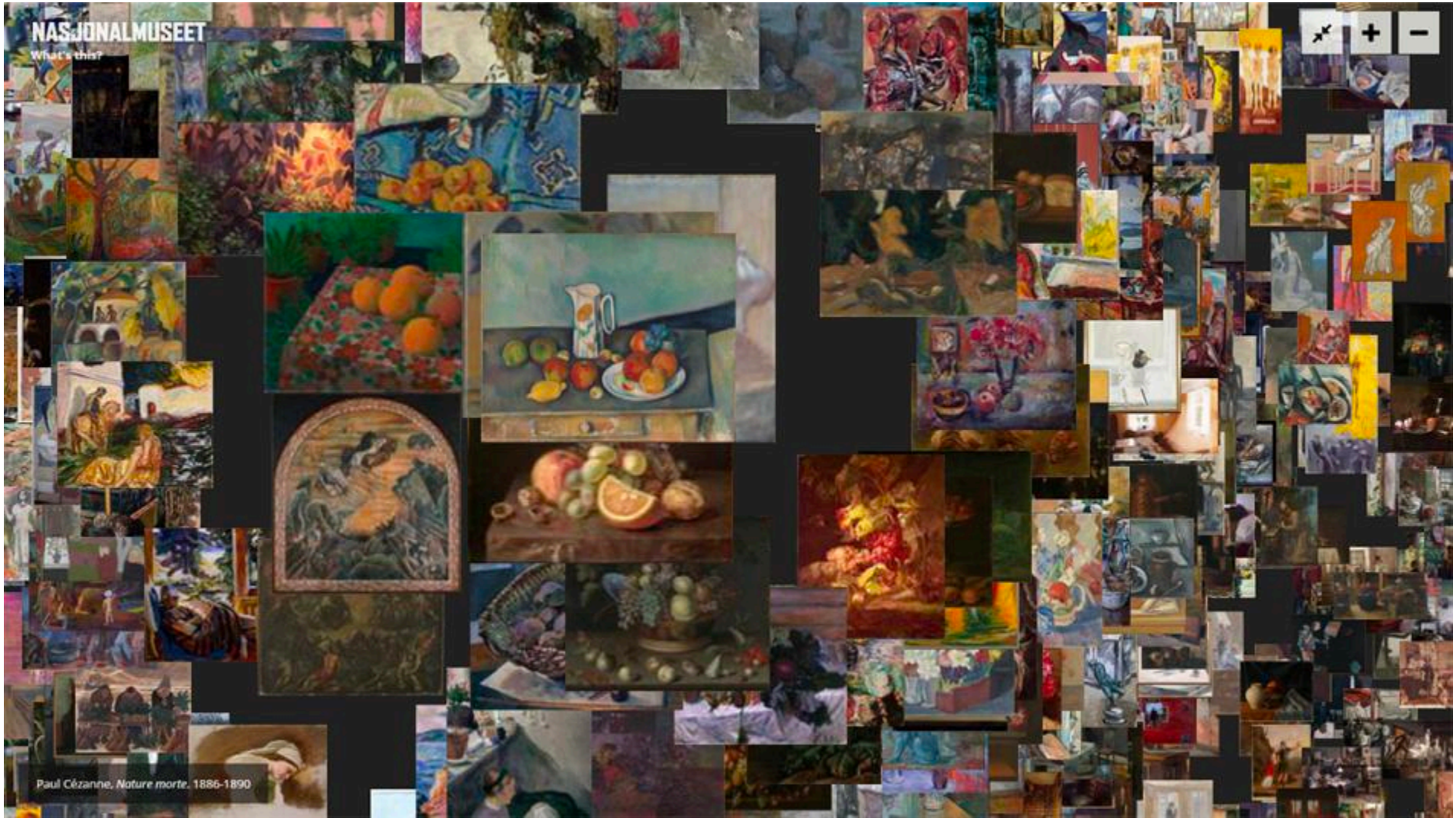
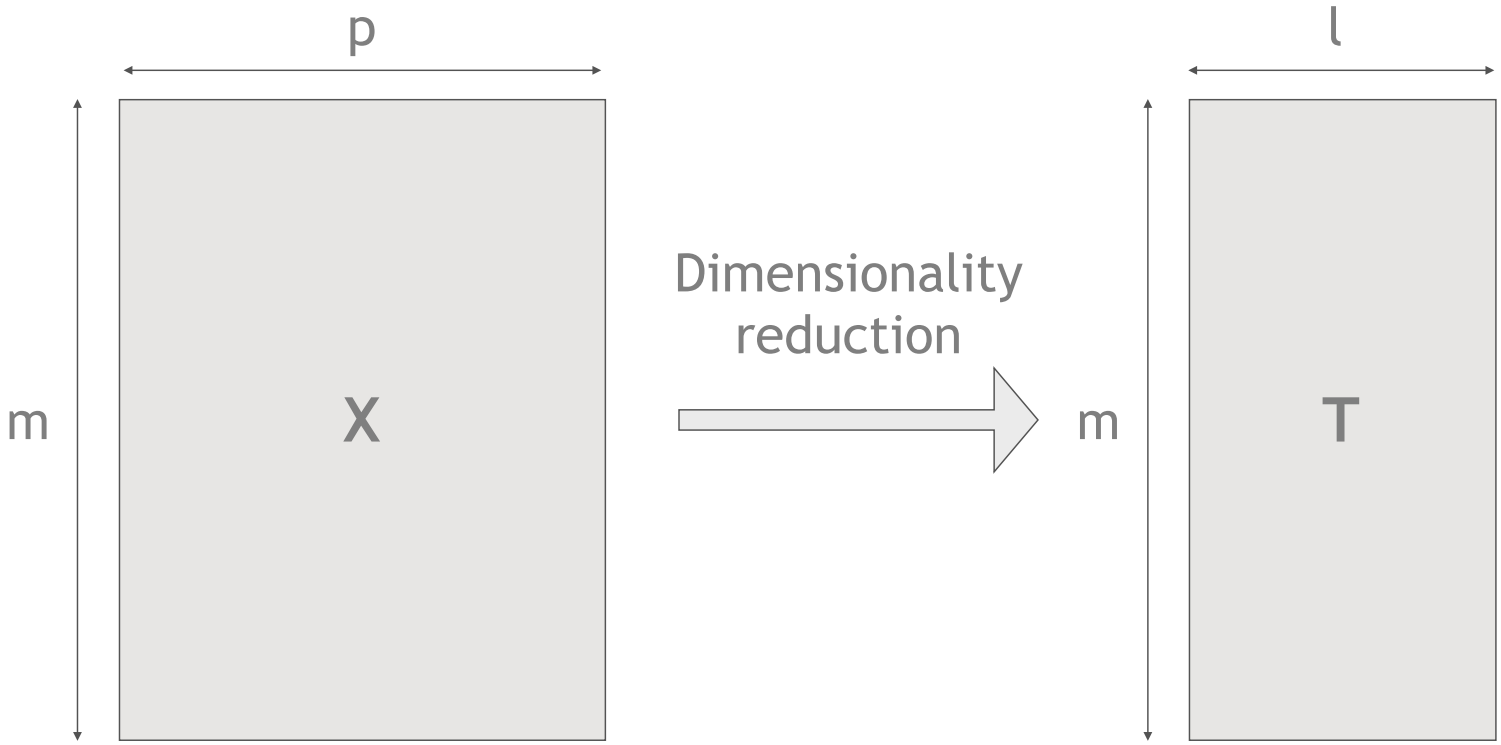

NASJONALMUSEET
What's this?
Paul Cézanne, *Nature morte*, 1886-1890

pca & mds

Dimensionality reduction

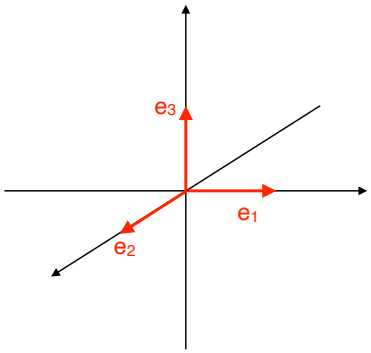


m = number of data samples

real vector spaces

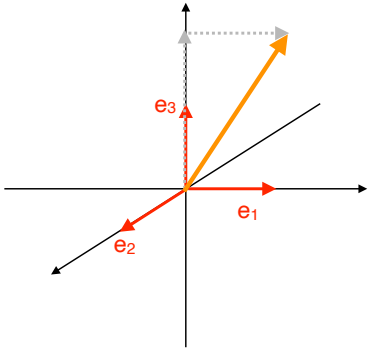
Basis

A basis B of a real vector space V is a subset of linearly independent elements of V that spans V, i.e. every element of V is a linear combination of elements of B.



example

$V=\mathbf{R}^3$ — basis $E=\{e_1=(1,0,0), e_2=(0,1,0), e_3=(0,0,1)\}$



coordinates

each vector in V can be written in a unique way as a linear combination of the basis element;
the coefficients are called coordinates:
$v=(7,-4,5)_E=7e_1+(-4)e_2+5(e_3)$

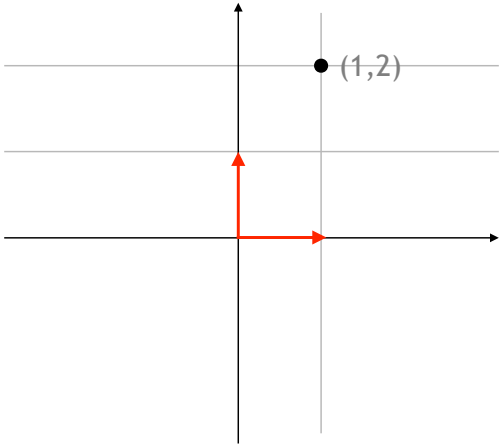
real vector spaces

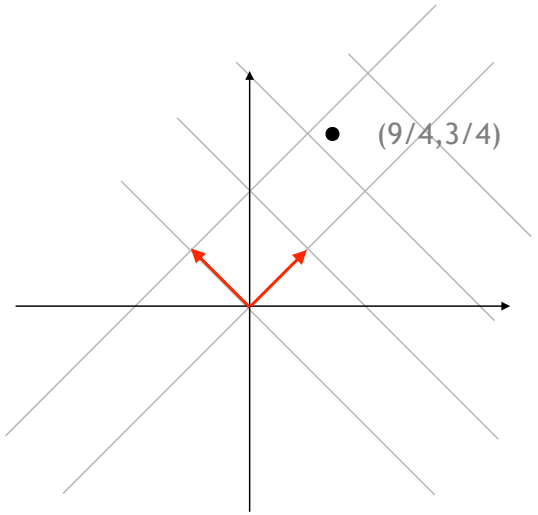
basis change

$B_1=\{v_1,\ldots,v_n\}$, $B_2=\{w_1,\ldots,w_n\}$ bases of V

$$w_j = \sum_{i=1}^{n} a_{ij} v_i \quad \text{if } v = \sum_{i=1}^{n} x_i v_i = \sum_{i=1}^{n} y_i w_i \qquad X = AY \quad A = (a_{ij}) \quad A \in \mathbb{R}^{nxn}$$

