

FURTHER REMARKS ON PRINCIPAL COMPONENT ANALYSIS PCA AND PROTEIN MD: ESSENTIAL DYNAMICS

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 20, Rome thu nov 9rd 2023

DIPARTIMENTO DI FISICA



SAPIENZA
UNIVERSITÀ DI ROMA

Outline L 20

- [geometric data analysis/dimensional reduction / classification/clustering)]
- PCA and PROTEIN MD
- ESSENTIAL DYNAMICS
- OUTLINE OF A MD PROJECT
- FURTHER LINKS:

PROTEOPEDIA eg: GTD_TS metric

https://proteopedia.org/wiki/index.php/Calculating_GDT_TS

DATI/METADATI/ONTOLOGIE

- DATI numeri (misure, valutazioni)
- DESCRITTORI DI DATI quale misura sto effettuando con quale protocollo?
- SIGNIFICATO emergere della semantica in un contesto operativo (ad. es. prendere decisioni)

ORGANIZZAZIONE BASE DI UN DATASET (base dati)

	TABLEAU	TABELLA
	DESCRITTORE 1 CP = parametro	DESCRITTORE 2 ... DESCR P
OGGETTO 1	X_{11}	$X_{12} \quad \dots \quad X_{1P}$
OGGETTO 2	X_{21}	$X_{22} \quad \dots \quad X_{2P}$
\vdots	\vdots	\vdots
OGGETTO N	X_{N1}	$X_{N2} \quad \dots \quad X_{NP}$

MATRICE $N \times P \quad \{X_{ij}\}_{\substack{j=1, \dots, P \\ i=1, \dots, N}}$

Esempi: ① acque minerali (residuo fisso, conducibilità elettrica)

② corpi umani in un contesto (es. leva militare) (Colesterolo, Trigliceridi, azotemia, Glicemia, Emocromo ...)

E3 ③ Volti umani parametru di distawza intra
v. slides segventi ---

E4 ④ amino acids
properties

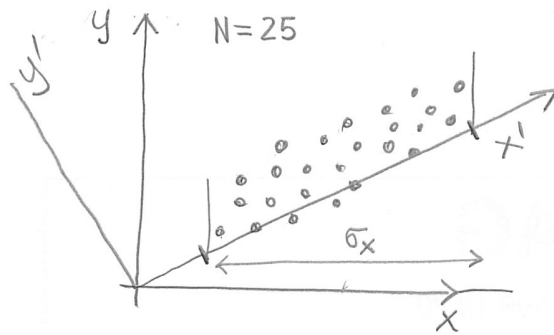
PCA (Principal Component Analysis) in a
nutshell (in estrema sintesi)

$i = 1, \dots, N$ Dati in una tabella (matrice
 $j = 1, \dots, P$ (base dati) $N \times P$)
oggetti, righe
quantità misurate, colonne

- Stiamo parlando di un insieme di dati
 P dimensionali, ogni riga, ogni oggetto
è rappresentato da un vettore P dimensionale

- IDEA della PCA
Cambiare sistema di riferimento
(rotazione di assi cartesiani)
in modo da esaltare la sensibilità
delle misure

- Es 2-Dim $X_i \ Y_i \ i = 1, 2, \dots, N$
(es. altezza e peso corporeo)



nel riferimento
ruotato le x sono
più sparpagliate =
abbiamo aumentato
la sensibilità
e forse anche il segnale

- Abbiamo cambiato assi, il riferimento (geometricamente, abbiamo cambiato 'BASE')
- L'idea è trovare, una volta presi i dati, una base ottimale (che massimizza qualcosa) che spingiamo esalti il segnale contenuto nei dati

• ALGORITMO PER LA PCA

0. PRENDI I DATI X_{ij} $i=1, \dots, N$ # oggetti
 $j=1, \dots, P$ # parametri

2. MEDIA DEI PARAMETRI
SUL CAMPIONE DI N oggetti

$$\mu_j = \frac{1}{N} \sum_{i=1}^N X_{ij}$$

3. DEFINISCI LA VARIABILITÀ SPECIFICA
DI OGNI PARAMETRO (DEVIAZIONE STD)

4. Z-TRASFORMATI DEI DATI INIZIALI
(STANDARDIZZA I DATI, PERCHÉ?) ③

$X \rightarrow Z$

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

σ_j DEVIATION STANDARD del
parametro j -mo

$$\sigma_j = \left[\frac{1}{N} \sum_{i=1}^N (X_{ij} - \mu_j)^2 \right]^{1/2} = \sqrt{\dots}$$

5 CAMBIA LA BASE (Cambia il riferimento per trovare il riferimento ottimale) [che massimizza la varianza contenuta nei dati]

Il 'segnale' contenuto nei dati è dato dalle correlazioni tra i valori assunti dai diversi parametri sui diversi oggetti, noi vogliamo esaltare queste correlazioni, se ci sono, trasformando i dati, cambiando sistema di riferimento

- Come si misurano le correlazioni?

R. Con coefficienti di Pearson

Introduciamo dunque, con la speranza di tirare fuori dai dati una 'struttura nascosta di correlazioni' la matrice di correlazione dei parametri sui dati originali (è una matrice $P \times P$)

$$C_{\ell k} = \frac{1}{N \sigma_{\ell} \sigma_k} \sum_{i=1}^N (x_{i\ell} - \mu_{\ell})(x_{ik} - \mu_k)$$

$\ell, k = 1, 2, \dots, P$

↑ ↑
{ somma sui N oggetti }

In termini delle variabili 'standardizzate'

$$C_{\ell k} = \frac{1}{N} \sum_{i=1}^N Z_{i\ell} Z_{ik}$$

Notare $C_{\ell k} = C_{k\ell}$; Perché?

Questa è una matrice reale e simmetrica e quindi è diagonalizzabile

Vuol dire questo: se C è una matrice reale e simmetrica allora esiste una matrice U che trasforma C in \tilde{C} con $\tilde{C} = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_p \end{pmatrix}$

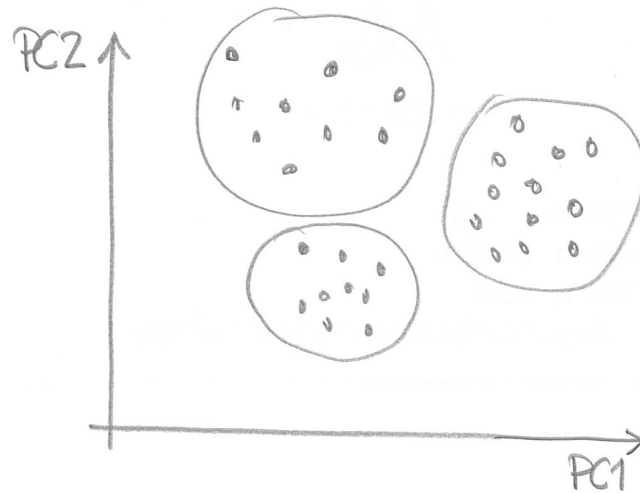
$$\tilde{C} = UCU^{-1} \quad U \cdot U^{-1} = I = \begin{pmatrix} 1 & & \\ & 1 & \\ & & \ddots \\ & & & 1 \end{pmatrix}$$

TBC

Consultare per esercizio il libro di
Paul Higgs e Teresa Attwood
Bioinformatics and Molecular Evolution
2.4, 2.5, 2.6, Box 2.2

CLUSTERING

Che cosa ci aspettiamo dalla PCA ?



