PCA (Karhunen - Loève Transform)

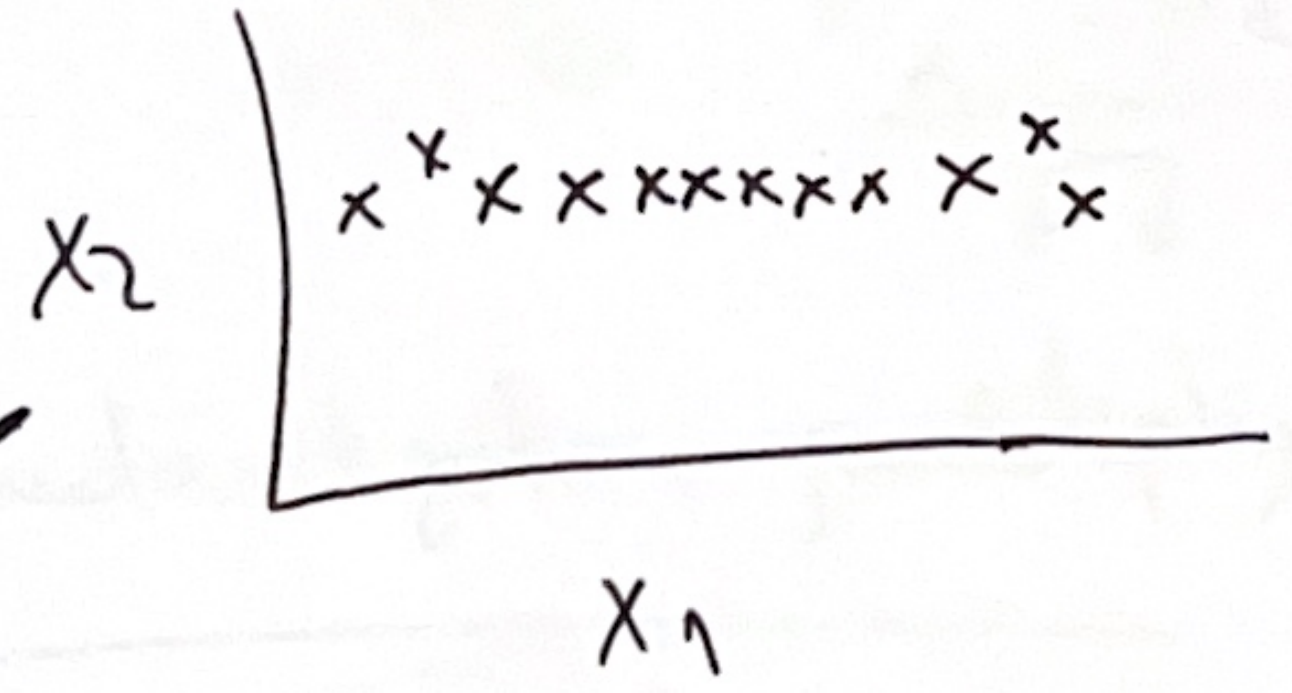
* high-dim data: images
  - hard to analyze
  - hard to interpret
  - hard to visualize
  - expensive to store

