

Multi-Omics Integration

Ahmed Karam Mohamed

Senior Bioinformatician in Proteomics and Metabolomics

Children's Cancer Hospital Egypt 57357

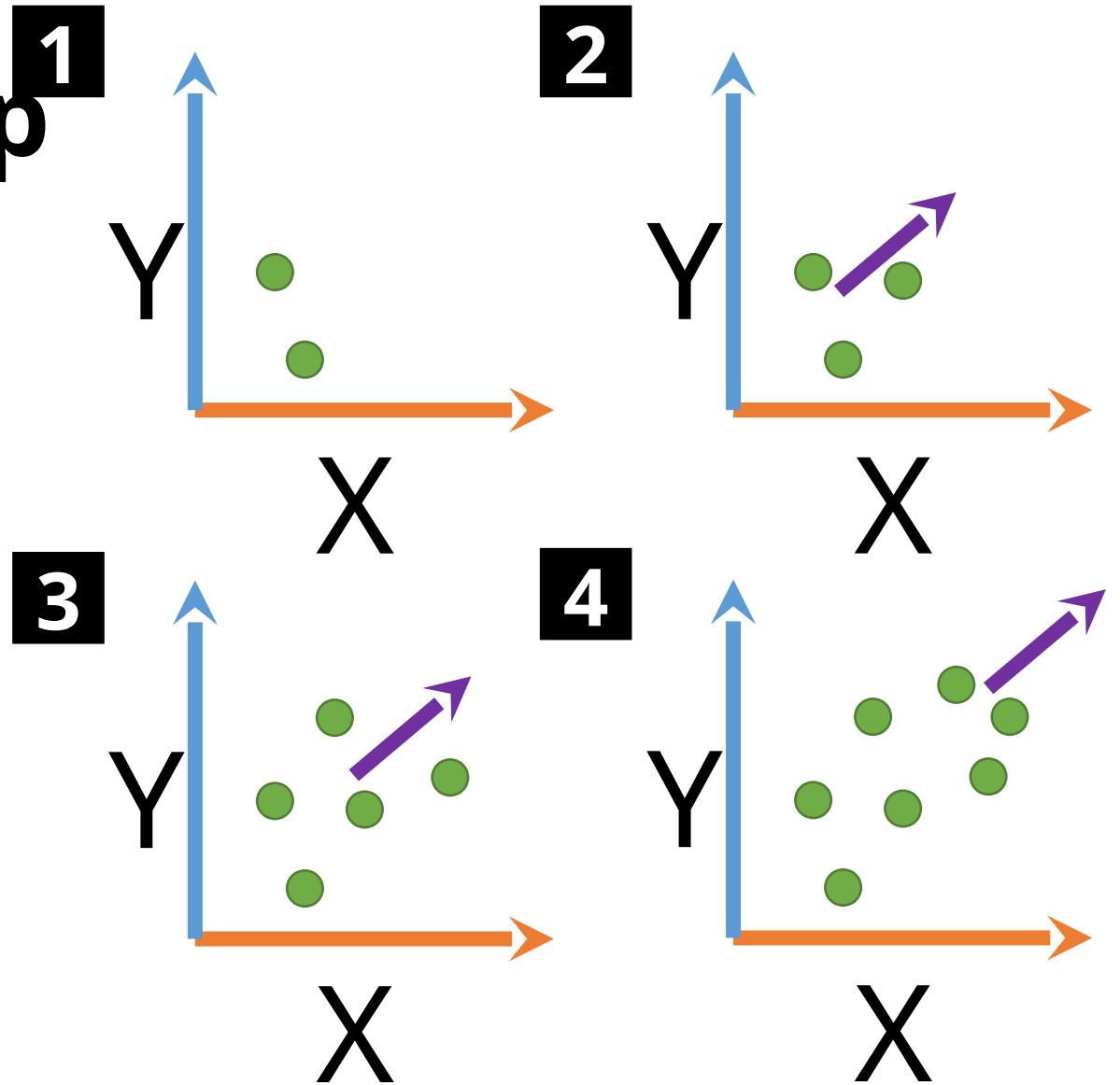
Outlines

- Linear Relationship.
- Least Squares.
- Principal Component Analysis (PCA).
- Factor Analysis (FA).
- Multiple Factor Analysis (MFA).
- Partial Least Squares (PLS).
- Discriminant Analysis (DA).
- Partial Least Squares-Discriminant Analysis (PLS-DA).
- General Multi-Omics Integration Protocol.

Linear Relationship

Linear Relationship¹

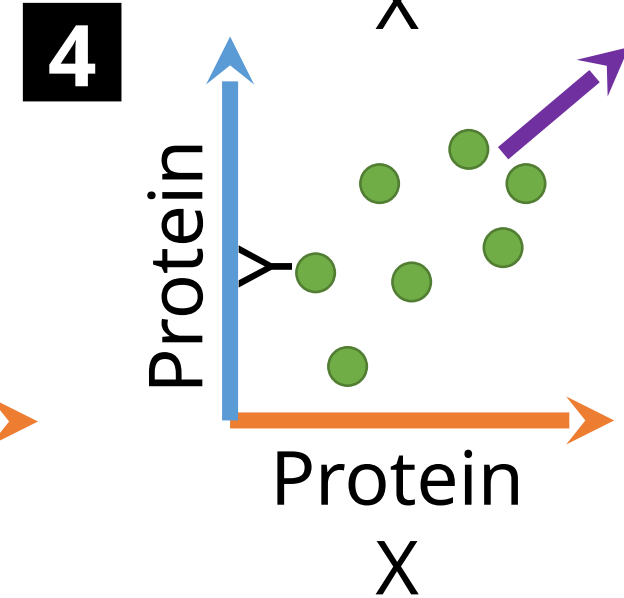
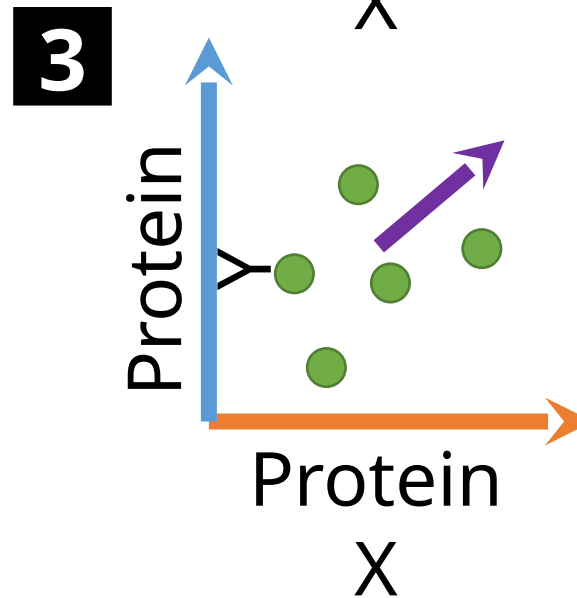
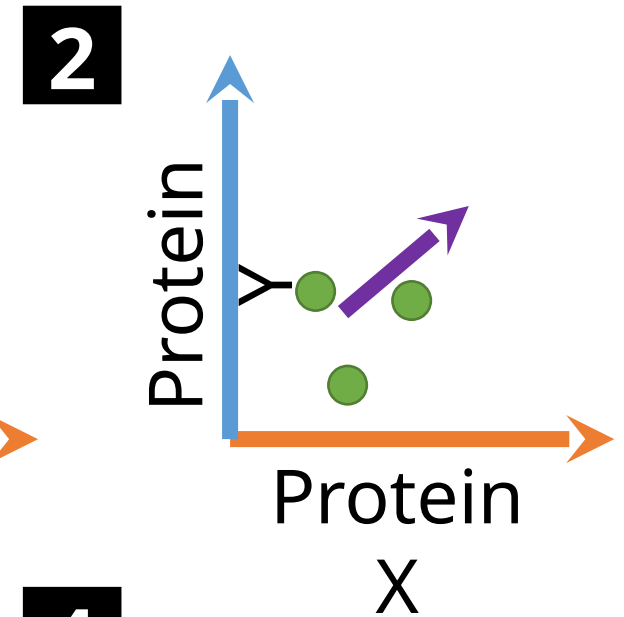
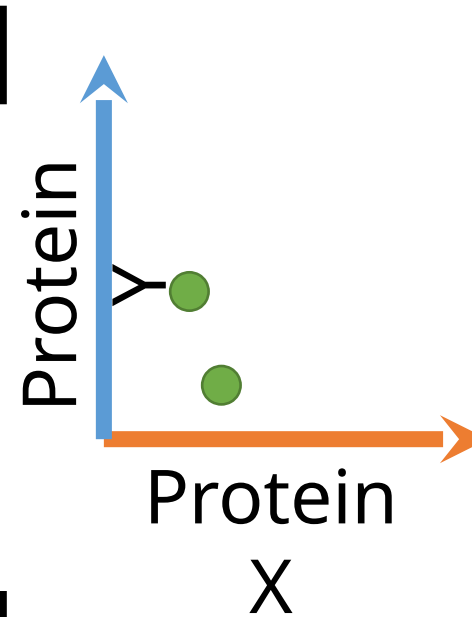
A linear relationship between two variables, say X and Y , means that there's a constant and predictable change in Y for every change in X .



● Sample

Linear Relationship¹

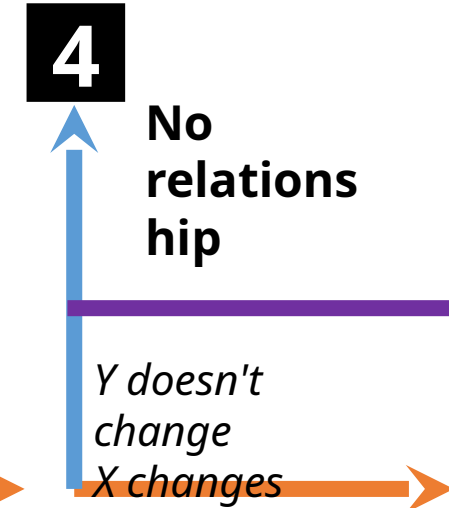
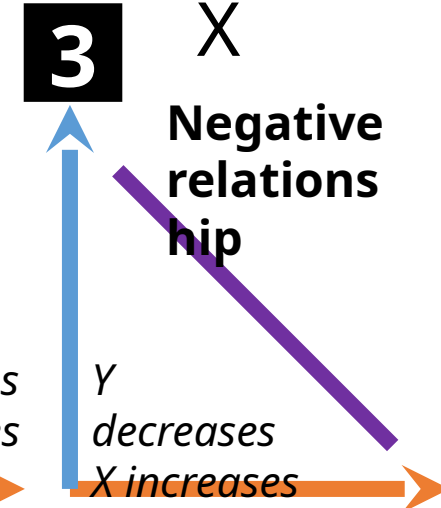
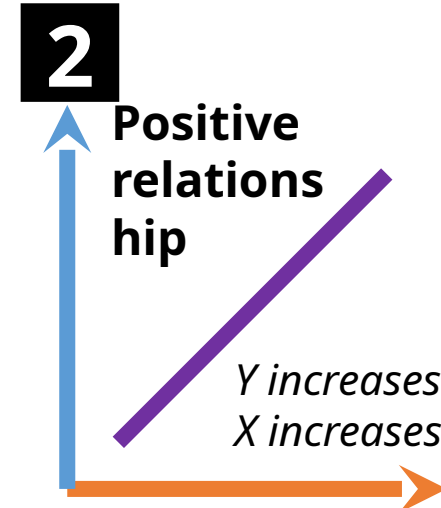
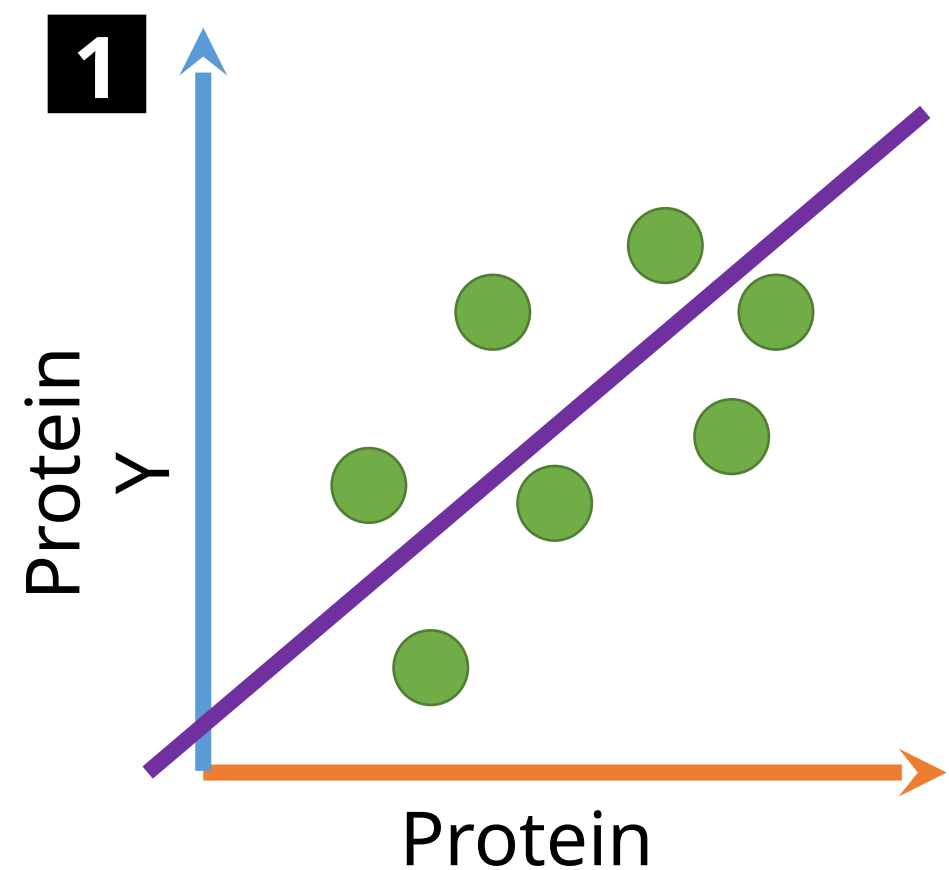
A linear relationship between two **proteins**, say **protein X** and **protein Y**, means that there's a constant and predictable change in **protein Y** for every change in **protein X**.



Linear Relationship

After you plot the data points, they fall along a straight line.

This line could be
(1) slanted upwards (positive relationship),
(2) slanted downwards (negative relationship), or
(3) perfectly horizontal (no relationship between X and Y).



Linear Relationship

A linear relationship can be expressed by a linear equation of the form:

$$Y = aX + b$$

where:

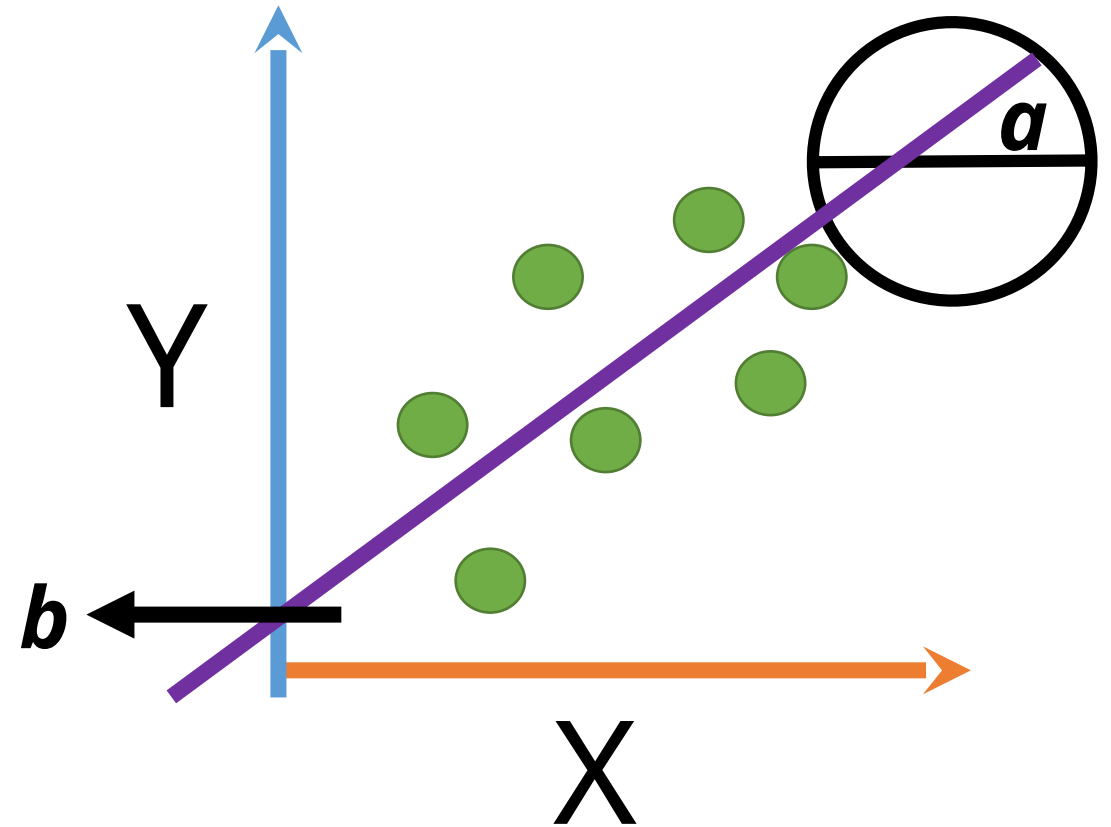
a is the slope of the line.

It tells you how much Y changes for every 1 unit change in X.

- Positive slope (***a*** > 0) means positive relationship.
- Negative slope (***a*** < 0) means negative relationship.
- Zero slope (***a*** = 0) means no relationship.

b is the y-intercept.

It's the point where the line crosses the Y-axis (X = 0).

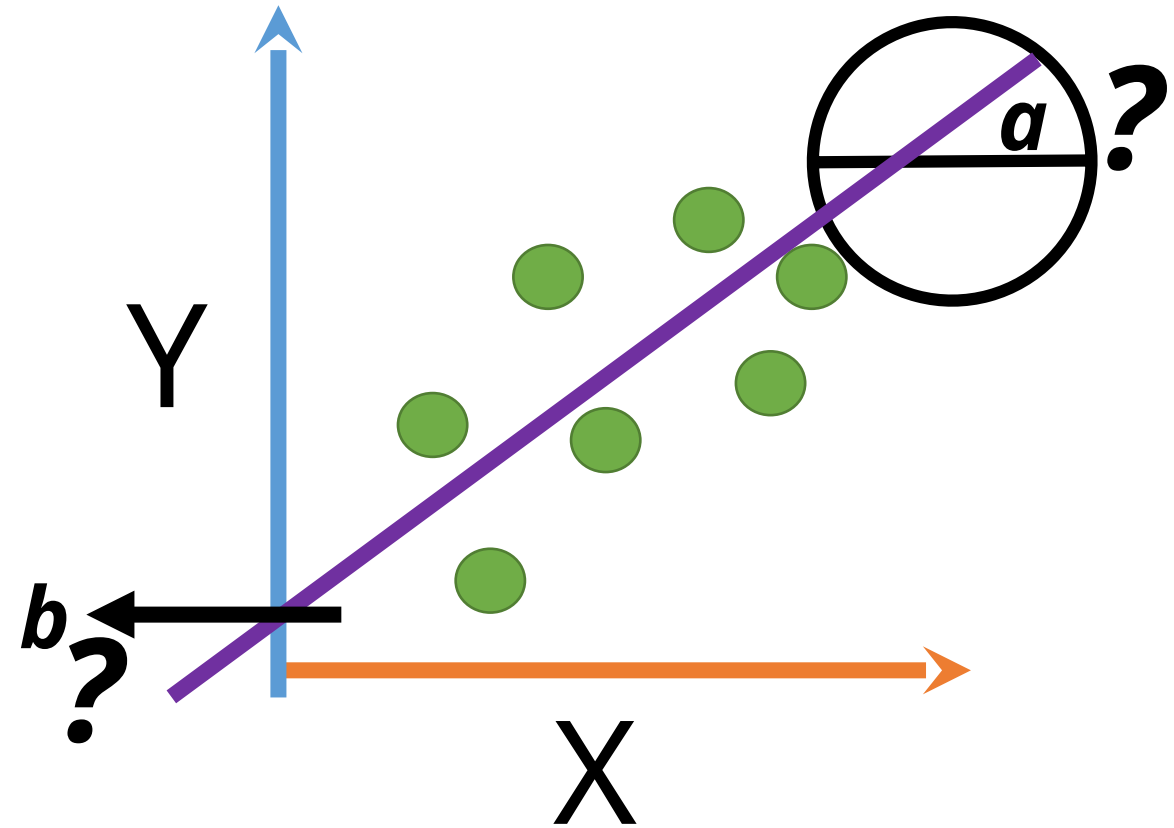


Linear Relationship

- Protein **Y** is measured.
- Protein **X** is measured.
- Individuals plotted to the graph.

But !

How we can know the ***a and b***
???

