Metoda glavnih osi (Principal Component Analysis- PCA)

Uvod

Metoda glavnih osi (PCA) je linearna transformacija (rotacija) *m*-dimenzionalnega merskega prostora. Prva koordinata zasukanega koordinatnega sistema (glavna os ali 1. PC) je tista smer, v kateri se celotna varianca sistema vseh obravnavanih podatkov najbolj zmanjša. Naslednja glavna os je pravokotna na prejšnjo in je zopet določena s tem, da opiše kar največ preostale variance, itd. Ker je tako velika večina celotne variance podatkov zbrana okog nekaj prvih novih osi, lahko ostale osi zanemarimo, posebej, če upoštevamo samo tiste, ki so skupno odgovorne za več kot npr. 90 % variance. Ker so v veliki večini kemijskih problemov spremenljivke med seboj odvisne, lahko že prvi dve glavni (novi) osi nosita več kot 75 % variance. Glavne osi so linearne kombinacije starih spremenljivk:

$$PC_1 = l_{11}x_1 + l_{12}x_2 + \ldots + l_{1i}x_i + \ldots + l_{1m}x_m = \sum_{i=1}^{m} l_{1i}x_i \rightarrow \quad prva\ glavna\ os$$

$$splošno: PC_j = \sum_{i=1}^{m} l_{ji}x_i \rightarrow \quad j-ta\ glavna\ os$$

$$\sum_{i=1}^{m} l_{ij}^2 = 1 \rightarrow za\ vsako\ os\ PC_j \qquad \sum_{i=1}^{m} l_{ij}l_{ik} = 0 \rightarrow za\ vsak\ poljubni\ par\ PC_j\ in\ PC_k$$

Koeficienti l_{ij} (*loadings*) povedo, koliko je vsaka originalna (stara) spremenljivka x_i udeležena v posamezni novi osi PC_j.

Ker je večina informacije (variance) skoncentrirana okoli prvih dveh novih osi, lahko PCA služi kot projekcija objektov iz *m*-dimenzionanega prostora v 2d prostor prvih dveh glavnih osi (projekcijo originalnih objektov lahko naredimo v 2d prostor katerih koli drugih dveh novih osi).

Izhodišče za izračun glavnih osi oziroma koordinatnega sistema v katerem so osi po vrsti usmerjene tako, da pojasnijo kar največ variance, je ***matrika korelacij*** med vsemi spremenljivkami. Če med spremenljivkami ni korelacij, torej v primeru, ko imamo opraviti z med seboj povsem neodvisnimi spremenljivkami, glavnih osi ne moremo izračunati!

Podatki za izračun glavnih osi

Čeprav ima danes večina statističnih programov, zlasti tistih, ki so namenjeni kemikom, metodo PCA že vgrajeno, si vseeno oglejmo matematični postopek izračuna glavnih osi. Določitev glavnih osi pričnemo z *nxm* veliko matriko, ki je izračunana na podlagi *n* izmerjenih podatkov ***X***. Posamezni objekt *k* je predstavljeni v *m*-dimenzionalnem prostoru kot vektor $X_k=(x_{k1}, x_{k2}, \ldots, x_{ki}, \ldots, x_{kj}, \ldots, x_{km})$. Za izračun PCA lahko matriko podatkov ***X*** predhodno preoblikujemo na tri načine. Preoblikovano matriko bomo pisali kot matriko A_i:

$X_1 \rightarrow x_{11}, x_{12}, \ldots, x_{1i}, \ldots, x_{1j}, \ldots, x_{1m}$

$X_2 \rightarrow x_{21}, x_{22}, \ldots, x_{2i}, \ldots, x_{2j}, \ldots, x_{2m}$

$X_k \rightarrow x_{k1}, x_{k2}, \ldots, x_{ki}, \ldots, x_{kj}, \ldots, x_{km}$

$X_n \rightarrow x_{n1}, x_{n2}, \ldots, x_{ni}, \ldots, x_{nj}, \ldots, x_{nm}$

$\bar{X} \rightarrow \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_i, \ldots, \bar{x}_j, \ldots, \bar{x}_m$

$\sigma_1, \sigma_2, \ldots, \sigma_i, \ldots, \sigma_j, \ldots, \sigma_m$

$v_i=\sigma^2_i$

$$A_1 = \|x_{ki}\| \rightarrow \textit{originalni podatki}$$

$$A_2 = \|x_{ki} - \bar{x}_i\| \rightarrow \textit{podatki premaknjeni za povprecje}$$

$$A_3 = \left\| \frac{x_{ki} - \bar{x}_i}{\sigma_i} \right\| \rightarrow \textit{normalizirani podatki } (\bar{x}_i = 0, \sigma = 1)$$

$$\text{varianca} \rightarrow v_i = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2}{n-1}$$

$$\text{kovarianca} \rightarrow v_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{n-1}$$

$$\textit{korelacija} \rightarrow r_{ij} = \frac{v_{ij}}{\sqrt{v_{ii} v_{jj}}} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2 \sum_{i=1}^{n} (x_{kj} - \bar{x}_j)^2}}$$

Z množenjem vsake od omenjenih treh matrik A_1, A_2, ali A_3 same s seboj v transponirani obliki (indeks ^{T}!), dobimo tri različne izhodne matrike. Iz produkta A^tA lahko z diagonalizacijo izračunamo glavne osi PC_i ($i = 1 \ldots m$).

$X = (A_1)^T A_1$ je matrika mešanih produktov faktorjev

$V = (A_2)^T A_2$ je matrika varianc in kovarianc

$R = (A_3)^T A_3$ je korelacijska matrika

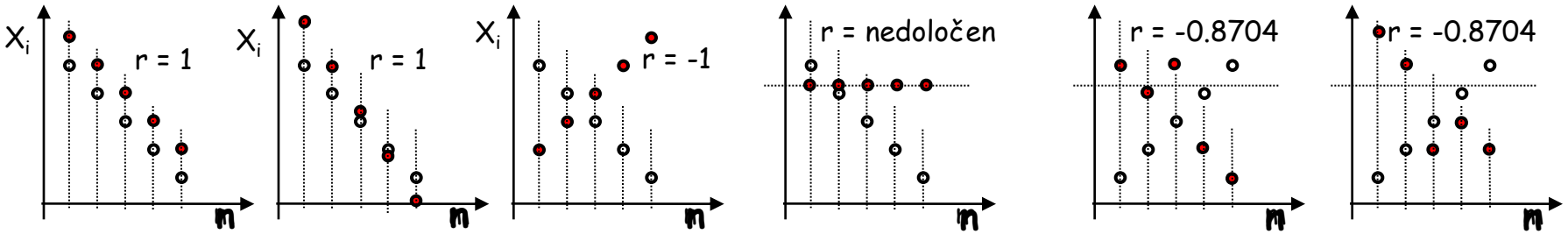
Izrazi za kovarianco in korelacijo med dvema nizoma vrednosti X_i in X_j

$$kovarianca\ v_{ij} = \frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{n-1} = \frac{\sum_{k=1}^{n}x_{ik}x_{jk} - \frac{1}{n}(\sum_{k=1}^{n}x_{ik})(\sum_{k=1}^{n}x_{jk})}{n-1}$$

$$korelacija\quad r_{ij} = \frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)^2}} = \frac{\sum_{k=1}^{n}x_{ik}x_{jk} - \frac{1}{n}(\sum_{k=1}^{n}x_{ik})(\sum_{k=1}^{n}x_{jk})}{\sqrt{[\sum_{k=1}^{n}x_{ik}^2 - \frac{1}{n}(\sum_{k=1}^{n}x_{ik})^2][\sum_{k=1}^{n}x_{jk}^2 - \frac{1}{n}(\sum_{k=1}^{n}x_{jk})^2]}}$$

Kovarianca v_{ij} in korelacija r_{ij} med dvema enako dolgima nizoma vrednosti: $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ik}, \ldots, x_{in})$ in $\boldsymbol{X}_j = (x_{j1}, x_{j2}, \ldots, x_{jk}, \ldots, x_{jn})$ se v statistiki računata dokaj pogosto. Obe vrednosti kažeta na medsebojno odvisnost obeh nizov števil. Veliko pogosteje kot kovarianca v_{ij} se uporablja korelacijski koeficient r_{ij}, ker je normaliziran in kot tak primerljiv med različnimi pari enako dolgih nizov.

Korelacijski koeficient r lahko zavzame vse realne vrednosti med -1 in 1. Če je $r = 0$, potem sta oba vektorja $\boldsymbol{X}_i$ in $\boldsymbol{X}_j$ med seboj neodvisna, če pa je $|r| = 1$, potem sta oba vektorja povezana z linearno funkcijo (enega lahko izračunamo iz drugega preko linearne enačbe): $\boldsymbol{X}_i = a\boldsymbol{X}_j + b$.



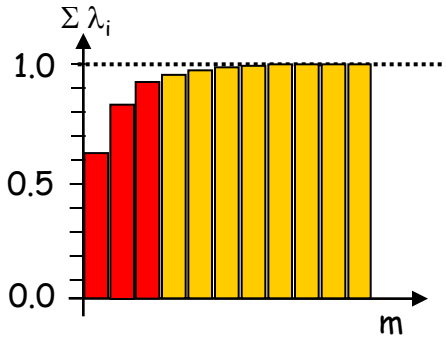
Izračun glavnih osi

Izračun glavnih osi je rotacija koordinatnega sistema. Rezultat PCA je transformacijska (rotacijska) matrika **L**, ki omogoča predstavitev objektov v novem koordinatnem sistemu. Matrika //**L**// definirana novi sistem tako, da je velikost deleža od celotne informacije (odstotek od celotne variance sistema vseh objektov) razporejen zaporedno po oseh: prva os prevzame največji, zadnja os pa najmanjši delež celotne variance (informacije).

$$(X^{new})^T = \|L\|\ X^{old} \qquad x_{ik}^{new} = \sum_{j=1}^{m} l_{jk} x_{ij}^{old}$$

Matriko //**L**// dobimo z diagonalizacijo ene od treh matrik, omenjenih na prejšnji strani: //**X**// (matrike mešanih produktov), //**V**// (matrike varianc) ali //**R**// (matrike korelacij). Z diagonalizacijo katerekoli od omenjenih treh $m \times m$ dimenzionalnih matrik, dobimo najprej m lastnih vrednosti $\lambda_1, \lambda_2, \ldots, \lambda_m$. Vsaki od lastnih vrednosti λ_i pripada lastni vektor $L_i = (l_{i1}, l_{i2}, \ldots, l_{ij}, \ldots, l_{im})$. Koeficiente l_{ij} imenujemo komponente lastnih vektorjev (loadings). Ko lastne vrednosti $\{\lambda_i\}$ uredimo po velikosti in v istem vrstnem redu razporedimo po vrsticah tudi lastne vektorje, dobimo transformacijsko matriko //**L**// za izračun glavnih osi.

$$(x_1^{new}, x_2^{new}, \ldots, x_m^{new}) = \begin{Vmatrix} l_{11} & l_{12} & \ldots & l_{1m} \\ l_{21} & l_{22} & \ldots & l_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ l_{m1} & l_{m2} & \ldots & l_{mm} \end{Vmatrix} \begin{vmatrix} x_1^{old} \\ x_2^{old} \\ \ldots \\ x_m^{old} \end{vmatrix}$$



Objekti X_i^{new}, ki so predstavljeni v koordinatnem sistemu prvih dveh glavnih osi (PC_1 in PC_2), so dejansko predstavljeni samo s prvima dvema novima komponentama $X_i^{new} = (x_1^{new}, x_2^{new}, 0, 0, \ldots, 0)$. Kolikor večji odstotek celotne informacije nosita prvi dve novi komponenti, toliko bolj verodostojna je preslikava objektov X_i v 2-dimenzionalni prostor PC_1/PC_2 .