* dimensionality reduction exploits
  - structure
  - correlation

  $\Downarrow$

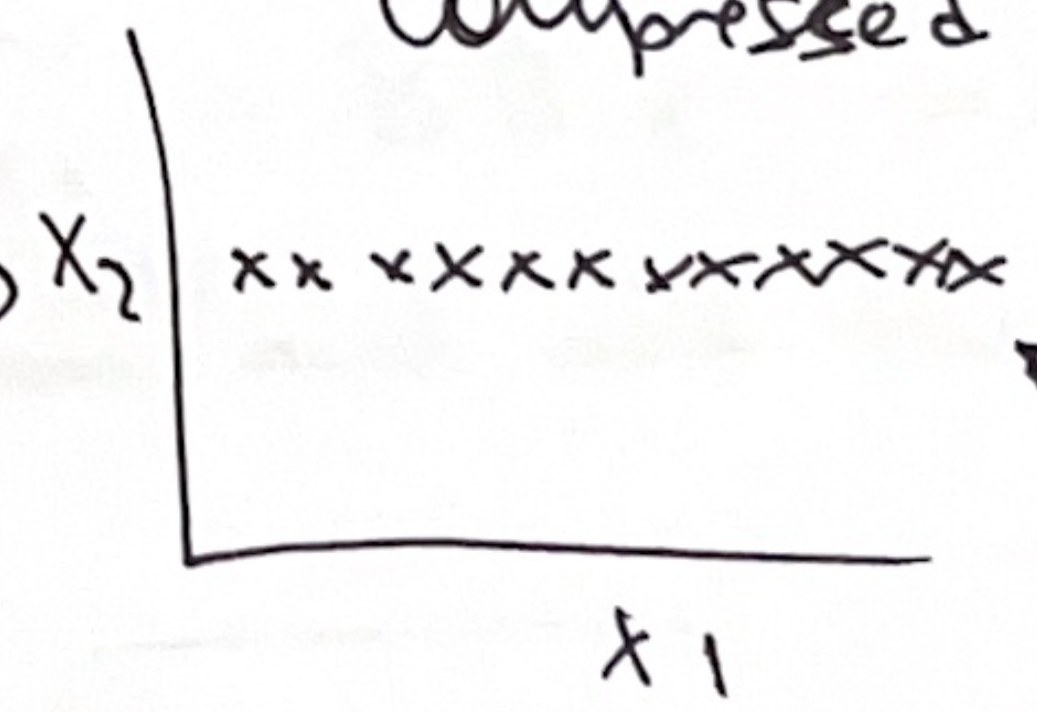
  compact/compressed representation of the data

