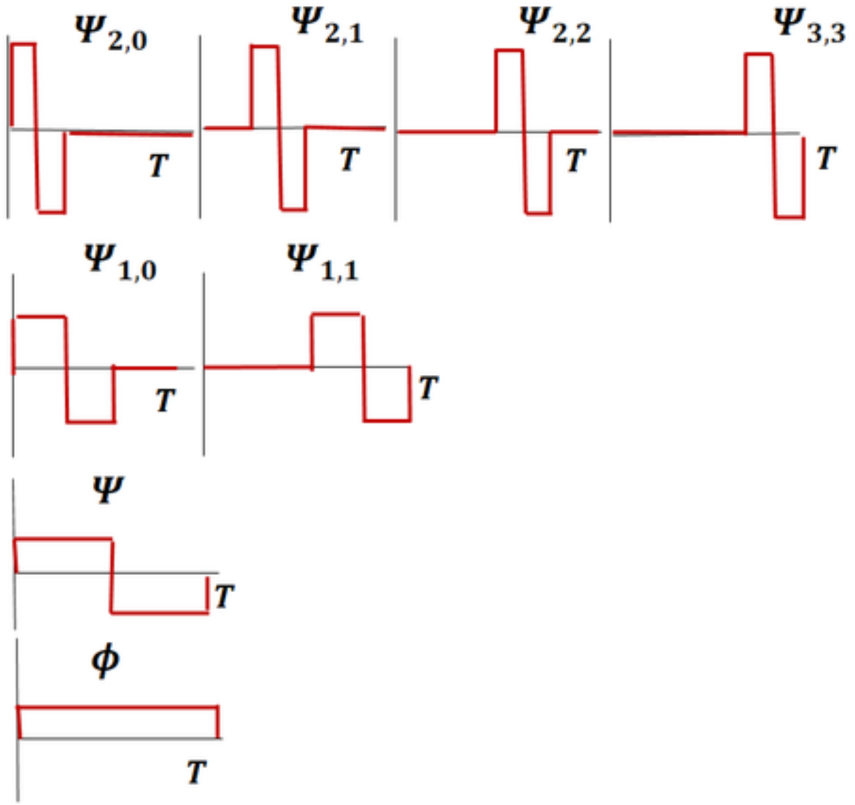
- DWT (Haar Wavelets)
  - Lo schema di calcolo è ricorsivo e parte calcolando le medie tra coppie di elementi successivi nella sequenza e poi andando a riaggregare le medie ad ogni livello fino a $k = 0$.

$$\psi_k^i = \left(\Phi_{k+1}^{2\cdot i-1} - \Phi_{k+1}^{2\cdot i}\right)/2$$
$$\Phi_k^i = \left(\Phi_{k+1}^{2\cdot i-1} + \Phi_{k+1}^{2\cdot i}\right)/2$$
$$\Phi_{\log_2 q}^i \equiv x^i$$

Portabilità tra tipologie di dati

- DWT (Haar Wavelets)

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$



Queste funzioni di base vanno bene per una sequenza di otto elementi

Portabilità tra tipologie di dati

- DWT (Haar Wavelets)
  - Il vettore **ψ** = DWT(**x**) e ha lunghezza *q*
  - La ricostruzione avviene attraverso la matrice dei vettori di base wavelet *W*

$$\boldsymbol{\psi} = \left[\psi^1_{\log_2 q-1}, \ldots, \psi^{q/2}_{\log_2 q-1}, \ldots, \psi^1_0, \phi^1_0\right]$$

$$W = \begin{bmatrix} \overline{W_1} \\ \vdots \\ \overline{W_q} \end{bmatrix}$$

$$\mathbf{x} = \sum_{i=1}^{q} \psi_i \cdot \overline{W_i}$$

Portabilità tra tipologie di dati

- Multidimensional Scaling (MDS)
  - Funziona per grafi pesati in cui il peso δ_{ij} tra due nodi abbia un significato di distanza o similarità tra essi.
  - Si assuma di voler rappresentare ogni nodo $i = 1, \ldots, n$ con un embedding k-dimensionale $X_i \in \mathbb{R}^k$
  - Assumendo di conoscere tutte le $\delta_{ij} = \delta_{ji}$ che sono $\binom{n}{2}$ si minimizzi il seguente funzionale

$$O = \sum_{i,j:i<j} \left(\| \overline{X_i} - \overline{X_j} \| - \delta_{ij} \right)^2$$

