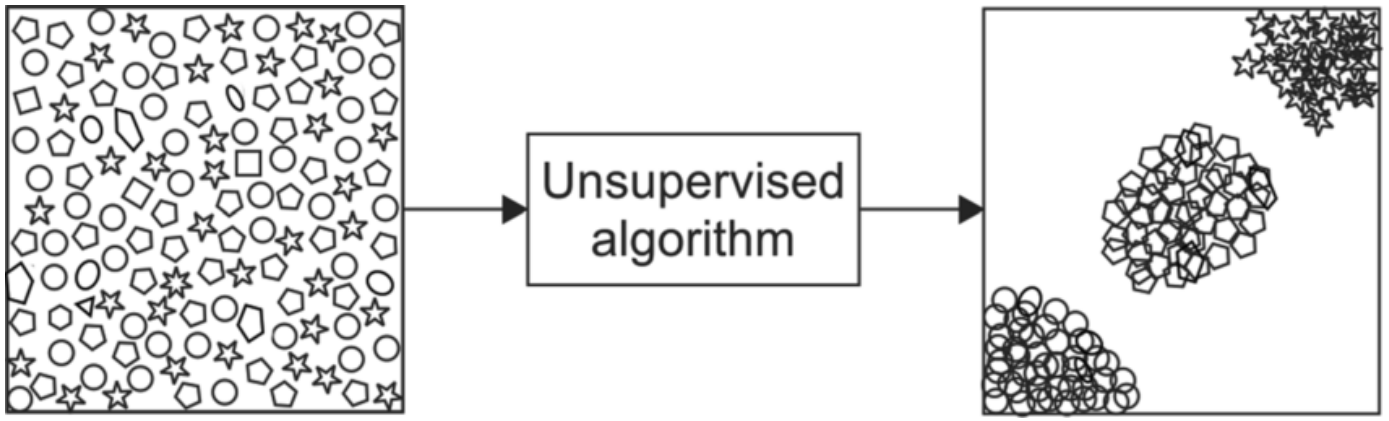
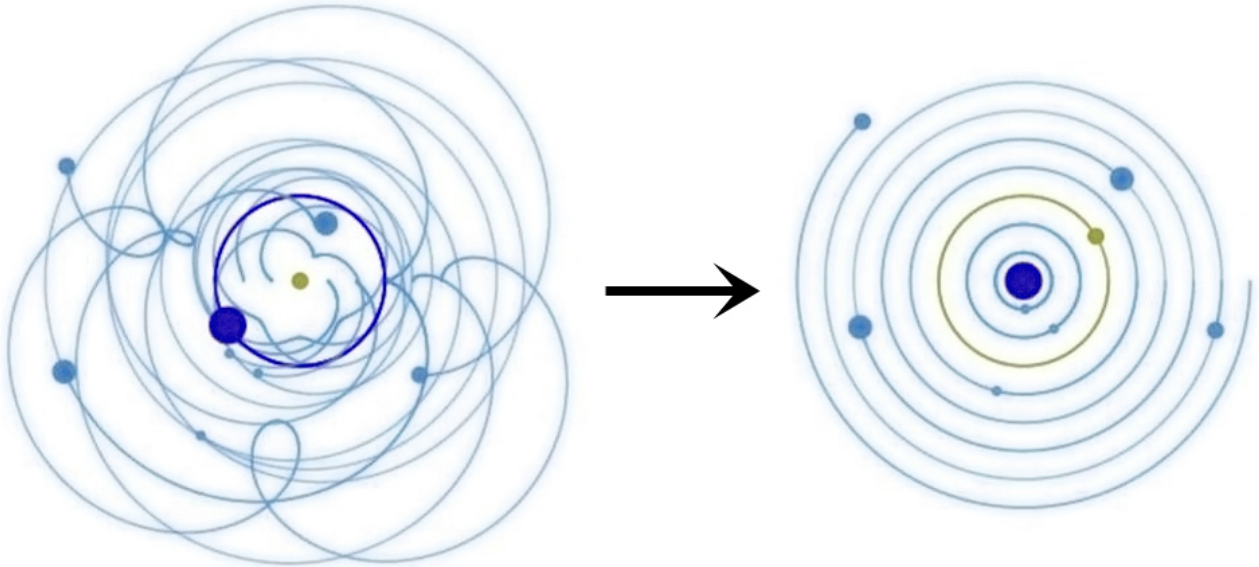
PCA, PLS, CCA, gCCA

