One-dimensional PCA

Archer Yang

McGill University

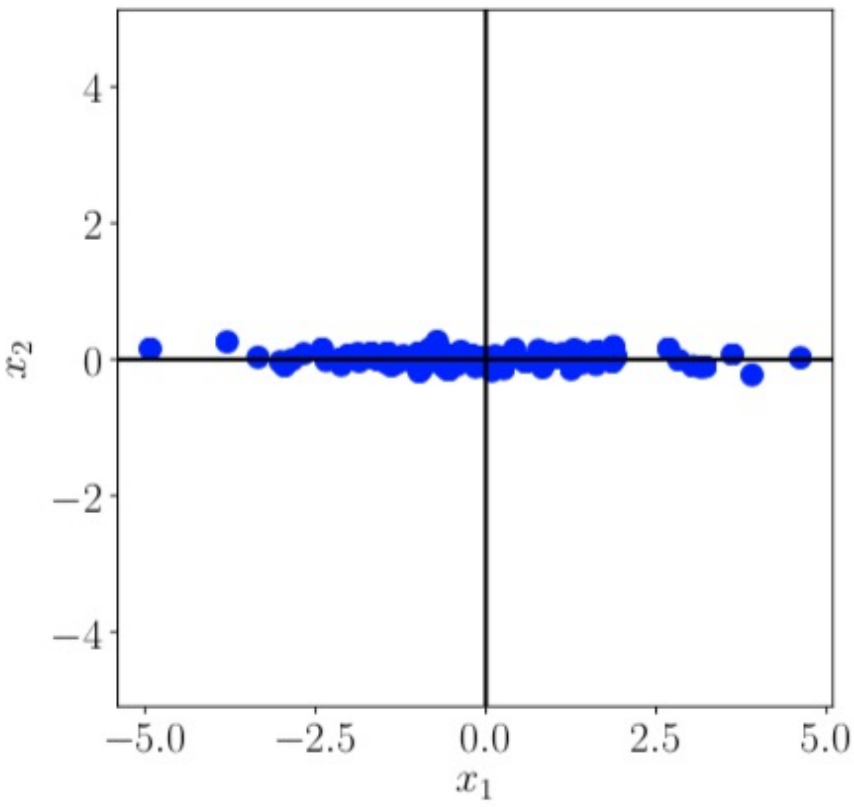
October 4, 2024

- **Reading assignment:** read chapter 10 of Mathematics for Machine Learning book.

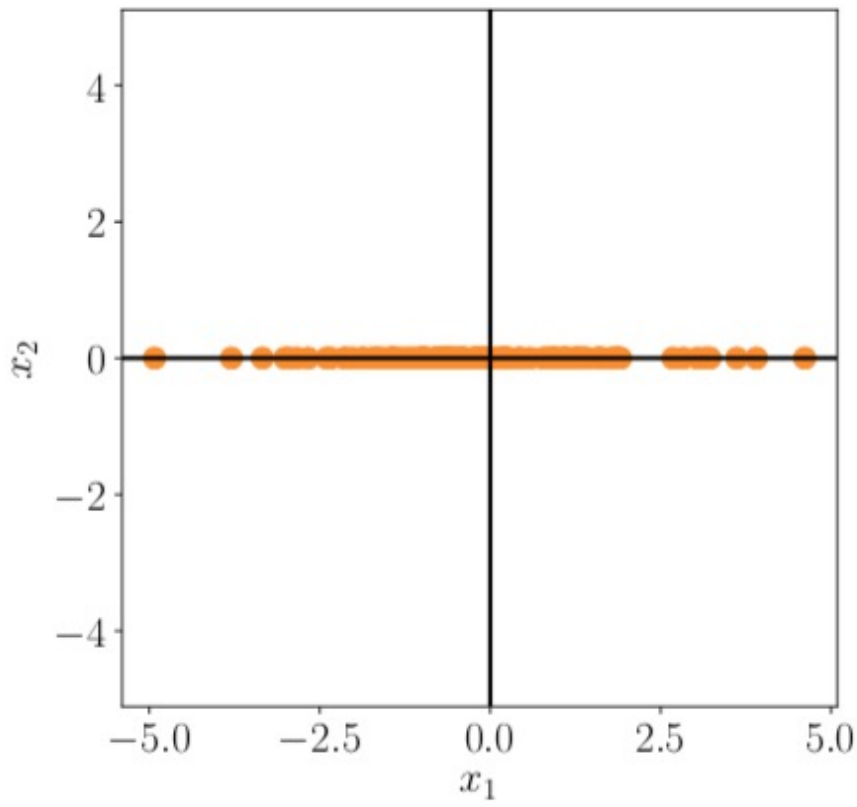
Dimension reduction

- High-dimensional data is very common in real-life applications, e.g. in genomics, finance, e-commerce etc.
- Dimensions in high-dimensional data are often **correlated** — it has an **embedded lower-dimensional structure**.
- Dimensionality reduction exploits such low dimensional structure
  - Compression
  - Visualization
  - Data generation
- We study **principal component analysis** (PCA) – linear dimensionality reduction by Pearson (1901) and Hotelling (1933).
- One of the most commonly used techniques for **compression** and **visualization**.