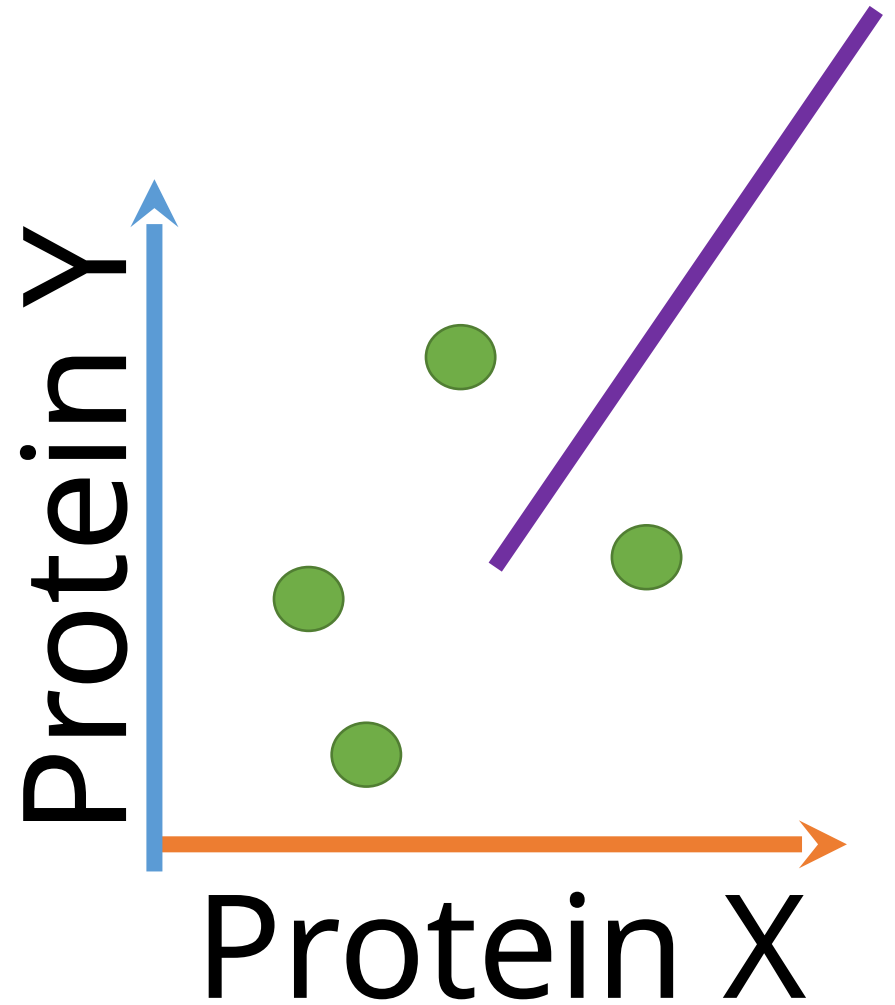


Least Squares

Least Squares

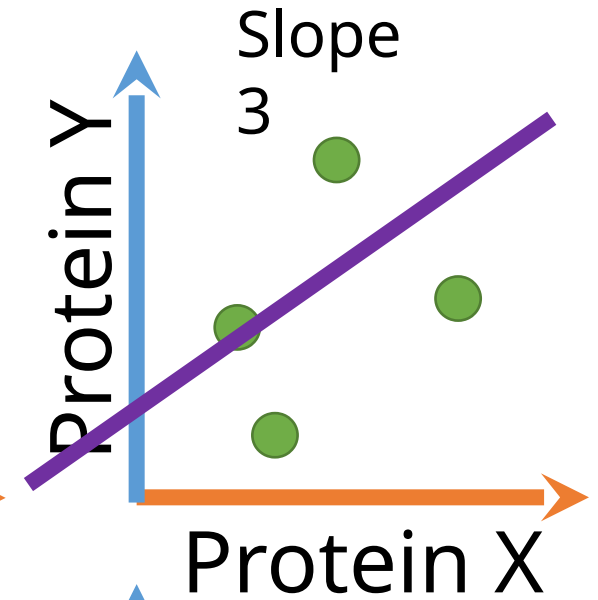
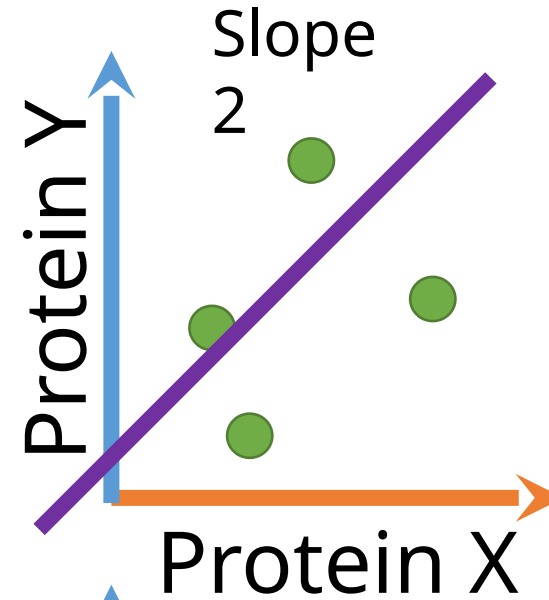
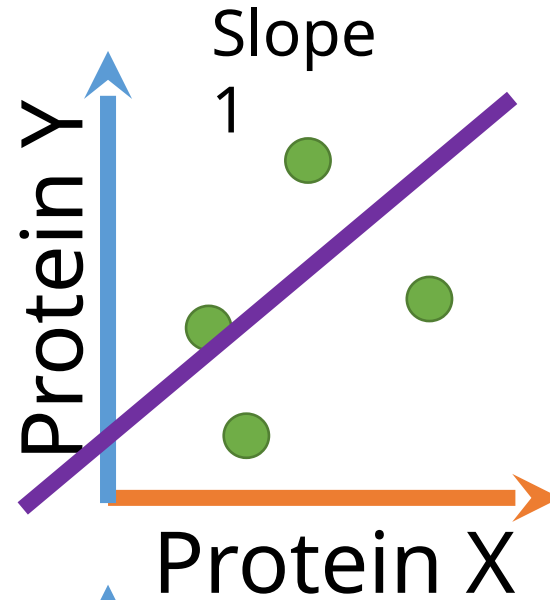
- Protein **Y** is measured.
- Protein **X** is measured.
- 4 Individuals plotted to the graph.

To draw a the representative line, (1) assume 3 random values for the slope and 2 random values for Y-intercept.

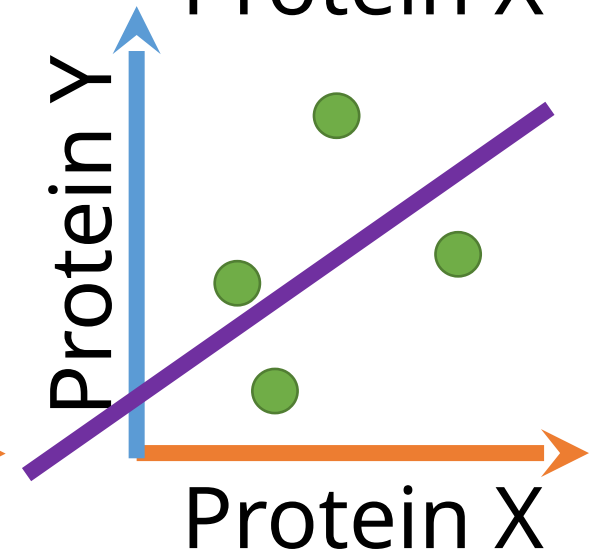
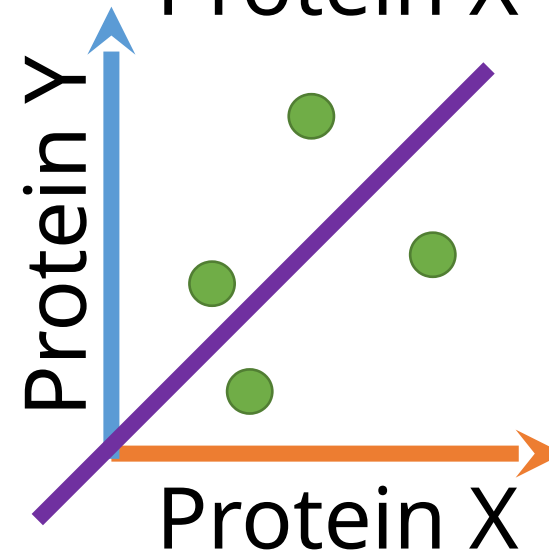
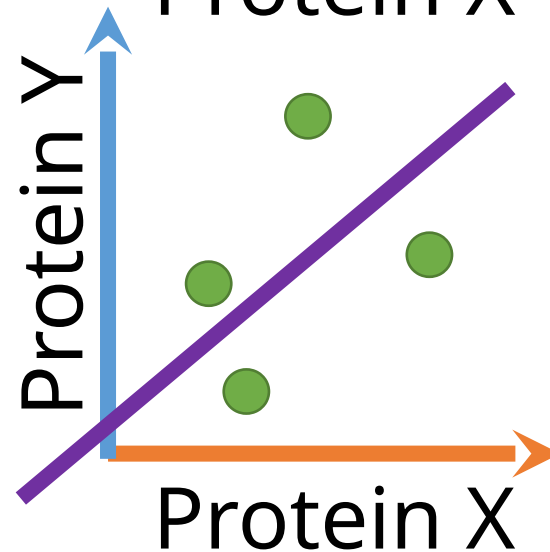


Least Squares

Y-intercept
1

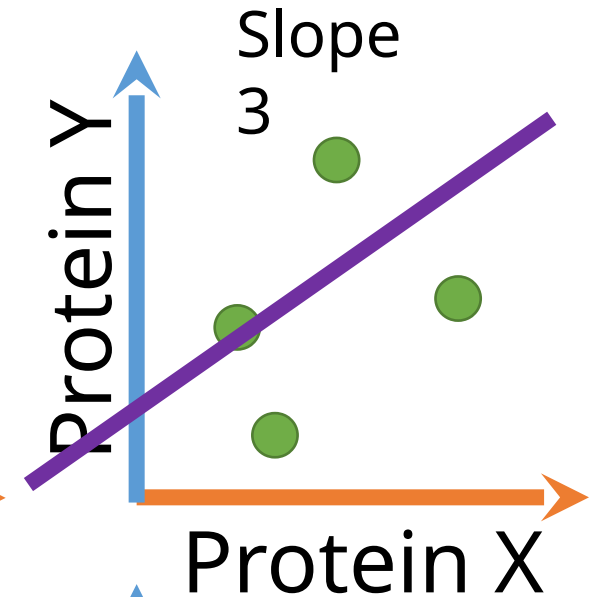
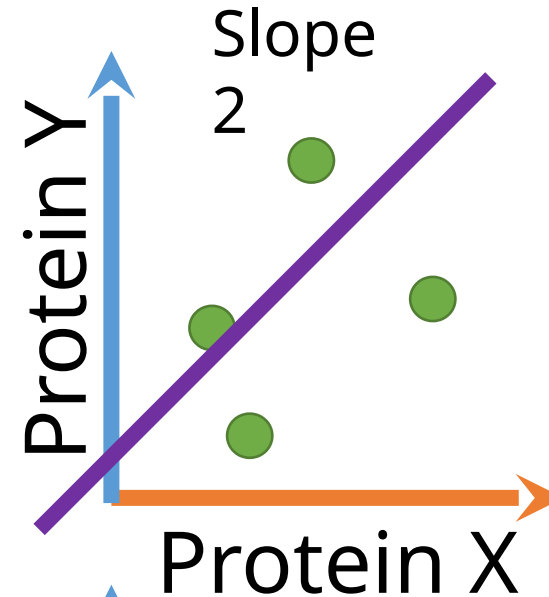
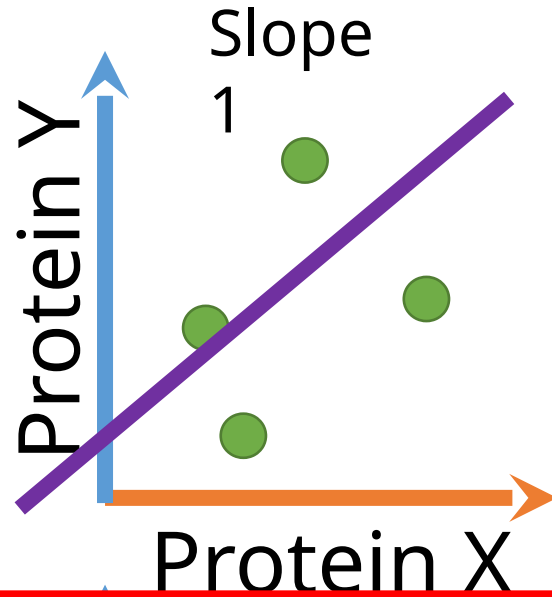


Y-intercept
2

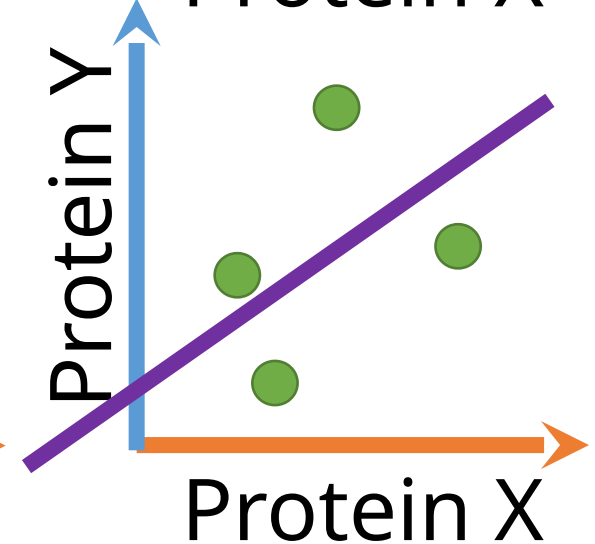
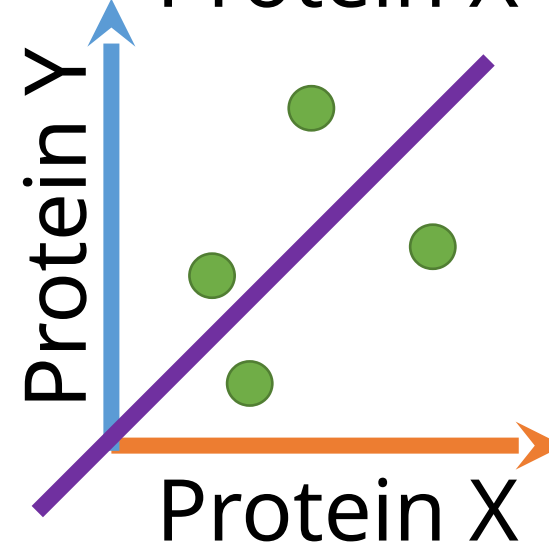
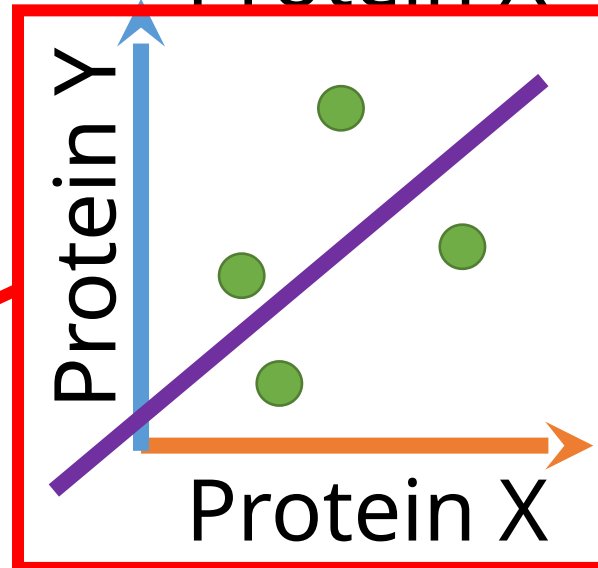


Least Squares

Y-intercept
1



Y-intercept
2

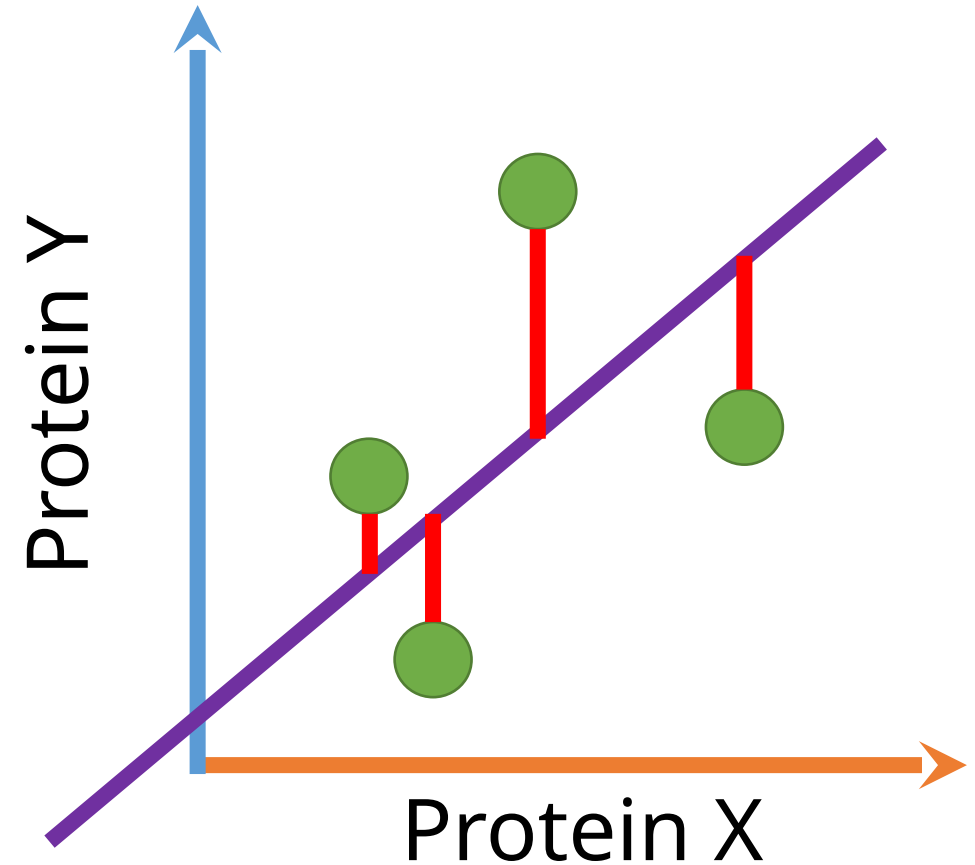


Transferred to the
next slide in
details

Least Squares

(2) Project each point to the line.

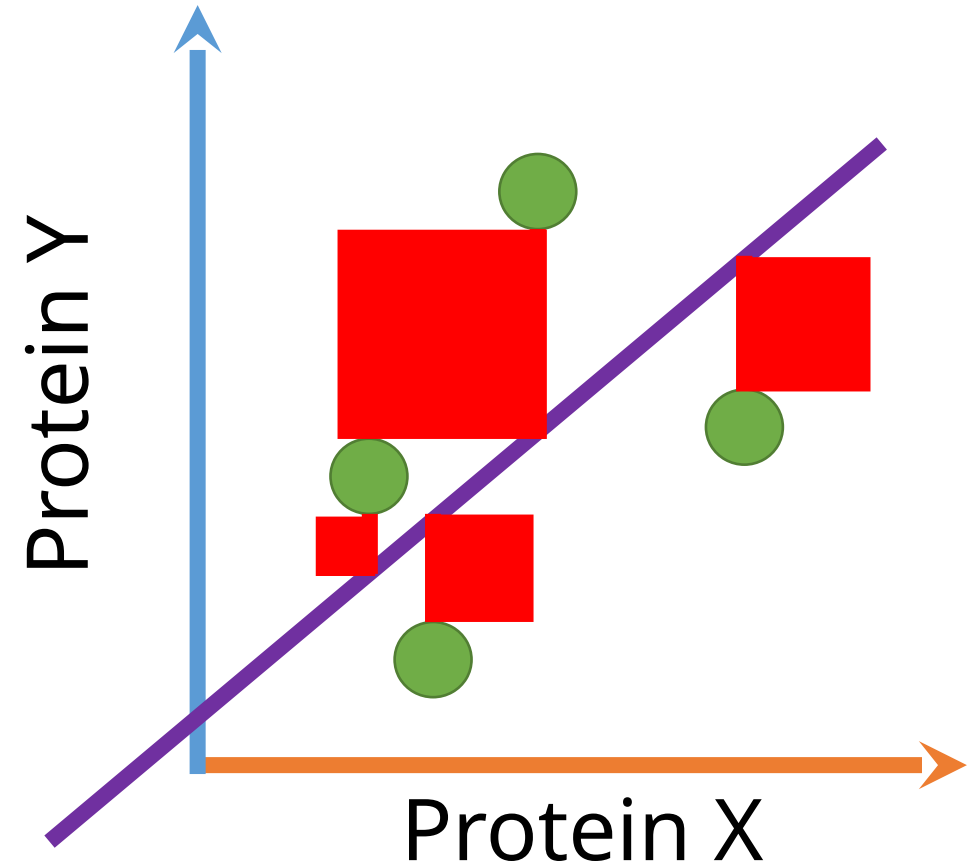
The red line is called “residual”.



Least Squares

(3) Square the residual length.

(4) Sum up the squares.