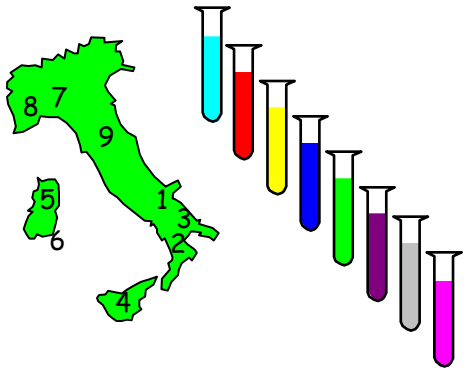
Primer: PCA analiza oljčnih olj

V danem primeru bomo s PCA obdelali 572 oljčnih olj, od katerih je vsako analizirano glede na vsebnost osmih maščobnih kislin. V spodnji tabeli so podani osnovni podatki: povprečja posameznih spremenljivk (koncentracije maščobnih kislin v 100 ppm), standardni odmik za vsako maščobno kislino in prvi dve vrstici izhodnih treh matrik ***//X//***, ***//V//*** in ***//R//***. Vse tri matrike imajo dimenzijo 572 x 8.

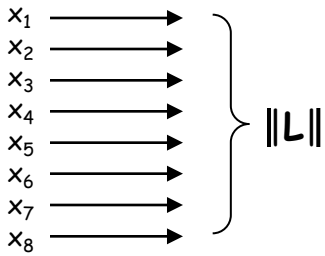
	Palmitic	Palmitoleic	Stearic	Oleic	Linoleic	Arachidic	Linolenic	Eicosenioc		
S. Apulija	1075	75	226	7823	672	36	60	29		X1
Kalabrija	1315	139	230	7299	832	42	60	32		X2
S. Apulija	-156,8	-51,1	-2,9	511,3	-308,5	4,0	1,8	12,7		V1=X1-XP
Kalabrija	83,3	12,9	1,1	-12,7	-148,5	10,0	1,8	15,7		V2=X2-XP
S. Apulija	-0,93	-0,97	-0,08	1,26	-1,27	0,32	0,08	0,90		R1=V1/s
Kalabrija	0,49	0,25	0,03	-0,03	-0,61	0,79	0,08	1,12		R2=V2/s
Povprečje	1231,8	126,1	228,9	7311,7	980,5	32,0	58,2	16,3		XP
std. odmik	168,6	52,5	36,7	405,8	242,8	12,7	21,8	14,1		s

A_1 A_2 A_3

Analize 572 oljčnih olj iz 9 italjanskih pokrajin.

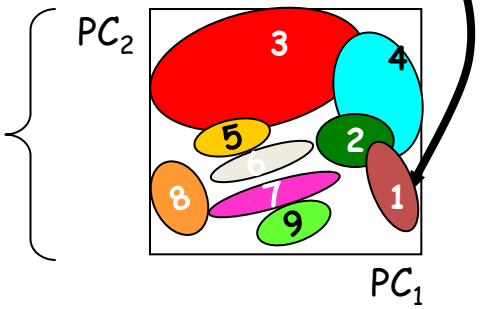


Vsako olje je opisano z osmimi podatki, ki jih PCA preslika v 2d ravnino PC_1/PC_2



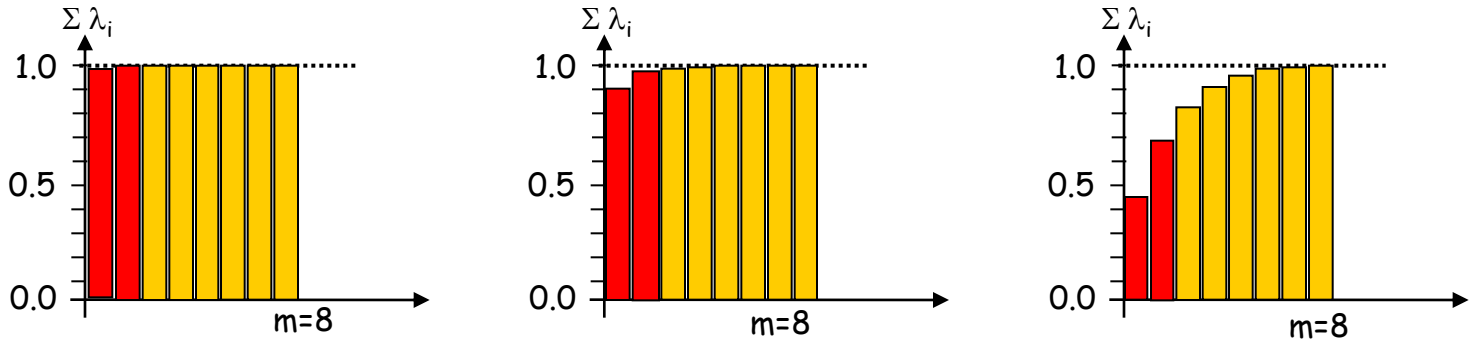
Analiza neznanega olja

2d preslikava olj



Po diagonalizaciji vseh treh matrik dobimo tri nize lastnih vrednosti in ustrezne lastne vektorje. Informacijske vsebnosti posameznih glavnih osi λ_i, glede na izhodiščno matriko (**X**, **V** ali **R**) so naslednje:

	$\mathbf{X}^T\mathbf{X}$	$\mathbf{V}^T\mathbf{V}$	$\mathbf{R}^T\mathbf{R}$
λ_1	99.76	89.71	46.49
λ_2	0.20	8.87	22.11
λ_3	0.04	0.80	12.62
λ_4	0.00	0.29	9.96
λ_5	0.00	0.24	4.13
λ_6	0.00	0.06	3.17
λ_7	0.00	0.02	1.49
λ_8	0.00	0.02	0.03

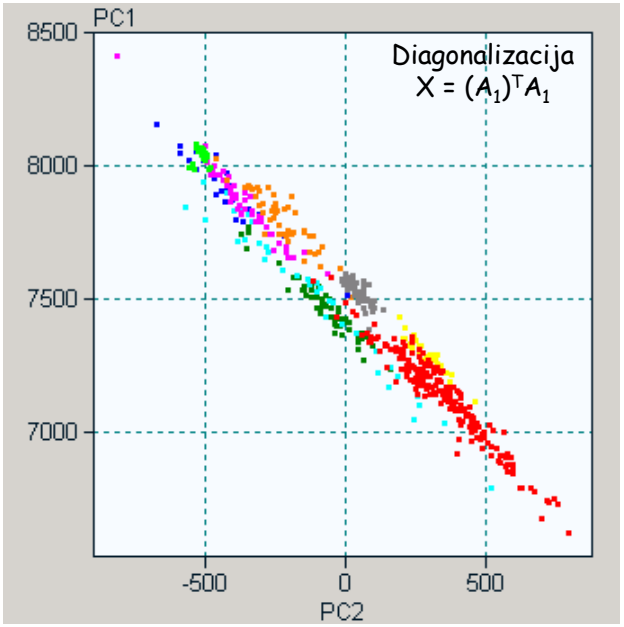


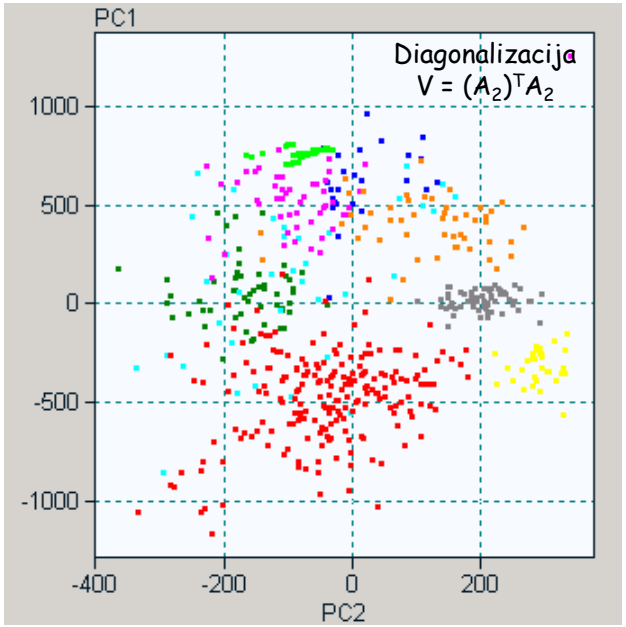
Razločno se vidi, da je velika večina informacij zbrana v nekaj prvih glavnih oseh. V prvem primeru (podatkovna matrika $\mathbf{X}^T\mathbf{X}$) je 99.76 % informacije zbrane okrog prve osi. V tretjem primeru (produkt korelacijskih matrik $\mathbf{R}^T\mathbf{R}$) pa je informacija porazdeljena bolj enakomerno. Katero od osnov izbrati za določen primer, odloča uspešnost projekcije objektov v PC_i/PC_j prostoru.