Dimension reduction



(a) Dataset with x_1 and x_2 coordinates.



(b) Compressed dataset where only the x_1 coordinate is relevant.

Derive PCA from First Principles

Problem Setting

- Consider an i.i.d. dataset

$$\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$$

  where $\boldsymbol{x}_n = (x_{n1}, \cdots, x_{np})^\top \in \mathbb{R}^p$ for $n = 1, \cdots, N$.

- For simplicity, we assume that the data is **centered** such that

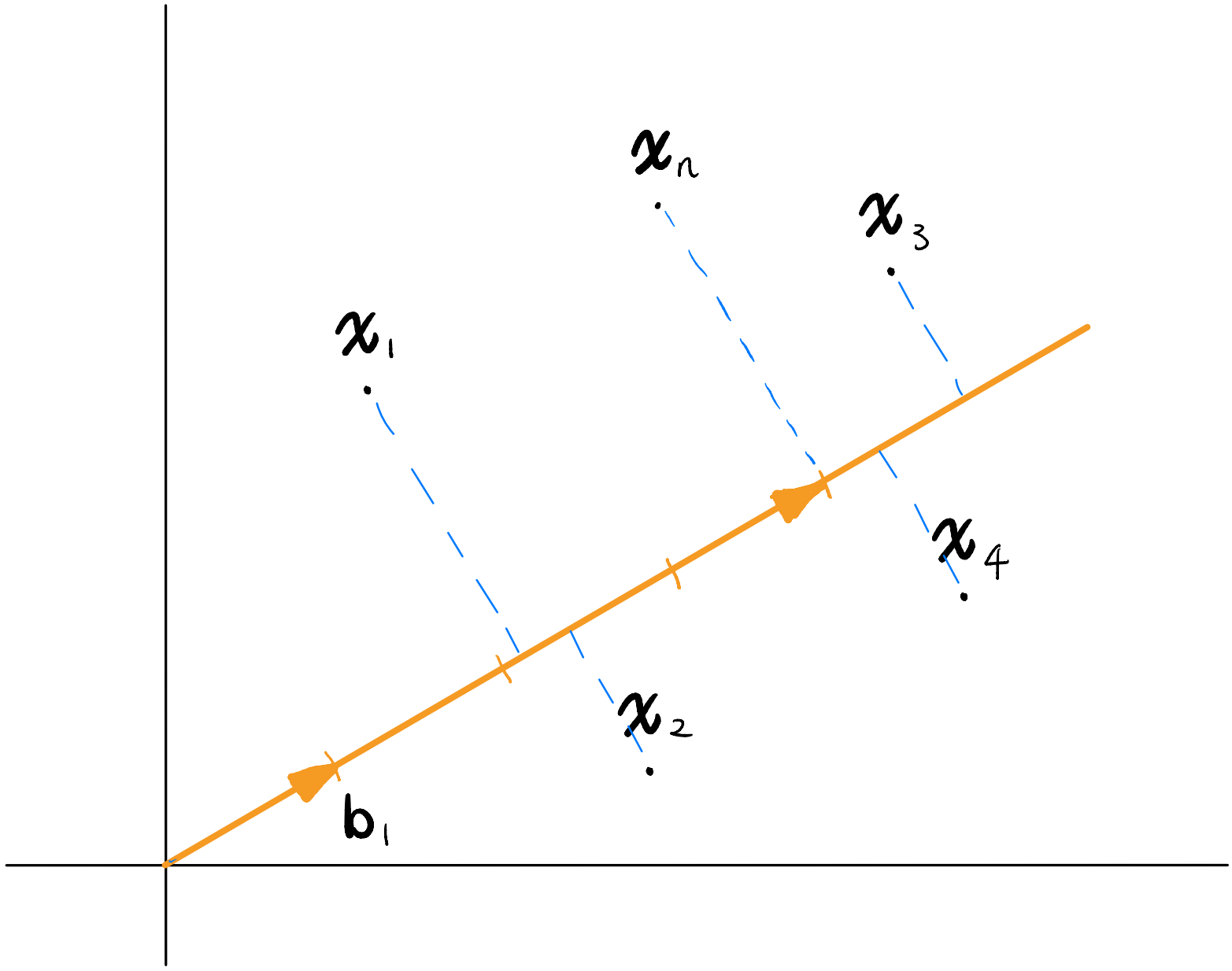
$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n = \mathbf{0}$$

- To achieve this, set

$$\boldsymbol{x}_n \leftarrow \boldsymbol{x}_n - \hat{\boldsymbol{\mu}}$$

- Find projections $\tilde{\boldsymbol{x}}_n$ of data points $\boldsymbol{x}_n$
  - Similar to the original data points as possible
  - But have a significantly lower intrinsic dimensionality.

