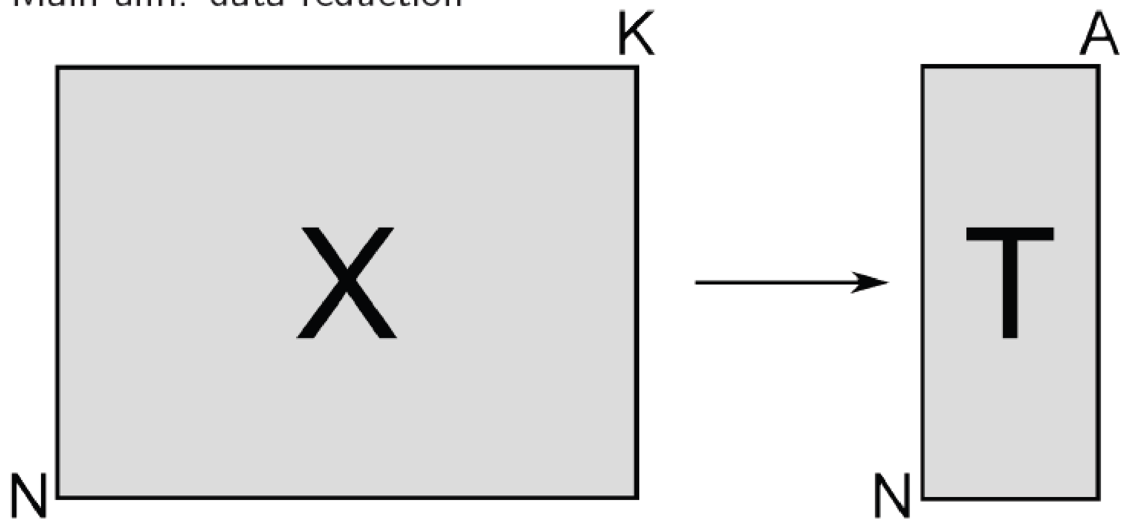
2. Principal Component Analysis

- Visualizing multivariate data
- Geometric interpretation of PCA
- Mathematical interpretation
- Example(s)

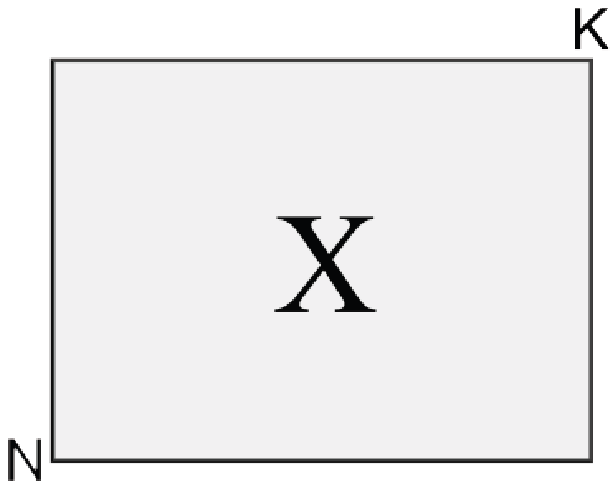
Principal Component Analysis

Main aim: data reduction



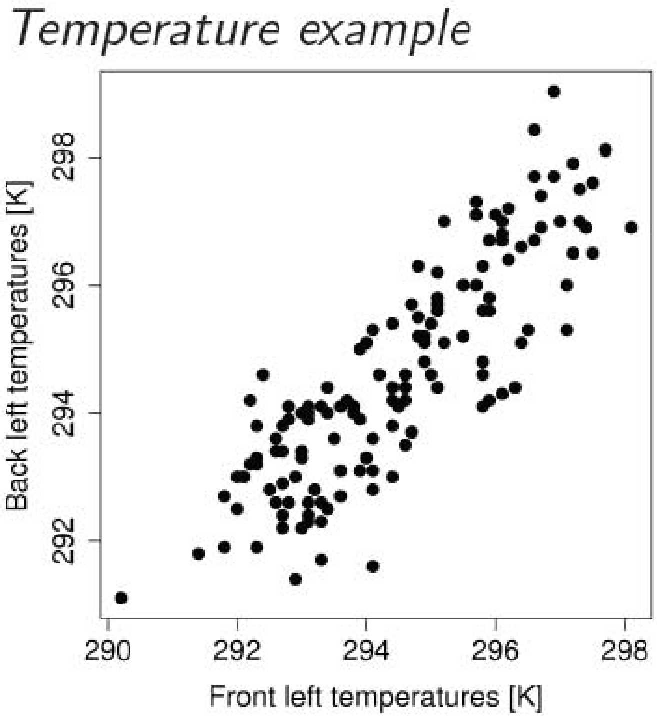
Visualizing Multivariate Data

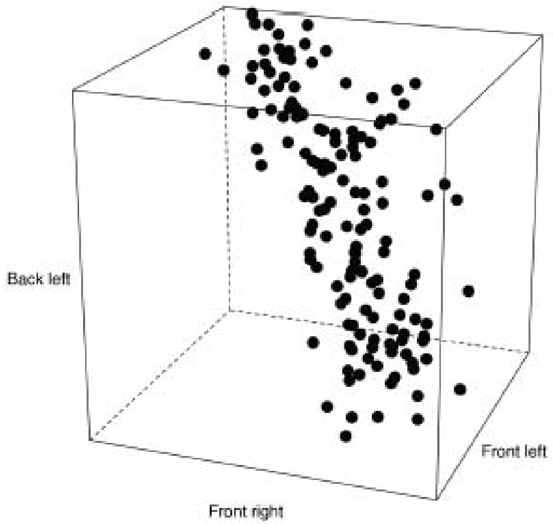
- PCA considers a single matrix: $\mathbf{X}$



- N observations
- K variables
- Which variables go in $\mathbf{X}$?

Visualizing Multivariate Data



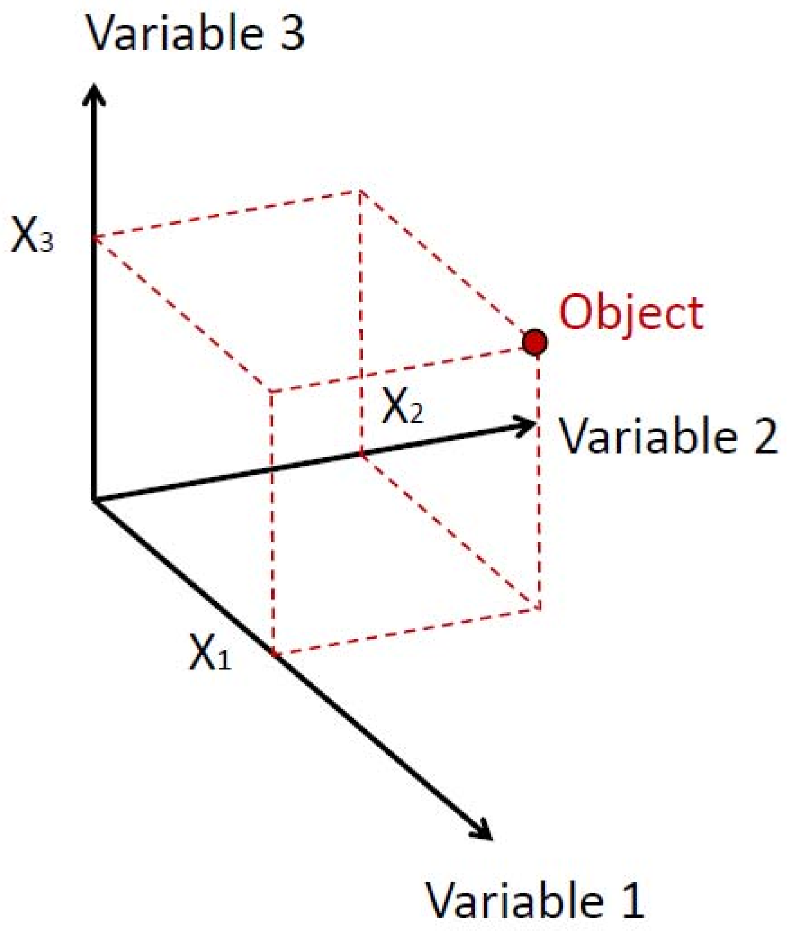


Geometric Interpretation

- Each variable defines an axis. A coordinate system can be made using all variables, called the variable space.

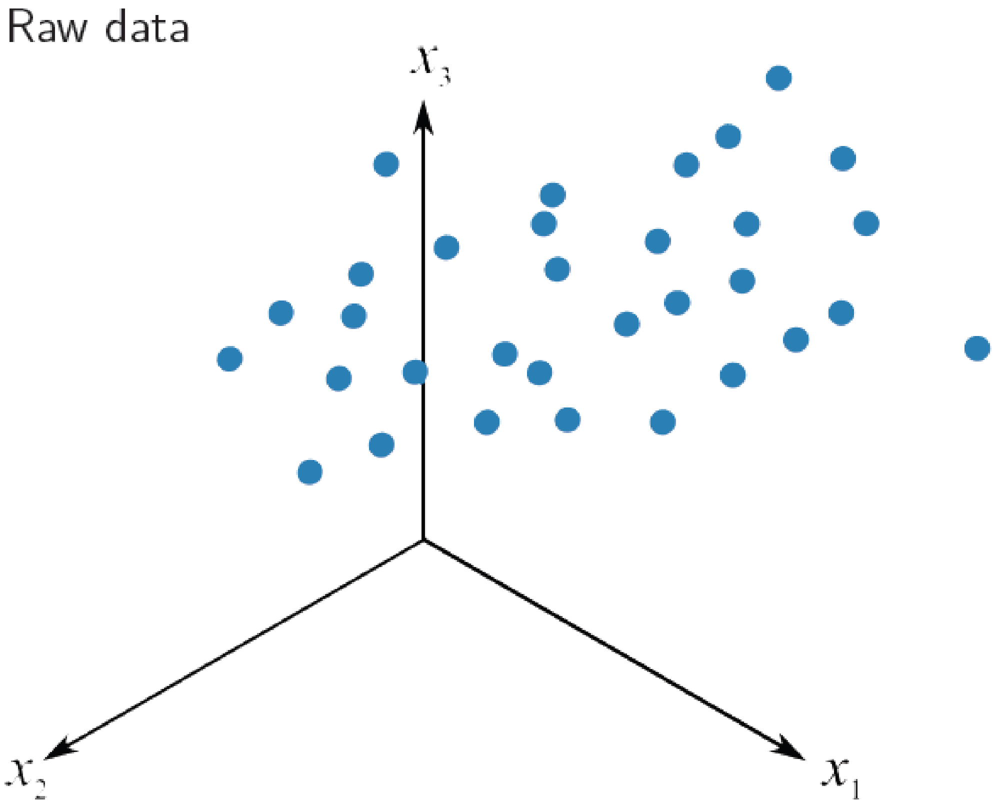
- Each object is a row in the data table (matrix) and is visualised as a point in the variable space.