* similar to (lossy) compression : jpg, mp3

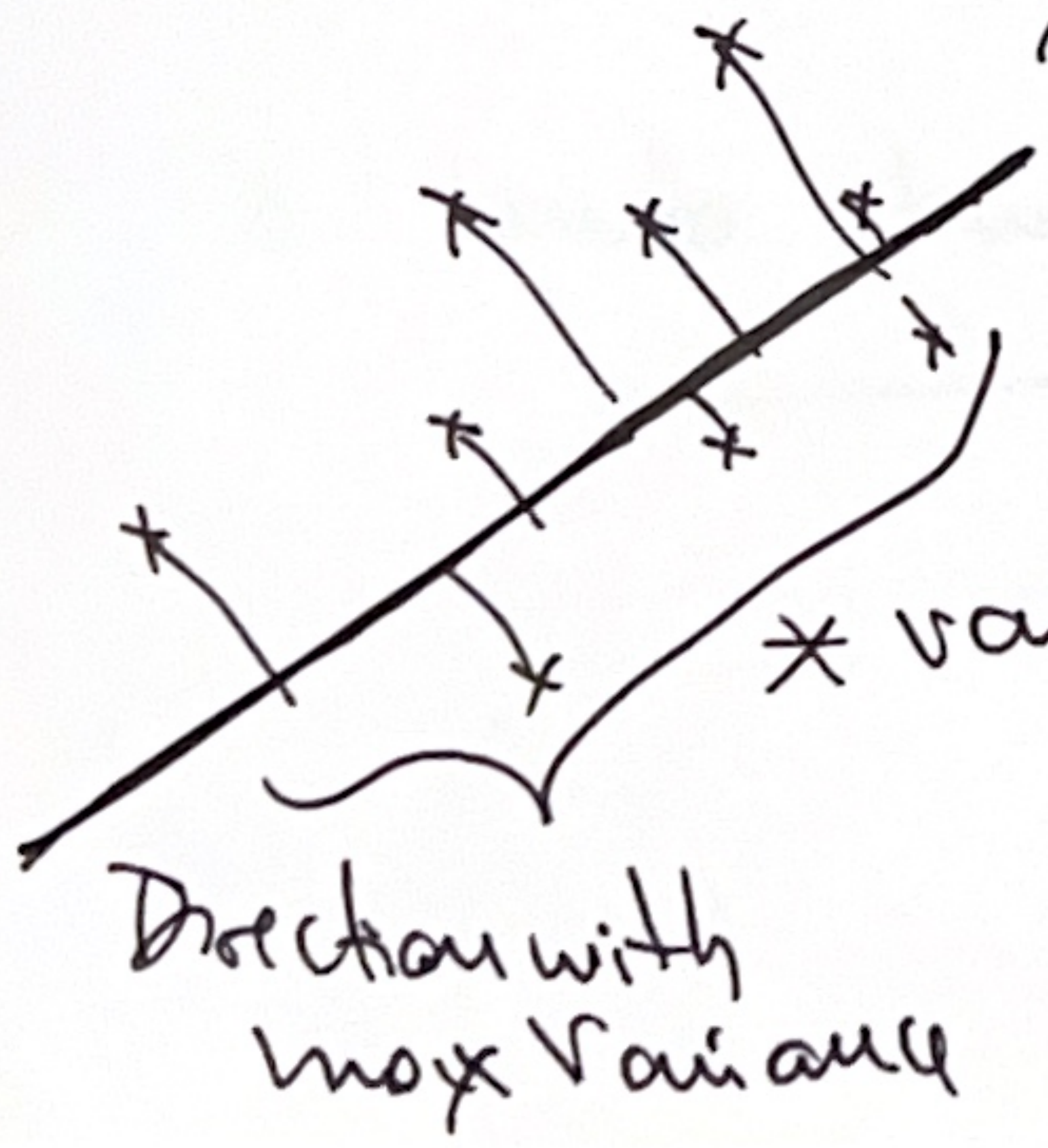


$\Longrightarrow$

compressed



data does not vary much on X_2 axis



* variance is an indicator of data spread $\Longrightarrow$ PCA maximizes the variance in the low-dim space to retain as much information as possible

PCA Steps

1. Mean subtraction to avoid numerical problems

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)}$$ and Normalization

2. Standardization $\Longrightarrow$ finds directions that maximize variance

$\Longrightarrow$ makes all vars have same variance $\simeq 1$

$z = \frac{x-\mu}{\sigma}$ $\Longrightarrow$ unit free

3. Eigen decomposition of the covariance Matrix

$$\text{cov } \cancel{\text{matrix}} = \begin{bmatrix} \underbrace{cov(x,x)}_{var(x)} & cov(x,y) \\ cov(y,x) & \underbrace{cov(y,y)}_{var(y)} \end{bmatrix}, \quad cov(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{N}$$