Portabilità tra tipologie di dati

- Generazione di grafi di similarità
  - Può essere conveniente cercare similarità a coppie per determinate applicazioni e alcuni algoritmi di analisi se ne avvantaggiano
  - Ogni oggetto O_i del data set è considerato un nodo multidimensionale
  - Se $d(O_i,O_j)$ è minore di una certa soglia si genera un arco (i,j) che viene pesato con un peso w_{ij} generato attraverso l'applicazione di un opportuno kernel

$$w_{ij} = e^{-\frac{d(O_i,O_j)^2}{t^2}}$$

Heat kernel

Data cleaning

- Gestione dei dati mancanti
  - Eliminazione dei record incompleti
  - Stima o imputazione dei dati mancanti
    - Può inficiare le prestazioni dell'algoritmo di analisi
  - Uso di algoritmi di analisi robusti rispetto ai dati mancanti
    - Classificazione come metodo di imputazione della feature mancante come «special feature» del dato rispetto alla quale appunto si classifica
    - Utility matrix imputation nei recommender systems

Data cleaning

- Gestione dei dati mancanti

	GLADIATOR	GODFATHER	BEN-HUR	GOODFELLAS	SCARFACE	SPARTACUS
U_1	1			5		2
U_2		5			4	
U_3	5	3		1		
U_4			3			4
U_5				3	5	
U_6	5		4			

(a) Ratings-based utility

	GLADIATOR	GODFATHER	BEN-HUR	GOODFELLAS	SCARFACE	SPARTACUS
U_1	1			1		1
U_2		1			1	
U_3	1	1		1		
U_4			1			1
U_5				1	1	
U_6	1		1			

(b) Positive-preference utility

Data cleaning

- Eliminazione dei dati inconsistenti o erronei
  - Analisi di dati inconsistenti tra flussi in arrivo da sorgenti diverse
    - Es. «John Fitzgerald Kennedy» vs. «JFK»
  - Uso della conoscenza di dominio per eliminare le inconsistenze
  - Analisi statistica dei dati per individuare il trend ovvero esplicito utilizzo di algoritmi di outlier detection

Data cleaning

- Scalatura e normalizzazione
  - Scalatura min-max

$$y_i^j = \frac{x_i^j - \min_j}{\max_j - \min_j}$$

  - Z-scaling

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j}$$

Riduzione della dimensionalità dei dati

- Campionamento
  - Campionamento statistico per ridurre la dimensionalità di un data set
  - Campionamento di sequenze
  - Campionamento «ottimo» da una distribuzione per la stima di una certa quantità desiderata
- Selezione di sottoinsiemi di feature rilevanti
- Riduzione della dimensionalità per *rotazione degli assi*
- Riduzione della dimensionalità per trasformazione della tipologia dei dati

Riduzione della dimensionalità dei dati

- Campionamento
  - Il campionamento può essere «stratificato» nel senso che si preferisce ricoprire il dominio della variabile con una serie di intervalli o *strati* e garantire la presenza di campioni in ogni strato

  - L'altra tecnica di campionamento è quella «per importanza» nella quale alcune parti dei dati sono certamente più rilevanti di altre e quindi esiste una distribuzione $p(x)$ che descrive la probabilità di campionare un certo valore x.