	Variable 1	Variable 2	Variable 3
Object 1			
Object 2			
Object 3			
Object 4			



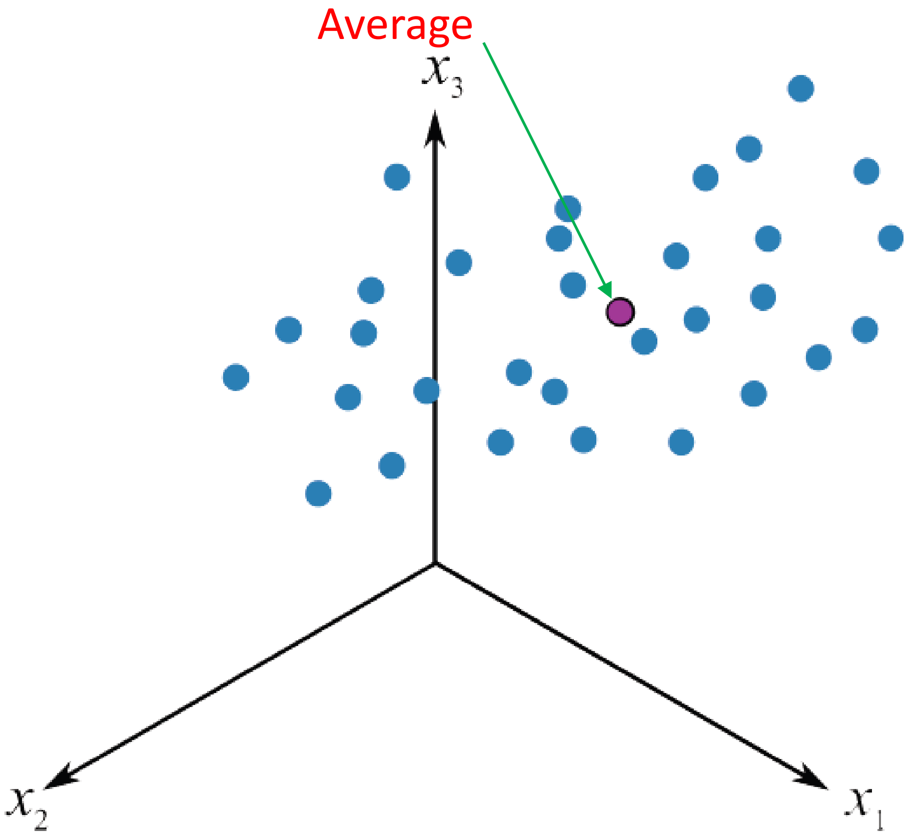
Geometric Interpretation

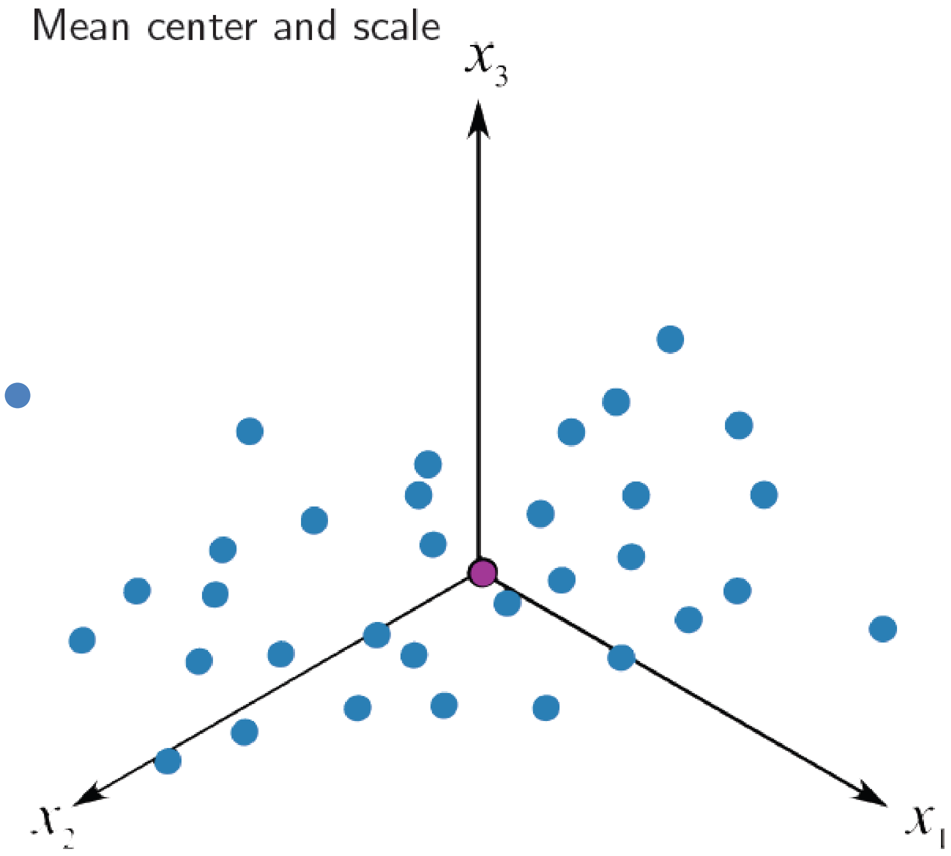
- Each sample / object is represented as a point in the variable space.
- The whole data table constitutes a swarm of points in the variable space.
- We would like to find out more about the structure of this swarm.



Geometric Interpretation

- In many cases, the swarm of points has a specific shape in some direction.

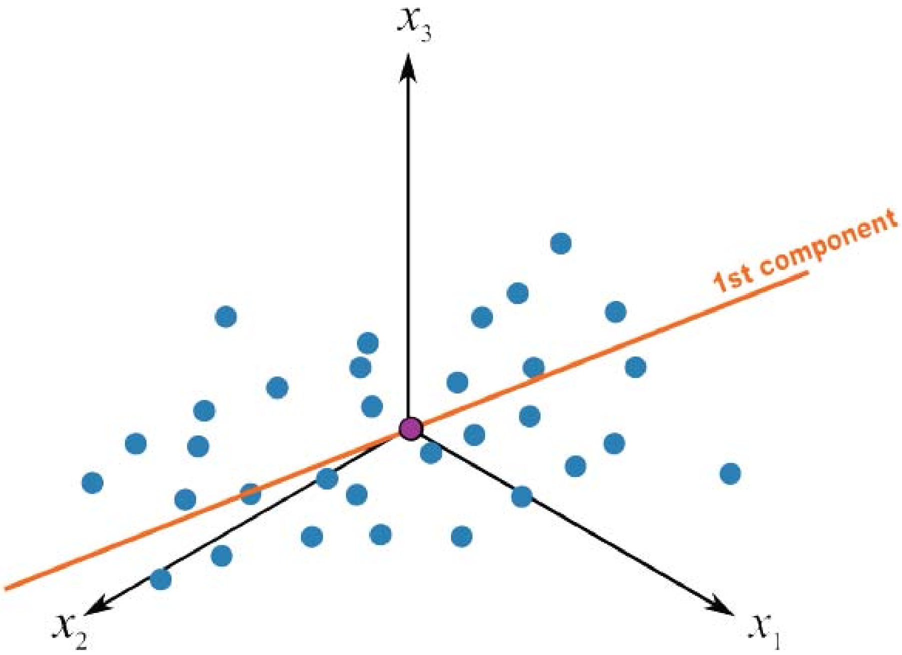




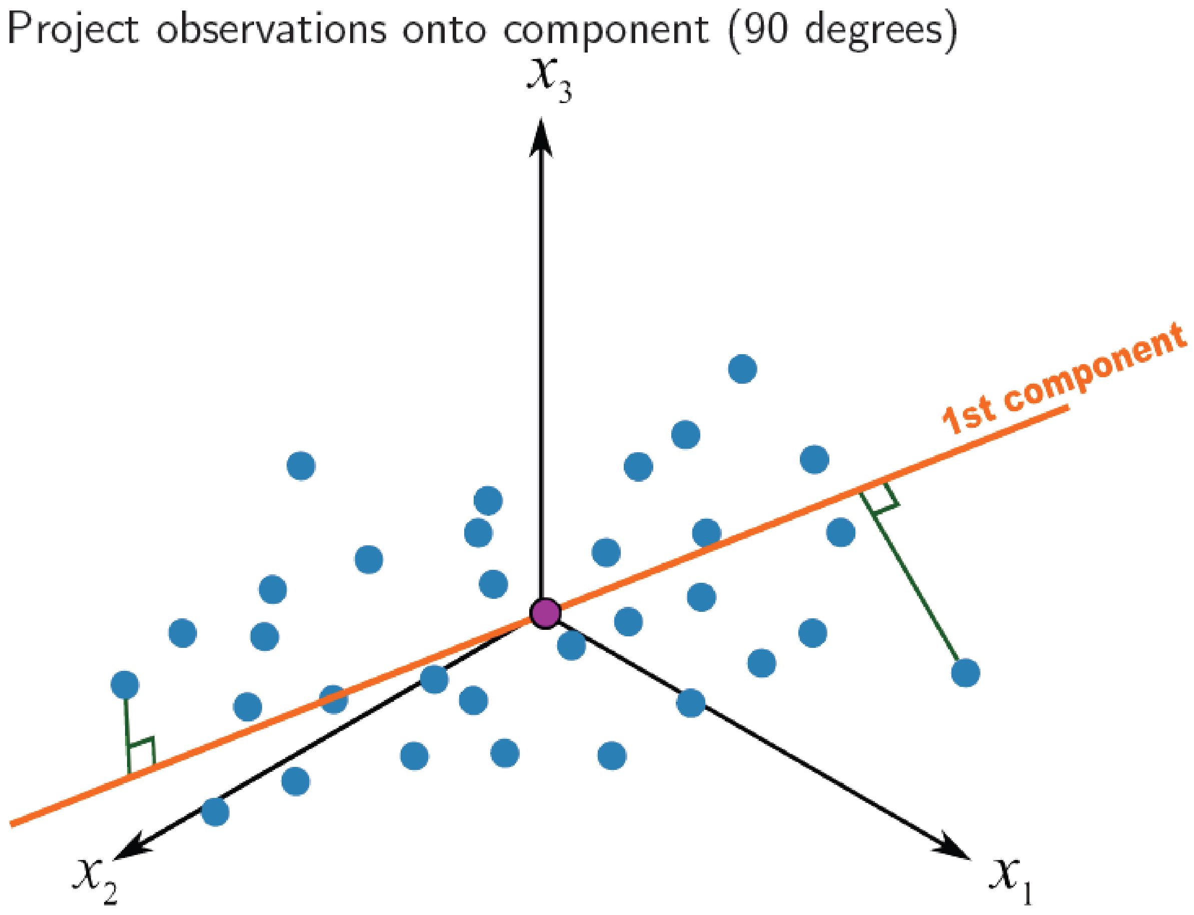
Mean center and scale

Geometric Interpretation

- By picking the direction of largest elongation, this direction will pass through the center of the swarm.
- This line is called the first principal component (PC1) and points in the direction of the maximum data variation.