Coordinates in basis B_1 — Coordinates in basis B_2

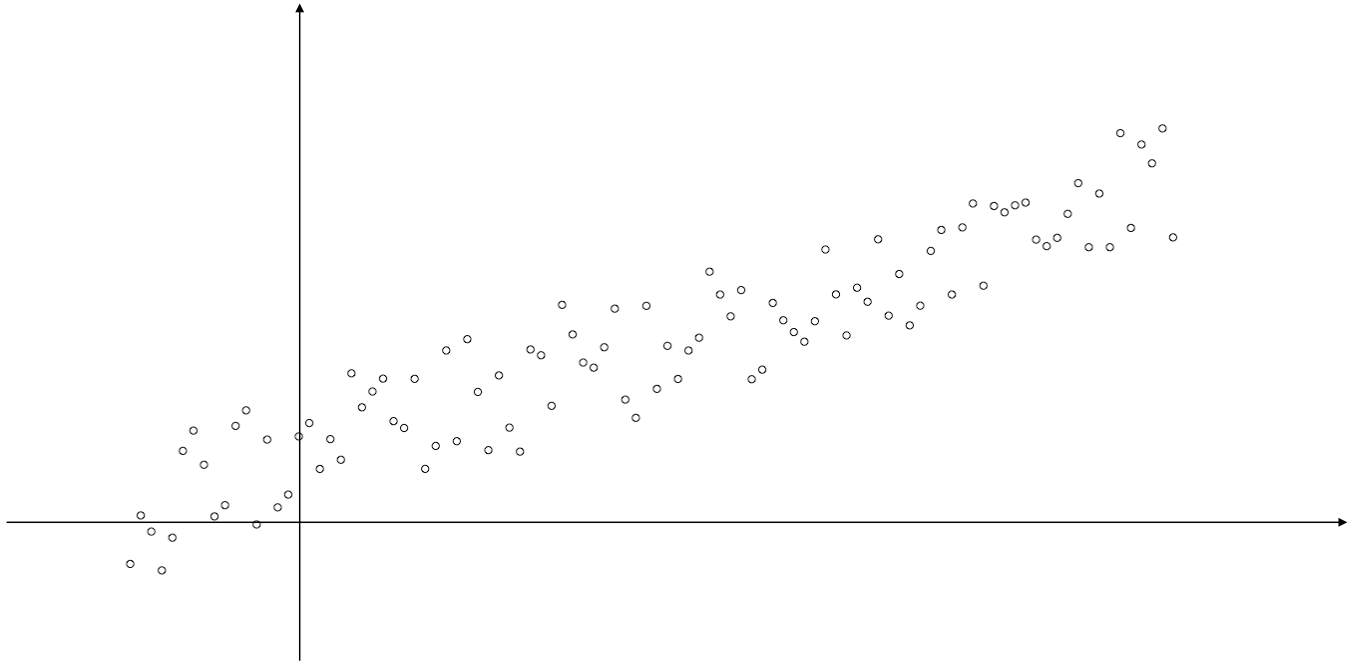




PCA - assumptions

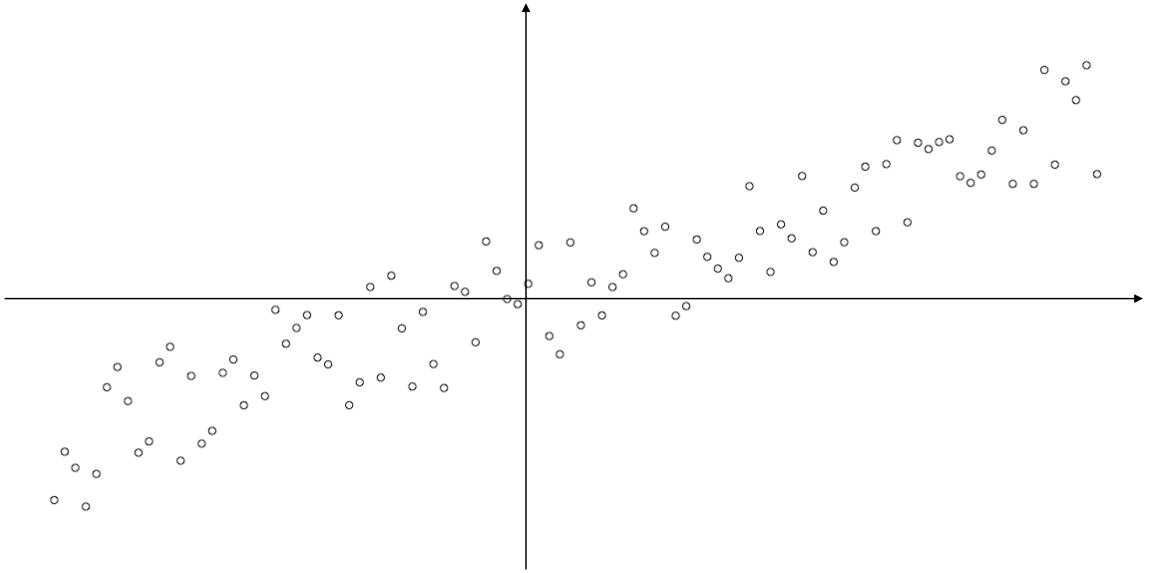
Dataset $X=\{x_1,\ldots,x_n\}$ of n p-dimensional points

All datapoints in $\mathbf{R}^p$ live close to a low-dimensional subspace $\mathbf{R}^l$ of dimension $l<p$



PCA - intuition

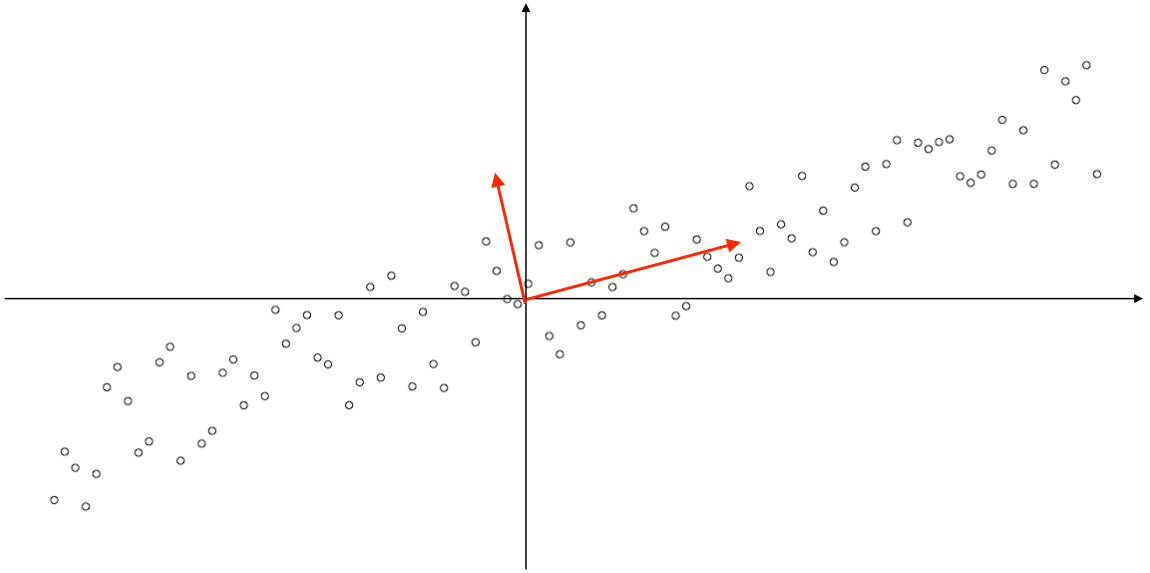
Center the data around zero



PCA - intuition

Center the data around zero

Transform the coordinate system such that the greatest variance of the data lies on the first coordinate of the new system, the second greatest variance on the second coordinate and so on

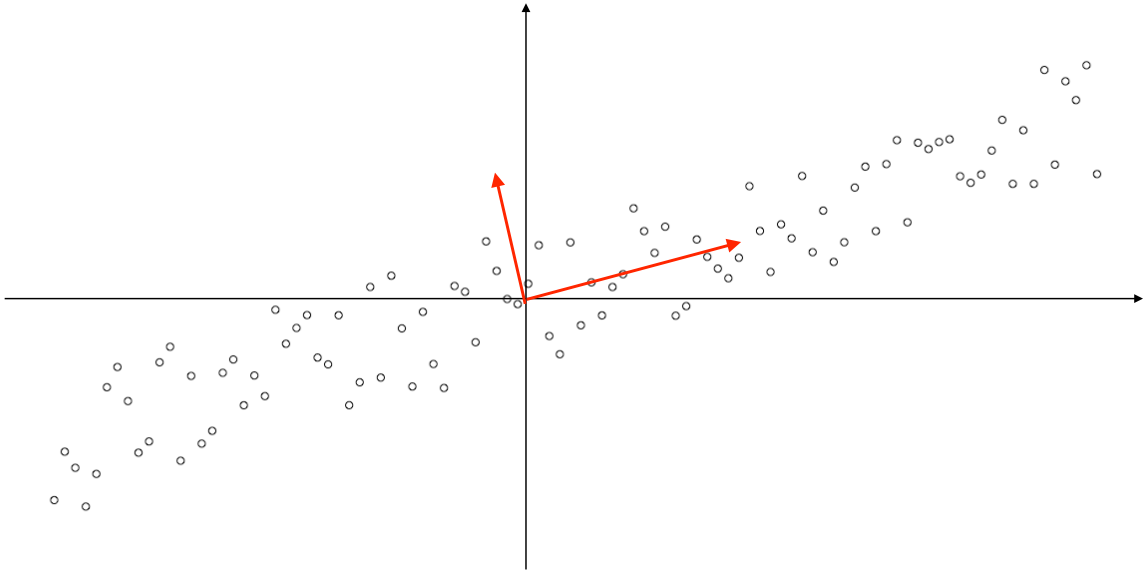


PCA - intuition

Center the data around zero

Transform the coordinate system such that the greatest variance of the data lies on the first coordinate of the new system, the second greatest variance on the second coordinate and so on

Each element can be written as a linear combination of the new system. The new coordinates are called principal components, or components of principal variation



PCA - intuition

Center the data around zero

Transform the coordinate system such that the greatest variance of the data lies on the first coordinate of the new system, the second greatest variance on the second coordinate and so on

Each element can be written as a linear combination of the new system. The new coordinates are called principal components, or components of principal variation.