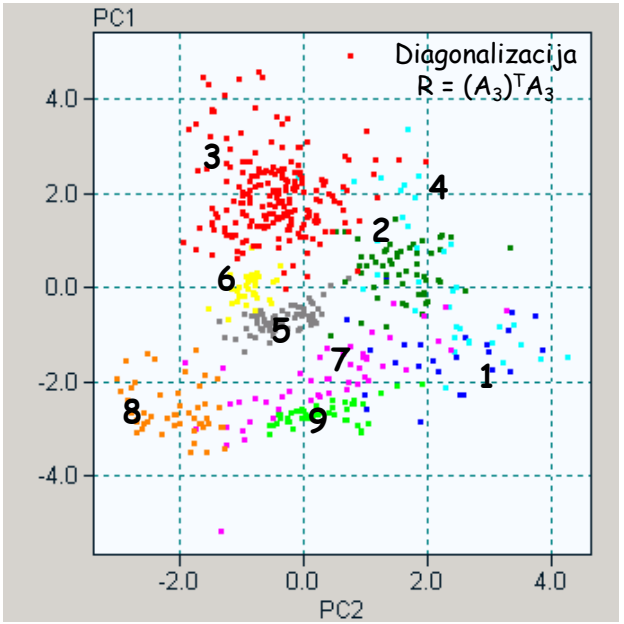
Na naslednji strani so projekcije PC_1/PC_2 prikazane za vse tri primere. Čeprav je v prvem in drugem primeru v obeh prvih oseh združenih blizu 100 % informacije, je v PC prostoru ločitev slabša kot v tretjem primeru. Vzrok zato so neprimerljive medsebojne absolutne velikosti posameznih koncentracij. Gre za navidezni vpliv absolutno večjih spremenljivk x_4, x_1, x_5 in x_2.

Projekcija analiz 8 maščobnih kislin v 572 olivnih oljih v prostor prvih dveh glavnih osi PC_1/PC_2

Vsaka točka predstavlja eno oljčno olje. Grupiranje je najboljše na skrajni desni. Številke utrezajo geografskim področjem.

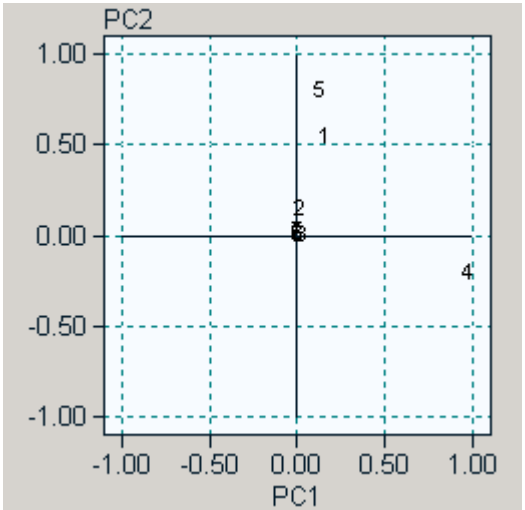


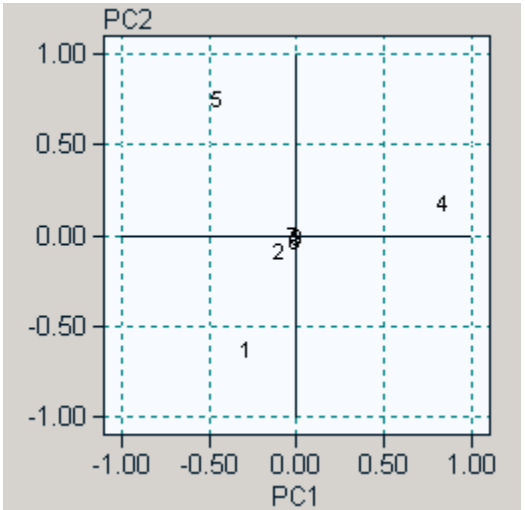


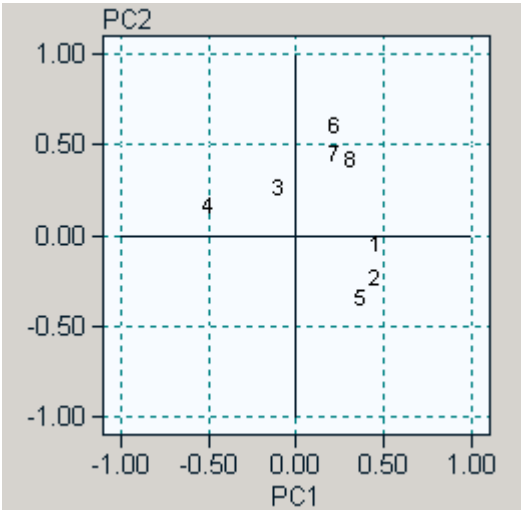


Vpliv prvotnih spremenljivk (koncentracij maščobnih kislin) na velikost PC_1 in PC_2

Številke označujejo koncentracije posameznih maščobnih kislin (1= palmitinska kislina …). Čim bliže je točka sredini koordinatnega sistema manjši vpliv ima v ustrzeni glavni osi.



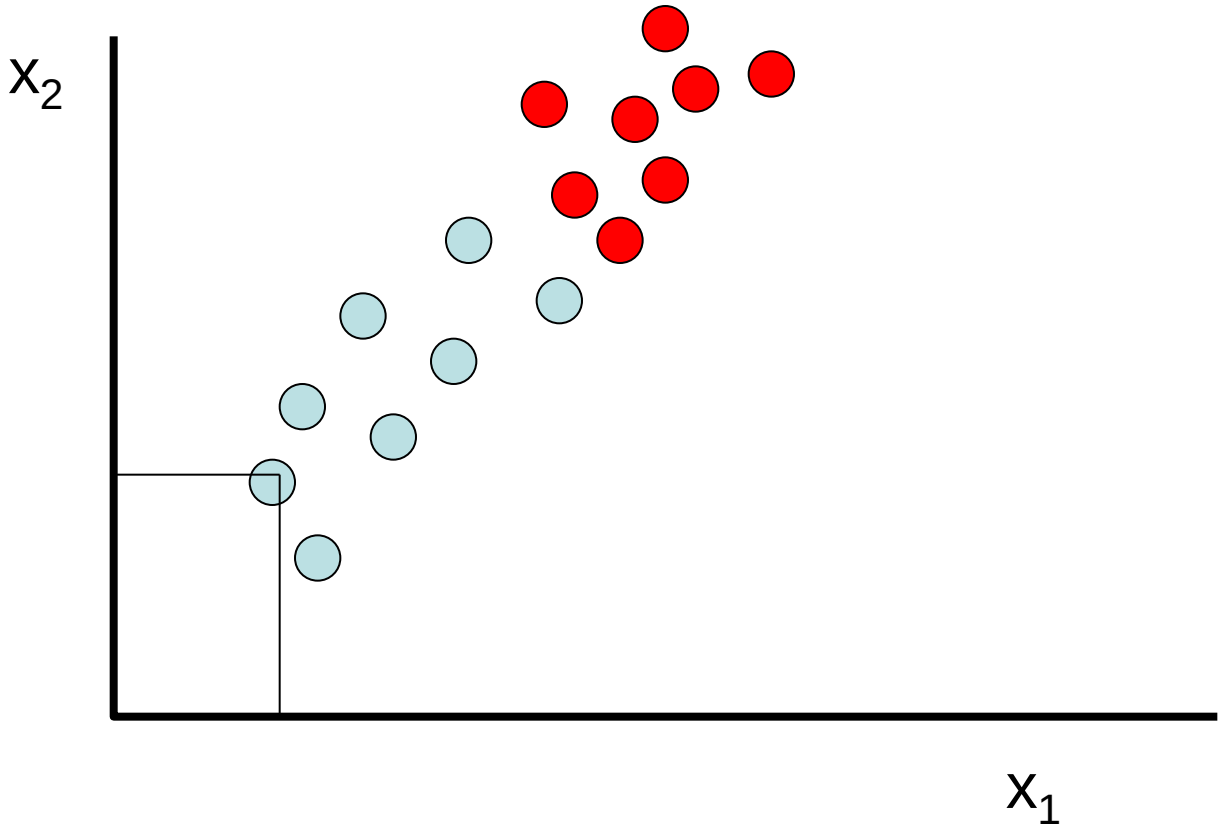




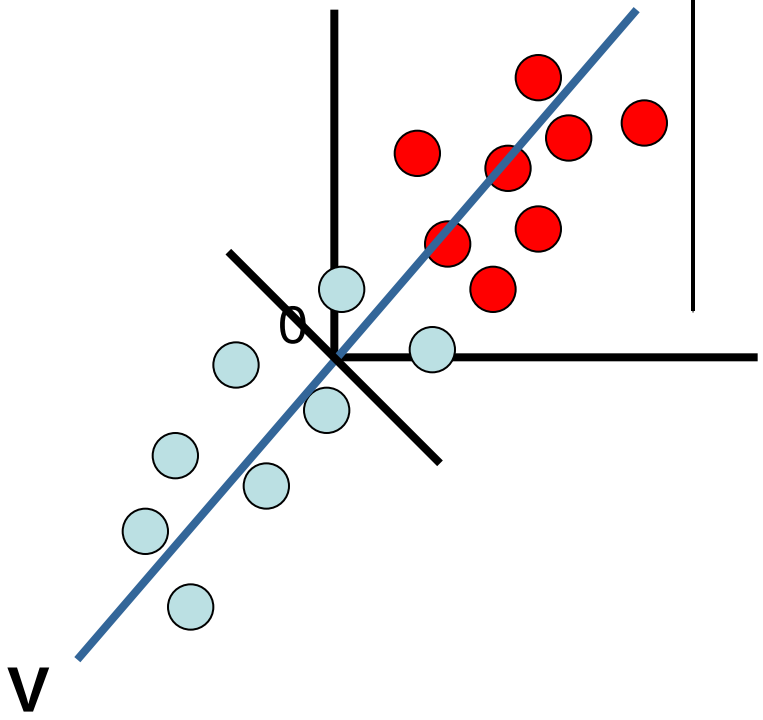
Basic Chemometrics in NIR Spectrometry
PCA – Principal Component Analysis

Marjana Novič
National Institute of Chemistry
Hajdrihova 19, Ljubljana, Slovenia

Variables > latent variables > model in latent space > classification in latent space