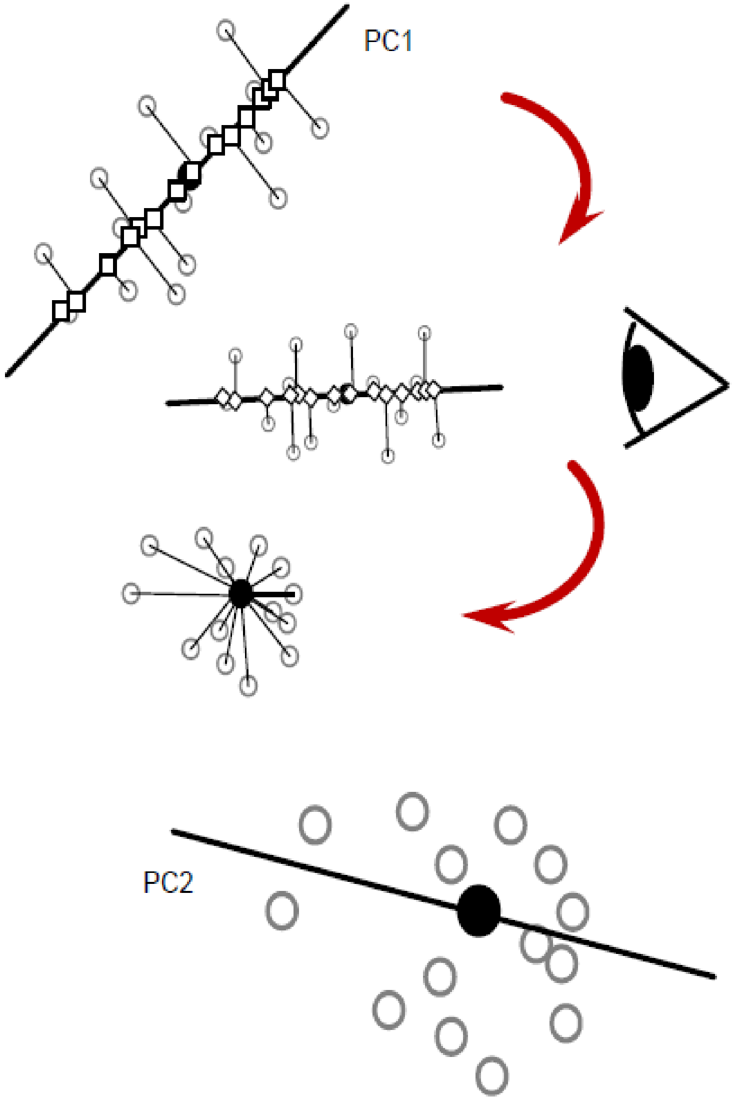


Geometric Interpretation



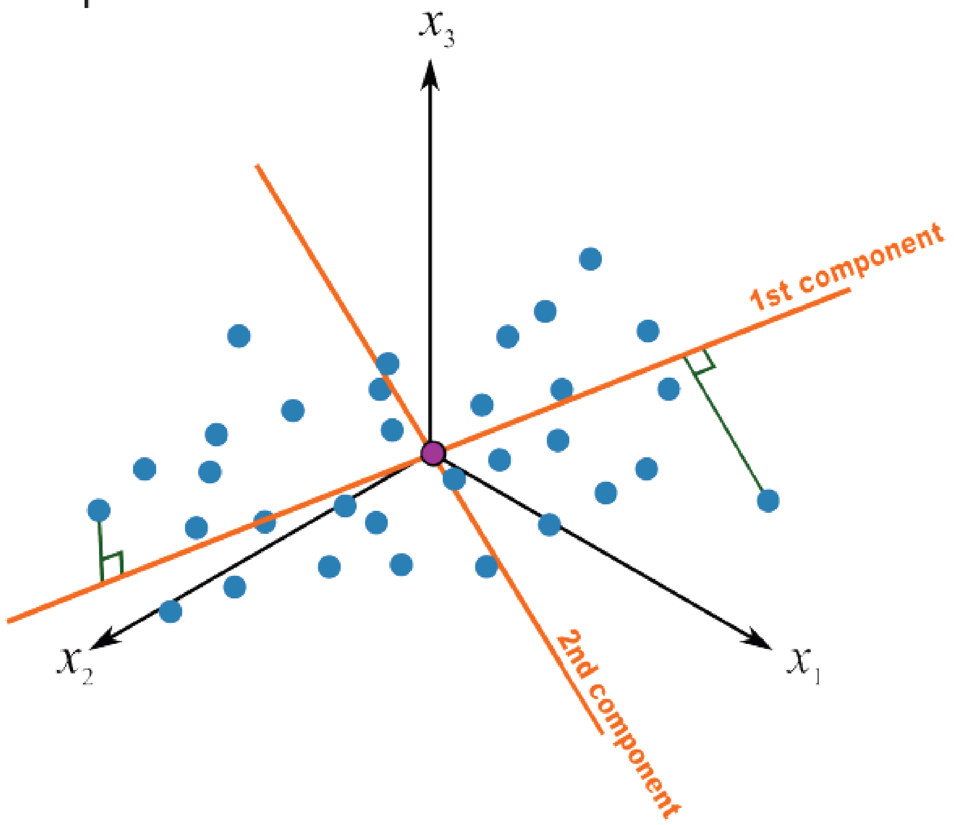
Geometric Interpretation

- The first principal component may not be enough to describe the data variation.
- By projecting the samples onto the new coordinate system, there is still unexplained variance (combination of the individual residuals)
- If we repeat the projection process in the remaining part of the space, we will find the second principal component (PC2).



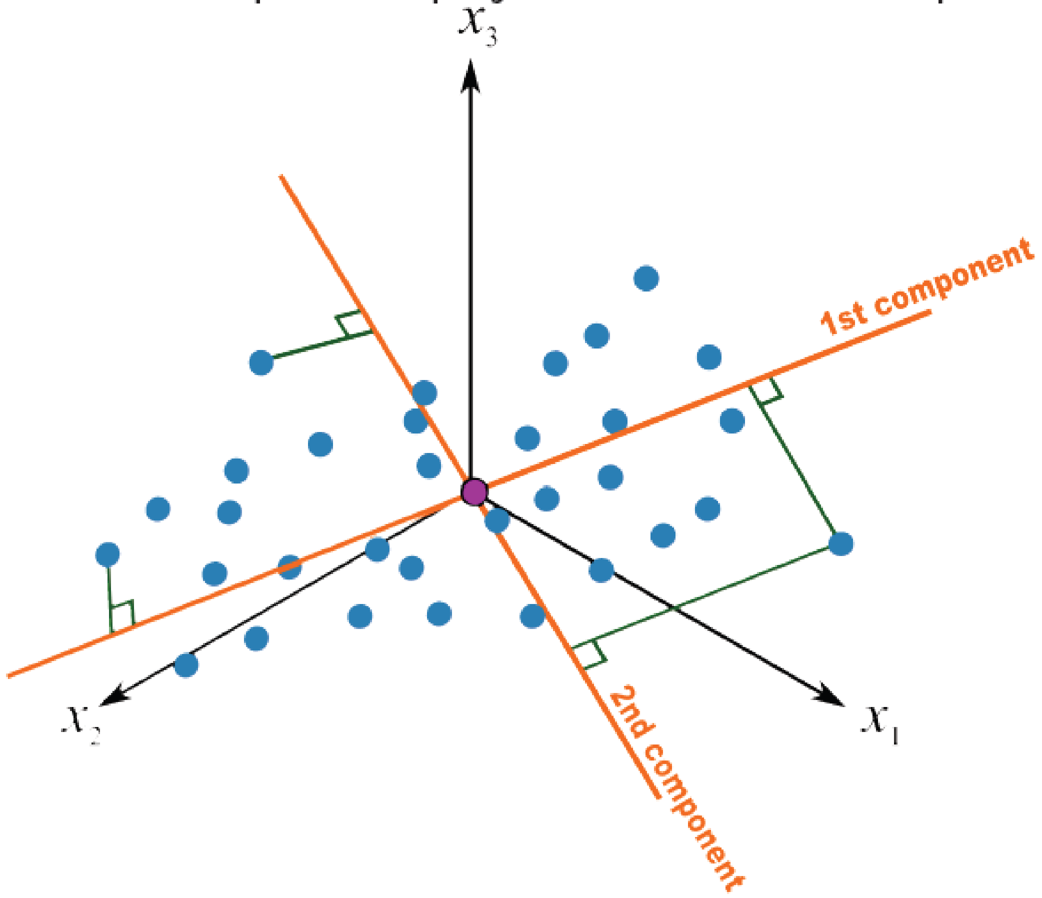
Geometric Interpretation

Second component: best-fit line; perpendicular to 1st component



Geometric Interpretation

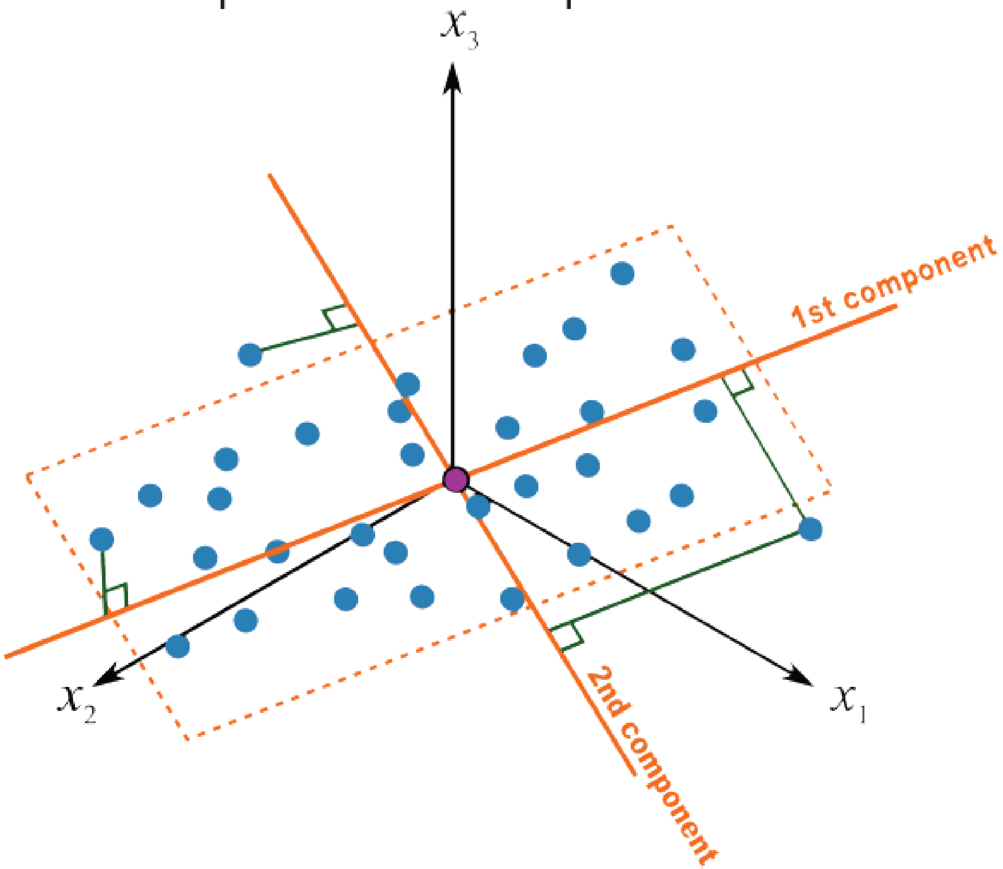
Second component: project onto second component