Data integration and Multi-omics Workshop

19/12/2025

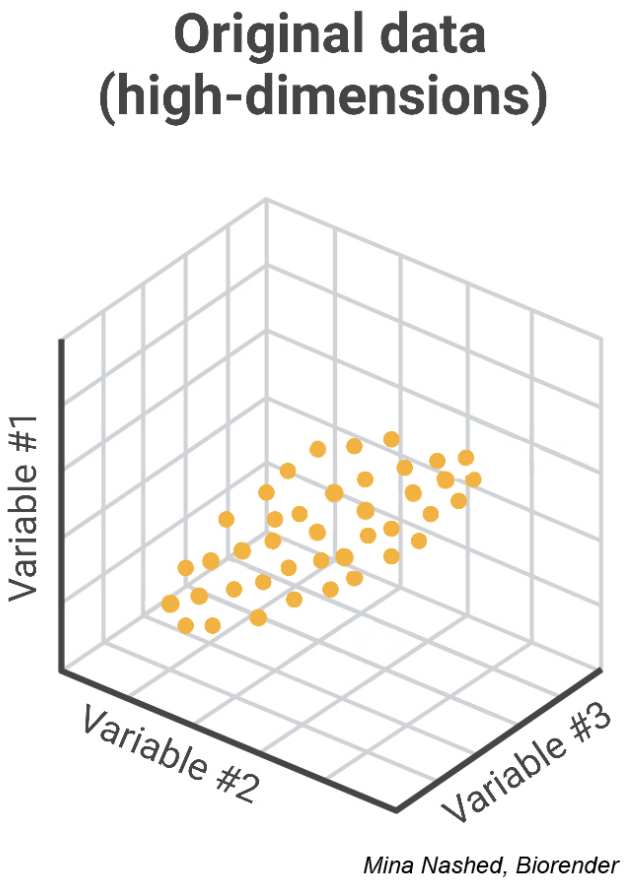
Motivation





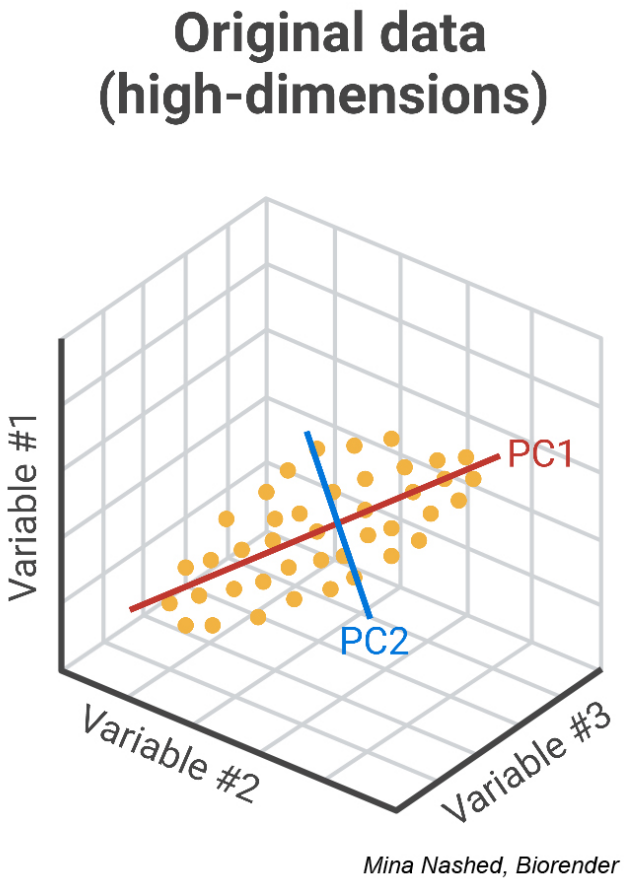
Original data

- Our dataset X has N observations (rows) and I variables (columns).
- Each point is an observation in the original coordinates system



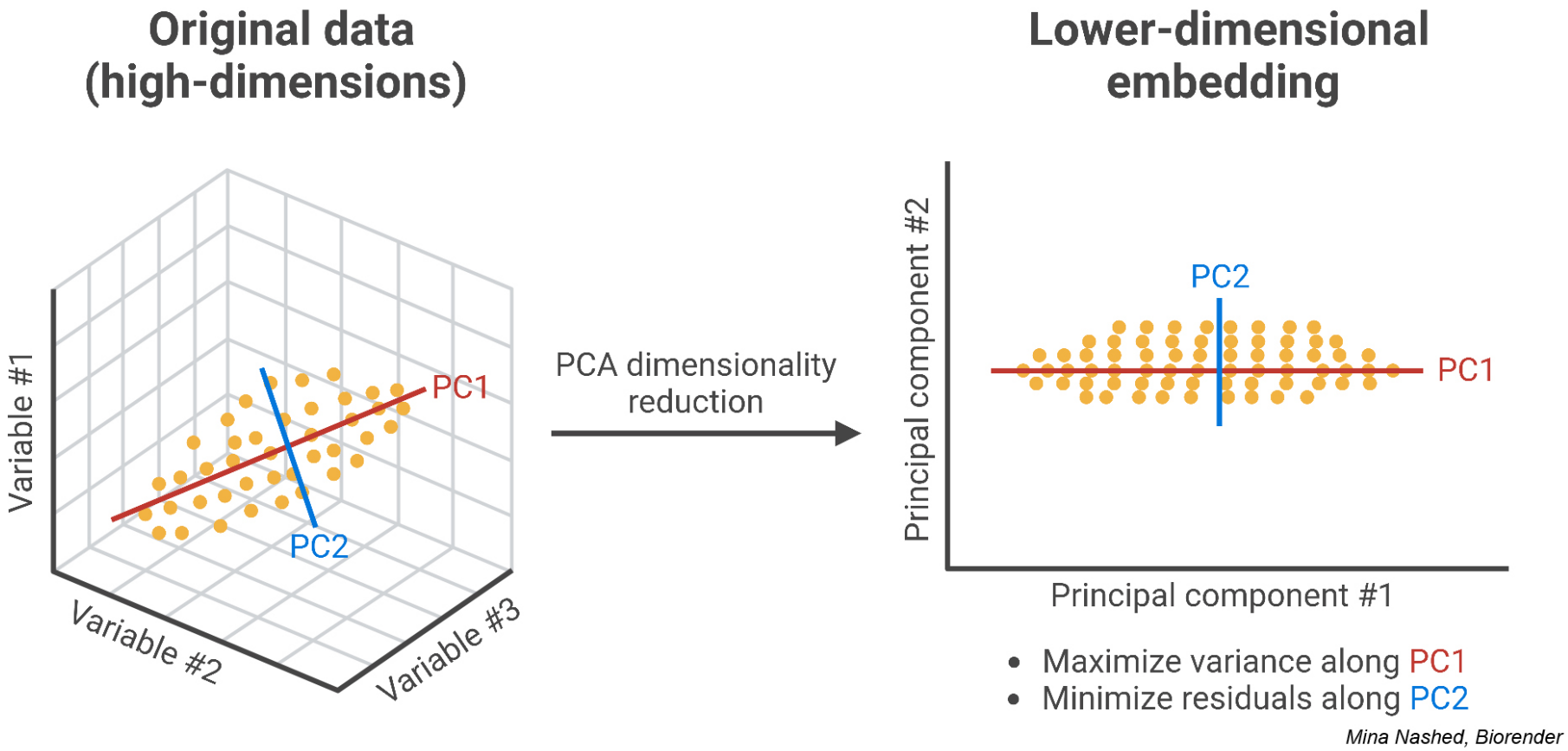
PCA

- What are the perpendicular directions of maximal *variance*?
  - ≈ Directions along which points spread the most



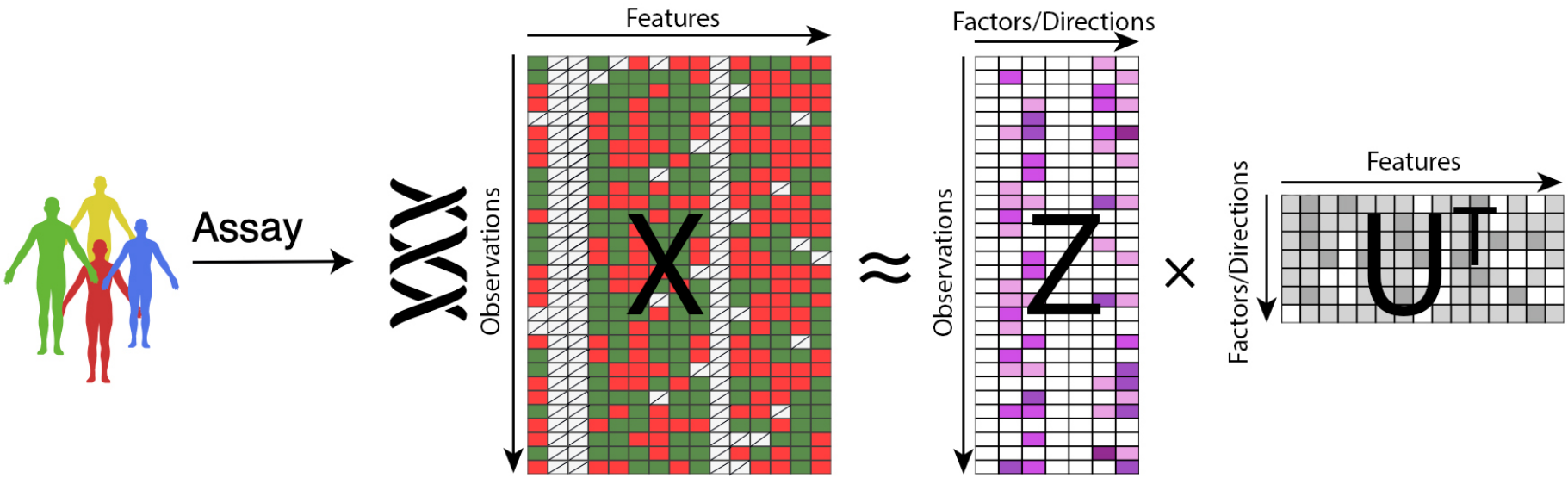
Projection

- Once we have those directions (or components), project the data on those new axes



