# Statistical Data Analysis

Dr. Jana de Wiljes

18. Januar 2022

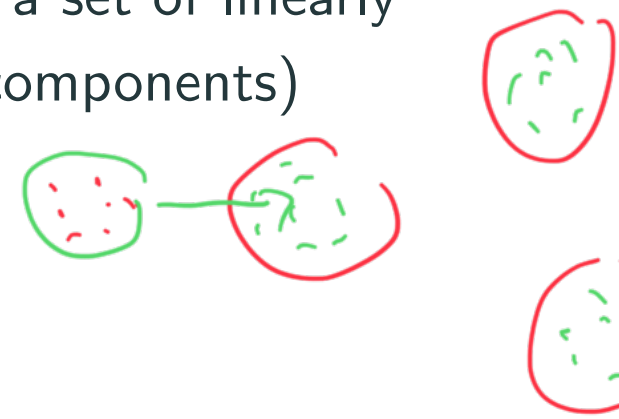
Universität Potsdam

# Principal Component Analysis

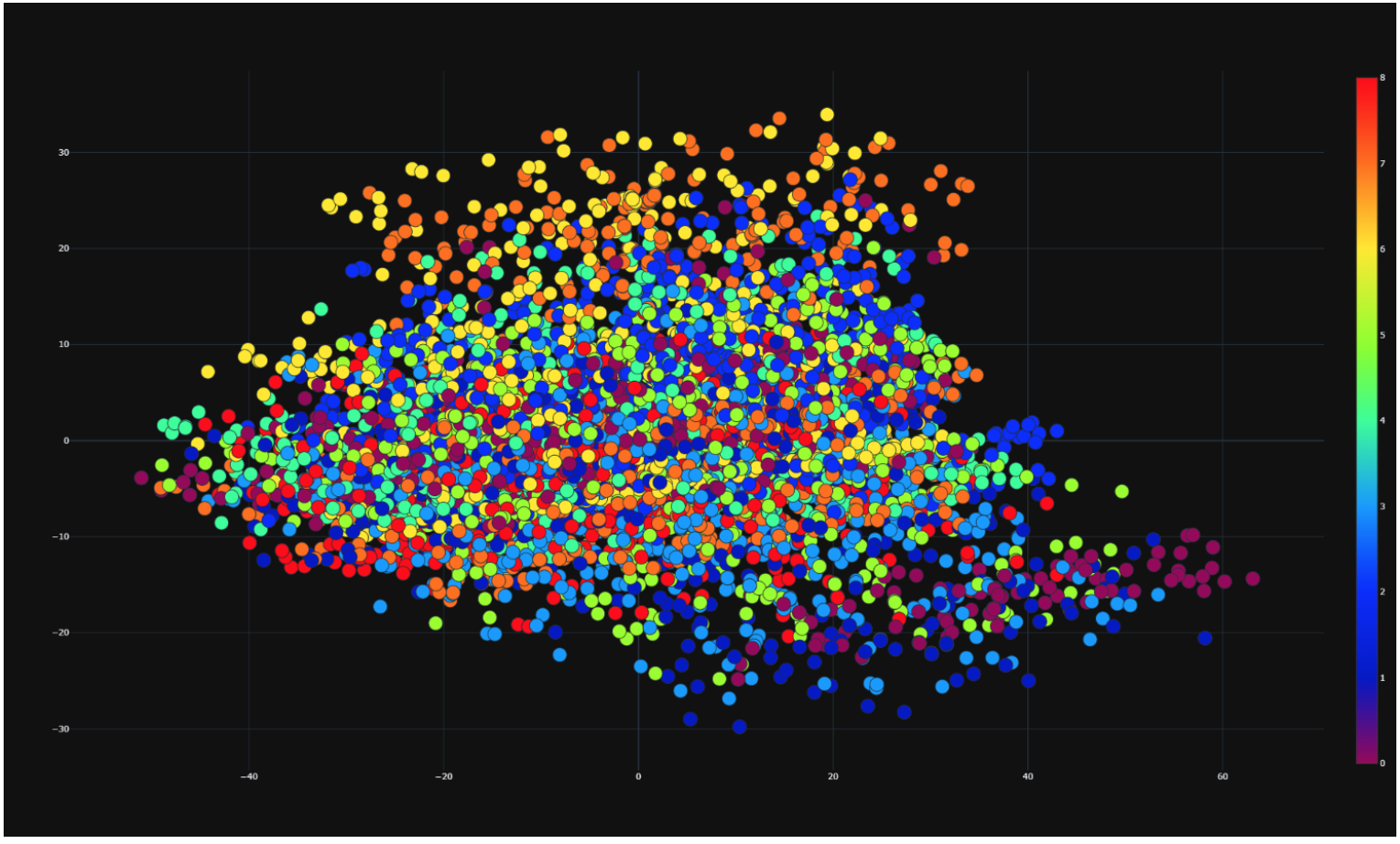
# Principal Component Analysis (PCA)

