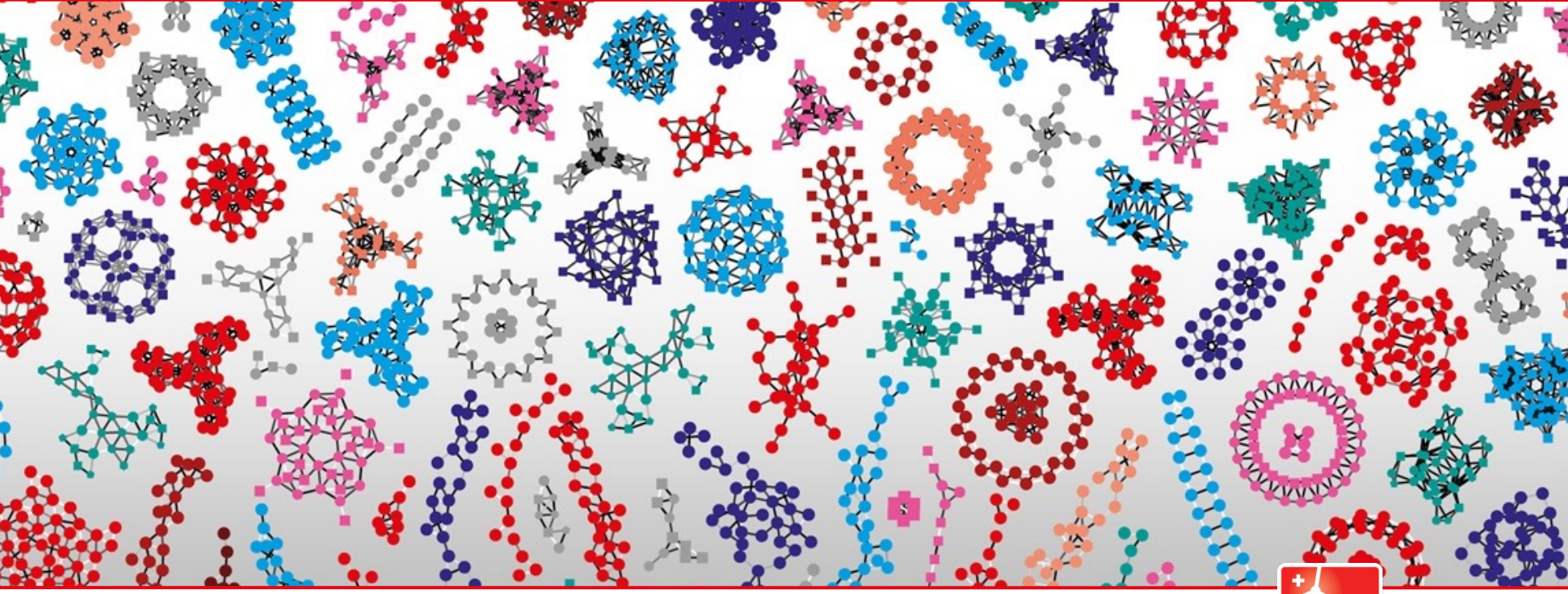


Swiss Institute of Bioinformatics

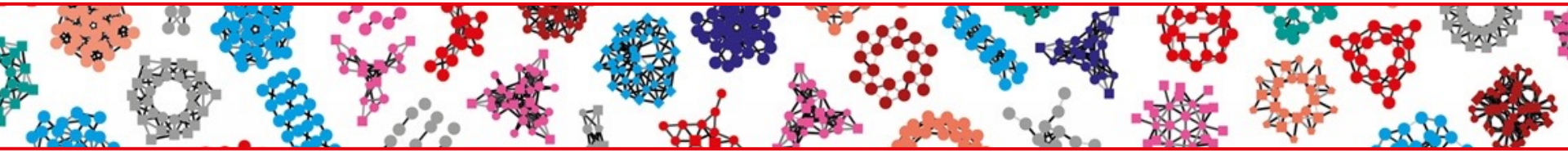
Dimensionality reduction

Van Du Tran

Vital-IT, SIB Swiss Institute of Bioinformatics

Geneva, 13-14 Mar 2024

Overview



01 Principal Component Analysis

02 Partial Least Squares

03 Canonical Correlation Analysis

04 Towards Nonlinearity

Questions in practice

- **A real-estate agency wants to see the similarity/difference between its properties**
  - Properties' characteristics: price, surface area, form, floors, bedrooms, bathrooms, entrances, garages, yards, etc.
  - Access to nearby facilities
  - Local living cost

- **Patients need stratifying for clinical trials**
  - Demographics
  - Lifestyle
  - Multi-omics patterns
  - etc.

What is PCA?

Several features (variables) to consider

- Relationships between features
- Risk of violation on assumptions of modeling
- Risk of overfitting the model to data

⇒ **Reduce the dimension of the feature space**

- *Feature selection*: find a subset of input features
- *Feature extraction*: project high-dimensional space into a space of fewer dimensions

PCA

Karl Pearson – mathematician & biostatistician (1901)

When is PCA used?

- Cannot identify features to eliminate
- Need *new* features independent of one another
- Accept that the *new* independent features are less interpretable

$\mathbf{Y}$: n samples $\times$ p features

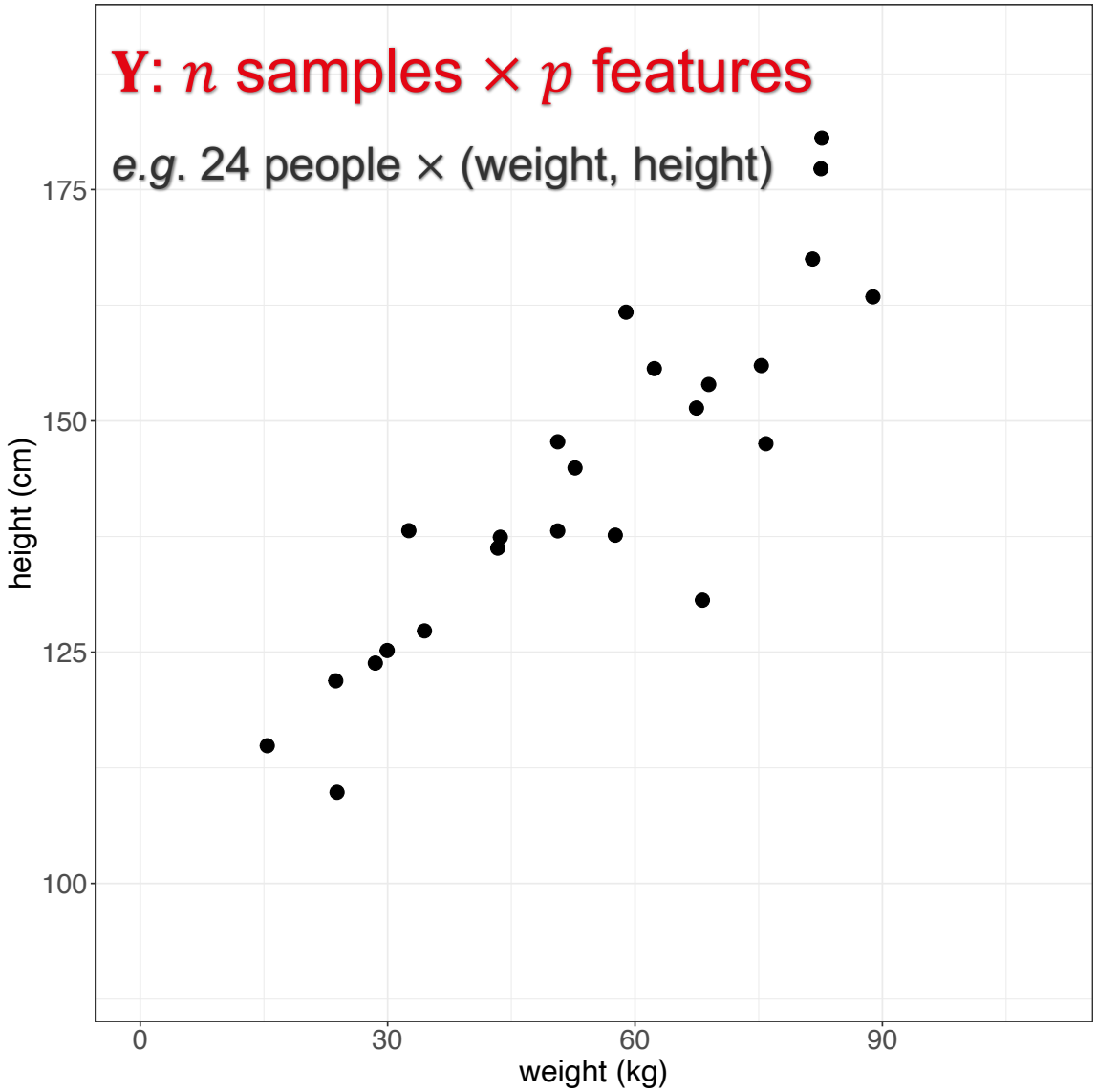
Samples

How are samples grouped together into subgroups by similarity?

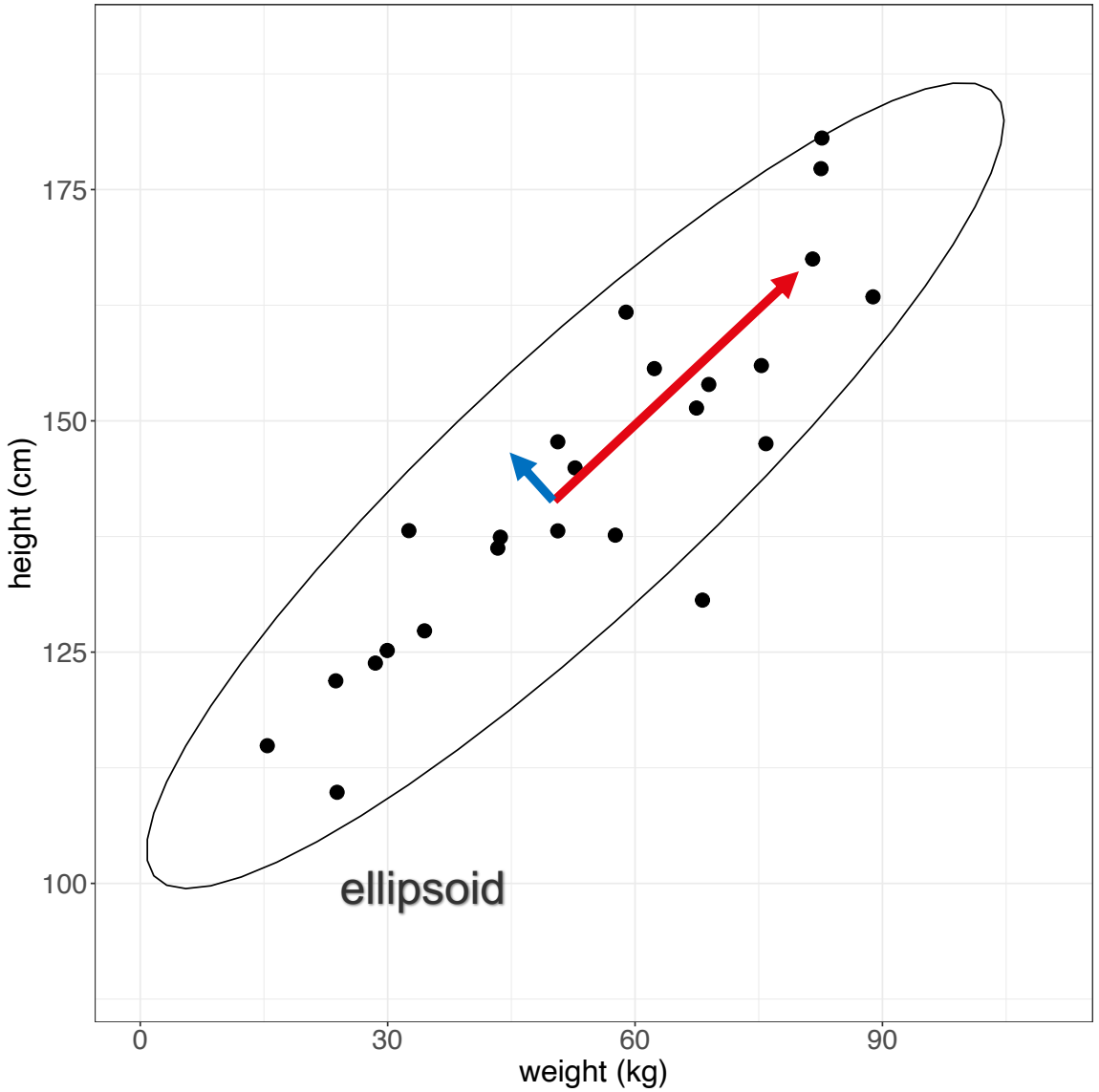
Features

What underlying factors influence the grouping?