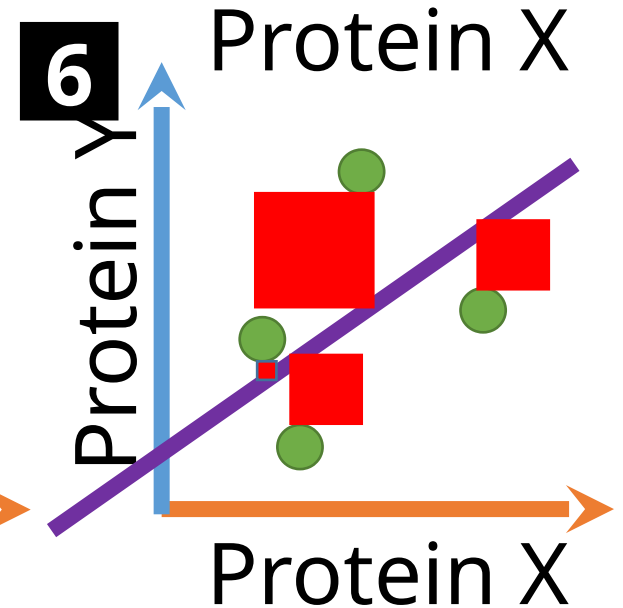
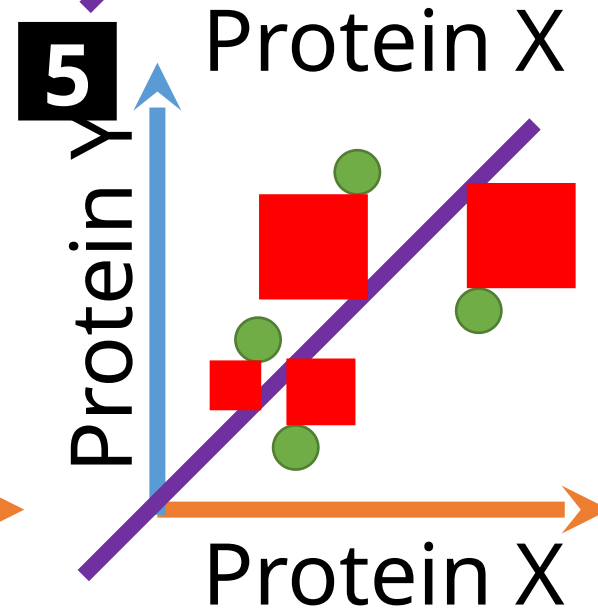
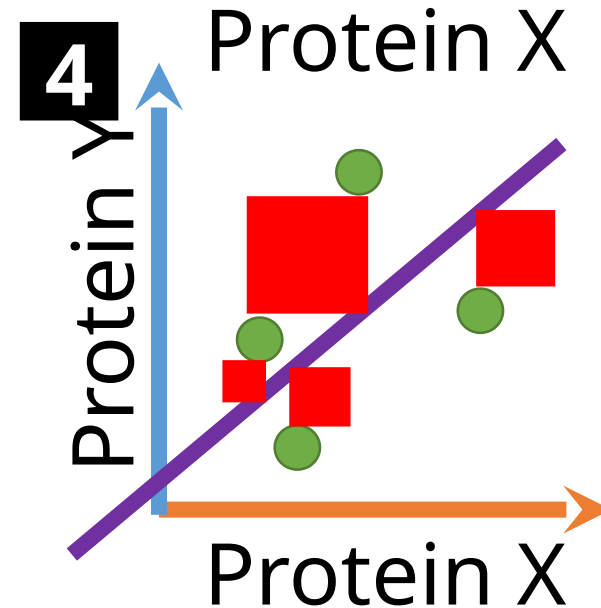
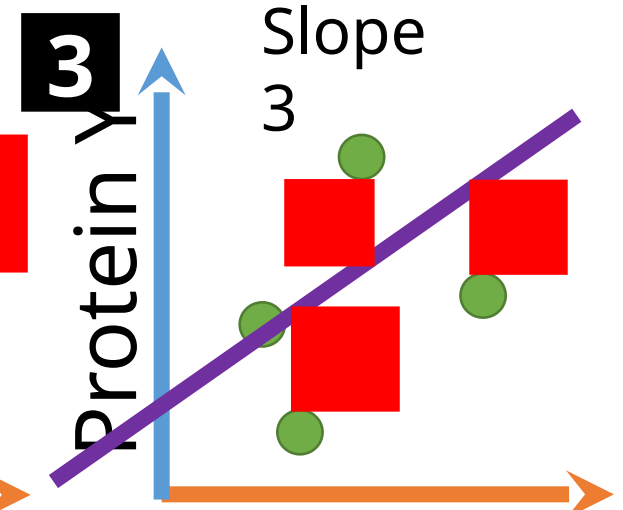
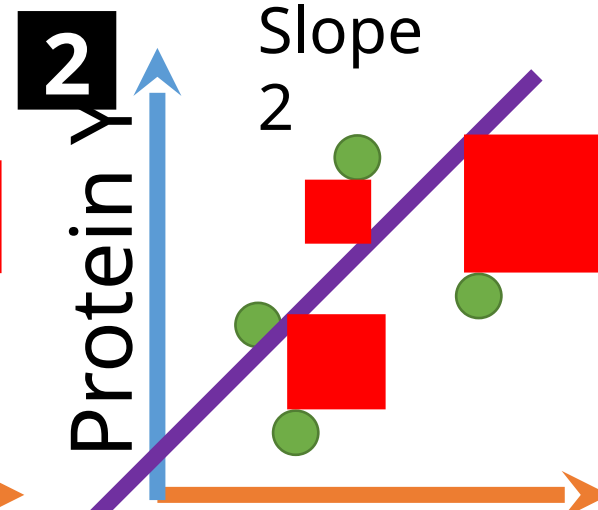
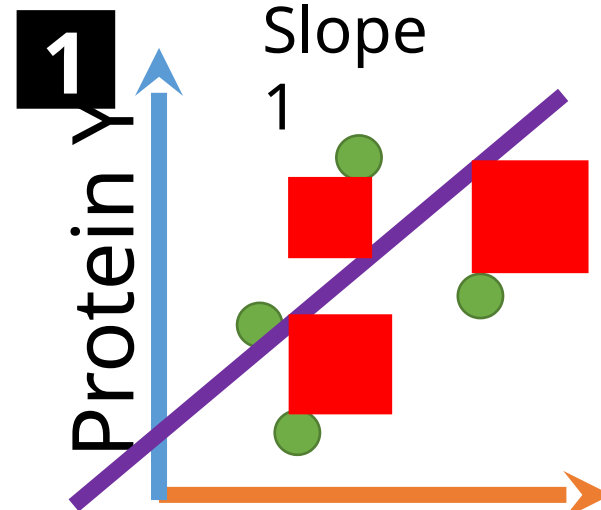


Least Squares

The minimum sum of squares the best fitted line.

Y-intercept
1

Y-intercept
2

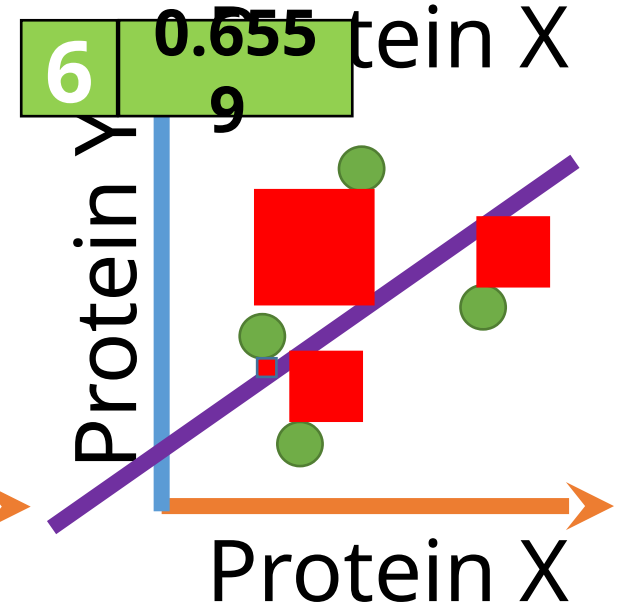
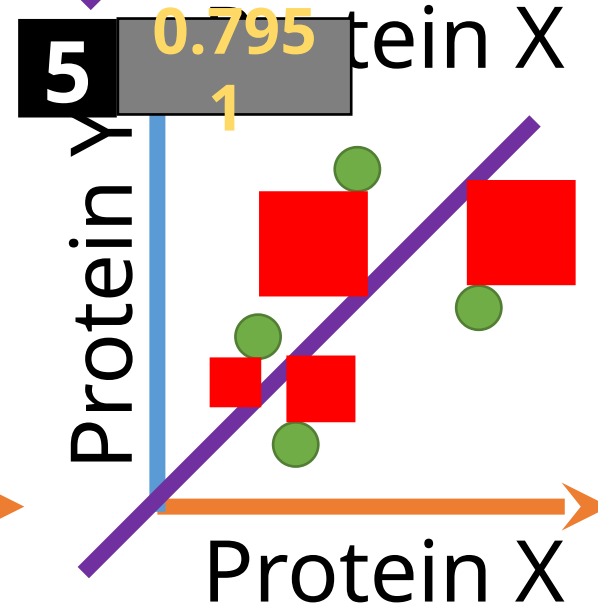
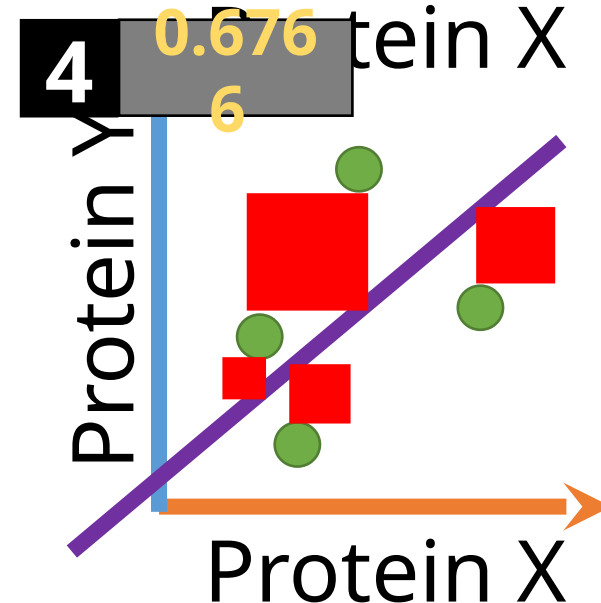
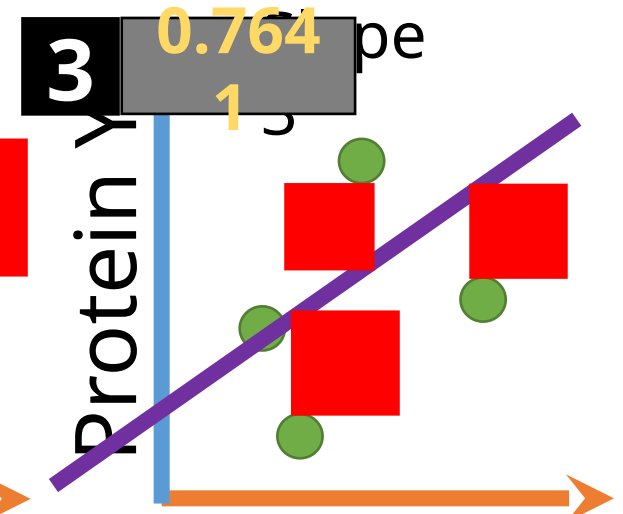
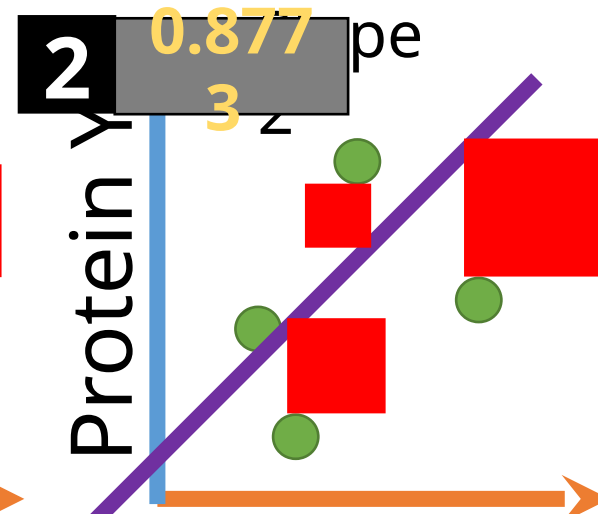
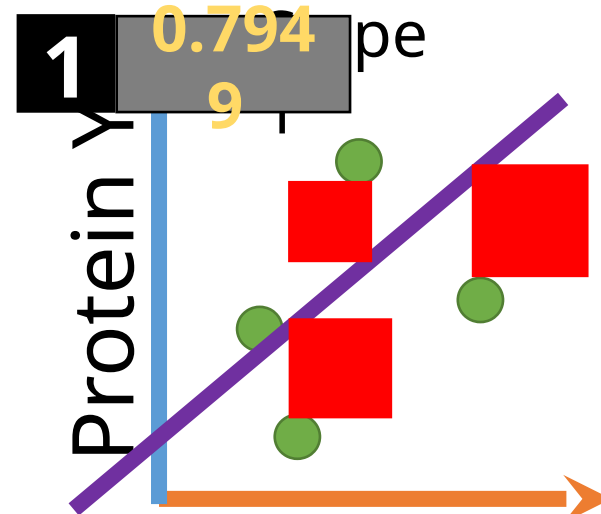


Least Squares

Y-intercept
1

The minimum
sum of
squares the
best fitted
line.

Y-intercept
2

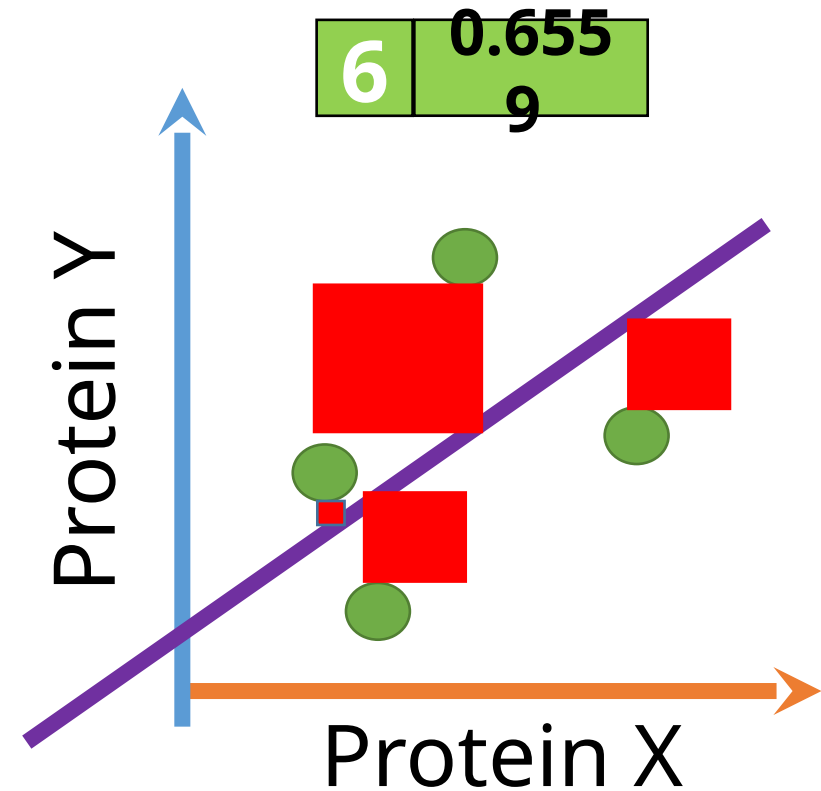


Least Squares

- Now ***a*** is known.
- Now ***b*** is known.

$$Y = aX + b$$

So, through the previous equation, the expression of protein ***Y*** can be predicted when the expression of protein ***X*** changes.



Least Squares

Mathematical example

- $(x_1, y_1) = (2, 3.2)$
- $(x_2, y_2) = (2.8, 1.2)$
- $(x_3, y_3) = (4, 6.4)$
- $(x_4, y_4) = (6.6, 3.8)$

Extracted
Information:

$$n = 4 ; i = [1, 2, 3, 4]$$

$$\sum x_i = 15.4$$

$$\sum y_i = 13.5$$

$$\sum x_i^2 = 71.4$$

$$\sum x_i y_i = 60.44$$

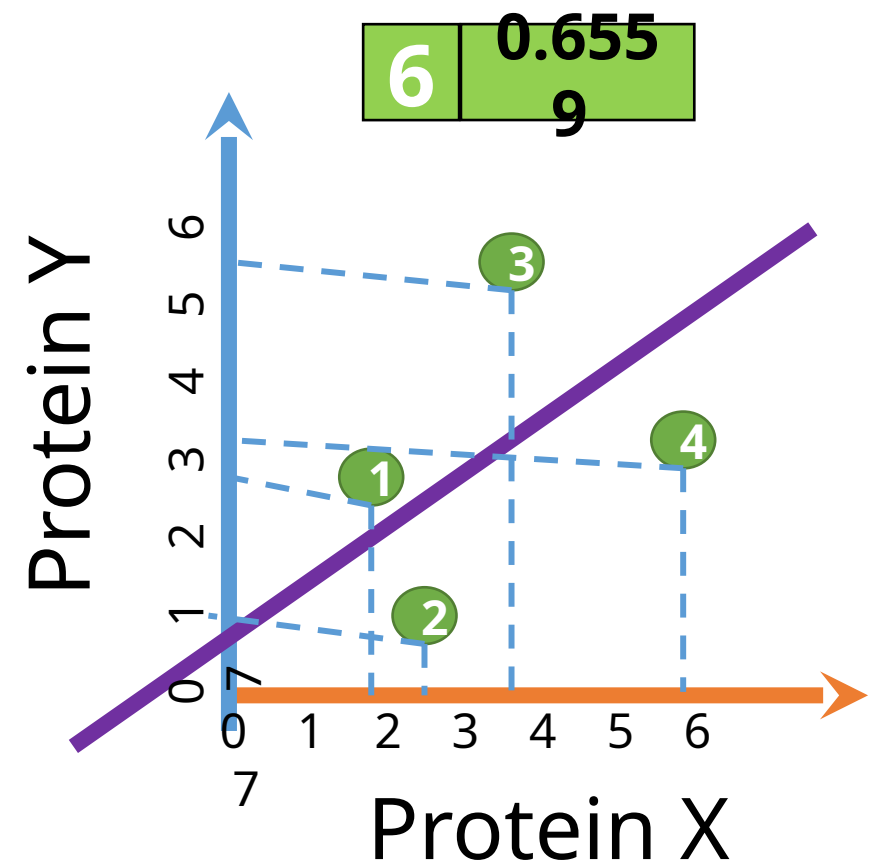
$$\Rightarrow \text{Eq1: } \mathbf{a} \sum x_i^2 + \mathbf{b} \sum x_i = \sum x_i y_i$$

$$\text{Eq2: } \mathbf{a} \sum x_i + \mathbf{b} n = \sum y_i$$

Simultaneous Equations Solver
<https://www.wolframalpha.com/>

$$\Rightarrow \begin{aligned} \mathbf{a} &= 1693/2422 = 0.69 \\ \mathbf{b} &= 1183/1730 = 0.68 \end{aligned}$$

a = Slope
b = Y-



Why is studying the relationship between features (genes, transcripts, proteins, metabolites, SNPs , or ...) important?

- Identify biomarkers: Molecules that can indicate the presence or severity of a disease.
- Develop new drugs: By targeting specific proteins or metabolic pathways involved in disease development.
- Understand cellular process: By studying how gene expression, protein and metabolite abundance change in response to different stimuli.
- Personalized medicine: By tailoring treatments to an individual's specific genetic makeup and molecular profile.

Revision (what did we learn?)

- The linear relationship between two measured variables.
- The types of linear relationships (positive, negative, and none).
- The parameters to fit a representative line to samples (slope and Y-intercept).
- The least squares to fit a line correctly.

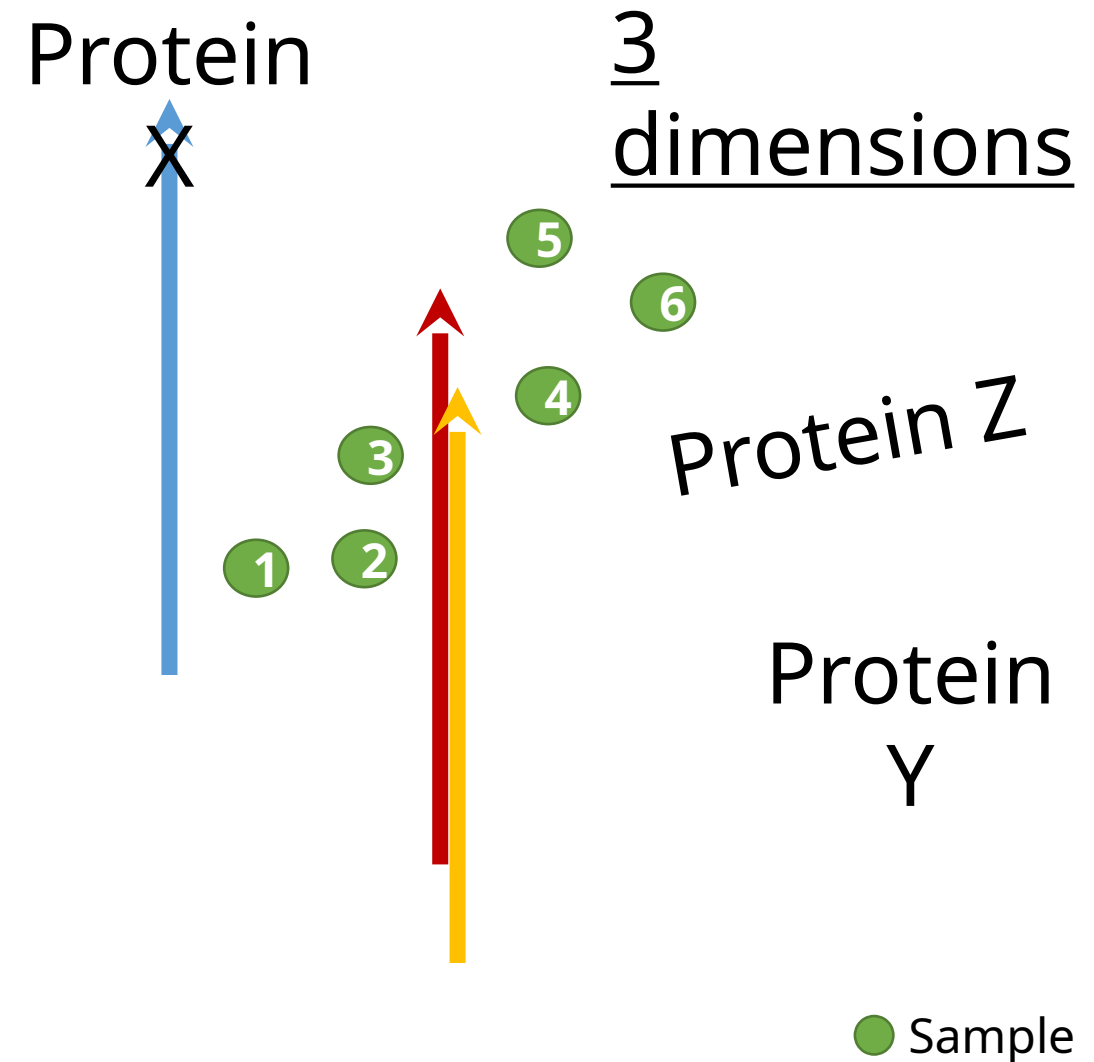
But !

What if we have 3 or more variables ???

Principal Component Analysis (PCA)

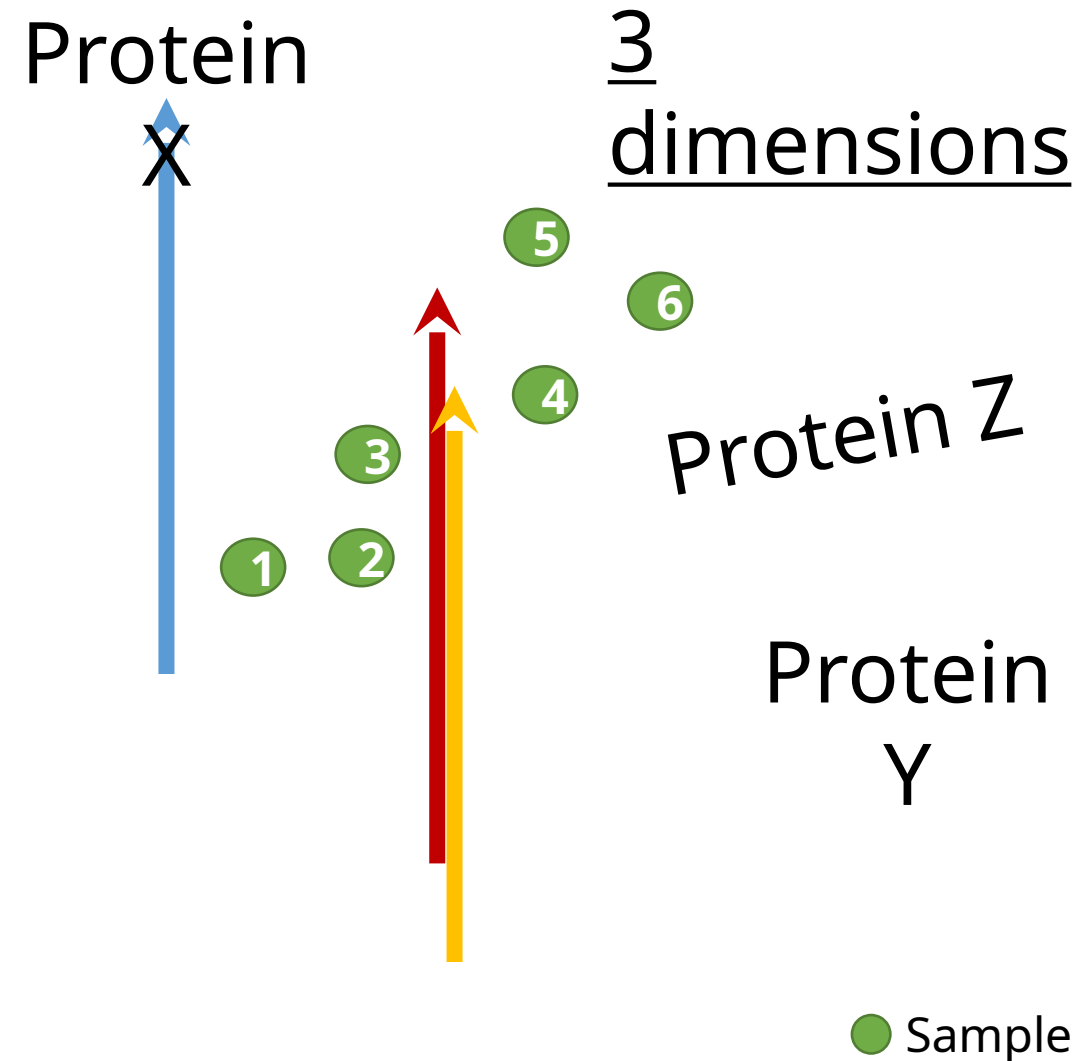
Principal Component Analysis (PCA)

(1) Plot the samples.