**Snapshot information:**

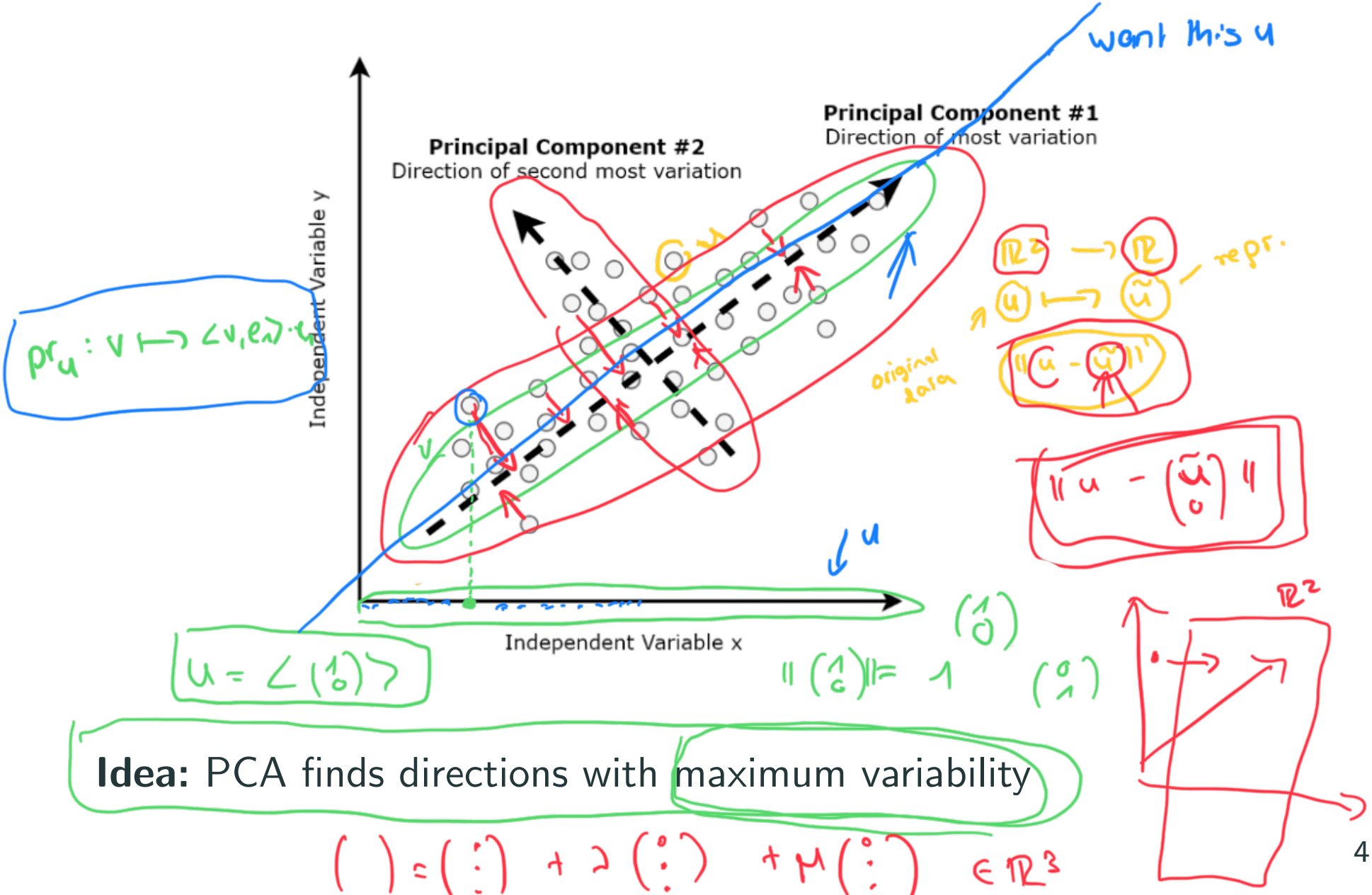
1. Dimensionality reduction, i.e., represent it in a more tractable, lower-dimensional form, without losing too much information
   - Data Compression (Save computation/memory) ~~Noise Reduction~~
   - Noise Reduction/ avoid overfitting to noise
   - Data Visualization (e.g., in two dimensions)
2. unsupervised learning algorithm (Pearson 1901)
3. Idea: uses an orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables (called principal components)
4. Known under many different names:
   - Karhunen-Love transformation
   - Hotelling transformation
   - empirical orthogonal functions