Projection Perspective: One-dimensional Projection


x_n
x_3
x_1
x_4
x_2
b_1

Minimizing reconstruction error

- We'll start by looking for a one-dimensional projection.
- Project N points $\boldsymbol{x}_n \in \mathbb{R}^p$, $n = 1, \cdots, N$ on to a line through the origin.
- We specify the line by a **unit column vector**

  $\boldsymbol{b}_1 = (b_{11}, \cdots, b_{1p})^\top \in \mathbb{R}^p$, with $||\boldsymbol{b}_1|| = 1$, which forms a subspace U

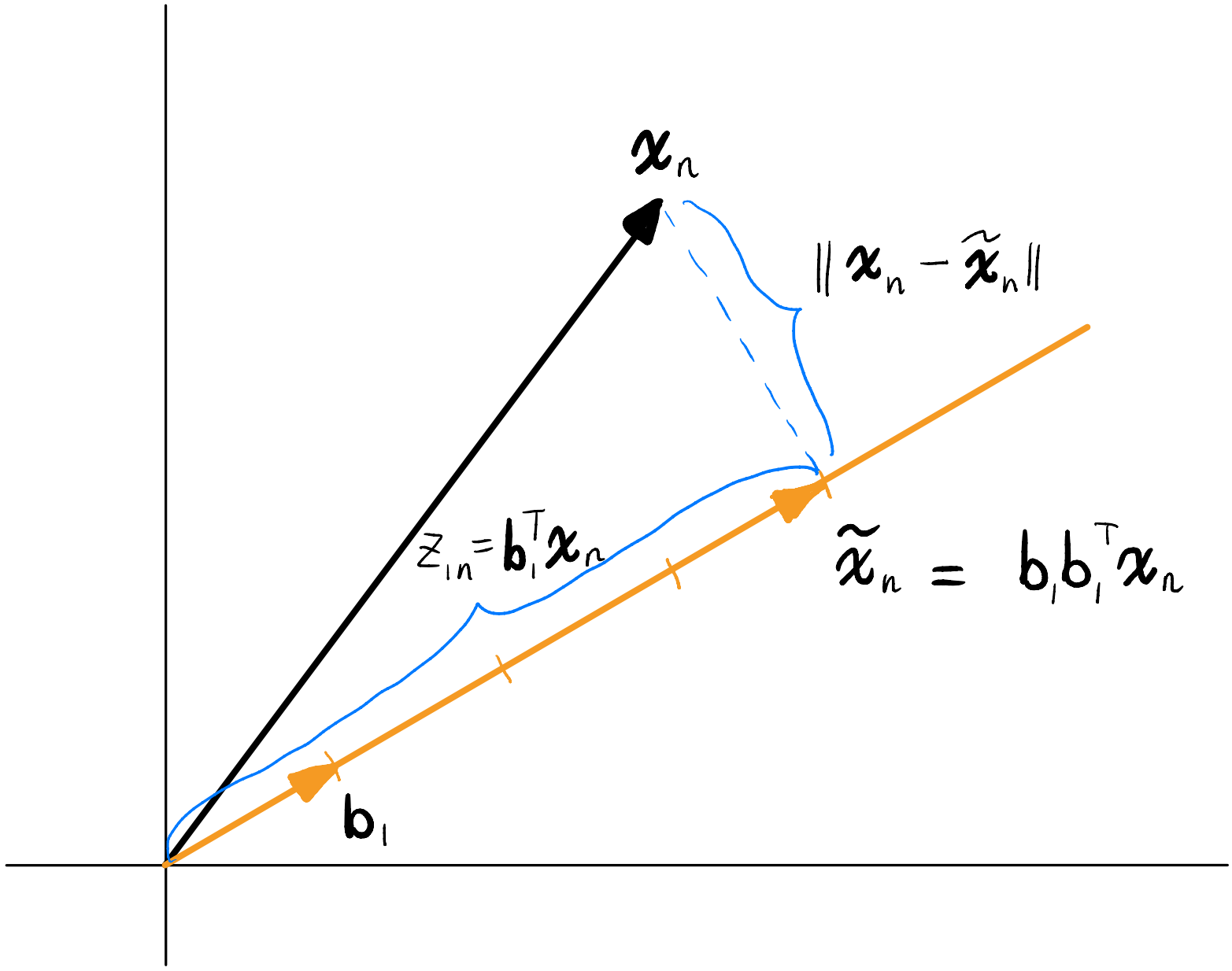
$$U = \{z_1 \boldsymbol{b}_1 : \forall z_1 \in \mathbb{R}\}$$

