Università degli Studi Di Padova

Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in Statistica per le Tecnologie e le Scienze

---

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

RELAZIONE FINALE

Approccio bayesiano all'analisi delle componenti principali

Relatore:
Prof.ssa Manuela Cattelan

Laureando:
Chiara Boraso
Matricola: 1227357

Anno Accademico 2021/2022

Introduzione

L'analisi delle componenti principali è una tecnica per ridurre la dimensionalità di dataset multivariati cogliendo più informazione possibile. Nel primo capitolo verranno quindi trattati dal punto di vista teorico 3 approcci per effettuarla sui dati: classico, probabilistico e Bayesiano. Ognuno di questi approcci ha dei vantaggi, degli svantaggi e dei contesti in cui è più efficace degli altri.
Nel secondo capitolo questi 3 approcci verranno applicati a 3 dataset diversi, ognuno con le sue particolarità, per osservare i risultati che si ottengono. In questa parte verrà evidenziata la capacità dell'approccio probabilistico e Bayesiano di fare stime consistenti per i dati mancanti. Questa funzionalità è evidenziata soprattutto con il primo dataset, **Metabolite**, relativo ai metaboliti prodotti dal corpo umano, in cui sono stati simulati dei dati mancanti dal dataset completo. Il dataset **MNIST** mostra l'efficacia della tecnica con applicazioni con molte variabili infatti ogni osservazione corrisponde a una cifra numerica scritta a mano trasformata in un quadrato di 28×28 pixel. L'ultimo dataset **Ozone** con dati mancanti reali serve a valutare se gli approcci riescono a cogliere le correlazioni tra i dati in questo caso relativi ai livelli di Ozono nell'aria con diverse condizioni atmosferische.

Indice

Capitolo 1

Approccio classico e Bayesiano all'analisi delle componenti principali

1.1 Analisi delle componenti principali

L'analisi delle componenti principali è una tecnica statistica per ridurre la dimensionalità di dataset mulivariati.
Il criterio di riduzione è basato sulla matrice di varianza e covarianza dei dati, per cui può essere applicato solo a variabili quantitative di cui è possibile calcolare la varianza. Il dataset a cui viene applicata è visualizzabile come una matrice $n \times p$ dove n corrisponde al numero di unità statistiche e p al numero di variabili.
Algebricamente si tratta di una specifica combinazione lineare di p variabili casuali del vettore $X = (X_1, X_2,..., X_p)$ che lo trasporta in un nuovo sistema di coordinate. I nuovi assi rappresentano la direzione con massima variabilità e forniscono una descrizione più parsimoniosa della struttura della covarianza $Var(X) = \Sigma$. La trasformazione lineare deve rispettare le seguenti proprietà:

- Le nuove variabili siano funzioni lineari di quelle di partenza;

- Le nuove variabili siano incorrelate tra loro.

Ridurre la dimensionalità significa che un numero $k < p$ di variabili tra le p variabili Y_1, Y_2,..., Y_p , combinazioni lineari di X, deve contenere più informazione possibile delle variabili del dataset di partenza. In particolare la prima componente è la combinazione lineare $a_i'X$ che massimizza $V(Y_i) = V(a_1'X) = a_1'\Sigma a_1$.

Teorema 1. *Data la matrice di covarianza Σ associata al vettore $X' = [X_1, X_2...X_p]$ con coppie di autovalori e autovettori di Σ uguali rispettivamente a (λ_1, e_1), (λ_2, e_2),..., (λ_p, e_p), dove $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, la $i-$esima componente principale è:*

$$Y_1 = e_i'X = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p, i = 1, 2, ..., p.$$

Allora:

$$\begin{array}{rcccl} Var(Y_i) & = & e_i'\Sigma e_i & = & \lambda_i \quad i = 1, 2, ..., p, \\ Cov(Y_i, Y_k) & = & e_i'\Sigma e_k & = & 0 \qquad i \neq k. \end{array} \tag{1.1}$$

Se vi sono dei λ_i uguali la scelta dei corrispondenti e_i non è unica.
[Johnson e Wichern, 2012]

Dal precedente risultato, si può dire che la proiezione lineare ottima per cui la varianza della matrice dei dati proiettata è massimizzata, è definita dai k autovalori $e_1, ..., e_k$ della matrice di covarianza Σ che corrispondono ai k più grandi autovalori $\lambda_1, ..., \lambda_k$. Si ricavano quindi alcune utili proprietà:

1. le varianze $V(Y_i)$ coincidono con gli autovalori. Da cui consegue che

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} V(X_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1} Var(Y_i),$$

   dove $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp}$ sono le varianze delle p-variabili di X ossia i valori che si trovano nella diagonale di Σ;

2. la correlazione ρ_{ik} tra Y_i e X_k è

$$\rho_{ik} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}.$$

Il segno della correlazione tra una componente principale e una variabile è determinato dall'autovettore corrispondente;

3. il rapporto
$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$
rappresenta la proporzione di varianza spiegata dalla k-esima componente;

4. la PCA non è invariante a standardizzazione; gli autovalori e gli autovettori calcolati sulla matrice di covarianza Σ o sulla matrice di correlazione ρ differiscono tra loro.

Standardizzazione

Le componenti principali possono essere ottenute anche da variabili standardizzate T di media nulla e varianza unitaria, calcolando gli autovettori e gli autovalori sulla matrice di correlazione. I risultati che si ottengono spesso non sono uguali poichè la varianza cambia in base all'unità di misura utilizzata per le variabili. Per cui se c'è molta differenza, la prima componente rappresenterà solo la variabile su scala maggiore. Abbiamo un ulteriore vantaggio interpretativo: la proporzione di varianza spiegata dalla $k-$sima componente diventerà
$$\frac{\lambda_k}{p}.$$

Numero di componenti

Uno degli elementi più importanti di questa tecnica è la scelta del numero di componenti principali, ossia trovare un numero $k < p$ abbastanza piccolo che non faccia perdere troppa informazione importante ma che comunque produca una significativa riduzione. Per esempio se si prendesse $k = 1$, la riduzione è portata al massimo ma si perderebbe probabilmente informazione importante. Per una prima idea, si può utilizzare un metodo grafico: lo scree plot. Si tratta di un grafico che ordina gli autovalori dal più grande al più

piccolo e che serve ad individuare il punto in cui gli autovalori si livellano e sono molto piccoli con valori simili. Il numero di componenti scelto sarà quindi uguale agli autovalori che si trovano prima di questo punto. Ci sono poi alcuni metodi numerici:

- scegliere una soglia (solitamente tra il 75% e l'85%) della percentuale cumulata della varianza totale che viene spiegata dalle componenti principali. Il numero di componenti usato è il più piccolo numero che permette di superare questa soglia;
- osservare la varianza spiegata dal singolo autovalore e tenere solo quelli che spiegano almeno più di una certa soglia;
- realizzare un algoritmo di convalidazione incrociata che consiste nel dividere il dataset in m campioni ed escludere iterativamente un gruppo alla volta. Sugli $m-1$ gruppi rimanenti si stima il modello e si utilizza il gruppo escluso per valutare l'adattamento con l'errore quadratico medio. Questa procedura va effettutata per tutti gli m gruppi per valutare il modello migliore. Lo scopo in questo caso è individuare il numero di componenti che riesce a cogliere la varianza tra le variabili evitando un sovradattamento.