Each component is defined by the rules:

- being *orthogonal* to the previous components
- having highest possible *variance*

PCA - intuition

Center the data around zero

Transform the coordinate system such that the greatest variance of the data lies on the first coordinate of the new system, the second greatest variance on the second coordinate and so on

Each element can be written as a linear combination of the new system. The new coordinates are called principal components, or components of principal variation.

Each component is defined by the rules:

- being *orthogonal* to the previous components
- having highest possible *variance*

Reduce dimensionality by considering only top L components

setup

$D=\{z_1,\ldots,z_n\}$ dataset of n samples in p variables: $z_i=(z_{i1},\ldots z_{ip})$

center variables: $x_{ij}=z_{ij}-(1/n)\sum_i z_{ij}$

obtain X in $\mathbf{R}^{nxp}$ — data matrix columnwise zero centered

goal: transform $X=\{x_1,\ldots,x_n\}$ into a new dataset $T=\{t_1,\ldots,t_n\}$ in l variables s.t. each sample (row) x_i is mapped into the sample (row) t_i by the matrix : $t_{ik}=x_{i.} \cdot w_{.k}$

$$\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1l} \\ t_{21} & t_{22} & \cdots & t_{2l} \\ \cdots & & & \\ t_{n1} & t_{n2} & \cdots & t_{nl} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1l} \\ w_{21} & w_{22} & \cdots & w_{2l} \\ \cdots & & & \\ w_{p1} & w_{p2} & \cdots & w_{pl} \end{pmatrix}$$

s.t. $t_{.1},\ldots,t_{.l}$ inherit the maximum possible variance from X and $w_{.k}=(w_{1k},\ldots,w_{pk})$ has norm one for each $k=1,\ldots,l$

caveat: pca depends on the chosen scaling

1st component

$$\begin{pmatrix} t_{11} \\ t_{21} \\ \cdots \\ t_{n1} \end{pmatrix} = \begin{pmatrix} x_{11}x_{12}\cdots x_{1p} \\ x_{21}x_{22}\cdots x_{2p} \\ \cdots \\ x_{n1}x_{n2}\cdots x_{np} \end{pmatrix} \cdot \begin{pmatrix} w_{11} \\ w_{21} \\ \cdots \\ w_{p1} \end{pmatrix}$$

$$t = Xw$$

$$\mathrm{Var}(t_{\cdot 1}) = \overline{t_{\cdot 1}^2} - (\overline{t_{\cdot 1}})^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} t_{i1}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}w_{j1}\right)^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}w_{j1}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{j=1}^{p} w_{j1}\sum_{i=1}^{n} x_{ij}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2$$

thus maximising variance means maximising $\sum_{i=1}^{n} t_{i1}^2$

$$\begin{pmatrix} t_{11} \\ t_{21} \\ \cdots \\ t_{n1} \end{pmatrix} = \begin{pmatrix} x_{11}x_{12}\cdots x_{1p} \\ x_{21}x_{22}\cdots x_{2p} \\ \cdots \\ x_{n1}x_{n2}\cdots x_{np} \end{pmatrix} \cdot \begin{pmatrix} w_{11} \\ w_{21} \\ \cdots \\ w_{p1} \end{pmatrix}$$

$$t = Xw$$

1st component

equation to be solved is thus:

$$w_{.1} = \arg \max_{||w_{.1}||=1} \sum_{i=1}^{n} t_{i1}^2 = \arg \max_{||w_{.1}||=1} \sum_{i=1}^{n} (x_{i.} \cdot w_{.1})^2 = \arg \max_{||w_{.1}||=1} ||Xw||^2 = \arg \max_{||w_{.1}||=1} wX^TXw$$

and, since ||w||=1

linear transformation
T:V→V
if T(v)=λv, v is an eigenvector and λ is an eigenvalue;
as matrices, (A-λI)v=0

$$w_{.1} = \arg \max \frac{w^TX^TXw}{w^Tw}$$

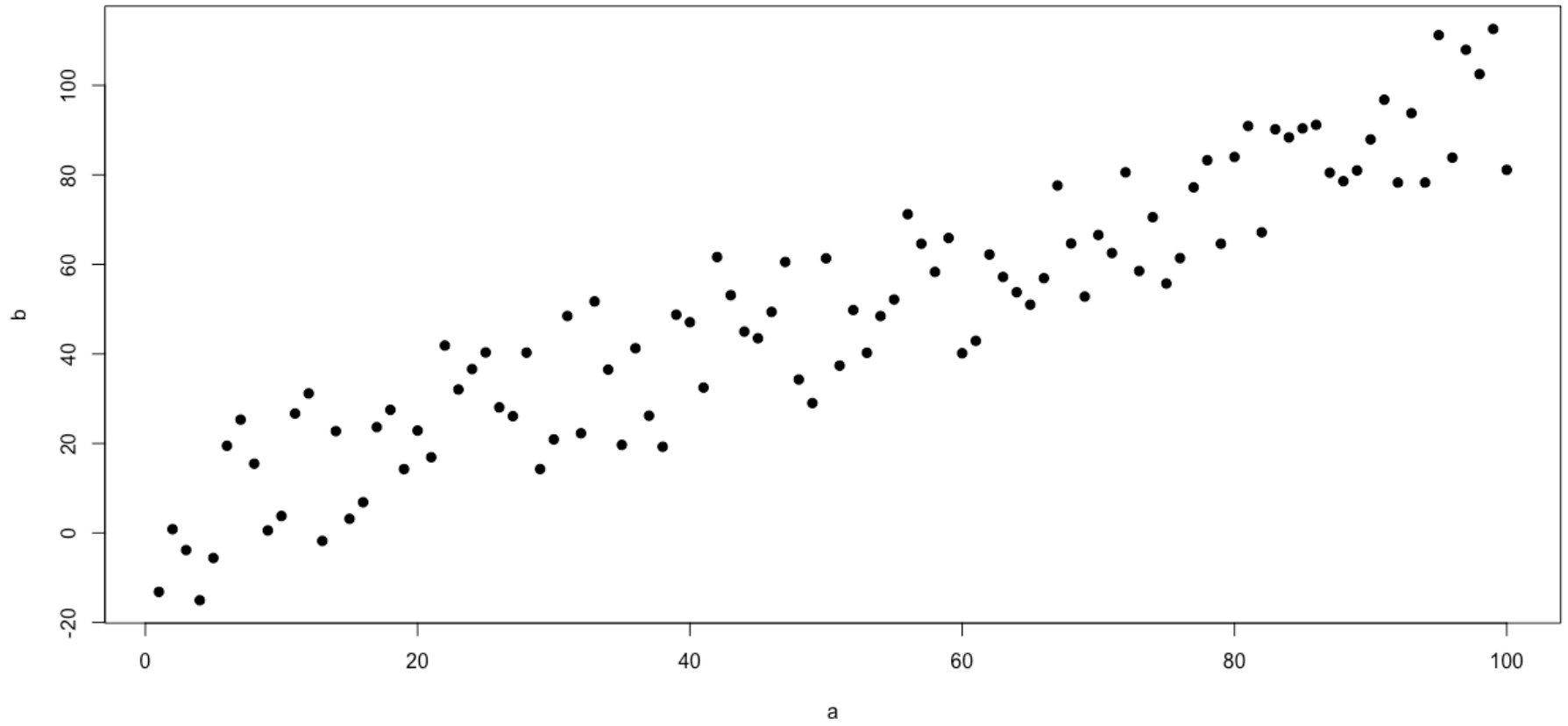
it is known that the solution corresponds to $w_{.1}$ being the eigenvector associated to the largest eigenvalue of the covariance matrix

thus the transformed datum is the real number $t_{i1} = x_{i.} \cdot w_{.1}$ in the new coordinates and the vector $t_{i1}\, w_{.1}$ in the old coordinates