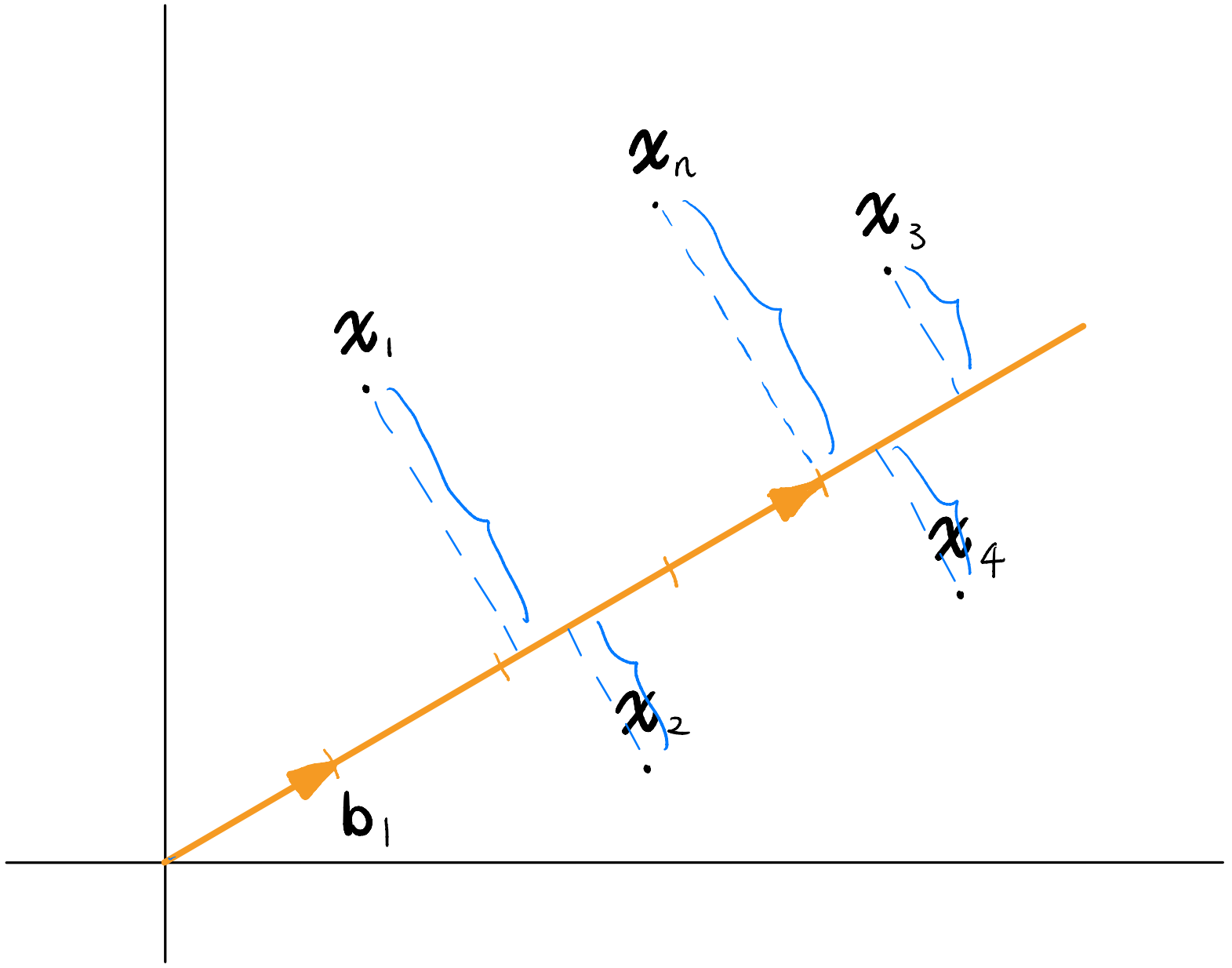
- The **coordinate** of the orthogonal projection of $\boldsymbol{x}_n$ onto the line U (one-dimensional subspace spanned by $\boldsymbol{b}_1$) is

$$z_{1n} = \boldsymbol{b}_1^\top \boldsymbol{x}_n$$

- The **orthogonal projection** $\tilde{\boldsymbol{x}}_n$ is

$$\tilde{\boldsymbol{x}}_n \equiv \pi_U(\boldsymbol{x}_n) = \underbrace{\boldsymbol{b}_1}_{\text{direction of proj.}} \cdot \underbrace{\boldsymbol{b}_1^\top \boldsymbol{x}_n}_{z_{1n}\text{:coordinate of proj.}}$$


x_n
$\|x_n - \tilde{x}_n\|$
$z_{1n} = b_1^T x_n$
$\tilde{x}_n = b_1 b_1^T x_n$
b_1


x_n
x_3
x_1
x_4
x_2
b_1

Minimizing reconstruction error

- If we approximate $\boldsymbol{x}_n$ using $\tilde{\boldsymbol{x}}_n$, the **squared approximation error** is

$$\|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \|\boldsymbol{x}_n - \pi_U(\boldsymbol{x}_n)\|^2 = \|\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2$$

