

Swiss Institute of
Bioinformatics

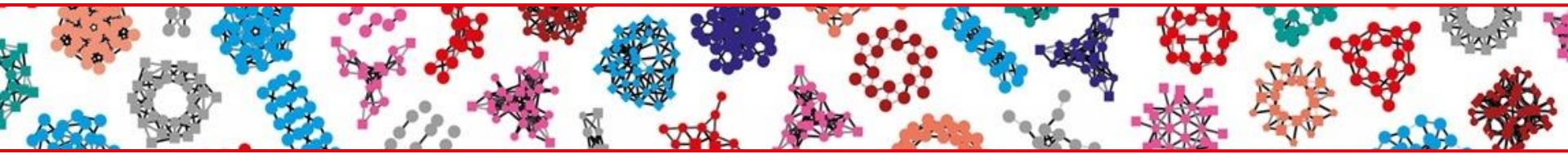
Dimensionality reduction

Van Du Tran

Vital-IT, SIB Swiss Institute of Bioinformatics

Nantes, 3-6 September 2025

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

Y: n samples \times p features

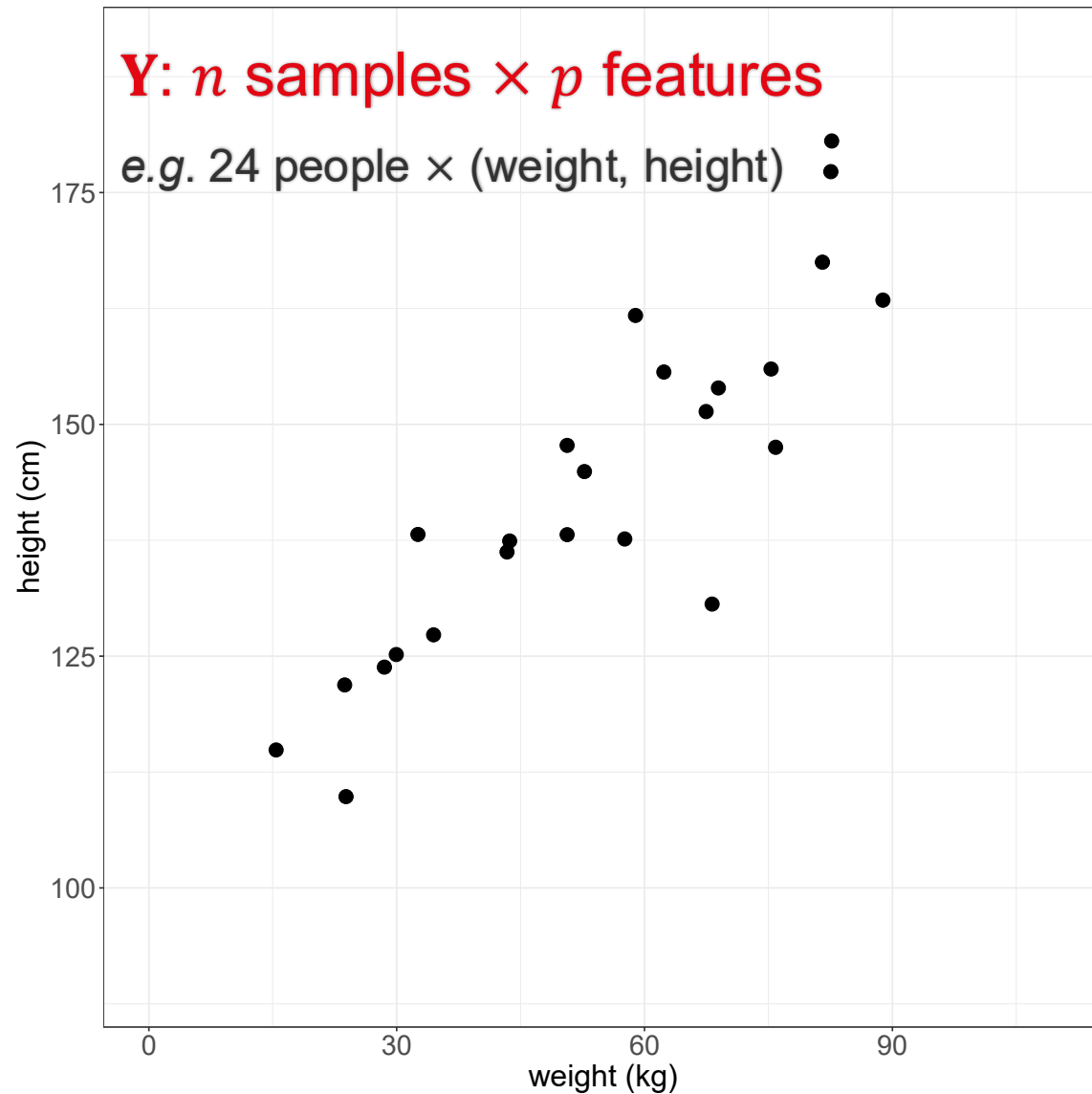
Samples

**How are samples
grouped together into
subgroups by
similarity?**

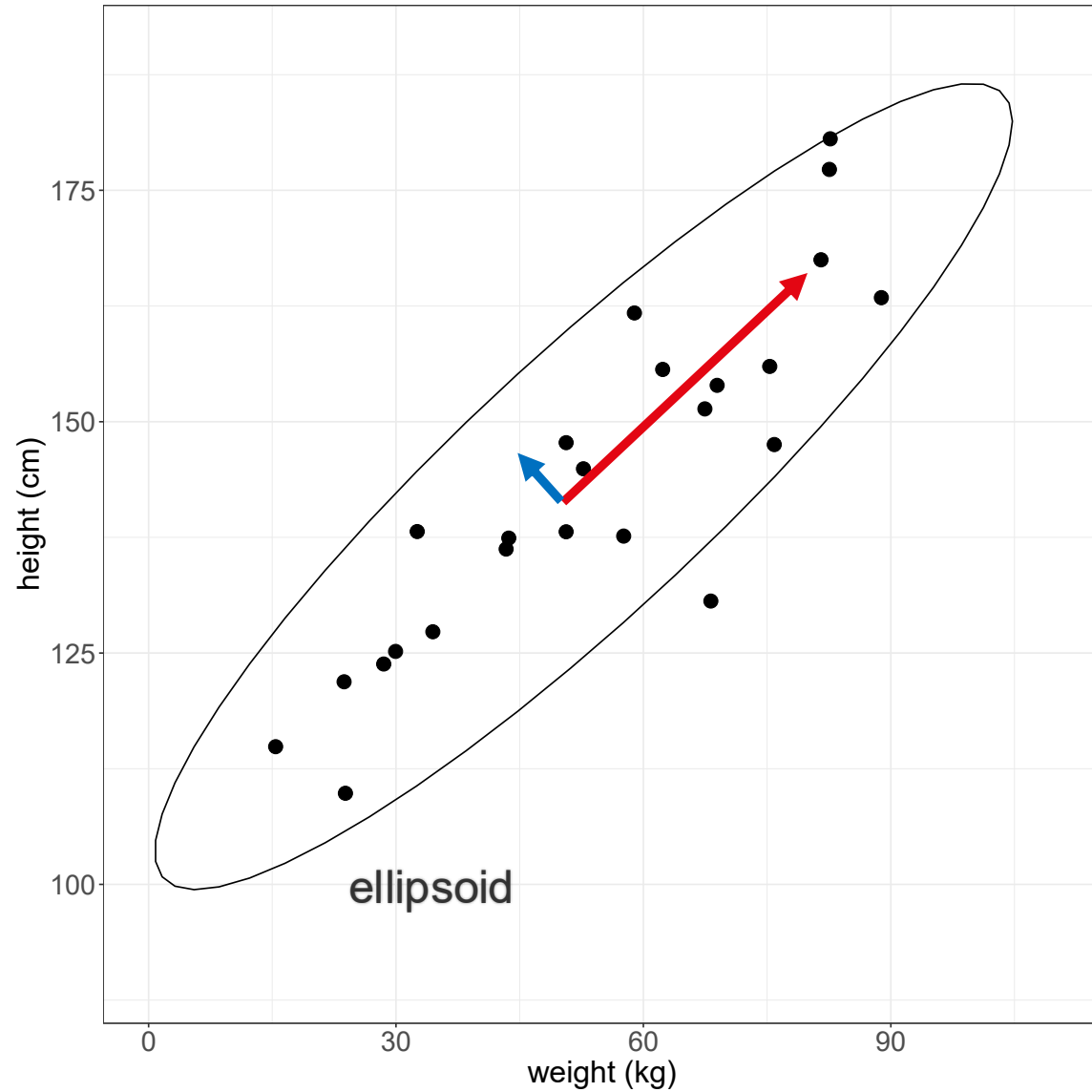
Features

**What underlying
factors influence the
grouping?**

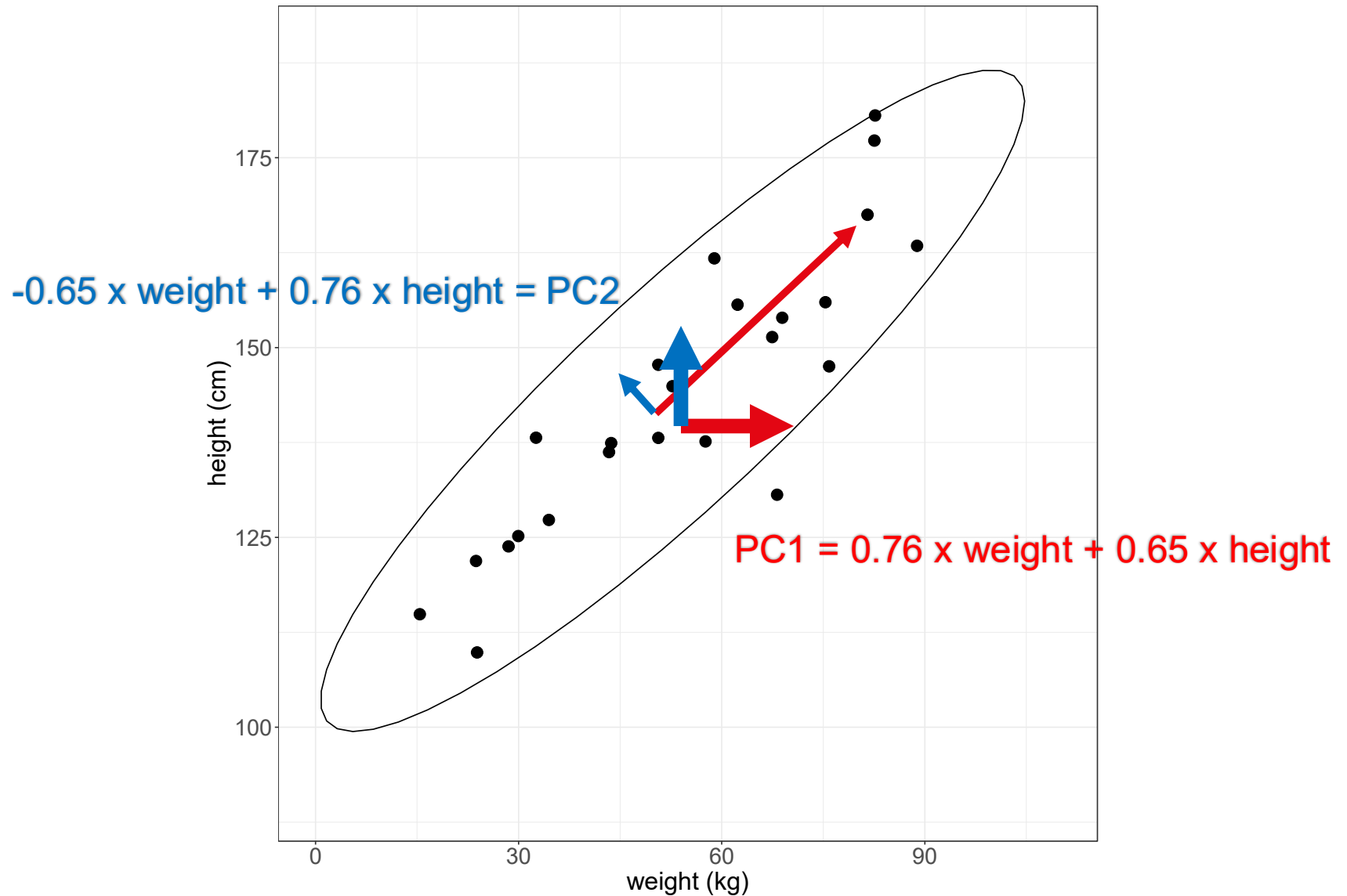
How does PCA work?