Geometric Interpretation

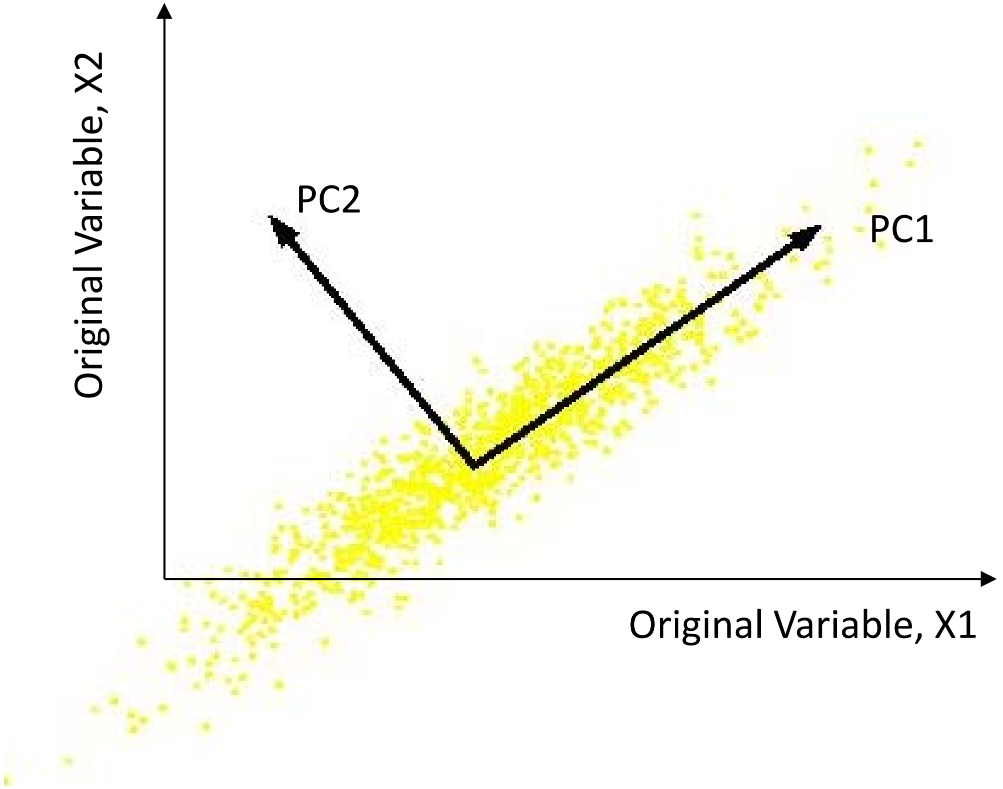
- With 2 principal components we can build a plane onto which the projections of the swarm lie closer yet to the original variable space.
- We have found the 2-dimensional 'window' which best describes the data.
- The process can be continued to find more PCs.

The 2 components create a plane



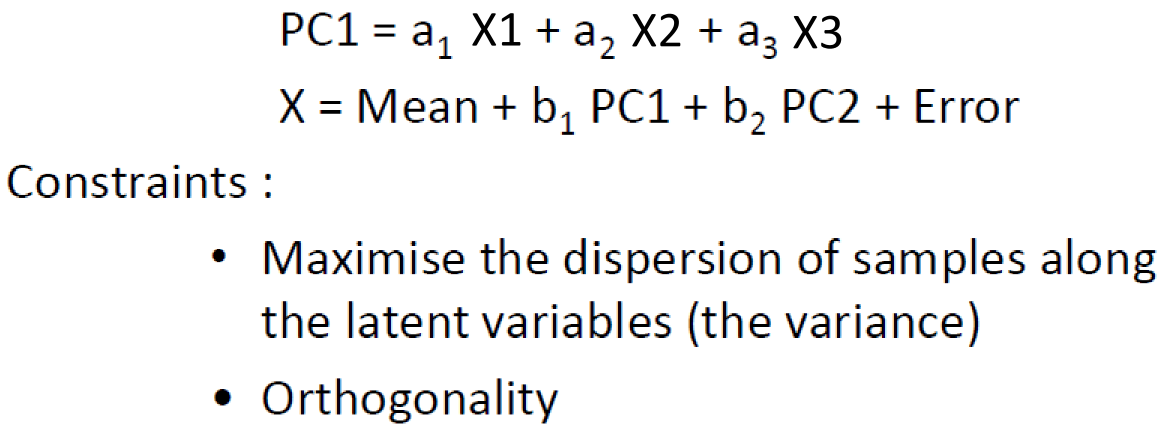
Geometric Interpretation

- In summary,
  - PCA finds a few orthogonal axes of greatest variance in data.



Geometric Interpretation

- New latent variables are linear combinations of the original variables.



$$PC1 = a_1\ X1 + a_2\ X2 + a_3\ X3$$

$$X = Mean + b_1\ PC1 + b_2\ PC2 + Error$$

Constraints :

- Maximise the dispersion of samples along the latent variables (the variance)
- Orthogonality

Mathematical Derivation

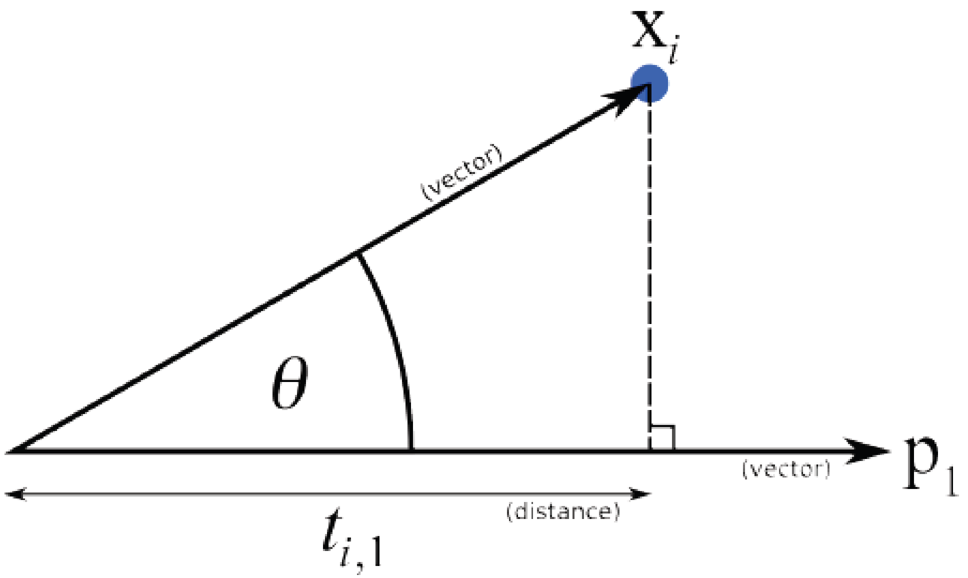
What has this done?

Break **X** down into 2 parts:

- projected points "*on the plane*"
- residual distance "*off the plane*"

Mathematical Derivation

- From linear algebra (or engineering mathematics),



$$\cos\theta = \frac{\text{adjacent length}}{\text{hypotenuse}} = \frac{t_{i,1}}{\|\mathbf{x}_i\|} \qquad \text{and also } \cos\theta = \frac{\mathbf{x}_i^T\mathbf{p}_1}{\|\mathbf{x}_i\|\|\mathbf{p}_1\|}$$

$$\begin{aligned} \frac{t_{i,1}}{\|\mathbf{x}_i\|} &= \frac{\mathbf{x}_i^T\mathbf{p}_1}{\|\mathbf{x}_i\|\|\mathbf{p}_1\|} \\ t_{i,1} &= \mathbf{x}_i^T\mathbf{p}_1 \\ (1\times 1) &= (1\times K)(K\times 1) \end{aligned}$$

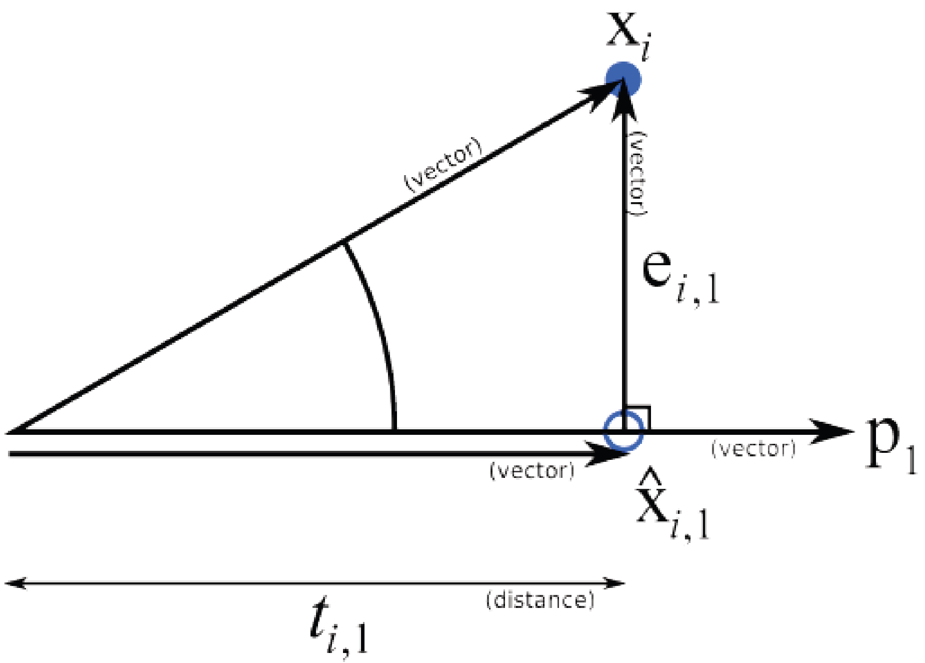
Mathematical Derivation

$$\begin{aligned} t_{i,1} &= \mathbf{x}_i^T \mathbf{p}_1 \\ &= x_{i,1}p_{1,1} + x_{i,2}p_{2,1} + \ldots + x_{i,k}p_{k,1} + \ldots + x_{i,K}p_{K,1} \end{aligned}$$

- K separate terms: added up (i.e. linear combination) to give t_1
- Entire data set: $\mathbf{T} = \mathbf{XP}$

Predicted value for each observations

- $\hat{\mathbf{x}}_i$:projected version of $\mathbf{x}_i$



$$\begin{aligned} \hat{\mathbf{x}}_{i,1}^T &= t_{i,1}\mathbf{p}_1^T \\ (1 \times K) &= (1 \times 1)(1 \times K) \end{aligned}$$

Predicted value for each observations

Residual vector:

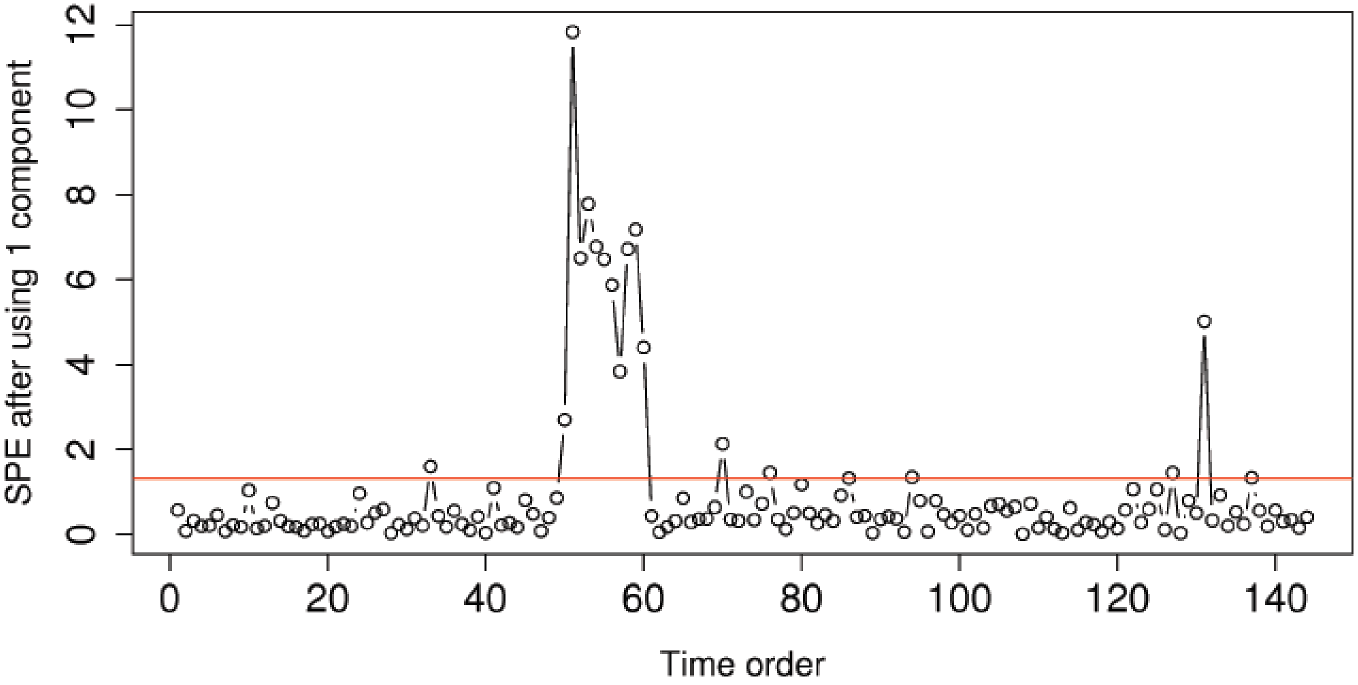
$$\begin{array}{rcl} \mathbf{e}_{i,A}^{T} & = & \mathbf{x}_{i}^{T} - \hat{\mathbf{x}}_{i,A}^{T} \\ (1 \times K) & = & (1 \times K) - (1 \times K) \end{array}$$

Residual distance:

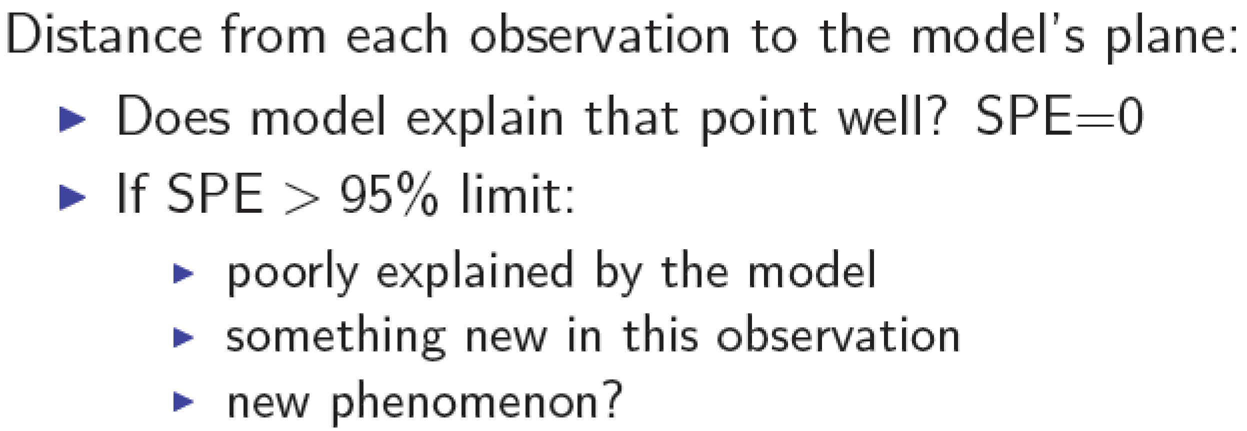
$$\begin{array}{rcl} \mathrm{SPE}_{i} & = & \sqrt{\mathbf{e}_{i,A}^{T}\mathbf{e}_{i,A}} \\ (1 \times 1) & = & (1 \times K)(K \times 1) \end{array}$$

Square Prediction Error

- $\mathbf{e}'_{i,A} = \mathbf{x}'_i - \widehat{\mathbf{x}}'_{i,A}$
- $\text{SPE}_i = \sqrt{e_{i,1}^2 + e_{i,2}^2 + \ldots + e_{i,K}^2}$
- Smallest SPE: $\text{SPE}_i = 0$
- Calculate 95% or 99% confidence limit



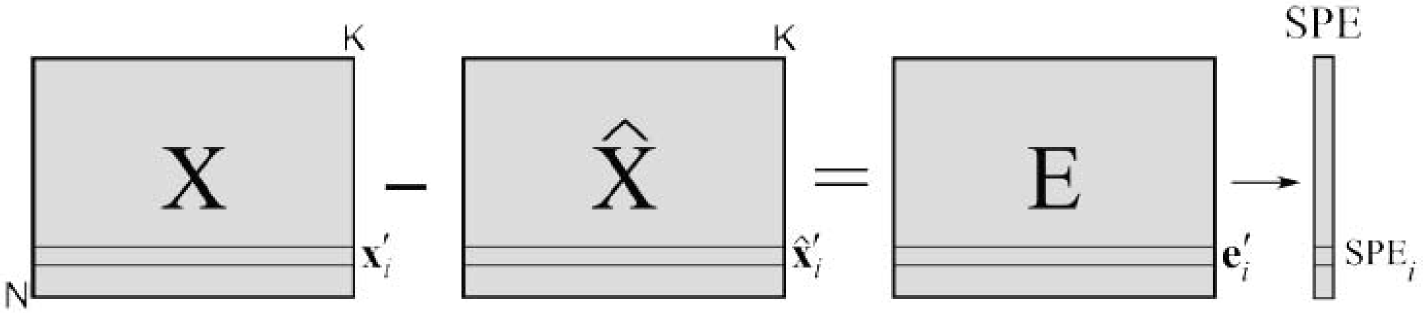
Square Prediction Error



Distance from each observation to the model's plane:

- Does model explain that point well? SPE=0
- If SPE $>$ 95% limit:
  - poorly explained by the model
  - something new in this observation
  - new phenomenon?

Square Prediction Error

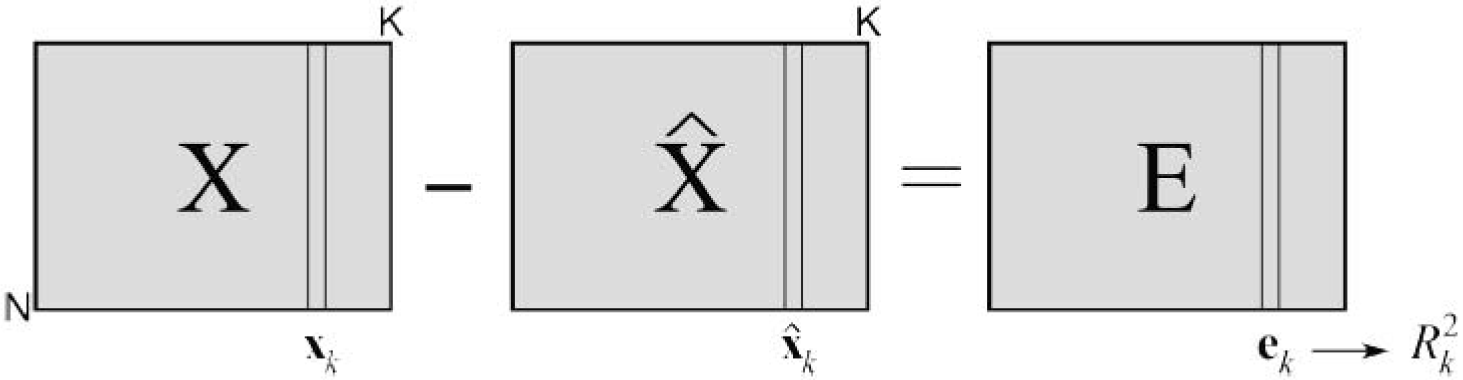


- $\mathbf{e}_i' = \mathbf{x}_i' - \widehat{\mathbf{x}}_i'$
- $\text{SPE}_i = \sqrt{e_{i,1}^2 + e_{i,2}^2 + \ldots + e_{i,K}^2}$

$$\mathbf{e}_i' = \left[(x_{i,1} - \hat{x}_{i,1}) \quad (x_{i,2} - \hat{x}_{i,2}) \quad \ldots \quad (x_{i,k} - \hat{x}_{i,k}) \quad \ldots \quad (x_{i,K} - \hat{x}_{i,K}) \right]$$

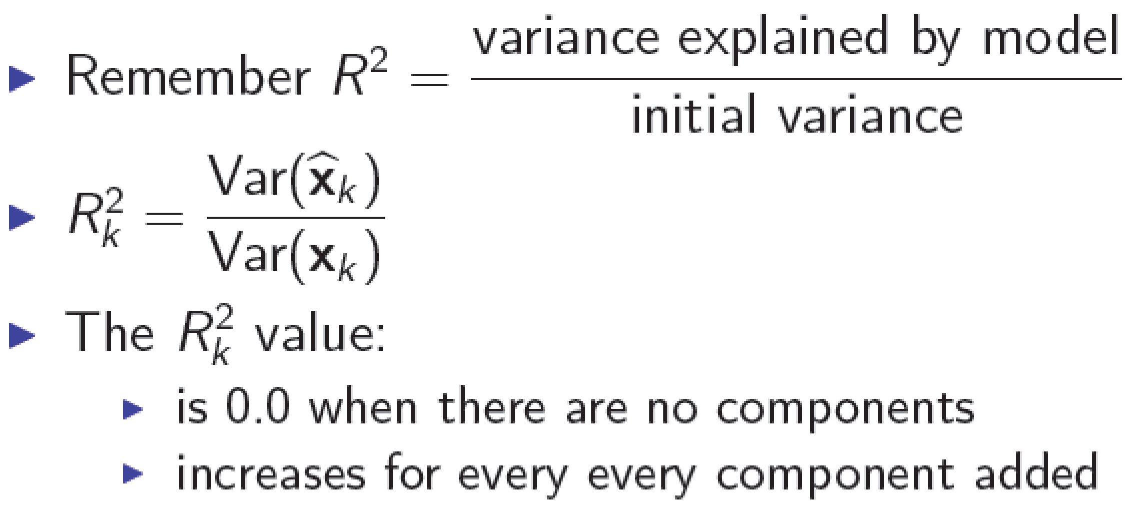
Column Residual

- SPE is the row residual for **X**
- Residuals also calculated for each column



- How well each column is explained by the model

Column Residual

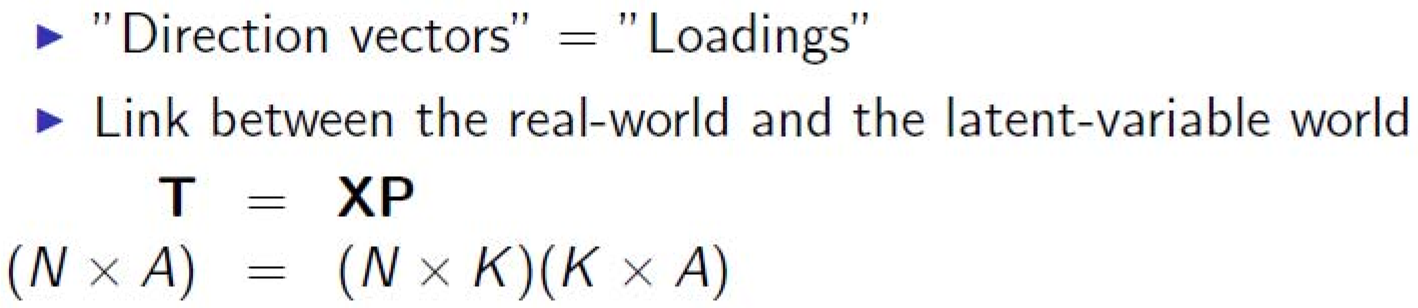


- Remember $R^2 = \dfrac{\text{variance explained by model}}{\text{initial variance}}$
- $R_k^2 = \dfrac{\mathrm{Var}(\widehat{\mathbf{x}}_k)}{\mathrm{Var}(\mathbf{x}_k)}$
- The R_k^2 value:
  - is 0.0 when there are no components
  - increases for every every component added