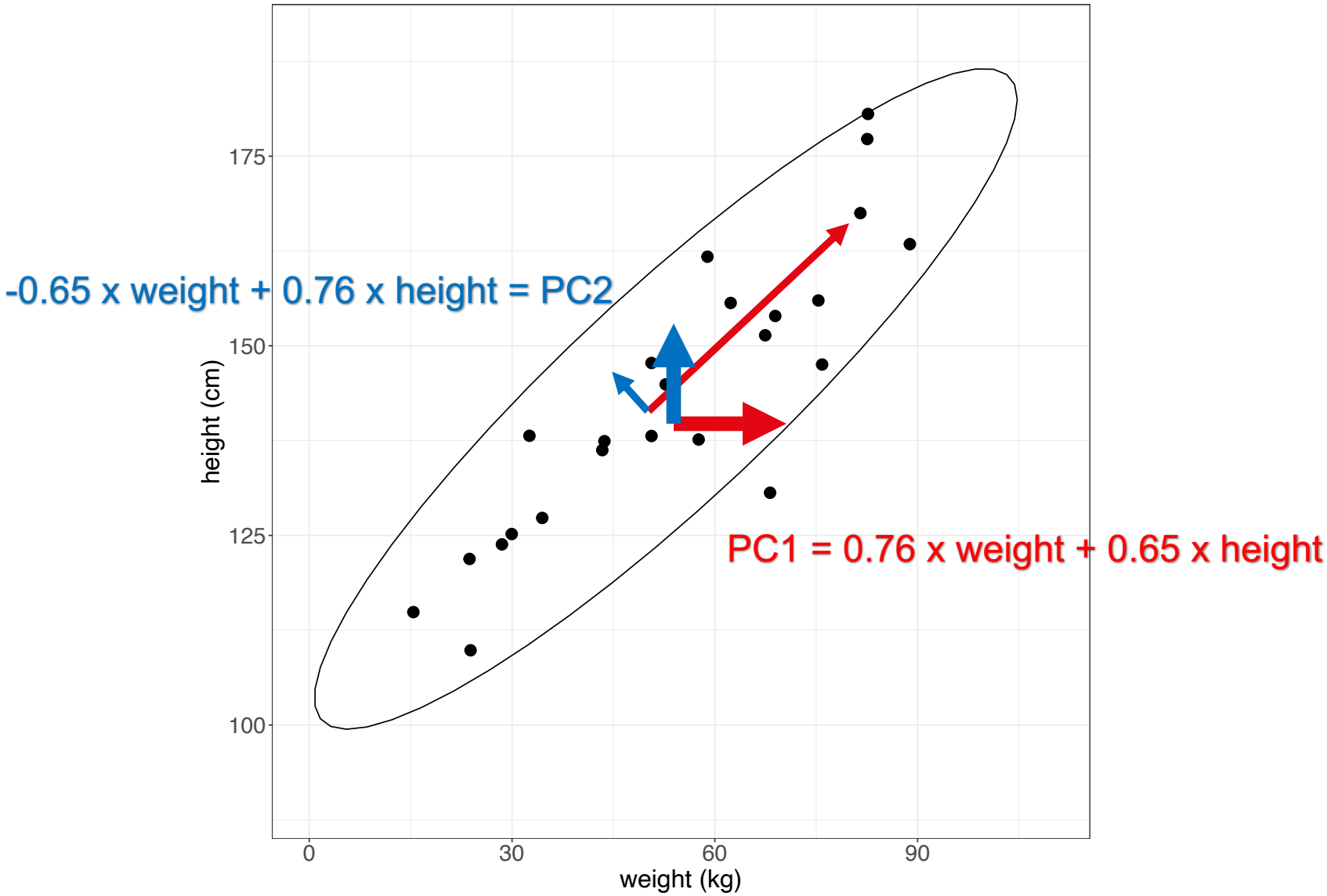
How does PCA work?



How does PCA work?



How does PCA work?



How does PCA work?

- $\mathbf{Y} = (\boldsymbol{Y}_1\ \boldsymbol{Y}_2 \ldots\ \boldsymbol{Y}_p)$
- REQUIRED: cleaning
- OPTIONAL: normalizing
- REQUIRED: centering

$$\text{mean}(\boldsymbol{Y}_i) = 0$$

- RECOMMENDED: appropriately scaling (after centering)

$$\text{var}(\boldsymbol{Y}_i) = 1, \text{ etc.}$$

Data cleaning

GARBAGE IN, GARBAGE OUT

- Missing values: biologically or technically unidentified

  $\Rightarrow$ *removal, imputation*
- Inconsistent data: qualitative, ill-formatted data

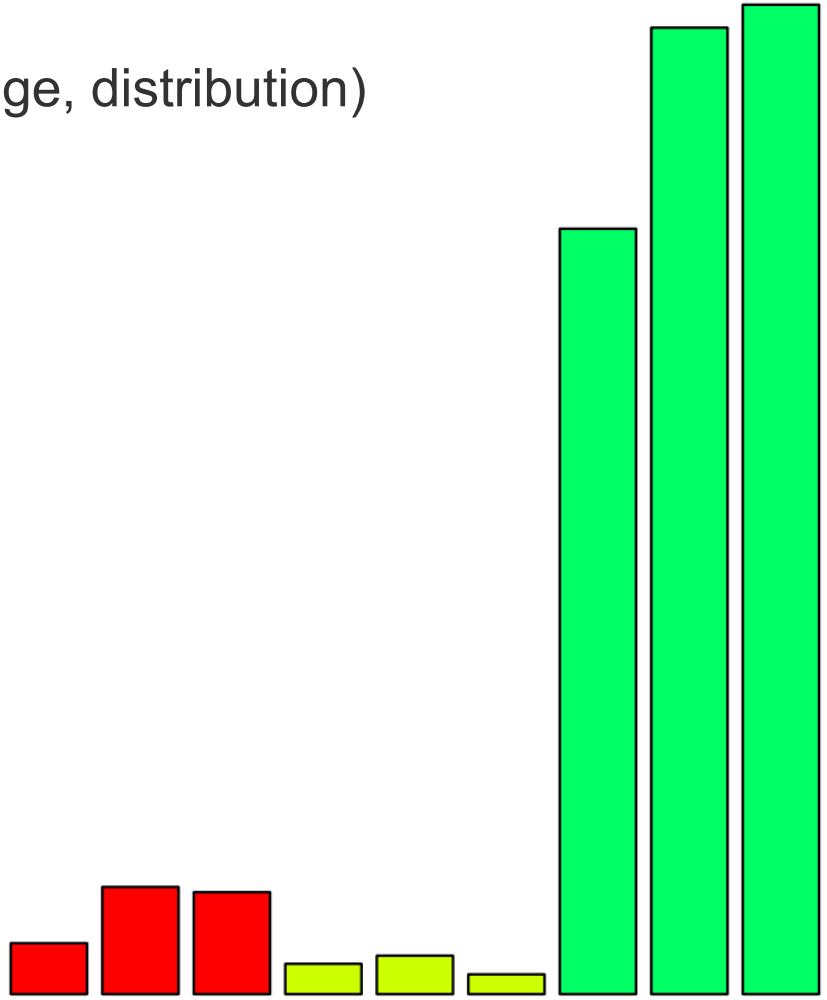
  $\Rightarrow$ *reformatting, correction*
- Outlier and noisy data

  $\Rightarrow$ *removal*
- Redundancy

  $\Rightarrow$ *removal*

Data normalization

- Account for biases (technical variation) from sample handling to instrumentation difference
- Make samples more comparable (range, distribution)
- Various normalization techniques



Data centering

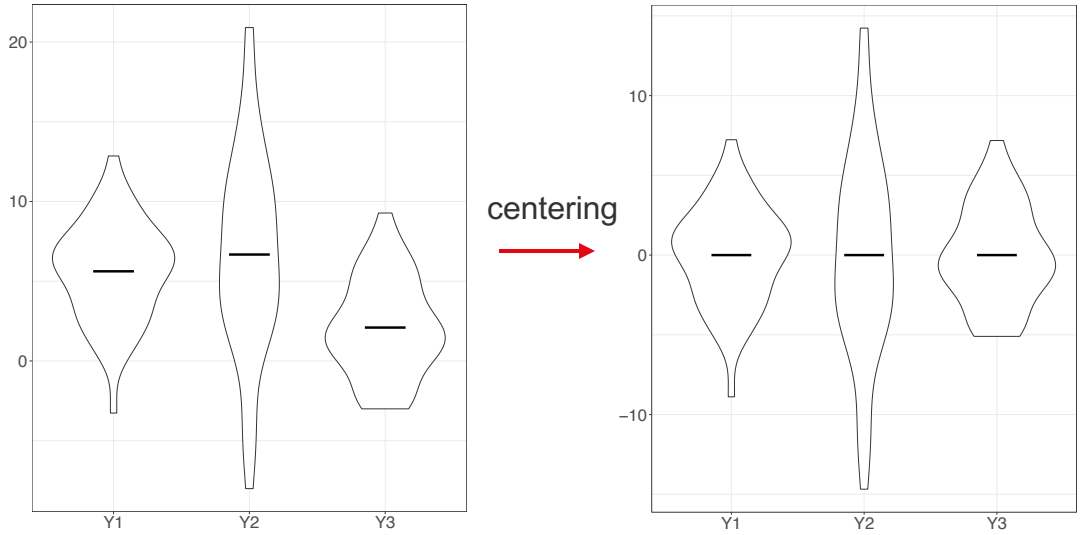
Each variable fluctuates around zero instead of its mean value

WHY: Offsets complicate models with more parameters, introduce algorithmic problems

HOW: For each variable

- Compute its mean
- Subtract the mean from all its values

$$\text{mean}(\boldsymbol{Y}_i) = 0$$



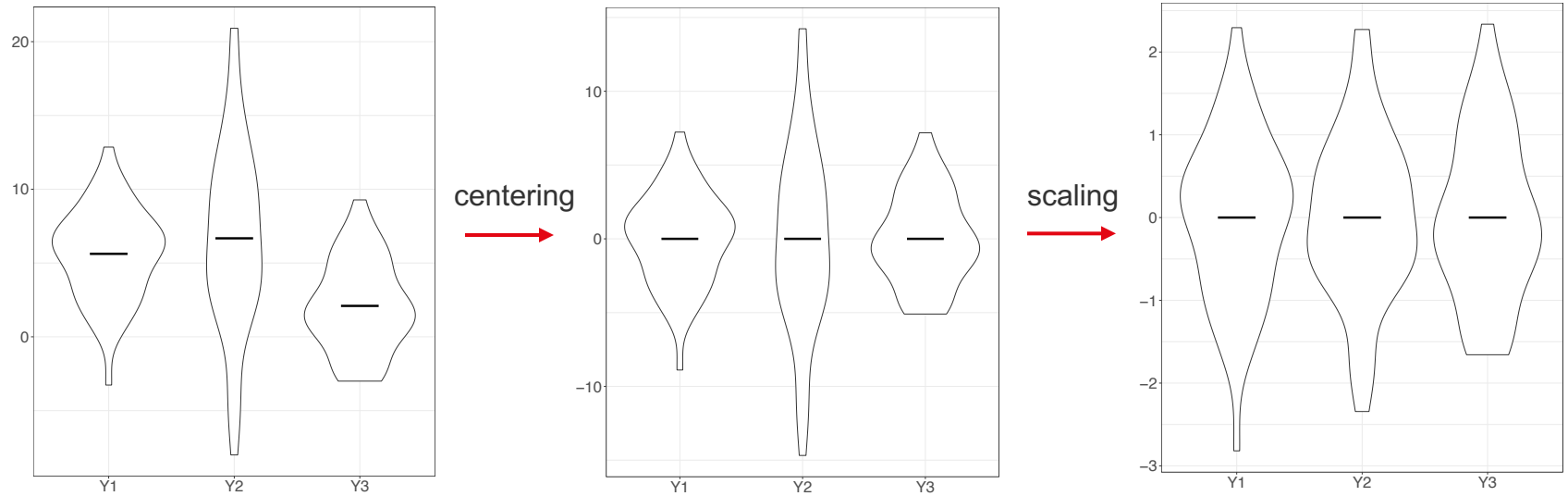
Data scaling

All variables are allocated an comparable importance.

WHY: Variables with higher variances dominate those of lower variances

HOW: For each centered variable, divide all variables centered values by the scaling factor

- Unit variance scaling: standard deviation $var(\boldsymbol{Y}_i) = 1$
- Pareto scaling: square root of standard deviation



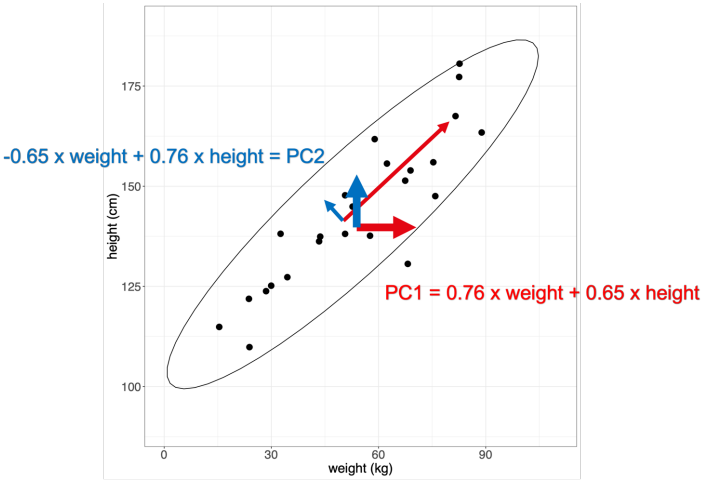
Principal component

Coordinate change: $(\boldsymbol{Y}_1\ \boldsymbol{Y}_2 \ldots\ \boldsymbol{Y}_p) \rightarrow (\boldsymbol{Z}_1\ \boldsymbol{Z}_2 \ldots\ \boldsymbol{Z}_p)$

Linear combination:

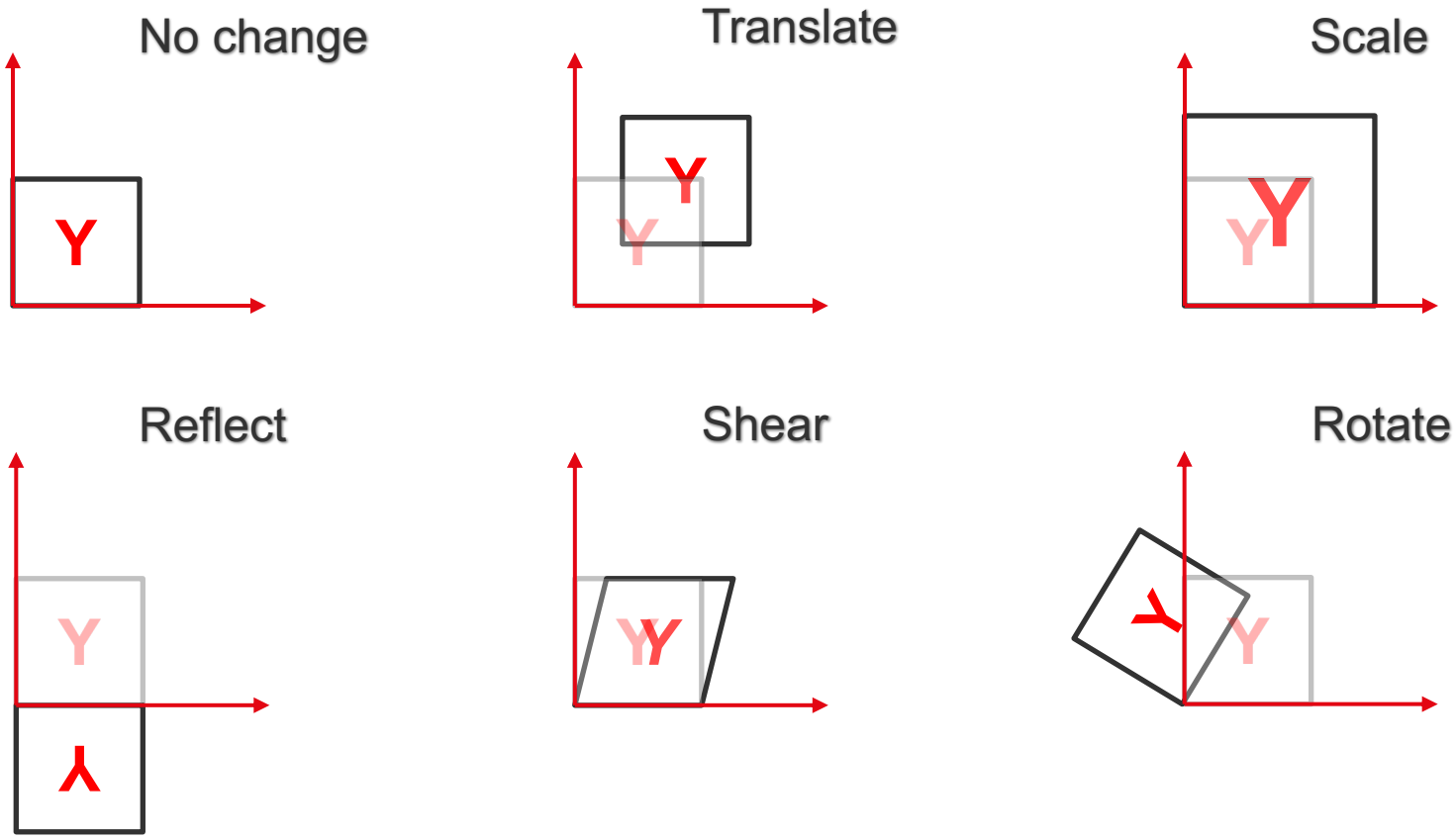
$$\begin{aligned} \boldsymbol{Z}_1 &= w_{11}\boldsymbol{Y}_1 + w_{12}\boldsymbol{Y}_2 + \cdots + w_{1p}\boldsymbol{Y}_p \\ \boldsymbol{Z}_2 &= w_{21}\boldsymbol{Y}_1 + w_{22}\boldsymbol{Y}_2 + \cdots + w_{2p}\boldsymbol{Y}_p \\ &\cdots \\ \boldsymbol{Z}_p &= w_{p1}\boldsymbol{Y}_1 + w_{p2}\boldsymbol{Y}_2 + \cdots + w_{pp}\boldsymbol{Y}_p \end{aligned}$$

$$\boldsymbol{Z}_1 = 0.8\,\boldsymbol{Y}_1 + 0.1\,\boldsymbol{Y}_2 + \cdots + 0.001\,\boldsymbol{Y}_p$$



Latent variables

Linear transformation



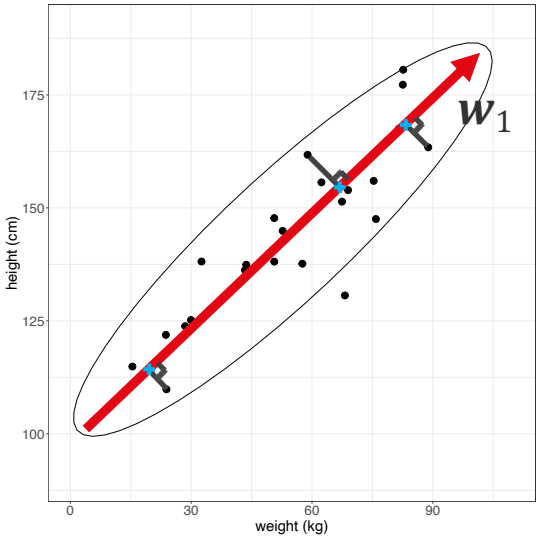
Principal component 1

Coordinate change: $(\boldsymbol{Y}_1\ \boldsymbol{Y}_2 \ldots\ \boldsymbol{Y}_p) \rightarrow (\boldsymbol{Z}_1\ \boldsymbol{Z}_2 \ldots\ \boldsymbol{Z}_p)$

Linear combination:

$$\boldsymbol{Z}_1 = w_{11}\boldsymbol{Y}_1 + w_{12}\boldsymbol{Y}_2 + \cdots + w_{1p}\boldsymbol{Y}_p = \mathbf{Y}\boldsymbol{w}_1$$

=> projection of $\mathbf{Y}$ onto vector $\boldsymbol{w}_1$



Find $\boldsymbol{w}_1$ ($\|\boldsymbol{w}_1\| = \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{w}_1 = 1$) to maximize:

$$\mathrm{var}(\boldsymbol{Z}_1) = \boldsymbol{w}_1^{\mathrm{T}}\,\mathrm{cov}(\mathbf{Y})\,\boldsymbol{w}_1$$

Variance-covariance matrix

$$\mathrm{cov}(\mathbf{Y}) = \begin{pmatrix} \mathrm{cov}(\boldsymbol{Y}_1, \boldsymbol{Y}_1) & \cdots & \mathrm{cov}(\boldsymbol{Y}_1, \boldsymbol{Y}_p) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\boldsymbol{Y}_p, \boldsymbol{Y}_1) & \cdots & \mathrm{cov}(\boldsymbol{Y}_p, \boldsymbol{Y}_p) \end{pmatrix}$$

Solution: max $\mathrm{var}(\boldsymbol{Z}_1)$ = max eigenvalue of $\mathrm{cov}(\mathbf{Y})$
at $\boldsymbol{w}_1$ = corresponding eigenvector

Hint: Lagrange multiplier + derivative

Principal component 2

Coordinate change: $(\boldsymbol{Y}_1\ \boldsymbol{Y}_2 \ldots\ \boldsymbol{Y}_p) \rightarrow (\boldsymbol{Z}_1\ \boldsymbol{Z}_2 \ldots\ \boldsymbol{Z}_p)$

Linear combination:

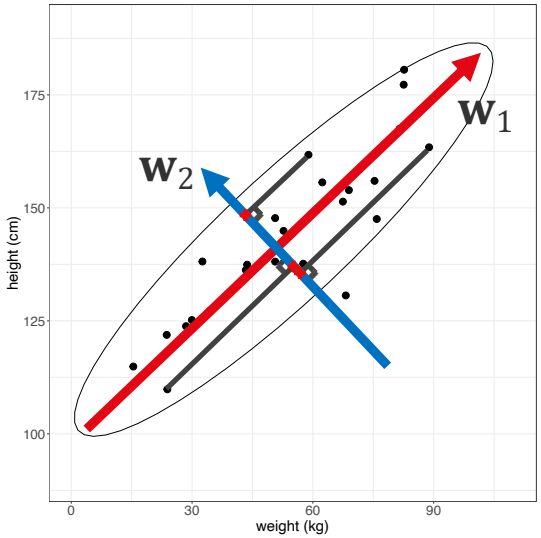
$$\boldsymbol{Z}_2 = w_{21}\boldsymbol{Y}_1 + w_{22}\boldsymbol{Y}_2 + \cdots + w_{2p}\boldsymbol{Y}_p = \mathbf{Y}\boldsymbol{w}_2$$

=> projection of $\mathbf{Y}$ onto vector $\boldsymbol{w}_2$

$\boldsymbol{w}_2 \perp \boldsymbol{w}_1$: independent projection

Find $\boldsymbol{w}_2$ ($\|\boldsymbol{w}_2\| = \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{w}_2 = 1$ and $\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{w}_1 = 0$) to maximize:

$$\mathrm{var}(\boldsymbol{Z}_2) = \boldsymbol{w}_2^{\mathrm{T}}\,\mathrm{cov}(\mathrm{Y})\,\boldsymbol{w}_2$$



Solution: max $\mathrm{var}(\boldsymbol{Z}_2)$ = 2nd max eigenvalue of $\mathrm{cov}(\mathbf{Y})$
at $\boldsymbol{w}_2$ = corresponding eigenvector

and so on

PCA implementation

$\mathbf{Y}$: n samples $\times$ p features

Input: $\mathbf{Y}$ or $\mathrm{cov}(\mathbf{Y})$

Output:

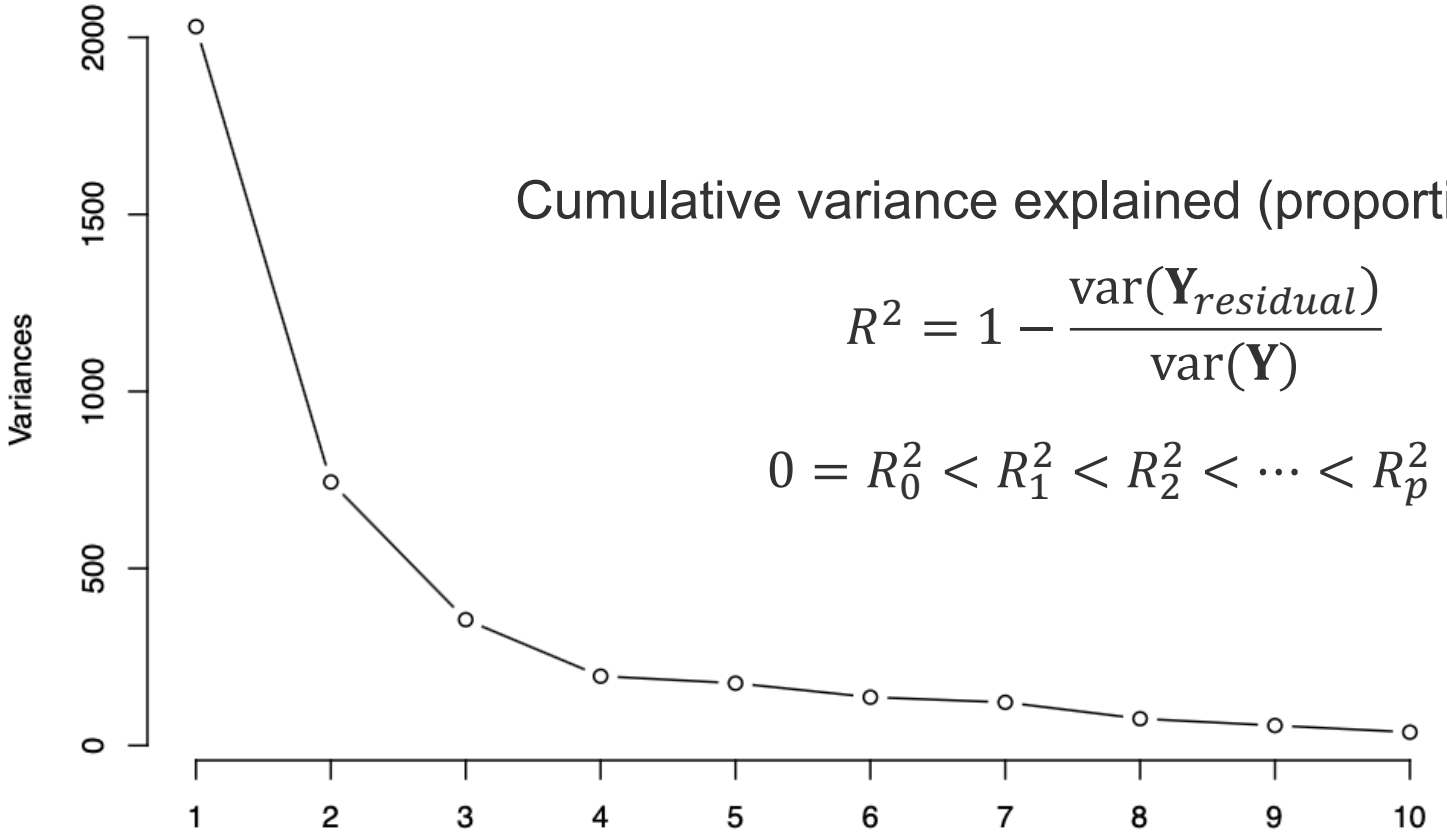
- sdev: square root of eigenvalues of $\mathrm{cov}(\mathbf{Y})$
- scores: ($\boldsymbol{Z}_1$ $\boldsymbol{Z}_2$ ... $\boldsymbol{Z}_p$)
- loadings/rotation: ($\boldsymbol{w}_1$ $\boldsymbol{w}_2$ ... $\boldsymbol{w}_p$)

Why does PCA work? Why should PCA be used?

- Covariance matrix: relation between features
- Eigenvectors of covariance matrix: directions of dispersion
- Eigenvalues of covariance matrix: importance of directions
- Assumption: variability ~ signal
- Application: data exploration, visualization of underlying patterns within correlated data sets, decorrelation, detection of outliers, data compression
- **Disadvantage**: linearity assumption, interpretability, sensitive to scaling and outliers

PCA: which number of principal components?