Matrix factorization

- Mathematically: $Z \approx XU$ (or $X \approx ZU^T$)
- Factors/Directions/Components are in fact latent variables:
  - → Clever recombinations of the base features (encoded in U here)



PCA - The eigenproblem

$$\underset{U^TU=I}{\arg\max}\, Var(Xu_i) \quad \forall i \quad \Leftrightarrow \nabla_U \left(U^T X^T XU - (U^TU-1)\Lambda\right) = 0$$

- $\nabla_U(...) = 0$: a maximum of a function is where its slope is null
- $Var(XU) \propto \sum_i u_i^T X^T X u_i = Trace(U^T X^T XU)$
- $-(U^TU-1)\Lambda$ (Lagrange multiplier) lowers the function we want to maximize only when $U^TU \neq I$

$$\Leftrightarrow \quad \underbrace{(X^TX)}_{Var(X)} U = U\Lambda$$

This is an eigenvalues equation

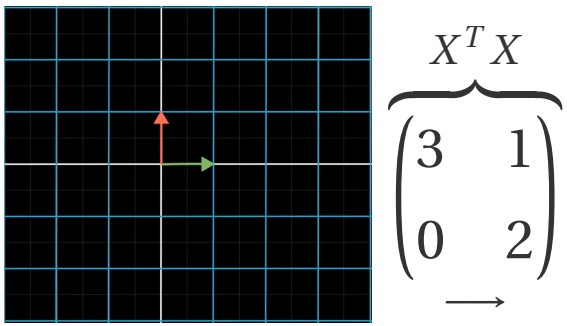
(Solvable numerically and interpretable geometrically)

Eigendecomposition

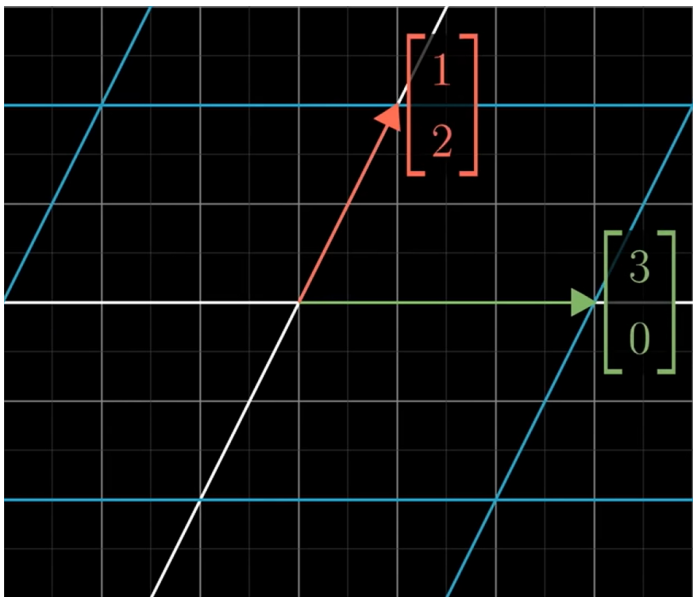
- Eigenvectors (u) give fundamental directions of $Var(X) = X^T X$, eigenvalues (λ) give the stretch along those directions

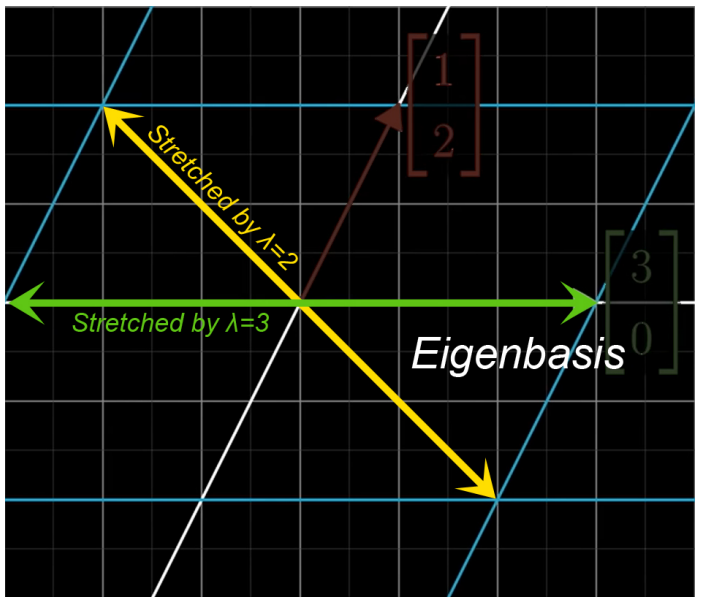
$$X^T X u = \lambda u$$

  - $X^T X$ being symmetric, all eigenvectors are perpendicular



$$\overbrace{\begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix}}^{X^T X}$$

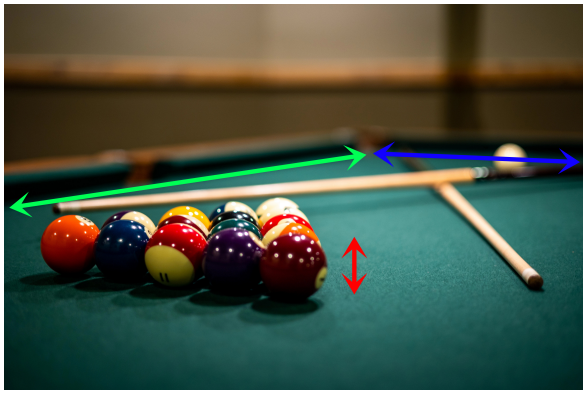




- 3Blue1Brown's "Eigenvectors and eigenvalues" on Youtube (17min)

Removing components

- Directions with negligible stretch ($\lambda_i \ll 1$) are irrelevant
  - → We can get rid of them (they play no role)
- Removing directions → Low-rank approximation