  - La probabilità $p(x)$ ottima per eccellenza è la stessa distribuzione di probabilità dei dati (anche se non la conosciamo)

Riduzione della dimensionalità dei dati

- Campionamento
  - Le sequenze sono campionate attraverso la creazione di un «magazzino» di campioni di dimensione *k*
  - Il campione *n*-esimo viene inserito nel magazzino con probabilità $p(x_n)=k/n$
  - Si elimina casualmente uno dei vecchi dati per far posto al nuovo campione, tenendo conto della probabilità
  - Si può dimostrare che la probabilità di essere inseriti nel magazzino dopo che sono arrivati *n* campioni è sempre *k*/*n*

Riduzione della dimensionalità dei dati

- Selezione di sottoinsiemi di feature rilevanti
  - tramite approccio non supervisionato → clustering **
  - tramite approccio supervisionato → classificazione

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi - PCA
  - La matrice di covarianza di una data set *D* di *n* record di dimensione *d* è

$$C = \frac{D^T D}{n} - \bar{\mu}^T \bar{\mu}$$

  - I suoi elementi si possono esprimere come

$$c_{ij} = \frac{\sum_{k=1}^{n} x_k^i x_k^j}{n} - \mu_i \mu_j \quad \forall i, j \in \{1 \dots d\}$$

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi - PCA

  - Per qualunque vettore $v \in \mathbb{R}^d$ è possibile calcolare la varianza unidimensionale della proiezione dei dati su *v*, *Dv* come:

$$\bar{v}^T C \bar{v} = \frac{(D\bar{v})^T D\bar{v}}{n} - (\bar{\mu}\bar{v})^2$$

  - Cerchiamo la base ortonormale di vettori su cui proiettare i dati per massimizzare la loro varianza

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi - PCA
  - Il problema di ottimizzazione può essere affrontato con i moltiplicatori di Lagrange

$$\bar{v}^T C \bar{v} + \lambda \left(1 - \|\bar{v}\|^2\right)$$

  - Se poniamo il gradiente del funzionale pari a 0, otteniamo che la base di vettori è costituita dagli autovettori di C e le varianze sono i corrispondenti autovalori

$$C\bar{v} = \lambda\bar{v} = 0 \qquad \bar{v}^T C \bar{v} = \bar{v}^T \lambda \bar{v} = \lambda$$

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi - PCA
  - C è semidefinita positiva e può essere diagonalizzata

$$C = P\Lambda P^T$$

  - P è la matrice degli autovettori di C, $P^T P = I$, mentre Λ è la matrice diagonale degli autovalori di C
  - Possiamo riordinare le righe di P in senso decrescente dal massimo al minimo autovalore (varianza dei dati lungo il corrispondente autovettore)

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi - PCA
  - Si può mostrare che i dati trasformati *D’ = DP*, con media *μP*, hanno una matrice di covarianza *Λ* :