[PC1, PC2; sono le direzioni più 'importanti' del nuovo sistema di riferimento]

... EMERGE UNA ONTOLOGIA

Table 2.2 Physico-chemical properties of the amino acids.

			Vol.	Bulk.	Polarity	pI	Hyd.1	Hyd.2	Surface area	Fract. area
Alanine	Ala	A	67	11.50	0.00	6.00	1.8	1.6	113	0.74
Arginine	Arg	R	148	14.28	52.00	10.76	-4.5	-12.3	241	0.64
Asparagine	Asn	N	96	12.28	3.38	5.41	-3.5	-4.8	158	0.63
Aspartic acid	Asp	D	91	11.68	49.70	2.77	-3.5	-9.2	151	0.62
Cysteine	Cys	C	86	13.46	1.48	5.05	2.5	2.0	140	0.91
Glutamine	Gln	Q	114	14.45	3.53	5.65	-3.5	-4.1	189	0.62
Glutamic acid	Glu	E	109	13.57	49.90	3.22	-3.5	-8.2	183	0.62
Glycine	Gly	G	48	3.40	0.00	5.97	-0.4	1.0	85	0.72
Histidine	His	H	118	13.69	51.60	7.59	-3.2	-3.0	194	0.78
Isoleucine	Ile	I	124	21.40	0.13	6.02	4.5	3.1	182	0.88
Leucine	Leu	L	124	21.40	0.13	5.98	3.8	2.8	180	0.85
Lysine	Lys	K	135	15.71	49.50	9.74	-3.9	-8.8	211	0.52
Methionine	Met	M	124	16.25	1.43	5.74	1.9	3.4	204	0.85
Phenylalanine	Phe	F	135	19.80	0.35	5.48	2.8	3.7	218	0.88
Proline	Pro	P	90	17.43	1.58	6.30	-1.6	-0.2	143	0.64
Serine	Ser	S	73	9.47	1.67	5.68	-0.8	0.6	122	0.66
Threonine	Thr	T	93	15.77	1.66	5.66	-0.7	1.2	146	0.70
Tryptophan	Trp	W	163	21.67	2.10	5.89	-0.9	1.9	259	0.85
Tyrosine	Tyr	Y	141	18.03	1.61	5.66	-1.3	-0.7	229	0.76
Valine	Val	V	105	21.57	0.13	5.96	4.2	2.6	160	0.86
Mean			109	15.35	13.59	6.03	-0.5	-1.4	175	0.74
Std. dev.			28	4.53	21.36	1.72	2.9	4.8	44	0.11

Vol., volume calculated from van der Waals radii (Creighton 1993); Bulk., bulkiness index (Zimmerman, Eliezer, and Simha 1968); Polarity, polarity index (Zimmerman, Eliezer, and Simha 1968); pI, pH of the isoelectric point (Zimmerman, Eliezer, and Simha 1968); Hyd.1, hydrophobicity scale (Kyte and Doolittle 1982); Hyd.2, hydrophobicity scale (Engelman, Steitz, and Goldman 1986); Surface area, surface area accessible to water in unfolded peptide (Miller *et al.* 1987); Fract. area, fraction of accessible area lost when a protein folds (Rose *et al.* 1985).

BOX 2.2

Principal component analysis in more detail

From the $N \times P$ data matrix, we can define a $P \times P$ matrix of correlation coefficients, C_{jk} , between the properties:

$$C_{jk} = \frac{1}{N\sigma_j\sigma_k} \sum_i (X_{ij} - \mu_j)(X_{ik} - \mu_k) = \frac{1}{N} \sum_i z_{ij}z_{ik}$$

The coefficients are always in the range -1 to 1 . If $C_{jk} > 0$, the two properties are positively correlated, i.e., they both tend to be large at the same time and small at the same time. If $C_{jk} < 0$, the properties are negatively correlated, i.e., one tends to be large when the other is small. The correlation matrix for the amino acid data looks like this.

The matrix is symmetric ($C_{jk} = C_{kj}$) and all the diagonal elements are 1.00 by definition. The values illustrate features of the data that are not easy to see in the original matrix. For example, volume has a strong positive correlation with surface area and bulkiness, and a fairly weak correlation with the other properties. The two hydrophobicity scales have strong positive correlation with each other and also with the fractional area property, but they have a significant negative correlation with the polarity scale.

It can be shown that the vectors \mathbf{v}_j that define the principal component axes are the eigenvectors of the correlation matrix, i.e., they satisfy the equation:

$$\sum_j v_{nj} C_{jk} = \lambda_n v_{nk}$$

where the λ_n are constants called eigenvalues. The first principal component (PC) vector is the eigenvector with the largest eigenvalue. Subsequent PCs can be listed in order of decreasing size of eigenvalue. The first two eigenvalues in this case are $\lambda_1 = 3.57$ and $\lambda_2 = 2.81$.

The variance along the n^{th} PC axis is equal to the corresponding eigenvalue:

$$\frac{1}{N} \sum_i y_{in}^2 = \frac{1}{N} \sum_i \sum_j \sum_k v_{nj} z_{ij} v_{nk} z_{ik} = \sum_j \sum_k v_{nj} C_{jk} v_{nk} = \sum_k \lambda_n v_{nk}^2 = \lambda_n$$

We know that the variance of each of the z coordinates is 1 , hence the total variance of all the coordinates is P . When we change the coordinates to the principal components, we just rotate the points in space, so the total variance in the PC space is still P . The fraction of the total variance represented by the first two PCs is therefore $(\lambda_1 + \lambda_2)/P$, which in our case is $(3.57 + 2.81)/8 = 0.797$. This is why it is useful to look at the data on the PC plot (as in Fig. 2.10). Roughly 80% of the variation in the positioning of the points in the original coordinates can be seen with just two PCs. When points appear close in the two-dimensional plot of the first two PCs, they really are close in the eight-dimensional space, because the remaining six dimensions that we can't see do not contribute much to the distance between the points. This means that if we spot patterns in the data in the PC plot, such as clusters of closely spaced points, then these are likely to give a true impression of the patterns in the full data.