# Visualisation

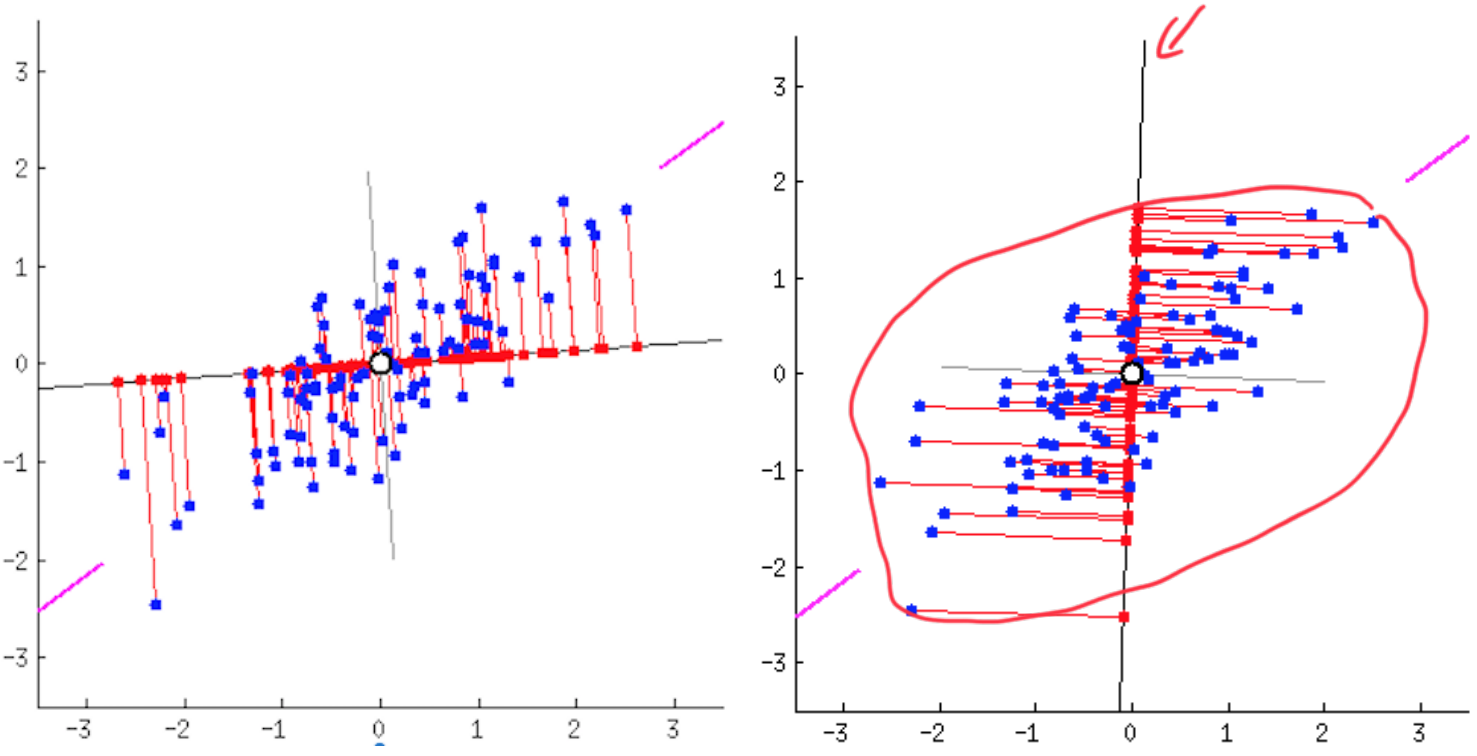


# Motivation behind PCA




**Idea:** PCA finds directions with maximum variability

# Motivation behind PCA




**PCA:**

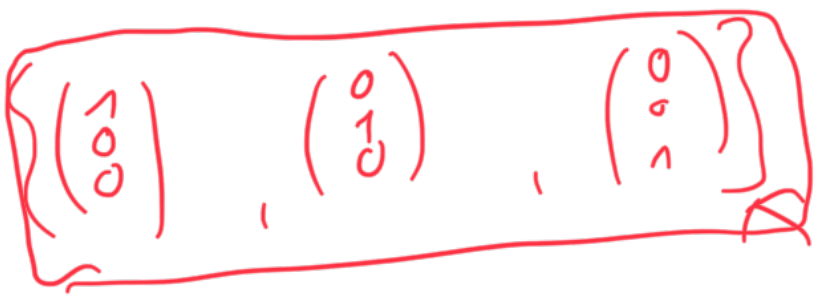
- Identify a Hyperplane that lies closest to the data
- Project the data onto the hyperplane.

# Background information

# Orthogonal basis

**Def:** Let $V$ be a vector space with scalar product $\langle \cdot, \cdot \rangle$ and $\{v_j\}_{j \in J}$ a family of vectors.

- $\{v_j\}_{j \in J}$ are a **orthogonal system**, if $\langle v_j, v_k \rangle = 0 \; \forall j \neq k \in J$ and $v_j \neq 0 \; \forall j \in J$.
- $\{v_j\}_{j \in J}$ is an **orthonormal system**, if additionally: $\langle v_j, v_j \rangle = 1 \; \forall j \; (\Leftrightarrow \|v_j\| = 1)$, in other words: $\langle v_j, v_k \rangle = \delta_{jk} \; \forall j, k \in J$.
- An Orthogonal- respectively. -normalsystem is called **orthogonal basis** bzw. **orthonormal basis**, if the die vectors of the systems form a basis.



# Representation with respect to an orthonormal basis

**Theorem**

*Let* $\{v_j\}_{j=1,\dots,n}$ *be an orthonormal basis of the vector space* $V$ *and* $w \in V$ *a second vector. Then the following holds:*

$$w = \langle v_1, w \rangle v_1 + \cdots + \langle v_n, w \rangle v_n.$$

# Proof

- Compute the difference $u = w - \sum_{j=1}^{n} \langle v_j, w \rangle v_j$
- apply the scalar product $\langle v_k, \cdot \rangle$ to it

$$\langle v_k, u \rangle = \langle v_k, w - \sum_{j=1}^{n} \langle v_j, w \rangle v_j \rangle$$

$$= \langle v_k, w \rangle - \sum_{j=1}^{n} \langle v_k, \underbrace{\langle v_j, w \rangle}_{\text{scalar}} v_j \rangle \qquad \langle x, \lambda y \rangle$$

$$= \langle v_k, w \rangle - \sum_{j=1}^{n} \langle v_j, w \rangle \langle v_k, v_j \rangle$$

$$= \langle v_k, w \rangle - \langle v_k, w \rangle = 0 \qquad \begin{cases} 1 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}$$