Variability explained with each component



Cumulative variance explained (proportion)

$$R^2 = 1 - \frac{\mathrm{var}(\mathbf{Y}_{residual})}{\mathrm{var}(\mathbf{Y})}$$

$$0 = R_0^2 < R_1^2 < R_2^2 < \cdots < R_p^2 = 1$$

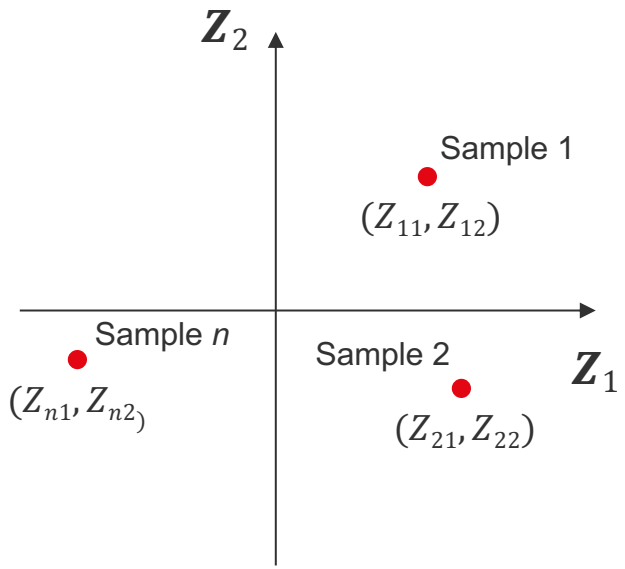
PCA score plot

Score: projection measures of samples in each principal component

=> Coordinate of samples on each axis PC

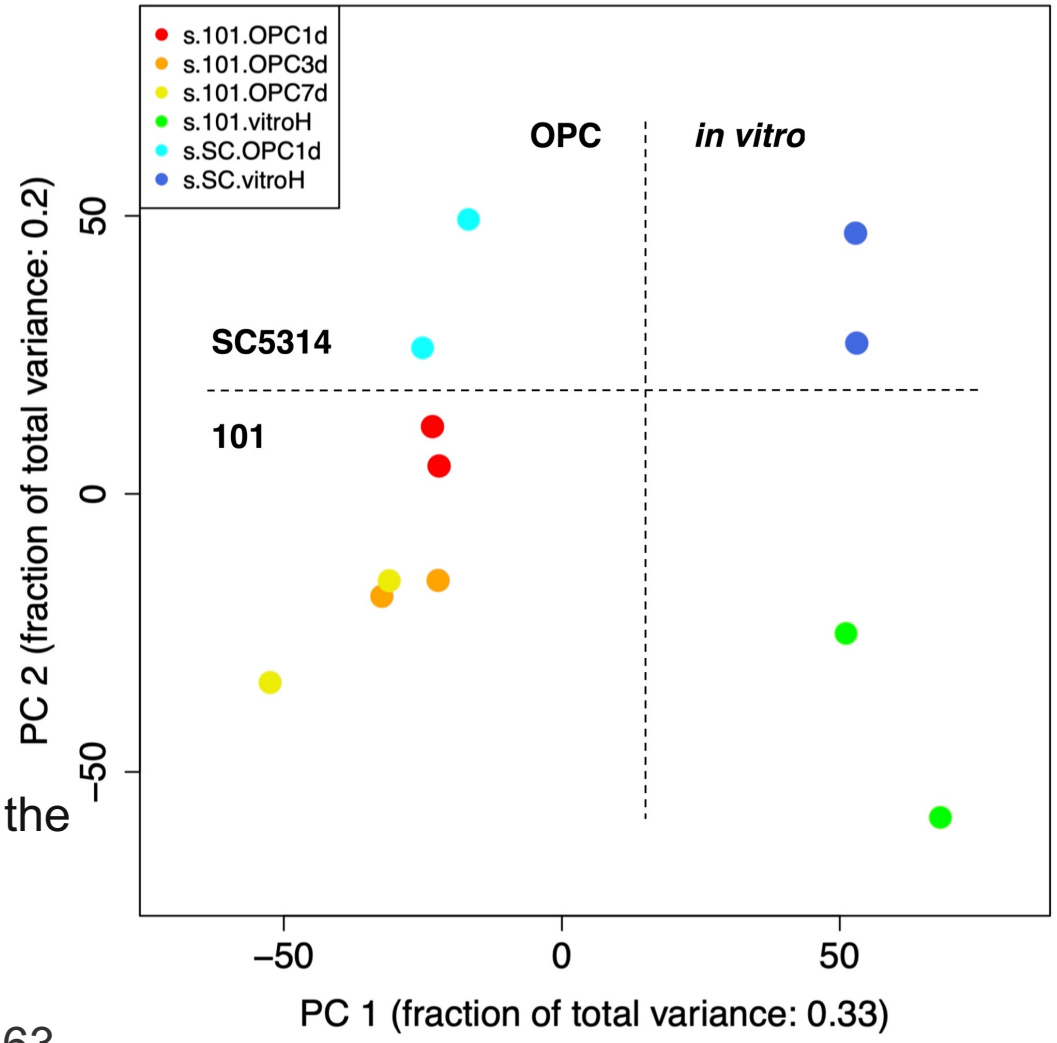
$$(\boldsymbol{Z}_1 \ \boldsymbol{Z}_2 \ \ldots \ \boldsymbol{Z}_p)$$

$$\begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \\ \\ Z_{n1} & Z_{n2} \end{pmatrix}$$



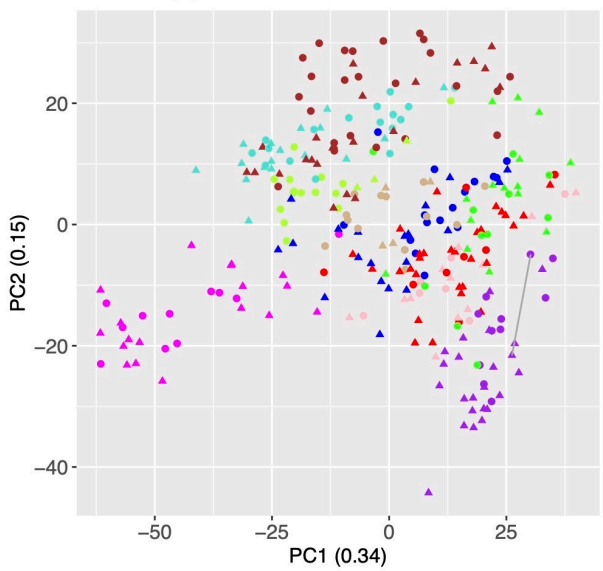
PCA score plot

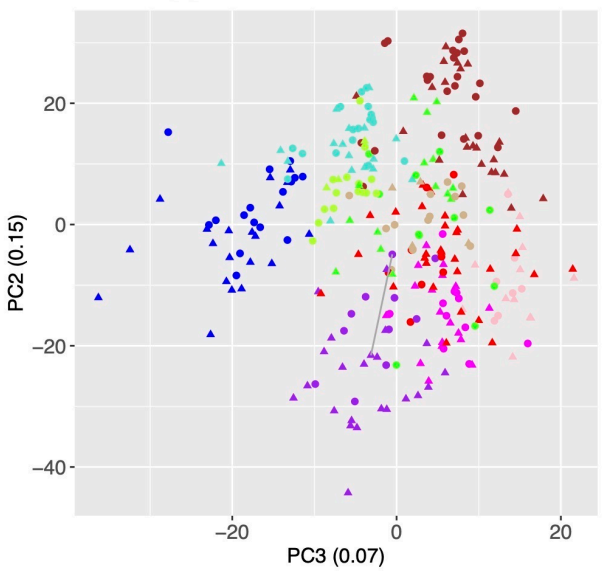
Transcriptomics dataset:
Gene expression profiles of different isolates of *Candida albicans in vitro* and during oral infection on mice

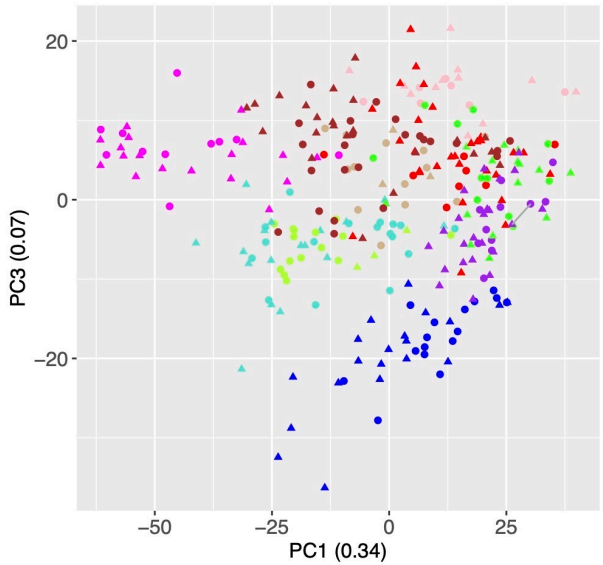


Candida albicans commensalism in the oral mucosa is favoured by limited virulence and metabolic adaptation
Lembert *et al*.
https://doi.org/10.1101/2021.10.11.463879

PCA score plot







Metagenomics dataset:
Microbiome composition in patients' lung post lung-transplantation

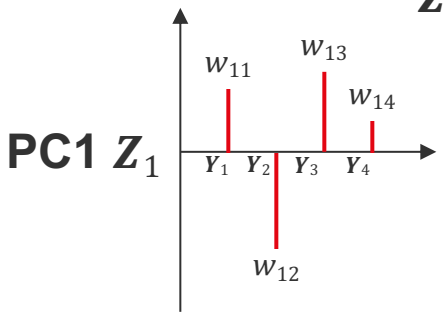
PCA loading plot

Loading: contribution of given features to each principal component

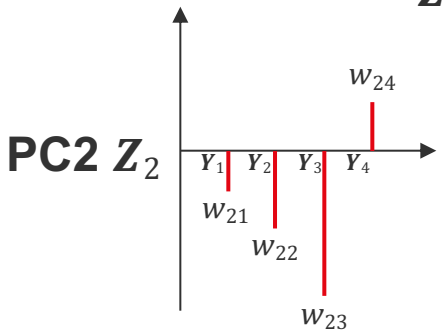
⇒ Coordinate of given features on each axis PC

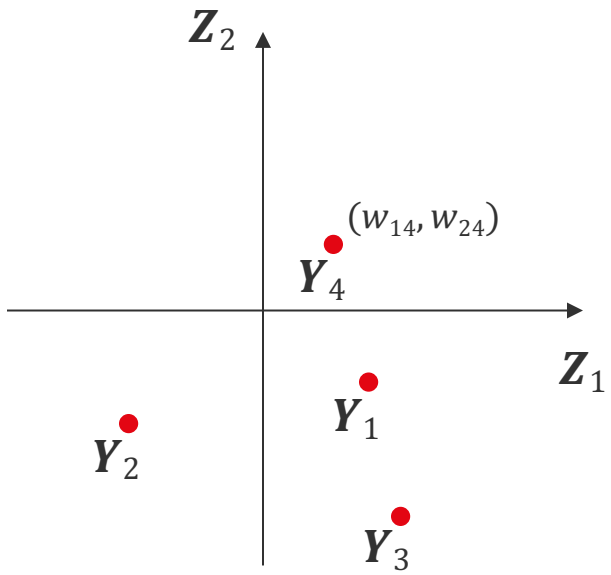
⇒ Highly correlated features: similar weights in the loading vectors; close together in the loading plots of all dimensions.

$$Z_1 = w_{11}Y_1 + w_{12}Y_2 + \cdots + w_{1p}Y_p$$



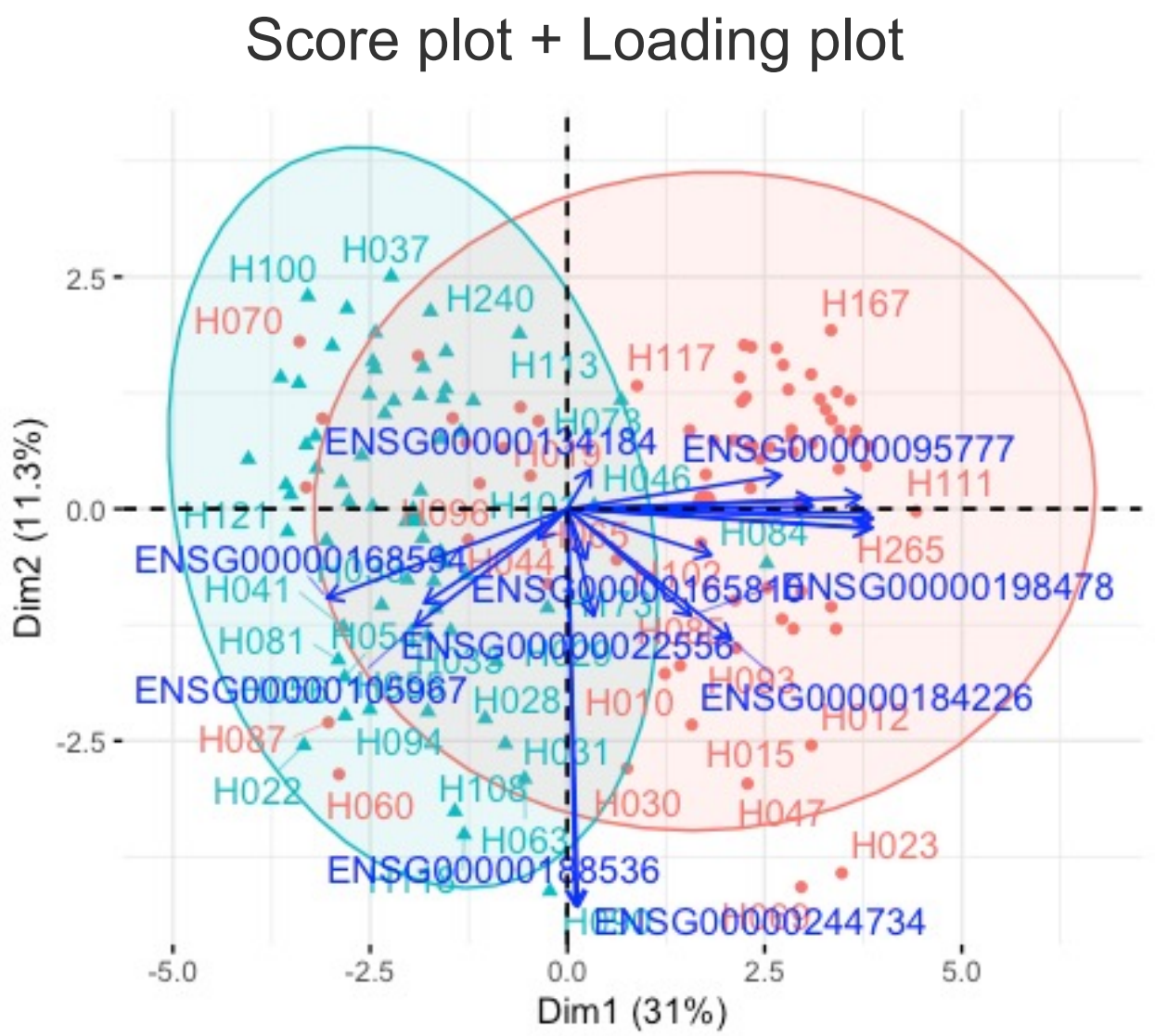
$$Z_2 = w_{21}Y_1 + w_{22}Y_2 + \cdots + w_{2p}Y_p$$