How does PCA work?



How does PCA work?



How does PCA work?

- $\mathbf{Y} = (\mathbf{Y}_1 \ \mathbf{Y}_2 \ \dots \ \mathbf{Y}_p)$

- **REQUIRED**: cleaning

- **OPTIONAL**: normalizing

- **REQUIRED**: centering

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

- **RECOMMENDED**: appropriately scaling (after centering)

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

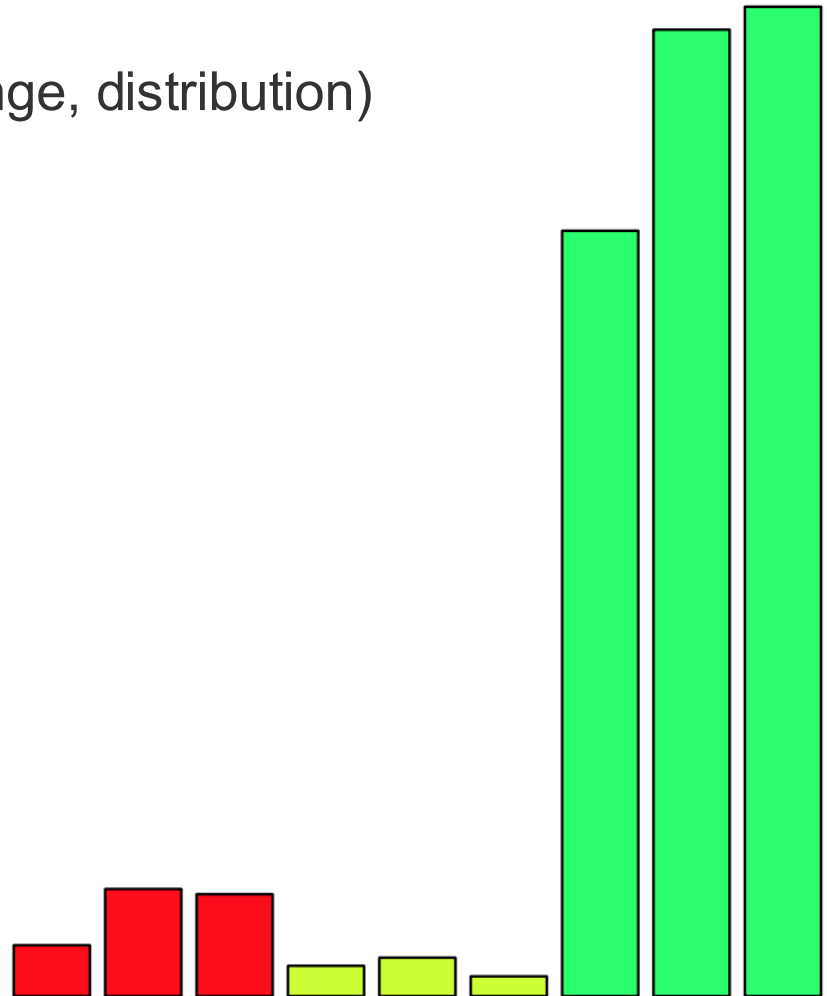
Data cleaning

GARBAGE IN, GARBAGE OUT

- Missing values: biologically or technically unidentified
⇒ *removal, imputation*
- Inconsistent data: qualitative, ill-formatted data
⇒ *reformatting, correction*
- Outlier and noisy data
⇒ *removal*
- Redundancy
⇒ *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

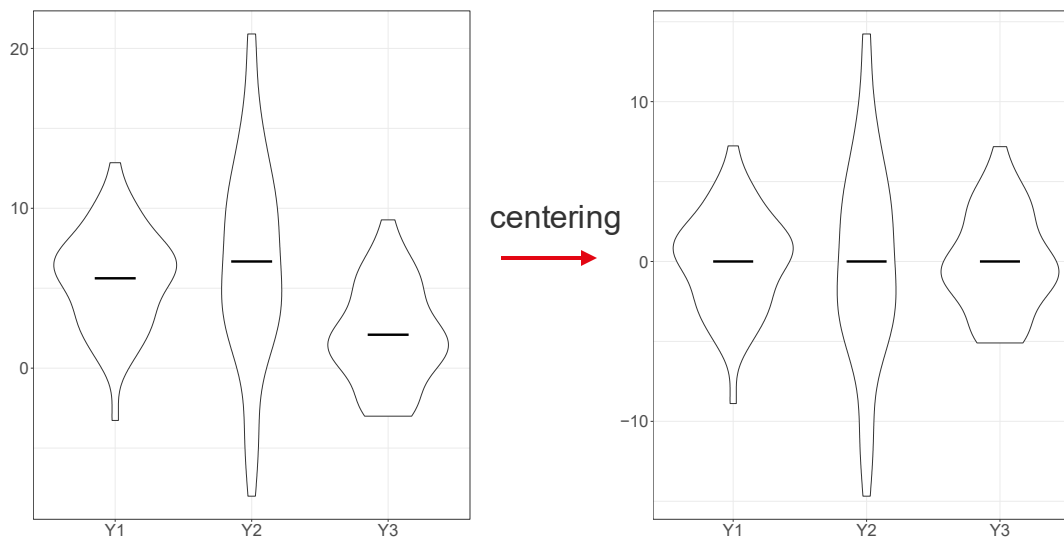
Each variable fluctuates around zero instead of its mean value

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

HOW: For each variable

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

$$\text{mean}(Y_i) = 0$$



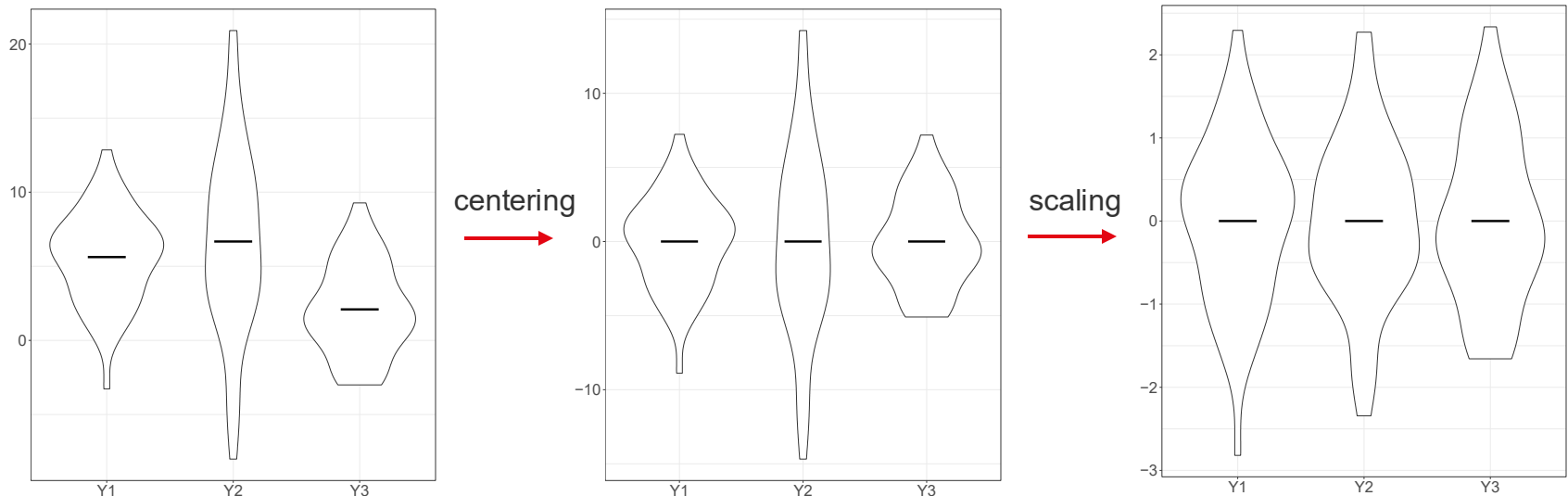
Data scaling

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 $var(Y_i) = 1$
- Pareto scaling: square root of standard deviation



Principal component

Coordinate change: $(Y_1 \ Y_2 \ \dots \ Y_p) \rightarrow (Z_1 \ Z_2 \ \dots \ Z_p)$

Linear combination:

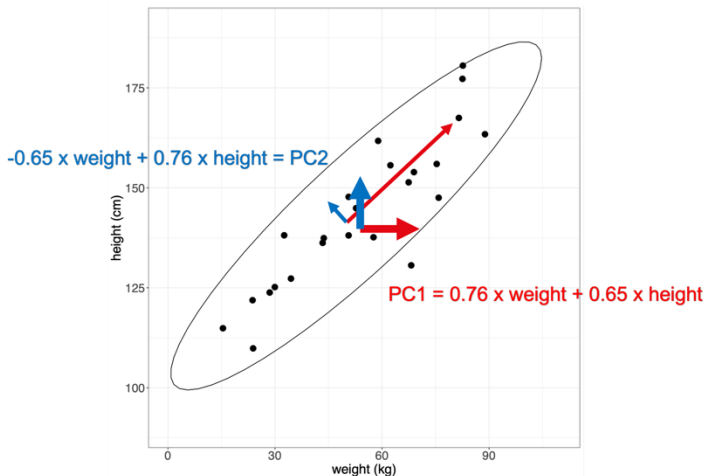
$$Z_1 = w_{11}Y_1 + w_{12}Y_2 + \dots + w_{1p}Y_p$$

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

...

$$Z_p = w_{p1}Y_1 + w_{p2}Y_2 + \dots + w_{pp}Y_p$$

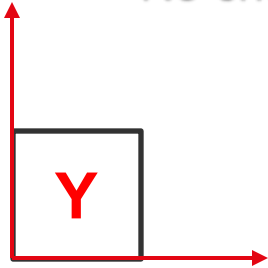
$$Z_1 = 0.8 Y_1 + 0.1 Y_2 + \dots + 0.001 Y_p$$