1.2 PCA probabilistica

L'analisi probabilistica delle componenti pincipali (PPCA) è una riformulazione del metodo classico in cui il risultato è espresso come la stima di massima verosomiglianza per il modello probabilistico di una variabile casuale latente. Il metodo è stato proposto per la prima volta indipendentemente da Tipping e Bishop (1997) e da Roweis (1998).
Per prima cosa viene introdotta una variabile latente esplicita Z di dimensione k che rappresenta il sottospazio delle componenti principali per cui si definisce una distribuzione gaussiana a priori con media nulla:

$$p(Z) \sim N(Z|0; I).$$

Poi si definisce una distribuzione gaussiana per le variabili osservate x condizionate ai valori z di Z:

$$p(x|z) \sim N(x|Wz + \mu; \sigma^2 I),$$

dove la variabile osservata x di dimensione p è definita come una trasformazione lineare

$$x = Wz + \mu + \epsilon,$$

con W una matrice $p \times k$ e ϵ una variabile gaussiana di disturbo con vettore delle medie nullo e matrice di covarianza $\sigma^2 I$.
Per calcolare una stima dei parametri usando il metodo di massima verosimiglianza è necessario conoscere la distribuzione marginale di x che è a sua volta una gaussiana

$$p(x) \sim N(x|\mu; C),$$

dove C è una matrice $p \times p$ definita come

$$C = WW^T + \sigma^2 I.$$

Si può anche ottenere la distribuzione a posteriori di z:

$$p(z|x) \sim N(z|M^{-1}W^T(x - \mu); \sigma^{-2}M),$$

dove M è una matrice $k \times k$ definita come

$$M = W^T W + \sigma^2 I.$$

1.2.1 Stime di massima verosimiglianza

Ora è quindi possibile calcolare le stime dei parametri W, μ, σ^2 che regolano x:

$$\begin{aligned} \ln p(X|\mu, W, \sigma^2) &= \sum_{i=1}^{n} \ln p(x_i|\mu, W, \sigma^2) \\ &= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|C| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T C^{-1}(x_i - \mu). \end{aligned} \tag{1.2}$$

Le stime che si ottengono sono:

1. $\hat{\mu}$=$\bar{x}$ ossia la media campionaria delle x_i, realizzazioni di X_i;

2. la matrice dei pesi
$$\hat{W} = U_k(L_k - \sigma^2 I)^{\frac{1}{2}} R$$
dove U_k è una matrice $p \times k$ le cui colonne sono un qualsiasi sottoinsiem di autovettori della matrice di covarianza Σ, L_k è una matrice $k \times k$ dei corrispondenti autovalori λ_k e R è una matrice arbitraria sempre di dimensione $k \times k$;

3. la varianza dell'errore
$$\hat{\sigma}^2 = \frac{1}{p-k} \sum_{i=k+1}^{p} \lambda_i,$$
che corrisponde alla media delle varianze associate ai parametri scartati dalla riduzione. Le stime così esposte sono presnti nel capitolo 12 di [Bishop, 2006]

Algoritmo EM

Le stime possono essere trovate anche utilizzando l'algoritmo iterativo Expectation-Maximisation. In generale, lo scopo dell'algoritmo EM è quello di aumentare, e possibilmente di massimizzare, la verosimiglianza dei parametri di un modello probabilistico rispetto ad un insieme di dati, risultati di un processo stocastico che coinvolge un processo non noto che in questo caso corrisponde alla distribuzione della variabile latente z.
L'algoritmo è composto da 2 passi che si applicano iterativamente fino al raggiungimento della convergenza quando l'aggiornamento dei parametri non incrementa più la verosimiglianza.

- Passo E: definisce il valore atteso della verosimiglianza rispetto alla distribuzione di densità condizionata $p(z|x)$ utilizzando le stime correnti dei parametri.

- Passo M: viene massimizzato il valore atteso trovato al passo precedente rispetto a W e σ^2 determinando i valori che lo massimizzano da usare come partenza per il passo E successivo.

L'algoritmo EM è particolarmente efficace quando abbiamo dataset di elevata dimensionalità o dati mancanti. In quest'ultimo caso infatti si può comunque usare la distribuzione marginale delle variabili non osservate.

1.2.2 Interpretazione del modello

La colonne della stima $\hat{W}$, quando $R = I$, corrispondono agli autovettori di X scalati da $\lambda_i - \sigma^2$. La varianza del modello nella direzione dell'i-esimo autovettore e_i, per convoluzioni di distribuzioni normali indipendenti, è la somma dei contributi $\lambda_i - \sigma^2$ dati dalla proiezione dallo spazio latente allo spazio dei dati di partenza della colonna di W corrispondenti e σ^2 che è la media di tutti gli autovalori scartati.
Si può quindi dire che il modello coglie la varianza lungo i k assi principali mentre nelle restanti direzioni la varianza è approssimata dal valore medio σ^2.

Identificabilità

Esiste però un problema nella determinazione di $p(x)$ poichè esiste un intera famiglia di matrici $\widetilde{W} = WR$ (con R matrice ortogonale) che producono la stessa distribuzione predittiva per x. Infatti, sfruttando la proprietà delle matrici ortogonali $RR^T = I$, si vede che:

$$\widetilde{W}\widetilde{W}^T = WRR^TW^T = WW^T.$$

Si nota però che la matrice C di covarianza è indipendente dalle rotazioni su W e quindi non viene modificata dalle rotazioni.

Gradi di libertà

L'analisi probabilistica delle componenti principali serve anche a ridurre i gradi di libertà di una gaussiana multivariata cogliendo comunque la correlazione tra le variabili. In generale la matrice di covarianza di una gaussiana multivariata ha $p(p+1)/2$ parametri indipendenti. Quando p è molto grande, il costo computazionale diventerebbe troppo elevato e sarebbe necessario considerare solo la matrice diagonale perdendo tutta la correlazione tra le variabili.
Con la PPCA si riesce a ridurre i gradi di libertà senza perdere l'informazine che non è sulla diagonale. La matrice C della varianza di x nella PPCA dipende da W con dimensione $p \times k$ e da σ^2 e per cui ha $pk + 1$ parametri. Inoltre nel paragrafo precedente abbiamo visto che è presente una ridondanza dovuta alla matrice ortogonale R di dimensioni $k \times k$. Ogni colonna i ha $k - i$ parametri incogniti con i=1, .., k. Per esempio la prima colonna, avrà $k - 1$ parametri indipendenti e così via per le successive. In totale la matrice R ha $k(k-1)/2$ parametri, di conseguenza i gradi di libertà della matrice C sono:

$$pk + 1 - \frac{k(k-1)}{2}$$

e il numero di parametri cresce linearmente con p, per k fissato.

1.3 PCA Bayesiana

Fino a ora è stato assunto un valore fissato k per il sottospazio latente delle componenti principali. Dopo aver definito un approccio probabilistico per le PCA, si può usare anche un approccio bayesiano.

Approccio bayesiano

Si tratta di un metodo di statistica inferenziale basato sul teorema di Bayes. L'obbiettivo è calcolare la distribuzione a posteriori dei parametri date le osservazioni partendo da:

- la distribuzione a priori dei parametri prima dell'osservazione dei dati;
- la funzione di verosimiglianza della distribuzione dei dati osservati condizionata ai parametri.