Relationship between given features

PCA biplot


Score plot + Loading plot
Dim2 (11.3%)
Dim1 (31%)
2.5
0.0
-2.5
-5.0
-2.5
0.0
2.5
5.0
H100
H037
H070
H240
H167
H117
H113
ENSG00000134184
ENSG00000095777
H046
H111
H121
H084
H265
ENSG00000168594
H041
ENSG00000198478
H081
ENSG00000022556
ENSG00000105967
H028
H010
ENSG00000184226
H012
H087
H094
H031
H015
H022
H060
H108
H030
H047
H023
H063
ENSG00000188536
ENSG00000244734

Exercise PCA

1. Load the `nutrimouse` data from the `mixOmics` R package and investigate its structure.

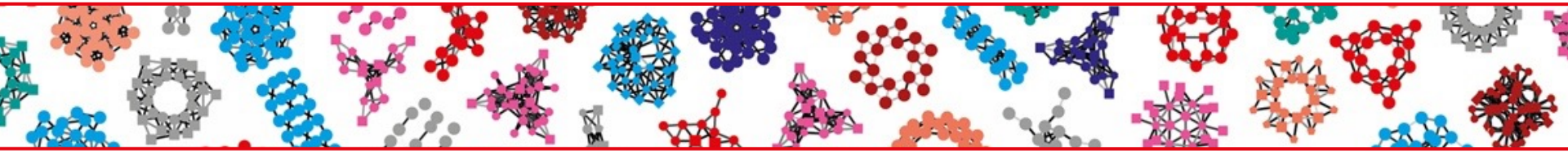
2. Take the gene expression dataset in *samples* x *variables* matrix format. Investigate their distribution.

3. Perform PCA and investigate variances, sample distribution and variable relationship with plots.

4. Visually investigate the sample distribution with coloring by metadata or expression of certain genes.

5. (Optional) PCA on lipidomic dataset

Overview



01 Principal Component Analysis

02 Partial Least Squares

03 Canonical Correlation Analysis

04 Towards Nonlinearity

Questions in practice

- **A real-estate agency wants to know why some properties were sold quickly and some not.**
  - Properties' characteristics: price, surface area, form, floors, bedrooms, bathrooms, entrances, garages, yards, etc; Access to nearby facilities; Local living cost
  - Properties' sold status

- **Relation between lifestyle and clinical measurements of patients**
  - Lifestyle
  - Clinical measurements

What is PLS?

Herman Wold – econometrician (1966)

Two sets of variables (features) to consider

- Many (predictors or independent/explanatory variables) vs One (response or dependent variable)
- Many vs Many

Reduce the dimension of the two variable spaces

- *Feature extraction*: project high-dimensional space into a space of fewer dimensions

Find the relation between the two sets of variables: covariance

PLS: also Projection to Latent Structures

When is PLS used?

- Cannot identify variables to eliminate
- Need *new* variables independent of one another
- Accept that the *new* independent variables are less interpretable

$\mathbf{X}$: n samples $\times$ p variables $\mathbf{Y}$: n samples $\times$ q variables

Samples

How is variation in both response and predictor

Variables

What underlying factors explain both variation?

Predict variables in $\mathbf{Y}$ using variables in $\mathbf{X}$

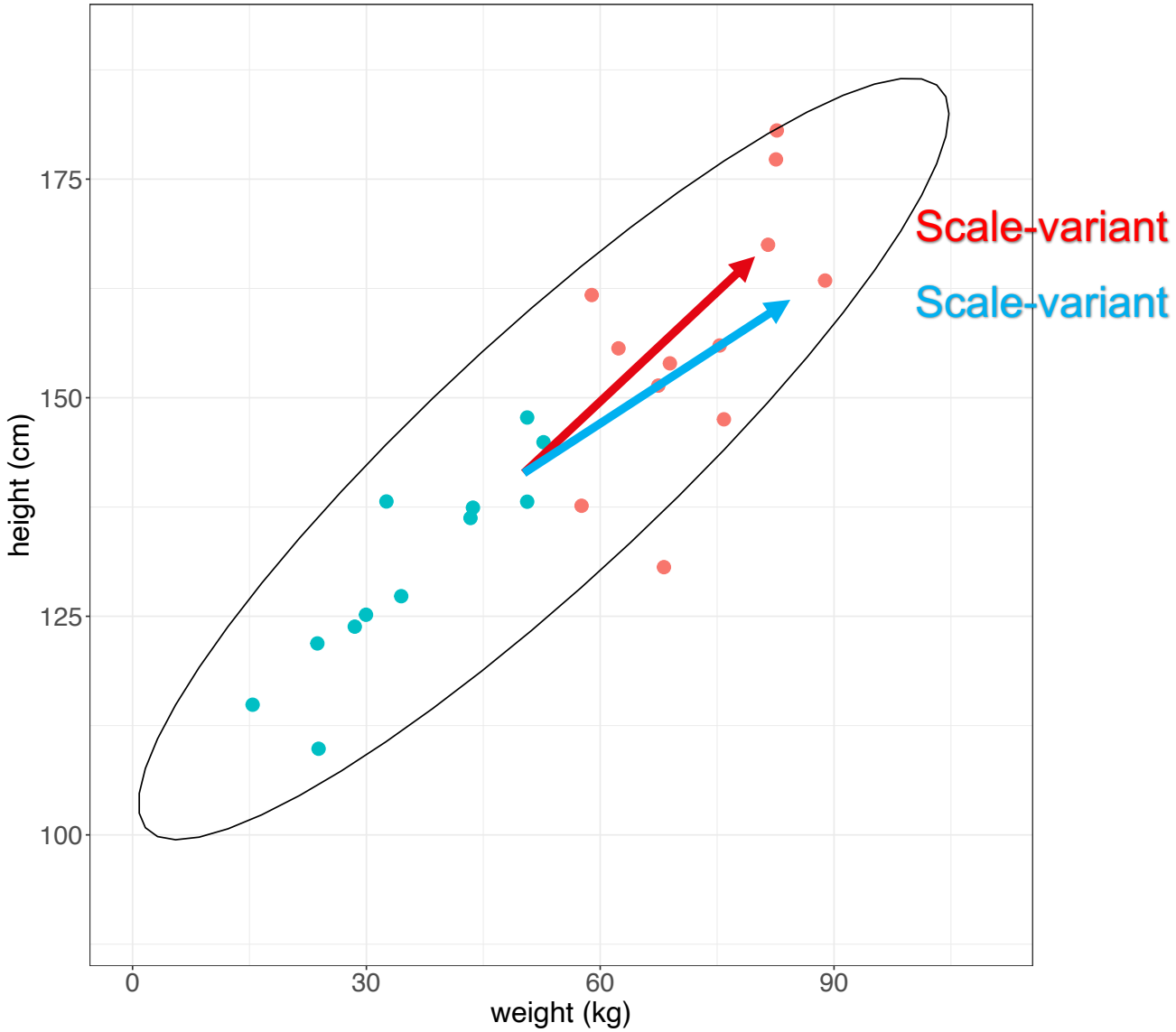
PLS Discriminant Analysis (PLS-DA)

$\mathbf{X}$: n samples $\times$ p variables

$\mathbf{Y}$: n samples $\times$ 1 categorical variable (labels)

- Supervised version of PCA
- Dimensionality reduction, feature selection, classification
- Maximize covariance between each latent variable and the labelling

PCA and PLS-DA



How does PLS work?

- **PCA objective: calculate latent variables**
  - best explaining variance in **X**

  $\Rightarrow$ Maximize variance in **X** latent variables

- **PLS objective: calculate latent variables**
  - best explaining variance in **X**
  - best explaining variance in **Y**
  - having greatest relationship between **X** and **Y**

  $\Rightarrow$ Maximize covariance between **X** and **Y** latent variables

Components 1

Coordinate change:

$$(\boldsymbol{X}_1\ \boldsymbol{X}_2 \dots\ \boldsymbol{X}_p) \rightarrow (\boldsymbol{U}_1\ \boldsymbol{U}_2 \dots\ \boldsymbol{U}_p)$$
$$(\boldsymbol{Y}_1\ \boldsymbol{Y}_2 \dots\ \boldsymbol{Y}_q) \rightarrow (\boldsymbol{V}_1\ \boldsymbol{V}_2 \dots\ \boldsymbol{V}_q)$$

Linear combination:

$$\boldsymbol{U}_1 = w_{11}\boldsymbol{X}_1 + w_{12}\boldsymbol{X}_2 + \cdots + w_{1p}\boldsymbol{X}_p = \mathbf{X}\boldsymbol{w}_1$$
$$\boldsymbol{V}_1 = c_{11}\boldsymbol{Y}_1 + c_{12}\boldsymbol{Y}_2 + \cdots + c_{1q}\boldsymbol{Y}_q = \mathbf{Y}\boldsymbol{c}_1$$

=> projection of $\mathbf{X}$ onto vector $\boldsymbol{w}_1$ and $\mathbf{Y}$ onto vector $\boldsymbol{c}_1$

Find $\boldsymbol{w}_1, \boldsymbol{c}_1$ ($\|\boldsymbol{w}_1\| = \|\boldsymbol{c}_1\| = \boldsymbol{w}_1^{\mathrm{T}}\ \boldsymbol{w}_1 = \boldsymbol{c}_1^{\mathrm{T}}\boldsymbol{c}_1 = 1$) to maximize:

$$\mathrm{cov}(\boldsymbol{U}_1, \boldsymbol{V}_1) = \mathrm{cor}(\boldsymbol{U}_1, \boldsymbol{V}_1)\sqrt{var(\boldsymbol{U}_1)}\sqrt{var(\boldsymbol{V}_1)}$$

- best explaining variance in **X**, given by $\sqrt{var(\boldsymbol{U}_1)} = \boldsymbol{U}_1^{\mathrm{T}}\ \boldsymbol{U}_1$
- best explaining variance in **Y**, given by $\sqrt{var(\boldsymbol{V}_1)} = \boldsymbol{V}_1^{\mathrm{T}}\ \boldsymbol{V}_1$
- having greatest relationship between **X** and **Y**, given by $\mathrm{cor}(\boldsymbol{U}_1, \boldsymbol{V}_1)$

Following components

- X-space: component 2 orthogonal to component 1
- Y-space:
  - Regression mode: not necessarily orthogonal ($\mathbf{Y}$ is deflated to the information from $\mathbf{X}$)
  - Canonical mode: orthogonal ($\mathbf{Y}$ is deflated to the information from $\mathbf{Y}$)

- **Algorithm**: iterative process
  - For component 1:

$\boldsymbol{w}_1$= eigenvector corresponding to max eigenvalue of $\mathbf{X^T Y Y^T X}$
$\boldsymbol{c}_1$= eigenvector corresponding to max eigenvalue of $\mathbf{Y^T X X^T Y}$

  - For component 2:

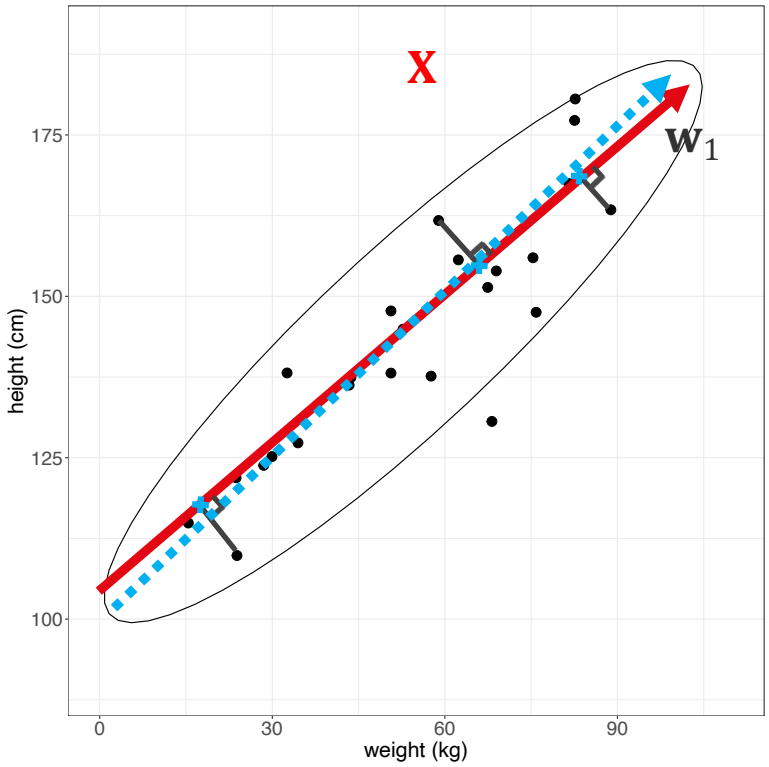
$\boldsymbol{w}_2$= eigenvector corresponding to max eigenvalue of
$\mathbf{X}^{\mathbf{T}}_{\mathbf{deflated2}}\mathbf{Y}_{\mathbf{deflated2}}\mathbf{Y}^{\mathbf{T}}_{\mathbf{deflated2}}\mathbf{X}_{\mathbf{deflated2}}$