Principal Component Analysis (PCA)

Here we will not fit a line now
but first we will
(2) Center the data by:

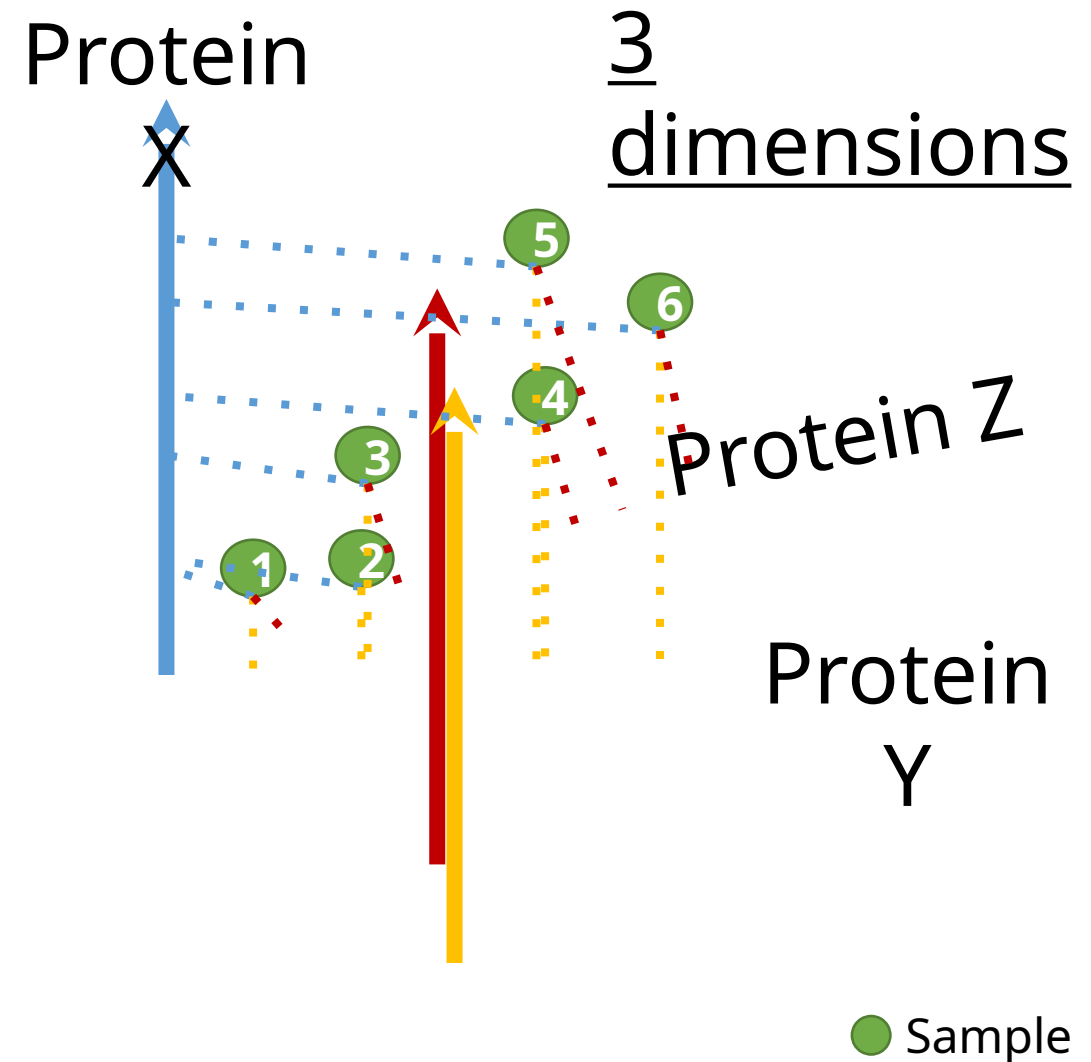


Principal Component Analysis (PCA)

Here we will not fit a line now
but first we will

(2) Center the data by:

(a) Projecting each
sample to all available
dimensions.



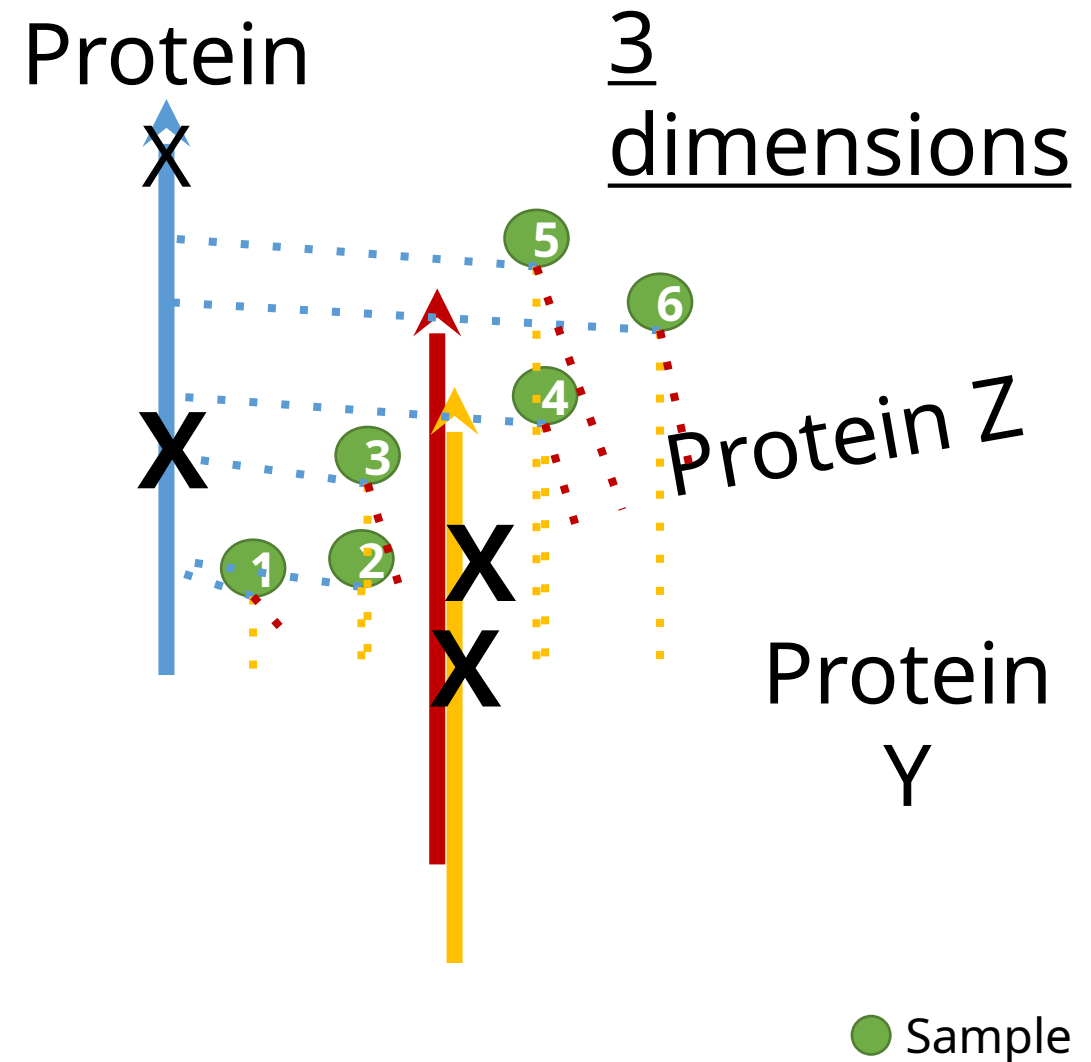
Principal Component Analysis (PCA)

Here we will not fit a line now
but first we will

(2) Center the data by:

(a) Projecting each
sample to all available
dimensions.

(b) Calculating the mean
value for each dimension.



Principal Component Analysis (PCA)

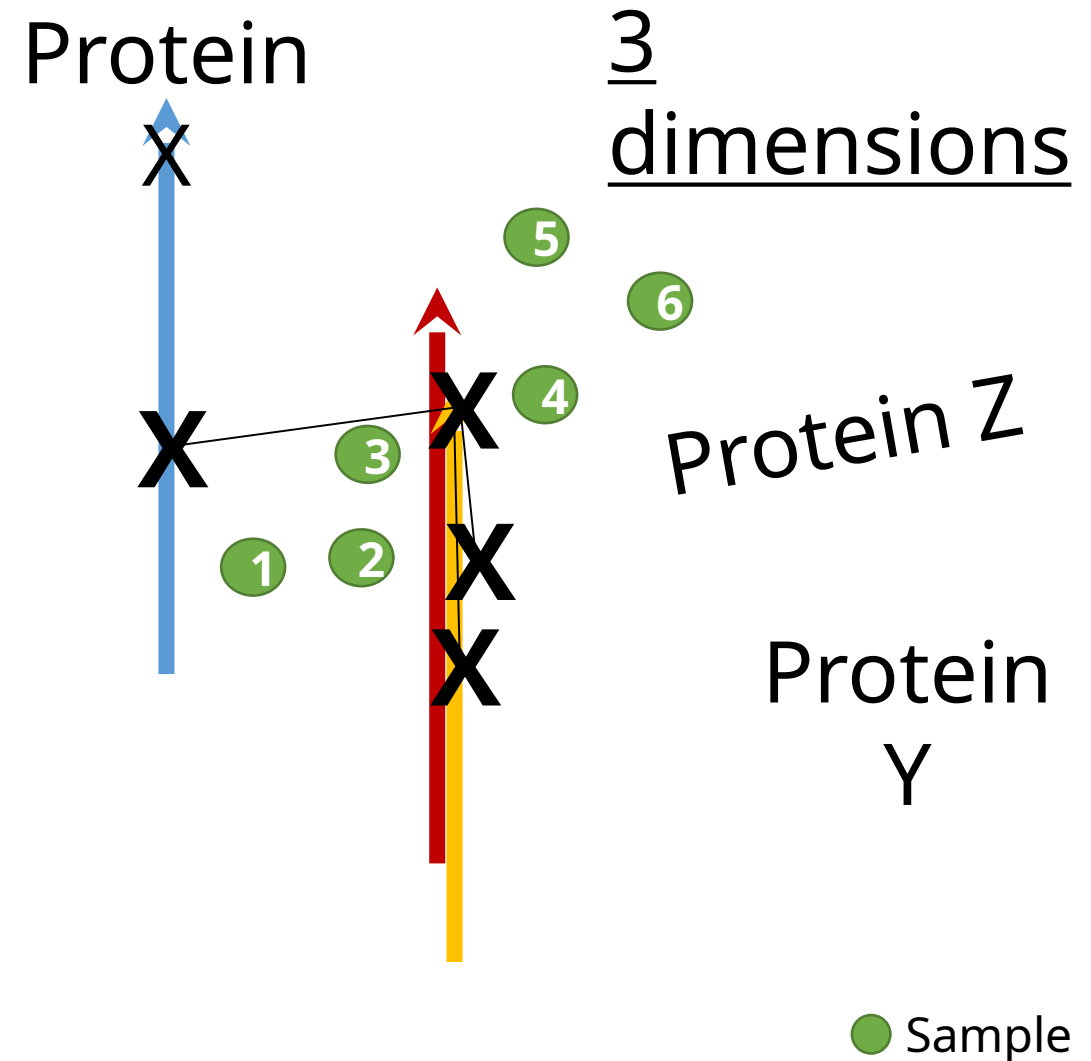
Here we will not fit a line now but first we will

(2) Center the data by:

(a) Projecting each sample to all available dimensions.

(b) Calculating the mean value for each dimension.

(c) Plot the means in the graph.



Principal Component Analysis (PCA)

Here we will not fit a line now but first we will

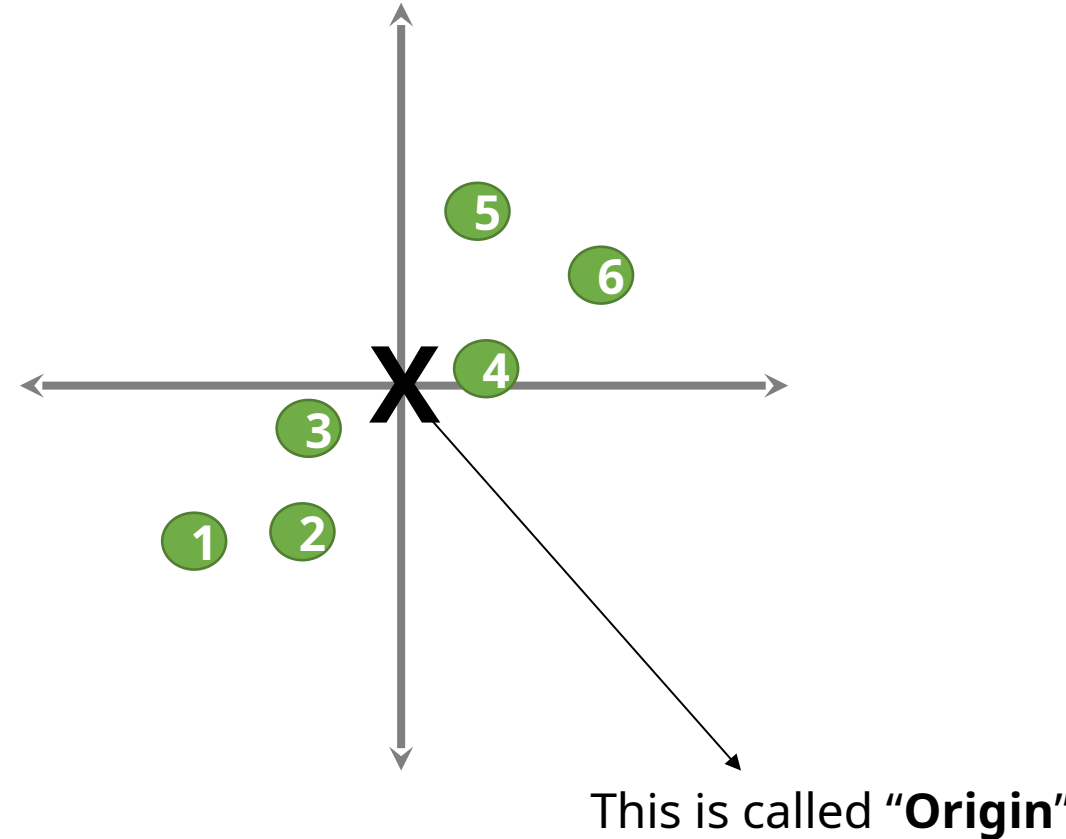
(2) Center the data by:

(a) Projecting each sample to all available dimensions.

(b) Calculating the mean value for each dimension.

(c) Plot the means in the graph.

(d) Create new X and Y axes from the plotted means.

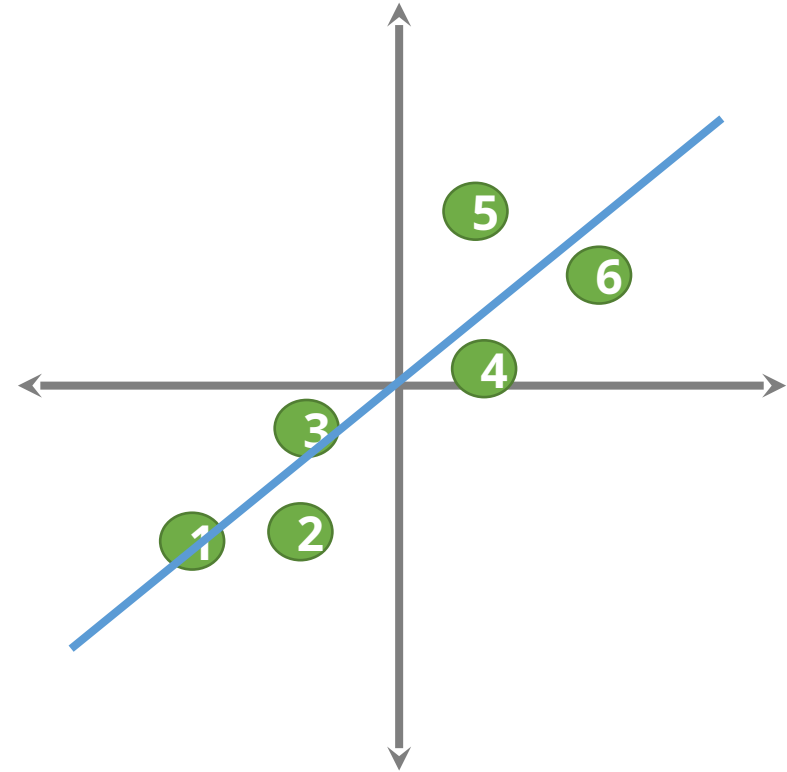


Principal Component Analysis (PCA)

(3) Calculate the least squares to fit a line.

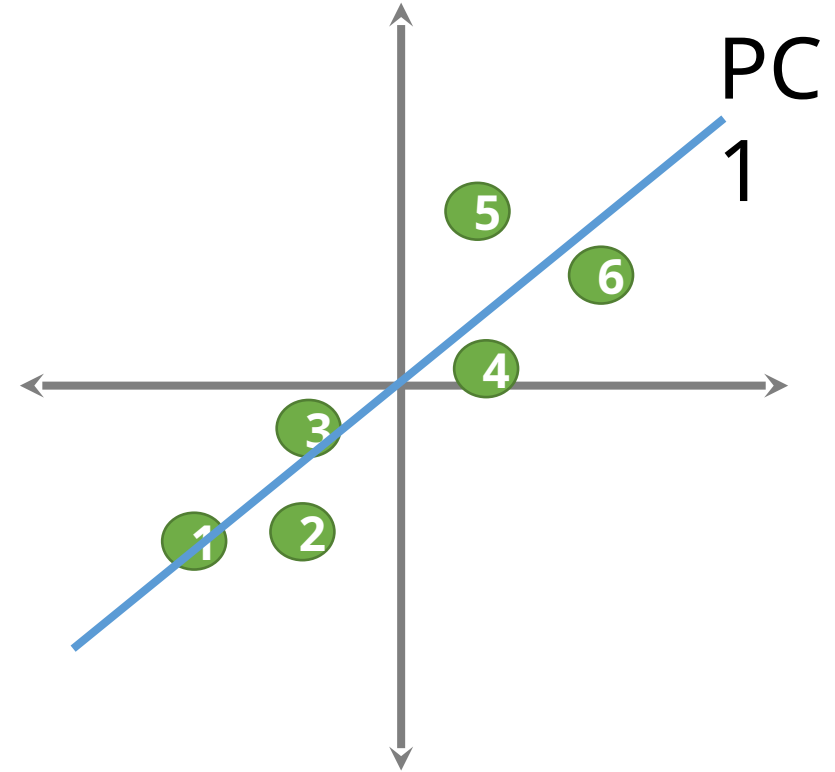
Note: the origin must be intercepted by the line.

There is no y-intercept.



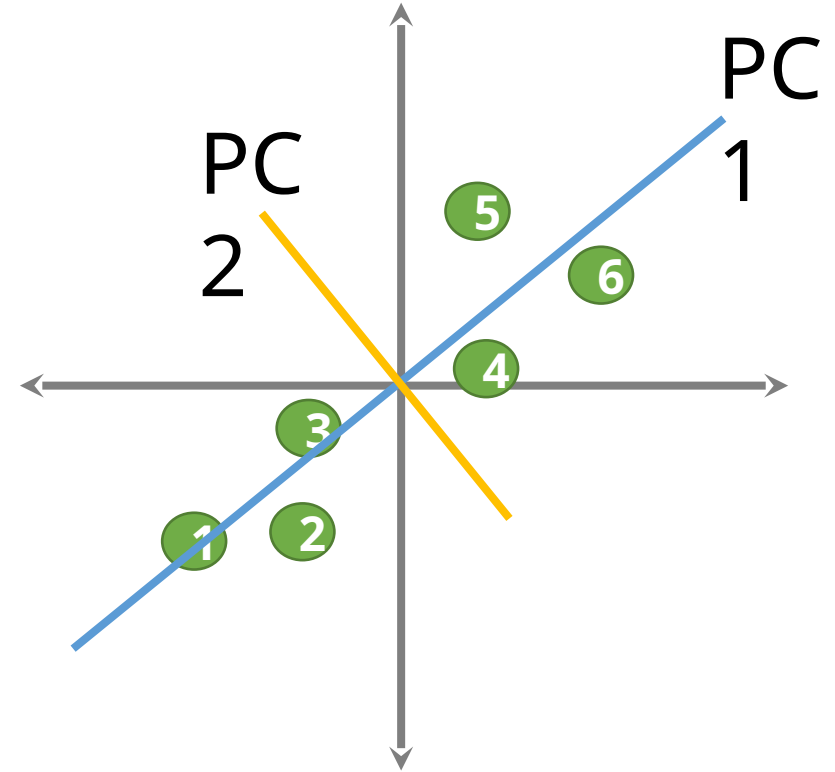
Principal Component Analysis (PCA)

The best fitted line is the Principal Component 1 (PC1).