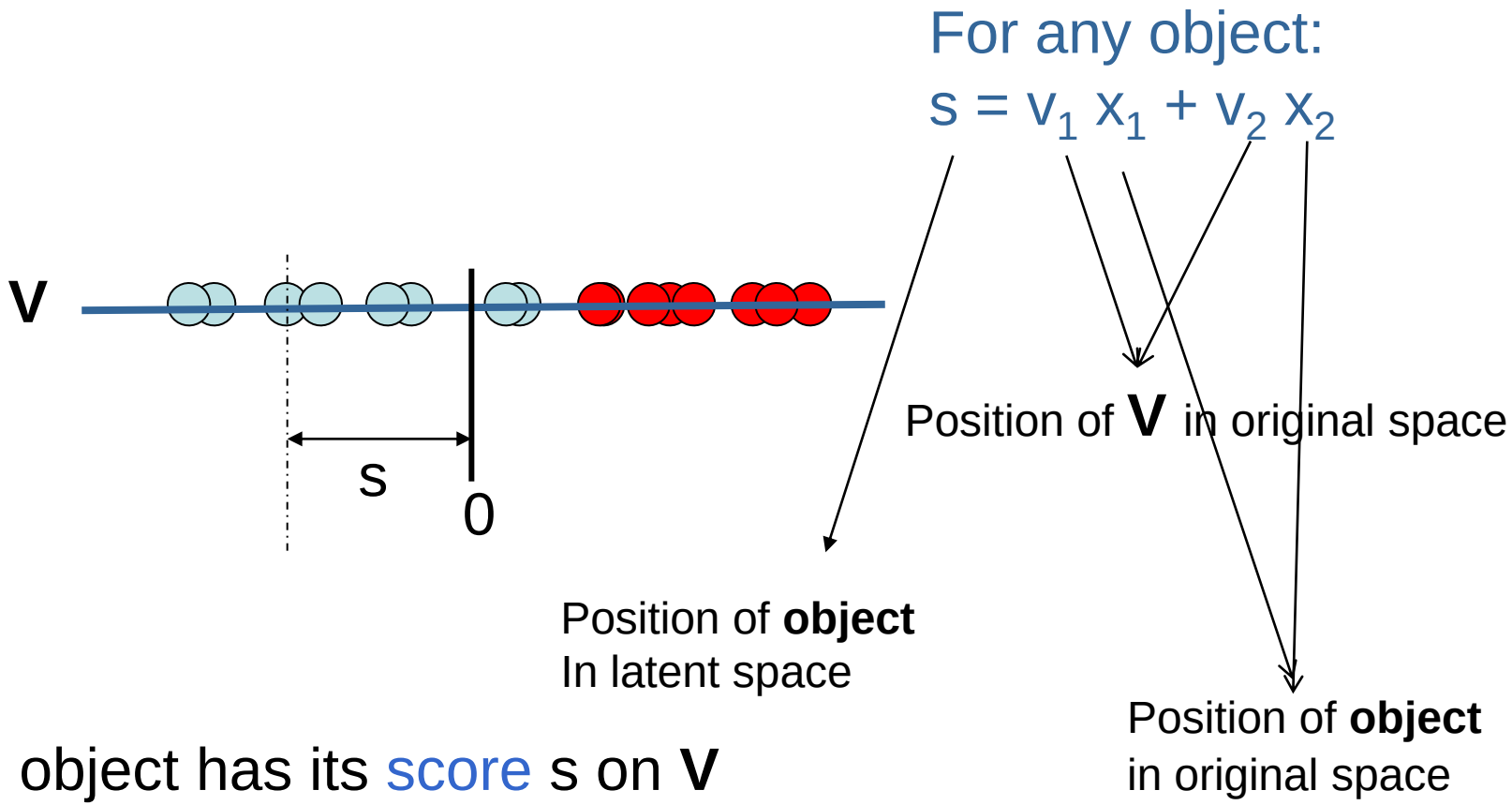


Variables > latent variables > model in latent space > classification in latent space



Sample	X1	X2	Class
N1	1.0	0.0	B
N2	0.0	0.5	B
N3	1.8	0.5	R
N4	1.2	0.8	R
N5	2.1	1.0	R
N6	1.9	2.0	R
N7	0.8	2.1	R
.	.	.	.
.	.	.	.

Variables > latent variables > model in latent space > classification in latent space



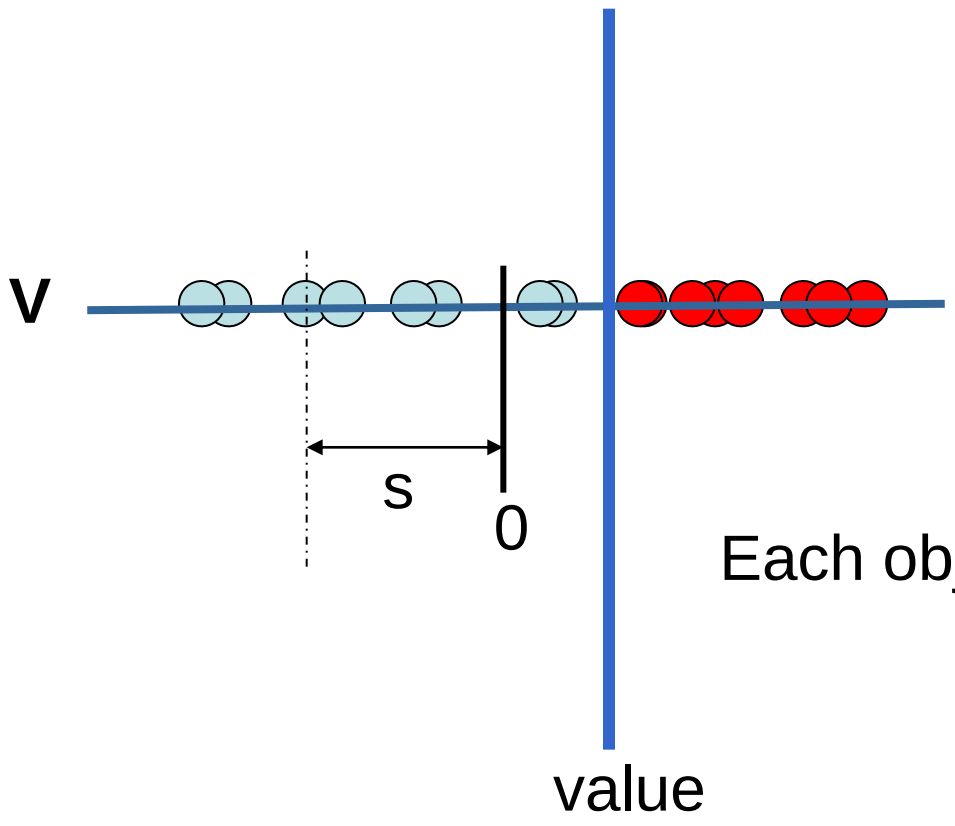
Each object has its score s on **V**

v_1 and v_2 are the loadings of the variables on **V**

V is called a principal component

For any object:

$s = v_1 x_1 + v_2 x_2$

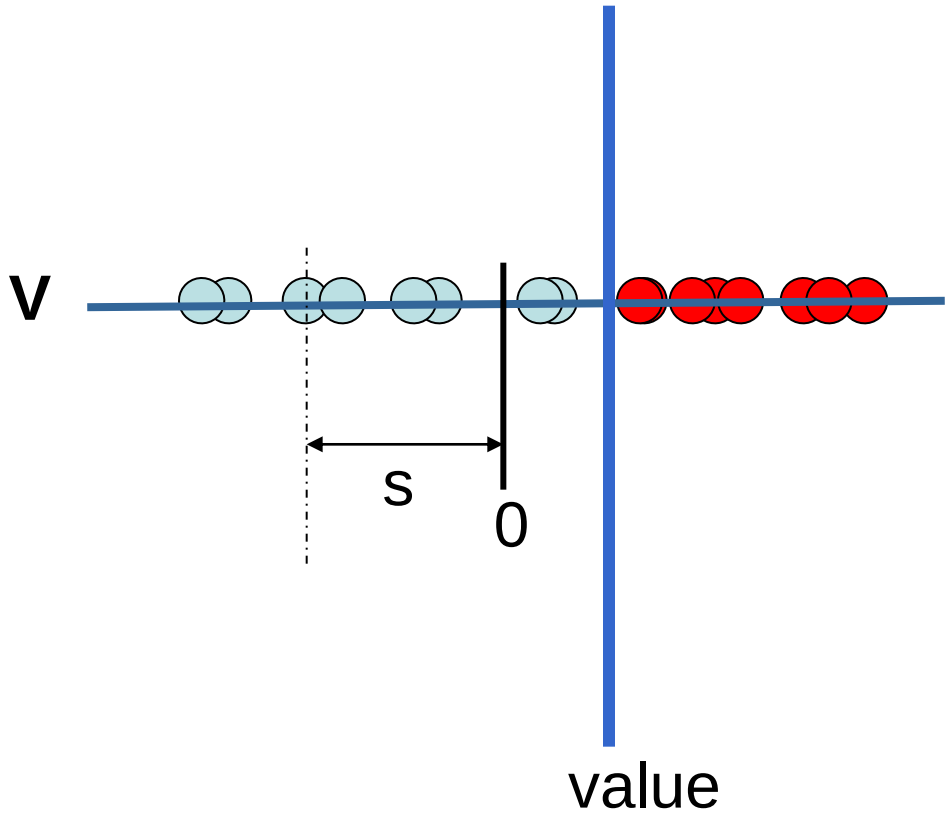


Each object has its score s

Variables > latent variables > model in latent space > classification in latent space

For any object:

$s = v_1 x_1 + v_2 x_2$

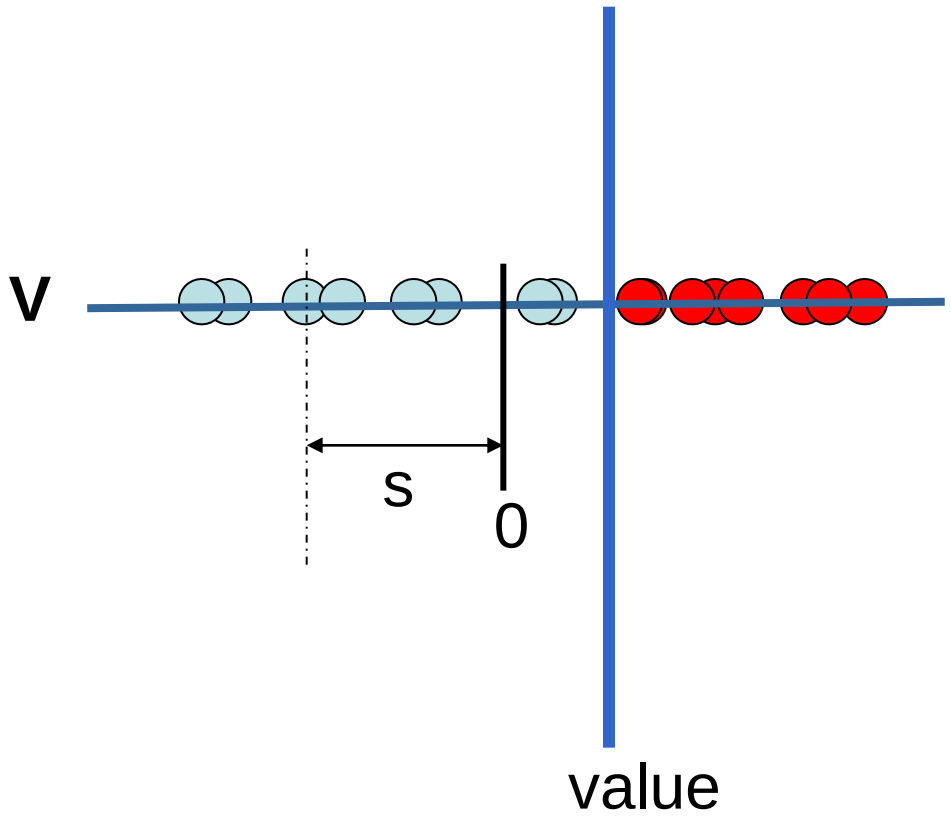


IF s < value THEN ○ Otherwise ●

Variables > latent variables > model in latent space > classification in latent space > specifications in variable space

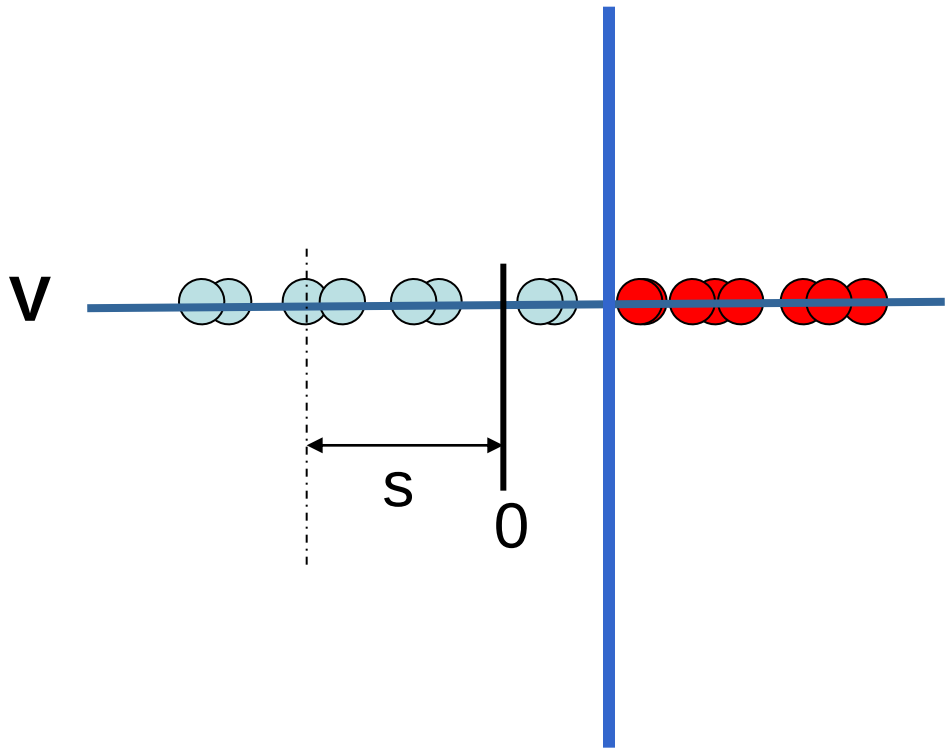
For any object:

$s = v_1 x_1 + v_2 x_2$



IF $v_1 x_1 + v_2 x_2$ < value THEN ○ Otherwise ●

Variables > latent variables > model in latent space > classification in latent space > specifications in variable space



For any object:

$s = v_1 x_1 + v_2 x_2$

IF $v_1 x_1 + \ldots + v_p x_p <$ value THEN ○ Otherwise ●

This concept is extendable to more manifest variables

Latent space < what it does < why we like it < why we hate it

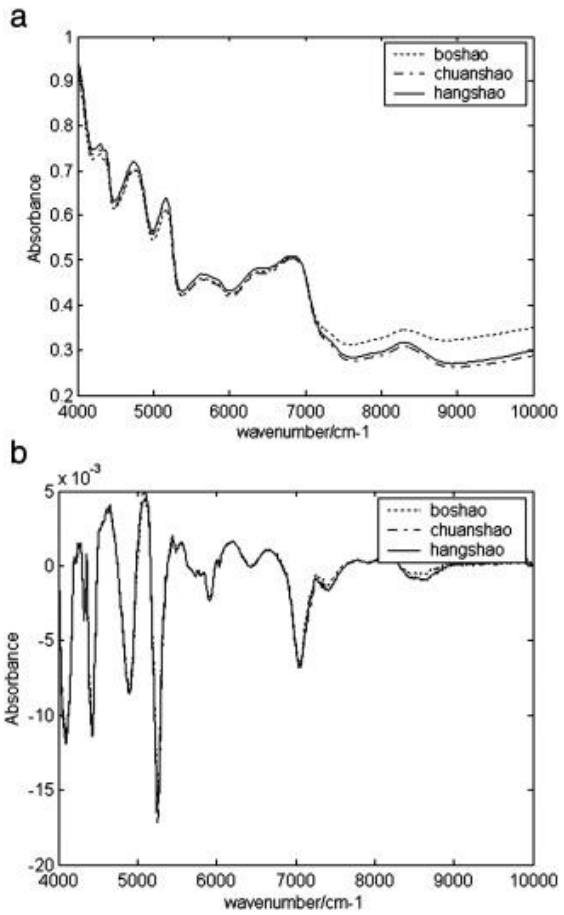
Latent space:

- transforms may manifest variables into a few latent variables
- each object is fully described by these few latent variables

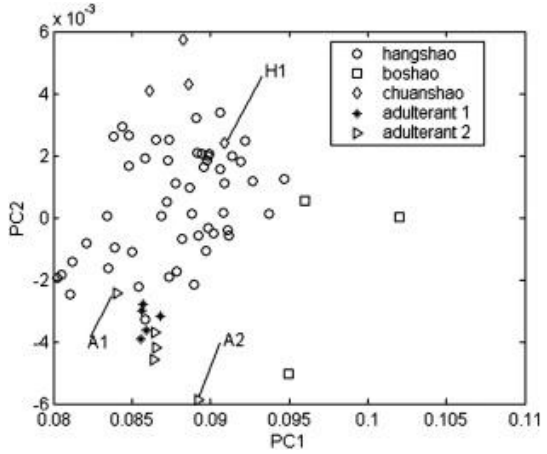
- latent variables are uncorrelated
- less objects required when modeling
- models are simpler

- latent variables are ‘abstract’
- number of latent variables
- back transform of latent models to manifest variables

X. Luo at al., *Microchemical Journal*, **90**, 2008, 8-12.
Rapid determination of Paeoniae Radix using near infrared spectroscopy



is the root of traditional Chinese Herb named Paeonia lactiflora Pallas, which is commonly used to treat liver diseases in China for centuries.

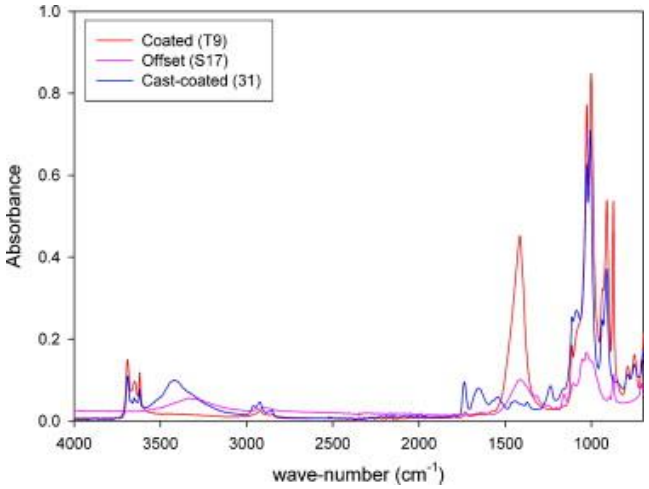


Two-dimension plot of PCA.

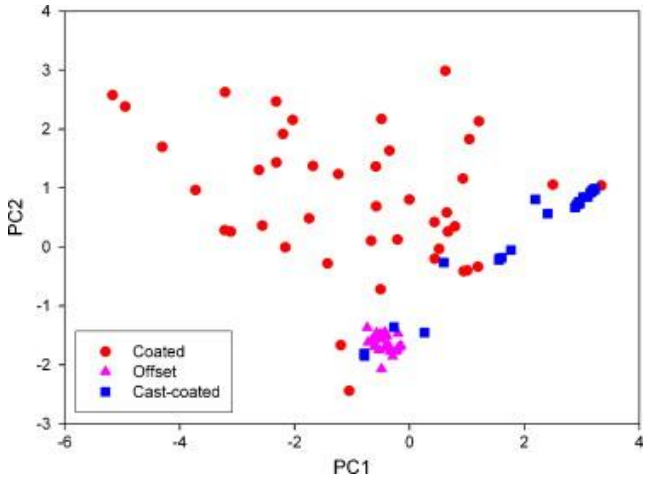
Mean NIR spectra for “hangshao”, “boshao” and “chuanshao” obtained from, (a)raw data, (b)with first derivative pretreatment.