Whole Matrix Residual

- $\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} = \widehat{\mathbf{X}} + \mathbf{E}$
- How well does the model fit the data?
- $R^2 = \dfrac{\text{Var}(\widehat{\mathbf{X}})}{\text{Var}(\mathbf{X})}$
  - $R^2 = 0.0$ when there are no components
  - R^2 increases with every component added
  - $R^2_{a=0} > R^2_{a=1} > R^2_{a=2} > \ldots > R^2_{a=A} = 1.0$

More about direction vectors

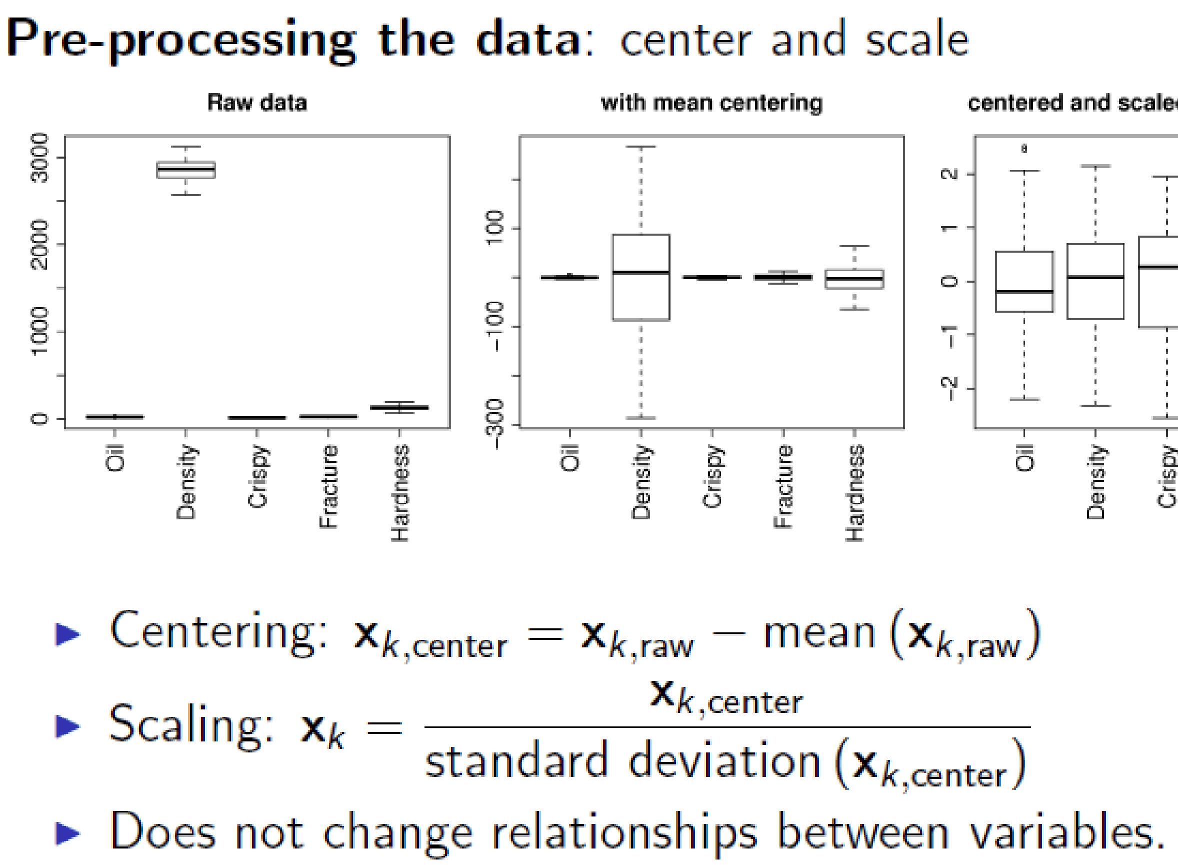


- "Direction vectors" = "Loadings"
- Link between the real-world and the latent-variable world

$$\begin{aligned} \mathbf{T} &= \mathbf{XP} \\ (N \times A) &= (N \times K)(K \times A) \end{aligned}$$

※ Statistically, loading vectors are eigenvectors of X^TX. Then how about eigenvalues?

Preprocessing



Pre-processing the data: center and scale

- Centering: $\mathbf{x}_{k,\text{center}} = \mathbf{x}_{k,\text{raw}} - \text{mean}\left(\mathbf{x}_{k,\text{raw}}\right)$
- Scaling: $\mathbf{x}_k = \dfrac{\mathbf{x}_{k,\text{center}}}{\text{standard deviation}\left(\mathbf{x}_{k,\text{center}}\right)}$
- Does not change relationships between variables.

Remember

$$Z = \frac{X - \mu}{\sigma}$$?

More on preprocessing Data

- Modifies the columns of X before building the model
- Center
- Scale
- Add transformations:
  - use log(T) instead of temperature, T
  - use 1/P instead of pressure, P
  - use sqrt(F) instead of flow, F

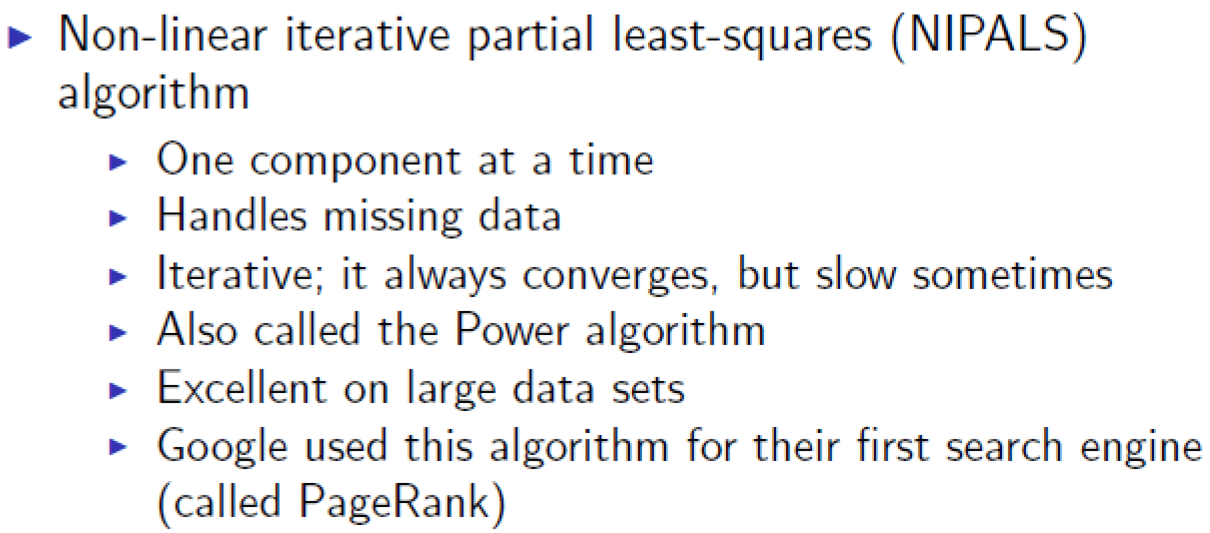
Add extra columns to X:

- heat balance
- dimensionless numbers
- square terms: $x_1^2, \quad x_2^2, \ldots$
- interaction terms: $x_1x_2, \quad x_1x_3, \quad x_2x_3 \quad \ldots$

How is PCA calculated?

- Eigenvalue decomposition
  - loadings are the eigenvectors of $\mathbf{X}'\mathbf{X}$.
  - once you have the eigenvectors, then $\mathbf{T} = \mathbf{XP}$
  - eigenvalues are the variances of the scores, s_a^2
- Singular value decomposition
  - $\mathbf{X} = \mathbf{U\Sigma V}' = \mathbf{TP}'$
  - scores, $\mathbf{T} = \mathbf{U\Sigma}$ and the loadings, $\mathbf{P} = \mathbf{V}$

How is PCA calculated?

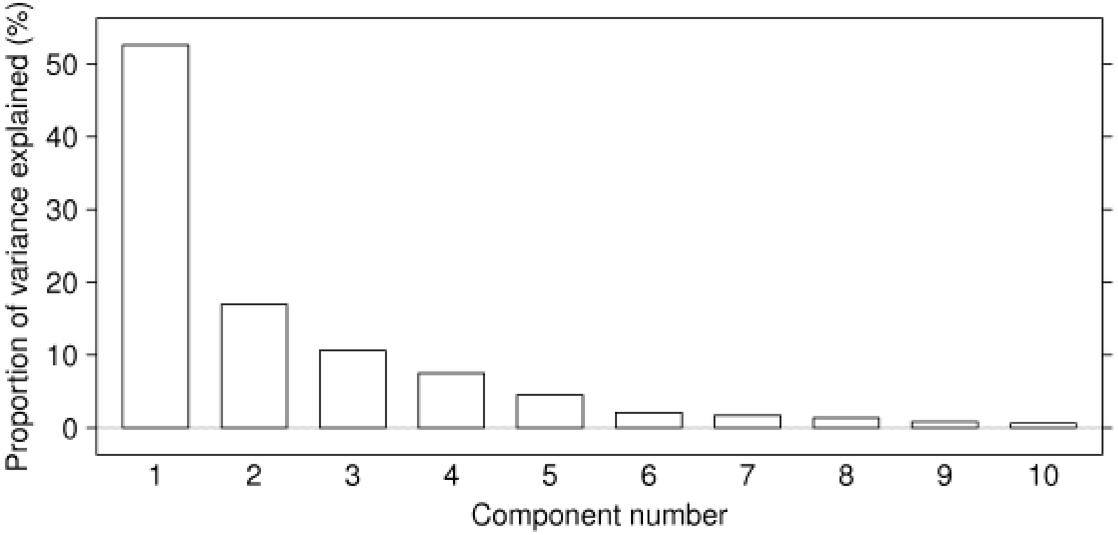


- Non-linear iterative partial least-squares (NIPALS) algorithm
  - One component at a time
  - Handles missing data
  - Iterative; it always converges, but slow sometimes
  - Also called the Power algorithm
  - Excellent on large data sets
  - Google used this algorithm for their first search engine (called PageRank)

More details next.

How many components?