- If average those $\|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2$ over all the points $n = 1, \cdots, N$, we get the average squared approximation error, i.e. **reconstruction error**

$$J_1 \equiv \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2 \quad \text{(reconstruction err.)}$$

- How to choose a good subspace U in PCA? – find $\boldsymbol{b}_1$ such that the reconstruction error is minimized

$$\min_{\boldsymbol{b}_1} J_1 \iff \min_{\boldsymbol{b}_1} \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2$$

Minimizing reconstruction error

- Here is an amazing result that bridges together two perspectives of PCA

$$\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n\|^2 - \frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{b}_1^\top\boldsymbol{x}_n\|^2$$

- which essentially claiming that

reconstruction err. = const. − projected variance

$$\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n\|^2 - \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$$

Proof

- The squared approximation error for point $\boldsymbol{x}_n$ can be rewritten as

$$
\begin{aligned}
\|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 &= \|\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n\|^2 \\
&= (\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n)^\top(\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n) \\
&= \boldsymbol{x}_n^\top\boldsymbol{x}_n - \boldsymbol{x}_n^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n - \boldsymbol{x}_n^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n + \boldsymbol{x}_n^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n \\
&\overset{(i)}{=} \boldsymbol{x}_n^\top\boldsymbol{x}_n - \boldsymbol{x}_n^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n \\
&= \|\boldsymbol{x}_n\|^2 - \|\boldsymbol{b}_1^\top\boldsymbol{x}_n\|^2 \\
&= \text{const} - z_{1n}^2
\end{aligned}
$$

where (i) follows by $\boldsymbol{x}_n^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n \overset{\boldsymbol{b}_1^\top\boldsymbol{b}_1=1}{=} \boldsymbol{x}_n^\top\boldsymbol{b}_1\boldsymbol{b}_1^\top\boldsymbol{x}_n$ and (ii) follows by $z_{1n} = \boldsymbol{b}_1^\top\boldsymbol{x}_n$.

Sample variance of projected coordinates

- (Homework) Show that the term $\frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$ indeed is equal to the **sample variance of the coodinates** $\{z_{11}, \cdots, z_{1N}\}$ of N projections.

$$\frac{1}{N}\sum_{n=1}^{N} z_{1n}^2 = \frac{1}{N}\sum_{n=1}^{N}\left(z_{1n} - \frac{1}{N}\sum_{n=1}^{N} z_{1n}\right)^2 \equiv \widehat{\mathrm{Var}}(z_{11}, \cdots, z_{1N})$$

- **Hint:** recall that the data is centered; also note that we use factor $\frac{1}{N}$ instead of $\frac{1}{N-1}$ in the sample variance.
- Note that $\widehat{\mathrm{Var}}$ stands for the sample variance, rather than the population variance Var.

- (Homework) Show that we also have

$$\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n\|^2 - \frac{1}{N}\sum_{n=1}^{N}\|\tilde{\boldsymbol{x}}_n\|^2$$

  i.e. norm of reconstruction is equal to norm of coordinate. (If you draw it, this is obvious).