T. Canals, Talanta, 77, 2008, 751-757

Characterization of paper finishes by use of infrared spectroscopy in combination with canonical variate analysis, T. Canals, J.R. Riba, R. Cantero, J. Cansino, D. Domingo and H. Iturriaga

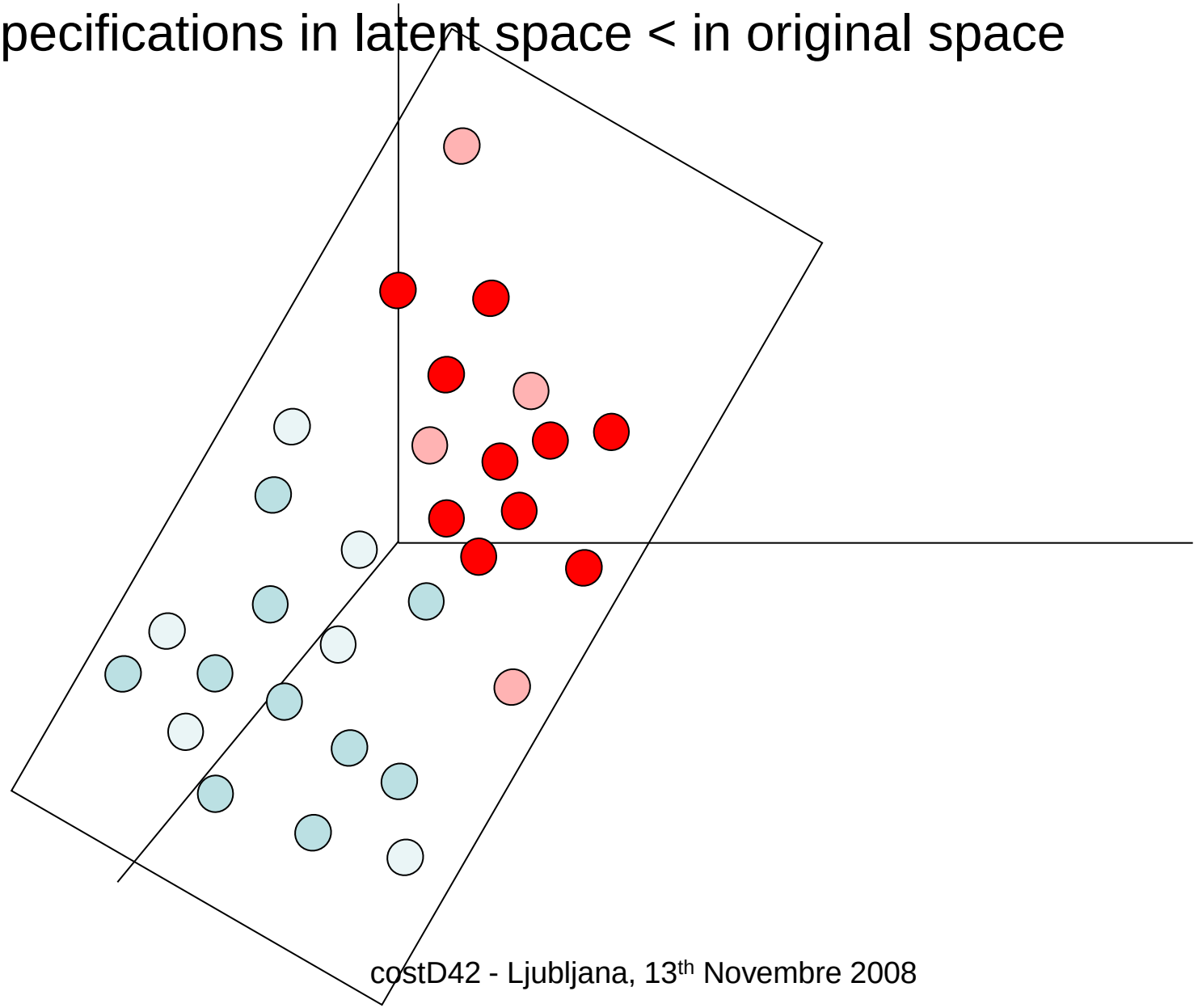


FTIR absorbance spectra for three selected paper samples representing each finish type.



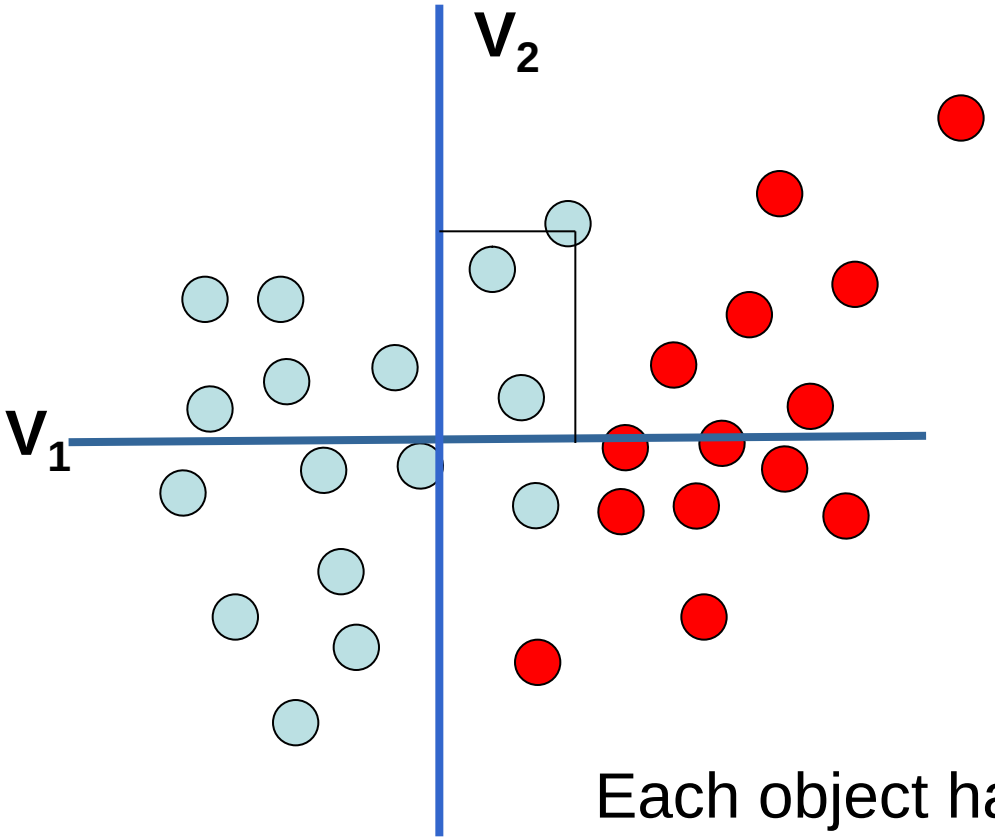
PCA of the FTIR absorbance spectra for the 92 samples as done on centred data.

Many manifest variables < many latent variables < model < specifications in latent space < in original space



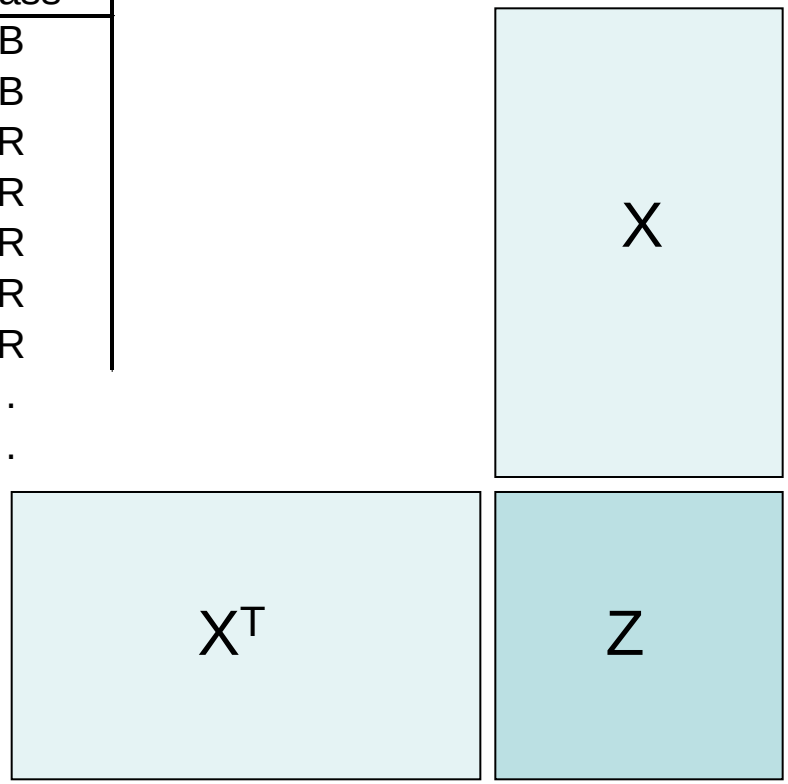
Many manifest variables < many latent variables < model < specifications in latent space < in original space

$x_1, x_2, x_3, \ldots x_p \longrightarrow s_1, s_2$



Each object has a score on V_1 and on V_2

Sample	X1	X2	Class
N1	1.0	0.0	B
N2	0.0	0.5	B
N3	1.8	0.5	R
N4	1.2	0.8	R
N5	2.1	1.0	R
N6	1.9	2.0	R
N7	0.8	2.1	R
.	.	.	.
.	.	.	.



Scattered matrix $\mathbf{Z} = \mathbf{X}^T\mathbf{X}$.

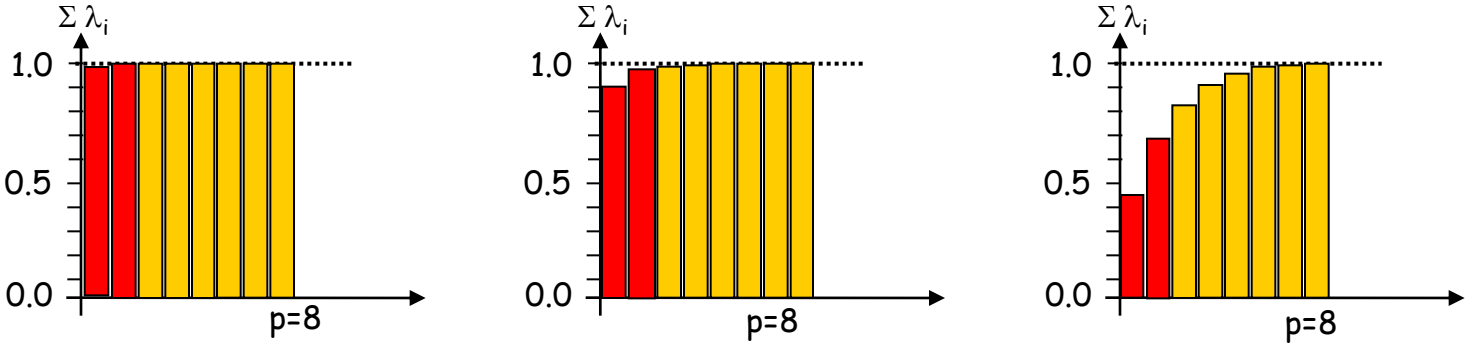
Matrix **Z** of dimension *p* by *p*, which can have at most *p* eigenvectors.