Project on X

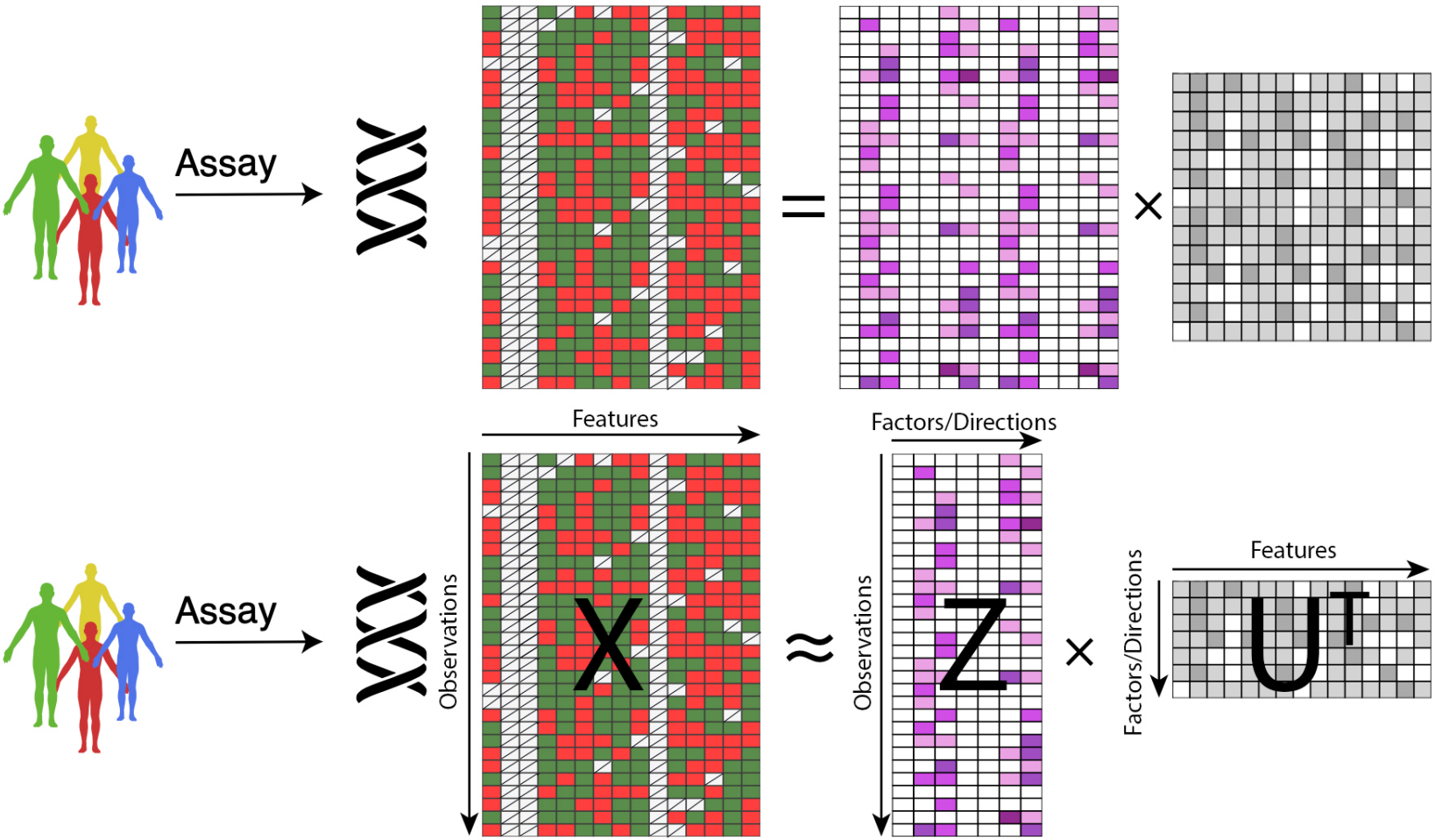
Project on X+Y

Project on X+Y+Z

Low-rank factorization

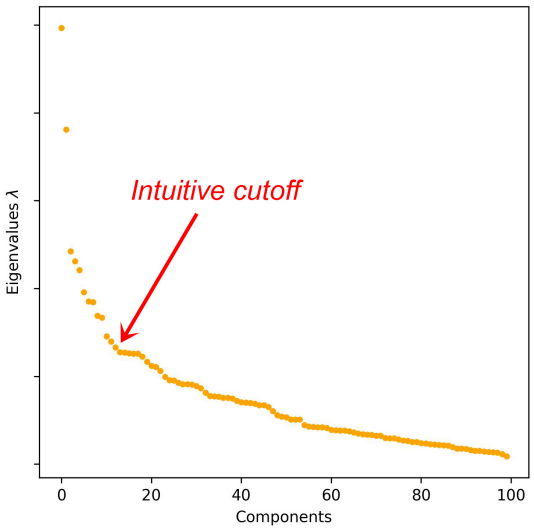
- Keep only the K "strongest" directions: $Z = XU$

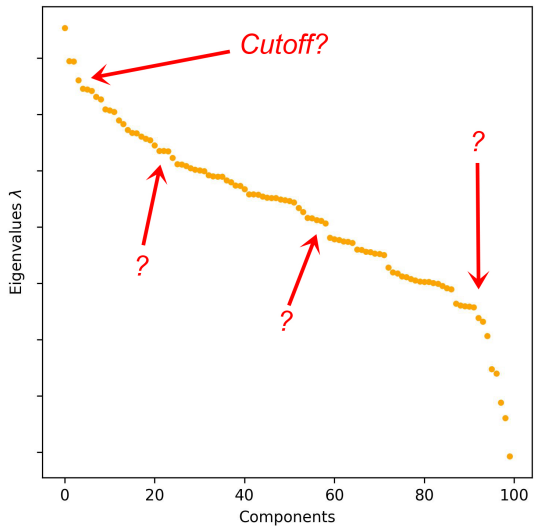
→ Removes noise in the data, helps the interpretation

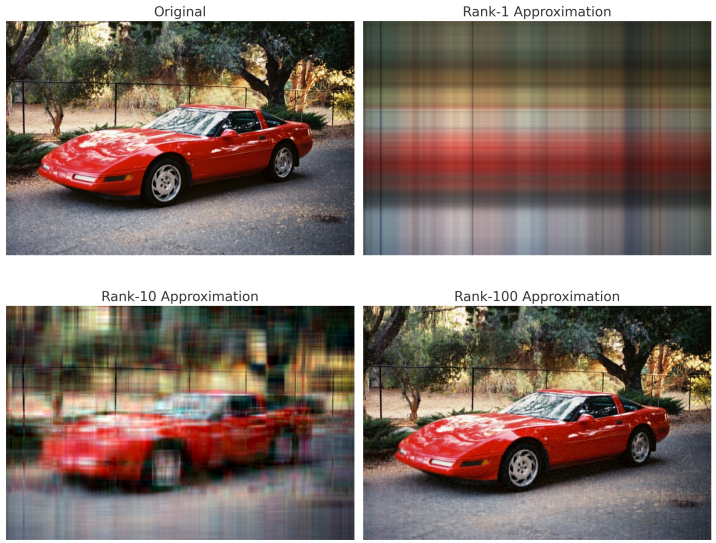


In practice

- In practice, finding the cutoff might not be easy:
  - Heuristics, cutoff criteria, cross-validation, ...