Yin and Yang 阴阳 – data's duality

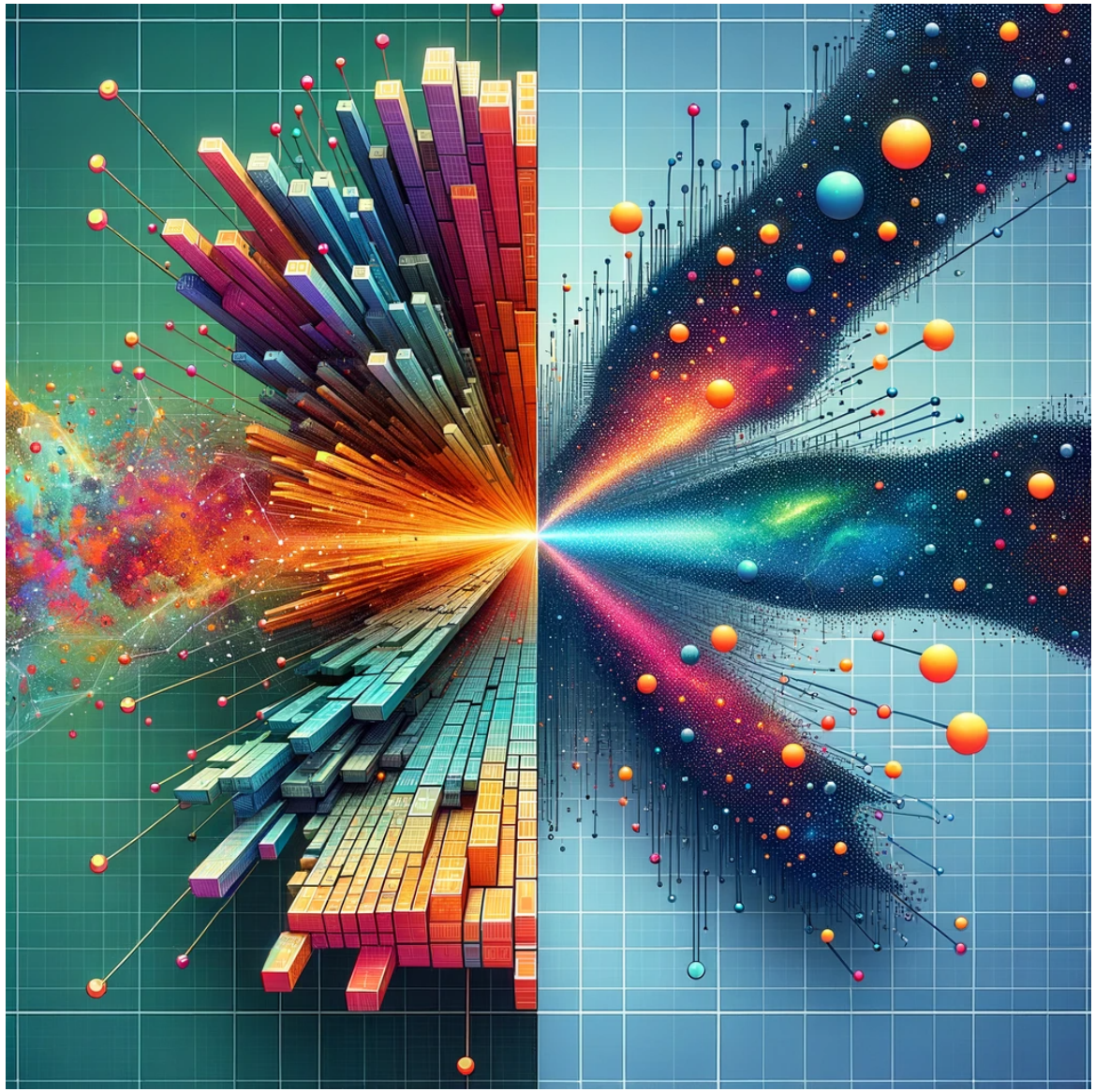
- Minimizing the average squared reconstruction error is equivalent to maximizing the variance of the projection.

$$\min_{\boldsymbol{b}_1} J_1 \iff \max_{\boldsymbol{b}_1} \widehat{\mathrm{Var}}(z_{11}, \cdots, z_{1N})$$

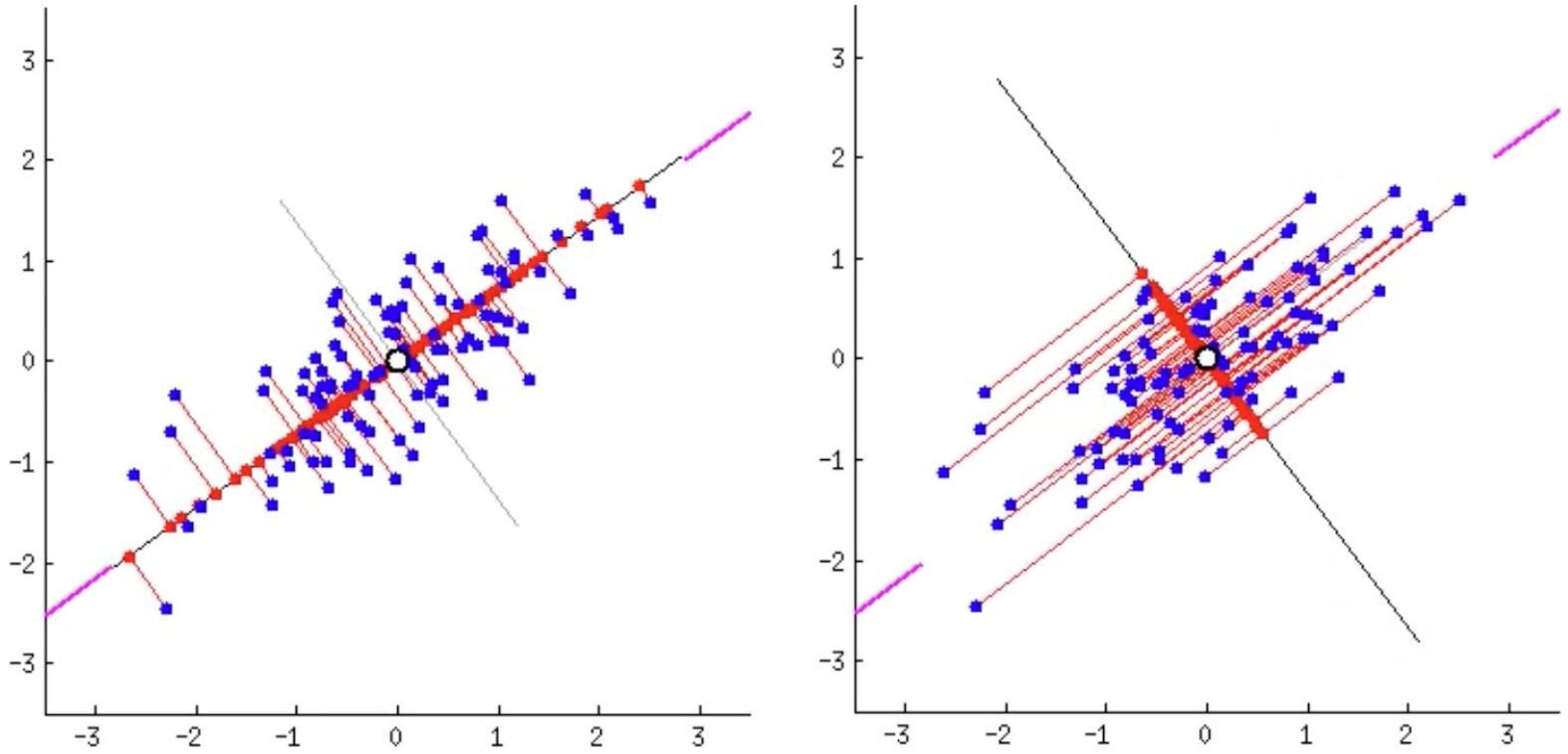
i.e.