These are the diagonal elements in the so called diagonalized matrix:

Singular value decomposition

$\mathbf{ZV} = \mathbf{V}\mathrm{diag}(\lambda 1, \ldots \lambda p)$
or
$\mathbf{Zv}_i = \lambda_1 v_i$
The eigen vectors v_i are rows of the loading matrix **V**.
The eigenvectors are orthogonal to each other, and the product $\mathbf{V}^T\mathbf{V}$ is the identity matrix **I** (**V** is an orthonormal matrix). The matrix **Z** can be expressed by its eigenvectors
$\mathbf{Z} = \mathbf{V}\mathrm{diag}(\lambda_1, \ldots \lambda_p)\mathbf{V}^T$
Note that the inverse matrix of **Z** can be written simply as
$\mathbf{Z}^{-1} = \mathbf{V}\mathrm{diag}(1/\lambda 1, \ldots 1/\lambda p)\mathbf{V}^T$

Influence of normalization (no, mean-centring, auto-scaling-standardize)

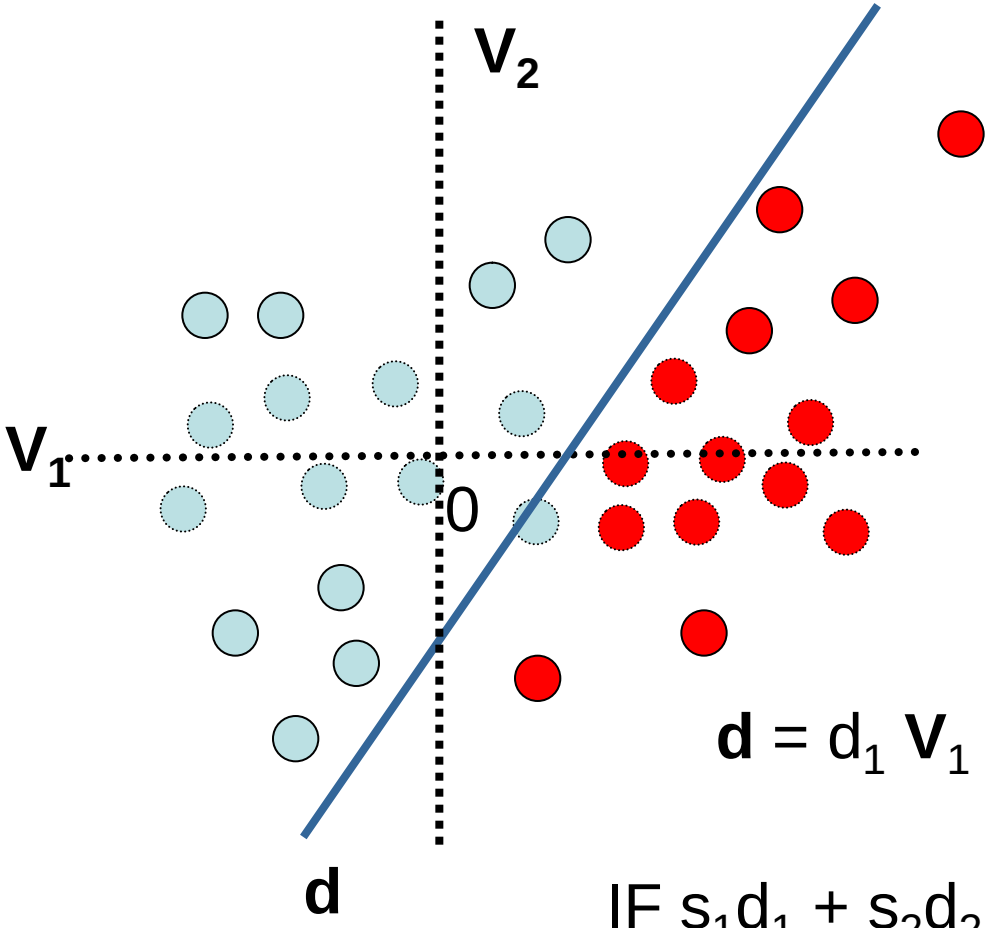


	X^TX	V^TV	R^TR
λ_1	99.76	89.71	46.49
λ_2	0.20	8.87	22.11
λ_3	0.04	0.80	12.62
λ_4	0.00	0.29	9.96
λ_5	0.00	0.24	4.13
λ_6	0.00	0.06	3.17
λ_7	0.00	0.02	1.49
λ_8	0.00	0.02	0.03

Total variance (sum of eigenvalues)

Many manifest variables < many latent variables < model < specifications in latent space < in original space

$x_1, x_2, x_3, \ldots x_p \longrightarrow s_1, s_2$



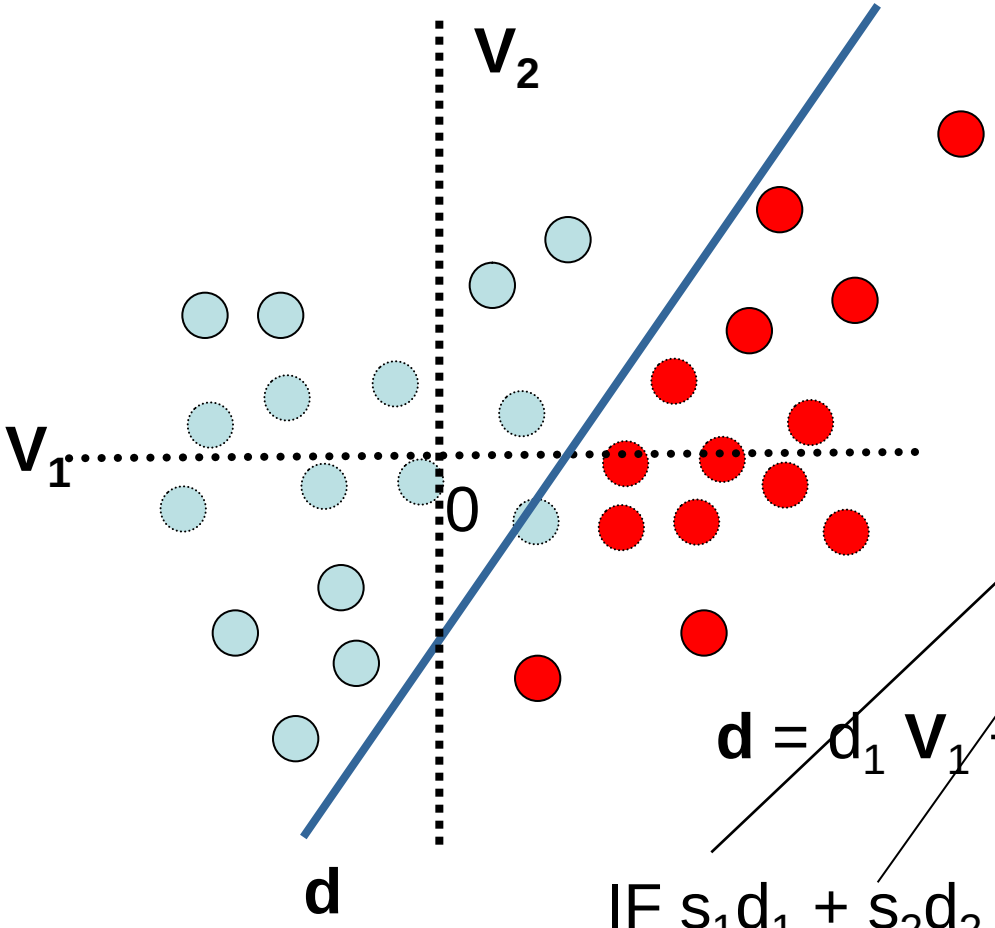
$\mathbf{d} = d_1 \mathbf{V}_1 + d_2 \mathbf{V}_2$

IF $s_1 d_1 + s_2 d_2 < 0$ THEN ●

otherwise ●

Many manifest variables < many latent variables < model < specifications in latent space < in original space

$x_1, x_2, x_3, \ldots x_p \longrightarrow s_1, s_2$



Transform scores to original variables (weighted sum)

$\mathbf{d} = d_1 \mathbf{V}_1 + d_2 \mathbf{V}_2$

IF $s_1 d_1 + s_2 d_2 < 0$ THEN ○

otherwise ●

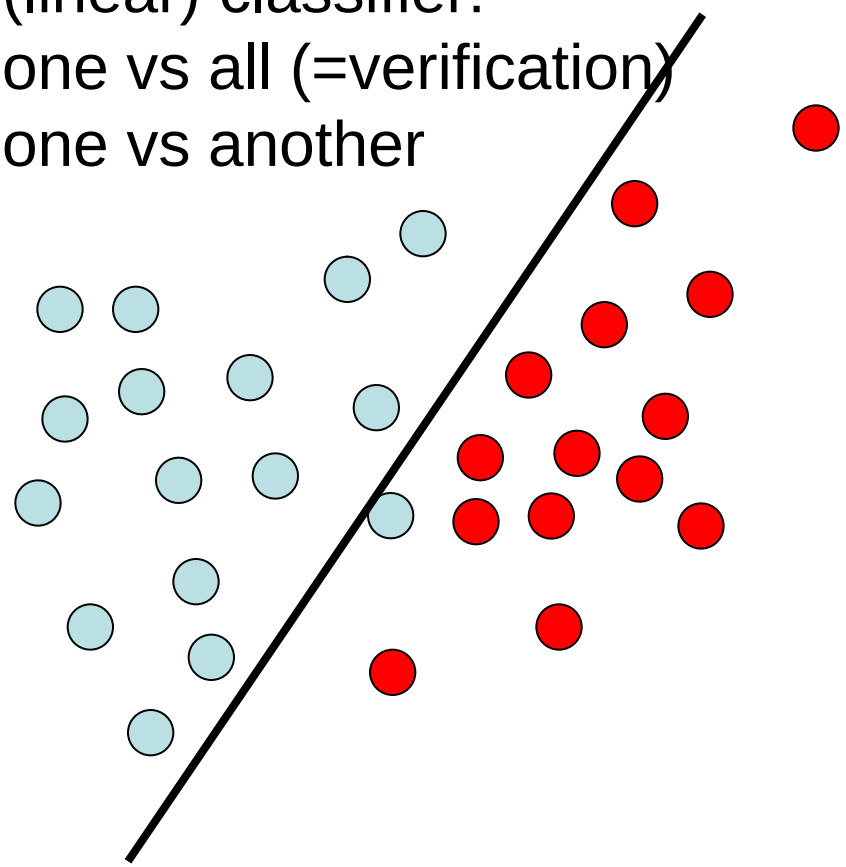
Models < discriminating models < local models < decision trees < Neural nets