- A quicker and more stable path is noting that:
  - $Var(X) = X^T X = U \Lambda U^T$
  - $X = V \sqrt{\Lambda} U^T$ is the Singular Value decomposition (SVD) of X

More than one dataset

- PCA: axes of max variance *within* one dataset, then cutoff
  - $\text{argmax}\, Var(Xu_i)\ \ s.t. ||u_i|| = 1, u_i \perp u_j\ \forall i, j$
- PLS: axes of max covariance *between* two datasets, then cutoff
  - $\text{argmax}\, Cov(Xu_i, Yv_i)\ \ s.t. ||u_i|| = 1, ||v_i|| = 1\ \forall i$

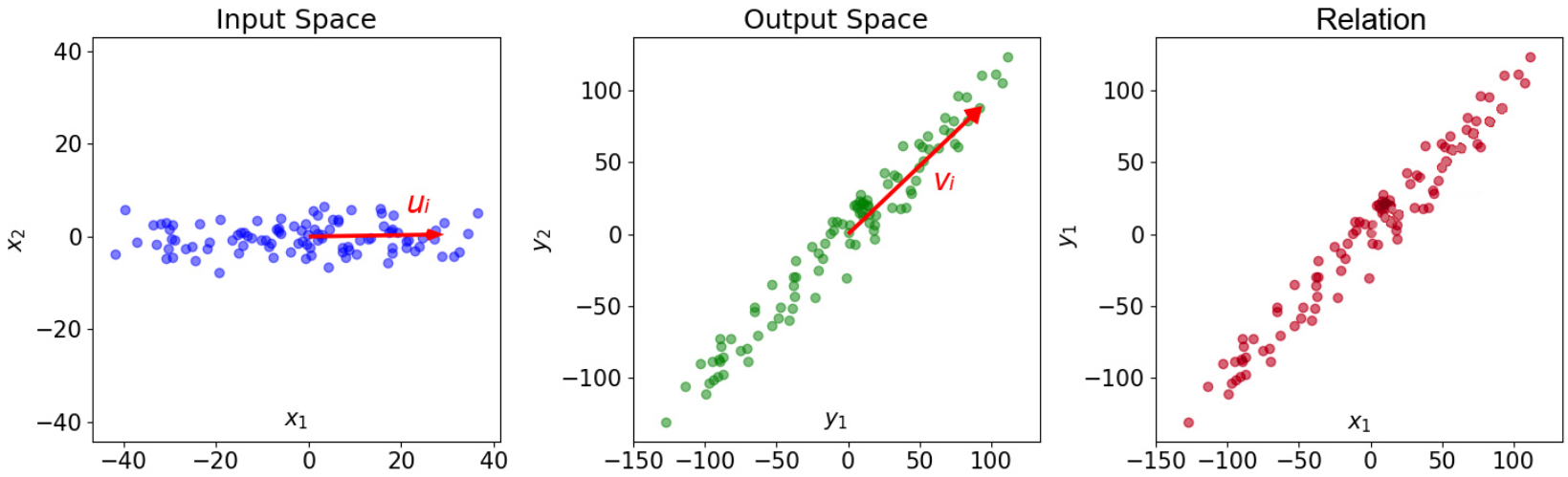


Figure 1: When x_1 increases (captured by u_1), y_1 and y_2 increase (capt. v_1).

PLS

- Same story as PCA, different objective $\underset{||u_i||=1,||v_i||=1}{\text{argmax}} \; u_i^T X^T Y v_i \; \forall i$
  - → Find that $v_i \propto Y^T X u_i$
  - → Rewrite the objective $\underset{||u_i||=1}{\text{argmax}} \, u_i^T X^T Y Y^T X u_i$ (looks like PCA)
  - → Find that $\left(\underbrace{X^T Y}_{Cov(X,Y)} \underbrace{Y^T X}_{Cov(Y,X)} \right) u_i = \lambda_i u_i$
  - → Remove the contributions of u_i, v_i (deflation), and search u_{i+1}, v_{i+1}

$$\Leftrightarrow \quad \left(X^T Y Y^T X\right) u_i = \lambda_i u_i$$

Another eigenproblem!

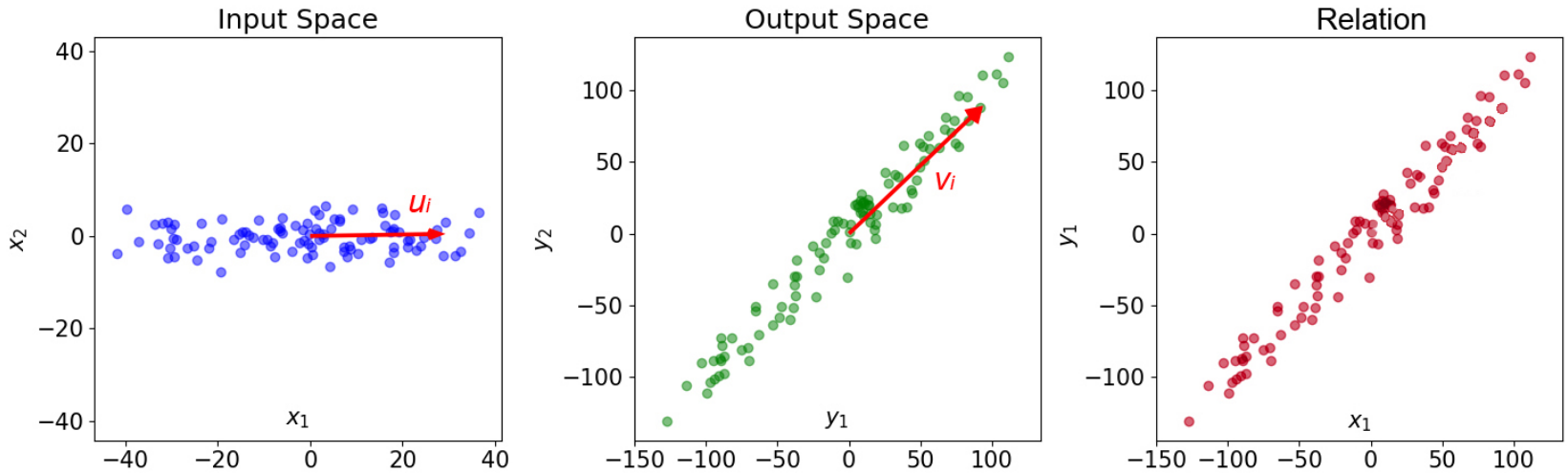
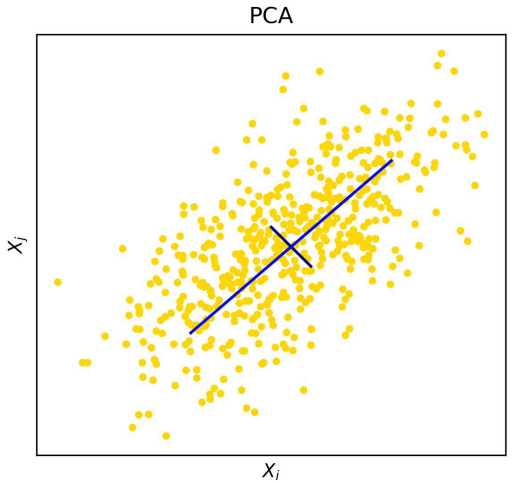