$$S = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$$

4. Projection

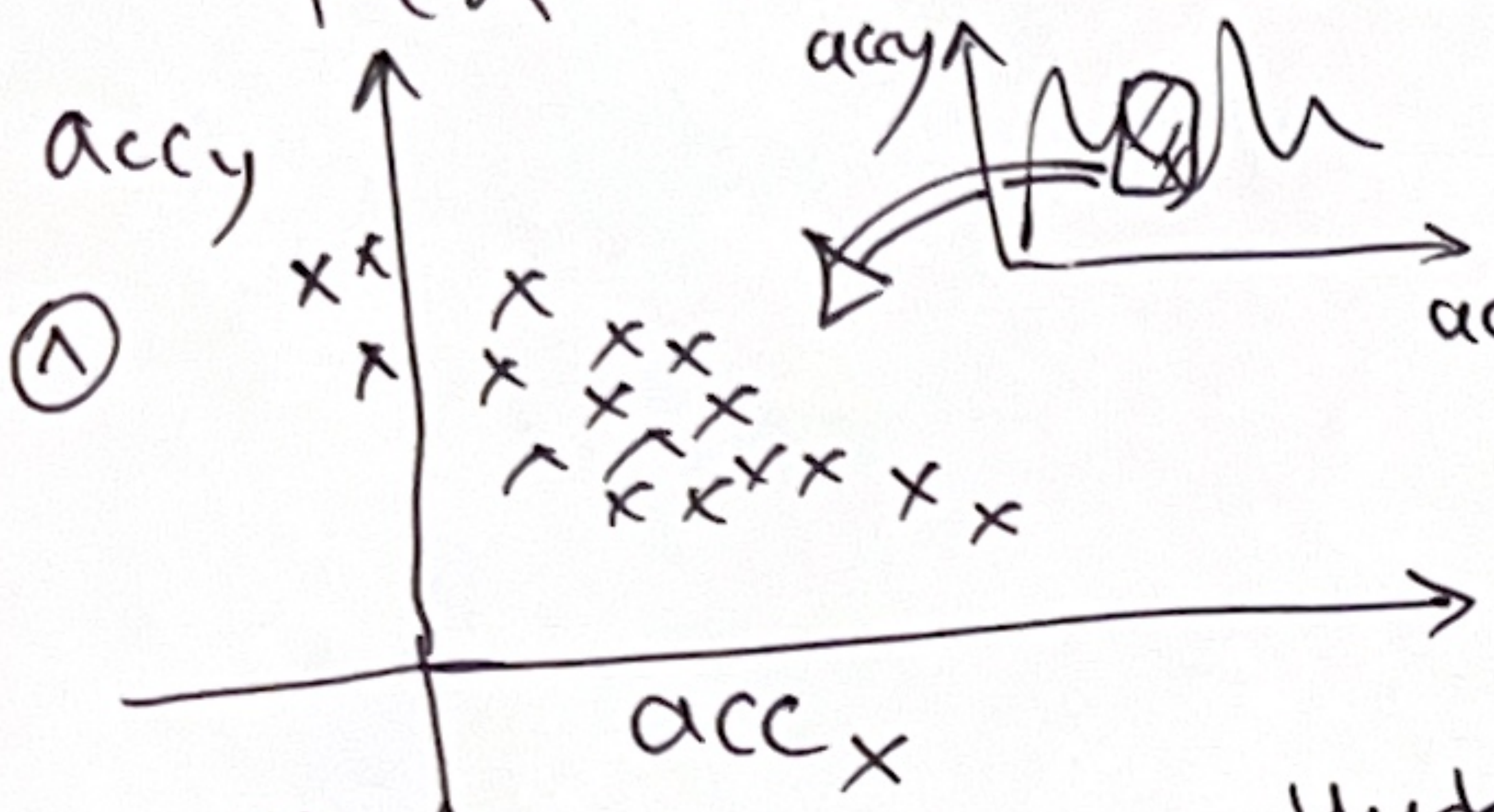
$$x_*^{(d)} \leftarrow \frac{x_*^{(d)} - \mu_d}{\sigma_d} \qquad \tilde{x}_* = BB^T x \implies \tilde{x}_*^{(d)} \leftarrow \tilde{x}_*^{(d)}\sigma_d + \mu_d$$