1st component

example

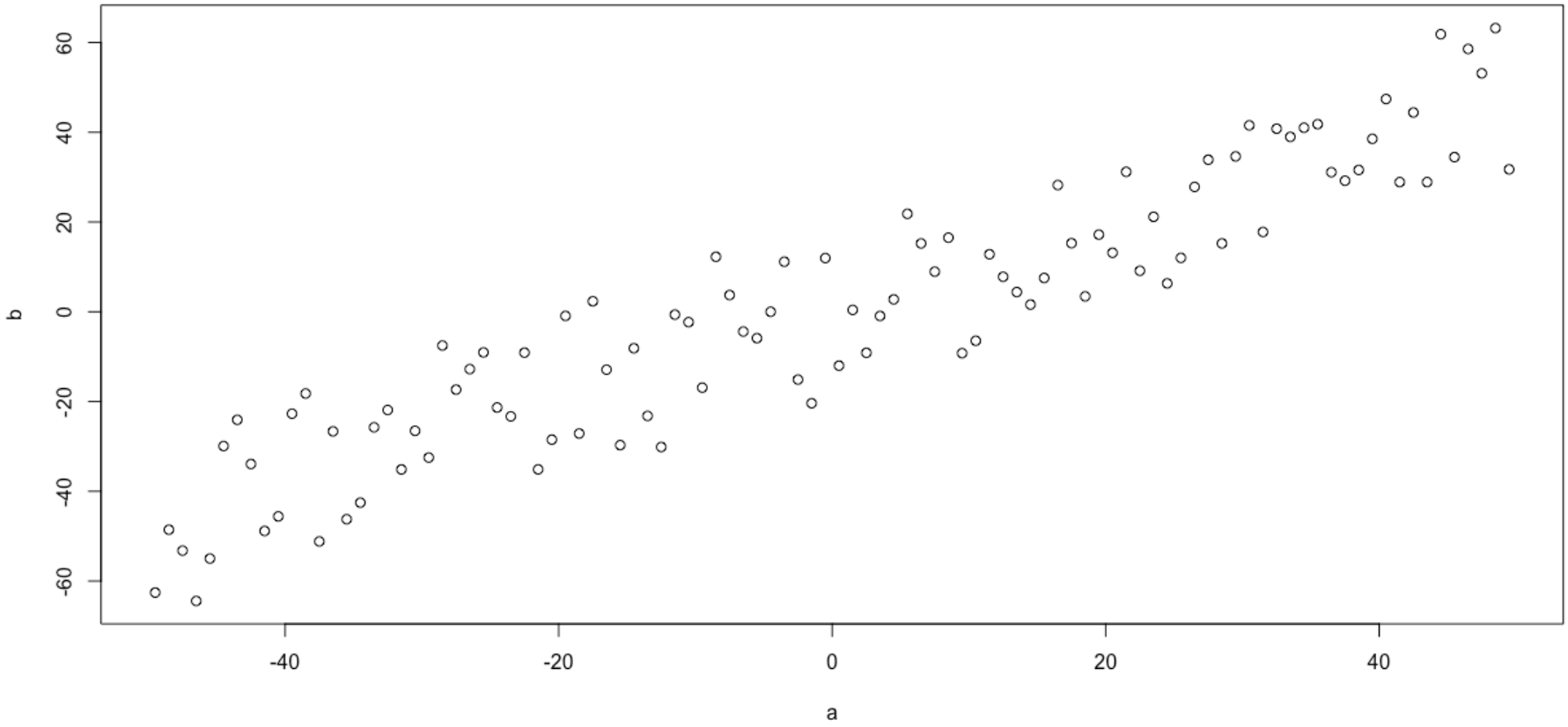
generate a dataset Z with n=100 samples in p=2 dim
distributed with little noise on y=x



1st component

example

center both variables to mean zero obtaining new dataset X

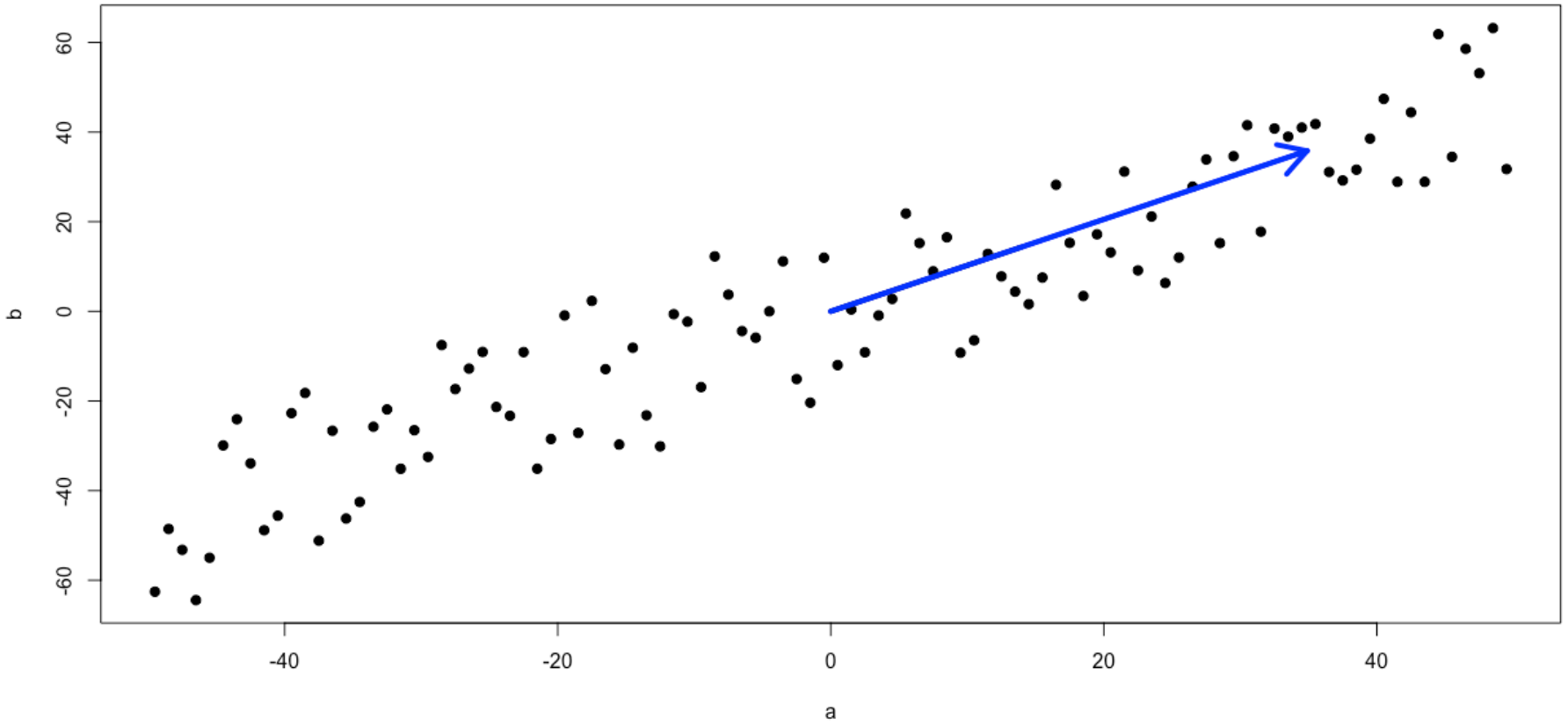


1st component

example

compute first component as the eigenvector corresponding to largest eigenvalue of X^TX:

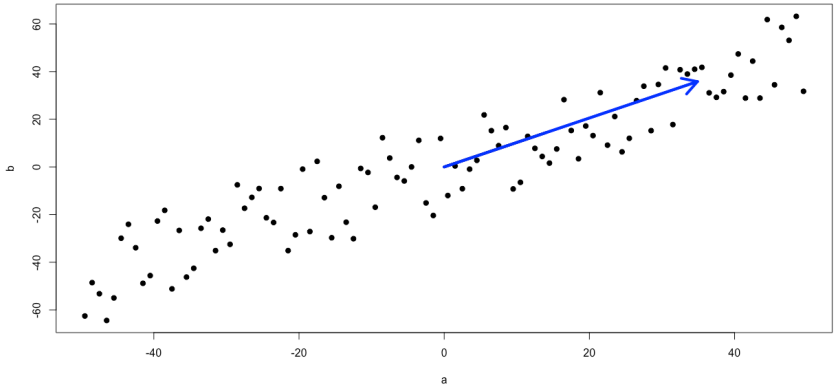
$w_{.1}=(0.6977871, 0.716305)$



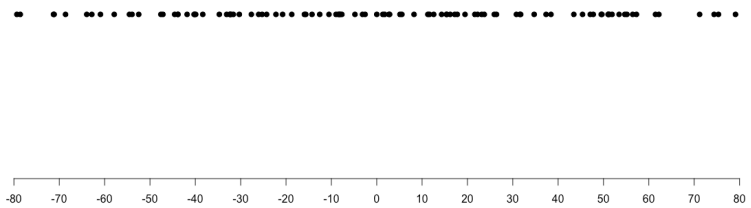
1st component

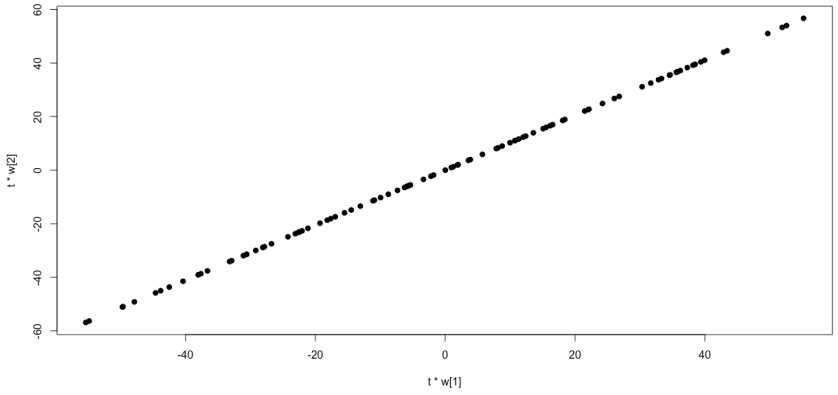
example

read samples on the w coordinates and on the x coordinates



original





$t_{i1} = x_{i.} \cdot w_{.1}$

$t_{i1}\ w_{.1}$

2nd to last components

subtract all previous components and proceed as in the previous case:

$$\hat{X}_k = X - \sum_{s=1}^{k-1} X w_{\cdot s} w_{\cdot s}^T$$

finding the vector with maximal variance:

$$w_{\cdot k} = \arg \max_{||w_{\cdot 1}=1||} ||\hat{X}_k w||^2 = \arg\max \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w}$$

solutions are exactly the remaining eigenvectors of X^TX

the full decomposition is thus given by T=XW with W in $\mathbf{R}^{pxp}$

in practice

three methods are used to compute the pca in practice:

1. covariance

better when data have similar scales

2. correlation

better when data have different scales

3. singular value decomposition

more general purpose

explained variance

the first principal component corresponds to a line that passes through the multidimensional mean and minimizes the sum of squares of the distances of the points from the line

each eigenvalue is proportional to the portion of the "variance" (more correctly of the sum of the squared distances of the points from their multidimensional mean) that is associated with each eigenvector

the sum of all the eigenvalues is equal to the sum of the squared distances of the points from their multidimensional mean

explained variance of k-th pca: $\lambda_k / \sum_i \lambda_i$

pca essentially rotates the set of points around their mean in order to align them with the principal components.
this moves as much of the variance as possible (using an orthogonal transformation) into the first few dimensions

dim reduction

for dimensionality reduction, only the top l components are used

$T_L = XW_L$ with $T_L \epsilon \mathbf{R}^{nxl}$ and $W_L \epsilon \mathbf{R}^{pxl}$

this maximises the variance in the original data that has been preserved, while minimising the total squared reconstruction error $||TW^T - T_L W_L^T||^2$

l=2 finds the two-dimensional plane through the high-dimensional dataset in which the data is most spread out;
if the data contains clusters these too may be most spread out, and therefore most visible to be plotted

also appropriate when the variables in a dataset are noisy:

- much of the signal is in the first few principal components
- later principal components may be dominated by noise, and thus discarded

several non-linear extensions of pca exist

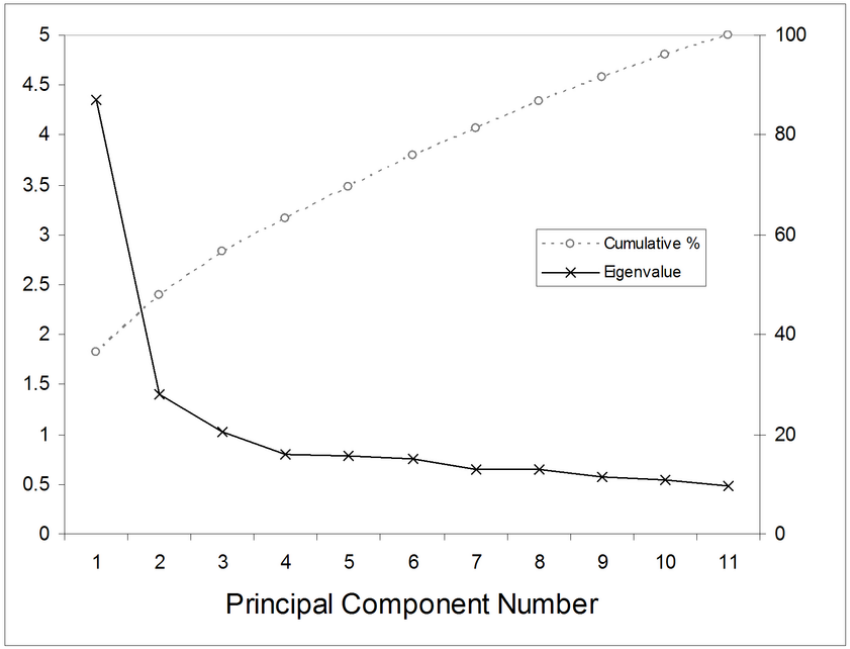
PCA - summary

Linear method

Preserve global structure (consider all data points simultaneously)

Assumptions: data centered around zero and with same scale

Quantifies “goodness” of reconstruction



MDS - multidimensional scaling

multidimensional scaling is a family of algorithms visualizing the level of similarity of individual cases of a dataset

in practice, mds finds an embedding of n objects into a r-dimensional euclidean space so to preserve as well as possible (a function of) the distances between original objects

mds can be distinguished in 3 kinds, depending on the objective function:

- classical mds - the obj fun is called **strain** and involves directly the original distances between objects
- metric mds - the obj fun is called **stress** and involves a function of the original distances
- non-metric mds - the original distances are dissimilarities, so the stress function finds a non-parametric monotonic relationship between the dissimilarities in the item-item matrix and the Euclidean distances between items, and defines the location of each item

Published as a separate and in *The Journal of Psychology*, 1954, **38**, 467-474.

DIMENSIONS OF COLOR VISION*

Department of Psychology, University of Stockholm

GOSTA EKMAN[1]

A. PROBLEM

There have been many theories concerned with color vision, most of them falling into two or three main categories. No one theory seems to account for all the facts of the field—the laws of color mixing, the types of color blindness, etc.

In recent years physiologists have obtained very promising results with methods of nerve fibre analysis. The work of Granit on animals is well known also to psychologists. It clearly demonstrates the existence (in some species) of at least four types (or groups) of receptors, approximately corresponding to the four "unique" colors. These physiological findings will be considered in a subsequent section.

This paper describes a new psychological approach to the problem of primary dimensions of color vision. It is psychological with regard to both problem and method, but an effort will be made to link the results up with the physiological findings.

B. METHOD

The method of *similarity analysis* was developed by the present writer for studies of the dimensionality of experience. Very briefly, its main features may be outlined as follows: Stimuli are presented two at a time. The subject is instructed to rate, on a suitable scale, the degree of subjective similarity between the stimuli. With n stimuli, a similarity matrix of the order $n \times n$ will thus be obtained. The entries of this matrix may be individual data or group data.

The theory of the method is based upon the reasonable assumption that the degree of perceived similarity is a function of the degree of overlap between those primary experiences (sensations, emotions) which are evoked by the stimuli. Under certain conditions the methods of factor analysis may be directly applied to the similarity matrix. The entries of the resulting factor

*Received in the Editorial Office on July 5, 1954, and published immediately at Provincetown, Massachusetts.

[1]This investigation was made possible by a research grant from Magnus Bergvalls Stiftelse. The writer is indebted to Mr. M. Björkman, Mr. S. Borg, and Mr. T. Künnapas for valuable assistance in laboratory and computational work.

example

eckman's colors

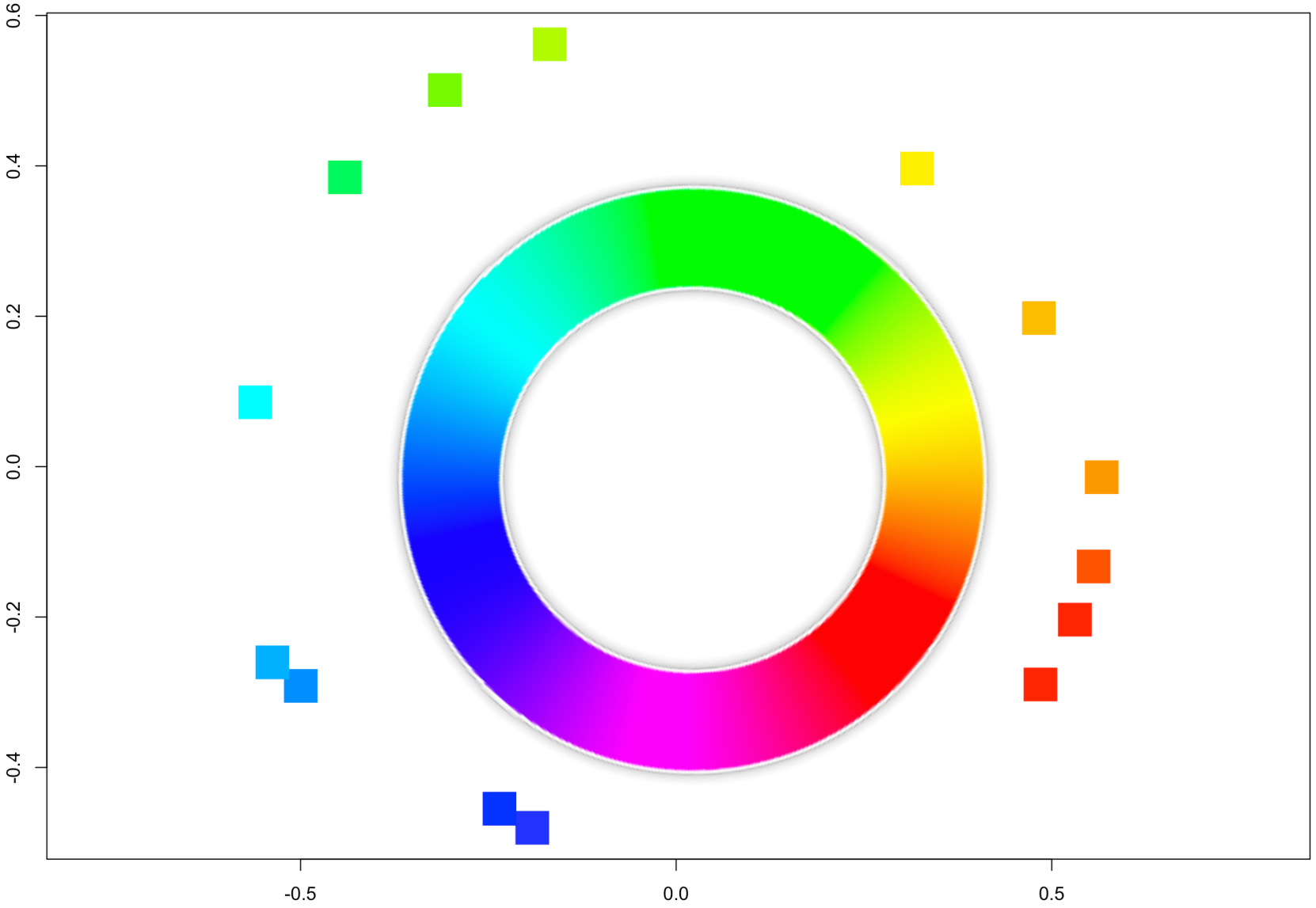
14 colors, different only in hue (wavelength), pairwise rated as similar in a 0-5 scale by 31 people