Distribuzione a priori

Per prima cosa sarebbe necessario ipotizzare una distribuzioni a priori per i parametri μ, W e σ^2. Nel nostro caso per semplicità trattiamo μ e σ^2 come parametri fissati da stimare e definiamo una distribuzione a priori gaussiana indipendente per ciascuna colonna di W, dove la varianza di ciascuna gaussiana dipende inversamente dall'iperparametro α_i tale per cui:

$$p(W|\alpha) = \prod_{i=1}^{k} \Big(\frac{\alpha_i}{2\pi}\Big)^{p/2} exp\Big\{-\frac{1}{2}\alpha_i w_i^T w_i\Big\}, \tag{1.3}$$

dove w_i è la i-esima colonna di W.
I valori di α_i sono trovati iterativamente dalla massimizzazione della verosimiglianza, integrata rispetto a W. Può risultare che alcuni α_i tendano a infinito e, visto il legame inverso della varianza, i w_i tendono a 0. Allora la PCA Bayesiana effettua automaticamente la scelta del numero di componenti considerando solo i vettori w_i corrispondenti ad α_i finiti, cercando un compromesso tra un buon adattamento e la complessità del modello.

Funzione di Verosimiglianza

Il secondo passo dell'approccio bayesiano è stimare W dalla log-verosimiglianza della distribuzione a posteriori di X condizionata ai parametri:

$$\ln p(W|X) = L - \frac{1}{2}\sum_{i=1}^{p-1} \alpha_i \|w_i\|^2 + const \tag{1.4}$$

dove L è la funzione verosimiglianza di $p(x)$ definita in formula 1.2.

Dalla massimizzazione della verosimiglianza marginale rispetto a W

$$p(X|\alpha, \mu, \sigma^2) = \int p(X|W, \mu, \sigma^2) p(W|\alpha) dW \tag{1.5}$$

si determinano gli α_i. L'integrale che si ottiene non è risolvibile analiticamente; si utilizza quindi l'approssimazione di Laplace. Le stime di α aggiornate con la distribuzione a posteriori sono :

$$\alpha_i^{agg} = \frac{p}{w_i^T w_i'}.$$

Gli α_i^{agg} possono essere utilizzati nell'algoritmo EM come valori di partenza nel passo E per trovare i valori di W.
La PCA bayesiana si dimostra essere un metodo funzionale per trovare il numero adatto di componenti da utilizzare perchè per dataset finiti individuerà i corretti gradi di libertà scartando quelle variabili per cui non c'è sufficiente supporto dai dati. La varianza nelle colonne scartate è definita dal singolo parametro σ^2; in modeli più complessi si può decidere anche di ipotizzare una distribuzione a priori anche per questo parametro.

1.4 Analisi fattoriale

L'analisi fattoriale è un metodo utilizzato di solito per cogliere quei fenomeni che non sono direttamente osservabili sulle x con delle variabili latenti z (fattori). Si può quindi vederla come un'estensione della PPCA; l'unica differenza è nella matrice di covarianza della distribuzione normale condizionata di x rispetto a z:

$$p(x|z) = N(x|Wz + \mu, \Psi),$$

dove Ψ è una matrice diagonale $p \times p$ che significa che si può avere una quantità di disturbo diversa in ogni direzione. Ciascuna valore della diagonale (fattore specifico) della matrice Ψ corrisponde alla varianza di una delle variabili mentre W, le cui colonne sono chiamate pesi fattoriali, coglie la correlazione tra le variabili. Si conclude che quando $\Psi \to \sigma^2 I$, l'analisi fattoriale corrisponde con la PPCA. Tutti i parametri μ, W e Ψ si possono stimare con il metodo di massimo verosimiglianza e poi risolvere con un algoritmo iterativo come quello EM.

1.5 Confronto tra gli approcci

La PCA probabilistica risulta particolarmente efficace quando ci sono dataset di grandi dimensioni mentre la PCA bayesiana quando abbiamo dataset con molti parametri ma poche osservazioni perchè supera il problema dei valori estremi (-1 e 1) della correlazione. Entrambe, al contrario della PCA classica, sono funzionali anche con dati mancanti, infatti le distribuzioni ipotizzate delle variabili z dello spazio latente si adattano comunque. Il vantaggio principale della PCA bayesiana è che non dobbiamo fare ipotesi sul numero di parametri dello spazio della variabile latente, perchè viene determinato automaticamente. Questa proprietà rende la BPCA anche un metodo per le stime dei valori mancanti.
È interessante notare anche come la pca classica è ottenibile dalla pca probabilistica quando per la variabile gaussiana latente $\sigma^2 \to 0$.
In generale quindi quando ci sono dati mancanti è preferibile utilizzare i metodi probabilistici che forniscono risultati migliori, riducendo i gradi di libertà ma senza perdere informazione sulla matrice di correlazione. Inoltre utilizza l'algoritmo EM che per dataset grandi ha un costo computazionale minore rispetto alla stima del valore che massimizza la verosimiglianza.

Capitolo 2

Implementazione nel software R

In questo capitolo, verranno presentate alcune delle librarie R che hanno al loro intorno delle funzioni per trattare la PCA classica, la PCA probabilistica e la PCA bayesiana. Le librerie delle prime 2 sezioni sono presenti nel CRAN mentre l'ultima si tratta di una libreria di Bioconductors.

2.1 Librerie R

2.1.1 Libreria base

prcomp

Si tratta di uno dei principali comandi per effettuare la PCA classica calcolandola con la decomposizione a valori singolari. Accetta nell'input un dataframe in cui è necessario definire come trattare la presenza di NA; di default i dati mancanti vengono trattati con il comando *na.omit*() che restituisce una matrice in cui sono stati rimosse le osservazioni incomplete.
Si può scegliere se effettuare il calcolo direttamente sulla matrice di covarianza o su quella di correlazione con l'argomento *scale*.

princomp

Questo comando calcola la PCA classica basandosi sugli autovalori e auttovettori che ricava dalla matrice di partenza. Anche in questo caso si puoò scegliere se effettuare il calcolo sulla matrice di covarianza o di correlazione ma con l'argomento *cor*. Come tutte le funzioni per il calcolo della PCA classica non accetta valori mancanti e di default li omette e riporta nell'output se la rimozione ha qualche tipo di rilevanza.

2.1.2 Rdimtools

Rdimtools è un pacchetto più generale con funzioni per applicare tecniche di riduzione lineari e non lineari. La principale problematica delle funzioni del pacchetto è che non accettano valori NA per cui in questo caso risulterà impossibile utilizzarle per verificare il buon adattamento dei dati.

do.ppca

La funzione calcola la PCA probabilistica ricevendo in input la matrice dei dati e il numero di componenti principali che vogliamo calcolare. Vengono calcolate le colonne W dello spazio latente e σ^2 con il metodo di massima verosimiglianza; non è specificato se venga usato l'algoritmo EM.

do.bpca

Il comando calcola la PCA bayesiana sulla matrice dei dati utilizzando l'algoritmo EM per ottenere la stima della funzione a posteriori. Per ciò nell'input è possibile inserire un parametro che limiti il numero massimo di iterazioni. L'output fornisce più precisamente le quantità teoriche del modello W e gli α. Il numero di componenti sarà allora dato dal numero di colonne di W che non tendono ad annularsi. Nell'output è presente anche la matrice Y dello spazio latente con un numero di colonne scelto dall'utente.