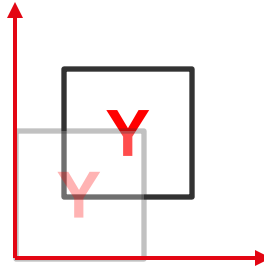
Latent variables

Linear transformation

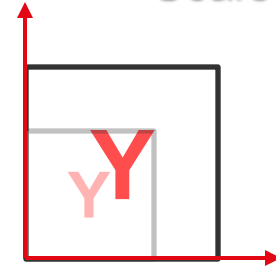
No change



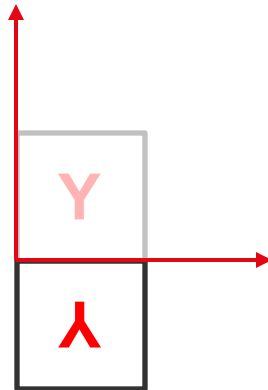
Translate



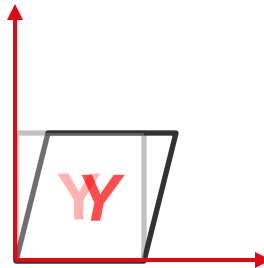
Scale



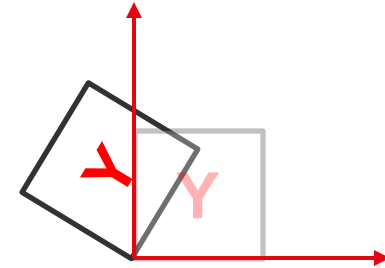
Reflect



Shear



Rotate



Principal component 1

Coordinate change: $(Y_1 \ Y_2 \dots Y_p) \rightarrow (Z_1 \ Z_2 \dots Z_p)$

Linear combination:

$$Z_1 = w_{11}Y_1 + w_{12}Y_2 + \dots + w_{1p}Y_p = \mathbf{Y}w_1$$

=> projection of \mathbf{Y} onto vector w_1

Find w_1 ($\|w_1\| = w_1^T w_1 = 1$) to maximize:

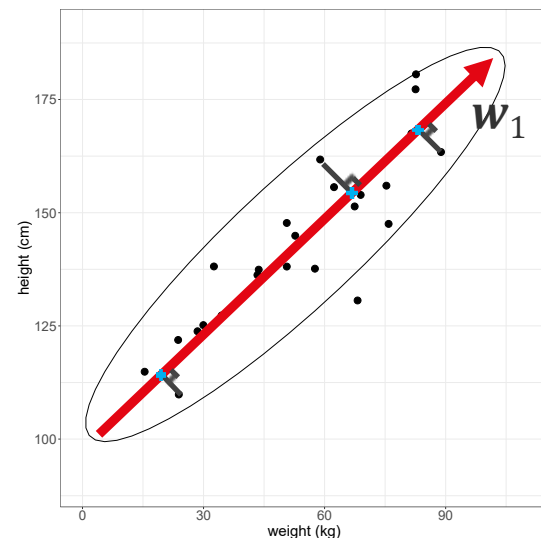
$$\text{var}(Z_1) = w_1^T \text{cov}(\mathbf{Y}) w_1$$

Variance-covariance matrix

$$\text{cov}(\mathbf{Y}) = \begin{pmatrix} \text{cov}(Y_1, Y_1) & \dots & \text{cov}(Y_1, Y_p) \\ \vdots & \ddots & \vdots \\ \text{cov}(Y_p, Y_1) & \dots & \text{cov}(Y_p, Y_p) \end{pmatrix}$$

Solution: $\max \text{var}(Z_1) = \max \text{eigenvalue of } \text{cov}(\mathbf{Y})$
at $w_1 = \text{corresponding eigenvector}$

Hint: Lagrange multiplier + derivative



Principal component 2

Coordinate change: $(Y_1 \ Y_2 \dots Y_p) \rightarrow (Z_1 \ Z_2 \dots Z_p)$

Linear combination:

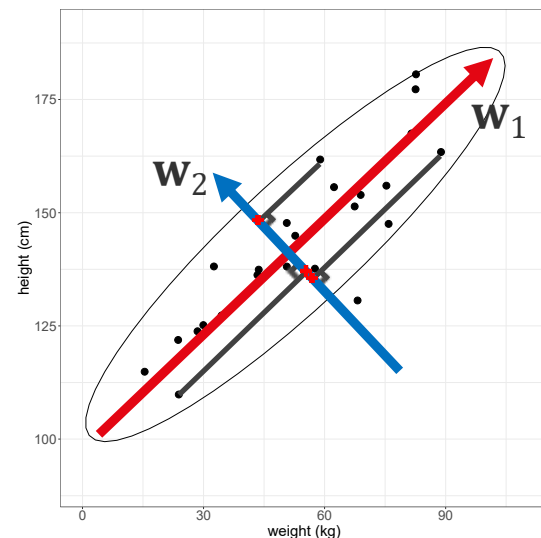
$$Z_2 = w_{21}Y_1 + w_{22}Y_2 + \dots + w_{2p}Y_p = \mathbf{Y}w_2$$

=> projection of \mathbf{Y} onto vector w_2

$w_2 \perp w_1$: independent projection

Find w_2 ($\|w_2\| = w_2^T w_2 = 1$ and $w_2^T w_1 = 0$) to maximize:

$$\text{var}(Z_2) = w_2^T \text{cov}(\mathbf{Y}) w_2$$



Solution: $\max \text{var}(Z_2) = 2^{\text{nd}} \text{ max eigenvalue of } \text{cov}(\mathbf{Y})$
at $w_2 = \text{corresponding eigenvector}$

and so on

PCA implementation

Y: n samples \times p features

Input: Y or cov(Y)

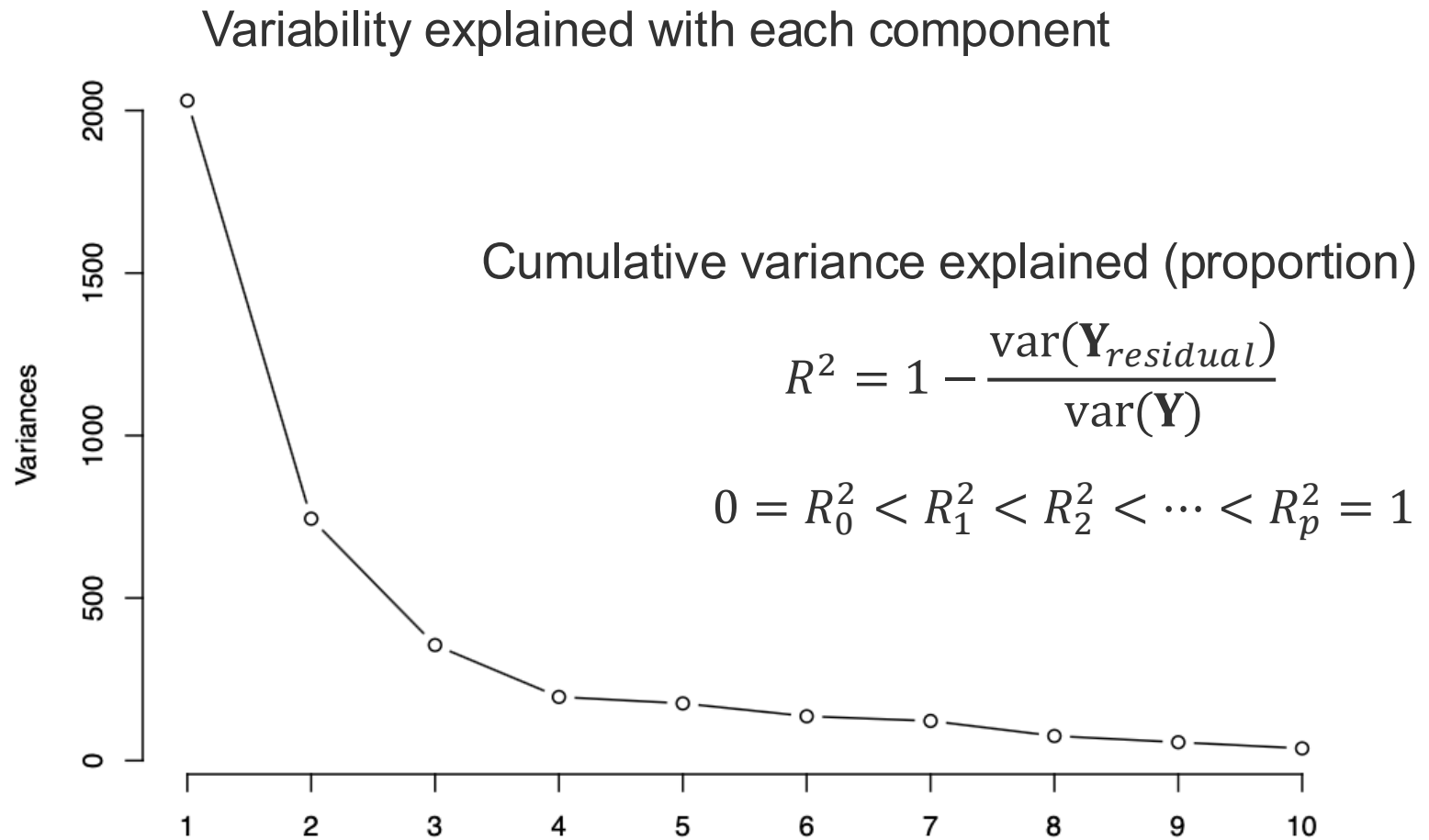
Output:

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



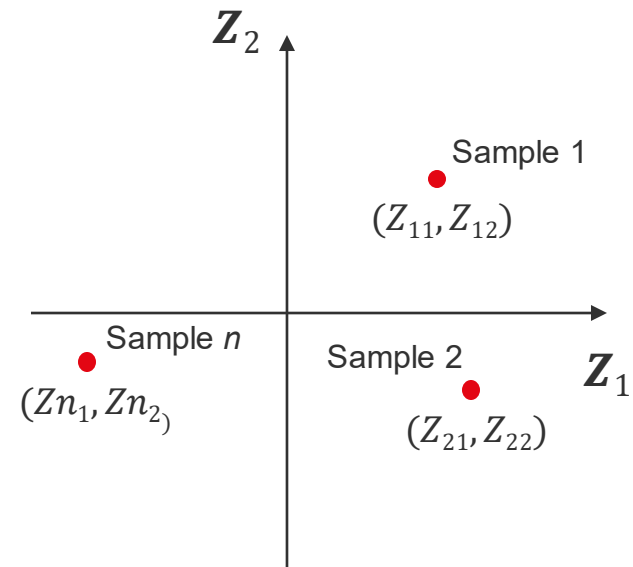
PCA score plot

Score: **projection measures of samples** in each principal component

=> Coordinate of samples on each axis PC

$$(\underbrace{Z_1 \ Z_2 \ \dots \ Z_p})$$

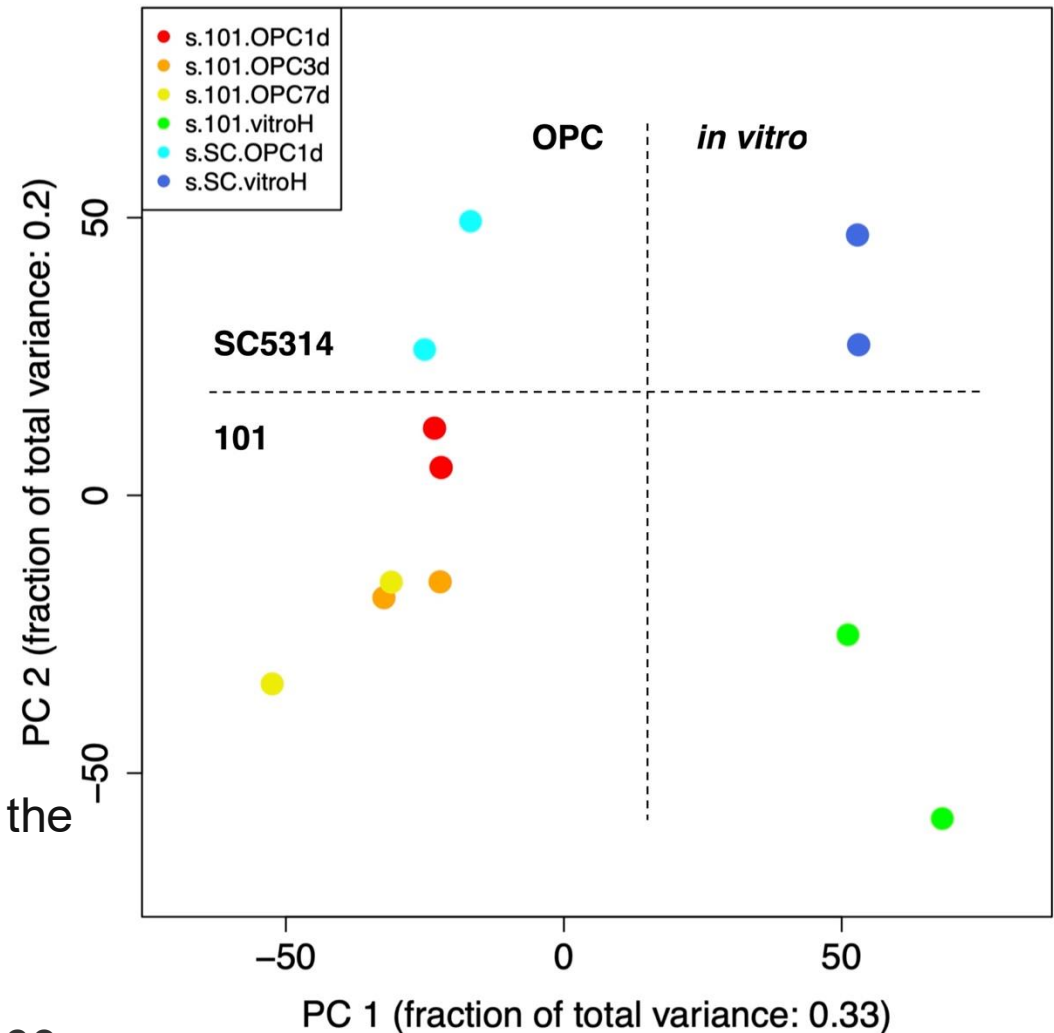
$$\begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \\ Z_{n1} & Z_{n2} \end{pmatrix}$$



PCA score plot

Transcriptomics dataset:

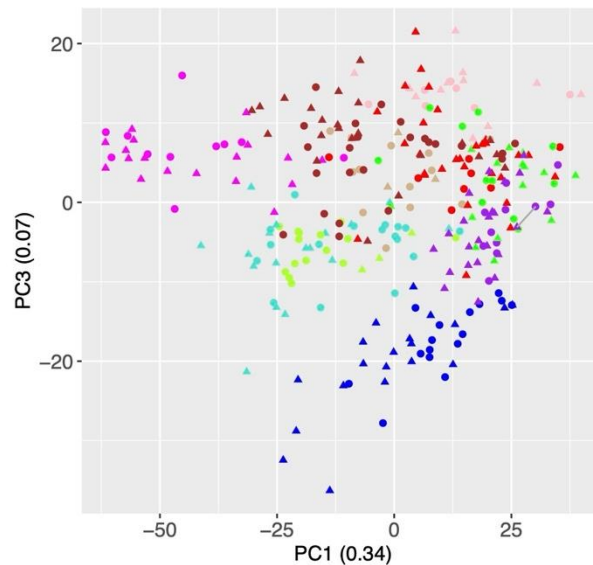
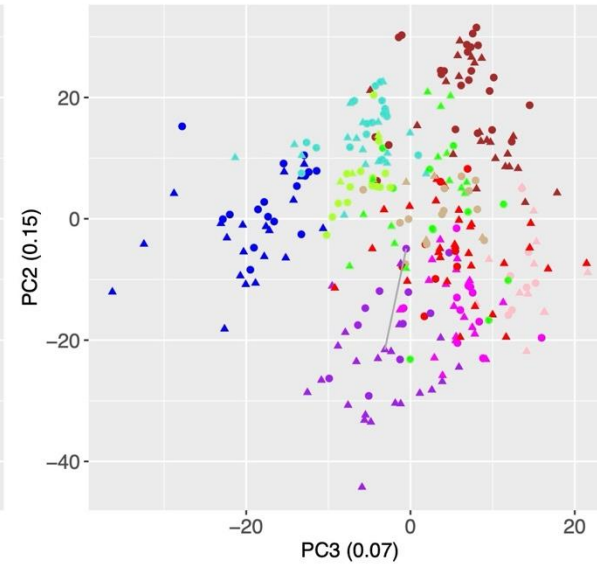
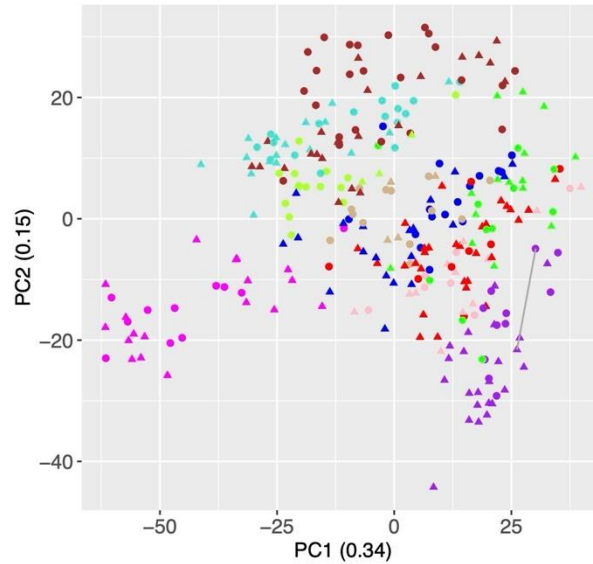
Gene expression profiles of different isolates of *Candida albicans* *in vitro* and during oral infection on mice



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

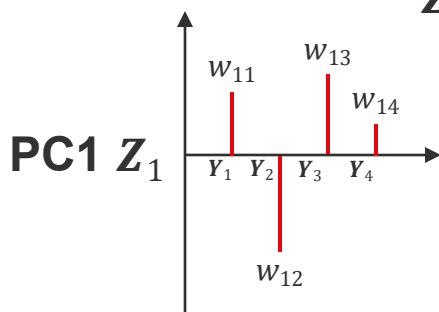
PCA loading plot

Loading: **contribution of given features** to each principal component

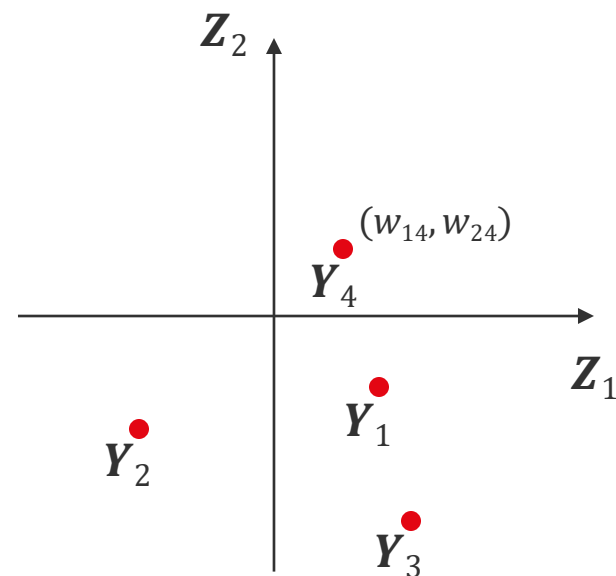
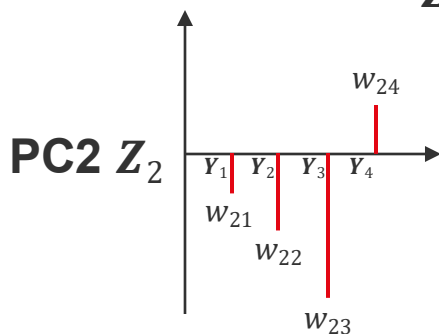
⇒ Coordinate of given features on each axis PC

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

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



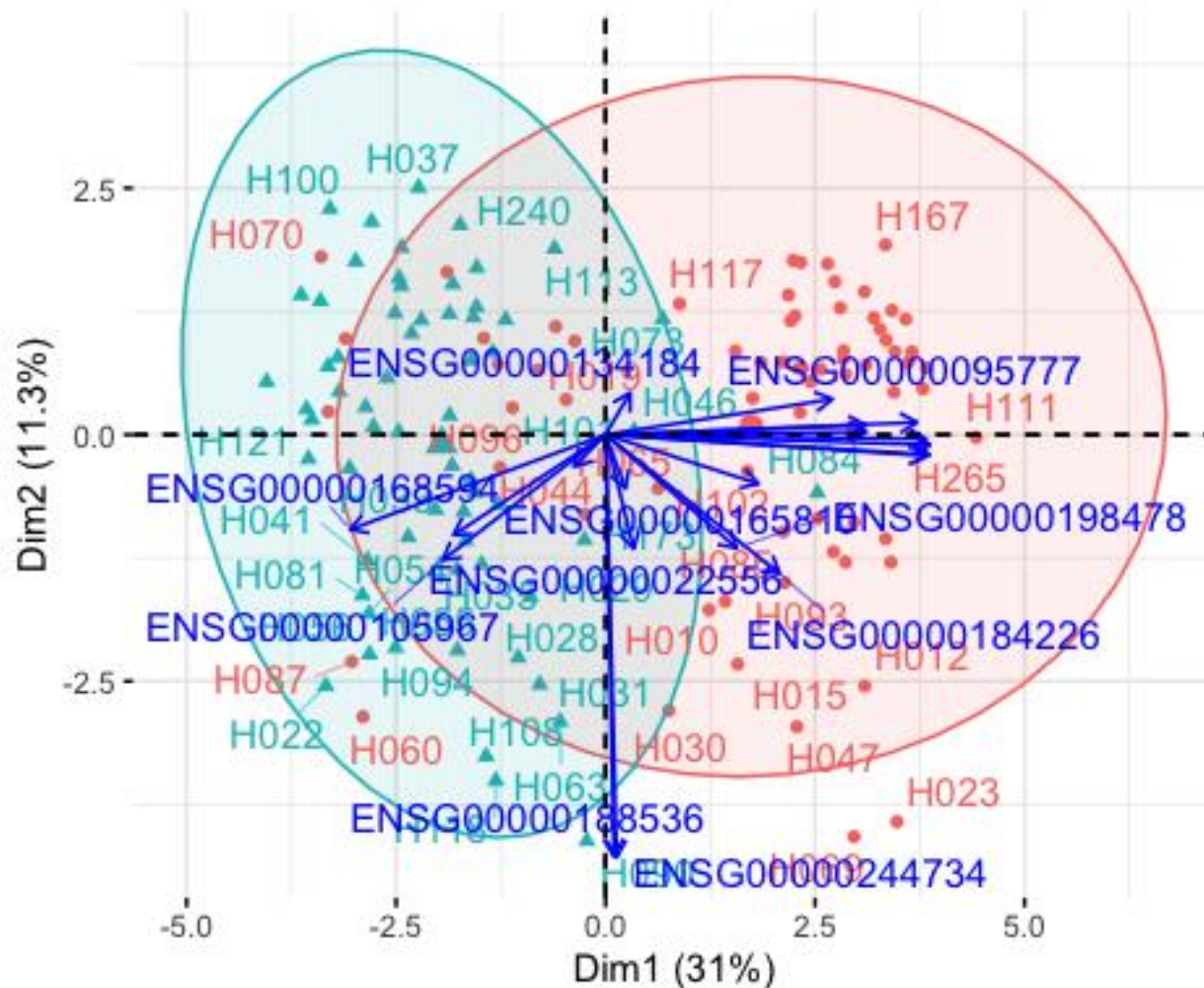
$$Z_2 = w_{21}Y_1 + w_{22}Y_2 + \dots + w_{2p}Y_p$$