- $\Rightarrow \langle u, v_k \rangle = 0 \quad \forall k \Rightarrow \langle u, \sum_{i=1}^{n} \lambda_i v_i \rangle = 0$

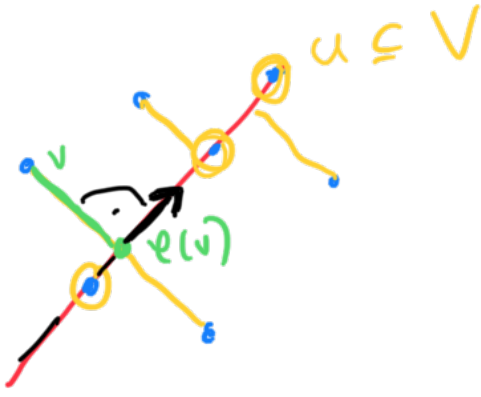
$u = \sum \lambda_i v_i$ — can find Basis

$\Rightarrow \langle u, u \rangle = 0 \Rightarrow \|u\| = 0 \Rightarrow u = 0$ □

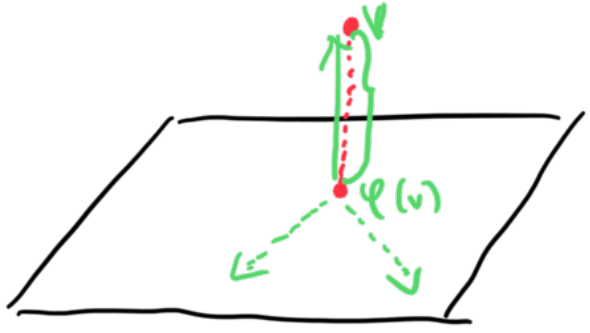
# Orthogonal Projection

**Def:** Let $U \subseteq V$ be a subspace. A map $\varphi\colon V \to V$ is called **projection of $V$ onto $U$**, if for every $u \in U$ gilt: $\varphi(u) = u$. A projection is called **orthogonal projection onto subspace $U$**, if for every vector $v \in V$ holds:

$$(\varphi(v) - v) \perp U.$$




Example: Hyperspace



# Orthonormalbasis

**Def:** Let $V$ be a $K$-vector space with scalar product and $U \subseteq V$ a finite-dimensional subspace. Furthermore let $\{u_1, \ldots, u_k\}$ be an orthonormal basis of $U$. Then the map

$$\mathrm{pr}_U : V \to U, \quad v \mapsto \sum_{j=1}^{k} \langle u_j, v \rangle u_j$$

is an orthogonal projection.

$\{u_1, \ldots, u_k\}$ linearly independent

then one can always extend to a basis of $V$

$\left\langle \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix} \right\rangle \begin{pmatrix}0\\0\\1\end{pmatrix}$

$e_1 \quad e_2 \quad e_3$

$\sum_{j=1}^{n} \langle u_j, v \rangle \cdot u_j \in U$?