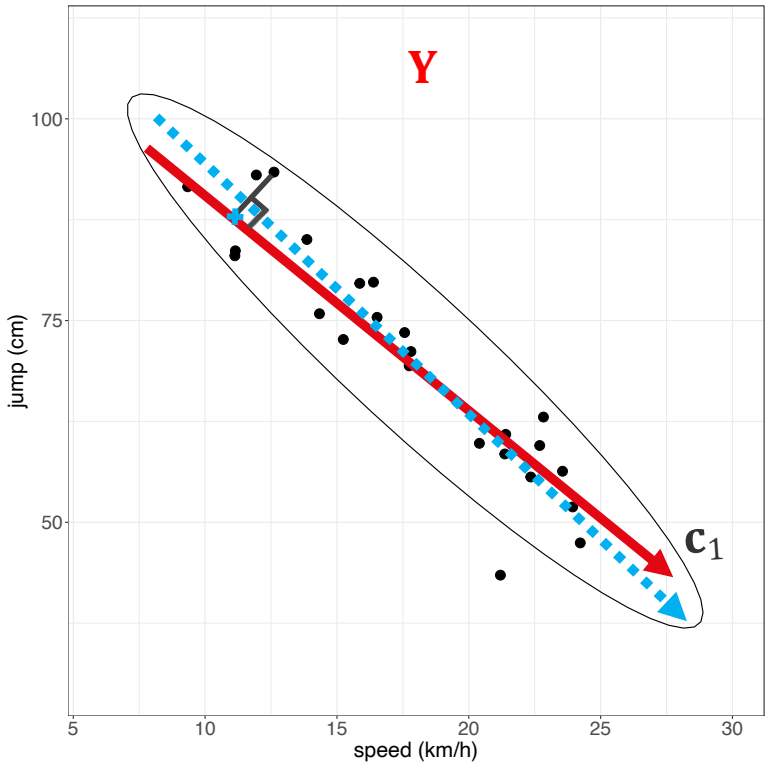
$\mathbf{X}_{\mathbf{deflated1}} = \mathbf{X}, \mathbf{X}_{\mathbf{deflated2}} = \mathbf{X} - \boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}}\mathbf{X}_{\mathbf{deflated1}}$, etc.

  - And so on, at most $\min(p, q)$

Scores

- Maximum covariance between X-space scores (e.g. projection of $\mathbf{X}$ onto vector $\boldsymbol{w}_1$) and Y-space scores (e.g. $\mathbf{Y}$ onto vector $\boldsymbol{c}_1$)
- Not necessarily identical to PCA components => affected scores
- Visualize clusters, outliers, interesting patterns in sample distribution





Loadings

- Highly correlated variables: similar weights in the loading vectors; close together in the loading plots of **all** dimensions.
- Loading plot: superimpose loading plots from X and Y

$\Rightarrow$ Relationship between X variables, between Y variables, between all variables

- **Loadings**: weights on *deflated (residual)* matrices
- **Loadings-star**: weights on input matrices

$$\boldsymbol{U}_1 = \mathbf{X}\boldsymbol{w}_\mathbf{1}^* = \mathbf{X}_\mathbf{deflated1}\boldsymbol{w}_1: \quad \mathbf{X}_\mathbf{deflated1} = \mathbf{X}, \quad \boldsymbol{w}_\mathbf{1}^* = \boldsymbol{w}_1$$
$$\boldsymbol{U}_2 = \mathbf{X}\boldsymbol{w}_\mathbf{2}^* = \mathbf{X}_\mathbf{deflated2}\boldsymbol{w}_2: \quad \mathbf{X}_\mathbf{deflated2} = \mathbf{X} - \boldsymbol{U}_1\boldsymbol{U}_1^{\mathrm{T}}\mathbf{X}_\mathbf{deflated1}, \boldsymbol{w}_\mathbf{2}^* \neq \boldsymbol{w}_2$$

=> Interpret **Loadings-star** rather than Loadings when investigating relationships in PLS

Variability explained with components

- Cumulative variance explained for each space

$$R^2_{\mathbf{X}} = 1 - \frac{\mathrm{var}\left(\mathbf{X}_{deflated}\right)}{\mathrm{var}(\mathbf{X})}$$

$$R^2_{\mathbf{Y}} = 1 - \frac{\mathrm{var}(\mathbf{Y}_{deflated})}{\mathrm{var}(\mathbf{Y})}$$

- Plot of $R^2_{\mathbf{X}}$ and $R^2_{\mathbf{Y}}$ for each variable

$$R^2_{\mathbf{X},k} = 1 - \frac{\mathrm{var}\left(\mathbf{X}_{deflated,k}\right)}{\mathrm{var}(\mathbf{X}_k)}$$

$$R^2_{\mathbf{Y},k} = 1 - \frac{\mathrm{var}(\mathbf{Y}_{deflated,k})}{\mathrm{var}(\mathbf{Y}_k)}$$

PLS implementation

$\mathbf{X}$: n samples × p variables $\mathbf{Y}$: n samples × q variables

Input: $\mathbf{X}, \mathbf{Y}$

Output:

- cor: correlations
- variates:$(\boldsymbol{U}_1\ \boldsymbol{U}_2 \ldots\ \boldsymbol{U}_{ncomp}), (\boldsymbol{V}_1\ \boldsymbol{V}_2 \ldots\ \boldsymbol{V}_{ncomp})$
- loadings: $(\boldsymbol{w}_1\ \boldsymbol{w}_2\ \ldots\ \boldsymbol{w}_{ncomp}), (\boldsymbol{c}_1\ \boldsymbol{c}_2\ \ldots\ \boldsymbol{c}_{ncomp})$
- loadings-star
- proportion of explained variance
- correlation between variates and input data

Orthogonal PLS

Johan Trygg and Svante Wold (2002)

Orthogonal: removes variation from predictors $\mathbf{X}$ that is

not correlated to responses $\mathbf{Y}$

- Maximize explained variance on the first component(s) of predictors
- Remaining components capturing variance that is orthogonal to responses

$\Rightarrow$ Model separately variations of $\mathbf{X}$ correlated and uncorrelated to $\mathbf{Y}$

$$\mathbf{X} = \mathbf{T}_p\mathbf{P}_p^{\mathbf{T}} + \mathbf{T}_o\mathbf{P}_o^{\mathbf{T}} + \mathbf{E}$$

Y = OPC *versus in vitro*

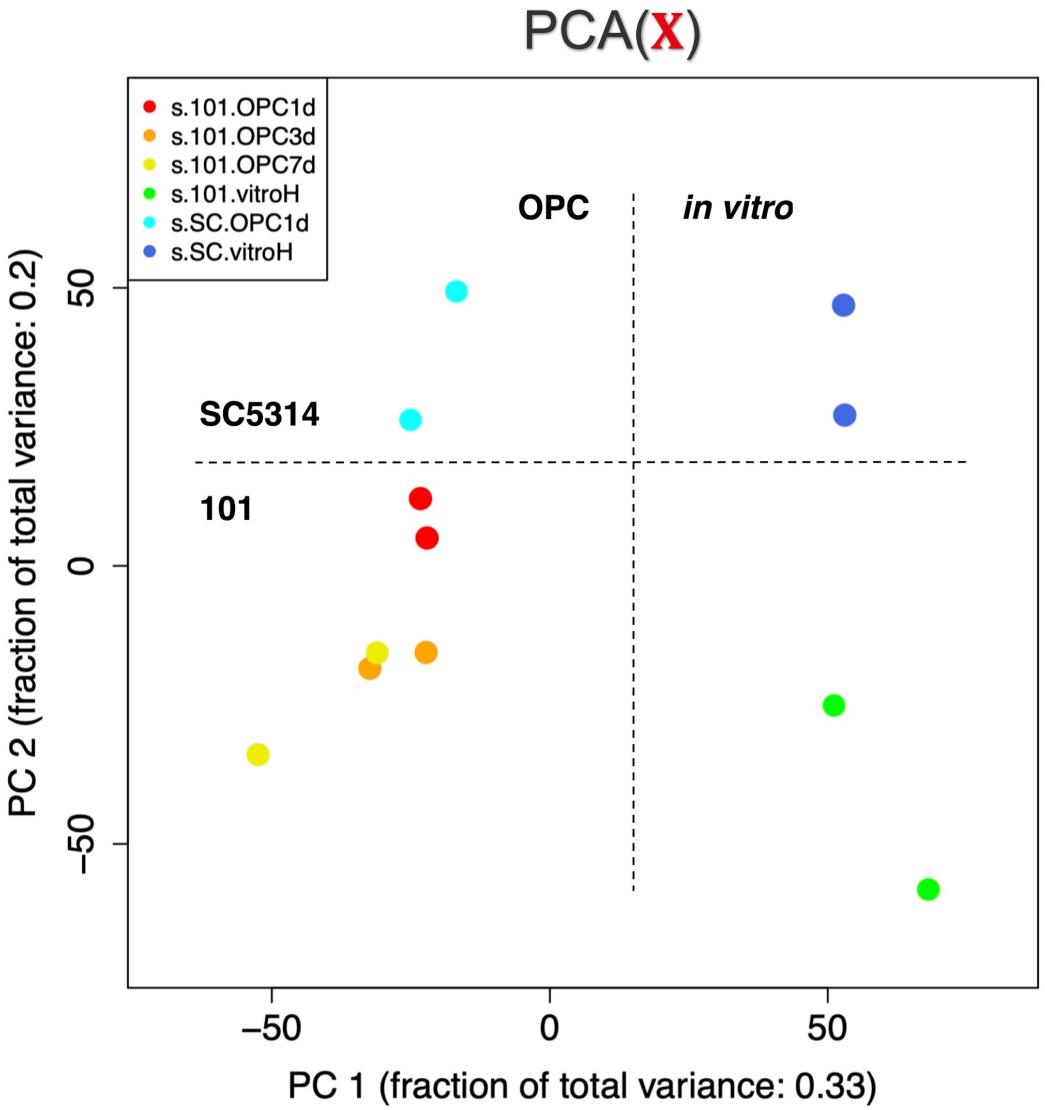
Predictive component ≅ PC1

Orthogonal component ≅ PC2

Y = 101 *versus* SC5314

Predictive component ≅ PC2

Orthogonal component ≅ PC1



Orthogonal PLS

- Reduce model complexity: lower the number of latent variables
- Allow identification and investigation of the source of orthogonal variation
- Interpret more easily
- Produce more efficient predictive model, particularly when structured noise dominates

Exercise PLS

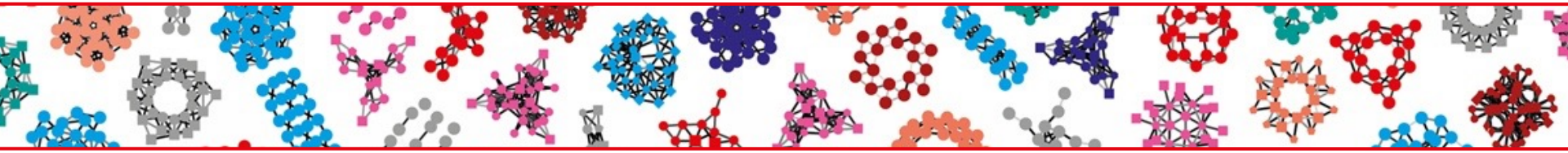
1. Perform PLS (`mixOmics::pls`) between `gene` and `lipid`. Investigate its output, sample distribution and variable relationship with plots.

2. Observe the difference between the two modes *regression* and *canonical* of PLS.

3. Perform PLS-DA (`mixOmics::plsda`) between `gene` and `genotype`. Redo PLS-DA using `mixOmics::pls` and compare the results.

4. Perform OPLS-DA (`ropls::opls`) between gene and genotype. Investigate its output, sample distribution, variable relationship and predictive performance.

Overview



01 Principal Component Analysis

02 Partial Least Squares

03 Canonical Correlation Analysis

04 Towards Nonlinearity

Questions in practice

- **Relation between lifestyle and clinical measurements of patients**
  - Lifestyle
  - Clinical measurements
- **Relation between two omics profiles**

CCA versus PLS

- PLS

$$\max_{w,c} \operatorname{cov}(\mathbf{X}\boldsymbol{w}, \mathbf{Y}\boldsymbol{c})$$

$$\text{subject to } \|\boldsymbol{w}\| = \|\boldsymbol{c}\| = \mathbf{1}$$

- CCA (Harold Hotelling – statistician/economic theorist (1936))

$$\max_{w,c} \operatorname{cor}(\mathbf{X}\boldsymbol{w}, \mathbf{Y}\boldsymbol{c})$$

$$\text{subject to } \operatorname{var}(\mathbf{X}\boldsymbol{w}) = \operatorname{var}(\mathbf{Y}\boldsymbol{c}) = \mathbf{1}$$

Components 1

Coordinate change:

$$(\boldsymbol{X}_1\ \boldsymbol{X}_2 \dots\ \boldsymbol{X}_p) \rightarrow (\boldsymbol{U}_1\ \boldsymbol{U}_2 \dots\ \boldsymbol{U}_p)$$
$$(\boldsymbol{Y}_1\ \boldsymbol{Y}_2 \dots\ \boldsymbol{Y}_q) \rightarrow (\boldsymbol{V}_1\ \boldsymbol{V}_2 \dots\ \boldsymbol{V}_q)$$

Linear combination:

$$\boldsymbol{U}_1 = w_{11}\boldsymbol{X}_1 + w_{12}\boldsymbol{X}_2 + \cdots + w_{1p}\boldsymbol{X}_p = \mathbf{X}\boldsymbol{w}_1$$
$$\boldsymbol{V}_1 = c_{11}\boldsymbol{Y}_1 + c_{12}\boldsymbol{Y}_2 + \cdots + c_{1q}\boldsymbol{Y}_q = \mathbf{Y}\boldsymbol{c}_1$$

=> projection of $\mathbf{X}$ onto vector $\boldsymbol{w}_1$ and $\mathbf{Y}$ onto vector $\boldsymbol{c}_1$

Find $\boldsymbol{w}_1, \boldsymbol{c}_1$ ($\mathrm{var}(\mathbf{X}\boldsymbol{w}_1) = \mathrm{var}(\mathbf{Y}\boldsymbol{c}_1) = \boldsymbol{w}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}_1 = \boldsymbol{c}_1^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\boldsymbol{c}_1 = 1$) to maximize:

$$\mathrm{cor}(\boldsymbol{U}_1, \boldsymbol{V}_1)$$

How does CCA works?