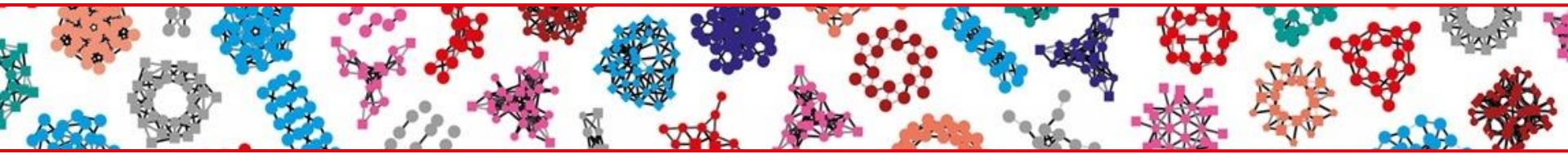
Relationship between given features

PCA biplot

Score plot + Loading plot



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

X: n samples \times p variables

Y: n samples \times q variables

Samples

Variables

**How are variations in
both response and
predictor?**

**What underlying
factors explain both
variations?**

Predict variables in **Y** using variables in **X**

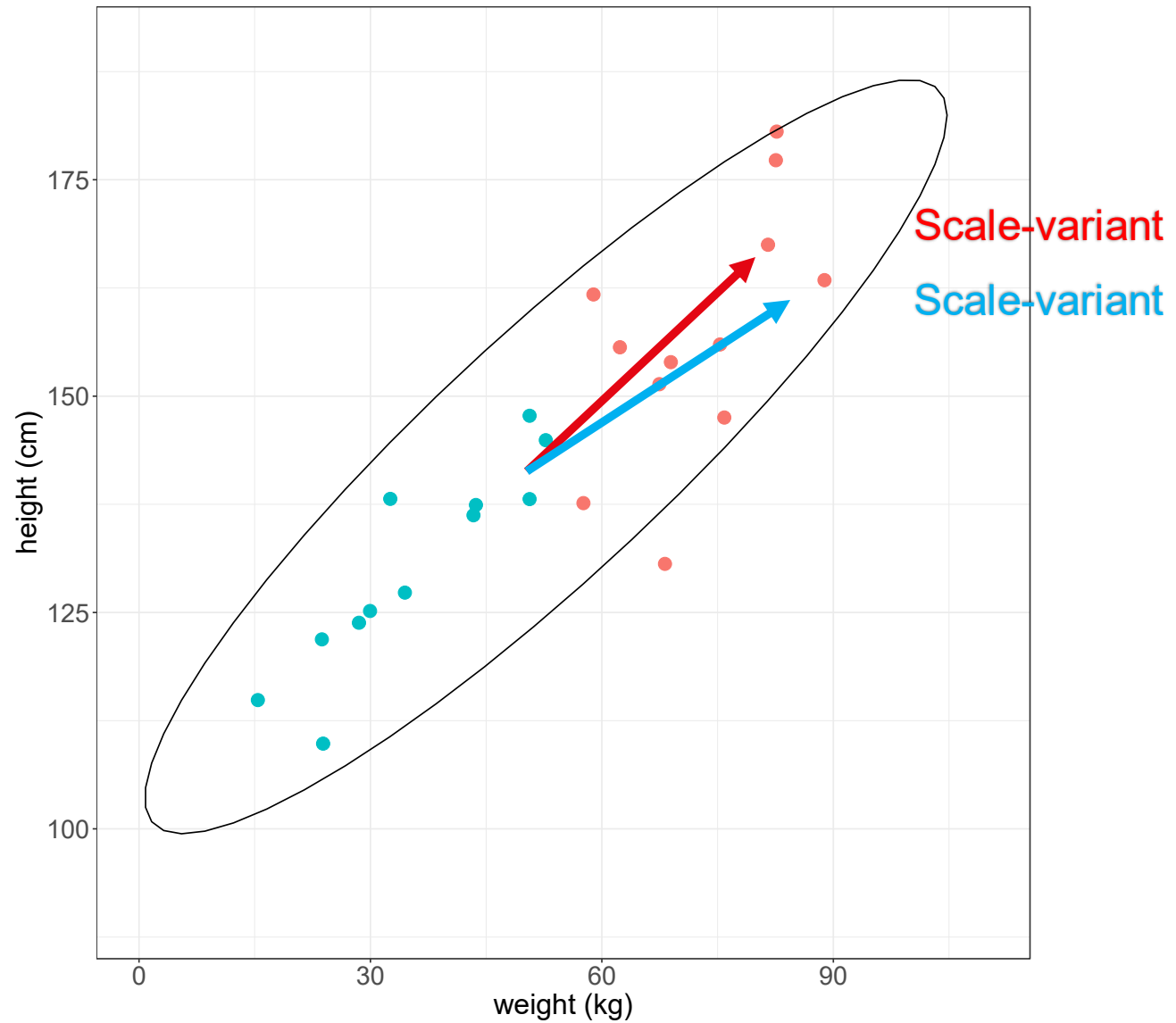
PLS Discriminant Analysis (PLS-DA)

X: n samples \times p variables

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**

⇒ 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**

⇒ Maximize covariance between **X** and **Y** latent variables

Components 1

Coordinate change:

$$\begin{aligned}(X_1 \ X_2 \ \dots \ X_p) &\rightarrow (U_1 \ U_2 \ \dots \ U_p) \\ (Y_1 \ Y_2 \ \dots \ Y_q) &\rightarrow (V_1 \ V_2 \ \dots \ V_q)\end{aligned}$$

Linear combination:

$$\begin{aligned}U_1 &= w_{11}X_1 + w_{12}X_2 + \dots + w_{1p}X_p = \mathbf{X}\mathbf{w}_1 \\ V_1 &= c_{11}Y_1 + c_{12}Y_2 + \dots + c_{1q}Y_q = \mathbf{Y}\mathbf{c}_1\end{aligned}$$

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

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

$$\text{cov}(U_1, V_1) = \text{cor}(U_1, V_1) \sqrt{\text{var}(U_1)} \sqrt{\text{var}(V_1)}$$

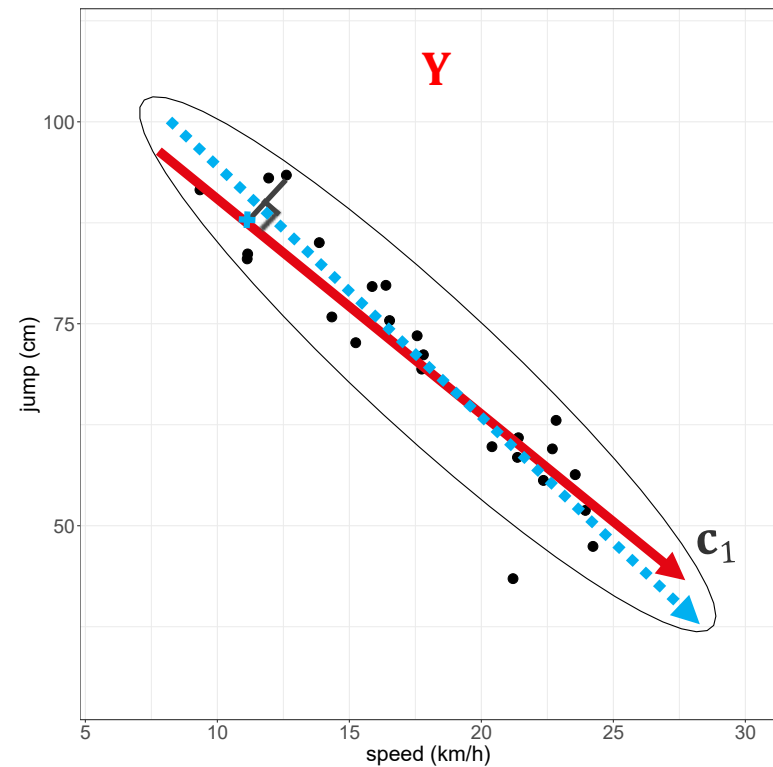
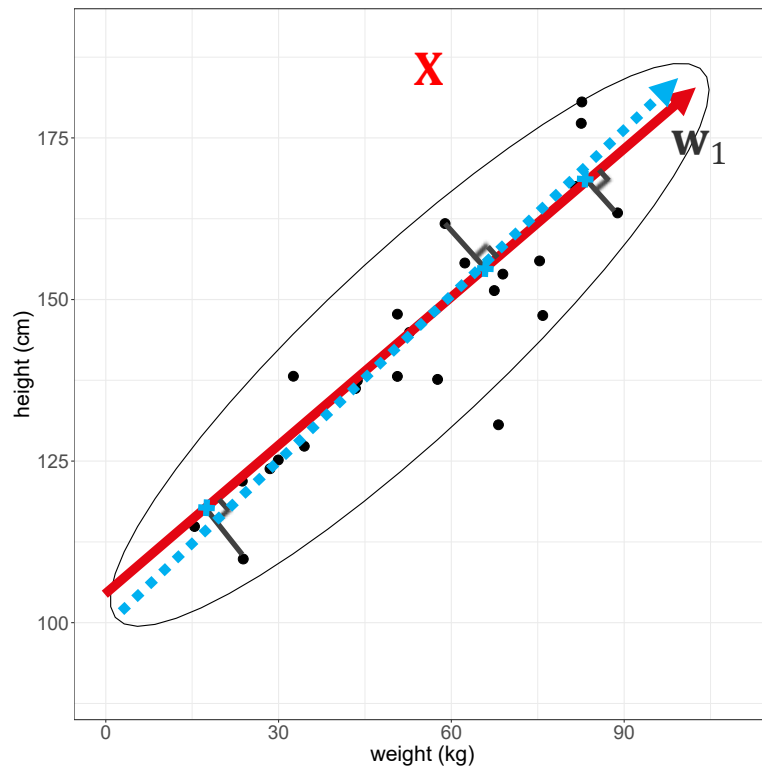
- best explaining variance in \mathbf{X} , given by $\sqrt{\text{var}(U_1)} = \mathbf{U}_1^T \mathbf{U}_1$
- best explaining variance in \mathbf{Y} , given by $\sqrt{\text{var}(V_1)} = \mathbf{V}_1^T \mathbf{V}_1$
- having greatest relationship between \mathbf{X} and \mathbf{Y} , given by $\text{cor}(U_1, V_1)$