scalar $\in U$

$= U$

$v = \sum_{i=1}^{3} \langle e_i, v \rangle e_i$

$v \approx \sum_{i=1}^{2} \langle e_i, v \rangle \cdot e_i$

$u, v \in U \Rightarrow u + v \in U$

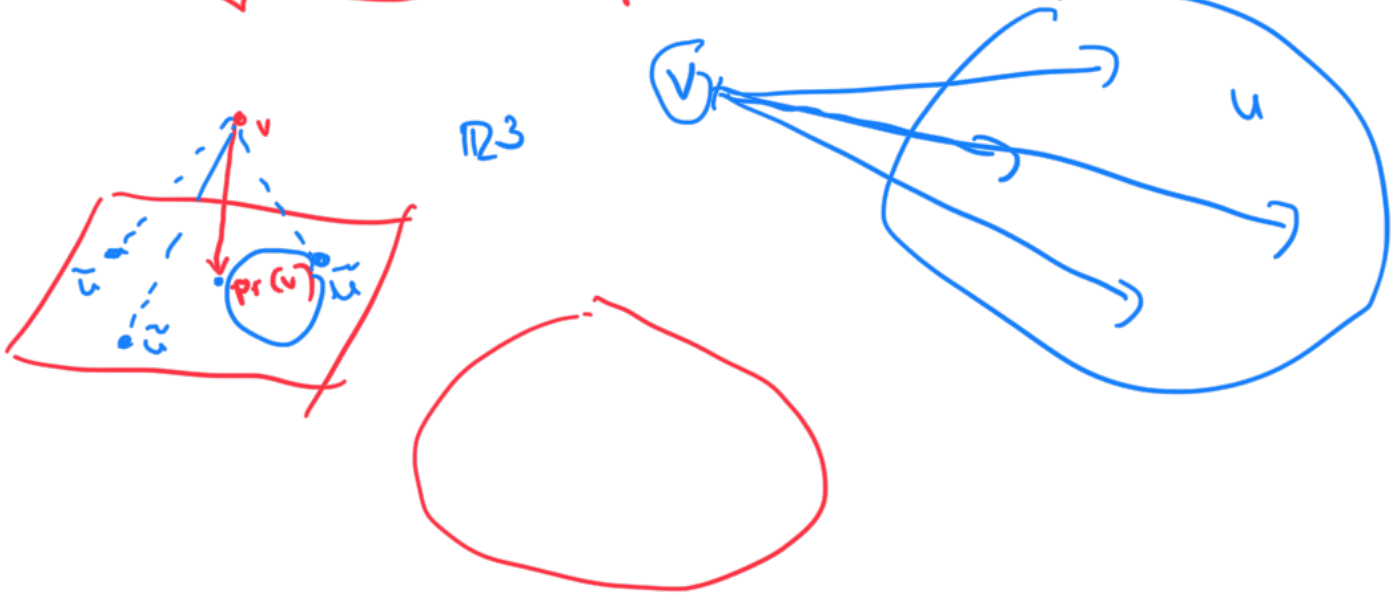
$u \in U \quad \lambda \in \mathbb{R} \quad \lambda \cdot u \in U$

# Approximation theorem

**Theorem**

*Let $V$ be a $\mathbb{R}$-vector space, with a scalar product and the corresponding norm $\|\cdot\|$. Let $U$ be a subspace of $V$. Then for every $v \in V$ $pr_U(v)$ is the best approximation of $v$ in $U$, i.e.,*

$$\|v - pr_U(v)\| < \|v - u\| \; \forall u \in U \text{ mit } u \neq pr_U(v).$$



# PCA

**Given:** data set of p dimensional vectors

$x_i \in \mathbb{R}^p$ want subspace $U \cong \mathbb{R}^q$

$X$ $n \times p$ dimensional

**Goal:** Want to project them to q- dimensional subspace ($q << p$)

$\|w\| = 1$

**Principal components:** q uncorrelated, orthogonal directions formed by projecting the original data


$y = \langle e_1 \rangle$

$U = \langle w \rangle$

$\|w\| = 1$

# Derivation

Special case: Data $x_i \in \mathbb{R}^p$ and there are $n$ data points

⚠ Assume: Data is normalised, i.e. $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$

Want to project to a line (i.e. $q = 1$)

$$\|x_i - \langle w, x_i \rangle w\|^2 = \langle x_i - \langle w, x_i \rangle w,\ x_i - \langle w, x_i \rangle w \rangle$$

$x_i$: Data point; $\langle w, x_i \rangle w$: $\text{pr}_u(x_i)$

$$= \langle x_i, x_i \rangle - \langle x_i, \langle w, x_i \rangle w \rangle - \langle \langle w, x_i \rangle w, x_i \rangle + \langle \langle w, x_i \rangle w,\ \langle w, x_i \rangle \cdot w \rangle$$