2.1.3 pcaMethods

PcaMethods è un pacchetto R che si trova su Bioconductor per effettuare la PCA su dataset incompleti e trovare stime per i dati mancanti.

ppca

Permette di calcolare le componenti principali probabilistiche anche in presenza di dati mancanti che devono essere denotati come NA ma si deve fornire la dimensione dello spazio latente o trovarlo con un algoritmo di cross-validation. Il comando inserisce nella matrice iniziale per i dati mancanti dei valori casuali scelti da una normale. La funzione utilizza per la stima dei parametri l'algoritmo EM e assume distribuzione normale multivariata per le variabili dello spazio latente e normale univariata per l'errore ϵ. Per verificare se la convergenza dell'algoritmo è buona, si può far girare l'algoritmo con diversi valori inseriti al posto di quelli mancanti e vedere se la varianza delle stime è piccola.

bpca

La funzione implementa la BPCA su dataset che possono contenere anche valori mancanti. In generale si ottengono risultati diversi dalla classica pca perchè non forza l'ortogonalità degli assi. Il comando utilizza l'algoritmo EM per aggiornare le stime e ha quindi un costo computazionale abbastanza elevato e si ferma quando raggiunge la convergenza ossia stima con una precisione di $1e^{-4}$, ossia la differenza nella funzione obbiettivo tra interazioni successive, o si supera il numero fissato di interazioni massime. Gli autori evidenziano come la differenza tra le stime e i veri valori degli autovalori aumenti quando il numero di osservazioni è minore, perchè manca l'informazione per determinare i veri autovettori.

Capitolo 3

Applicazioni sul software R

In questo capitolo verranno analizzati alcuni dataset con le funzioni del capitolo precedente per cercare di individuare i punti di forza dei diversi approcci.

3.1 MetaboliteData

Il primo dataset considerato è contenuto nel pacchetto "pcaMethods" ed è composto da 154 osservazioni di 52 metaboliti ossia i prodotti intermedi o finali del processo del metabolismo per rendere la sostanza assorbibile dall'organismo. Avendo il campione completo, è possibile controllare l'efficacia dei 3 metodi, togliendo in maniera casuale circa il 5% delle osservazioni e sostituendole con NA. Dalla figura 3.1 i valori in rosso corrispondono ai valori mancanti nel dataset e sembrano distibuiti casualmente. Questa analisi verrà effettuata con le funzioni dal pacchetto pcaMethods.

PCA

Il metodo classico per l'analisi delle componenti principali non accetta valori mancanti nel suo input; dobbiamo quindi eliminare tutte le osservazioni che hanno un dato mancante.
In questo caso si procede adattando la trasformazione lineare alla matrice

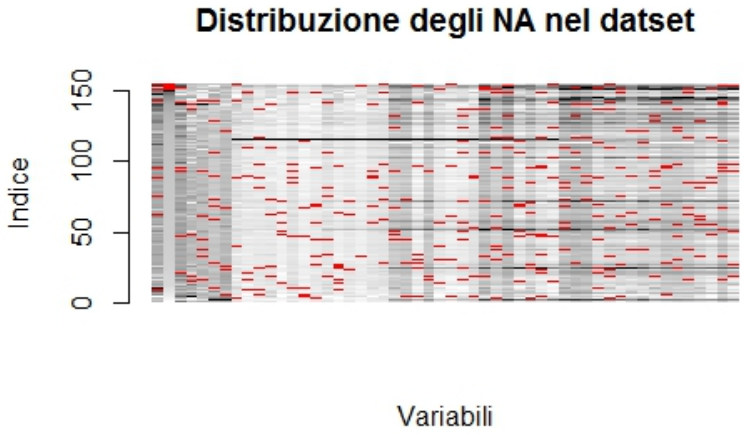


Figura 3.1: NA nel dataset

di correlazione del dataset completo che abbiamo a disposizione. Con 4 componenti principali si coglie l'84.0% della varianza totale.

PC1	PC2	PC3	PC4
0.6031	0.7336	0.79951	0.84018

Tabella 3.1: Varianza cumulata spiegata dalle prime 4 componenti principali del dataset completo

La PCA classica in questo caso senza dati mancanti risulta essere comunque un ottimo metodo per cogliere la covarianza tra le variabili. Il grafico in figura 3.2 rappresenta un biplot delle prime 2 componenti: si può dire che la prima componente sia formata da molti metaboliti mentre la seconda sia rappresentativa principalmente dei pochi metaboliti in direzione verticale. L'osservazione isolata in basso a destra corrisponde all'osservazione relativa alla Metossiammina maltosa che avrà per tutte le osservazioni valori molto alti poichè la sua proiezione sui vettori si discosta molto dall'origine come in partenza i suoi valori si discostavano da quelli medi calcolati per le altre va-

riabili. Quando questo avviene, significa che il metodo ha colto correttamente la matrice di correlazione.

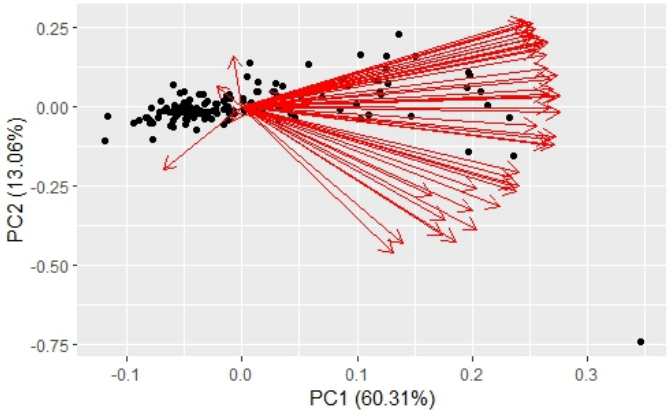


Figura 3.2: Biplot delle componenti principali sul dataset completo

PPCA

La pca probabilistica è stata applicata sul dataset incompleto per confrontare se le stime dei dati mancanti sono buone. La PCA applicata al dataset ricostruito e scalato da risultati molto similicome si vede confrontando la tabella 3.2 con la tabella della PCA classica 3.2.

PC1	PC2	PC3	PC4
0.6091	0.7402	0.80547	0.84454

Tabella 3.2: Varianza cumulata spiegata dalle prime 4 componenti principali del dataset ricostruito con PPCA

L'errore quadratico medio tra i valori nel dataset completo e quelli nel dataset ricostruito è 0.0259 e un buon adattamento dei dati si può vedere anche dagli assi stimati per le componenti principali (loadings); se i punti si trovano sulla bisettrice, significa che le componenti principali assumano lo stesso valore nei 2 casi. Il maggior scostamento si puo notare nella quarta

componente che è appunto quella che ha risutati più distanti del valore di varianza spiegata.