	Vol	Bulk.	Pol.	pl	Hyd.1	Hyd.2	S.A.	Fr.A.
Vol.	1.00	0.73	0.24	0.37	-0.08	-0.16	0.99	0.18
Bulk.	0.73	1.00	-0.20	0.08	0.44	0.32	0.64	0.49
Pol.	0.24	-0.20	1.00	0.27	-0.69	-0.85	0.29	-0.53
pl	0.37	0.08	0.27	1.00	-0.20	-0.27	0.36	-0.18
Hyd.1	-0.08	0.44	-0.67	-0.20	1.00	0.85	-0.18	0.84
Hyd.2	-0.16	0.32	-0.85	-0.27	0.85	1.00	-0.23	0.79
S.A.	0.99	0.64	0.29	0.36	-0.18	-0.23	1.00	0.12
Fr.A.	0.18	0.49	-0.53	-0.18	0.84	0.79	0.12	1.00

BOX 2.2

Principal component analysis in more detail

From the $N \times P$ data matrix, we can define a $P \times P$ matrix of correlation coefficients, C_{jk} , between the properties:

$$C_{jk} = \frac{1}{N\sigma_j\sigma_k} \sum_i (X_{ij} - \mu_j)(X_{ik} - \mu_k) = \frac{1}{N} \sum_i z_{ij}z_{ik}$$

The coefficients are always in the range -1 to 1 . If $C_{jk} > 0$, the two properties are positively correlated, i.e., they both tend to be large at the same time and small at the same time. If $C_{jk} < 0$, the properties are negatively correlated, i.e., one tends to be large when the other is small. The correlation matrix for the amino acid data looks like this.

The matrix is symmetric ($C_{jk} = C_{kj}$) and all the diagonal elements are 1.00 by definition. The values illustrate features of the data that are not easy to see in the original matrix. For example, volume has a strong positive correlation with surface area and bulkiness, and a fairly weak correlation with the other properties. The two hydrophobicity scales have strong positive correlation with each other and also with the fractional area property, but they have a significant negative correlation with the polarity scale.

It can be shown that the vectors \mathbf{v}_j that define the principal component axes are the eigenvectors of the correlation matrix, i.e., they satisfy the equation:

$$\sum_j v_{nj} C_{jk} = \lambda_n v_{nk}$$

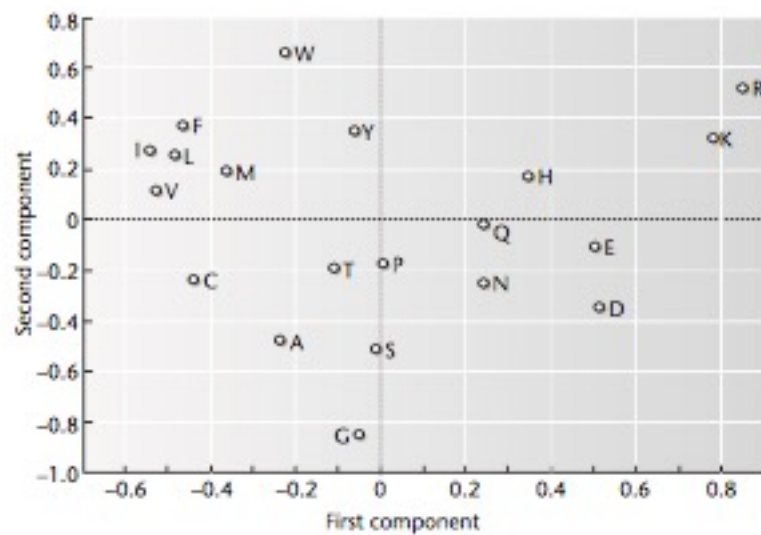
where the λ_n are constants called eigenvalues. The first principal component (PC) vector is the eigenvector with the largest eigenvalue. Subsequent PCs can be listed in order of decreasing size of eigenvalue. The first two eigenvalues in this case are $\lambda_1 = 3.57$ and $\lambda_2 = 2.81$.

The variance along the n^{th} PC axis is equal to the corresponding eigenvalue:

$$\frac{1}{N} \sum_i v_{in}^2 = \frac{1}{N} \sum_i \sum_j \sum_k v_{nj} z_{ij} v_{nk} z_{ik} = \sum_j \sum_k v_{nj} C_{jk} v_{nk} = \sum_k \lambda_n v_{nk}^2 = \lambda_n$$

We know that the variance of each of the z coordinates is 1 , hence the total variance of all the coordinates is P . When we change the coordinates to the principal components, we just rotate the points in space, so the total variance in the PC space is still P . The fraction of the total variance represented by the first two PCs is therefore $(\lambda_1 + \lambda_2)/P$, which in our case is $(3.57 + 2.81)/8 = 0.797$. This is why it is useful to look at the data on the PC plot (as in Fig. 2.10). Roughly 80% of the variation in the positioning of the points in the original coordinates can be seen with just two PCs. When points appear close in the two-dimensional plot of the first two PCs, they really are close in the eight-dimensional space, because the remaining six dimensions that we can't see do not contribute much to the distance between the points. This means that if we spot patterns in the data in the PC plot, such as clusters of closely spaced points, then these are likely to give a true impression of the patterns in the full data.

	Vol	Bulk.	Pol.	pl	Hyd.1	Hyd.2	S.A.	Fr.A.
Vol.	1.00	0.73	0.24	0.37	-0.08	-0.16	0.99	0.18
Bulk.	0.73	1.00	-0.20	0.08	0.44	0.32	0.64	0.49
Pol.	0.24	-0.20	1.00	0.27	-0.69	-0.85	0.29	-0.53
pl	0.37	0.08	0.27	1.00	-0.20	-0.27	0.36	-0.18
Hyd.1	-0.08	0.44	-0.67	-0.20	1.00	0.85	-0.18	0.84
Hyd.2	-0.16	0.32	-0.85	-0.27	0.85	1.00	-0.23	0.79
S.A.	0.99	0.64	0.29	0.36	-0.18	-0.23	1.00	0.12
Fr.A.	0.18	0.49	-0.53	-0.18	0.84	0.79	0.12	1.00



Neutral, nonpolar	W, F, G, A, V, I, L, M, P
Neutral, polar	Y, S, T, N, Q, C
Acidic	D, E
Basic	K, R, H

Loadings

	Vol	Bulk.	Pol.	pI	Hyd.1	Hyd.2	S.A.	Fr.A.
Comp. 1	(0.06,	-0.22,	0.44,	0.19,	-0.49,	-0.51,	0.10,	-0.45)
Comp. 2	(0.58,	0.48,	0.10,	0.25,	0.03,	-0.03,	0.56,	0.17)

DETTAGLI PCA

Dalla matrice dei dati $N \times P$ si ottiene la matrice di covarianza tra descrittori (parametri) $P \times P$

$$C_{jk} = \frac{1}{N \sigma_j \sigma_k} \sum_{i=1}^N (X_{ij} - \mu_j)(X_{ik} - \mu_k)$$

$$= \frac{1}{N} \sum_{i=1}^N z_{ij} z_{ik} \quad (\text{in termini di variabili trasformate } z)$$

oss.