	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434	0	0,86	0,42	0,42	0,18	0,06	0,07	0,04	0,02	0,07	0,09	0,12	0,13	0,16
445	0,86	0	0,5	0,44	0,22	0,09	0,07	0,07	0,02	0,04	0,07	0,11	0,13	0,14
465	0,42	0,5	0	0,81	0,47	0,17	0,1	0,08	0,02	0,01	0,02	0,01	0,05	0,03
472	0,42	0,44	0,81	0	0,54	0,25	0,1	0,09	0,02	0,01	0	0,01	0,02	0,04
490	0,18	0,22	0,47	0,54	0	0,61	0,31	0,26	0,07	0,02	0,02	0,01	0,02	0
504	0,06	0,09	0,17	0,25	0,61	0	0,62	0,45	0,14	0,08	0,02	0,02	0,02	0,01
537	0,07	0,07	0,1	0,1	0,31	0,62	0	0,73	0,22	0,14	0,05	0,02	0,02	0
555	0,04	0,07	0,08	0,09	0,26	0,45	0,73	0	0,33	0,19	0,04	0,03	0,02	0,02
584	0,02	0,02	0,02	0,02	0,07	0,14	0,22	0,33	0	0,58	0,37	0,27	0,2	0,23
600	0,07	0,04	0,01	0,01	0,02	0,08	0,14	0,19	0,58	0	0,74	0,5	0,41	0,28
610	0,09	0,07	0,02	0	0,02	0,02	0,05	0,04	0,37	0,74	0	0,76	0,62	0,55
628	0,12	0,11	0,01	0,01	0,01	0,02	0,02	0,03	0,27	0,5	0,76	0	0,85	0,68
651	0,13	0,13	0,05	0,02	0,02	0,02	0,02	0,02	0,2	0,41	0,62	0,85	0	0,76
674	0,16	0,14	0,03	0,04	0	0,01	0	0,02	0,23	0,28	0,55	0,68	0,76	0

example

non-metric mds

eckman's colors



distances

let $O=\{o_1,\ldots,o_n\}$ be a dataset of n objects with a dissimilarity measure d_{ij}

mds aims at finding n points $x_1,\ldots,x_n$ in $\mathbf{R}^p$ so that $\|x_i-x_j\| \cong d_{ij}$

If d satisfies:

1. $d(x,y)>0$ for $x\neq y$
2. $d(x,y)=0 \iff x=y$
3. $d(x,y)=d(y,x)$ for all x,y
4. $d(x,y) \leqq d(x,z)+d(z,y)$ for all x,y,z

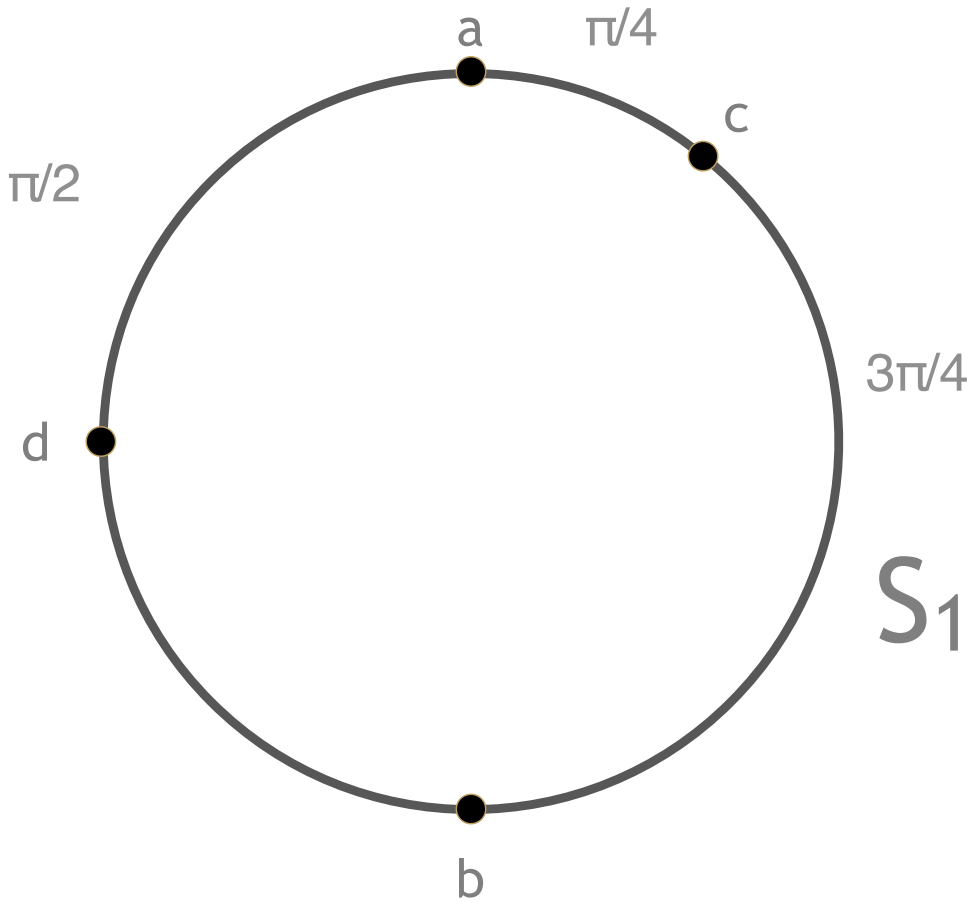
the dissimilarity d is called a distance, or a metric

if for some p $\|x_i-x_j\|=d_{ij}$, d is called an euclidean distance

if no such p exists but d is a metric, d is called a non-euclidean distance

distances

a non-euclidean distance



S_1

d(x,y) is the length of the shortest arc connecting x and y (or the absolute value of the smaller arc measure between x and y in [0,π])

not embeddable in any $\mathbf{R}^p$

π ·

	a	b	c	d
a	0	1	1/4	1/2
b	1	0	3/4	1/2
c	1/4	3/4	0	3/4
d	1/2	1/2	3/4	0

classical mds algorithm

suppose d_{ij} is a (euclidean) distance: find $X=[x_1,\ldots,x_n]\epsilon\mathbf{R}^{nxq}$ so that $\|x_i-x_j\|=d_{ij}$

X is not unique: for any $c\epsilon\mathbf{R}^q$, X+c (rowwise) is also a solution,
since $\|(x_i+c)-(x_j+c)\|= \|x_i-x_j\|=d_{ij}$

usually the centered configuration constraint is added:
$\sum_i x_{ik}=0$ for all columns k (**1**)

(**1**) $\sum_i x_{ik}=0$ for all columns k

(**2**) $b_{ij} = -1/2\ (d^2_{ij} -b_{ii}-b_{jj})$

let $B=XX^T\epsilon\mathbf{R}^{nxn}$ be the Gram matrix
$b_{ij}=x_{i1}x_{j1}+x_{i2}x_{j2}+\ldots+x_{in}x_{jn}=\langle x_i,x_j\rangle$

now $d^2_{ij}=\|x_i-x_j\|^2=\langle x_i-x_j,x_i-x_j\rangle=\langle x_i,x_i\rangle+\langle x_j,x_j\rangle-2\langle x_i,x_j\rangle=b_{ii}+b_{jj}-2b_{ij}$
thus $b_{ij} = -1/2\ (d^2_{ij} -b_{ii}-b_{jj})$ (**2**)

classical mds algorithm

call $T=tr(B)=\sum_k b_{kk}$

because of (**1**), $\sum_i b_{ij}=\sum_i\sum_k x_{ik}\ x_{jk}=\sum_k x_{jk}\sum_i x_{ik}=0$ (**3**)

by (**2**), $\sum_i d^2_{ij}= \sum_i (b_{ii}+b_{jj}-2b_{ij})= \sum_i b_{ii} + \sum_i b_{jj} - 2\sum_i b_{ij}$
by (**3**) = $T + nb_{jj} - 0$,
thus $\sum_i d^2_{ij} = T+nb_{jj}$ (**4**)

(**1**) $\sum_i x_{ik}=0$ for all columns k

(**2**) $b_{ij} = -1/2\ (d^2_{ij} -b_{ii}-b_{jj})$

(**3**) $\sum_i b_{ij}=\sum_i\sum_k x_{ik}\ x_{jk}=\sum_k x_{jk}\sum_i x_{ik}=0$

(**4**) $\sum_i d^2_{ij} = T+nb_{jj}$

(**5**) $\sum_j d^2_{ij} = T+nb_{ii}$

(**6**) $\sum_j\sum_i d^2_{ij} = 2nT$

by symmetry $\sum_j d^2_{ij} = T+nb_{ii}$ (**5**)

by (**2**), $\sum_j\sum_i d^2_{ij}= \sum_j\sum_i (b_{ii}+b_{jj}-2b_{ij})= \sum_j\sum_i b_{ii} + \sum_j\sum_i b_{jj} - 2\sum_j\sum_i b_{ij}$
by (**3**) = $nT + nT - 0$,
thus $\sum_j\sum_i d^2_{ij} = 2nT$ (**6**)

classical mds algorithm

(**1**) $\sum_i x_{ik}=0$ for all columns k

(**2**) $b_{ij} = -1/2\ (d^2_{ij} - b_{ii} - b_{jj})$

(**3**) $\sum_i b_{ij} = \sum_i \sum_k x_{ik}\ x_{jk} = \sum_k x_{jk} \sum_i x_{ik} = 0$

(**4**) $\sum_i d^2_{ij} = T + n b_{jj}$

(**5**) $\sum_j d^2_{ij} = T + n b_{ii}$

(**6**) $\sum_j \sum_i d^2_{ij} = 2nT$

now plugging (**4**), (**5**) & (**6**) in (**2**):

$$
\begin{aligned}
b_{ij} &= -\frac{1}{2}\left(d_{ij}^2 - b_{ii} - b_{jj}\right) \\
&= -\frac{1}{2}\left(d_{ij}^2 - \frac{\sum_{j=1}^{n} d_{ij}^2 - T}{n} - \frac{\sum_{i=1}^{n} d_{ij}^2 - T}{n}\right) \\
&= -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n}2T\right) \\
&= -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right) \qquad (7)
\end{aligned}
$$

classical mds algorithm

find now a matrix form for (**7**) by defining the matrices $D_2=(d^2_{ij})$
and **O** in $\mathbf{R}^{nxn}$ and **1** in $\mathbf{R}^{nx1}$

$$\mathbb{O}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ & & \cdots & \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} \cdot (11\cdots1) = \mathbf{1} \cdot \mathbf{1}^T$$

then (**7**) reads:
$B=(-1/2)(D_2 - (1/n)\mathbf{O}D_2 - (1/n)D_2\mathbf{O} + (1/n)\mathbf{O}D_2(1/n)\mathbf{O})$

that, introducing the centering matrix $C_n=\mathbf{I}_n-(1/n)\mathbf{1}\mathbf{1}^T$, becomes
$B=(-1/2)\ C_nD_2C_n$ (**8**)