$u = \langle w \rangle$ ↑ Orthonormal basis vector

$\langle w, x_i \rangle$: scalar; $\langle w, x_i \rangle$: scalar

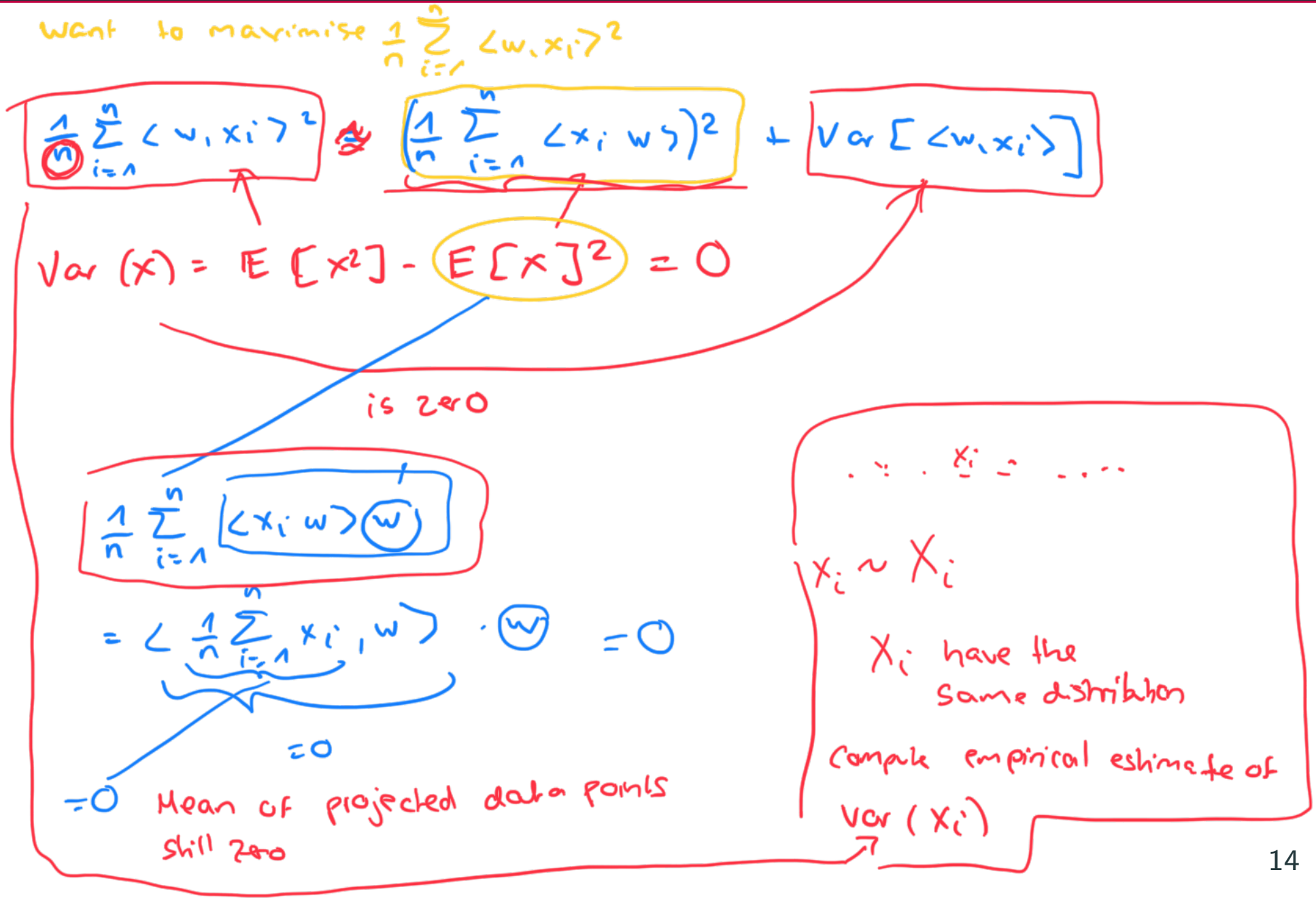
$$= \|x\|^2 - 2\langle w, x_i \rangle^2 + \langle w, x_i \rangle^2 \cdot \underbrace{\langle w, w \rangle}_{1}$$

$$= \|x\|^2 - \langle w, x_i \rangle^2$$

$$\text{MSE}(w) = \frac{1}{n}\sum_{i=1}^{n} \left( \|x_i\|^2 - \langle w, x_i \rangle^2 \right)$$