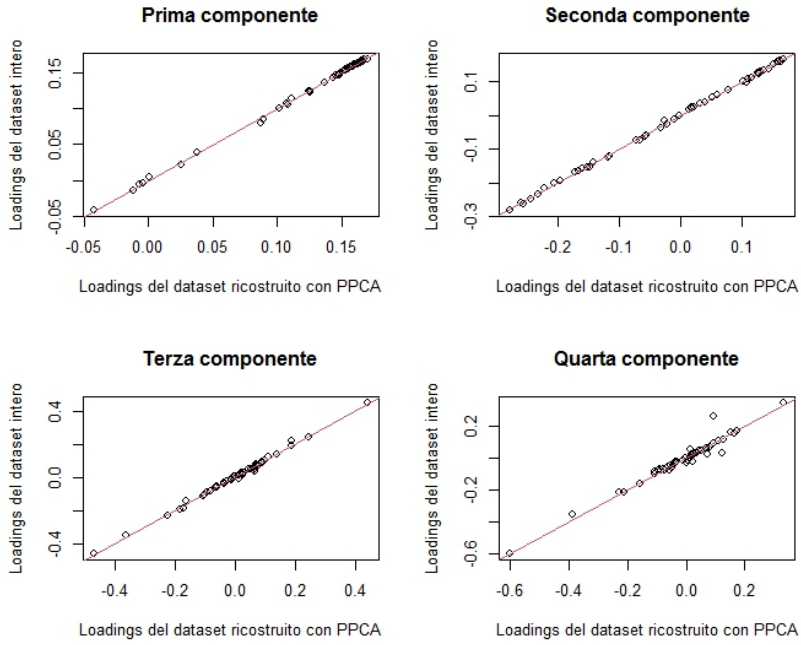


Figura 3.3: Loadings a confronto

BPCA

Ora vediamo invece le stime calcolate con l'approccio bayesiano; in questo caso non è necessario ipotizzare il numero di componenti dello spazio latente perchè sono individuate automaticamente dal processo bayesiano. In questo caso i risultati ottimi si ottengono con 4 o 5 componenti principali. Senza inserire nessun numero massimo di iterazioni, l'algoritmo si ferma dopo 50 iterazioni perchè raggiunge il livello di precisione. L'errore quadratico medio delle stime dei valori assenti è 0.0262 di poco superiore a quello della ppca. Queste variazioni possono essere anche dovute alla scelta del *seed*: per quest'analisi è stato impostato uguale a 3. Con la PCA applicata alla dataset ricostruito otteniamo dei valori della varianza cumulata da ciascuna compo-

nente principale simili a quelli evidenziati nella tabella 3.1 e nella tabella 3.2. Come precedentemente, si verifica una leggera sovrastima di quanto spieghino le componenti:

PC1	PC2	PC3	PC4
0.6094	0.7405	0.80545	0.84447

Tabella 3.3: Varianza cumulata spiegata dalle prime 4 componenti principali del dataset ricostruito con bpca

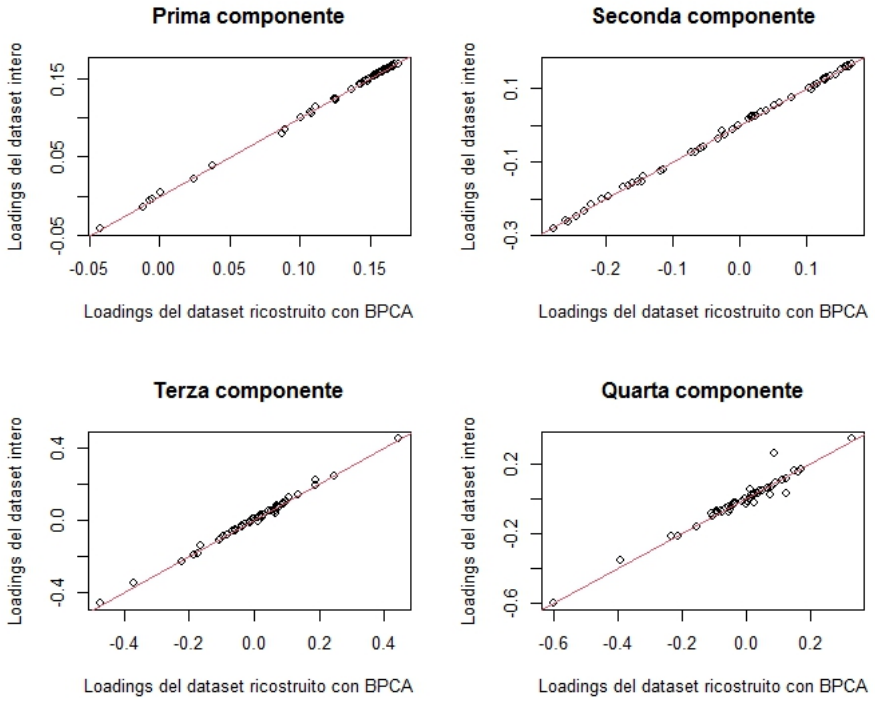


Figura 3.4: Loadings a confronto

Anche in questo caso per controllare un buon adattamento grafico possiamo utilizzare il confronto dei loadings e verificare se seguono la bisettrice ossia

se i valori si corrispondono. In generale l'adattamento sembra leggermente meno accurato soprattutto nella parte inferiori dei grafici in figura 3.4.

Commenti

Questo dataset con dati mancanti artificialmente e casualmente ci ha permesso di vedere che i 2 metodi probabilistici per la pca forniscono una buona stima per i dati mancanti. Si nota anche che si sono ottenuti valori di poco distinti ma in entrambi i casi ottimi. Si è condotto uno studio di simulazione, creando 50 dataset per ciascuna percentuale di dati manacanti e valutando poi la media degli errori quadrati medi. Si nota che il MSE cresce all'aumentare delle percentuali per entrambi i metodi. Nella PCA l'errore è sempre più piccolo rispetto alla BPCA a parità di dati mancanti. Questa differenza fino a un 30% di dati mancanti non sembra essere significativa, ma poi l'errore della BPCA esplode e raggiunge valori molto alti, quasi vicino al doppio dell'errore della PCA classica.

3.2 MINST dataset

Il secondo dataset considerato contiene 60000 quadrati di 28x28 pixel di cifre scritte a mano. A ogni quadrato è associata una variabile che indica quale è la cifra rappresentata. Si è ridotto il dataset a solo le osservazioni riguardanti il numero 3 ossia 6131 quadrati. Infatti l'obbiettivo è osservare come si comportano i diversi metodi e se l'approssimazione rimane buona diminuendo le componenti principali. Per le analisi è stato utilizzato il pacchetto R *Rdimtools*.

PCA

Avendo il dataset molte variabili saranno necessarie più componenti principali; infatti per superare le soglie di varianza cumulata del 70%, dell'80% e del 90% servono rispettivamente 22, 39 e 80 componenti. La PCA è stata

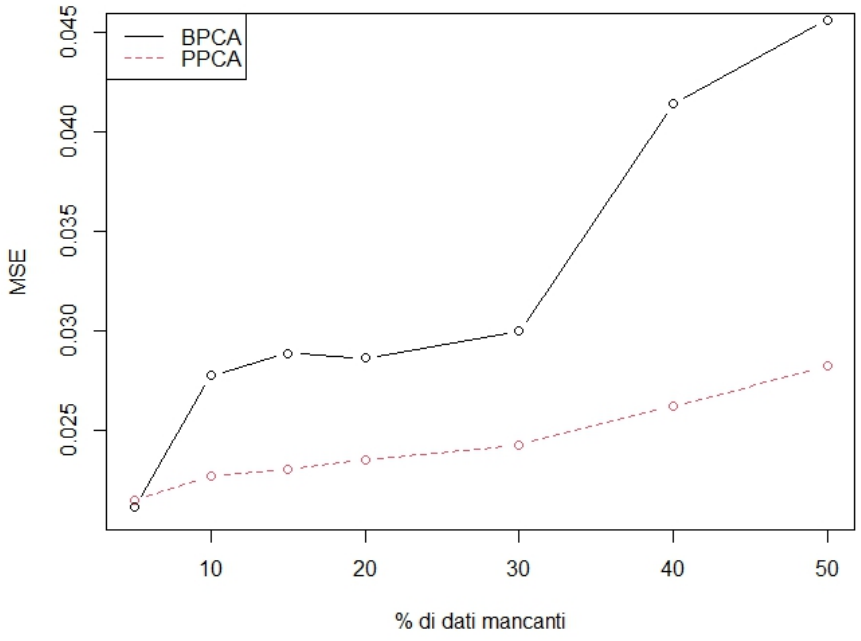


Figura 3.5: Distribuzione dei valori del MSE al variare della percentuale di dati mancanti

applicata alla matrice di covarianza perchè molte variabili specialmente quelle riferite ai pixel più esterni assumono valore costante 0 e varianza nulla. Questo però non causa grossi problemi interpretativi poichè tutte le variabili sono misurate con la stessa unità di misura. Dal grafico in figura 3.7 si vede che dopo circa 300 componenti, l'aggiunta di nuovi fattori non apporterebbe nessun miglioramento.