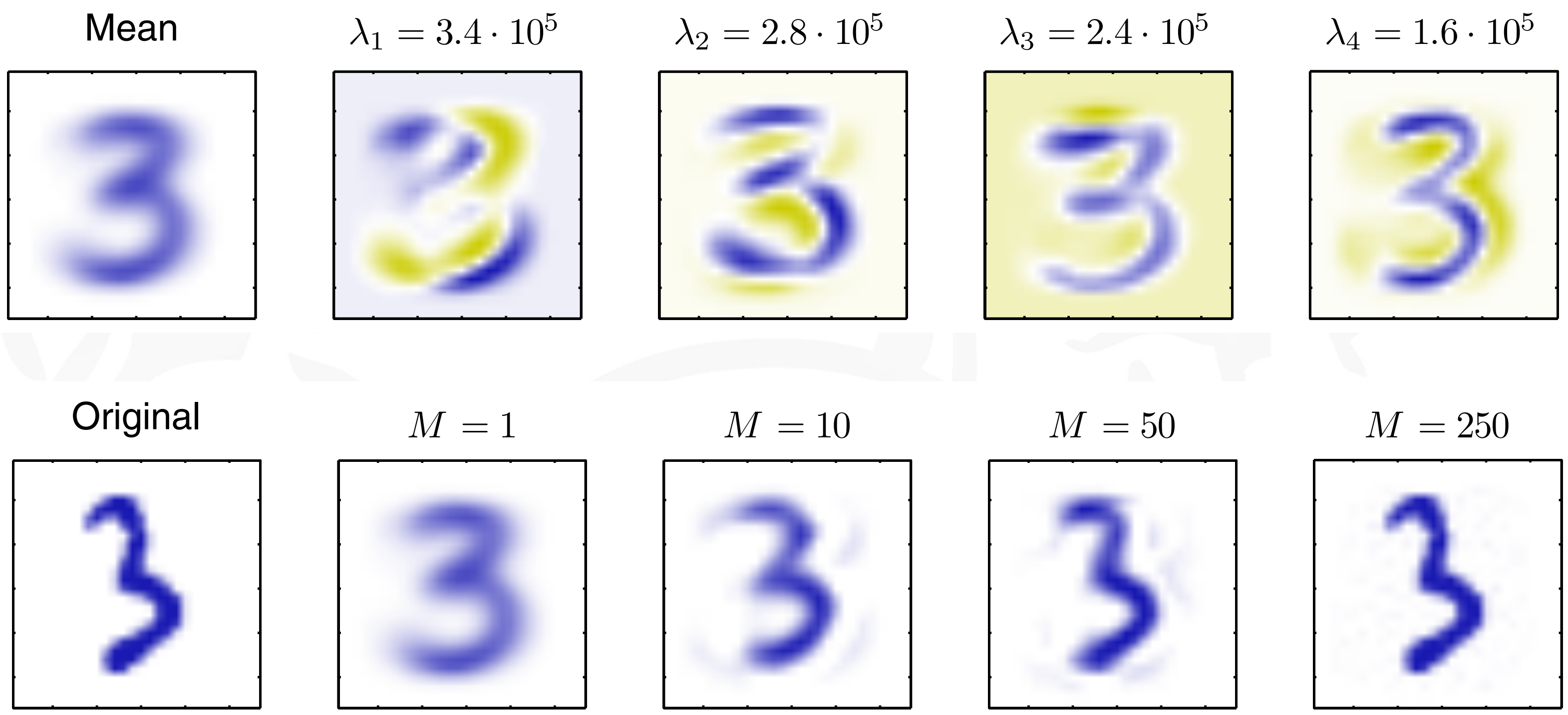
$$\frac{(DP)^T DP}{n} - (\bar{\mu}P)^T(\bar{\mu}P) =$$

$$\frac{P^T D^T DP}{n} - P^T \bar{\mu}^T \bar{\mu} P =$$

$$P^T C P \qquad\qquad P^T C P = P^T P \Lambda P^T P = I \Lambda I = \Lambda$$

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi - PCA



Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi – SVD
  - È una fattorizzazione della matrice dei dati di dimensione *n* x *d*

$$D = Q\Sigma P^T$$

  - *Q* ha dimensione *n* x *n* e le sue colonne formano una base ortonormale denominata *left singular vectors* → $Q^TQ=I$
  - *Σ* è una matrice diagonale di dimensione *n* x *d* e contiene i *singular values*, non negativi; gli elementi delle righe oltre *n* sono tutti nulli. I singular values sono le «variabili latenti» di questa rappresentazione dei dati
  - *P* ha dimensione *d* x *d* e le sue colonne formano una base ortonormale denominata *right singular vectors* → $P^TP=I$