5. Undo standardization multiply by ~~mean~~ std.dev and add the mean

Loadings = Eigenvectors $\cdot$ $\sqrt{\text{Eigenvectors}}$

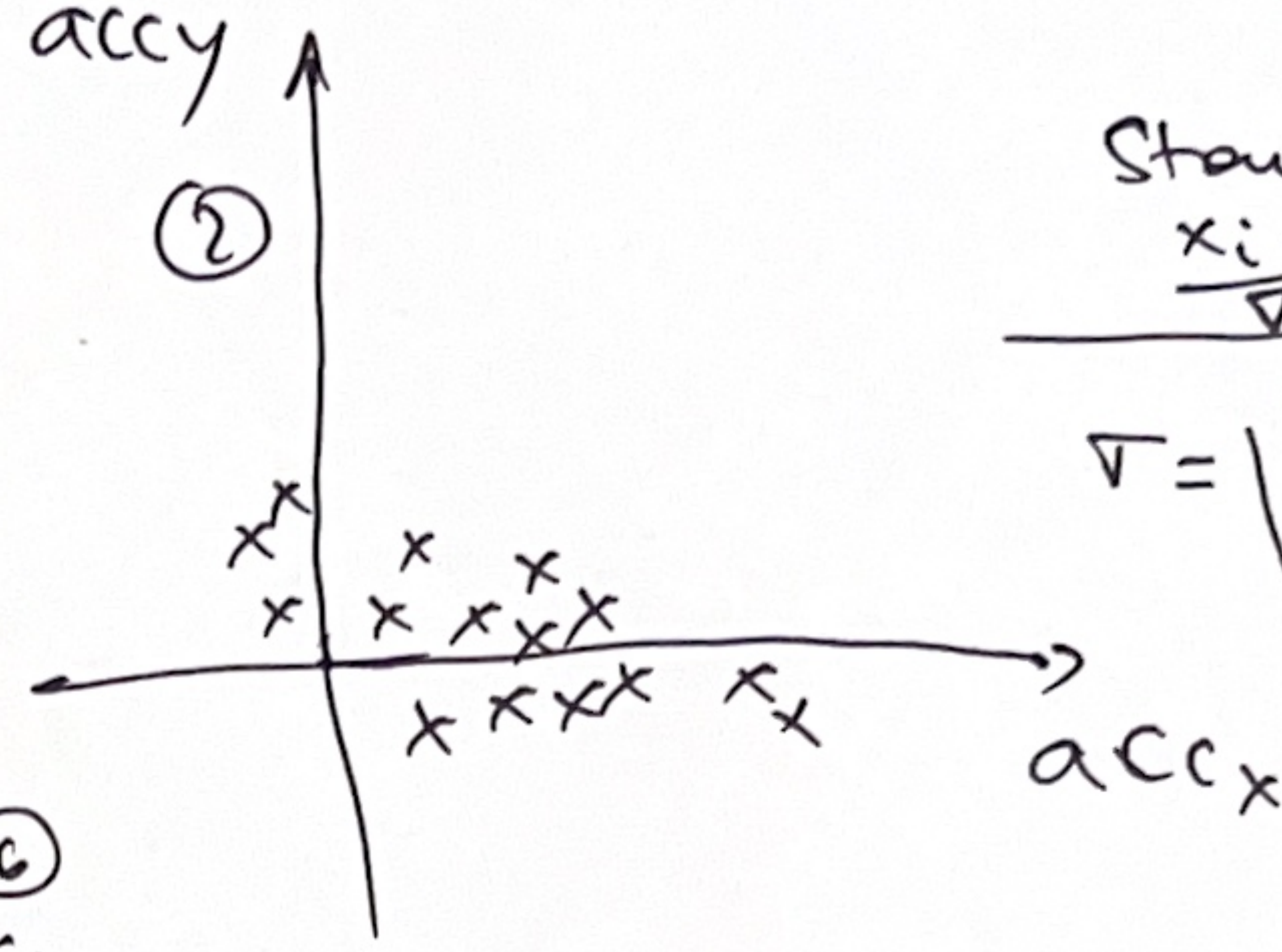
Covariance Matrix $\xrightarrow{\text{decomposition}}$ Scale part (eigenvalues) / direction part (eigenvectors)