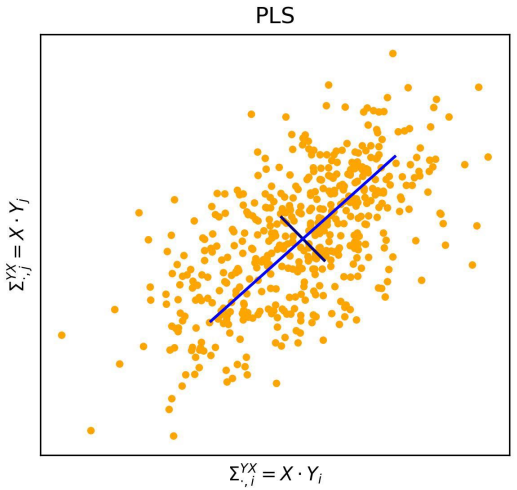
Figure 2: When x_1 increases (captured by u_1), y_1 and y_2 increase (capt. v_1).

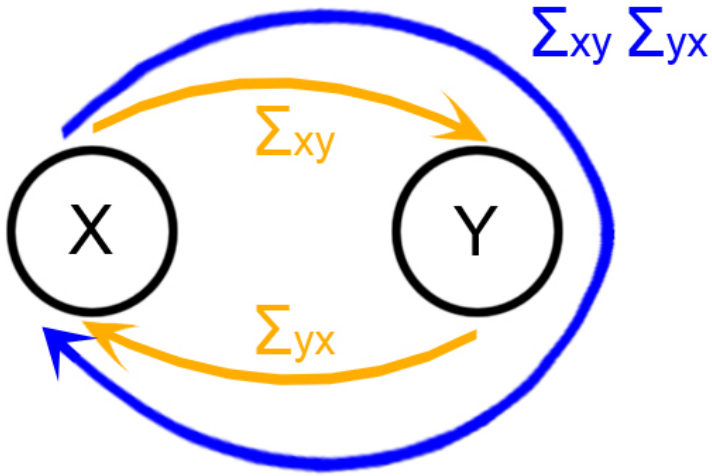
PLS – Directions of what

- PCA searched for principal components of $X^T X = \Sigma_{XX}$
- PLS searches for principal components of $(X^T Y)(Y^T X) = \Sigma_{XY}\Sigma_{YX}$
  - Σ_{YX} quantifies how well Y predicts X
  - $\Sigma_{XY}\Sigma_{YX} \sim Var(\Sigma_{YX})$

  → Principal components are the most expressive directions of a round trip $X \overset{\Sigma_{XY}}{\rightarrow} Y \overset{\Sigma_{YX}}{\rightarrow} X$, and conversely

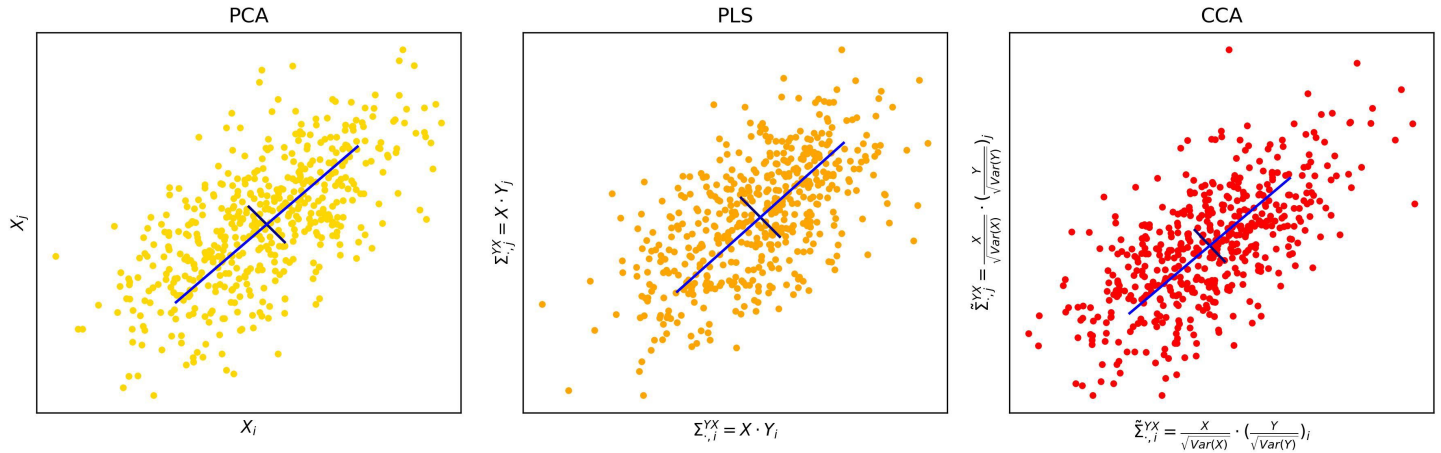






PLS vs CCA

- PLS maximizes covariance $Cov(X, Y)$; CCA maximizes $Corr(X, Y)$
  - $Corr$ is a normalized Cov : $Corr(X, Y) \sim Cov(X, Y)/(||X|| \cdot ||Y||)$
- PLS maximizes covariance $Cov(X, Y)$:
  - × No uniquely defined solution (deflation strategies), sensitive to scale
  - ✓ Robust against colinearities
- CCA maximizes correlation $Corr(X, Y)$:
  - ✓ Uniquely defined solution (*canonical*), no scale considerations
  - × Sensitive to colinearities
- Colinearity: when vars are redundant (if $x_2 = 5x_1$, we only need x_1)
  - Makes $Var(X)$ "zero-like", so $Corr(X, Y)$ is not defined