- Eigenvalues
  - sum of the eigenvalues $= \sum_{a}^{a=K} \lambda_a = K$
  - keep adding components as long as $\lambda_a > 1$
- Plot R^2 for each component

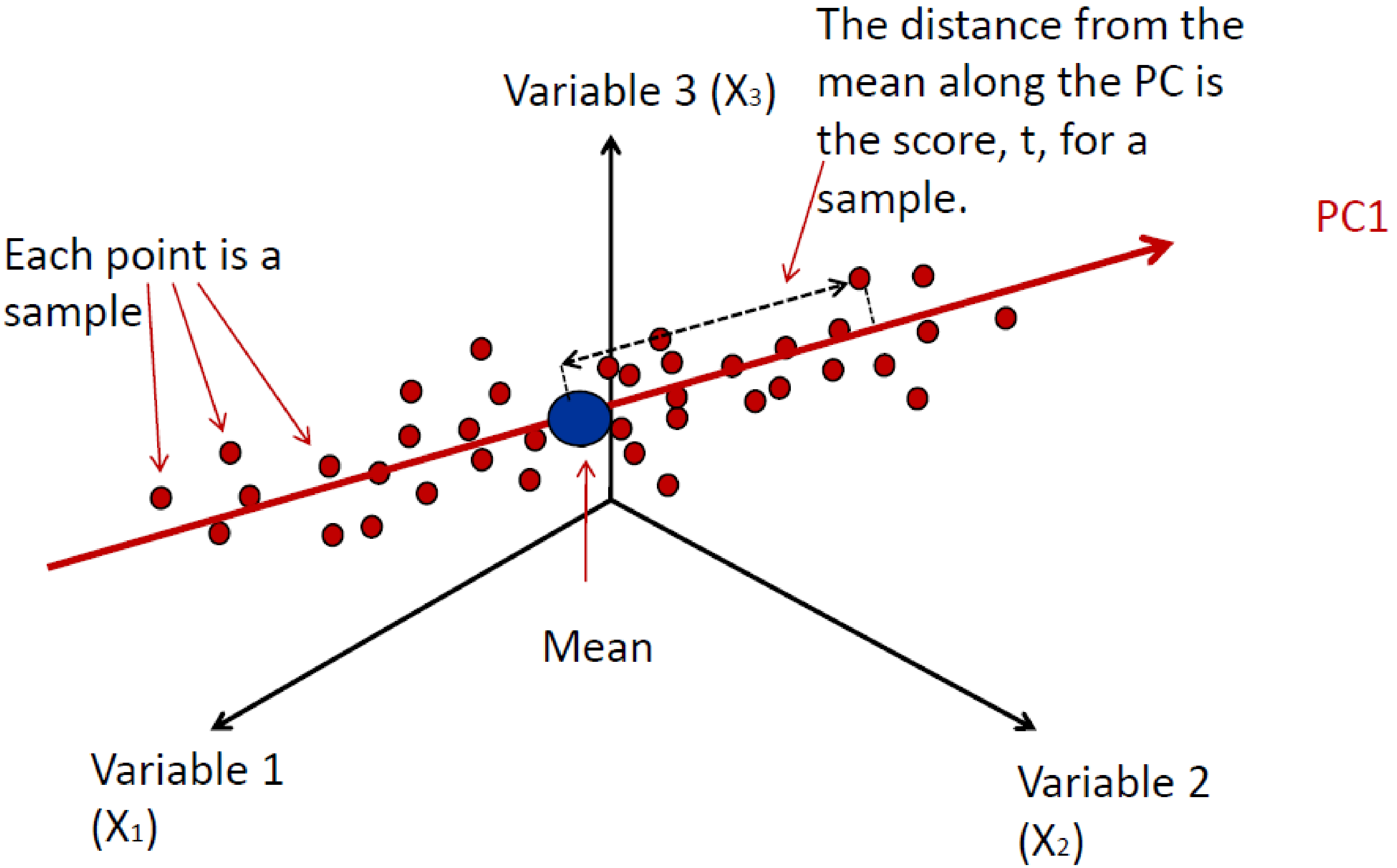


Called scree plot

- Use cross-validation

Review of PCA

- What is score?



Variable 3 (X_3)

The distance from the mean along the PC is the score, t, for a sample.

PC1

Each point is a sample

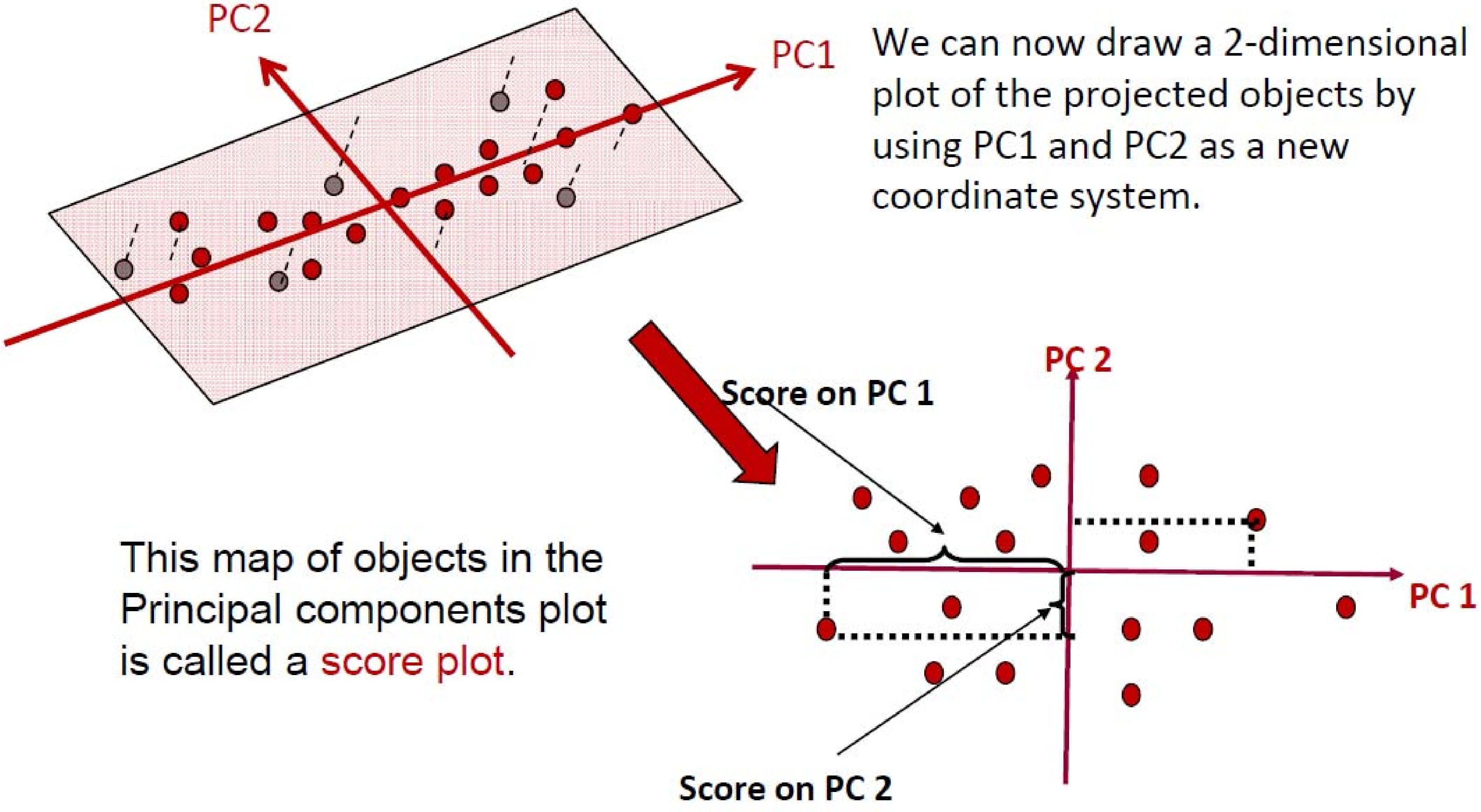
Mean

Variable 1 (X_1)

Variable 2 (X_2)

Review of PCA

- Score plot – low dimensional summary of samples



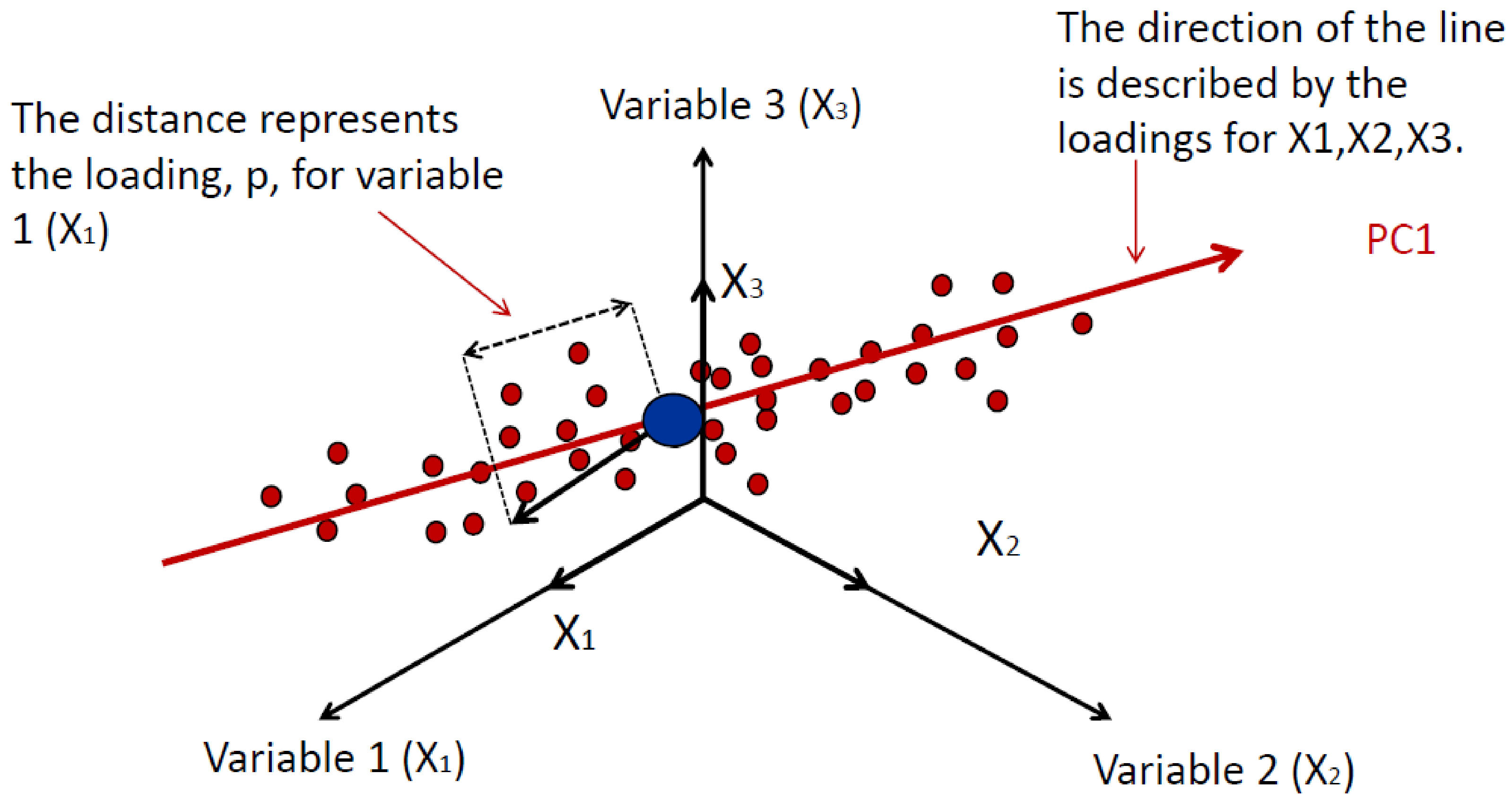
We can now draw a 2-dimensional plot of the projected objects by using PC1 and PC2 as a new coordinate system.