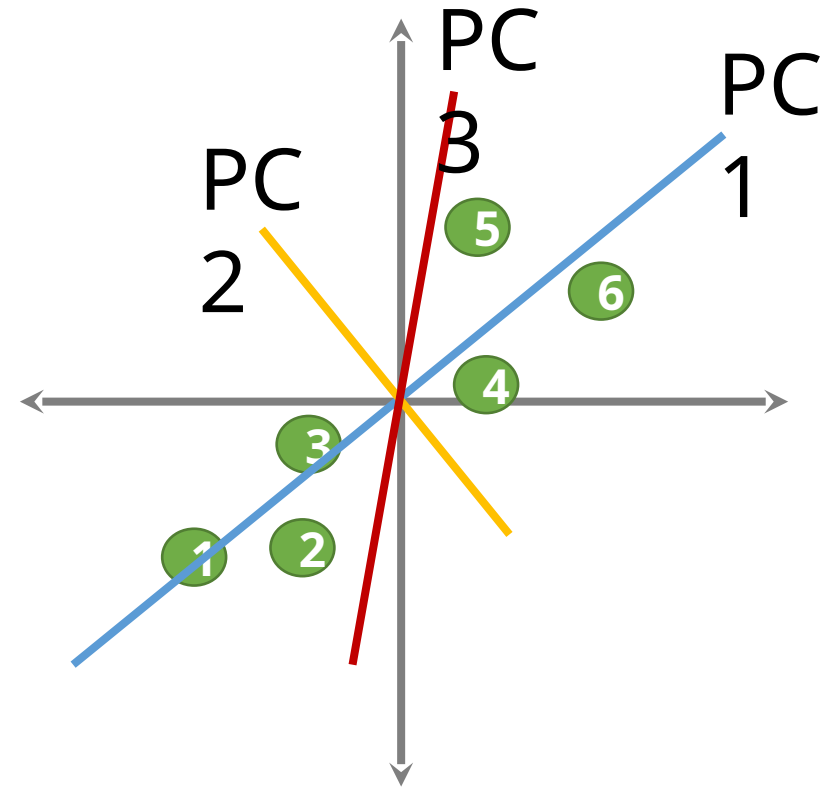
Principal Component Analysis (PCA)

The Principal Component 2 (PC2) will be perpendicular to PC1.



Principal Component Analysis (PCA)

The Principal Component 3 (PC3) will be perpendicular to PC1 & PC2.



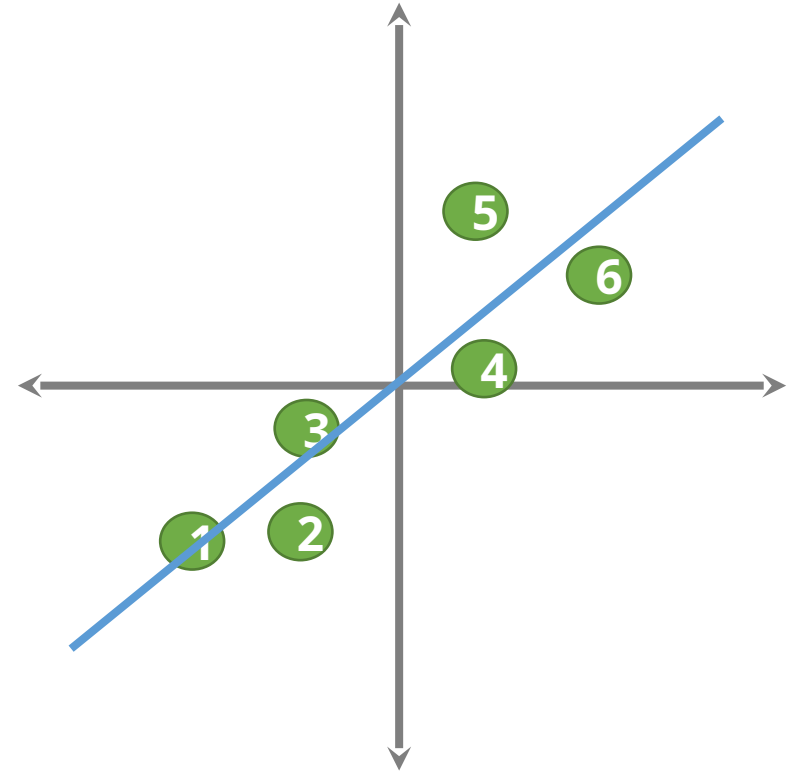
Principal Component Analysis (PCA)

Back again here

(3) Calculate the least squares to fit a line.

Note: the origin must be intercepted by the line.

There is no y-intercept.

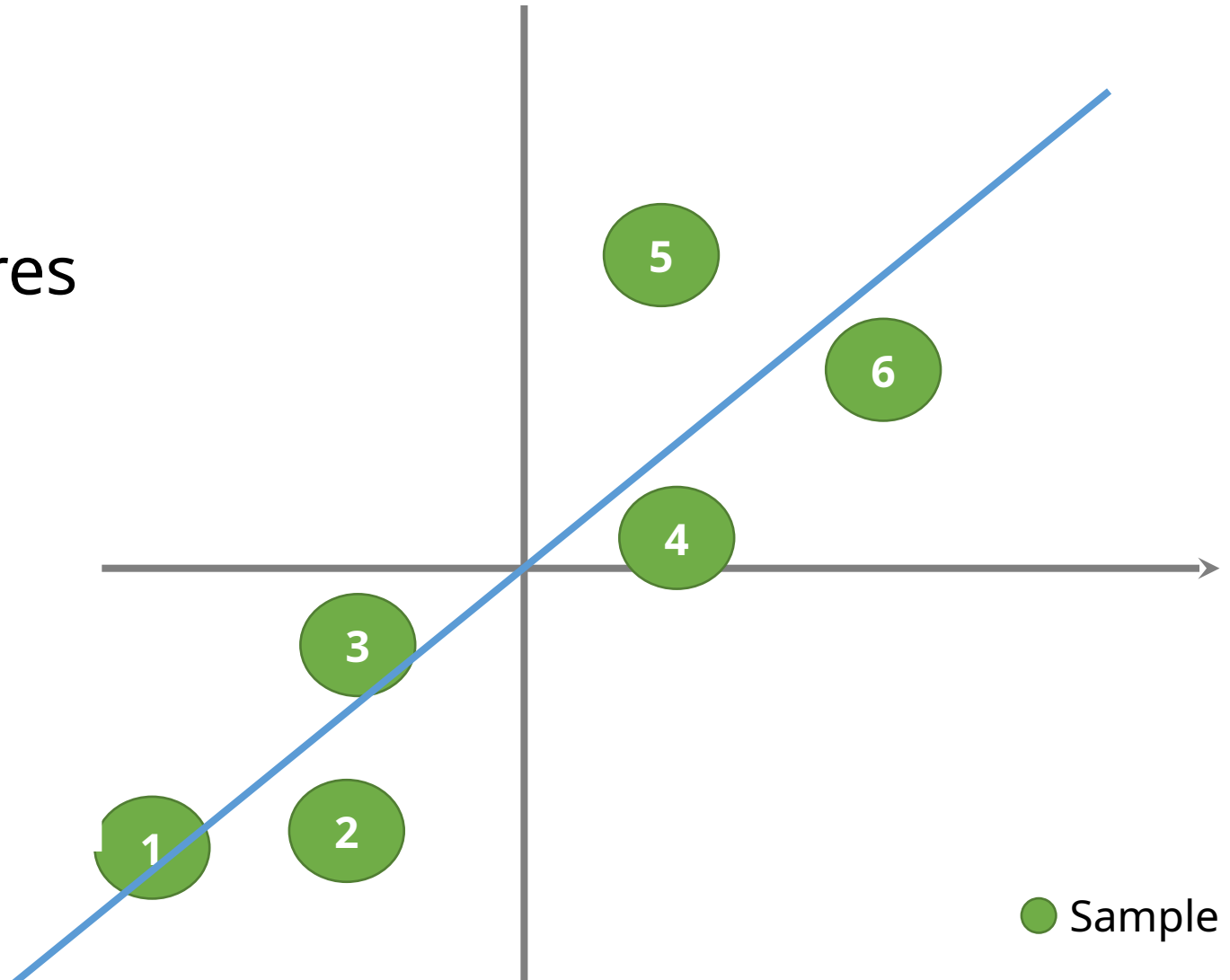


Principal Component Analysis (PCA)

ZOOOOM

(3) Calculate the least squares to fit a line.

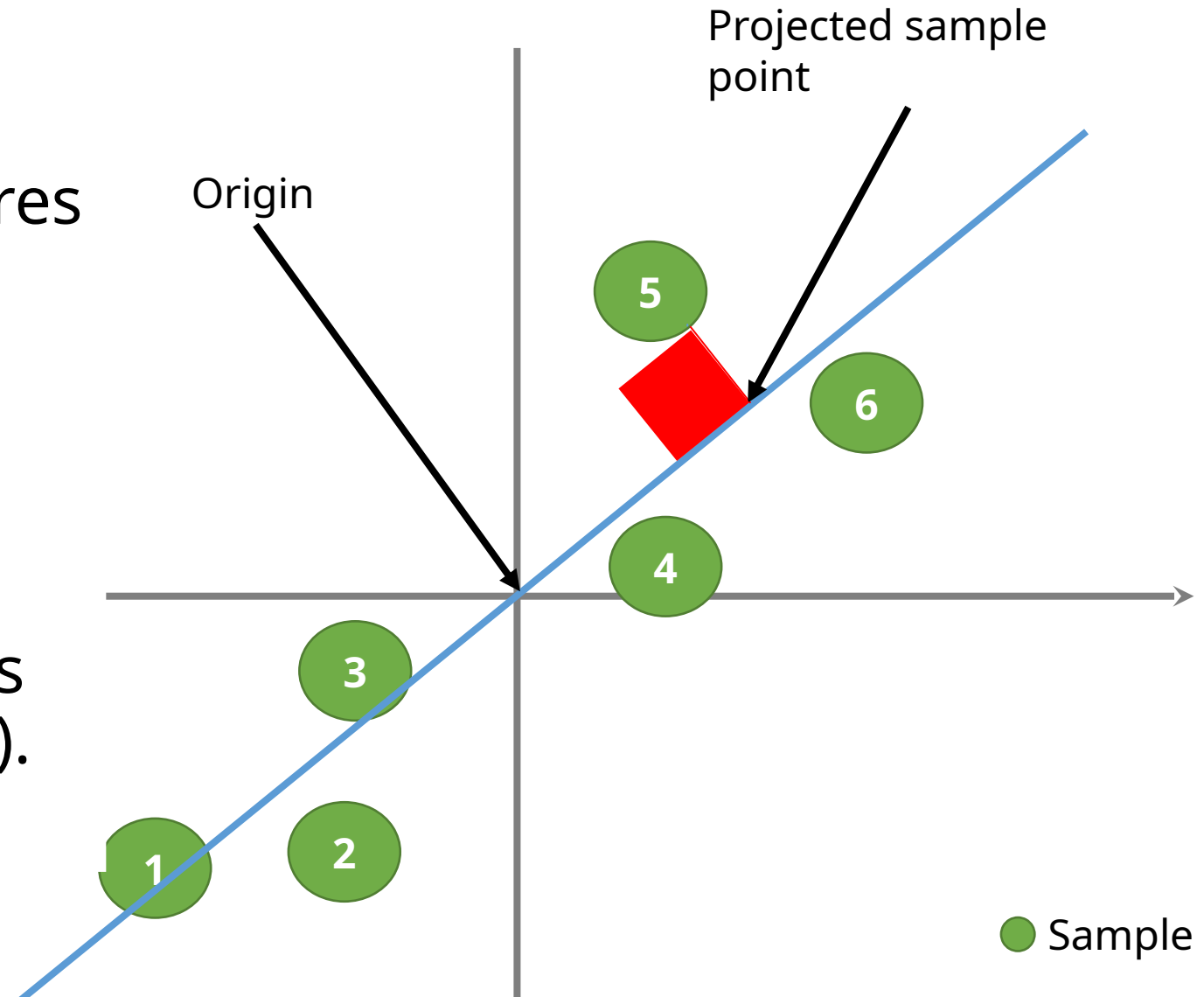
Note: the origin must be intercepted by the line.
There is no y-intercept.



Principal Component Analysis (PCA)

(3) Calculate the least squares to fit a line.

We minimize the square distance (red square) between the line and the sample (remember that this distance is called “residual”).

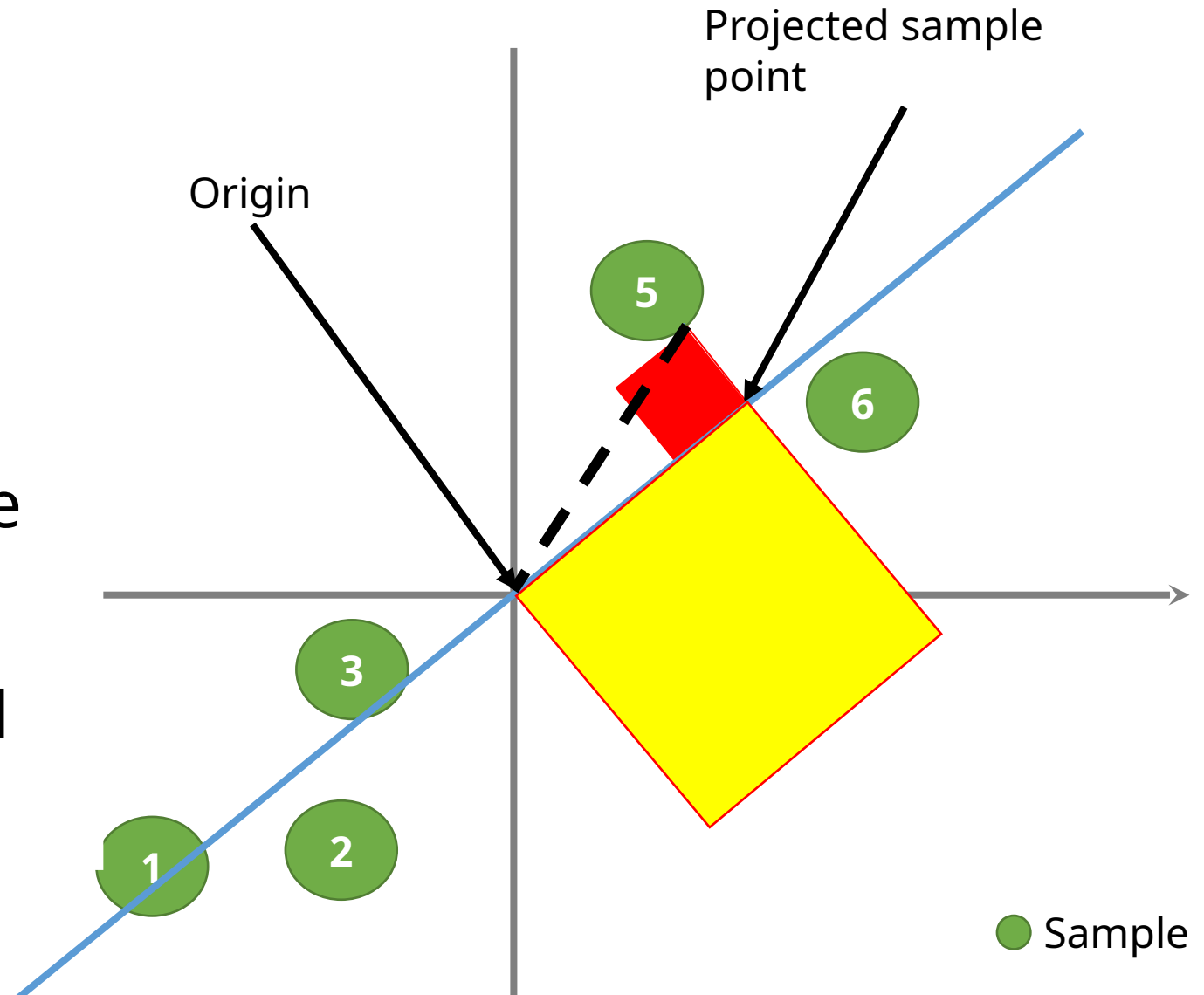


Principal Component Analysis (PCA)

Another way to fit a line.

(3) Calculate the largest squares to fit a line.

In other word, we maximize the square distance (yellow square) between the projected sample point and the origin.

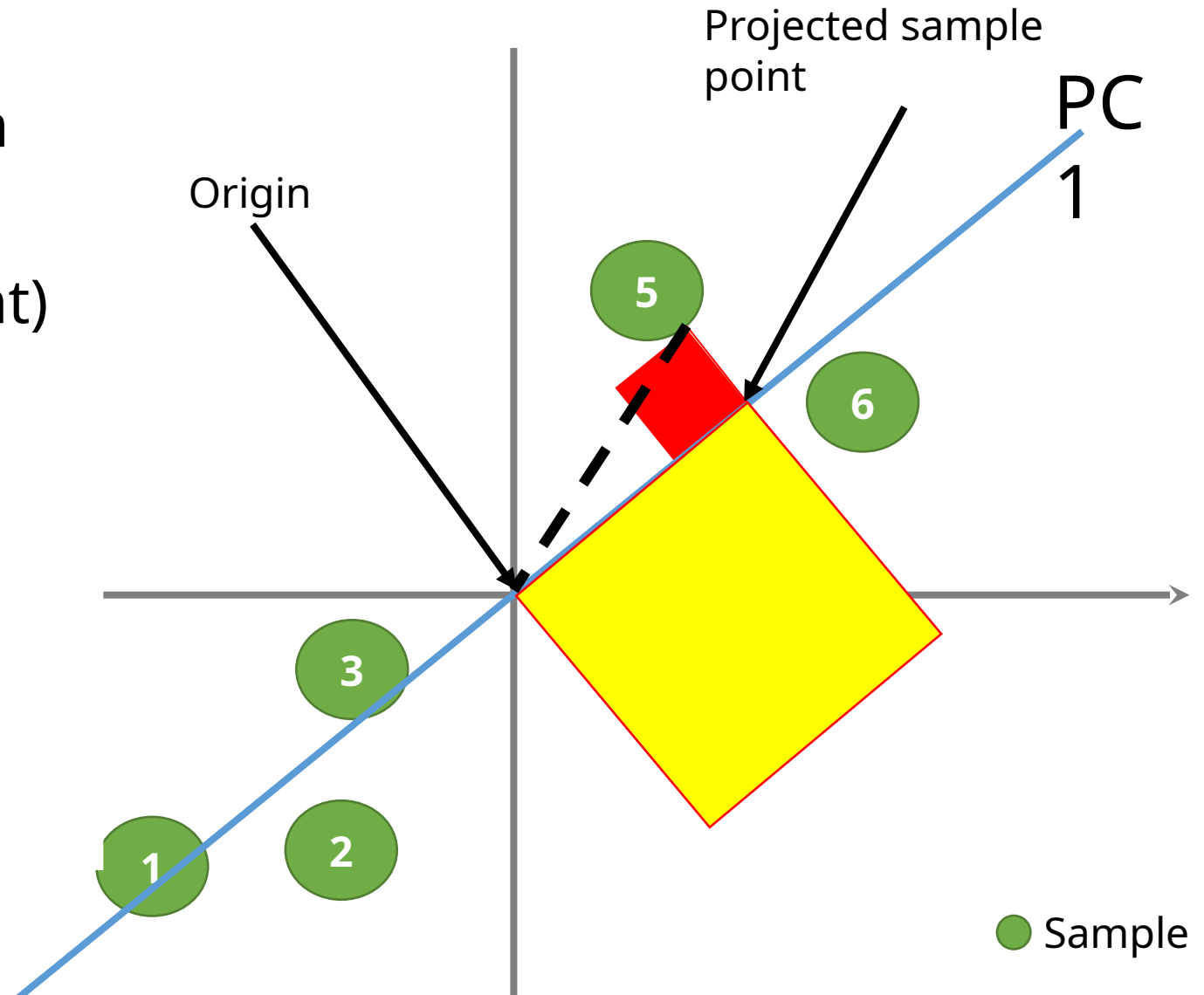


Principal Component Analysis (PCA)

The best fitted line (minimum sum of red squares OR maximum sum of yellow squares for each sample point) is the Principal Component 1 (PC1).

Obtain the slope of PC1.

We haven't finished explaining PCA yet.



Revision (what did we learn?)

- Reducing dimensions or variables through data centering.
- The 3D-graph (3 proteins) has been converted to 2D-graph.

But !

What happened in the PCA at the level of variables? In other words, where did the proteins go?

And, what does component mean?

Factor Analysis

Factor analysis

Factor analysis is a **multivariate method** that can be used for analyzing large data sets

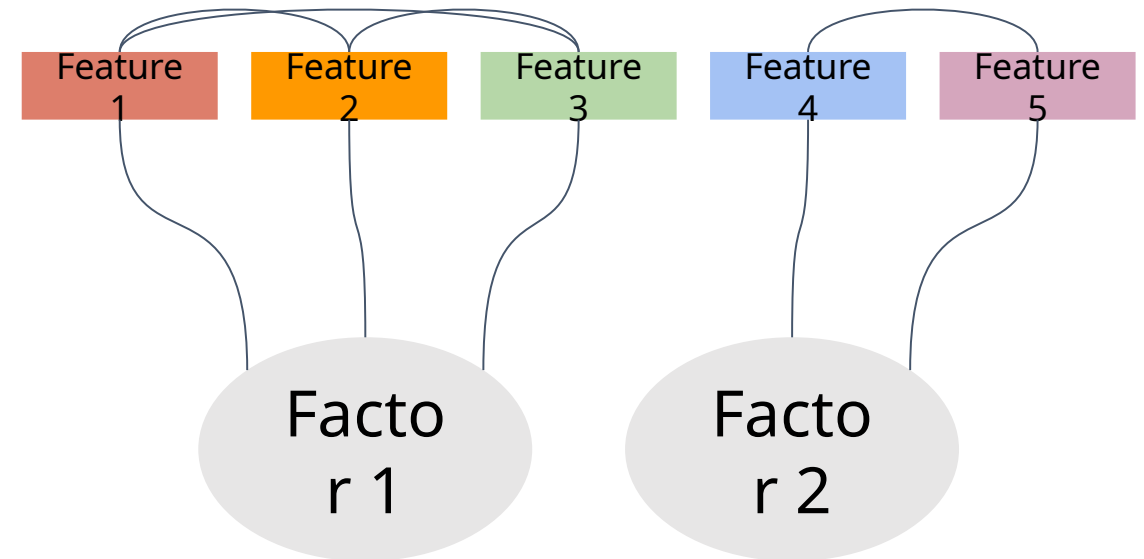
With two main goals:

to **reduce** a (1) large number of (2) correlating features to a fewer number of factors,

to **structure** the data with the aim of **identifying dependencies** between correlating features and examining them for common causes (factors) in order to generate a new construct (factor) on this basis.

Analysis conditions

(1) large number of (2) correlating features



Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores

1. Calculating the correlation matrix

| | Feature 1 | Feature 2 | Feature 3 | Feature 4 | Feature 5 |
|-----------|--------------|--------------|-----------|--------------|-----------|
| Feature 1 | 1 | | | | |
| Feature 2 | 0.712 | 1 | | | |
| Feature 3 | 0.961 | 0.704 | 1 | | |
| Feature 4 | 0.109 | 0.138 | 0.078 | 1 | |
| Feature 5 | 0.044 | 0.067 | 0.024 | 0.983 | 1 |

Note: Standardization (scaling) just simplifies the calculation of the correlation matrix.