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi – SVD

$$DD^T = Q\Sigma(P^TP)\Sigma^TQ^T = Q(\Sigma\Sigma^T)Q^T$$

Q → autovettori di DD^T
$\Sigma\,\Sigma^T$ → autovalori di DD^T

$$D^TD = P\Sigma^T(Q^TQ)\Sigma P^T = P(\Sigma^T\Sigma)P^T$$

P → autovettori di D^TD
$\Sigma^T\Sigma$ → autovalori di D^TD

$$\boldsymbol{\mu} = \mathbf{0} \Rightarrow C = \frac{D^TD}{n}$$

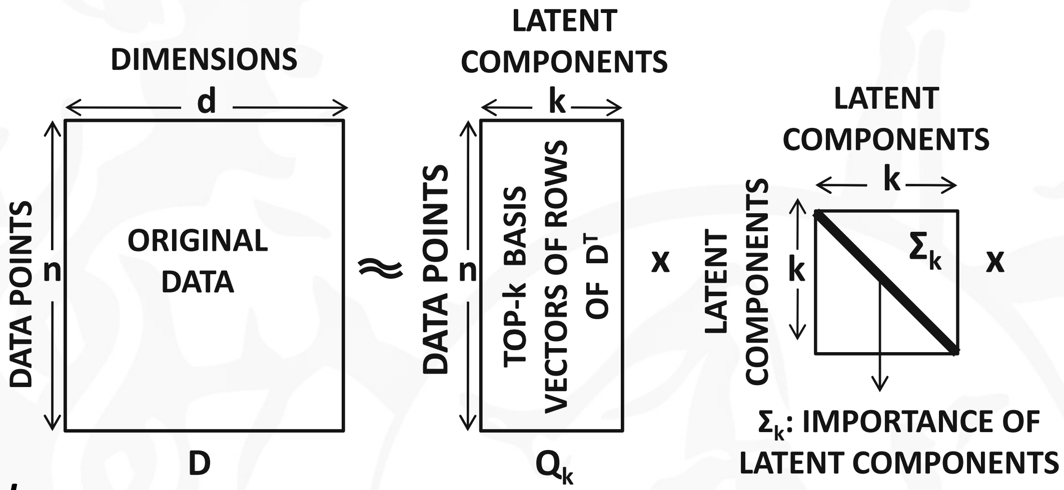
P → autovettori di *C*
$\Sigma^T\Sigma$ → autovalori di *C* moltiplicati per *n*

SVD e PCA sono la stessa rotazione degli assi per dati centrati rispetto alla propria media

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi – SVD
  - La SVD troncata rappresenta il data set con i primi *k* valori singolari

$$D \approx Q_k \Sigma_k P_k^T$$



  - $k << n,d$
  - Le d - k colonne restanti dei dati ricostruiti $D'=DP$ sono effettivamente nulle

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi – SVD
  - La SVD troncata rappresenta il data set con i primi *k* valori singolari

$$D \approx Q_k \Sigma_k P_k^T$$

  - Richiamando le considerazioni sulla PCA che massimizza la varianza (somma dei quadrati delle distanze euclidee dalla media) la *SVD massimizza l'energia associata ai dati* in termini di somma dei quadrati delle distanze dall'origine che si ottiene da $D^T D$
  - Tale energia, nel caso troncato, è data dalla *somma dei quadrati dei primi k valori singolari*; dato un right singular vector *v* e il corrispondente valore singolare σ:

$$(D\bar{v})^T (D\bar{v}) = \bar{v}^T \left(D^T D\right) \bar{v} \equiv \bar{v}^T \bar{v} \sigma^2 \bar{v} \bar{v}^T = \sigma^2$$

Riduzione della dimensionalità dei dati

- Riduzione della dimensionalità per rotazione degli assi – LSA
  - SVD troncata applicata ad una matrice D di valori TF-IDF relativi all'occorrenza di n termini globalmente in d documenti
  - D è molto sparsa e di elevatissime dimensioni
  - Non viene fatta la normalizzazione rispetto alla media perché la sparsità garantisce una matrice di covarianza comunque approssimativamente proporzionale a $D^T D$