Solution:

max $\mathrm{cor}(\boldsymbol{U}_1, \boldsymbol{V}_1)$ = *sqrt*(max eigenvalue of $\mathrm{cov}(\mathbf{X})^{-1}\mathrm{cov}(\mathbf{X},\mathbf{Y})\mathrm{cov}(\mathbf{Y})^{-1}\mathrm{cov}(\mathbf{Y},\mathbf{X})$)

= *sqrt*(max eigenvalue of $\mathrm{cov}(\mathbf{Y})^{-1}\mathrm{cov}(\mathbf{Y},\mathbf{X})\mathrm{cov}(\mathbf{X})^{-1}\mathrm{cov}(\mathbf{X},\mathbf{Y})$)

at $\boldsymbol{w}_1$, $\boldsymbol{c}_1$ = corresponding eigenvectors

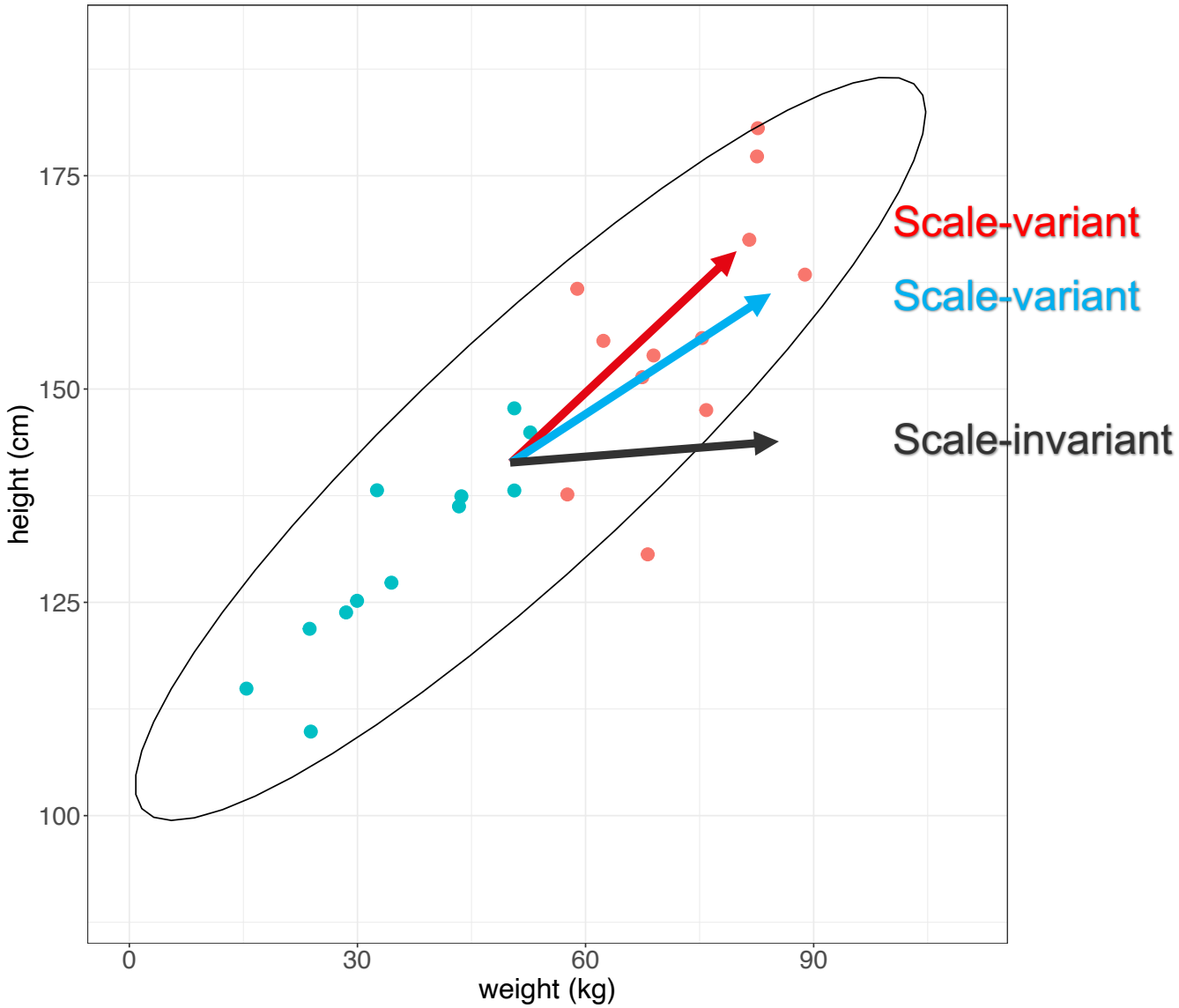
Hint: Lagrange multiplier + derivative

max $\mathrm{cor}(\boldsymbol{U}_2, \boldsymbol{V}_2)$ = 2nd max eigenvalue

at $\boldsymbol{w}_2$, $\boldsymbol{c}_2$ = corresponding eigenvectors

and so on

CCA versus PLS versus PCA



CCA implementation

$\mathbf{X}$: n samples × p variables $\mathbf{Y}$: n samples × q variables

Input: $\mathbf{X}, \mathbf{Y}$

Output:

- cor: correlations
- variates:$(\boldsymbol{U}_1\ \boldsymbol{U}_2 \ldots\ \boldsymbol{U}_{ncomp}), (\boldsymbol{V}_1\ \boldsymbol{V}_2 \ldots\ \boldsymbol{V}_{ncomp})$
- loadings: $(\boldsymbol{w}_1\ \boldsymbol{w}_2\ \ldots\ \boldsymbol{w}_{ncomp}), (\boldsymbol{c}_1\ \boldsymbol{c}_2\ \ldots\ \boldsymbol{c}_{ncomp})$
- proportion of explained variance
- correlation between variates and input data

Regularized CCA

Solution:

max $\mathrm{cor}(\boldsymbol{U}_1, \boldsymbol{V}_1)$ = *sqrt*(max eigenvalue of $\mathrm{cov}(\mathbf{X})^{-1}\mathrm{cov}(\mathbf{X},\mathbf{Y})\mathrm{cov}(\mathbf{Y})^{-1}\mathrm{cov}(\mathbf{Y},\mathbf{X})$)

= *sqrt*(max eigenvalue of $\mathrm{cov}(\mathbf{Y})^{-1}\mathrm{cov}(\mathbf{Y},\mathbf{X})\mathrm{cov}(\mathbf{X})^{-1}\mathrm{cov}(\mathbf{X},\mathbf{Y})$)

What if $\mathrm{cov}(\mathbf{X})$ or $\mathrm{cov}(\mathbf{Y})$ is not invertible? (*singularity* problem)

- Constant variables

$\Rightarrow$ remove

- Multicollinear variables (esp. when more variables than samples ($p > n$))

$\Rightarrow$ regularized CCA: $\mathrm{cov}(\mathbf{X}) \leftarrow \mathrm{cov}(\mathbf{X}) + \boldsymbol{\lambda}_{\mathbf{X}}\mathbf{I}$, $\mathrm{cov}(\mathbf{Y}) \leftarrow \mathrm{cov}(\mathbf{Y}) + \boldsymbol{\lambda}_{\mathbf{Y}}\mathbf{I}$

$\Rightarrow$ tuning for optimal values of $\boldsymbol{\lambda}_{\mathbf{X}}$, $\boldsymbol{\lambda}_{\mathbf{Y}}$

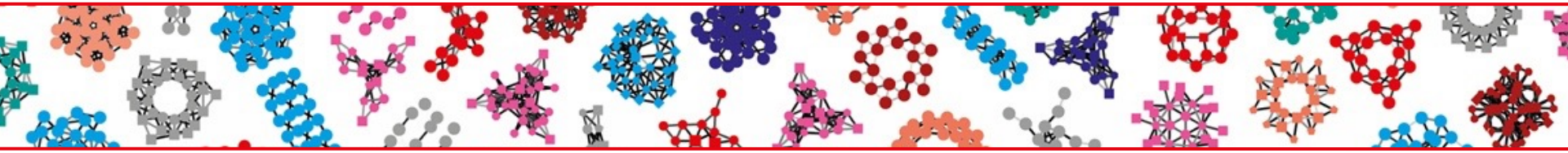
Exercise CCA

1. Perform CCA (`mixOmics::rcc`) between 20 genes and all lipids. Investigate correlations, sample distribution and variable relationship with plots.

2. Perform CCA with scaled datasets and observe the difference

3. Perform regularized CCA with all genes and lipids.

Overview



01 Principal Component Analysis

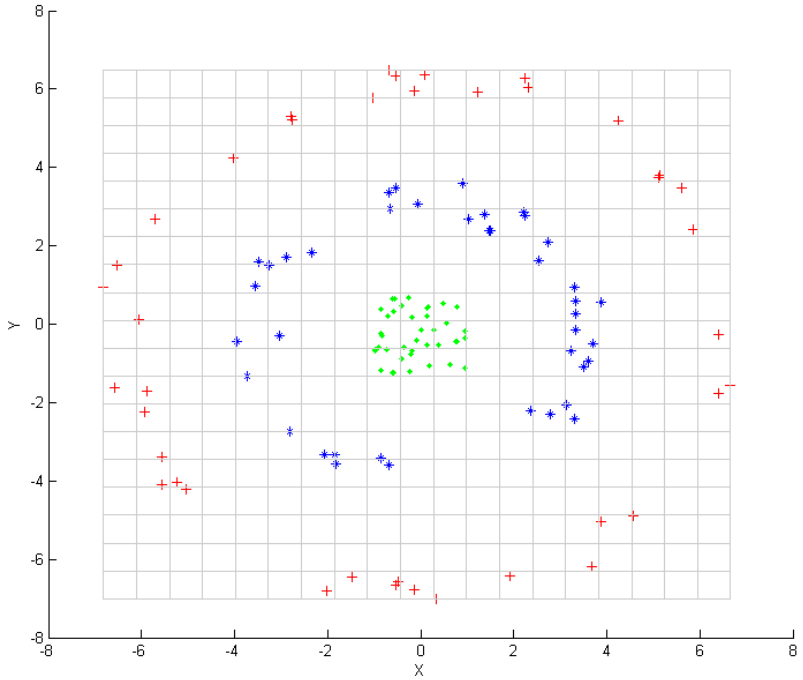
02 Partial Least Squares

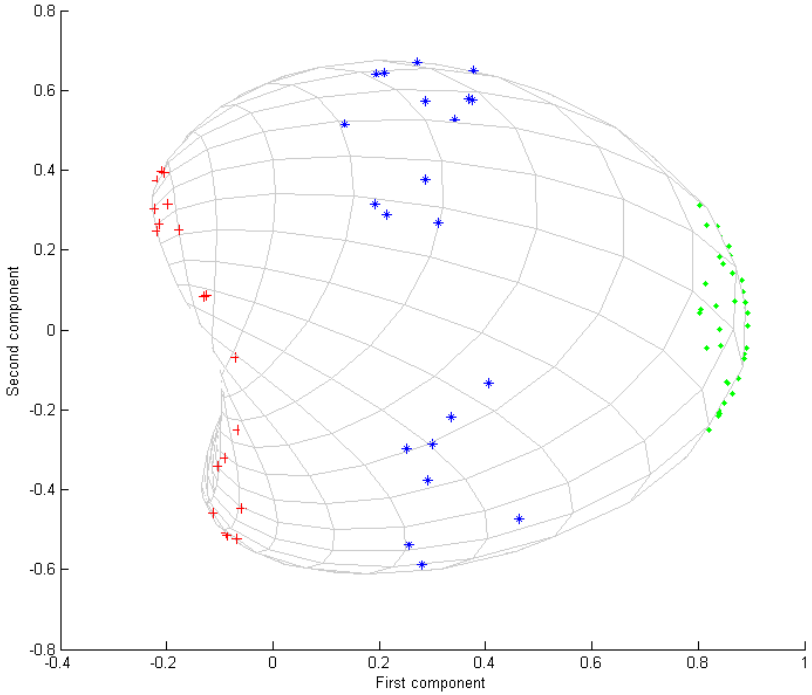
03 Canonical Correlation Analysis

04 Towards Nonlinearity

Towards Nonlinearity

Map data points that cannot be linearly separated into a space of higher dimension: Kernel PCA