C_n is called centering matrix because for any Y, C_nYC_n has rowwise and columnwise sum equal to zero

classical mds algorithm

but $B=XX^T$, so it is real symmetric: $B=B^T$;
let $v \neq 0$ be an eigenvector for the eigenvalue $\lambda \epsilon \mathbf{C}$, so $Bv=\lambda v$

$$\bar{v}^T Bv = \bar{v}^T(Bv) = \bar{v}^T(\lambda v) = \lambda(\bar{v}^T \cdot v)$$
$$=$$
$$\bar{v}^T Bv = (B\bar{v})^T v = (\bar{\lambda}\bar{v})^T v = \bar{\lambda}(\bar{v}^T \cdot v)$$

then λ is real; let $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_n$ the n real eigenvalues of B and $z_1,\ldots,z_n$ the corresponding eigenvectors

then B can be *diagonalised*, i.e. written as
$B = V \Lambda V^T$
where Λ is the diagonal matrix of the eigenvalues, and V is the orthogonal matrix ($V^{-1}=V^T$) with the corresponding eigenvectors as columns.

classical mds & dr

thus we have
$(-1/2)\ C_n D_2 C_n = B = XX^T = V\ \Lambda V^T$

which yields
$X = V\ \Lambda^{1/2}$

the solution is then unique, and the above algorithm is constructive

for non-euclidean distances, only approximation if possible

for dimensionality reduction, since the first $p<q$ components best preserve the distances among all other p-dim reductions, one can use

$X_{(p)} = V_{(p)}\ \Lambda^{1/2}_{(p)}$

where $\Lambda^{1/2}_{(p)}$ is the pxp diagonal matrix with the p largest √eigenvalues of B, and $V_{(p)}$ is collected through the first p columns of V

example 1

unitary tetrahedron

distance matrix of the four vertices

$$D_{(2)} = D = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$B = (1/8) \cdot \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

$$\Lambda = (1/2) \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

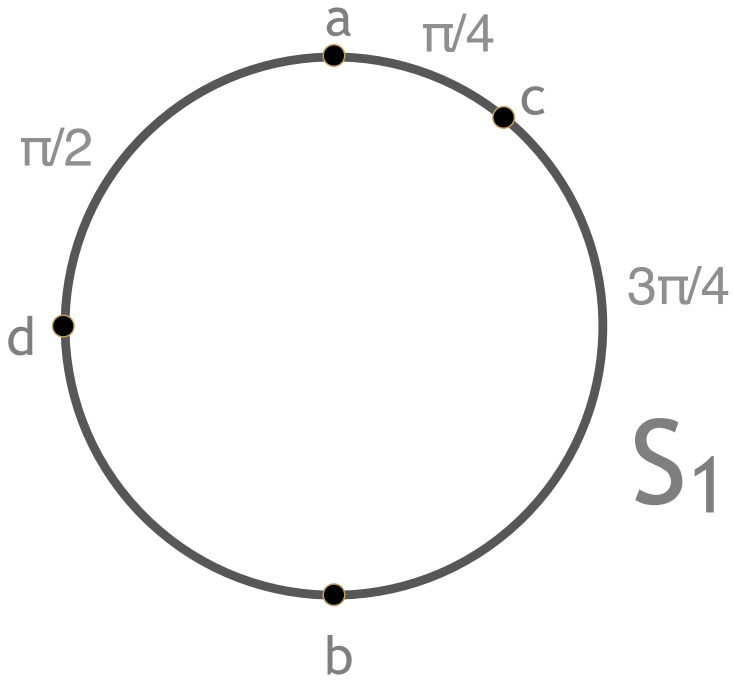
restricting to $\mathbf{R}^3$

$$X_{(3)} = V_{(3)}\Lambda_{(3),\frac{1}{2}} = \begin{pmatrix} 0.0000000 & 0.6123724 & 0.0000000 \\ -0.1893048 & -0.2041241 & 0.5454329 \\ -0.3777063 & -0.2041241 & -0.4366592 \\ 0.5670111 & -0.2041241 & -0.1087736 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

A,B,C,D are exactly the vertices of regular tetrahedron of edge 1 in $\mathbf{R}^3$

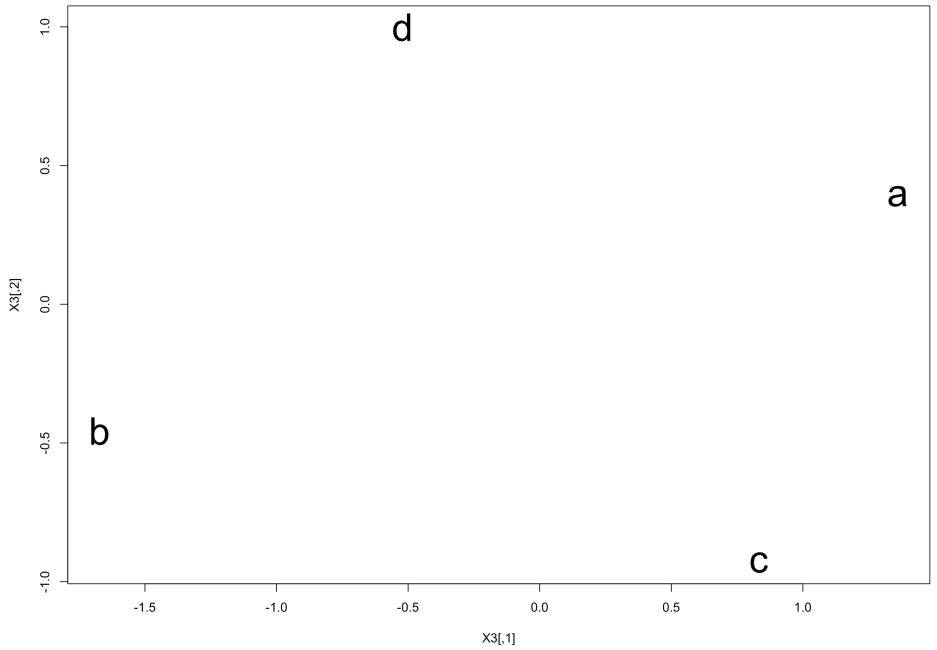
example 2

circular distance



$D=\pi \cdot$

	a	b	c	d
a	0	1	1/4	1/2
b	1	0	3/4	1/2
c	1/4	3/4	0	3/4
d	1/2	1/2	3/4	0



spec(B)={5.6117, 2.2234 ,0,–1.2039)

thus we need to restrict to $X_{(2)}$

$dist=\pi \cdot$

	a	b	c	d
a	0	1,00	0,46	0,63
b	1,00	0	0,81	0,59
c	0,46	0,81	0	0,75
d	0,63	0,59	0,75	0

alternative mds algorithms

relax the objective $\|x_i - x_j\| \cong d_{ij}$ to $\|x_i - x_j\| \cong f(d_{ij})$

metric/non-metric mds depending on d being quantitative or not (e.g. ordinal)

becomes an optimisation process aimed at minimising a stress function, solved by iterative algorithms

metric mds

given a dimension p and a monotone function f, metric mds aims at finding $X=\{x_1,\ldots,x_n\}$ such that $\|x_i-x_j\| \cong f(d_{ij})$ as close as possible, i.e. minimising a chosen loss function (stress) $\mathscr{L}$

example 1:

$f(d_{ij})=\alpha d_{ij}+\beta$ and $\mathscr{L}= (\sum_{i<j} (\|x_i-x_j\|-(\alpha d_{ij}+\beta))^2 / \sum_{i<j} d^2_{ij})^{(1/2)}$

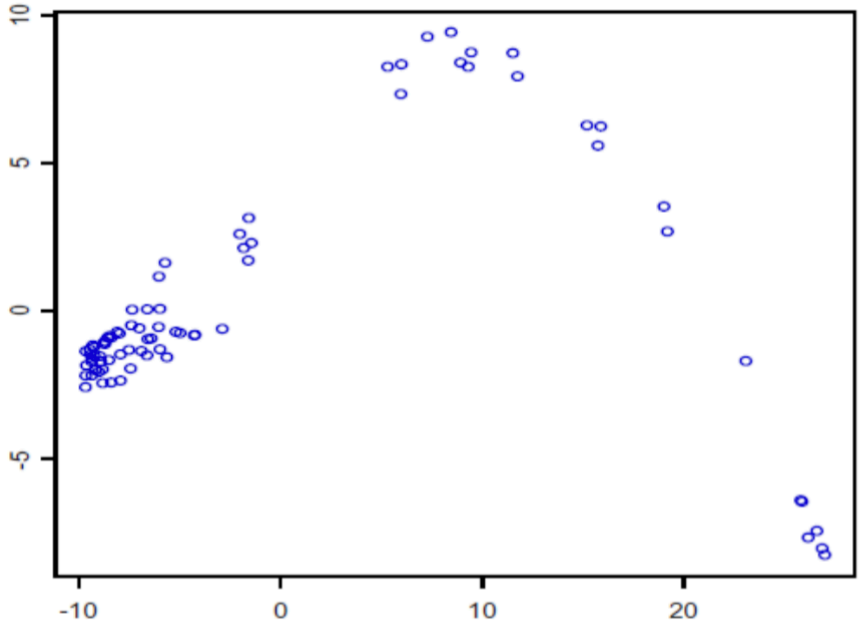
note that $\alpha=1$ and $\beta=0$ gives the classical case, with a different solution