2. Assessing the suitability of the data (e.g. Kaiser–Meyer–Olkin criterion; Recommended KMO value ≥ 0.8)

Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

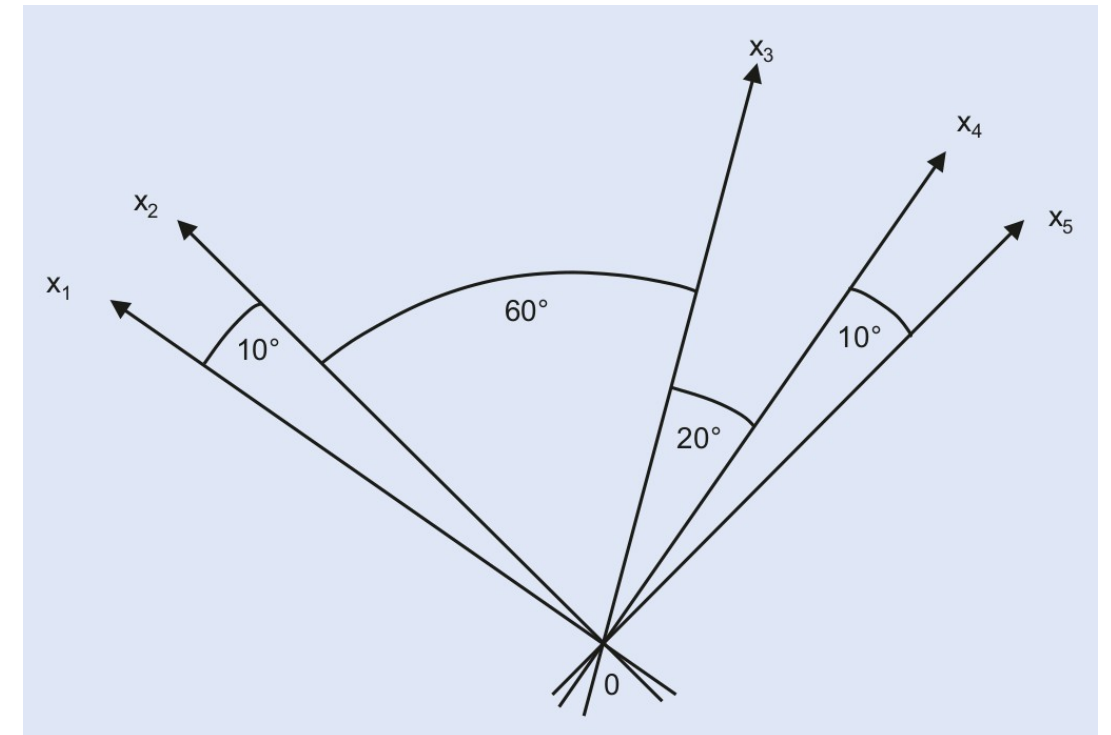
Step 4: Determining the factor scores

Inverse Cosine

$$\mathbf{R} = \begin{pmatrix} 1 & & \\ 0.8660 & 1 & \\ 0.1736 & 0.6428 & 1 \end{pmatrix} \text{ which is equal to } \mathbf{R} = \begin{pmatrix} 0^\circ & & \\ 30^\circ & 0^\circ & \\ 80^\circ & 50^\circ & 0^\circ \end{pmatrix} \blacktriangleleft$$

1. Converting correlation to angle (\cos^{-1})

The smaller the angle, the higher the correlation between two features.



The center of gravity determines the first factor.

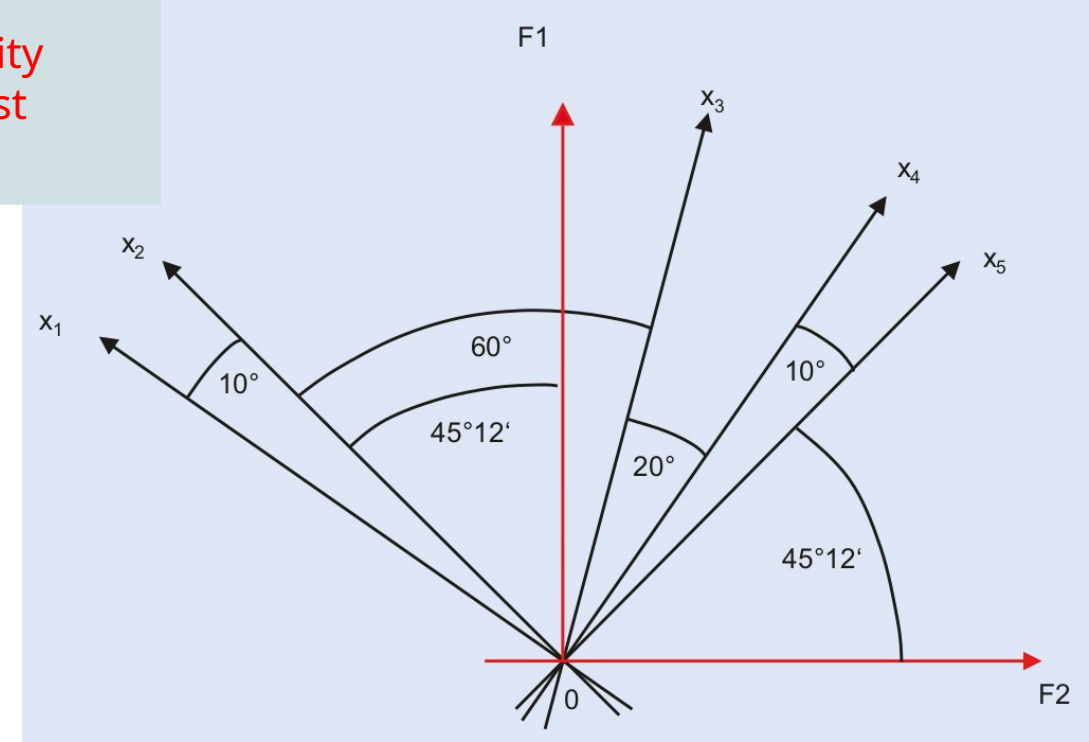
Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores



1. Extracting Factors (Graphical Method)

Factor analysis searches for factors that are independent (uncorrelated), a second factor should be orthogonal to the first factor.

The center of gravity determines the first factor.

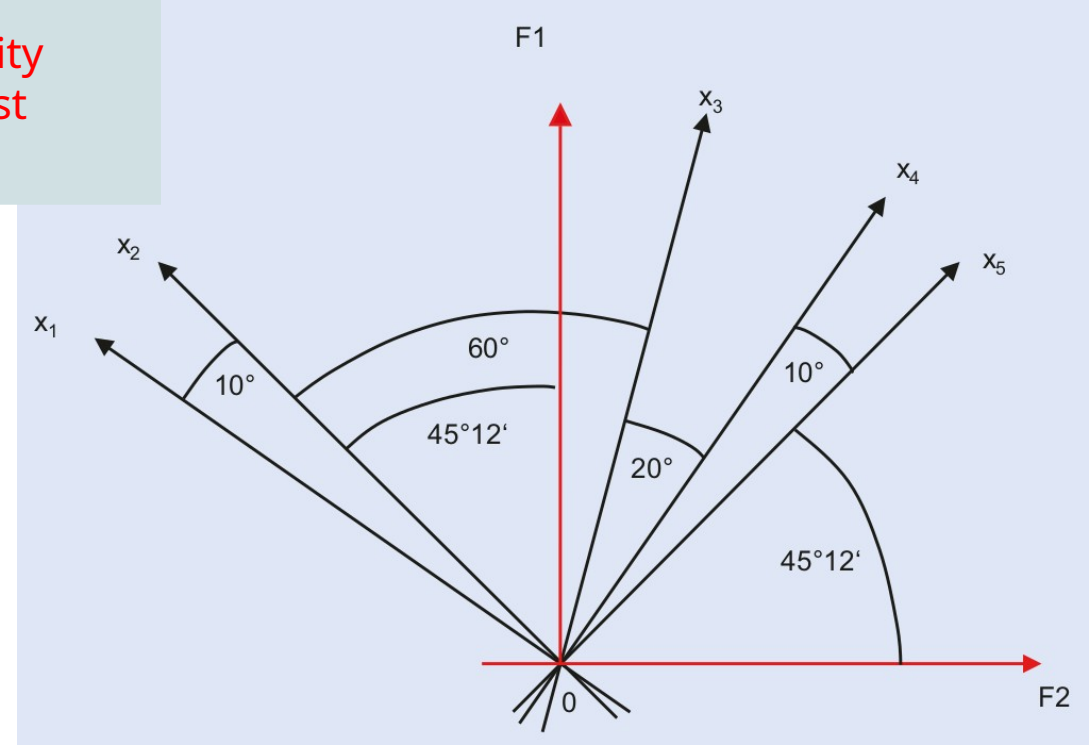
Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores



1. Extracting Factors (Graphical Method)

The angle between the first factor and x_1 equals $55^{\circ}12'$ ($= 45^{\circ}12' + 10^{\circ}$), which corresponds to a factor loading of 0.571.

The absolute value of a factor loading should be greater than 0.5 ($< 60^{\circ}$) to be relevant for a factor.

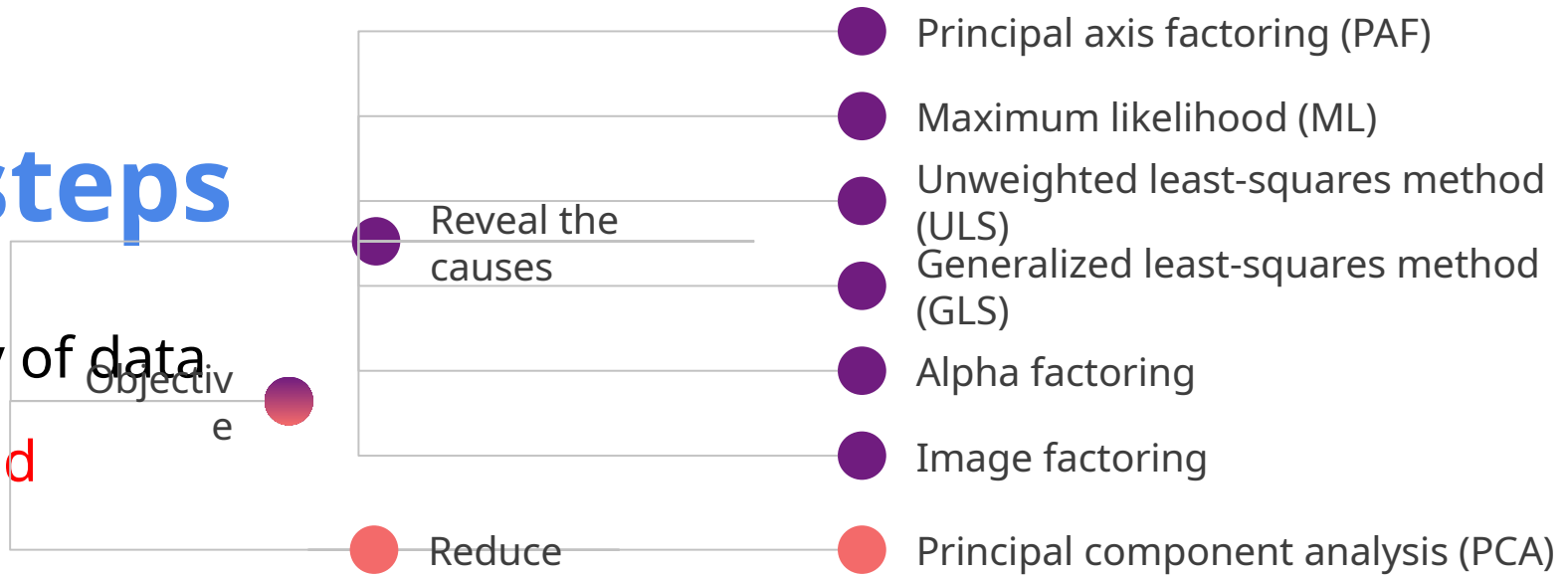
Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores



1. Extracting Factors (Mathematical Methods)

For **reveal the causes objective**, we use factor analysis (FA). The factors are interpreted as the causes of the observed features and their correlations.

For **reduce objective** we use principal component analysis (PCA). We look for a small number of factors (principal components) that preserve a maximum of the variance (information) contained in the features.

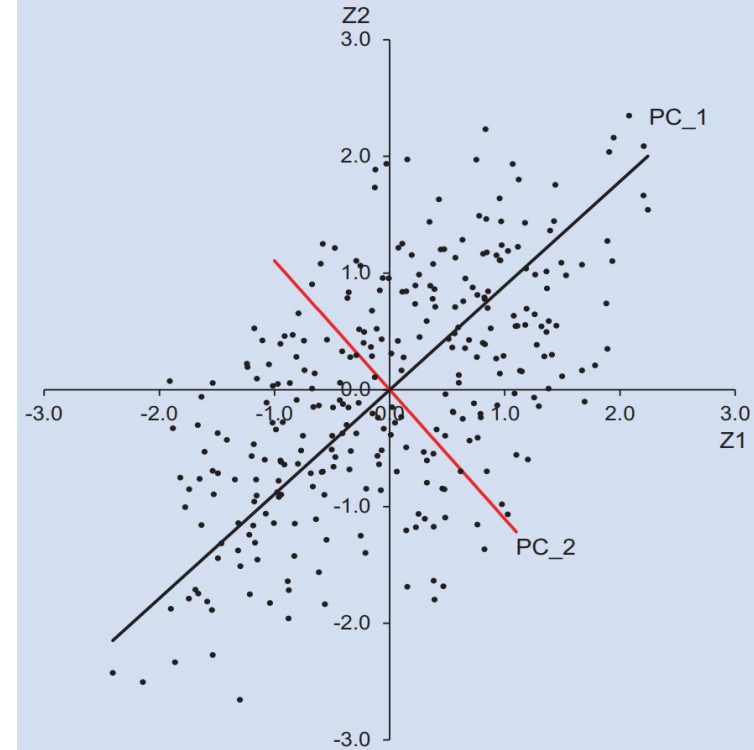
Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores



2. Extracting Factors (Mathematical Methods)

If a variable has a variance of zero, it does not contain any information. Otherwise, after standardization, each feature has a variance of 1.

Ex. **The variance of PC1 = $s^2 = 1.596$.**

Z1 (standardized feature) variance = 1; Z2 variance = 1; so, the **total variance of the data is 2**.

Variance explained by PC = $(s^2 / \text{total variance}) * 100$.

PC1 variance explained = $1.596 / 2 = 0.80 * 100 = 80\%$.

Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores

The Eigenvalue of the components

Square the loadings

| Component loadings | | | | | |
|--------------------|-------|--------|--------|--------|--------|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0.937 | -0.229 | -0.223 | -0.138 | 0.017 |
| 2 | 0.843 | -0.160 | 0.514 | 0.004 | 0.004 |
| 3 | 0.929 | -0.254 | -0.233 | 0.137 | -0.015 |
| 4 | 0.342 | 0.936 | -0.001 | -0.026 | -0.079 |
| 5 | 0.277 | 0.957 | -0.028 | 0.029 | 0.078 |

| | 1 | 2 | 3 | 4 | 5 | Communalities |
|---|-------|-------|-------|-------|-------|---------------|
| 1 | 0.879 | 0.053 | 0.050 | 0.019 | 0.000 | 1.000 |
| 2 | 0.710 | 0.026 | 0.264 | 0.000 | 0.000 | 1.000 |
| 3 | 0.862 | 0.064 | 0.054 | 0.019 | 0.000 | 1.000 |
| 4 | 0.117 | 0.876 | 0.000 | 0.001 | 0.006 | 1.000 |
| 5 | 0.077 | 0.915 | 0.001 | 0.001 | 0.006 | 1.000 |
| 6 | 2.645 | 1.934 | 0.369 | 0.039 | 0.013 | 5.000 |

2. Extracting Factors (Mathematical Methods)

The **communality** measures a feature's variance (information) that the components can explain.

The sum of loadings per feature must be equal to 1 (the variance).

The sum of the squared loadings over the features yields the **eigenvalue** of a component.

The **eigenvalue** can be interpreted as a measure of the information contained in a

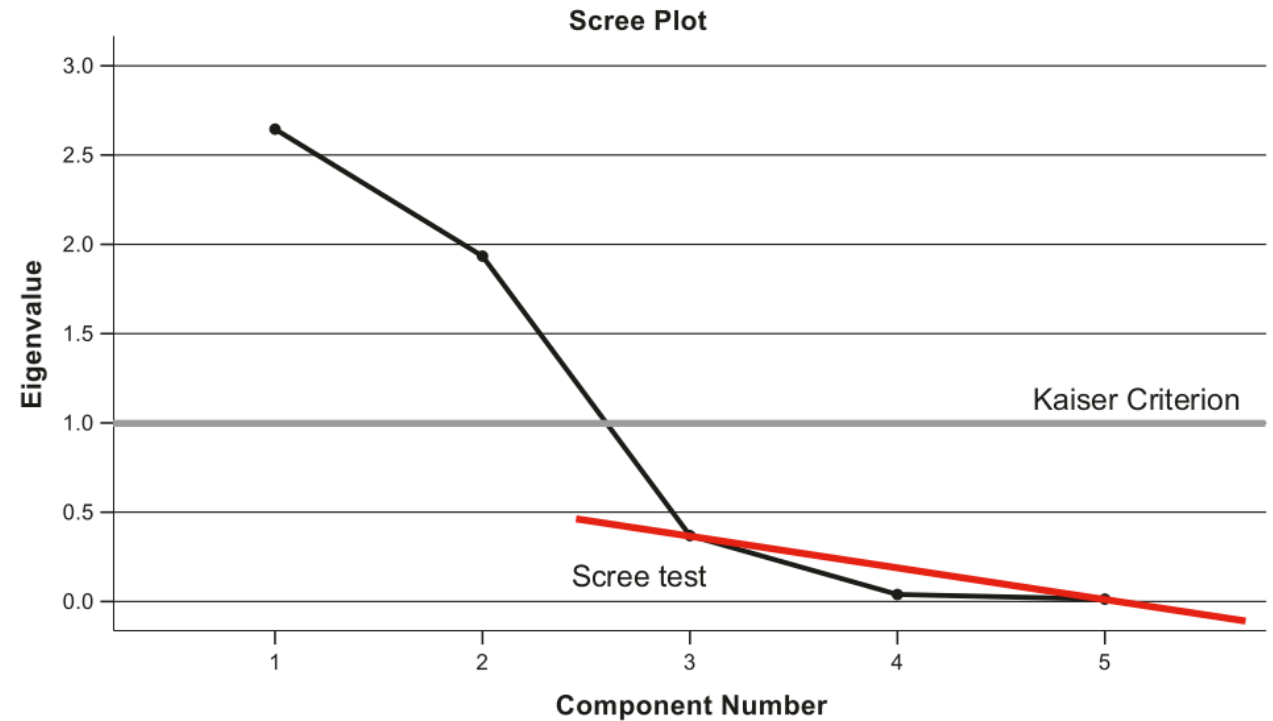
Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores



1. Number of Factors
 - a. Eigenvalue or Kaiser criterion
 - b. Scree test

Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

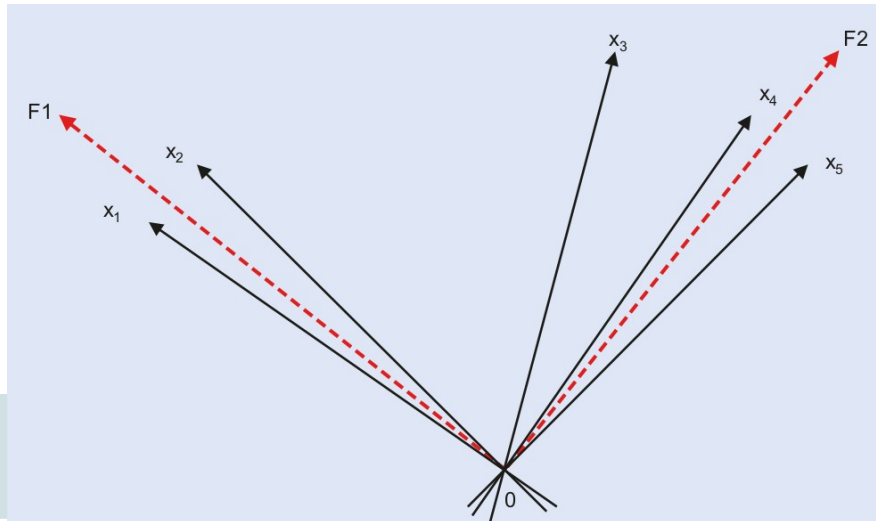
Step 3: Interpreting the factors

Step 4: Determining the factor scores

The interpretation of the factors requires a **high level of expertise and some creativity** on the part of the user.

Difficulties and solutions:

1. Cross-loadings.
High correlations on multiple factors.
Solution: select factor loading > 0.5 .
2. Ambiguous assignment of features to a factor.
Solution: Factor rotation.



Factor rotation

Factor analysis steps

Step 1: Evaluating the suitability of data

Step 2: Extracting the factors and determining their number

Step 3: Interpreting the factors

Step 4: Determining the factor scores

One of three approaches could be used:

1. Surrogates.
2. Summated scales.
3. Regression analysis.

All three approaches rely on the factor loadings but view these values differently.