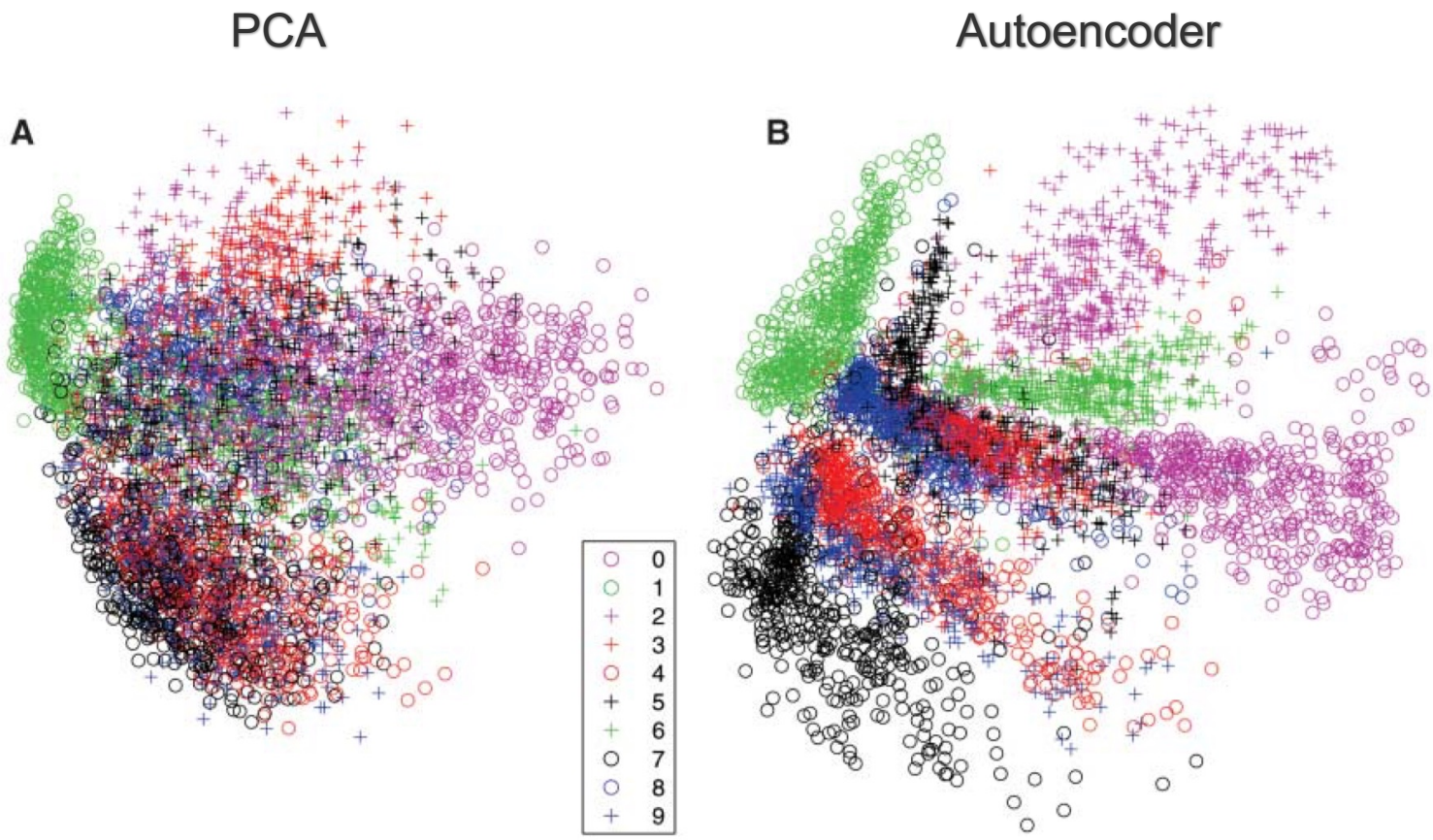


$$\boldsymbol{k(x,y)} = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$$

https://commons.wikimedia.org/wiki/File:Kernel_pca_input.png
https://commons.wikimedia.org/wiki/File:Kernel_pca_output.png
https://commons.wikimedia.org/wiki/File:Kernel_pca_output_gaussian.png

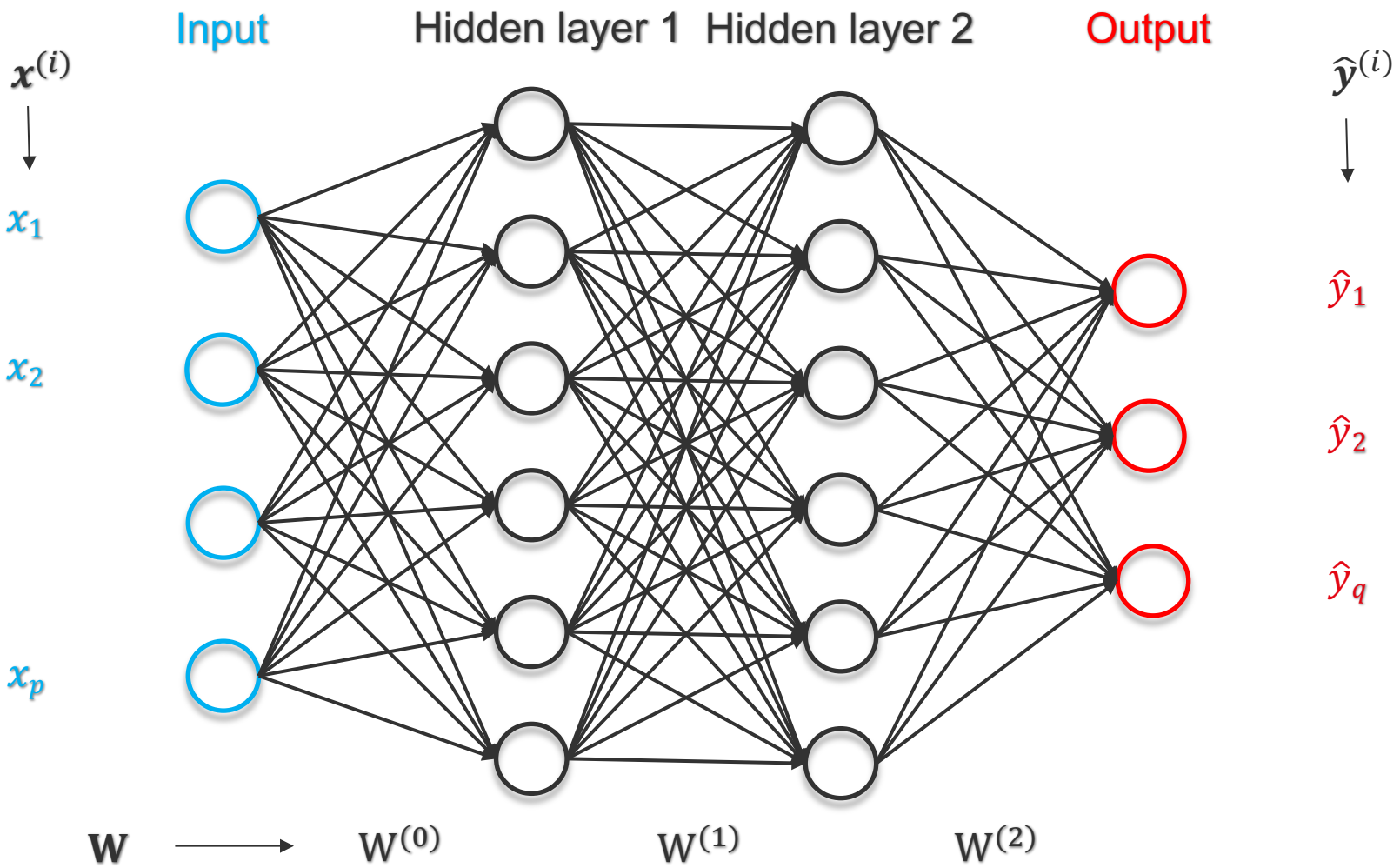
Towards Nonlinearity

Generalize principal components from straight lines to curves:
Autoencoder



Reducing the Dimensionality of
Data with Neural Networks, Hinton & Salakhutdinov, Science (2006)

Neural Networks

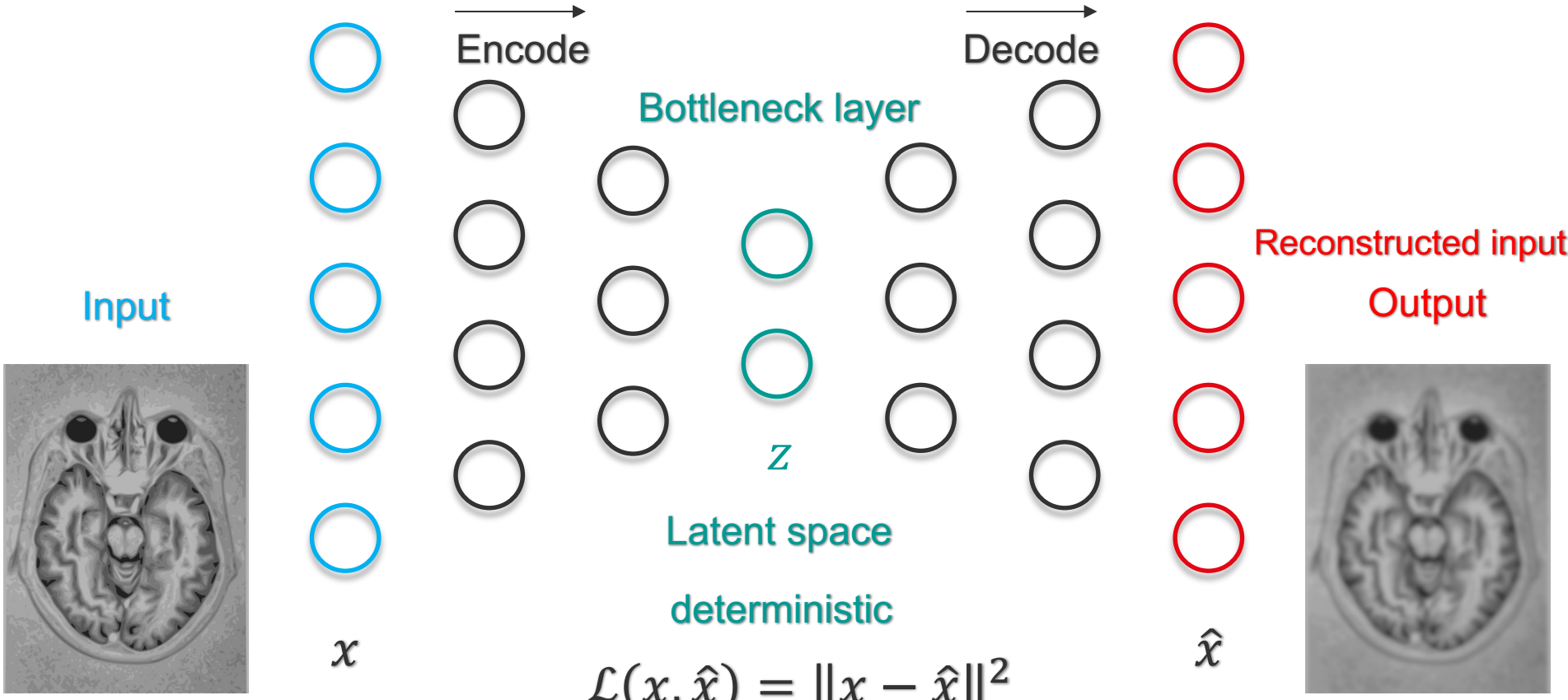


Activation functions on linear regressions

Loss optimization $\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(\boldsymbol{x}^{(i)}, \mathbf{W}), \boldsymbol{y}^{(i)})$

Autoencoder

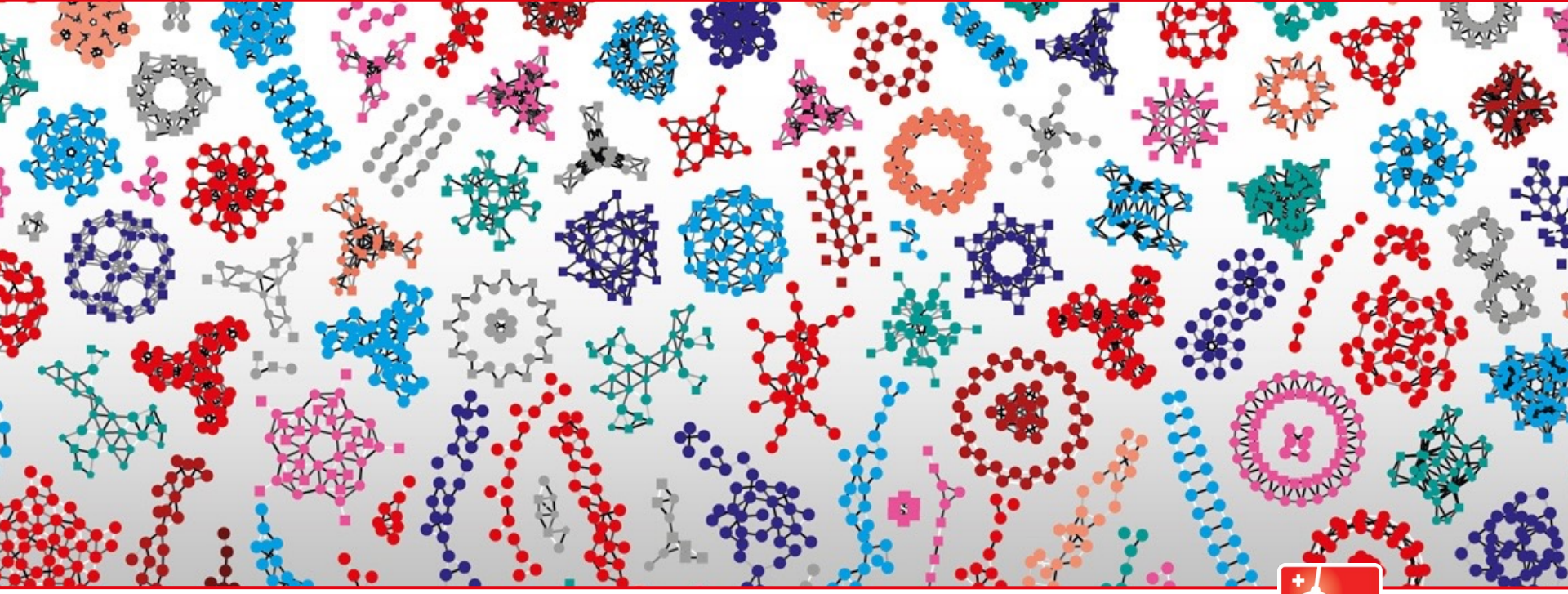
- Learning a lower-dimensional feature representation (compression) from unlabeled training data and learning a reconstruction back



Autoencoder application

- Image compression, denoising and generation, recommendation system, anomaly detection, feature extraction

- Life sciences
  - dimensionality reduction (clustering) in sequencing data
  - multi-omics and biomedical data integration



SIB
Swiss Institute of Bioinformatics

Thank you

elixir SWITZERLAND

www.sib.swiss