CCA

- Objective:

$$\underset{Var(XU)=Var(YV)=I}{\operatorname{argmax}} \; u_i^T X^T Y v_i / \sqrt{Var(Xu_i)Var(Yv_i)} \quad \forall i$$
$$= \underset{Var(XU)=Var(YV)=I}{\operatorname{argmax}} \; u_i^T X^T Y v_i \quad \forall i$$

- Resolution: same as PLS
  - → Find that $V \propto Var(Y)^{-1} Y^T X U$
  - → Inject the expression of V in the objective
  - → Find that $\left((X^TX)^{-\frac{1}{2}} X^T Y (Y^TY)^{-1} Y^T X (X^TX)^{-\frac{1}{2}}\right) U' = U'\Lambda$
  - → More simply $\left(\tilde{\Sigma}_{YX}^T \tilde{\Sigma}_{YX}\right) U' = \Lambda U'$ with $U' = \sqrt{Var(X)}U$ → Eigenproblem

$$\Leftrightarrow$$

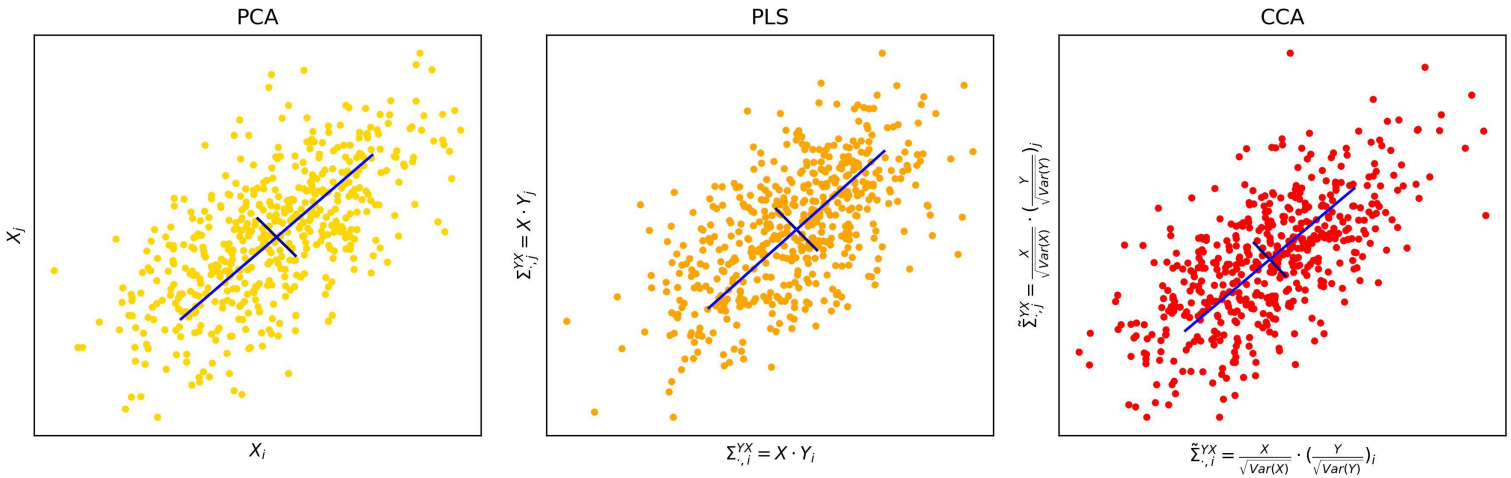
$$\left(\underbrace{(X^TX)^{-\frac{1}{2}} X^T Y (Y^TY)^{-\frac{1}{2}}}_{Corr(X,Y)} \underbrace{(Y^TY)^{-\frac{1}{2}} Y^T X (X^TX)^{-\frac{1}{2}}}_{Corr(Y,X)} \right) U' = U'\Lambda$$

Eigenproblem again!

CCA – Directions of what

- PCA searched for principal components of $X^T X = \Sigma_{XX}$
- PLS searched for the PC of $(X^T Y)(Y^T X) = \Sigma_{XY}\Sigma_{YX}$
- CCA for the PC of $\tilde{\Sigma}_{YX}^T \tilde{\Sigma}_{YX}$, with $\tilde{\Sigma}_{YX}$ is the whitened Σ_{YX}
  - $\tilde{\Sigma}_{YX}$ quantifies how well scale-free and uncorrelated Y predicts X

  → Principal components are the most expressive scale-free and uncorrelated directions of a round trip $X \overset{\tilde{\Sigma}_{YX}^T}{\rightarrow} Y \overset{\tilde{\Sigma}_{YX}}{\rightarrow} X$, and conversely

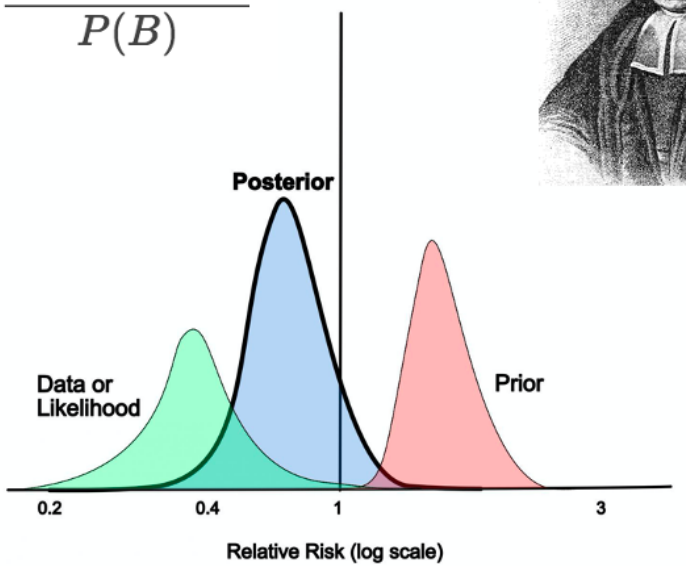


Probabilistic versions

- Probabilistic formulations of these statistic tools exist:
  - p-PCA : $z \sim \mathcal{N}(0, I); \quad X = Uz + \mathcal{N}(0, \sigma^2 I)$
  - p-FA : $z \sim \mathcal{N}(0, I); \quad X = Uz + \mathcal{N}(0, \vec{\psi}_x)$
  - p-PLS : $z \sim \mathcal{N}(0, I); \quad X = Uz + \mathcal{N}(0, \vec{\psi}_x); \quad Y = Vz + \mathcal{N}(0, \vec{\psi}_y)$
  - p-CCA : $z \sim \mathcal{N}(0, I); \quad X = Uz + \mathcal{N}(0, \vec{\psi}_x); \quad Y = Vz + \mathcal{N}(0, \vec{\psi}_y)$