Multiple Factor Analysis (MFA)



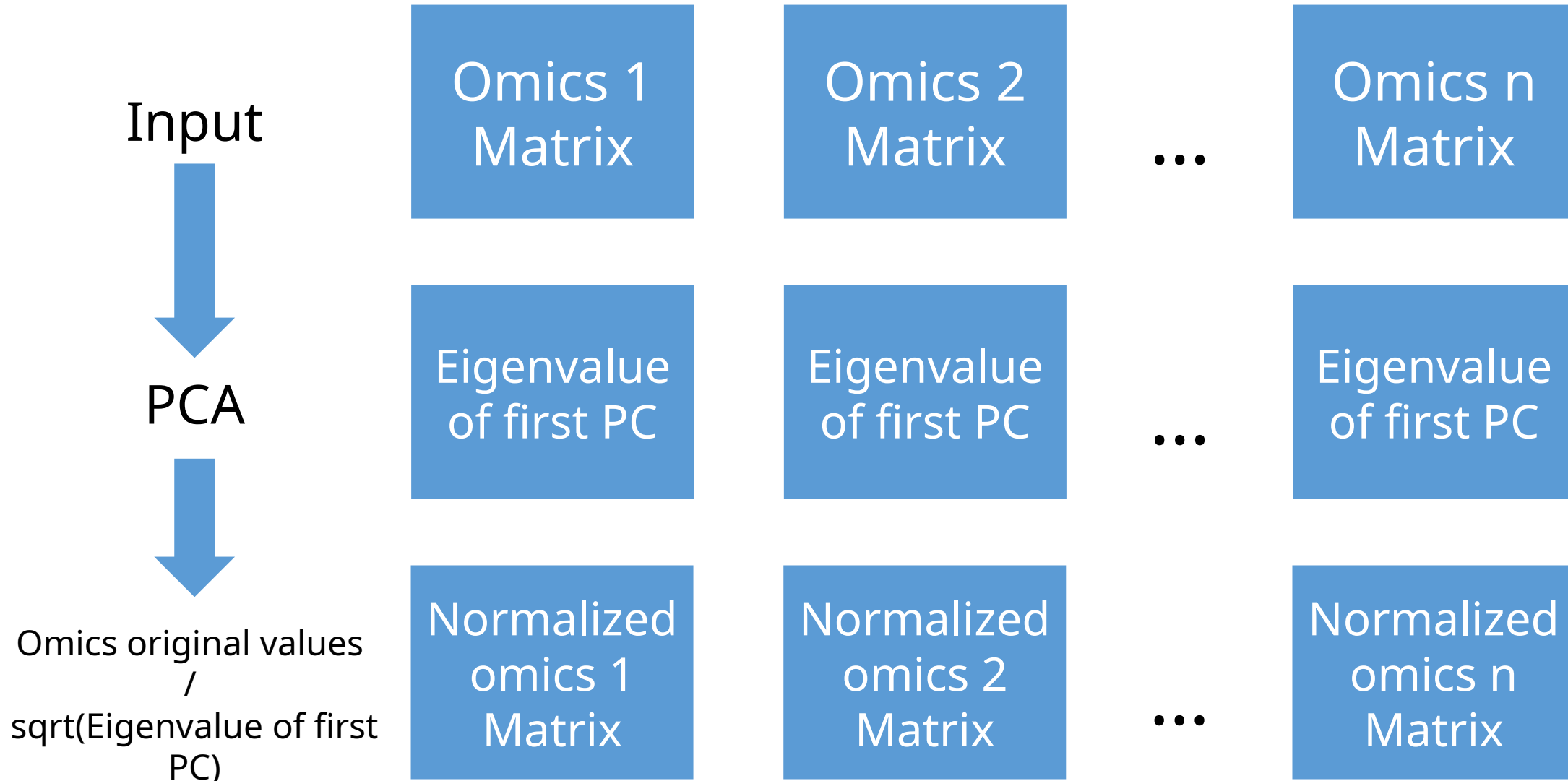
Multiple Factor Analysis



MFA is performed in two steps.

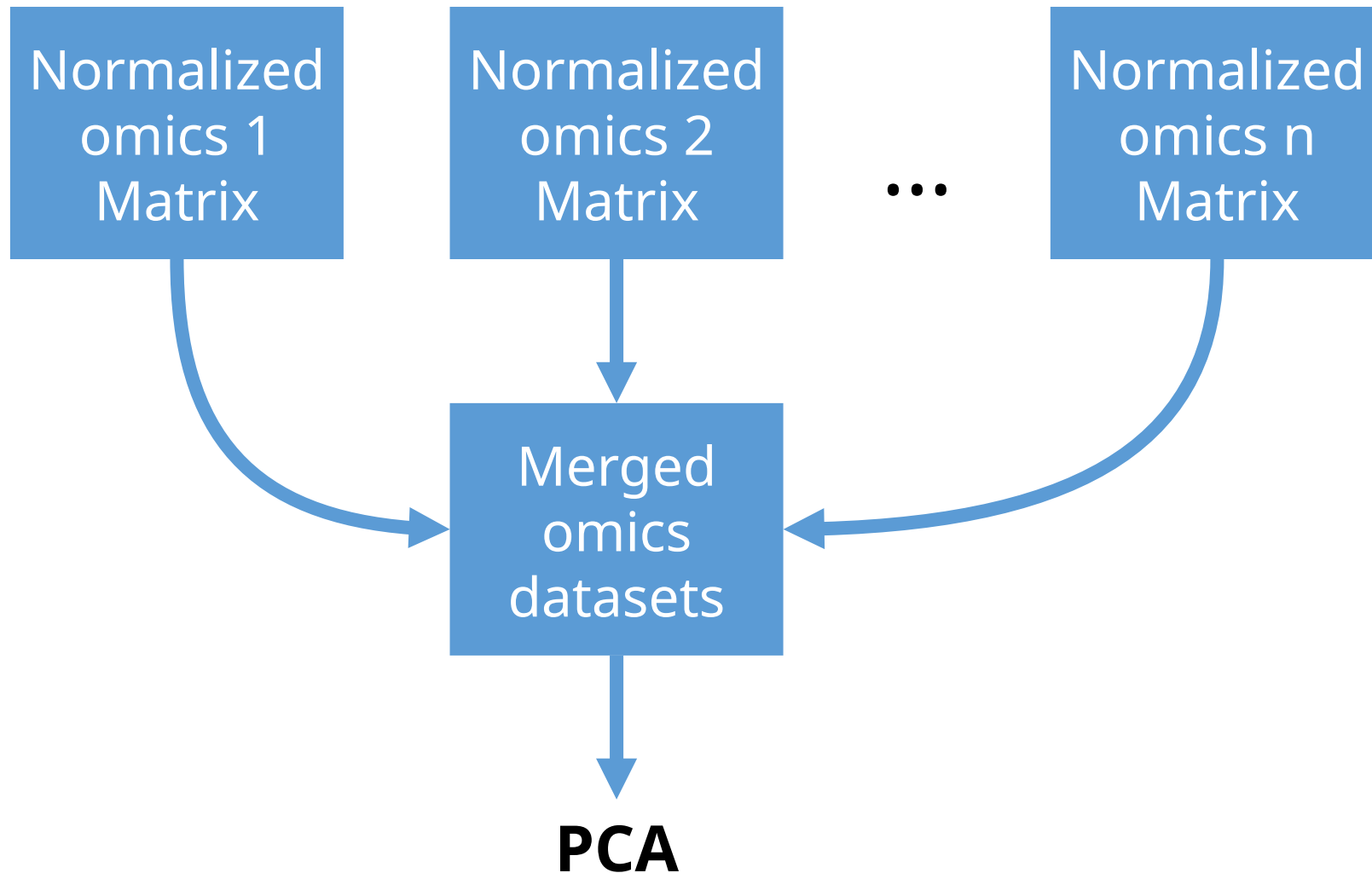


Multiple Factor Analysis (first step)





Multiple Factor Analysis (second step)

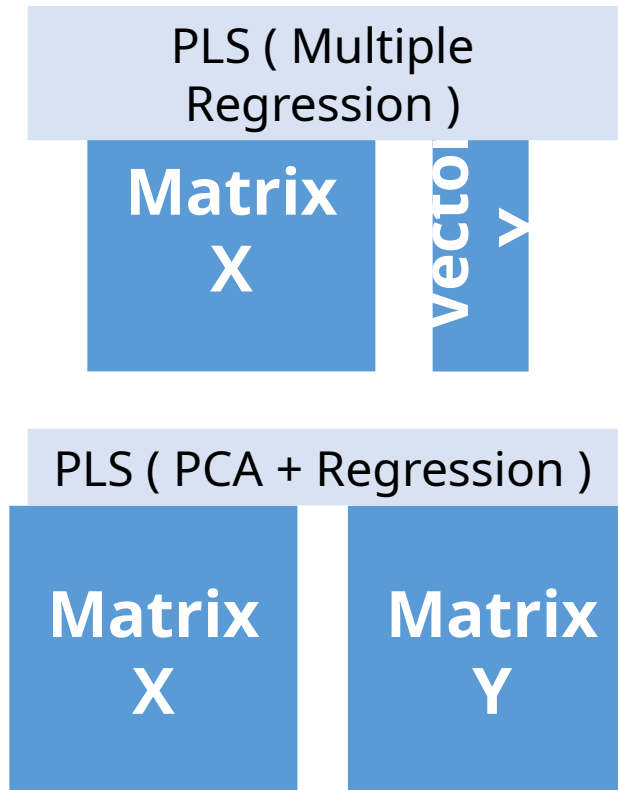


Partial Least Squares (PLS)



Partial Least Squares

- PLS is useful when we need to predict *a set of dependent variables* from a (very) large set of independent variables (i.e., predictors).
- PLS regression generalizes and combines features from principal component analysis and multiple regression.





Partial Least Squares

- The covariance maximization is done using linear combination of components.
- PLS matrix is a diagonal matrix with the “regression weights” as diagonal elements.
- PLS matrix can contain any numbers of components (*e.g.* the regression weights of 2 components).


So we can predict the Y features using X features.




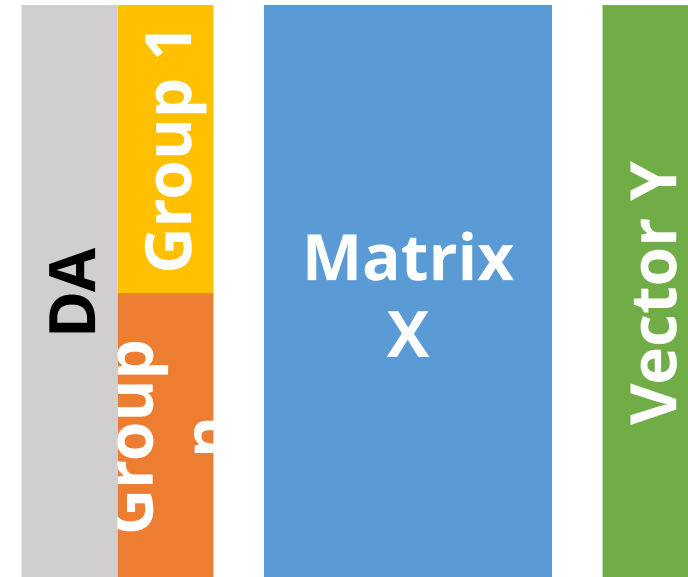
Discriminant Analysis (DA)

Discriminant Analysis

- Discriminant Analysis (DA) undertakes the same task by predicting an outcome.
- DA is useful when we need to predict the dependent variable (outcome) that is of **categorical data** from the independent variables (predictors) that is of **continuous data**.
- DA is useful when we have **different predefined groups** and we want to maximize the separation between them.

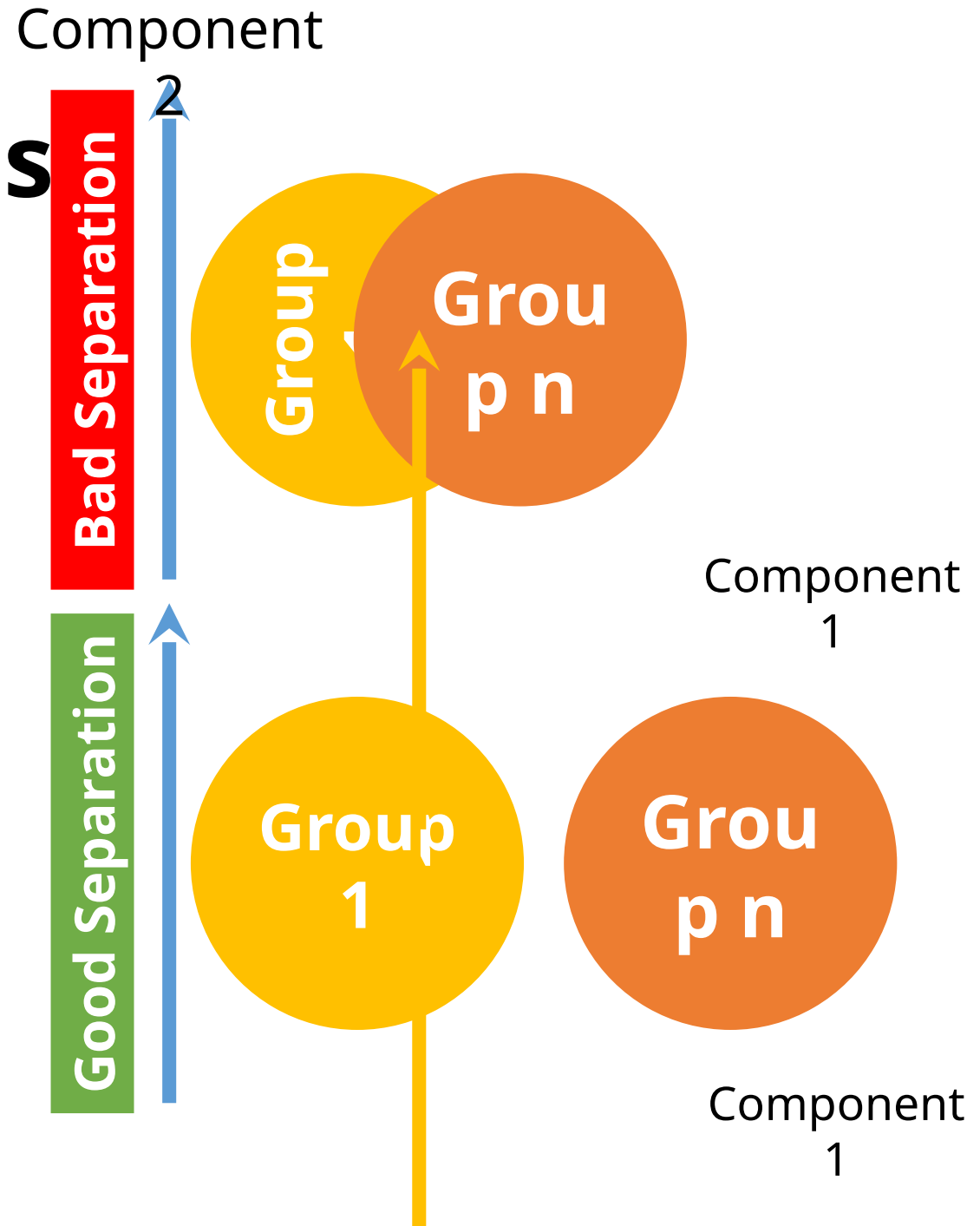
Continuous data 

Categorical data 



Discriminant Analysis

- BUT the objective of DA is to *find linear combinations of features that best discriminate between groups not to find latent components or factors.*
- So, DA maximizes between-group variance while minimizing within-group variance.
- The components that extracted from DA describe the correlation between the features and the component in the separation power.

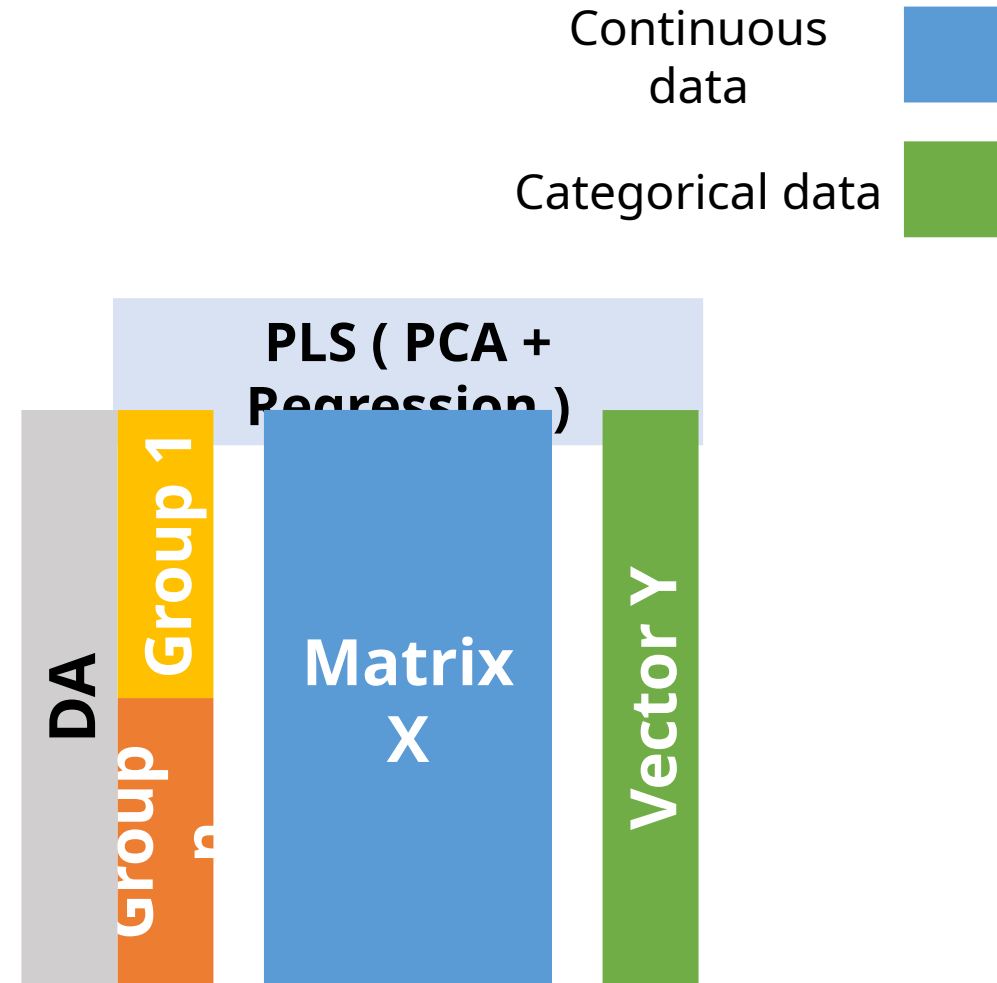


Partial Least Squares – Discriminant Analysis (PLS-DA)

Partial Least Squares - Discriminant Analysis

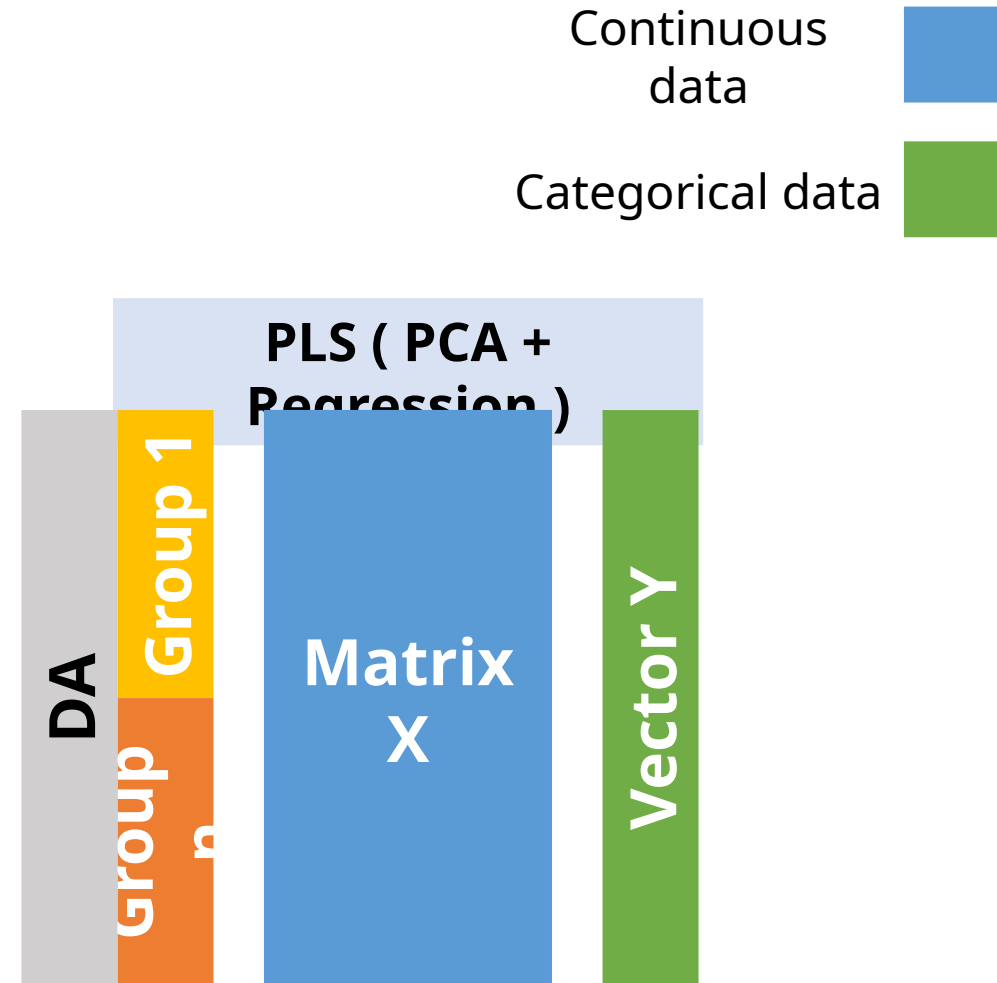
PLS-DA is a linear multivariate model which performs *classification* tasks and is able to *predict* the class of new samples.

As the name suggests, the method *extends PLS* from integrating *two continuous* data matrices to integrating a *continuous* data matrix X *with* a *categorical* outcome variable.



Partial Least Squares - Discriminant Analysis

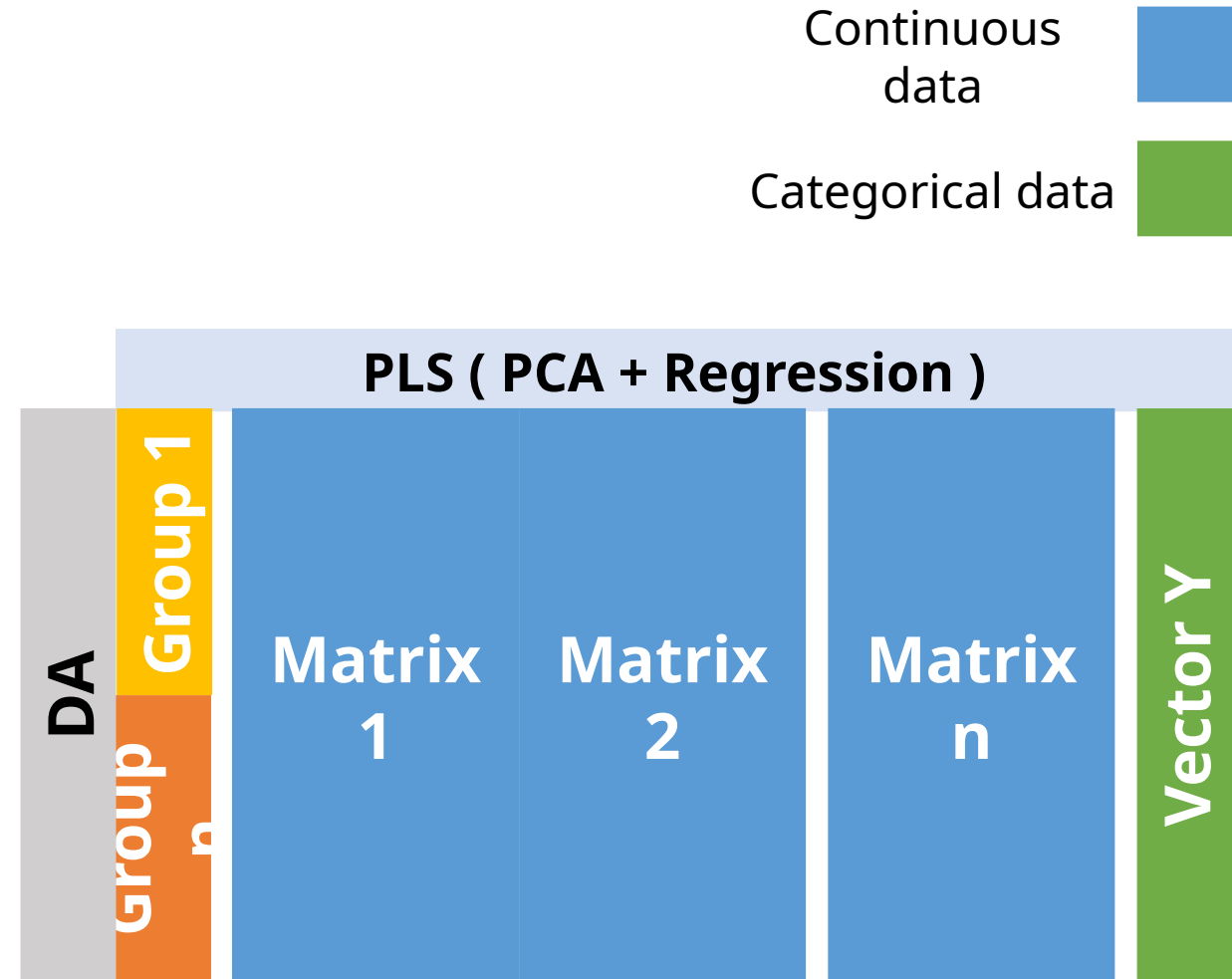
- PLS-DA seeks for **components** (PLS) that best **separate** (DA) the sample groups.