$-1 \leq C_{jk} \leq 1$, la matrice C è diagonale

Es. Quanto vale C_{ii} ($i = 1, \dots, P$)?

$$\text{rem. } \sigma_j \equiv \left(\frac{1}{N} \sum_{i=1}^N (x_{ij} - \mu_j)^2 \right)^{1/2}$$

DF. Le componenti principali sono gli autovettori \underline{V} di C ordinati secondo la grandezza dell'autovalore corrispondente

\underline{V}_j $j = 1, 2, \dots, P$ è un insieme di P vettori di dimensione P che sono una base

$$\underline{V}_j = \{V_{j1}, V_{j2}, \dots, V_{jP}\}$$

Originariamente avevamo N vettori di dimensione P Z_{ij}

$$\begin{matrix} i = 1, \dots, N \\ j = 1, \dots, P \end{matrix}$$

Perché i P vettori $\{V_j\}_{j=1}^P$ formino una base è necessario che siano ORTONORMALI

$$\sum_{k=1}^P V_{jk}^2 = 1 \quad \sum_{k=1}^P V_{jk} V_{sk} = \delta_{ks} \quad \forall k, s = 1, \dots, P$$

TRASFORMAZIONE DI COORDINATE

$Z_{ij} \longrightarrow Y_{ij}$
 vecchie \quad nuove
 i indice di osservazione $Y_{ij} = \sum_{k=1}^P V_{jk} Z_{ik}$
 j indice di descrizione \quad nuova base \quad vecchie coord
 per $i = 1, 2, \dots, N$

Ma la base nuova la scegliamo con un criterio che lungo ogni nuova direzione, definita dai nuovi vettori di base, la varianza nelle nuove coordinate sia massima.

$$\frac{1}{N} \sum_{i=1}^N Y_{ij}^2$$

Varianza lungo la direzione 1 nella nuova base

Equazione per gli autovalori e autovettori

$$\sum_{j=1}^P V_{mj} C_{jk} = \lambda_n V_{nk} \quad \begin{matrix} n=1, \dots, P \\ k=1, \dots, P \end{matrix}$$

$\lambda_1, \lambda_2, \dots, \lambda_P$ sono P costanti dette
autovalori, nella PCA questi vengono
ordinati $\lambda_1 > \lambda_2 > \lambda_3 \dots \lambda_P$

Es. La varianza dei dati y_{i1} $i=1, \dots, N$
lungo la direzione del primo autovettore è

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N y_{i1}^2 &= \left(\frac{1}{N} \right) \sum_{i=1}^N \sum_{j=1}^P \sum_{k=1}^P V_{ij} Z_{ij} \underbrace{V_{1k} Z_{ik}} \\ &= \sum_{j=1}^P \sum_{k=1}^P \underbrace{V_{ij} C_{jk} V_{1k}} = \sum_{k=1}^P \lambda_1 V_{1k}^2 \\ &= \lambda_1 \quad \text{Questo è il significato} \\ &\quad \text{degli autovalori in PCA} \end{aligned}$$

La somma $\lambda_1 + \lambda_2 + \dots + \lambda_P$ è la varianza
totale rappresentata nelle nuove variabili;

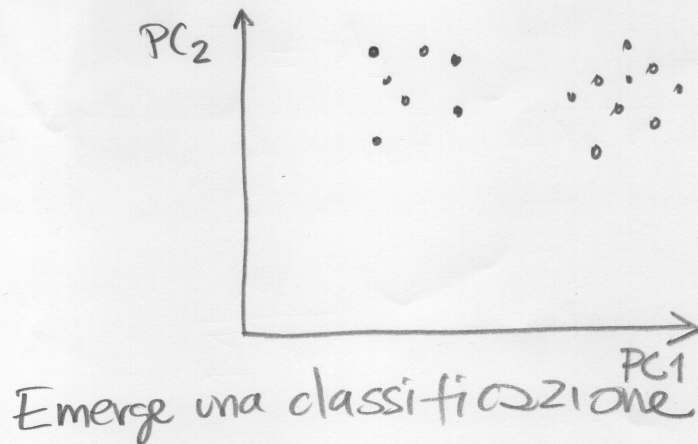
Quindi $\frac{\lambda_1}{\sum_{k=1}^P \lambda_k}$ è la 'percentuale' di varianza
rappresentata, proiettata
nella direzione 'nuova' 1

Quando 'funziona' la PCA ?

Quando $\frac{\lambda_1 + \lambda_2}{\sum_{k=1}^p \lambda_k}$ è 70, 80 % della
varianza

Geometricamente vuol dire che
Se ci limitiamo alle prime due
componenti perdiamo poco, della
variabilità (informazione) originale.
MA! abbiamo operato una
RIDUZIONE DIMENSIONALE

COMPRESSIONE

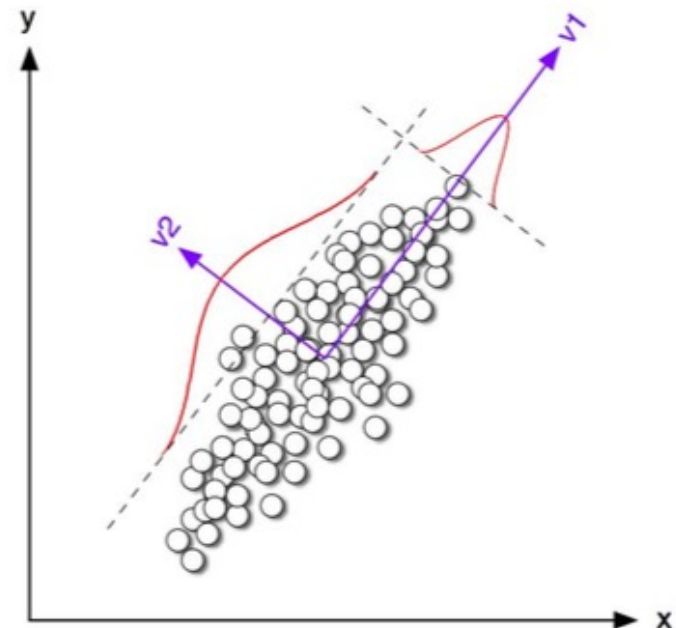
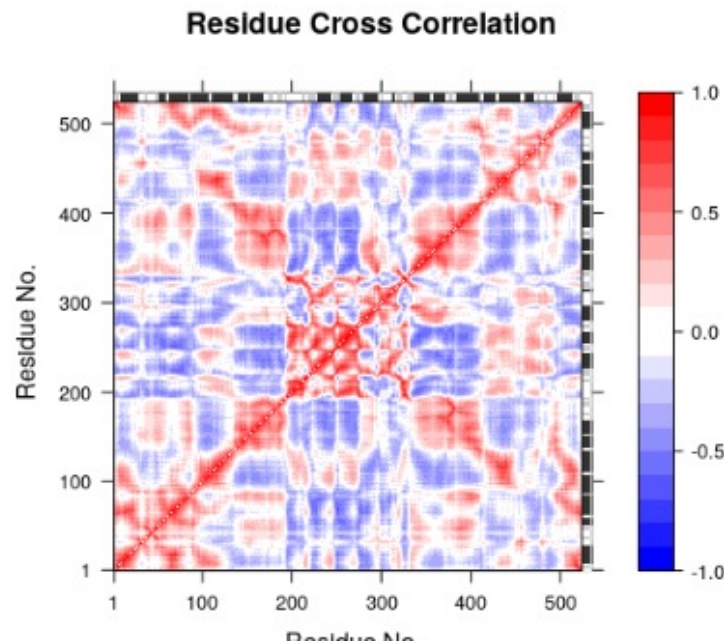