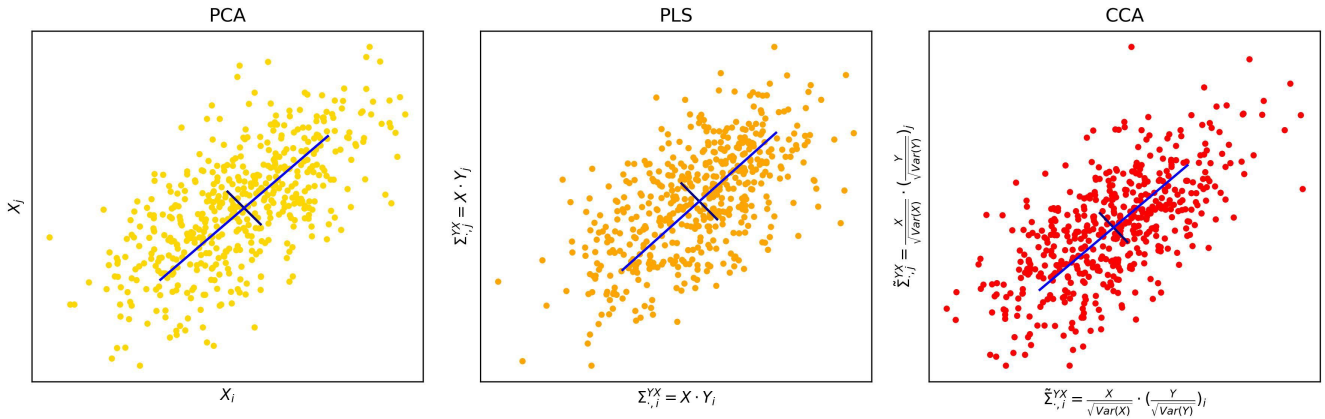
- Useful to:
  - Handle missing data
  - Uncertainty estimation
  - Generate observations
  - Estimate distributions
  - Bayesian frameworks
  - ...

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Summary

- PCA: variance in **one** dataset
  - → "Which parts of X vary the most?"
  - → Directions in $Var(X)$
- PLS: covariance between **two** datasets
  - → "Which parts of X helps predict Y the most?"
  - → Directions in $Var(X^T Y)$ and $Var(Y^T X)$
- CCA: correlation between **two** datasets
  - → "Which standardized parts of X helps predict Y the most?"
  - → Directions in $Var((X^T X)^{\frac{1}{2}} X^T Y (Y^T Y)^{\frac{1}{2}})$ and conversely
- More than two datasets?



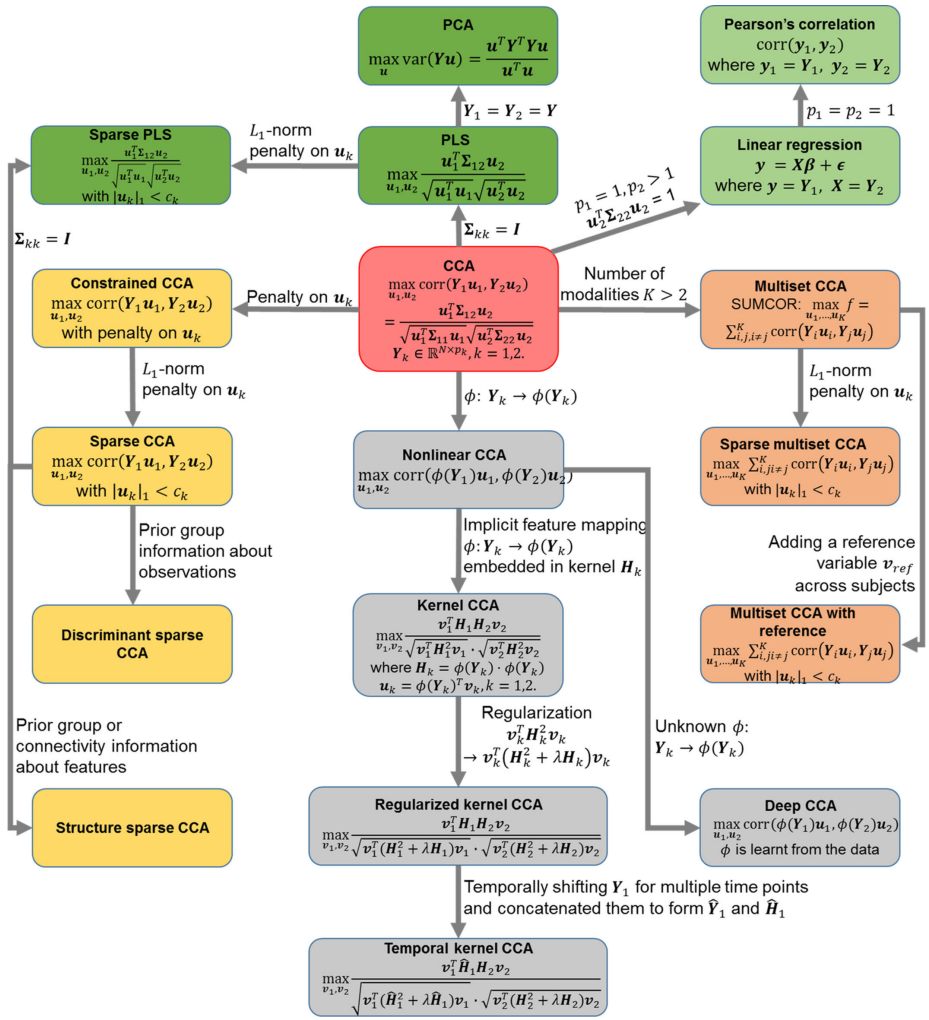
gCCA

- We might want to relate more than two datasets at once:
  → Ex: metabolic profile, lifestyle habits, environmental exposures
- Problem: $Cov(A, B, C)$ does not exist.
  No definition of covariance for more than two variables
- We can use tricks to force the covariance to accommodate more than three datasets $X_1, ..., X_n$:
  - SUMCORR: $\text{argmax}_{w_i, w_j} \Sigma_{i<j}^{n} Corr(X_i w_i, X_j w_j)$ (pariwise)
  - SSQCOR: $\text{argmax}_{w_i, w_j} \Sigma_{i<j}^{n} Corr(X_i w_i, X_j w_j)^2$ (pariwise)
  - MAXVAR: $\text{argmax}_{z, w_i} \Sigma_{i<j}^{n} Corr(X_i w_i, z)$ (shared representation)

$$\Sigma = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X,Y) & \text{Cov}(X,Z) \\ \text{Cov}(X,Y) & \text{Var}(Y) & \text{Cov}(Y,Z) \\ \text{Cov}(X,Z) & \text{Cov}(Y,Z) & \text{Var}(Z) \end{bmatrix}$$

More iterations

- Many iterations:
  - Regularization
  - Sparsity
  - Supervised learning
  - Block handling
  - Nonlinearity
  - ...
- Widely used state-of-the-art models as a combination of those:
  - MOFA: FA + sparsity + shared and private latent factors.
  - DIABLO: SUMCORR-gCCA + sparsity + supervision



Thanks for your attention!