Multiblock sPLS-DA (DIABLO)

DIABLO

DIABLO seeks for *latent components – linear combinations of variables*, from each omics data set so that the sum of covariances between each pair of data sets, including the outcome, is maximized.



DIABLO (Feature selection)

Examining the components after dimension reduction only provides the first level of understanding of the data at the sample level.

Feature selection is a method that extract additional information concerning ***which features are key to explain the variance or covariance of the data.***

Multi-block PLS-DA
for *N*-Integration only

Multi-block sparse PLS-DA (sPLS-DA)
for *N*-integration and feature selection

DIABLO (Feature selection)

Loading vectors (also called Factor loadings or weights) are coefficients that *reflect the importance* of *each feature* in defining each *component*.

Lasso penalization or *sparse* method is the method that *used to identify features* that are *influential* in defining the *components*.

Lasso shrinks many of the features' coefficients in the loading vectors to exactly *zero*.

Since components are extracted separately for each dataset, how can we ensure that important features in a component do not affect other components?

Revision



Dimensionality Reduction Techniques

- Multi-block sPLS-DA.
- Multiple Factor Analysis.

| Tool Name | Technique | Mathematical Base |
|-----------|--------------------------|-------------------|
| DIABLO | Multi-block sPLS-DA | PLS-DA |
| MOFA | Multiple Factor Analysis | PCA |

- PLS-DA = Partial Least Squares Discriminant Analysis
- PCA = Principle Component Analysis

Multi-Omics Integration Protocol

Protocol Summary

The protocol covers three key components:

1. Single-omics data analysis.
2. Knowledge-driven integration using biological networks.
3. Data-driven integration through joint dimensionality reduction.

1 Single-omics data analysis

Input:

Transcriptomics / proteomics data analysis using
ExpressAnalyst.

Metabolomics / Lipidomics data analysis using
MetaboAnalyst.

Output: *Significant features* (mRNA, Proteins, Lipids, ...).

2 Knowledge-driven integration

Input: *Significant features*.

Knowledge-driven integration using OmicsNet.

Output: *key functional features or activity hotspots*.

3 Data-driven integration

Input: *Omics datasets and/or metadata*

Data-driven integration using MOFA and/or DIABLO.

Output: *Influential features*.

Knowledge-driven vs. Data-driven

| | Knowledge-driven | Data-driven |
|---------------------|------------------------------|----------------------------------|
| Dependencies | Database(s) | Data itself |
| Limitation | Limited | Not limited |
| Strategy | Biological networks | Dimensionality reduction |
| Results | Representing facts | Shared patterns and correlations |
| Suitable for | Well-studied model organisms | All organisms |

General workflow for single-omics analysis

| | Next-generation sequencing (NGS) | Mass spectrometry (MS) |
|---------------------------------------|--|--|
| Omics | Genomics, epigenomics or transcriptomics | Proteomics, metabolomics or lipidomics |
| Output file | FASTQ | MS spectra |
| Raw data processing pipelines | Well-established | Still under active development |
| Result from raw omics data processing | High-dimensional table containing abundance measures for different molecules | |

General workflow for single-omics analysis

Bioinformatics workflow:

1. Data quality improvement.

Normalization, filtration, imputation

*Normalization steps are employed to correct for systemic and technical factors across samples.
(such as differences in volume or sequencing depth).*

Data transformation makes feature distributions more comparable and roughly normally distributed.

2. Comparative analysis, also known as differential analysis.

Aiming to identify molecular features that are significantly associated with the phenotypes of interest

3. Functional analysis.

Connects changes of individual molecules with biological activities

Additional processing steps for multi-omics

A filtration step must be done for the omics layer whose features are greater than the others. (e.g. MOFA results will be nearly the same as if the other omics types were not included at all)

Scaling must be done on the omics obtained from different platforms. Since covariance is unscaled, the omics layer with a greater range of values will dominate the analysis. Scaling each omics layer to have comparable distributions can address these issues.

Common functional analysis types

| | Overrepresentation analysis (ORA) | Gene set enrichment analysis (GSEA) |
|----------------------|---|--|
| Omics type | All | All (<i>Except metabolomics and lipidomics</i>) |
| Input gene list | Differentially expressed molecules only | All measured molecules (<i>but ranked based p_value & fold change</i>) |
| Background gene list | List of tissue molecules | |
| Statistical method | Contingency table & fisher's exact test | Enrichment score & empirical phenotype-based permutation test |
| False | Yes | No |

Functional analysis for multi-omics data

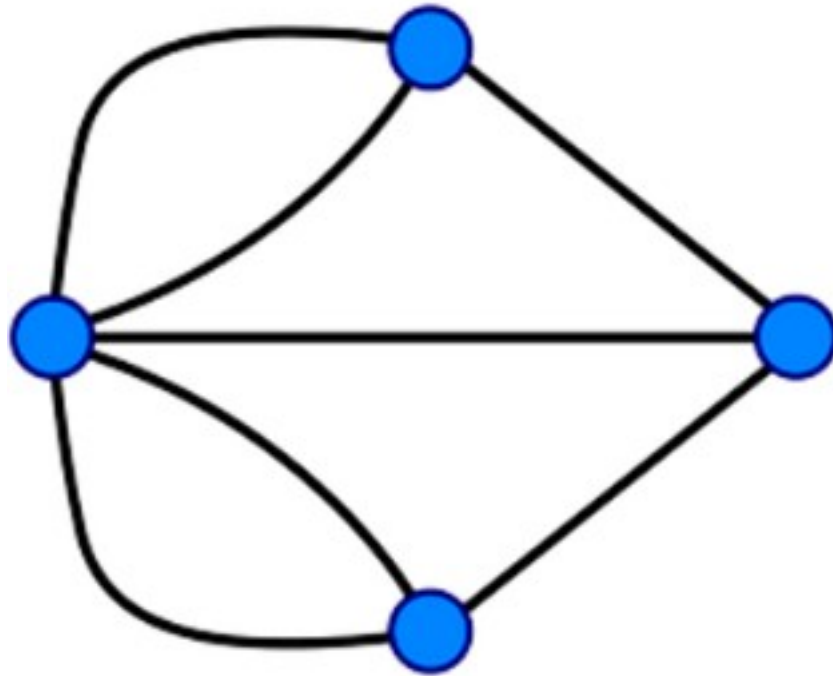
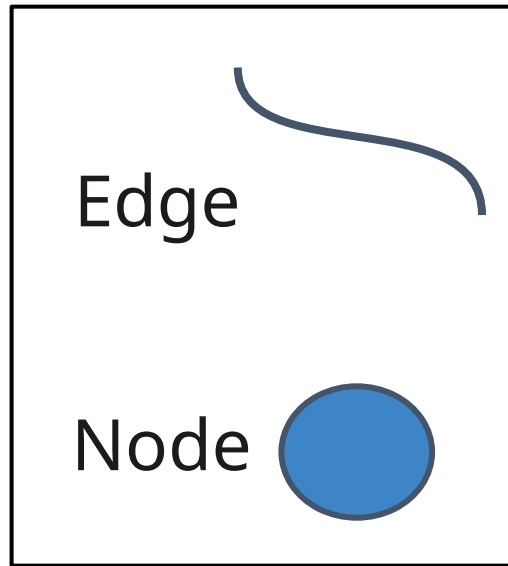
A typical single-omics workflow performs functional analysis after statistical analysis.

For multi-omics data, we may take three general different approaches:

- Independent statistical and functional analysis
is performed separately for each omics type.
- Independent statistical analysis and integrated functional analysis
in case the database contains different types of molecules, this will improve statistical power.
- Integrated statistical analysis with independent functional analysis
DIABLO compute a set of new components that maximize covariance within and across omics layer *then functional analysis can be performed on each set of loadings independently.*

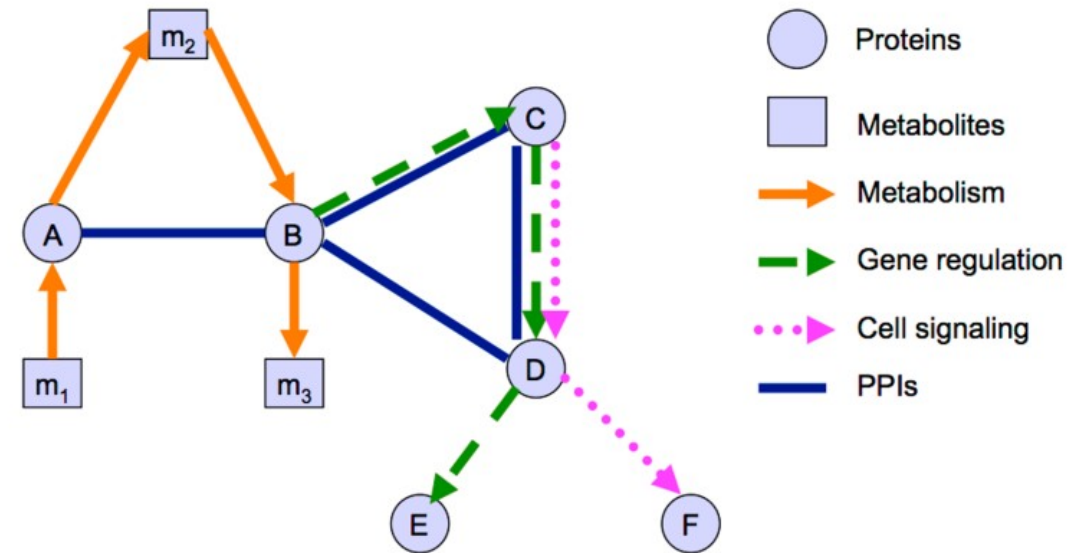
Using biological networks for knowledge-driven

Networks are natural representations of our knowledge.



Some of the most common types of biological networks

- Protein-protein interaction network
- Metabolic networks.
- Genetic interaction networks.
- Gene / transcriptional regulatory networks.
- Cell signalling networks.



The sources of data underlying biological networks

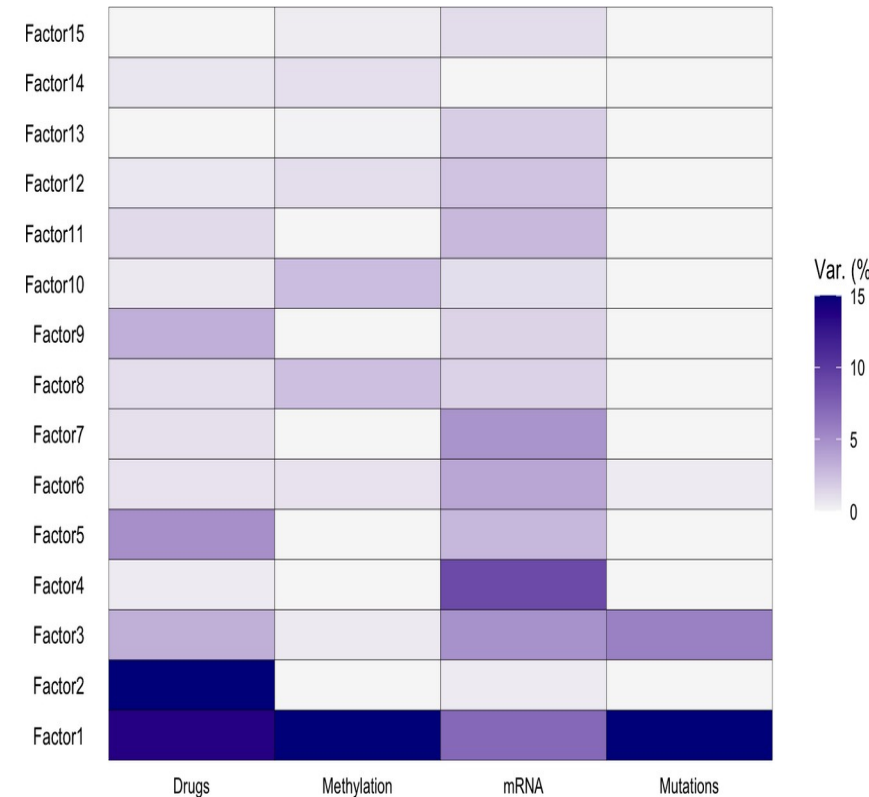
- Manual curation of scientific literature. (*but the size of the datasets is limited*)
- High-throughput datasets. (*but different technique and quality*)
- Computational predictions. (*increases the coverage but the datasets produced are the noisiest*)
- Literature text-mining. (*increases the coverage but the datasets produced are the noisiest*)

Using dimensionality reduction for data-driven

Computes a low-dimensional representation that captures the main characteristics of the high-dimensional data.

There are two important assumptions:

1. The high-dimensional data contains redundant information (*molecules involved in the same biological processes*) so it can be greatly compressed (*because they are often linearly correlated*) without losing important messages.
2. Typical statistical summaries such as variance can adequately capture key characteristics of the data.

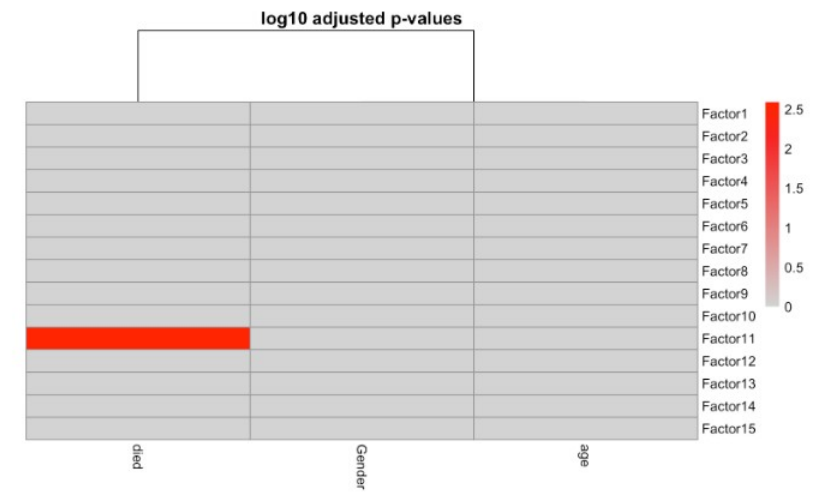
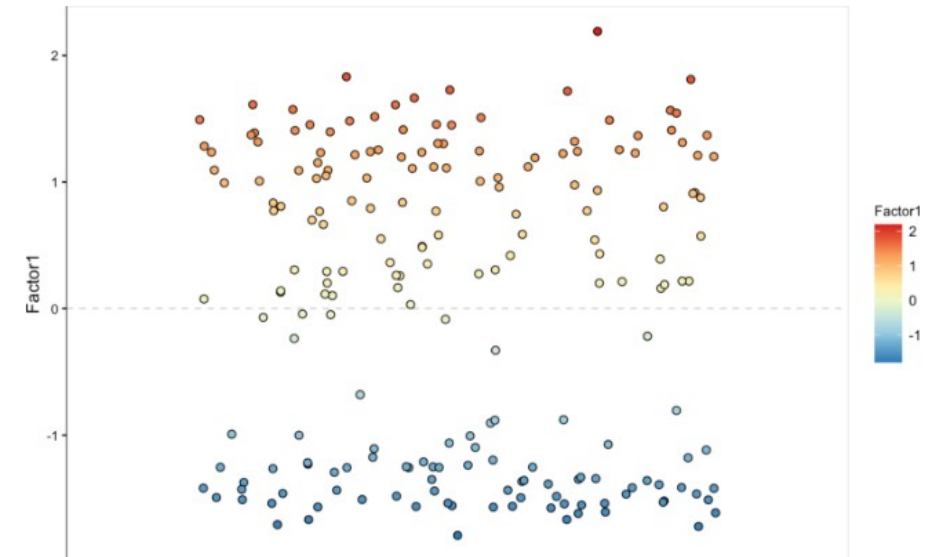


Using dimensionality reduction for data-driven

DR methods compute loadings (feature coefficients) that are multiplied with the omics data to obtain scores (samples in the low-dimensional space defined by the components).

We can annotate scores with different metadata labels.

Once we have identified the components with interesting patterns, we can inspect the corresponding loadings to see which features are influential, and further explore their functional implications.



DR methods

The most commonly used DR methods are PCA and partial least square–discriminant analysis (PLS–DA).

PCA aims to identify a few new components that explain the most **variance** of the data.

PLS–DA aims to identify a few new components that explain the most **covariance** between data and phenotypes of interest.

Shared versus complementary trends

In general, *we cannot directly merge our omics matrices, even after scaling*. This is because scaling does not address the *'shape'* of the data distributions.

Solutions:

DIABLO *identify components separately* but *simultaneously in each layer*, by maximizing a term that includes variance of each data and correlation across data. This finds a balance between components that both explain a substantial proportion of the variability within each layer and are shared across layers.

MOFA first performs an additional *normalization* step to correct for systematic differences in 'shape'. Then, all omics features are directly merged into the same matrix, and subject to PCA.

Supervised versus unsupervised

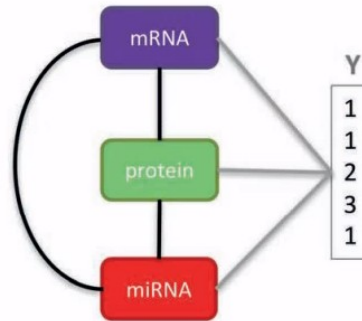
Both **MOFA** is **unsupervised** as they only consider omics data, while **DIABLO** is **supervised** as it also considers the variance of a single metadata variable.

Supervised versus unsupervised

DIABLO has a covariance parameter (between 0 and 1) that determines how much the covariance between omics layers is weighted relative to the variance of the specified metadata.

Lower values of the parameter make the '*supervised*' component more influential. A value of **1** will maximize covariance across omics layers (i.e., **does not consider the metadata at all**), making it very similar to MOFA.

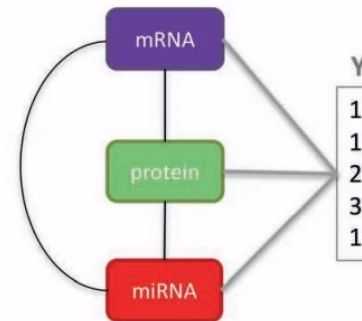
Full design



Design matrix:

| | mRNA | protein | miRNA | Y |
|---------|------|---------|-------|---|
| mRNA | 0 | 1 | 1 | 1 |
| protein | 1 | 0 | 1 | 1 |
| miRNA | 1 | 1 | 0 | 1 |
| Y | 1 | 1 | 1 | 0 |

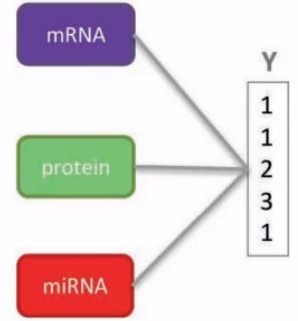
Full weighted design



Design matrix:

| | mRNA | protein | miRNA | Y |
|---------|------|---------|-------|---|
| mRNA | 0 | 0.1 | 0.1 | 1 |
| protein | 0.1 | 0 | 0.1 | 1 |
| miRNA | 0.1 | 0.1 | 0 | 1 |
| Y | 1 | 1 | 1 | 0 |

Null design



Design matrix:

| | mRNA | protein | miRNA | Y |
|---------|------|---------|-------|---|
| mRNA | 0 | 0 | 0 | 1 |
| protein | 0 | 0 | 0 | 1 |
| miRNA | 0 | 0 | 0 | 1 |
| Y | 1 | 1 | 1 | 0 |

Summary

- Least squares is used to extract components in PCA.
- PCA is used as a dimensionality reduction step in factor analysis, PLS, PLS-DA, and DIABLO.
- PLS is used to predict continuous outcome features using continuous predictor features.
- PLS-DA is used to predict categorical outcome using continuous predictor features.
- DIABLO is used to predict categorical outcome using multiple block of continuous predictor features.

Summary

| | DIABLO | MOFA |
|------------------------------------|--|---|
| Basic Principal | Maximize the covariance within and across the omics layers. | Maximize the variance of merged omics layers. |
| Component extraction method | Separately but simultaneously in each omics layer. Balanced components. | Shared and complementary. Combined components. |
| Machine learning type | Supervised or unsupervised | Unsupervised |
| Dimensionality reduction technique | Multi-block sPLS-DA. PLS-DA. | Multiple factor analysis. PCA. |
| Input data type | Continuous and categorical | Continuous |

References

- Lê Cao, K.A. and Welham, Z.M., 2021. **Multivariate data integration using R**: methods and applications with the mixOmics package. Chapman and Hall/CRC.
- Backhaus, K., Erichson, B., Gensler, S., Weiber, R. and Weiber, T., 2021. **Multivariate analysis**. Springer Books, 10, pp.978-3.
- Ewald, J.D., Zhou, G., Lu, Y., Kolic, J., Ellis, C., Johnson, J.D., Macdonald, P.E. and Xia, J., 2024. **Web-based multi-omics integration using the analyst software suite**. Nature Protocols, pp.1-31.
- Abdi, H., 2010. **Partial least squares regression and projection on latent structure regression (PLS Regression)**. Wiley interdisciplinary reviews: computational statistics, 2(1), pp.97-106.
- Abdi, H. and Valentin, D., 2007. **Multiple factor analysis (MFA)**. Encyclopedia of measurement and statistics, pp.657-663.

Thank You