PCA in der Praxis



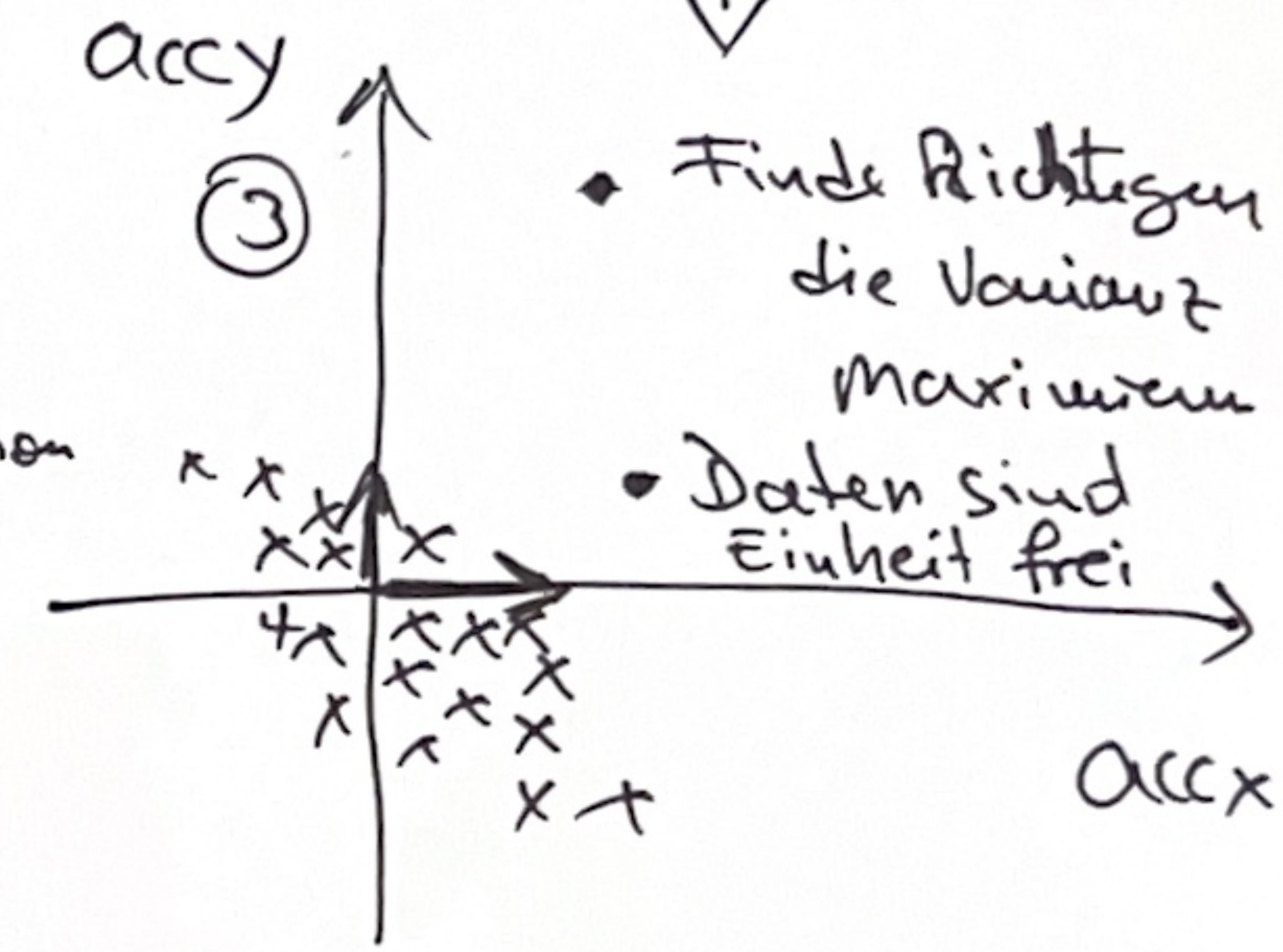
Centering $x_i - \mu$

$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$



Standardization $\frac{x_i - \mu}{\sigma}$

$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)}$



- Finde Richtungen die Varianz maximum
- Daten sind Einheit frei

Eigendecomposition

* Elipse ist Kovarianz Matrix

$$\mathrm{Cov} = \begin{bmatrix} \overbrace{\mathrm{cov}(x,x)}^{\mathrm{var}(x)} & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & \underbrace{\mathrm{cov}(y,y)}_{\mathrm{var}(y)} \end{bmatrix}$$

* Compute Eigenvectors & Eigenvalues (Pfeile)

$$S = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T = \frac{1}{N} X X^T$$