Following components

- X-space: component 2 orthogonal to component 1
- Y-space:
 - Regression mode: not necessarily orthogonal (**Y** is deflated to the information from **X**)
 - Canonical mode: orthogonal (**Y** is deflated to the information from **Y**)
- **Algorithm:** iterative process
 - For component 1:
 w_1 = eigenvector corresponding to max eigenvalue of $X^T Y Y^T X$
 c_1 = eigenvector corresponding to max eigenvalue of $Y^T X X^T Y$
 - For component 2:
 w_2 = eigenvector corresponding to max eigenvalue of $X_{\text{deflated}2}^T Y_{\text{deflated}2} Y_{\text{deflated}2}^T X_{\text{deflated}2}$
 $X_{\text{deflated}1} = X, X_{\text{deflated}2} = X - U_1 U_1^T X_{\text{deflated}1}$, etc.
 - And so on, at most $\min(p, q)$

Scores

- Maximum covariance between X-space scores (e.g. projection of \mathbf{X} onto vector \mathbf{w}_1) and Y-space scores (e.g. \mathbf{Y} onto vector \mathbf{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
- ⇒ Relationship between X variables, between Y variables, between all variables
- **Loadings**: weights on *deflated (residual)* matrices
 - **Loadings-star**: weights on input matrices

$$\begin{aligned} U_1 &= Xw_1^* = X_{\text{deflated1}}w_1: & X_{\text{deflated1}} &= X, & w_1^* &= w_1 \\ U_2 &= Xw_2^* = X_{\text{deflated2}}w_2: & X_{\text{deflated2}} &= X - U_1U_1^T X_{\text{deflated1}}, & w_2^* &\neq w_2 \end{aligned}$$

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

Variability explained with components

- Cumulative variance explained for each space

$$R_{\mathbf{X}}^2 = 1 - \frac{\text{var}(\mathbf{X}_{deflated})}{\text{var}(\mathbf{X})}$$

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

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

$$R_{\mathbf{X},k}^2 = 1 - \frac{\text{var}(\mathbf{X}_{deflated,k})}{\text{var}(\mathbf{X}_k)}$$

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

PLS implementation

X: n samples \times p variables **Y: n samples \times q variables**

Input: X, Y

Output:

- cor: correlations
- variates: $(U_1 \ U_2 \ \dots \ U_{ncomp}), (V_1 \ V_2 \ \dots \ V_{ncomp})$
- loadings: $(w_1 \ w_2 \ \dots \ w_{ncomp}), (c_1 \ c_2 \ \dots \ c_{ncomp})$
- 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 **X** that is
not correlated to responses **Y**

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

⇒ Model separately variations of **X** correlated and uncorrelated to **Y**

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

Y = OPC *versus in vitro*

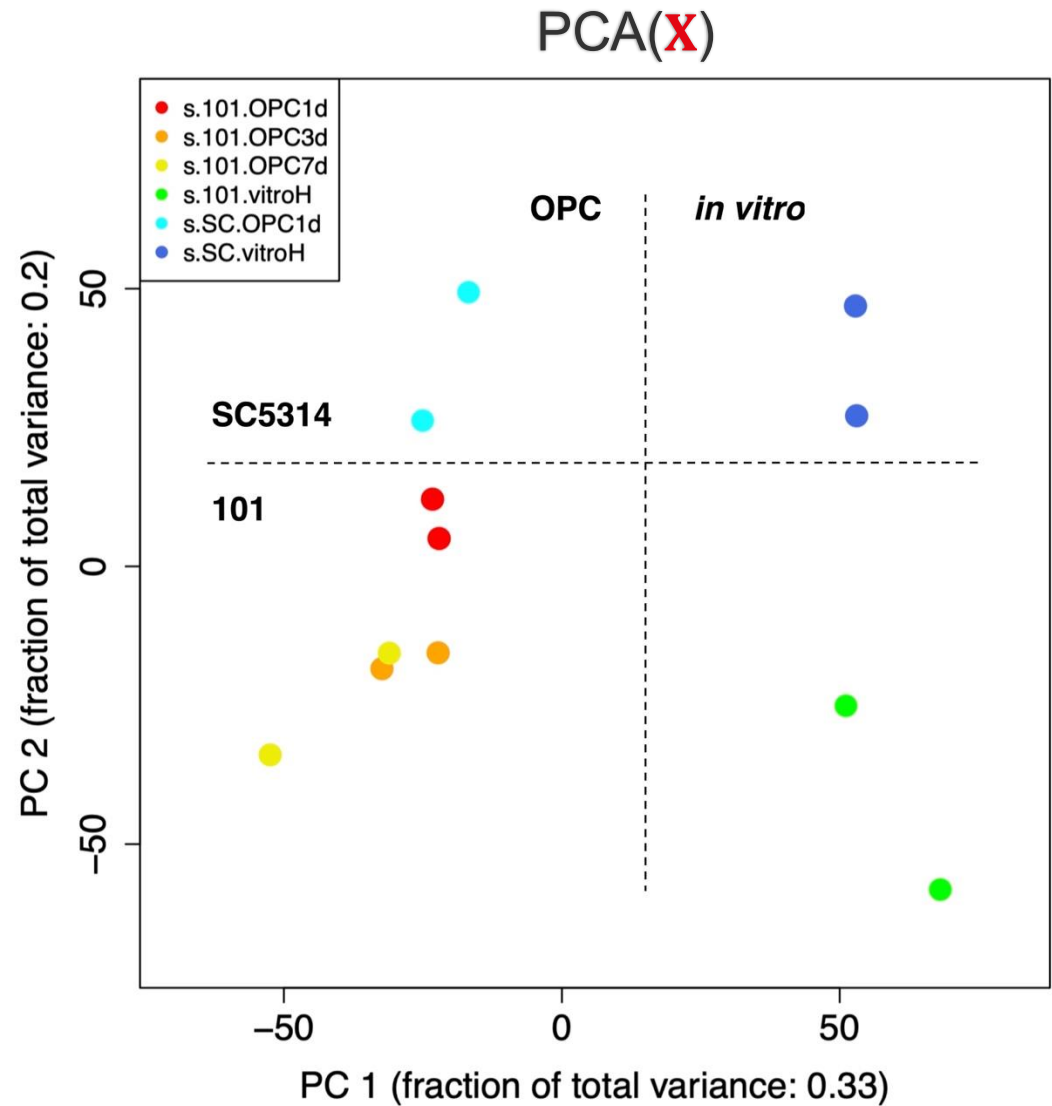
Predictive component \cong PC1

Orthogonal component \cong PC2

Y = 101 *versus* SC5314

Predictive component \cong PC2

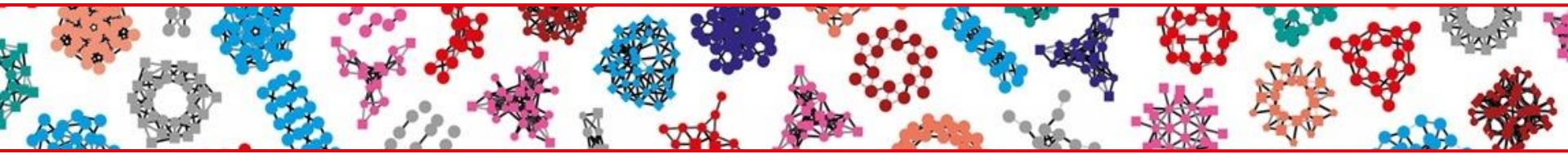
Orthogonal component \cong PC1



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

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

$$\begin{aligned} & \max_{w,c} \text{cov}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c}) \\ & \text{subject to } \|\mathbf{w}\| = \|\mathbf{c}\| = 1 \end{aligned}$$

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

$$\begin{aligned} & \max_{w,c} \text{cor}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c}) \\ & \text{subject to } \text{var}(\mathbf{X}\mathbf{w}) = \text{var}(\mathbf{Y}\mathbf{c}) = 1 \end{aligned}$$

Components 1

Coordinate change:

$$\begin{aligned} (X_1 \ X_2 \ \dots \ X_p) &\rightarrow (U_1 \ U_2 \ \dots \ U_p) \\ (Y_1 \ Y_2 \ \dots \ Y_q) &\rightarrow (V_1 \ V_2 \ \dots \ V_q) \end{aligned}$$

Linear combination:

$$\begin{aligned} U_1 &= w_{11}X_1 + w_{12}X_2 + \dots + w_{1p}X_p = \mathbf{X}\mathbf{w}_1 \\ V_1 &= c_{11}Y_1 + c_{12}Y_2 + \dots + c_{1q}Y_q = \mathbf{Y}\mathbf{c}_1 \end{aligned}$$

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

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

$$\text{cor}(\mathbf{U}_1, \mathbf{V}_1)$$

How does CCA works?

Solution:

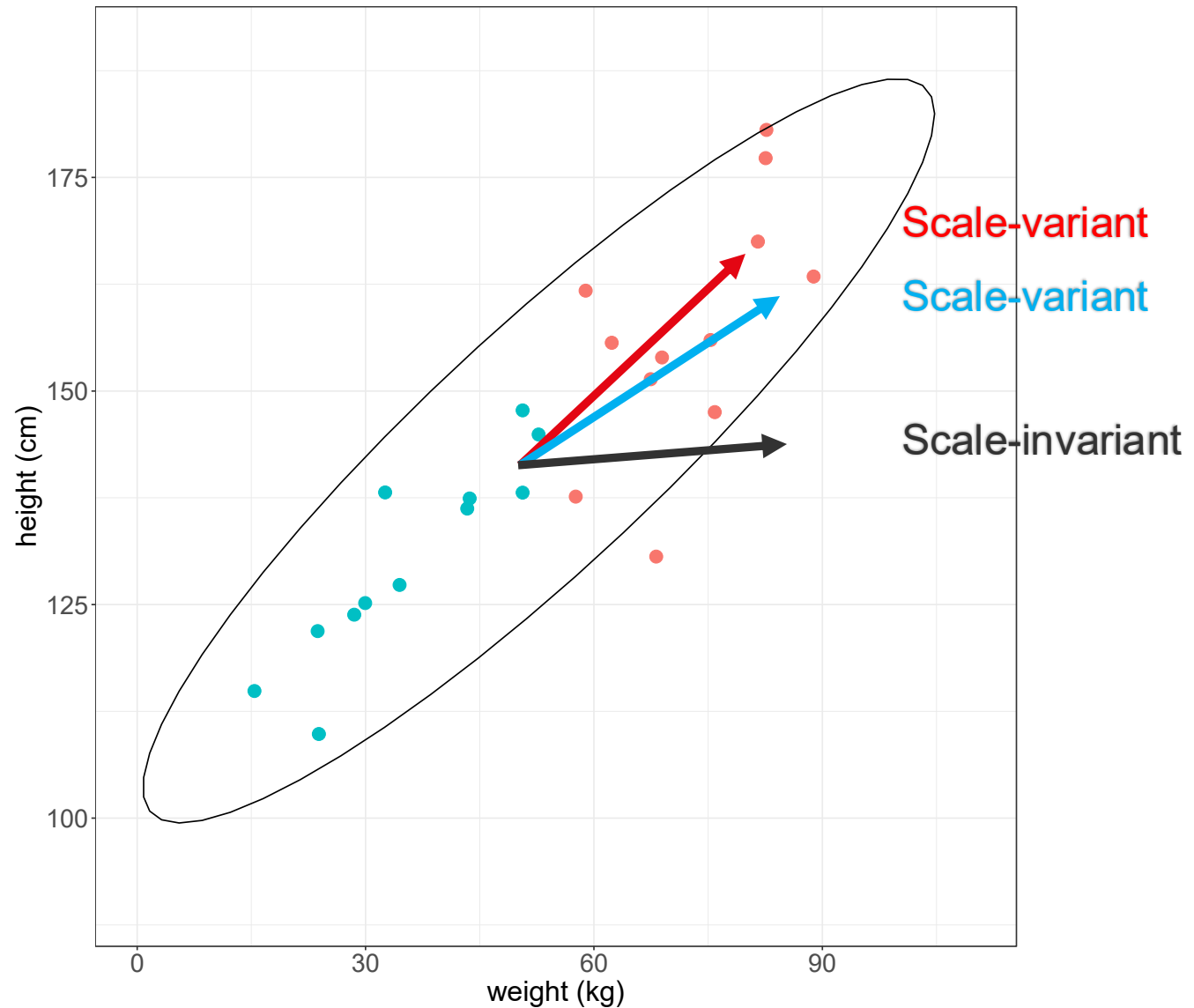
$$\begin{aligned}\max \text{cor}(\mathbf{U}_1, \mathbf{V}_1) &= \text{sqrt}(\max \text{ eigenvalue of } \text{cov}(\mathbf{X})^{-1} \text{cov}(\mathbf{X}, \mathbf{Y}) \text{cov}(\mathbf{Y})^{-1} \text{cov}(\mathbf{Y}, \mathbf{X})) \\ &= \text{sqrt}(\max \text{ eigenvalue of } \text{cov}(\mathbf{Y})^{-1} \text{cov}(\mathbf{Y}, \mathbf{X}) \text{cov}(\mathbf{X})^{-1} \text{cov}(\mathbf{X}, \mathbf{Y})) \\ &\quad \text{at } \mathbf{w}_1, \mathbf{c}_1 = \text{corresponding eigenvectors}\end{aligned}$$