Nella figura successiva viene mostrato il primo numero 3 del dataset rappresentato prima con 1 una sola componente, poi con 10, 50 e 250. Già dal primo grafico si può riconoscere la cifra ma diventa sempre più comprensibile aumentando le componenti. Con 50 componenti principali raggiungiamo la soglia dell'80% della varianza spiegata mentre con 250 come si vedeva nel grafico in figura 3.7 si arriva quasi a spiegare tutta la varianza ed infatti la

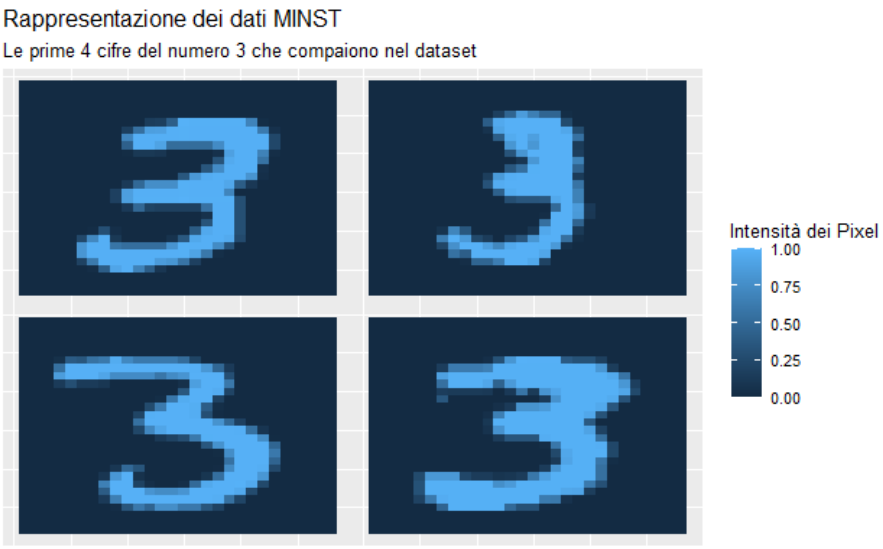


Figura 3.6: Visualizzazione della cifra del dataset MINST

cifra rappresentata è praticamente uguale a quella di partenza. Anche se con 10 componenti non si raggiunge neppure il 70% di varianza spiegata, il miglioramento rispetto all'utilizzo della solo la prima componente è evidente: infatti le prime componenti contengono sempre pù informazione. Con questo dataset risulta quindi utile la PCA per ridurre la dimensionalità perchè le cifre rimangono comunque riconoscibili.

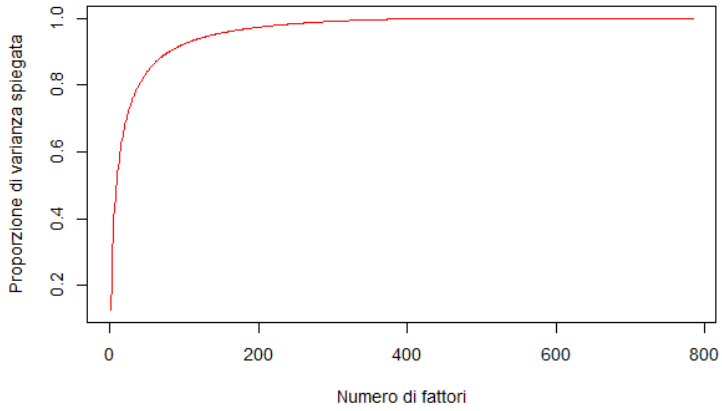
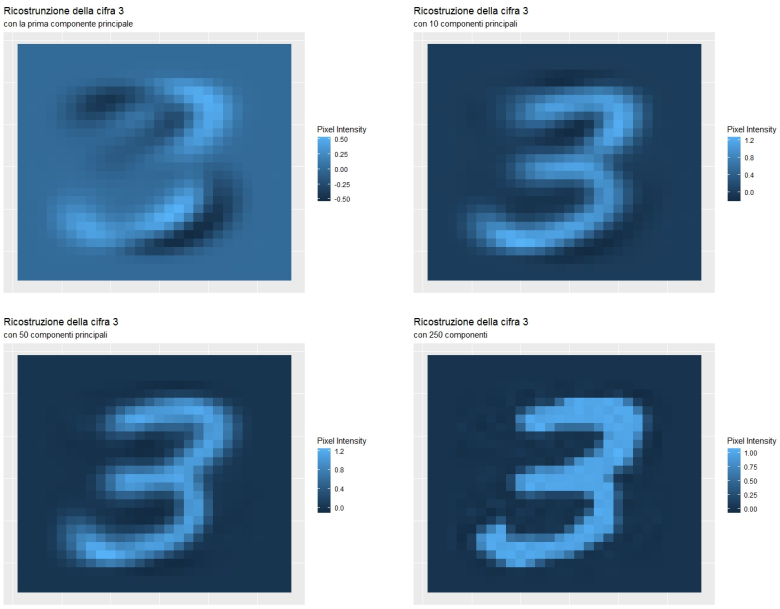
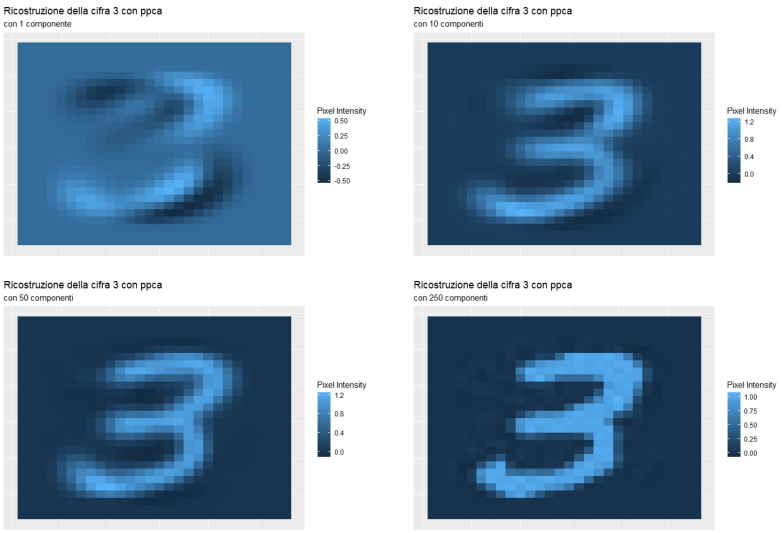


Figura 3.7: Varianza cumulata



PPCA

Ora vediamo come la Pca probabilistica riesce a cogliere le informazioni per rappresentare il numero 3 al variare delle dimensioni dello spazio latente.