example 2: sammon mapping

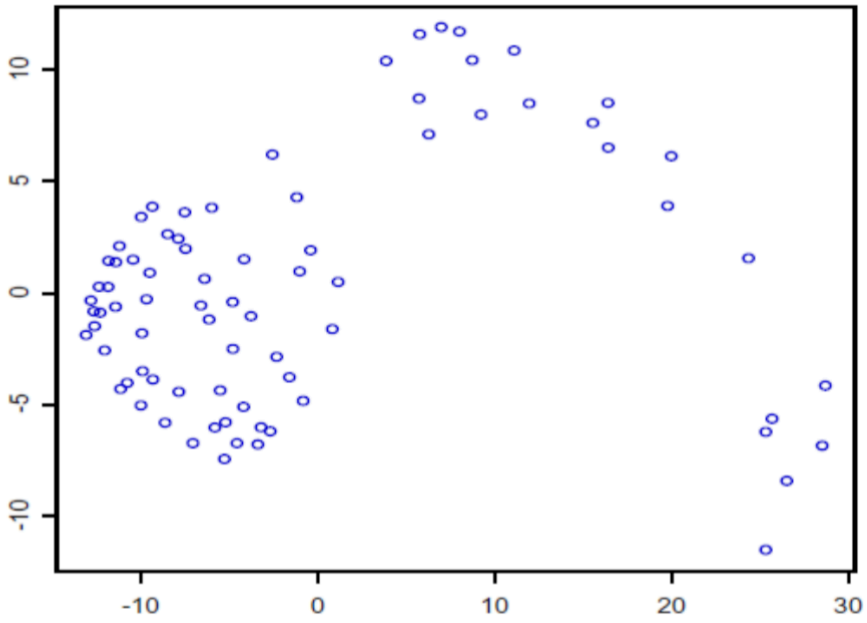
$$\mathscr{L} = \frac{1}{\sum_{l<k} d_{lk}} \sum_{i<j} \frac{(||x_i - x_j|| - d_{ij})^2}{d_{ij}}$$

sammon mapping preserves the small d_{ij}, giving them a greater degree of importance in the fitting procedure than for larger values of d_{ij}

sammon mapping



classical



sammon

sammon mapping better preserves inter-distances for smaller dissimilarities, while proportionally squeezes the inter-distances for larger dissimilarities.

non-metric mds

when dissimilarities are known only by their rank order, and the spacing between successively ranked dissimilarities is of no interest or is unavailable

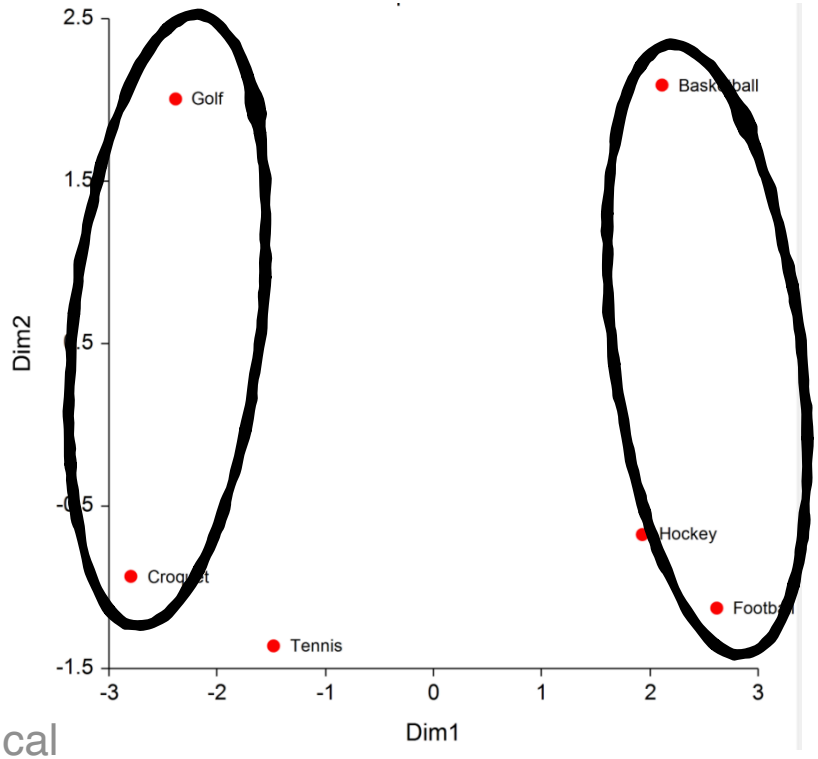
here f is only implicitly defined as a regression curve, and only preserves the order of d, that is $f(d_{ij})<f(d_{kl})$ if $d_{ij}<d_{kl}$
thus only the order of d is needed, not the actual values

most common algorithm is the Kruskal mds

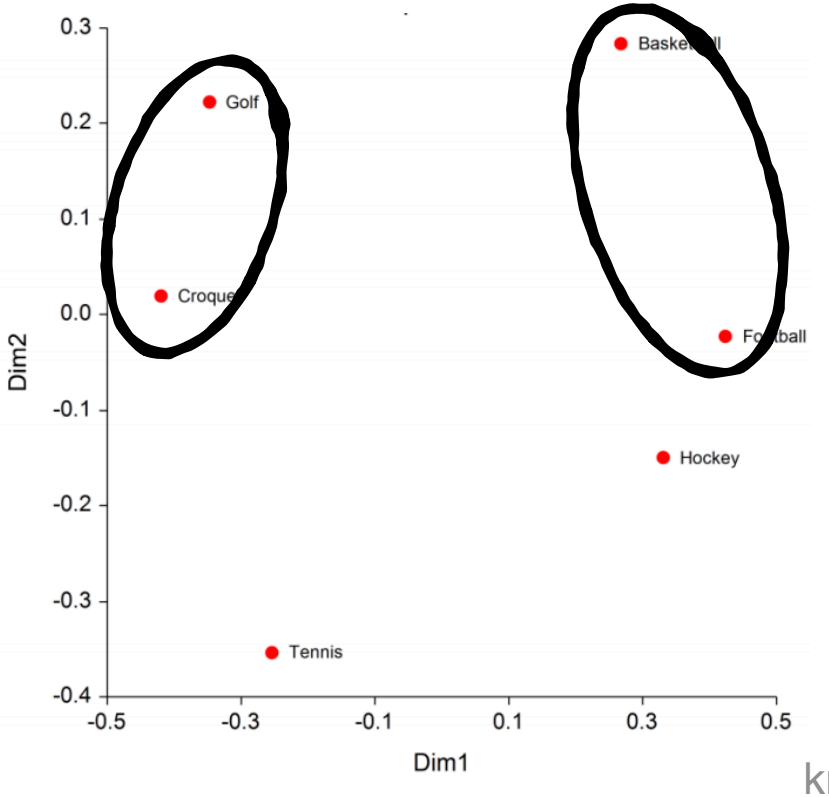
kruskal mds

dissimilarity rating between sports

Sport	Hockey	Football	Basketball	Tennis	Golf	Croquet
Hockey	0	2	3	4	5	5
Football		0	3	5	6	5
Basketball			0	5	4	6
Tennis				0	4	3
Golf					0	2
Croquet						0



classical



kruskal

MDS - summary

Linear or Non-linear method

Preserve local structure (consider pairwise distances between observations)

Assumptions: euclidean distance for classical mds

Isomap

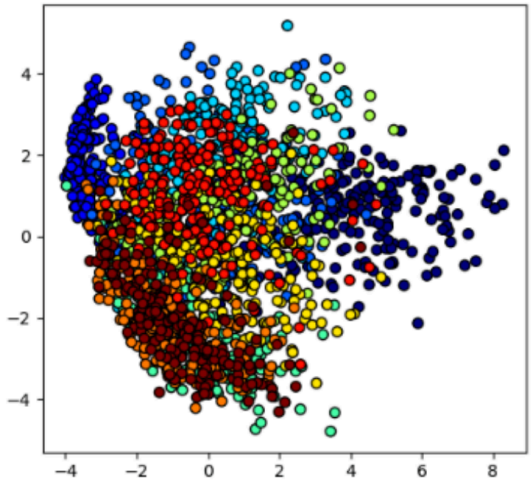
Extension of mds

Uses geodesic distances instead of euclidean distances

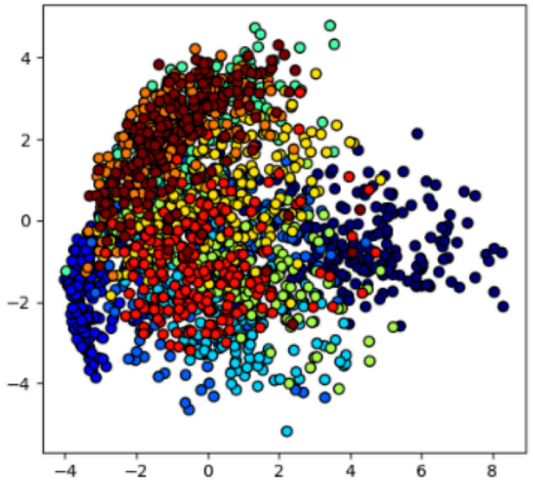
1) Construct a graph based on nearest neighbours
2) Computes shortest path (geodesic distance) between graph nodes
3) Applies mds using the geodesic distances as dissimilarities

MNIST dataset

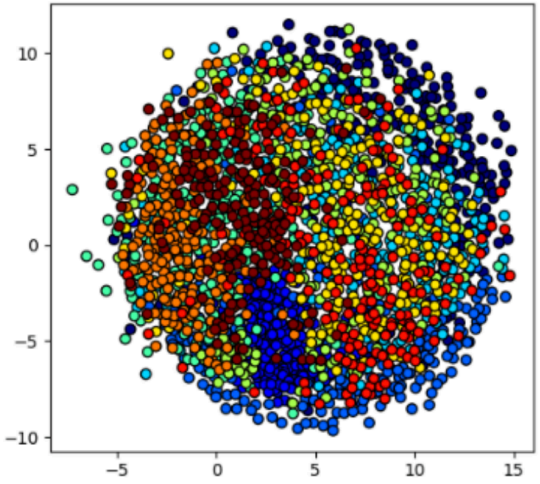
PCA



Classical MDS



Sammon MDS



Isomap