Hint: Lagrange multiplier + derivative

$$\begin{aligned}\max \text{cor}(\mathbf{U}_2, \mathbf{V}_2) &= 2^{\text{nd}} \max \text{ eigenvalue} \\ &\quad \text{at } \mathbf{w}_2, \mathbf{c}_2 = \text{corresponding eigenvectors}\end{aligned}$$

and so on

CCA versus PLS versus PCA



CCA implementation

X: n samples \times p variables **Y: n samples \times q variables**

Input: X, Y

Output:

- cor: correlations
- variates: $(U_1 \ U_2 \ \dots \ U_{ncomp}), (V_1 \ V_2 \ \dots \ V_{ncomp})$
- loadings: $(w_1 \ w_2 \ \dots \ w_{ncomp}), (c_1 \ c_2 \ \dots \ c_{ncomp})$
- proportion of explained variance
- correlation between variates and input data

Regularized CCA

Solution:

$$\begin{aligned}\max \text{cor}(\mathbf{U}_1, \mathbf{V}_1) &= \text{sqrt}(\text{max eigenvalue of } \text{cov}(\mathbf{X})^{-1} \text{cov}(\mathbf{X}, \mathbf{Y}) \text{cov}(\mathbf{Y})^{-1} \text{cov}(\mathbf{Y}, \mathbf{X})) \\ &= \text{sqrt}(\text{max eigenvalue of } \text{cov}(\mathbf{Y})^{-1} \text{cov}(\mathbf{Y}, \mathbf{X}) \text{cov}(\mathbf{X})^{-1} \text{cov}(\mathbf{X}, \mathbf{Y}))\end{aligned}$$

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

- Constant variables

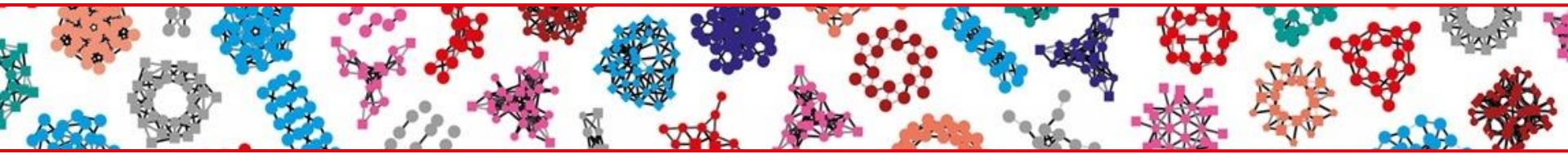
⇒ remove

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

⇒ regularized CCA: $\text{cov}(\mathbf{X}) \leftarrow \text{cov}(\mathbf{X}) + \lambda_{\mathbf{X}} \mathbf{I}$, $\text{cov}(\mathbf{Y}) \leftarrow \text{cov}(\mathbf{Y}) + \lambda_{\mathbf{Y}} \mathbf{I}$

⇒ tuning for optimal values of $\lambda_{\mathbf{X}}$, $\lambda_{\mathbf{Y}}$

Overview



01 —• **Principal Component Analysis**

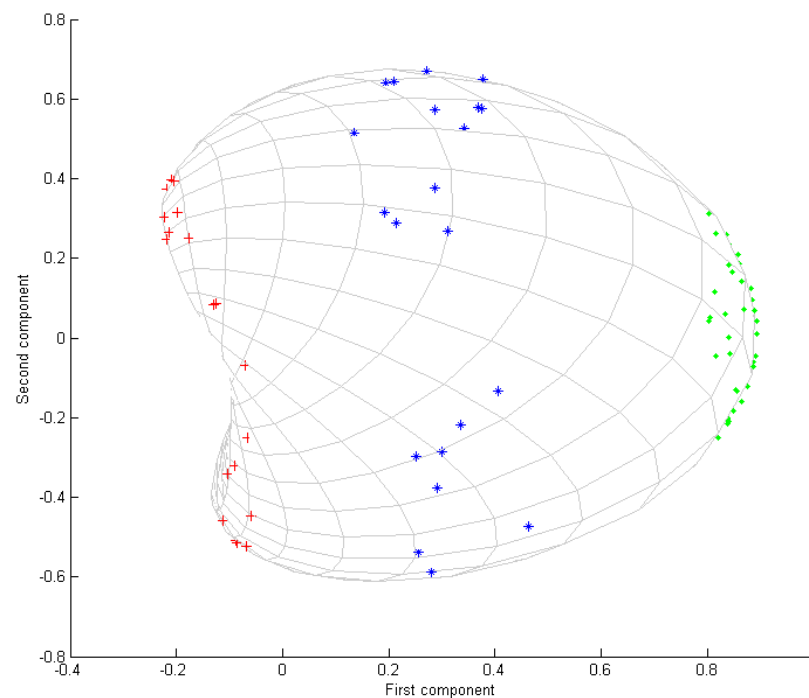
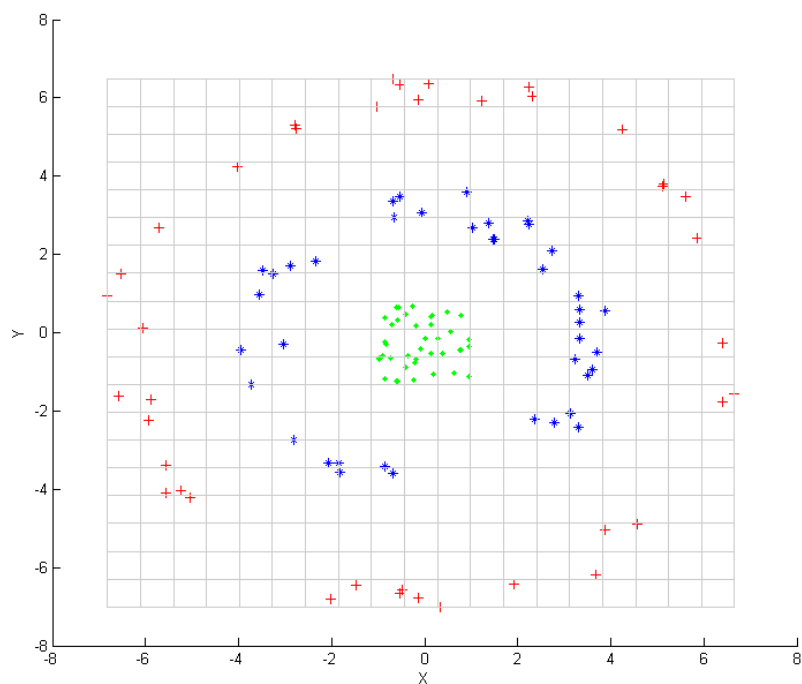
02 —• **Partial Least Squares**

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**



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

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

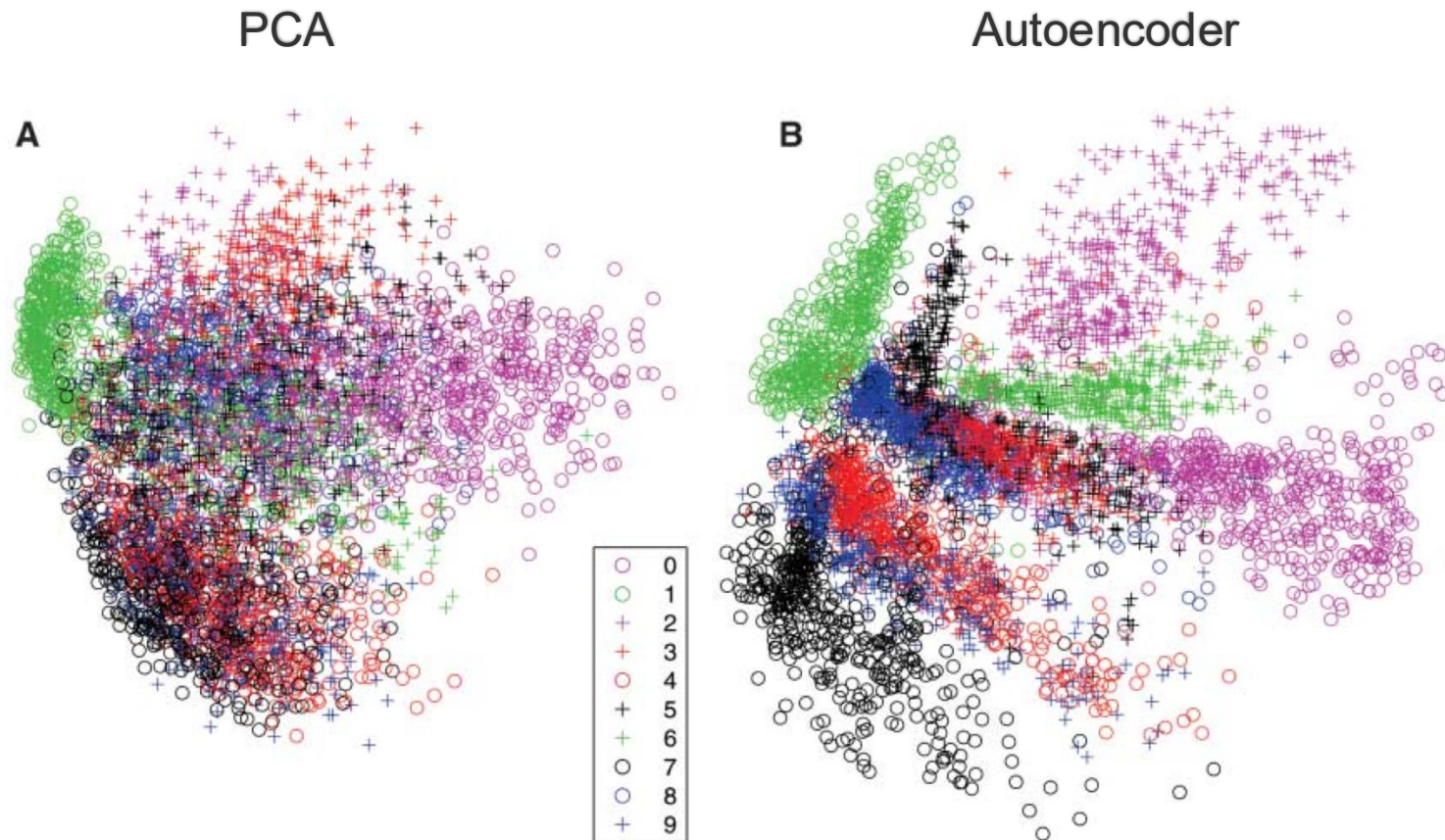
https://commons.wikimedia.org/wiki/File:Kernel_pca_output.png