This map of objects in the Principal components plot is called a score plot.

Review of PCA

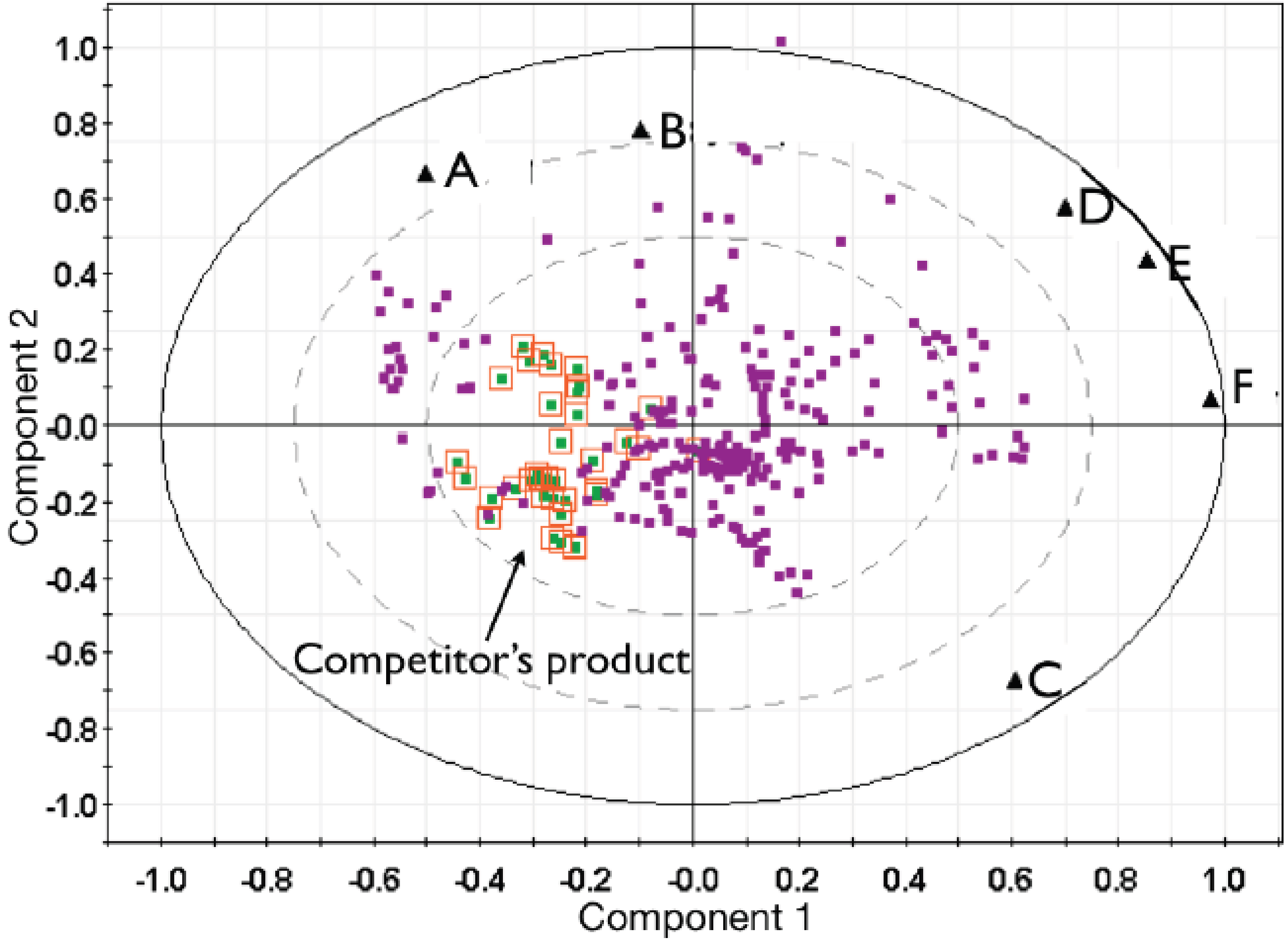
- What is loading?

Coefficients in the linear combination $PC1 = a_1\ X1 + a_2\ X2 + a_3\ X3$


The distance represents the loading, p, for variable 1 (X_1)
Variable 3 (X_3)
The direction of the line is described by the loadings for X1,X2,X3.
PC1
X_3
X_2
X_1
Variable 1 (X_1)
Variable 2 (X_2)

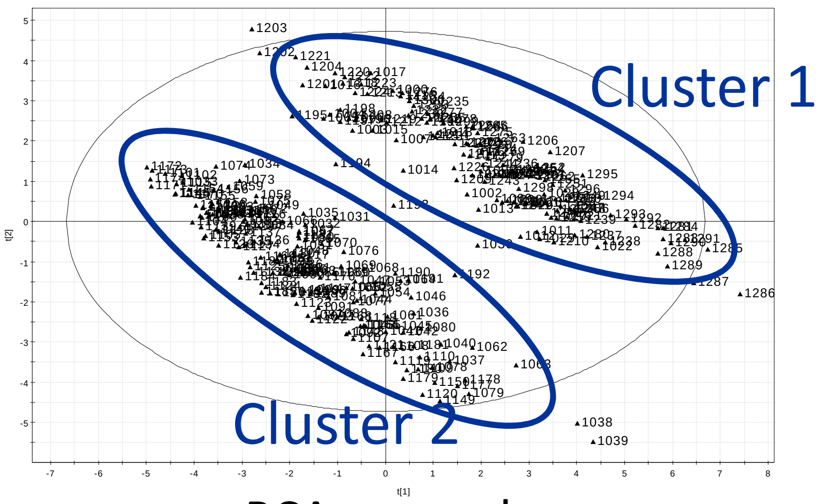
Use of PCA

- Improved Process Understanding

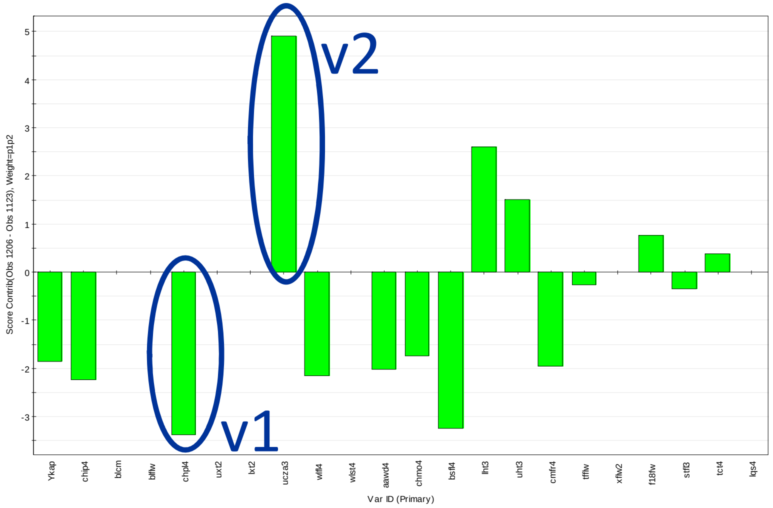


Use of PCA

- Troubleshooting Process Problem



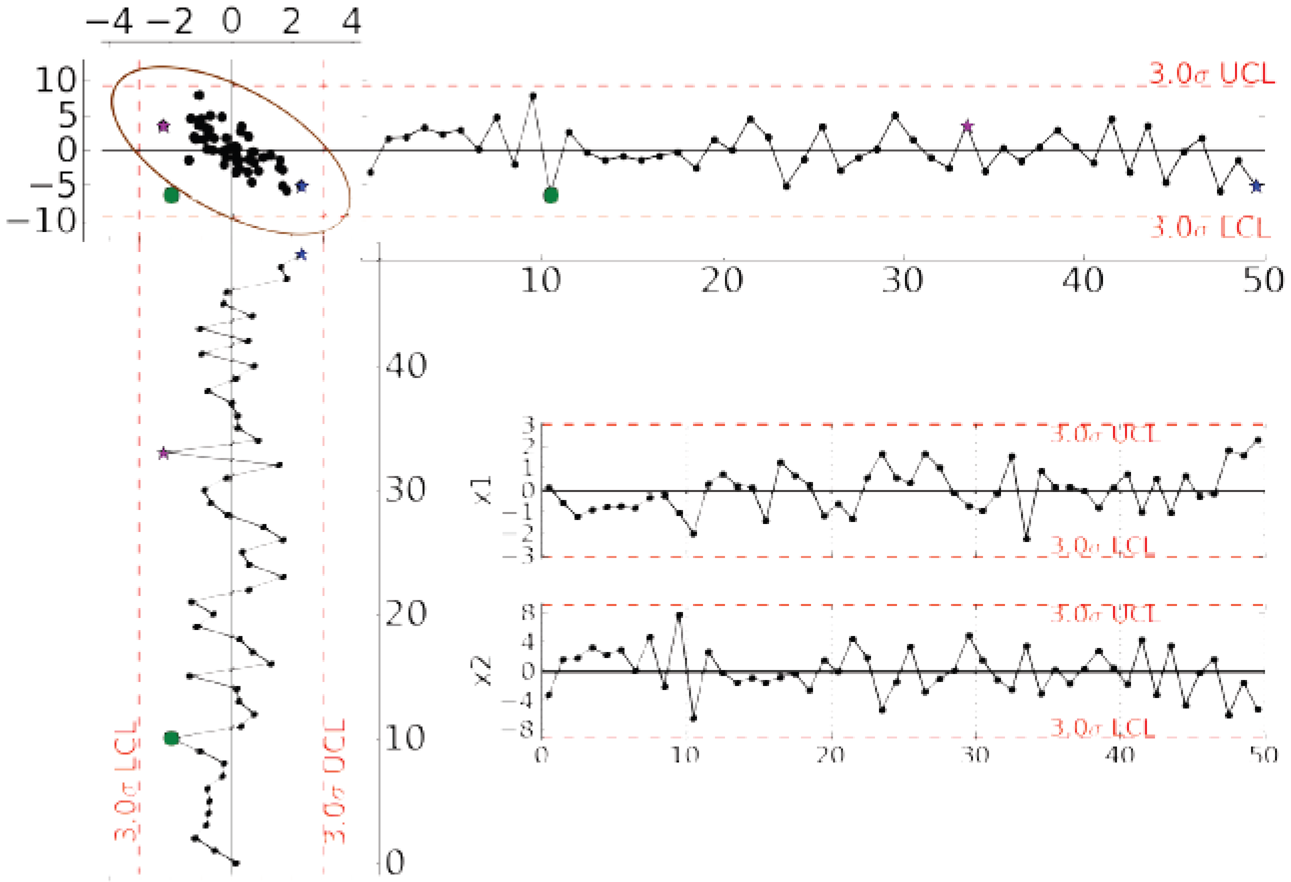
PCA score plot



PCA Contribution plot

Use of PCA

- Multivariate Stastiscal Process Control (MSPC)



In the next lecture

- Tutorials
- NIPALS algorithm
- A bit more on PCA
- Assignment #1