$$\min_{\boldsymbol{b}_1} \frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{b}_1 \boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2 \iff \max_{\boldsymbol{b}_1} \frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2$$

- Choosing a subspace to maximize the projected variance, or minimize the reconstruction error, is called principal component analysis (PCA).



ChatGPT: (experience PCA in a dual light) – one side radiates with the expansion of variance, while the other elegantly narrows down reconstruction errors. A visual symphony of data's duality.



Archer: (poor man's version?) – these figures show that maximizing the variance of the projection is equivalent to minimizing the average squared reconstruction error.

Maximum Variance Perspective

Direction with maximal variance

- Accordingly, let's find the projection direction $\boldsymbol{b}_1$ that **maximizes the variance**!

$$V_1 \equiv \widehat{\mathrm{Var}}(z_{11}, \cdots, z_{1N}) = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2 = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2$$

- The corresponding optimization problem is

$$\max_{\boldsymbol{b}_1} V_1 \equiv \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2 \tag{1}$$

Direction with maximal variance

- For simplicity, rewrite the problem in a matrix format

$$
\begin{aligned}
V_1 &= \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{b}_1^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{b}_1 = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{b}_1^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{b}_1 \\
&= \boldsymbol{b}_1^\top \left(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \right) \boldsymbol{b}_1 \\
&= \boldsymbol{b}_1^\top \left(\frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \right) \boldsymbol{b}_1 \\
&= \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1
\end{aligned}
$$

Direction with maximal variance

- Here S is the **sample (data) covariance matrix**

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top = \frac{1}{N}\boldsymbol{X}^\top \boldsymbol{X}$$

where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

Direction with maximal variance

- Written in the matrix format, the optimization problem (1) becomes

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$

- Note that arbitrarily increasing the magnitude of the vector $\boldsymbol{b}_1$ increases $\boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$. Thus we have to restrict $\boldsymbol{b}_1$ to be **a unit vector**, i.e. $||\boldsymbol{b}_1|| = 1$

$$\begin{aligned} &\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 \\ &\text{subject to } ||\boldsymbol{b}_1|| = 1 \end{aligned} \tag{2}$$

- To solve problem (2), we introduce the **Lagrangian**

$$\mathcal{L}(\boldsymbol{b}_1, \lambda_1) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

- The partial derivatives of $\mathcal{L}$ are

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top \qquad \frac{\partial \mathcal{L}}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1$$

- Setting these partial derivatives to $\mathbf{0}$ gives

$$\boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1 \qquad\qquad \boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1$$

- We see that
  - $\boldsymbol{b}_1$ is an **eigenvector** of the sample covariance matrix $\boldsymbol{S}$.
  - The Lagrange multiplier λ_1 is the corresponding **eigenvalue**, it is also the variance of the resulting projected coordinates

$$V_1 = \boldsymbol{b}_1^\top \boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1^\top \boldsymbol{b}_1 = \lambda_1$$

- Therefore, to maximize the variance V_1, we choose the basis vector $\boldsymbol{b}_1$ associated with the **largest eigenvalue** of the data covariance matrix $\boldsymbol{S}$.
- The eigenvector $\boldsymbol{b}_1$ is called the 1st **principal component**.