D-PLS (regression model)

binary (linear) classifier:

one vs all (=verification)

one vs another



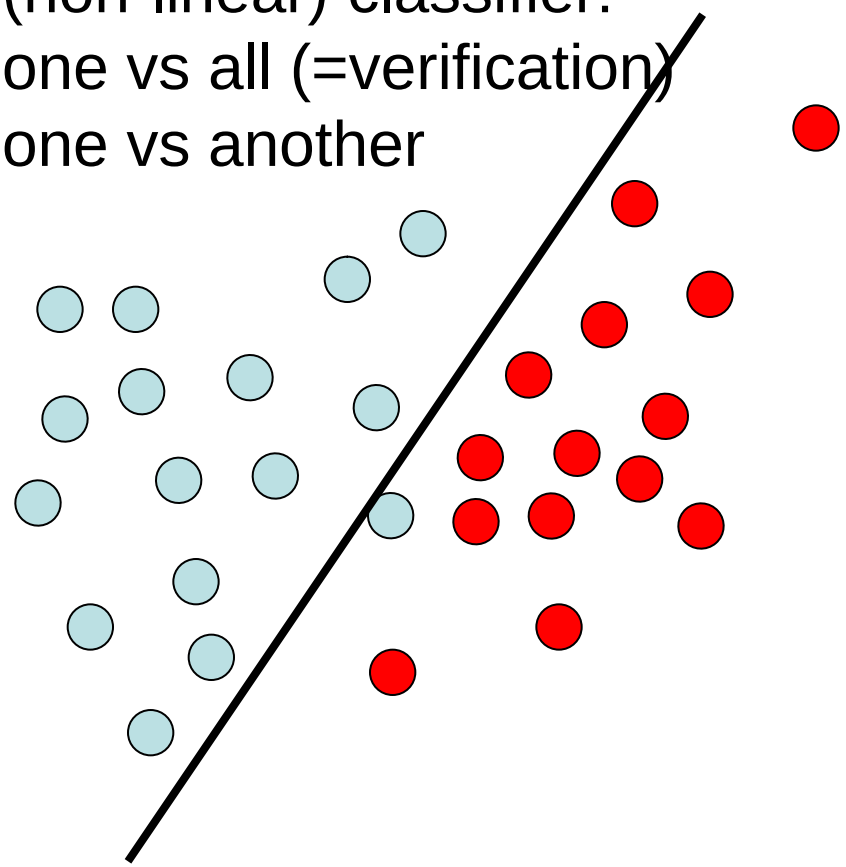
Models < discriminating models < local models < decision trees < Neural nets

Support Vector Machine (regression model)

binary (non-linear) classifier:

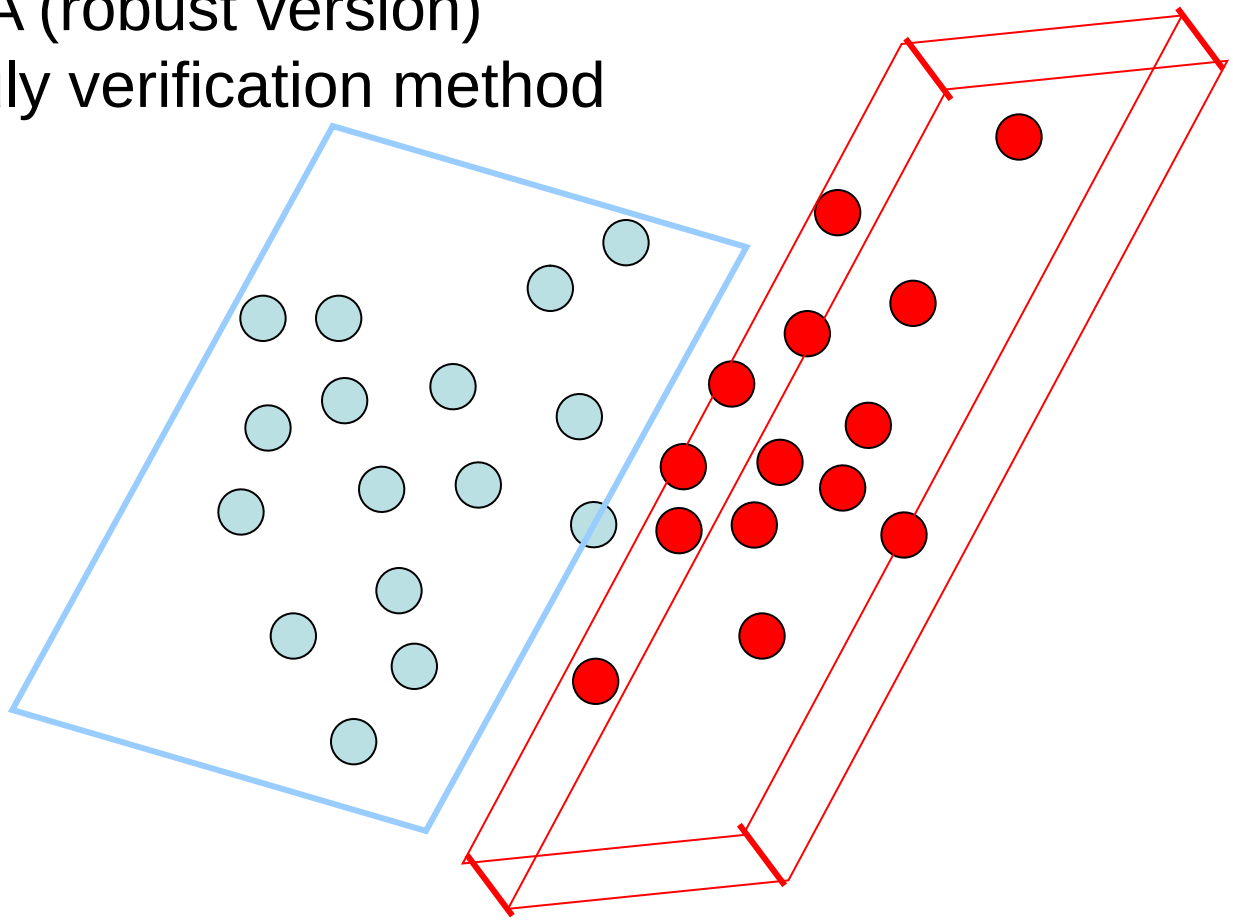
one vs all (=verification)

one vs another



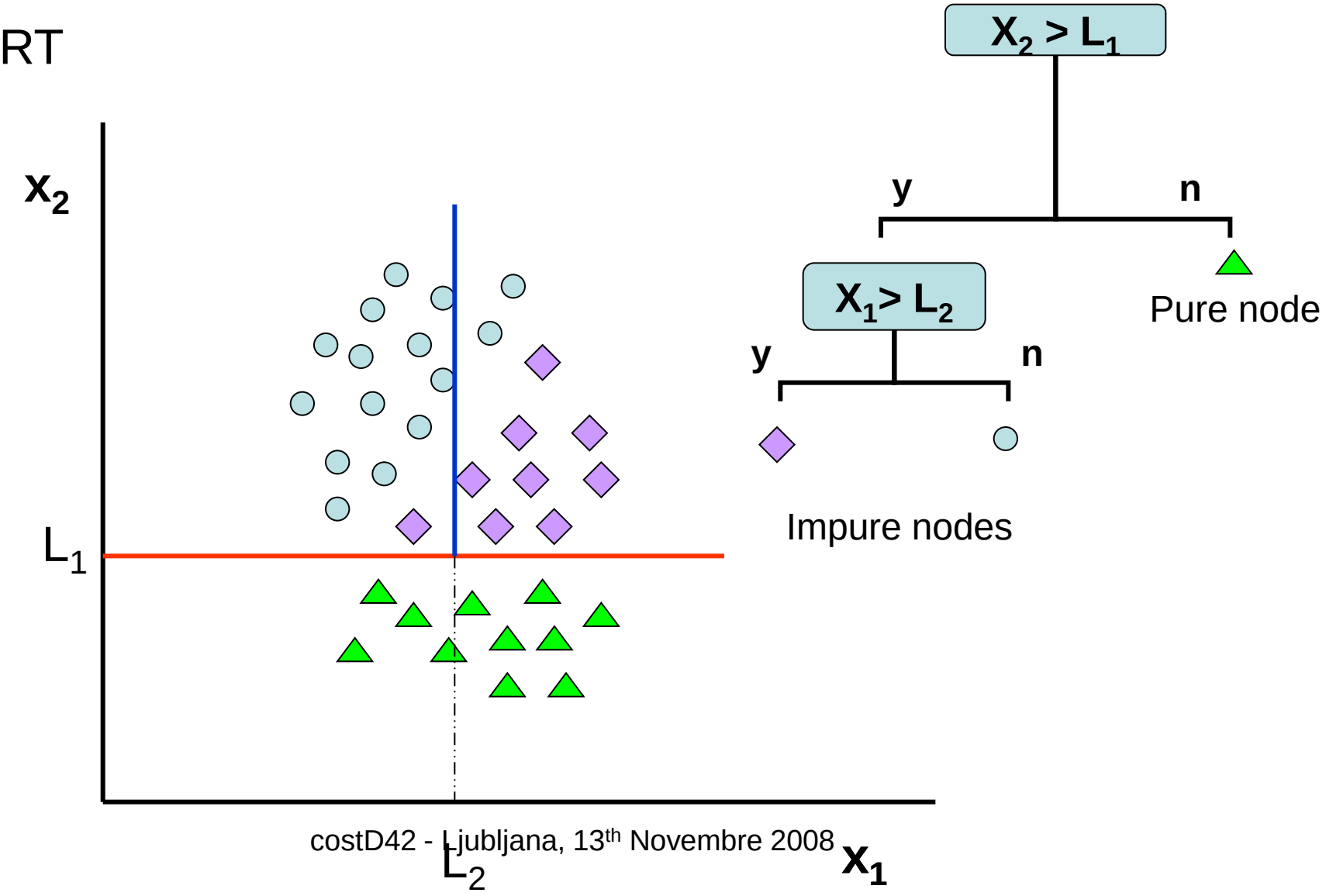
Models < discriminating models < local models < decision trees < Neural nets

SIMCA (robust version)
= a truly verification method



Models < discriminating models < local models < decision trees < Neural nets

CART



Models < discriminating models < local models < decision trees < Neural nets

Original variable space
(or latent space)

Kohonen Map