SVD on S to get v and λ

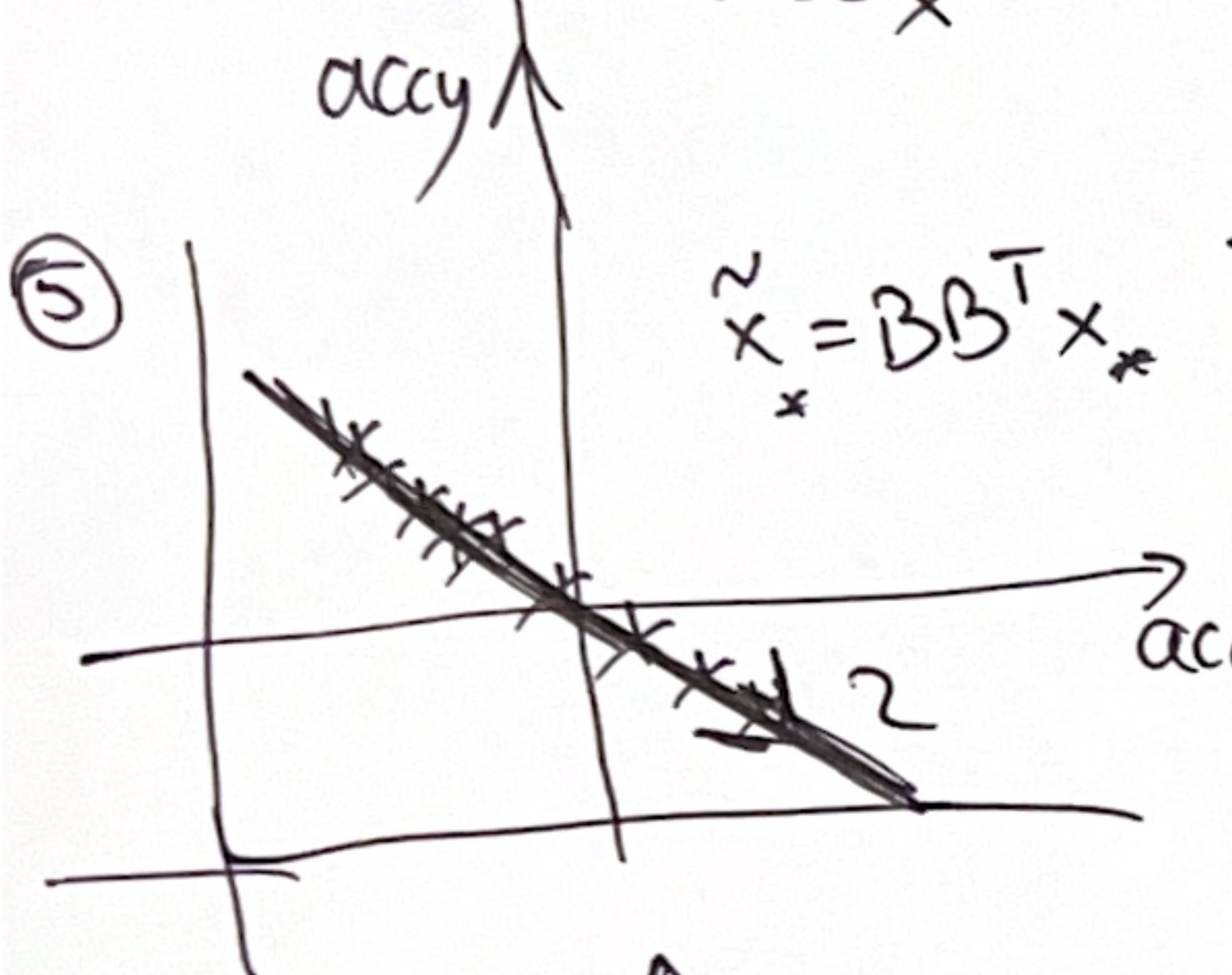
$$\begin{cases} S v_1 = \lambda_1 v_1 \\ S v_2 = \lambda_2 v_2 \end{cases}$$

B matrix ← eigenvectors with highest eigenvalues (principal subspace)

Select $m < d$

Projection

$\tilde{x}_* = B B^T x_*$



undo standardization
multiply σ and add μ