In generale i risultati appaiono simili soprattutto perchè non ci sono dati mancanti.
Definendo uno spazio latente di dimensione $k = 22$ come il numero di componenti principali necessarie per spiegare almeno il 70% della varianza cumulata, la funzione resituisce la stima della matrice W di dimensione 784×22 dove ciascuna colonna rappresenta la direzione della componente corrispondente e la stima di σ^2 che è uguale a 0.0177 che coincide con la varianza in tutte le altre direzioni che non entrano nella matrice W. Per quanto detto precedentemente, ipotizzando la distribuzione a priori nello spazio latente, la varianza processo è:

$$C = WW^T + \sigma^2 I.$$

La matrice C dovrebbe quindi rappresentare la vera varianza del dataset e infatti l'errore quadratico medio dei valori delle 2 matrici è 1.185.

Commenti

In questo esempio abbiamo visto che i 2 modelli danno risultati molto simili. L'algoritmo per la bpca è poco funzionale per questi dati con molte osservazioni perchè utilizza l'algoritmo EM. La PPCA risulta efficace anche a livello di tempo impiegato infatti l'intero algoritmo calcola le componenti

principali in 9.74 secondi contro i 32.32 secondi della PCA classica. Per la BPCA non è calcolabile nemmeno il tempo di realizzazione.

3.3 OzoneNA dataset

Il terzo dataset contiene 112 osservazioni di 12 variabili raccolte a Rennes nel 2001 relative al livello massimo di ozono nell'aria con la variabile **maxO3** per il giorno corrente e **maxO3v** per il giorno antecedente. Sono presenti poi variabili metereologiche nel momento della registrazione e nei giorni precedenti. Le variabili **T** indicano la temperatura, le variabili **Vx** la proiezione del vettore della velocità del vento e le variabili **Ne** la presenza di nuvole tutte raccolte a 3 orari: 9, 12, 15. La scelta di questo dataset è motivata dalla mancanza di numerose osservazioni che non ci permettono di utilizzare la pca classica.

Analisi descrittiva

La percentuale di dati mancanti è del 23.07% e ci sono solo 13 osservazioni complete, quindi con tutte le variabili rilevate. Nella figura 3.8 sono rappresentate tutte le celle della matrice dei dati: i rettangoli rossi indicano i dati mancanti mentre negli altri il dato è rappresentato seguendo la scala di grigi. Si nota che le mancanze non sono casuali; per esempio quando non è stata rilevata la temperatura alle 9, manca più frequentemente anche il dato relativo agli orari successivi. Questo andamento seppur meno evidente si può notare anche nelle altre 2 variabili registrate in 3 orari.
Dalla figura sembra anche che quando la variabile temperatura viene rilevata, si hanno valori più grandi della nebulosità.

Ci sono poi alcune variabili che sono state registrate meno frequentemente; per la variabile nuvolosità alle 12 (Ne12) il dato manca nel 37.5% delle osservazioni mentre la temperatura non ha, in media tra i 3 orari, il 31.84% dei dati.

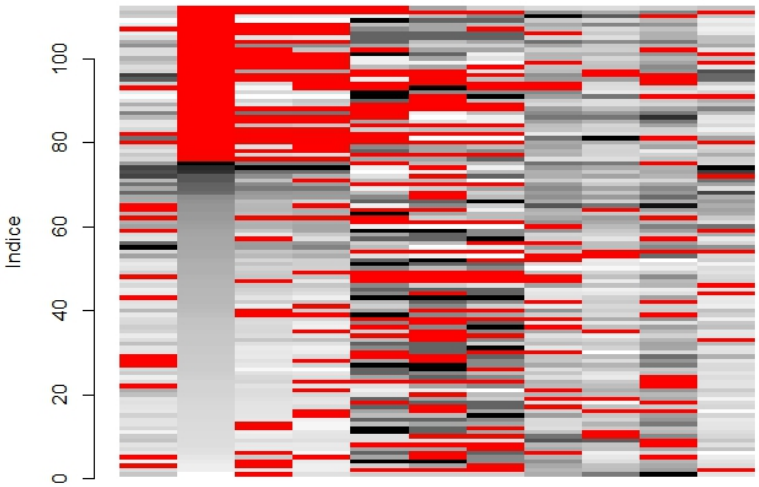


Figura 3.8: Dati mancanti

PPCA

Proviamo ad applicare la PCA probabilistica con entrambe le funzioni a nostra disposizione. A differenza dei primi 2 esempi, la dimensione ottima dello spazio latente è stata scelta con un metodo di convalida incrociata che ha dato come risultato 5. Dopo aver applicato la funzione PPCA e ricostruito il dataset con le osservazioni stimate, troviamo le componenti principali con il metodo classico sui dati standardizzati. Con 5 componenti si spiega il 92.8% della varianza cumulata che è un valore molto alto; per ridurre di più la dimensionalità ci potrebbero bastare 3 componenti che spiegano comunque l'84.6%.
Dal biplot delle prime 2 componenti in figura 3.9 si nota che:

- Tutte le osservazioni della stessa variabile nei diversi orari sono fortemente correlate positivamente tra loro;

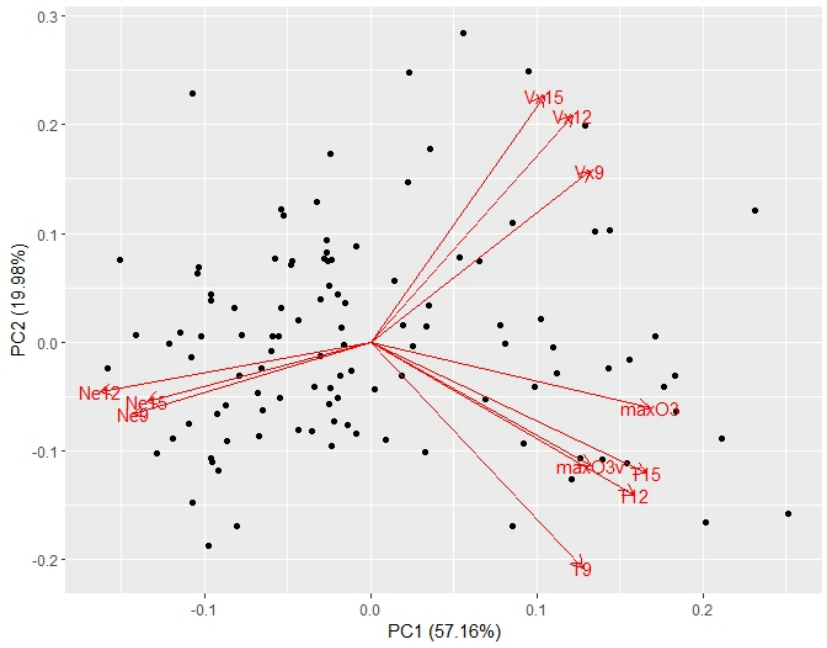


Figura 3.9: Biplot delle prime 2 componenti sul dataset Ozone ricostruito con ppca

- I livelli di ozono sono correlati alla temperatura positivamente e negativamente con nuvolosità;
- La correlazione con la velocità del vento è bassa con quasi tutte le variabili ma specialmente con la quantità di ozono, formando un angolo di circa 90 gradi.

La prima componente rappresenta la nuvolosità e i livelli di ozono il giorno corrente mentre la seconda rappresenta più la condizione climatica: con vento forte e temperature alte siamo appunto agli estremi della situazione. La rotazione degli assi di partenza nella seconda e nella terza componente coglie anche la variabile categoriale che rappresenta la direzione del vento, evidenziando dei raggruppamenti soprattutto per la direzione est (a destra) e ovest (sinistra) come vediamo nella figura 3.10. Le osservazioni relative ai giorni in cui il vento soffia da Nord sono le uniche poco raggruppate ma

sparse nel grafico.