Riduzione della dimensionalità dei dati

- Utilizzo della Haar Wavelet Transform
  - La ricostruzione di una serie temporale si può riscrivere

$$T = \sum_{i=1}^{q} a_i \overline{W_i} = \sum_{i=1}^{q} \left(a_i ||\overline{W_i}||\right) \frac{\overline{W_i}}{||\overline{W_i}||}$$

  - I vettori di base normalizzati sono ortonormali
  - Si può dimostrare che troncando la HWT con i primi *k* coefficienti normalizzati si minimizza l'errore di ricostruzione

Riduzione della dimensionalità dei dati

- Multi Dimensional Scaling (MDS)

$$O = \sum_{i,j:i<j} \left(\| \overline{X_i} - \overline{X_j} \| - \delta_{ij} \right)^2$$

  - Piuttosto che effettuare una minimizzazione si affronta il problema come decomposizione SVD
  - Nota la matrice $\Delta = [\delta_{ij}^2]_{nxn}$, cerchiamo la matrice $D = [X_i]_{nxk}$ dei dati che abbia Δ come insieme delle distanze e tale che $k \ll n$
  - $X_i \rightarrow$ *embedding* k-dimensionale del nodo *i*

Riduzione della dimensionalità dei dati

- Multi Dimensional Scaling (MDS)
  - Dalla legge del coseno ($a^2 = b^2 + c^2 - 2bc\cos(\alpha)$) estesa ai vettori n-dimensionali possiamo ricavare la matrice dei prodotti scalari $S = [S_{ij} \triangleq X_i \cdot X_j]$ a partire da Δ

$$\overline{X_i} \cdot \overline{X_j} = -\frac{1}{2}\left[\left\|\overline{X_i} - \overline{X_j}\right\|^2 - \left(\left\|\overline{X_i}\right\|^2 + \left\|\overline{X_j}\right\|^2\right)\right] \quad \forall i, j \in \{1 \ldots n\}$$

$$S = -\tfrac{1}{2}(I - U/n)\Delta(I - U/n) \equiv DD^T$$

  - U è la matrice di tutti 1 di ordine n

Riduzione della dimensionalità dei dati

- Multi Dimensional Scaling (MDS)
  - La SVD troncata di $S \approx Q_k \Sigma^2{}_k Q_k{}^T$ fornisce una sua fattorizzazione per ricavare D

$$S \approx Q_k \Sigma_k^2 Q_k^T = \left(Q_k \Sigma_k\right)\left(Q_k \Sigma_k\right)^T$$

$$D = Q_k \Sigma_k$$

Riduzione della dimensionalità dei dati

- Trasformazione spettrale per embedding di grafi
  - Un grafo non orientato G=(N,A) è caratterizzato da una matrice di pesi W=[w_{ij}] che esprimono *similarità* (non distanze) tra i nodi → etichette degli archi
  - Ricerchiamo l'embedding di nodi D = $[X_i]_{nxk}$ che minimizzi

$$O = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left\| \bar{X}_i - \bar{X}_j \right\|^2$$

  - Useremo la matrice laplaciana L di G

Riduzione della dimensionalità dei dati

- Trasformazione spettrale per embedding di grafi
  - Vediamo il caso monodimensionale cioè embedding scalare per ogni nodo

$$O = \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} w_{ij} \left(y_i - y_j\right)^2$$

Matrice Laplaciana

$$L = \Lambda - W, \ \Lambda_{ii} = \Sigma_{j=1}^{n} w_{ij}$$

$$O = 2\bar{y}^T L \bar{y}, \ \bar{y}^T \Lambda \bar{y} = 1, \ \bar{y} = (y_1, \ldots y_n)^T$$

$$\Lambda^{-1} L \bar{y} = \lambda \bar{y}$$

- La soluzione è data dai k più piccoli autovettori della matrice $\Lambda^{-1}L$