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

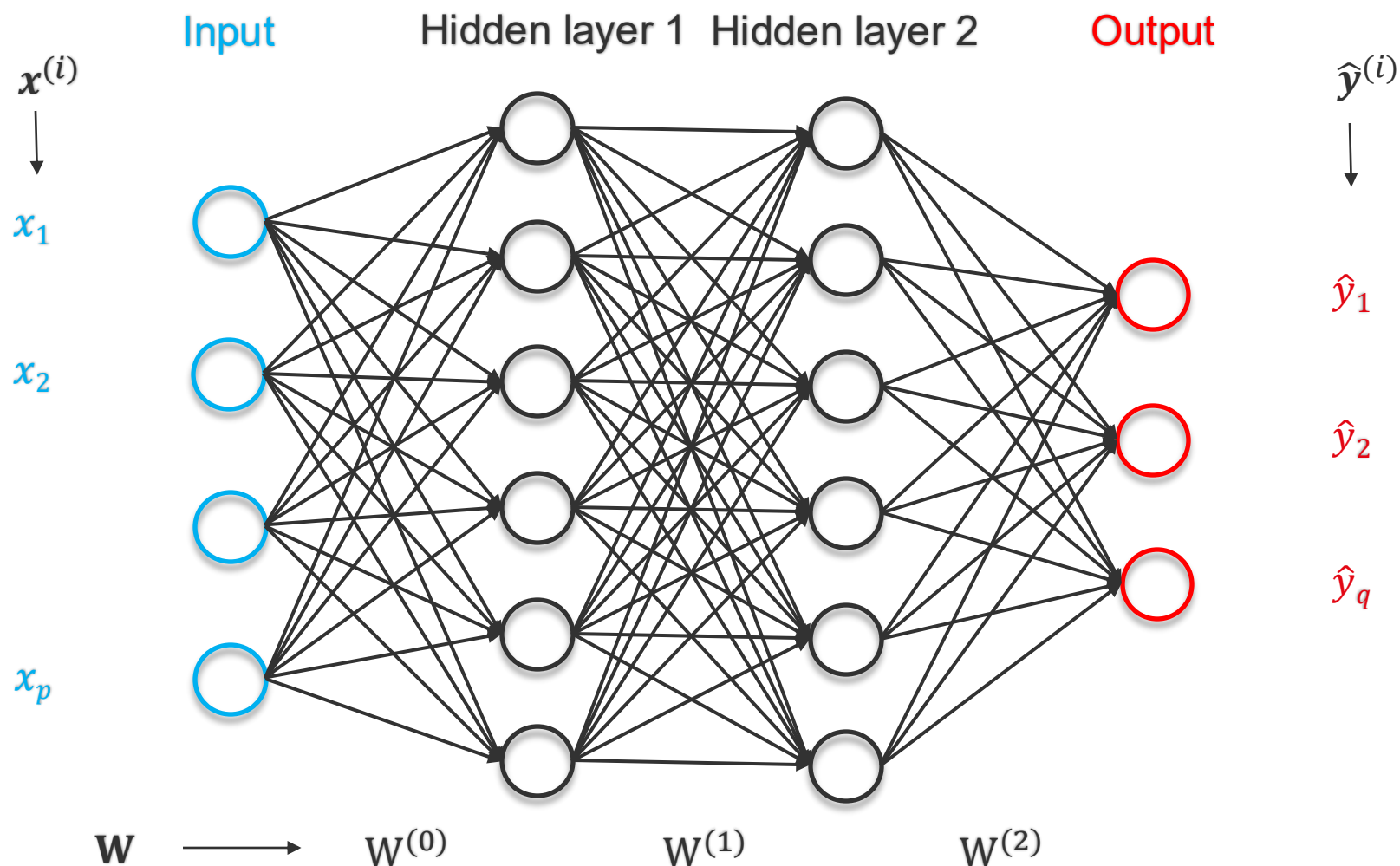
Towards Nonlinearity

Generalize principal components from straight lines to curves:

Autoencoder



Neural Networks

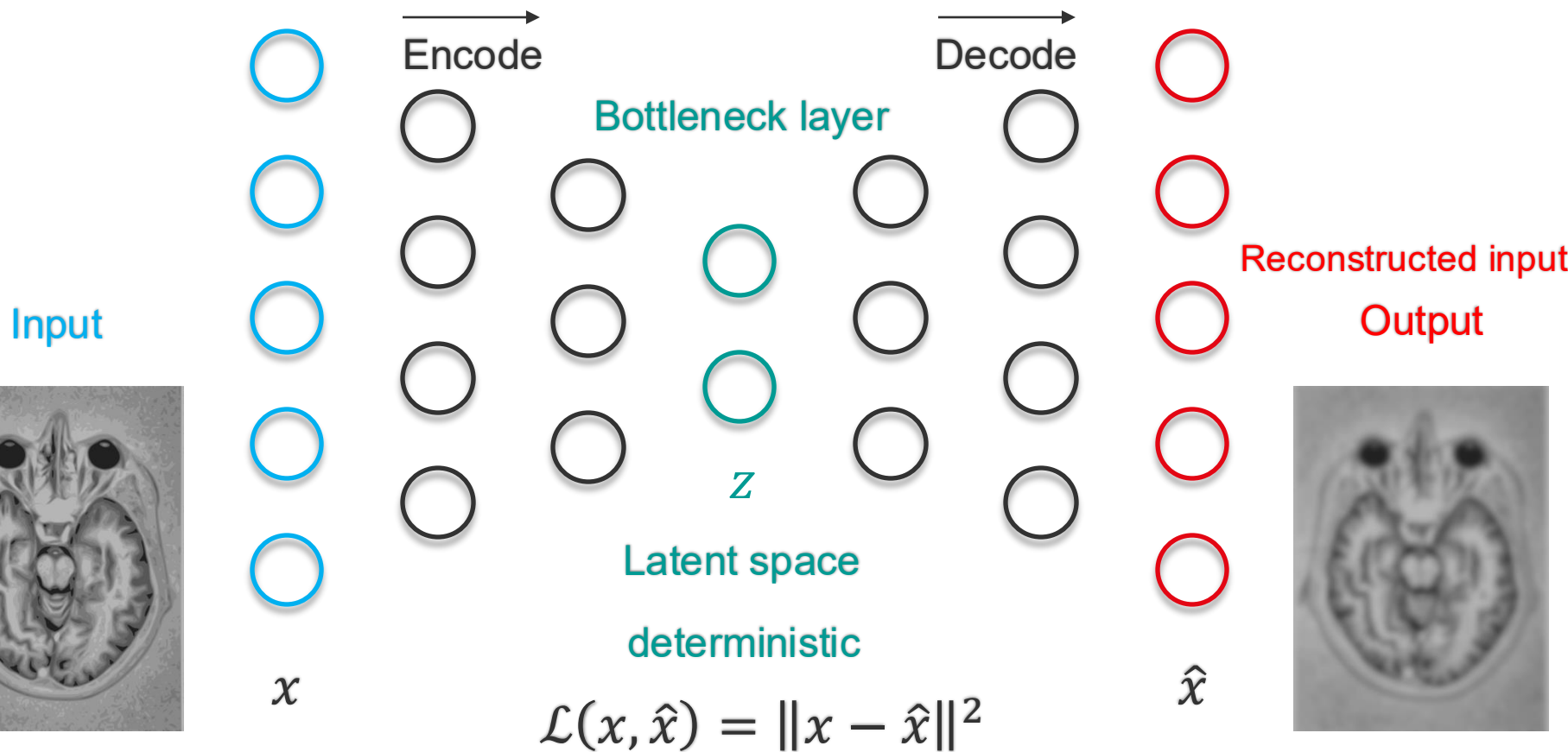


Activation functions on linear regressions

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

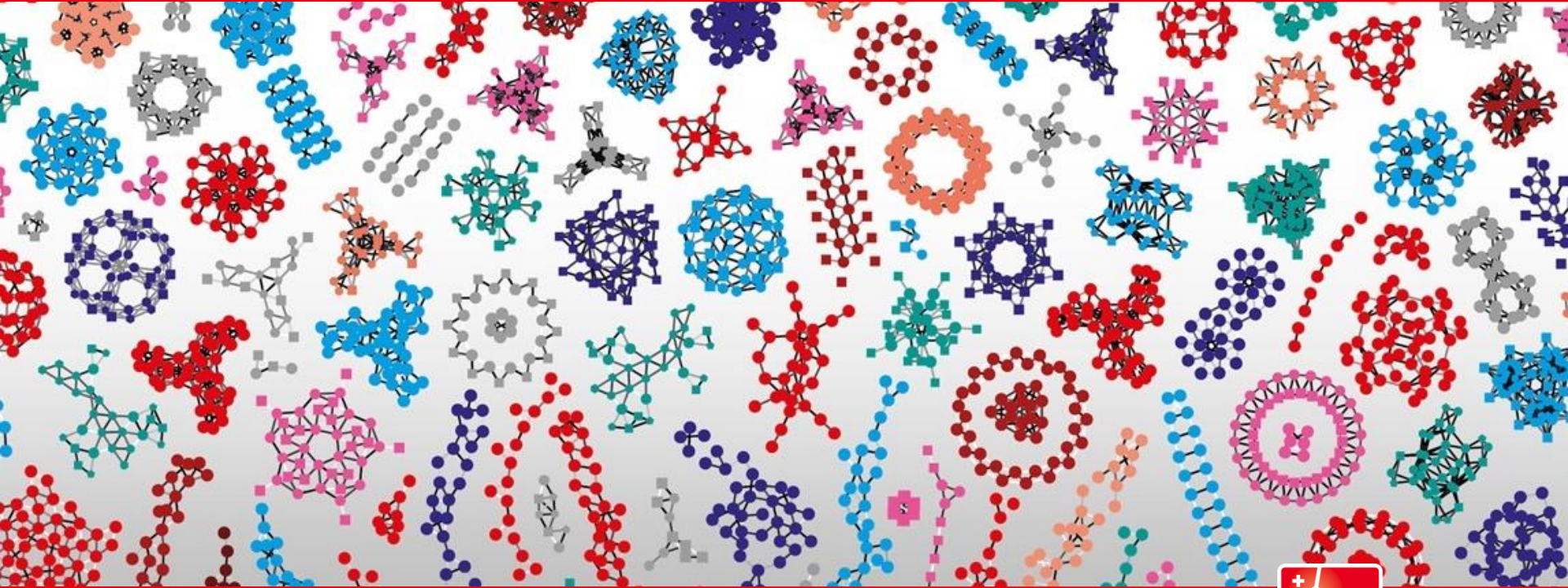
Autoencoder

- 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



Swiss Institute of
Bioinformatics

Thank you