Essential Dynamics Simulation

Collective coordinates, as obtained by a principal component analysis of atomic fluctuations, are commonly used to predict a low-dimensional subspace in which essential protein motion is expected to take place.

Conformational transitions in proteins are essential for their function, such as substrate binding and product release, allosteric regulation, and many others.

The two most widely used computational methods to identify collective motions are normal mode analysis (NMA) and principal component analysis (PCA)



Essential Dynamics Simulation

PCA is a multivariate statistical analysis that involves diagonalization of a correlation matrix for a set of observables to yield collective variables.

PROTEINS: Structure, Function, and Genetics 17:412–425 (1993)

Essential Dynamics of Proteins

Andrea Amadei, Antonius B.M. Linssen, and Herman J.C. Berendsen

Department of Biophysical Chemistry and BIOSON Research Institute, the University of Groningen, 9747 AG Groningen, The Netherlands

Essential Dynamics Simulation

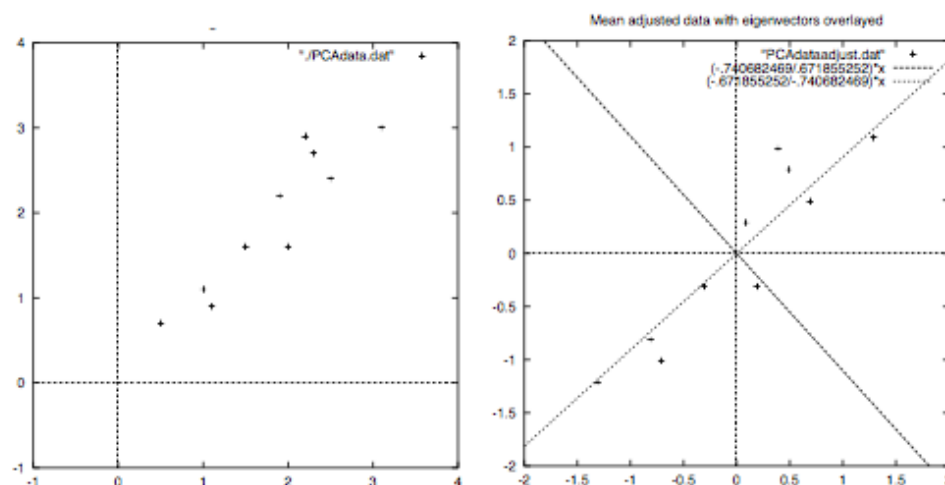
PCA is a multivariate statistical analysis that involves diagonalization of a correlation matrix for a set of observables to yield collective variables.

THEORETICAL FOUNDATION

- We consider the dynamics of a protein in equilibrium in a given environment at a temperature T
- We first eliminate the overall translational and rotational motion because these are irrelevant for the internal motion we wish to analyze
- The internal motion is now described by a trajectory $\mathbf{x}(t)$, where \mathbf{x} is a $3N$ -dimensional vector of all atomic coordinates, represented by a column vector.
- The correlation between atomic motions can be expressed in the covariance matrix \mathbf{C} of the positional deviations:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)}$$

$$\mathbf{C} = \begin{pmatrix} \text{cov}(x, x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{cov}(y, y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{cov}(z, z) \end{pmatrix}$$



Essential Dynamics Simulation

PCA is a multivariate statistical analysis that involves diagonalization of a correlation matrix for a set of observables to yield collective variables.

THEORETICAL FOUNDATION

The total positional fluctuation can be thought to be built up from the contributions of the eigenvectors:

$$\begin{aligned}\sum_i \langle (\mathbf{x}_i - \langle \mathbf{x}_i \rangle)^2 \rangle &= \langle (\mathbf{x} - \langle \mathbf{x} \rangle)^T (\mathbf{x} - \langle \mathbf{x} \rangle) \rangle = \\ \langle \mathbf{q}^T T^T T \mathbf{q} \rangle &= \langle \mathbf{q}^T \mathbf{q} \rangle = \sum_i \langle q_i^2 \rangle = \sum_i \lambda_i.\end{aligned}$$

We choose to sort λ_i in order of decreasing value. Thus the first eigenvectors represent the largest positional deviation, and most of the positional fluctuations reside in a limited subset of the first n eigenvalues, where n is small compared to a total of $3N$.

We now divide the total q -space in an essential subspace:

$$q(1), \dots, q(n),$$

and the remaining space

$$q(n+1) \dots, q(3N)$$

We denote coordinates in the essential subspace by \mathbf{n} and coordinates in the remaining subspace by \mathbf{s} .

It is possible to separate the configurational space into two subspaces:

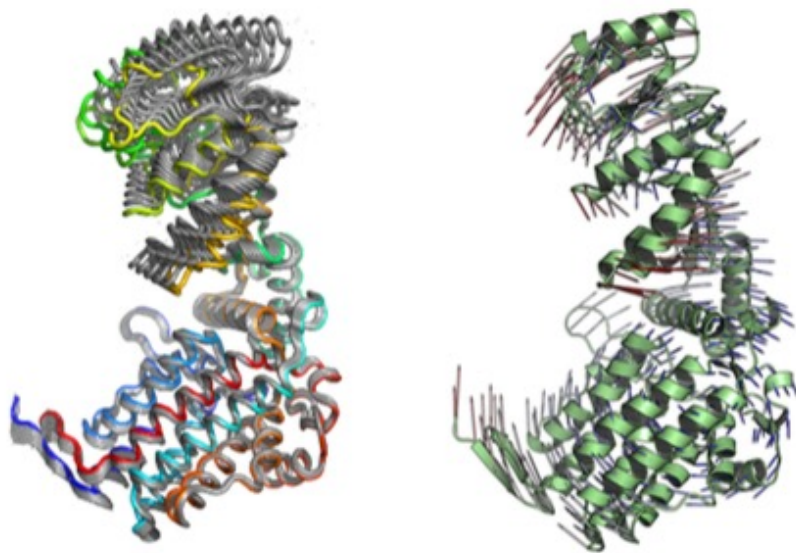
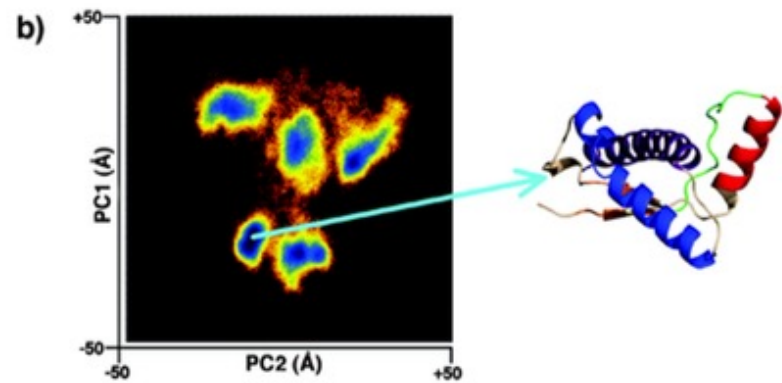
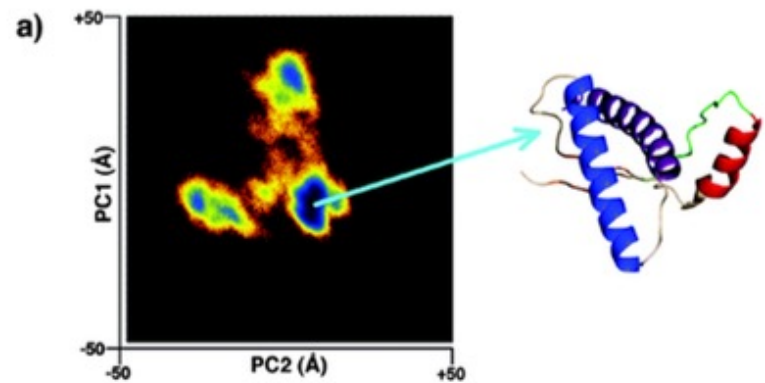
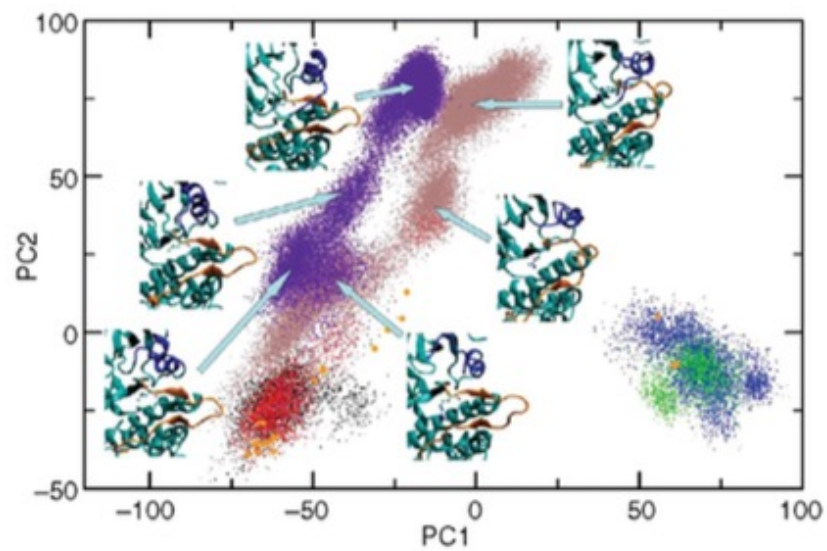
- (1)** an “essential” subspace containing only a few degrees of freedom in which anharmonic motion occurs that comprises most of the positional fluctuations; **(e)**
- (2)** the remaining space in which the motion has a narrow Gaussian distribution and which can be considered as “physically constrained.” **(s)**

The **s-coordinates** behave effectively as constraints: they have narrow Gaussian distributions with zero mean and do not contribute significantly to the positional fluctuations. Thus they behave as harmonic oscillators with a large force constant.

Thus they behave as harmonic oscillators with a large force constant.

This means that the mechanics in the essential subspace can be approximated by setting all $s = 0$, the approximation becoming exact if the **force constants of the s-coordinates tend to infinity.**

The forces in s-space then vanish since the basically independent Gaussian distributions found for the s-coordinates imply that V can be approximated as



Principal component analysis (PCA)

- Purpose of PCA
- Covariance and correlation matrices
- PCA using eigenvalues
- PCA using singular value decompositions
- Selection of variables
- Biplots
- References
- Exercises

Purpose of PCA

The main idea behind the principal component analysis is to represent multidimensional data with fewer number of variables retaining main features of the data. It is inevitable that by reducing dimensionality some features of the data will be lost. It is hoped that these lost features are comparable with the “noise” and they do not tell much about underlying population.

The method PCA tries to project multidimensional data to a lower dimensional space retaining as much as possible variability of the data.

This technique is widely used in many areas of applied statistics. It is natural since interpretation and visualisation in a fewer dimensional space is easier than in many dimensional space. Especially if we can reduce dimensionality to two or three then we can use various plots and try to find structure in the data.

Principal components can also be used as a part of other analysis.

Its simplicity makes it very popular. But care should be taken in applications. First it should be analysed if this technique can be applied. For example if data are circular then it might not be wise to use PCA. Then transformation of the data might be necessary before applying PCA.

PCA is one of the techniques used for dimension reductions.

Covariance and Correlation matrices

Suppose we have $n \times p$ data matrix X :

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

Where rows represent observations and columns represent variables. Without loss of generality we will assume that column totals are 0. If it would not be the case then we could calculate column averages and subtract them from each column. Covariance matrix is calculated using (when column averages are 0):

$$S = \frac{1}{n-1} X^T X = \frac{1}{n-1} \begin{pmatrix} \sum_{i=1}^n x_{i1}x_{i1} & \dots & \sum_{i=1}^n x_{i1}x_{ip} \\ \dots & \dots & \dots \\ \sum_{i=1}^n x_{ip}x_{i1} & \dots & \sum_{i=1}^n x_{ip}x_{ip} \end{pmatrix} = \begin{pmatrix} s_{11} & \dots & s_{1p} \\ \dots & \dots & \dots \\ s_{p1} & \dots & s_{pp} \end{pmatrix}$$

Correlation matrix is calculated using:

$$R = \begin{pmatrix} 1 & \dots & \frac{s_{1p}}{\sqrt{s_{11}s_{pp}}} \\ \dots & \dots & \dots \\ \frac{s_{p1}}{\sqrt{s_{11}s_{pp}}} & \dots & 1 \end{pmatrix} = \text{diag}(S)^{-1/2} S (\text{diag}(S))^{-1/2}$$

I.e. by normalisation of covariance matrix by its diagonals. Both these matrices are symmetric and non-negative.