↑ want a $w$ that maximizes this term here

# Derivation


want to maximise $\frac{1}{n}\sum_{i=1}^{n}\langle w, x_i\rangle^2$
$\frac{1}{n}\sum_{i=1}^{n}\langle w, x_i\rangle^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\langle x_i, w\rangle\right)^2 + \mathrm{Var}[\langle w, x_i\rangle]$
$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 0$
is zero
$\frac{1}{n}\sum_{i=1}^{n}\langle x_i, w\rangle w$
$= \left\langle \frac{1}{n}\sum_{i=1}^{n} x_i, w \right\rangle \cdot w = 0$
$= 0$
$= 0$ Mean of projected data points still zero
$x_i \sim X_i$
$X_i$ have the same distribution
Compute empirical estimate of $\mathrm{Var}(X_i)$

# Derivation

Maximises: $Var[\langle w_1, x_i \rangle]$ $\forall i$ (X) $n \times p$

$x_i \sim X_i$ but all $x_i$ have distribution

$$= \frac{1}{n} \sum_{i=1}^{n} \langle w, x_i \rangle^2$$

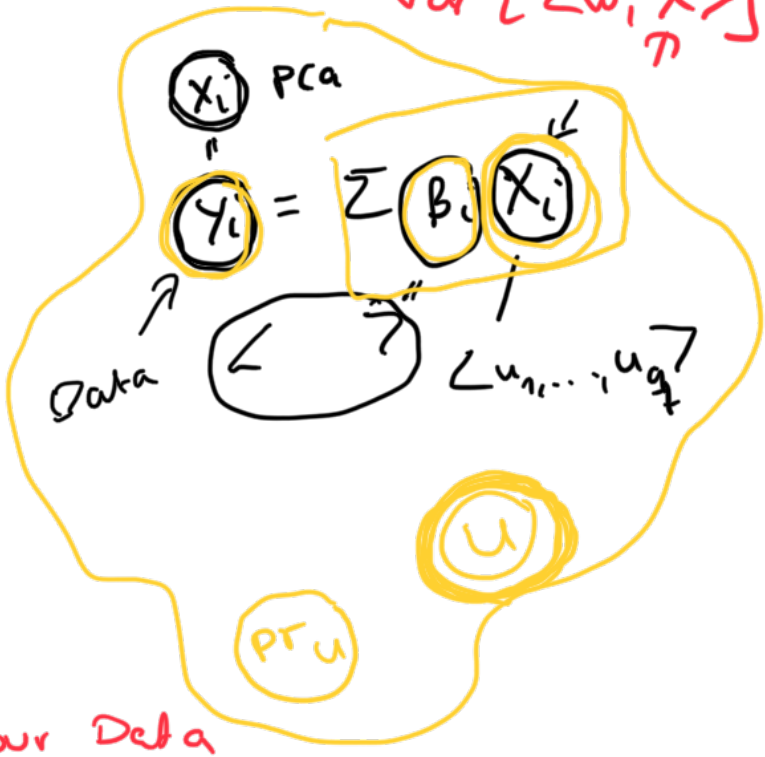
$Var[\langle w_1 X \rangle]$

$$Var(w, X) = \frac{1}{n} \sum \langle x_i \cdot w \rangle^2$$

$$= \frac{1}{n} (Xw)^T (Xw)$$

$$= \frac{1}{n} w^T X^T X w$$

$$= w^T \frac{X^T X}{n} w$$




emp. Covariance of our Data

$\|w\| = 1$

$$\Rightarrow w^T w = 1 \; ; \; \frac{X^T X}{n} w = \lambda w$$

$$\mathcal{L}(w, \lambda) = w^T \frac{X^T X}{n} w - \lambda (w^T w - 1)$$

larger value

# Derivation

# Derivation