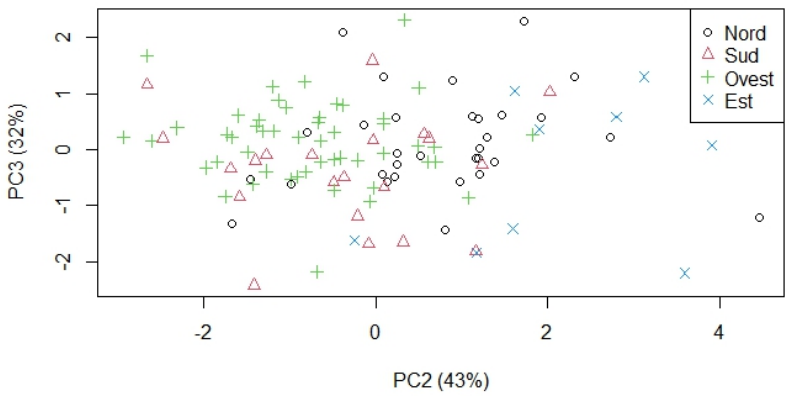


Figura 3.10: Divisione rispetto alla direzione del vento

BPCA

Ora vediamo il comportamento della PCA bayesiana. Questa volta utilizziamo sempre 5 come dimensione dello spazio latente per rendere più facile il confronto. Sui dati ricostruiti si può applicare sia la pca classica che la pca bayesiana con l'algoritmo "do.bpca". Con la classica funzione i contributi delle variabili alle componenti sembrano rimanere circa uguali. Cambiano leggermente le percentuali di varianza spiegata dalle singole componenti: la prima nel modello bayesiano spiega il 58.3% della varianza contro un 57.2% nella probabilistica. Nel complesso però con 3 componenti, la varianza cumulata è più alta nel primo modello ossia 84.6% contro 83.9%.
La funzione *do.bpca* riporta nell'output, oltre alla matrice delle osservazioni dello spazio latente di dimensione scelta k=5, le stime di:

- $\hat{\sigma}^2$=1733.96, stimata con l'algoritmo EM.

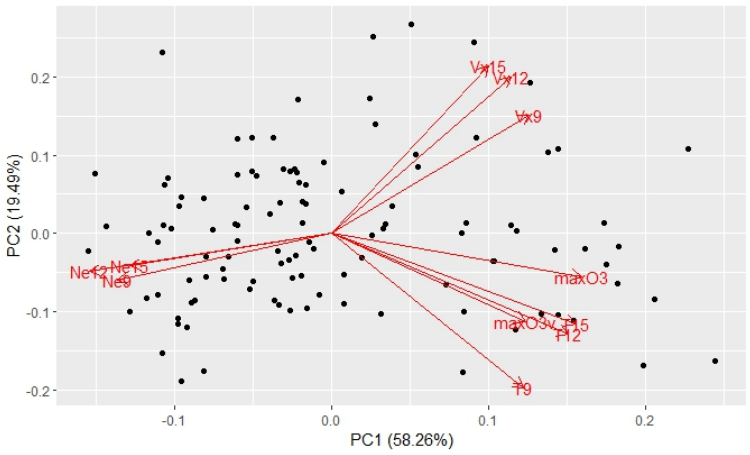


Figura 3.11: Biplot delle prime 2 componenti con Datset ricostruito con bpca

- $\hat{W}$, la matrice che rappresenta i vettori dello sottospazio di dimensione $p \times p - 1$.
- $\hat{\alpha}_i$ ossia gli iperparametri associati alle colonne di W. Per ciascun $\alpha_i \to \infty$ la corrispondente colonna di $w_i \to 0$ e quindi non verrà utilizzata nel modello. In questo esempio ci sono alcuni valori di α_i in ordini mateamatici diversi. I valori sull'ordine di circa 10^{80}, che nell'esempio corrispondono al quinto e al decimo valore, saranno quelli che non apporterano sufficiente informazione e la rispettiva colonna di W non sarà conteggiata.

La proiezione dei dati visibile in figura 3.12 non riesce a cogliere molto bene le divisioni in base alla direzione del vento e questo si vede dal fatto che i punti si sovrappongono mentre il biplot in figura 3.11 sembra uguale a livello interpretativo.

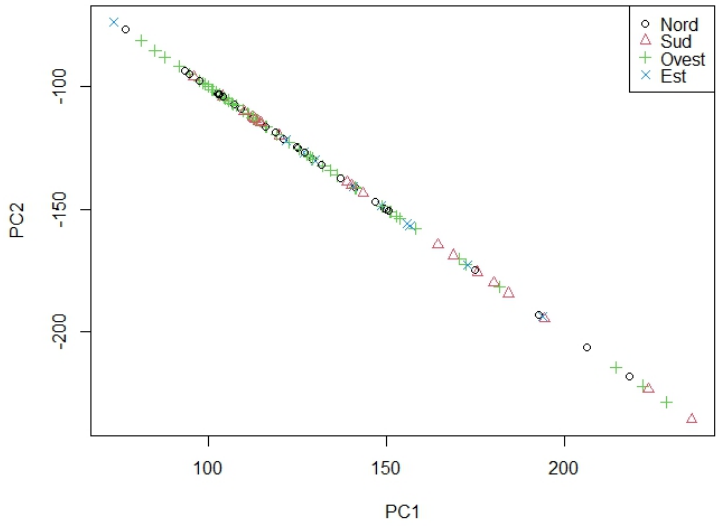


Figura 3.12: Rappresentazioni delle proiezioni con metodo bpca

Conclusioni

In conclusione, osservando i risultati ottenuti, si può dire che sia la PPCA che la BPCA sono 2 metodi utili per sostituire la PCA classica quando mancano delle osservazioni. Anche se poi non si volesse effettuare una riduzione del dataset, sono entrambi buoni metodi per stimare i valori mancanti quantitativi. Si è visto che la PPCA realizza delle stime buone anche con il 30% dei dati assenti come nel datset Metabolite e anche aumentando il numero di dati mancanti si ottengono buone previsioni.
Un altro vantaggio è sicuramente la capacità di cogliere la correlazione tra le variabili limitando il numero di parametri. Anche rispetto all'analisi fattoriale, utilizza un unico parametro per la varianza negli assi che spiegano poco invece che assegnarne uno ad ogni asse. Per quanto riguarda la BPCA, con gli algoritmi disponibili su R, si dimostra un metodo un pò più lento perchè utilizza l'algoritmo EM per stimare W e α_i e ogni volta deve confrontare i risultati di ogni step e aggiornarli. Dal dataset Ozone si è visto che rispetto alla PCA classica ha colto meno le differenze categoriale quando è stato applicato al dataset ricostruito. In generale per la stima dei valori mancanti i 2 metodi probabilistici hanno poche differenze, ma nell'applicazione per trovare le componenti sul dataset ricostruito la PCA classica rimane una buona soluzione.

Bibliografia

Bishop, Christopher M.; Tipping, Micheal E., Bayesian PCA. Microsoft Research, Cambridge, UK, 1999.

Bishop, Christopher M., Pattern Recognition And Machine Learning (capitolo 12). Springer-Nature New York Inc, 2006.

Johnson, Richard A.; Wichern, Dean W., Applied Multivariate Statistical Analysis. Pearson Education Limited, 2014, Sixth Edition.

Roweis, Sam, EM Algorithms for PPCA and SPCA. NIPS, 1998