Principal components as linear combination of original parameters

Let us assume that we have a random vector \mathbf{x} with p elements (variables). We want to find a linear combination of these variables so that variance of the new variable is large. I.e. we want to find new vector y :

$$y = \sum_{i=1}^p a_i x_i$$

so that it has maximum possible variance. It means that this variable contains maximum possible variability of the original variables. Without loss of generality we can assume that mean values of the original variables are 0. Then for variance of y we can write:

$$\text{var}(y) = \text{var}\left(\sum_{i=1}^p a_i x_i\right) = E\left(\sum_{i=1}^p a_i x_i\right)^2 = \sum_{i=1}^p a_i a_j \text{var}(x_i x_j) = \sum_{i=1}^p a_i a_j s_{ij}$$

Thus the problem reduces to finding maximum of this quadratic form.

If found this new variable will be the first principal component.

PCA using eigenvalues

We can write the above problem in a matrix-vector form:

$$\sum_{i=1, j=1}^{p, p} s_{ij} a_i a_j = \mathbf{a}^T \mathbf{S} \mathbf{a} \rightarrow \max$$

But by multiplying to a scalar value this expression (quadratic form) can be made as large as desired. Then we require that length of the vector is unit. I.e. desired vector is on the unit sphere (p-dimensional) that satisfies the condition:

$$\sum_{i=1}^p a_i a_i = \mathbf{a}^T \mathbf{a} = 1$$

Now if we use Lagrange multipliers technique then it reduces to unconditional maximisation of:

$$\mathbf{a}^T \mathbf{S} \mathbf{a} + \lambda(1 - \mathbf{a}^T \mathbf{a}) \rightarrow \max$$

If we get derivative of the left side and equate to 0 we have:

$$\frac{d}{d\mathbf{a}} (\mathbf{a}^T \mathbf{S} \mathbf{a} + \lambda(1 - \mathbf{a}^T \mathbf{a})) = \mathbf{S} \mathbf{a} - \lambda \mathbf{a} = 0 \Leftrightarrow \mathbf{S} \mathbf{a} = \lambda \mathbf{a}$$

Thus the problem of finding unit length vector with largest variance reduces to finding the largest eigenvalue and corresponding eigenvector. If we have largest eigenvalue and corresponding eigenvector then we can find second largest eigenvalue and so on. Finding principal components reduces to finding all eigenvalues and eigenvectors of the matrix S.

PCA and eigenvalues/eigenvectors

Note that since matrix S is symmetric and non-negative definite all eigenvalues are non-negative and eigenvectors are orthonormal (\mathbf{v} -s are the eigenvectors). I.e.:

$$\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

\mathbf{v}_j -s contain coefficient of principal components. They are known as factor loadings.

The $\text{var}(\mathbf{v}_j \mathbf{x}) = \lambda_j$ holds, I.e. variance of the i -th component is i -th eigenvalue. First principal component accounts the largest amount of the variance in the data. $\mathbf{X} \mathbf{v}_j$ gives scores of the n individuals (observation vectors) on this principal component. Relation:

$$\sum_{i=1}^p \lambda_i = \text{tr}(\mathbf{\Lambda}) = \text{tr}(\mathbf{V} \mathbf{\Lambda} \mathbf{V}^T) = \text{tr}(\mathbf{S}) = \sum_{i=1}^p s_{ii}$$

shows that sum of the eigenvalues is equal to the total variance in the data. Where $\mathbf{\Lambda}$ is the diagonal formed by eigenvalues and \mathbf{V} is the matrix formed by the eigenvectors of the covariance (correlation) matrix. Columns of this matrix is called loadings of principal components that is the amount of each variables contribution to the principal component.

When the correlation matrix is used then the total variance is equal to the dimension of the original variables, that is p . Variance of i -th principal component is λ_i . It is often said that this components accounts $\lambda_i / \sum_j \lambda_j$ proportion of the total variance.

Plotting the first few principal components together with observations may show some structure in the data.

PCA using SVD

Since we know that principal component analysis is related with eigenvalue analysis we can use similar techniques available in linear algebra. Suppose that \mathbf{X} is mean centered data matrix. Then we can avoid calculating covariance matrix by using singular value decomposition. If we have the matrix $n \times p$ we can use SVD:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where \mathbf{U} is $n \times n$ \mathbf{V} is $p \times p$ orthogonal matrices. \mathbf{D} is $n \times p$ matrix. p diagonal elements contains square root of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ and all other elements are 0. Rows of \mathbf{V} contains coefficients of the principal components. $\mathbf{U}\mathbf{D}$ contains scores of the principal components that is amount of each observations contribution to the principal components.

Some statistical packages use eigenvalues for principal component analysis and some use SVD.

Another way of applying SVD is using decomposition:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

Where \mathbf{U} is $n \times p$ matrix \mathbf{D} is $p \times p$ diagonal singular values matrix containing square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ and \mathbf{V} is $p \times p$ orthogonal matrix that contains coefficients of principal components. This decomposition is used for bi-plots to visualise data in an attempt to find structure in them.

Scaling

It is often the case that different variables have completely different scaling. For examples one of the variables may have been measured in meters and another one in centimeters (by design or accident). Eigenvalues of the matrix is scale dependent. If we would multiply one column of the data matrix \mathbf{X} by some scale factor (say s) then variance of this variable would increase by s^2 and this variable can dominate whole covariance matrix and hence whole eigenvalue and eigenvectors. It is necessary to take precautions when dealing with the data. If it is possible to bring all data to the same scale using some underlying physical properties then it should be done. If scale of the data is unknown then it is better to use correlation matrix instead of the covariance matrix. It is in general recommended option in many statistical packages.

It should be noted that since scale affects eigenvalues and eigenvectors then interpretation of the principal components derived by these two methods can be completely different. In real life application care should be taken when using correlation matrix. Outliers in the observation can affect covariance and hence correlation matrix. It is recommended to use robust estimation for covariances (in a simple case by rejecting of outliers). When using robust estimates covariance matrix may not be non-negative and some eigenvalues might be negative. In many applications it is not important since we are interested in the principal components corresponding to the largest eigenvalues.

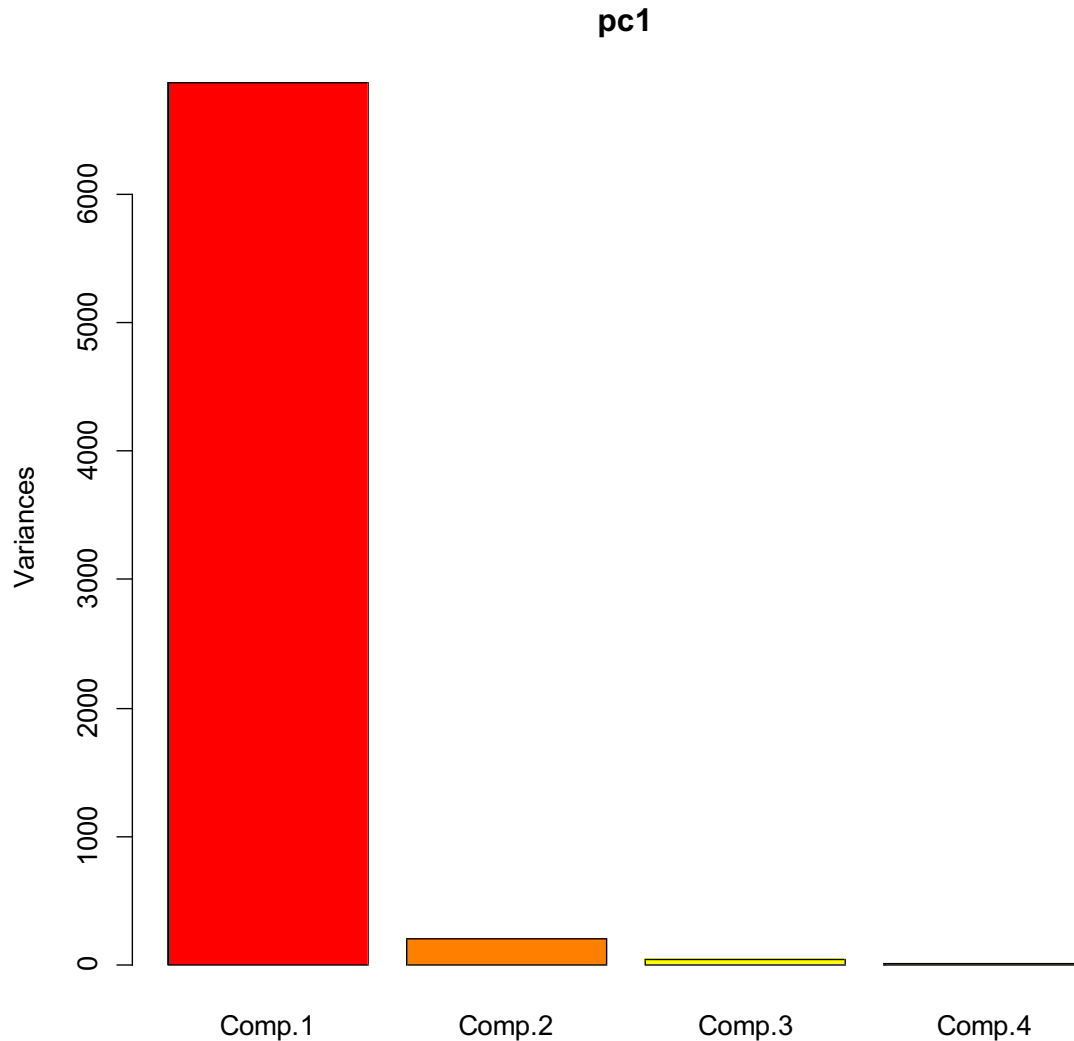
Standard packages allow using covariance as well as correlation matrices. R allows input the data, the correlation or the covariance matrices.

Screeplot

Scree plot is the plot of the eigenvalues (or variances of principal components) against their indices. For example plot given by R.

When you see this type of plot with one dominant eigenvalue (variance) then you should consider

scaling.



Dimension selection

There are many recommendations for the selection of dimension. Few of them are:

1. The proportion of variances. If the first two components account for 70%-90% or more of the total variance then further components might be irrelevant (Problem with scaling)
2. Components below certain level can be rejected. If components have been calculated using correlation matrix often those components with variance less than 1 are rejected. It might be dangerous. Especially if one variable is independent of the others then it might give rise the component with variance less than 1. It does not mean that it is uninformative.
3. If accuracy of the observations is known, then components with variances less than that, certainly can be rejected.
4. Scree plot. If scree plots show elbow then components with variances less than this elbow can be rejected.
5. There is cross-validation technique. One value of the observation is removed (x_{ij}) then using principal components this value is predicted and it is done for all data points. If adding the component does not improve prediction power then this component can be rejected. This technique is computer intensive.

Prediction error calculated using: $PRESS(m) = \frac{1}{np} \sum_{i=1, j=1}^{n,p} (\hat{x}_{ij} - x_{ij})^2$

It is PREdiction Sum of Squares and is calculated using first m principal components.

$$W_m = \frac{PRESS(m-1) - PRESS(m)}{PRESS(m)} \frac{p(n-1)}{n+p-2m}$$

If this value is 1 (some authors recommend 0.9) then only $m-1$ components are selected.

Biplots

Biplots are a useful way of displaying whole data in a fewer dimensional space. It is the projection of observation vectors and variables to $k < p$ dimensional space. How does it work? Let us consider PCA with SVD

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

If we want 2 dimensional biplot then we equate all elements of the \mathbf{D} to 0 but the first two. Denote it by \mathbf{D}^* . Now we have the reduced rank representation of \mathbf{X} :

$$\mathbf{X}^* = \mathbf{U}\mathbf{D}^*\mathbf{V}^T$$

Now we want to find \mathbf{GH}^T representation of data matrix where the rows of \mathbf{G} and the columns of \mathbf{H}^T are scores of the rows and the columns of the data matrix. We can choose them using:

$$\mathbf{G} = \mathbf{U}(\mathbf{D}^*)^\alpha \quad \text{and} \quad \mathbf{H}^T = (\mathbf{D}^*)^{1-\alpha} \mathbf{V}^T$$

The rows of \mathbf{G} and \mathbf{H} are then plotted in biplot. It is usual to take $\alpha=1$. In this case \mathbf{G} and \mathbf{H} are scores of observations on and contribution of variables to principal components. It is considered to be most natural biplot. When $\alpha=0$ then vector lengths corresponding to the original variables are approximately equal to their standard deviations.

R commands for PCA

First decide what data matrix we have and prepare data matrix. Necessary commands for principal component analysis are in the package called mva (in newer version it is in stats package). This package contains many functions for multivariate analysis. First load this package using

library(mva) – loads the library mva

data(USArrests) – loads data

pc1 = princomp(data,cor=TRUE) - It does actual calculations. if **cor** is absent then PCA is done with covariance matrix.

summary(pc1) - gives standard deviations and proportion of variances

pc1\$scores -gives scores of the observation vectors on principal components

pc1\$loadings

screeplot(pc1) - gives scree plot. It plots the values of eigenvectors vs their number

biplot(pc1) – gives biplot.

It would be recommended to use correlation and for quick decision use biplot

References

- 1) Krzanowski WJ and Marriot FHC. (1994) Multivariate analysis. Vol 1. Kendall's library of statistics
- 2) Rencher AC (1995) Methods of multivariate analysis
- 3) Mardia, KV, Kent, JT and Bibby, JM (2003) Multivariate analysis
- 4) Jolliffe, IT. (1986) Principal Component Analysis

Exercises 4

- a) Take data USArrests in R. Use principal component analysis with covariance and correlation matrices. Then try to give interpretation.