- We can get the coordinate of the projection

$$z_{1n} = \boldsymbol{b}_1^\top \boldsymbol{x}_n \in \mathbb{R}$$

- We can also get the approximation of $\boldsymbol{x}_n$ by mapping the coordinate z_{1n} back into data space

$$\tilde{\boldsymbol{x}}_n = \boldsymbol{b}_1 z_{1n} = \boldsymbol{b}_1 \boldsymbol{b}_1^\top \boldsymbol{x}_n \in \mathbb{R}^p$$

  which gives us the projected data point $\tilde{\boldsymbol{x}}_n$ in the original data space.

- **Remark:** Although $\tilde{\boldsymbol{x}}_n$ is a p-dimensional vector, it only requires a single coordinate z_{1n} to represent it with respect to the basis vector $\boldsymbol{b}_1 \in \mathbb{R}^p$.

- If we project the data to the first principal component $\boldsymbol{b}_1$, combining equation

$$J_1 \equiv \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n\|^2 - \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$$

and

$$\frac{1}{N}\sum_{n=1}^{N} z_{1n}^2 = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2 = \boldsymbol{b}_1^\top \boldsymbol{S}\boldsymbol{b}_1 = \lambda_1$$

we see that the corresponding reconstruction error is

$$J_1 = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n\|^2 - \lambda_1$$

Data Centering and Scaling

Data centering and scaling

- In PCA, the **centering** and **scaling** of data are important steps.
- But their necessity depends on the context and nature of your data.
- The original data matrix is

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

- **Centering** is almost always necessary in PCA.
- Involves substracting each entry by the corresponding column mean:

$$\begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{N1} - \bar{x}_1 & x_{N2} - \bar{x}_2 & \cdots & x_{Np} - \bar{x}_p \end{bmatrix}$$

- The column mean of each column **after centering** becomes zero.
- Ensures that each variable contributes equally to the analysis so that PCA focuses on the variance of the data.
- Without, PCA might be influenced by variables that are on a larger scale.

- **Scaling** - aka **standardization**
- Involves dividing each entry by the corresponding column standard deviation:

$$\begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\widehat{\mathrm{Var}}(x_{11},\cdots,x_{N1})}} & \frac{x_{12}-\bar{x}_2}{\sqrt{\widehat{\mathrm{Var}}(x_{12},\cdots,x_{N2})}} & \cdots & \frac{x_{1p}-\bar{x}_p}{\sqrt{\widehat{\mathrm{Var}}(x_{1p},\cdots,x_{Np})}} \\ \frac{x_{21}-\bar{x}_1}{\sqrt{\widehat{\mathrm{Var}}(x_{11},\cdots,x_{N1})}} & \frac{x_{22}-\bar{x}_2}{\sqrt{\widehat{\mathrm{Var}}(x_{12},\cdots,x_{N2})}} & \cdots & \frac{x_{2p}-\bar{x}_p}{\sqrt{\widehat{\mathrm{Var}}(x_{1p},\cdots,x_{Np})}} \\ \vdots & \vdots & & \vdots \\ \frac{x_{N1}-\bar{x}_1}{\sqrt{\widehat{\mathrm{Var}}(x_{11},\cdots,x_{N1})}} & \frac{x_{N2}-\bar{x}_2}{\sqrt{\widehat{\mathrm{Var}}(x_{12},\cdots,x_{N2})}} & \cdots & \frac{x_{Np}-\bar{x}_p}{\sqrt{\widehat{\mathrm{Var}}(x_{1p},\cdots,x_{Np})}} \end{bmatrix}$$

- The column variance of each column **after centering** becomes one.
- Crucial when the variables in data are **on different scales** (e.g., kg, km, F, C) or have different units of measurement.
- Scaling ensures that PCA gives equal weight to each variable, preventing variables with larger scales from dominating the PCA.

- Always center the data;
- Scale the data when variables are on different scales or units;
  - In imaging — where all coordinates are in the same units, namely pixel intensities — there is no need to do such coordinate scaling.
- Be cautious with scaling for interpretability.

Why use Orthogonal Projection in Reconstruction?

Why orthogonal projection is optimal?

- Recall that **orthogonal** projection is adopted to get reconstruction $\tilde{\boldsymbol{x}}_n$ in

$$J_1 \equiv \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{b}_1\boldsymbol{b}_1^\top \boldsymbol{x}_n\|^2$$

- But **why** use orthogonal projection?
- Consider an unknown linear projection method, the resulting coordinate of the projection for data point $\boldsymbol{x}_n$ is z_{1n}
- The corresponding reconstruction error is

$$J_1 = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{b}_1 z_{1n}\|^2$$

- Assume $\boldsymbol{b}_1$ is given, find z_{1n} that minimize the reconstruction error

$$\min_{z_{11},\cdots,z_{1N}} \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{b}_1 z_{1n}\|^2$$

- To solve this problem, compute the partial derivative and set it to zero

$$\frac{\partial J_1}{\partial z_{1n}} = -\frac{2}{N}(\boldsymbol{x}_n - \boldsymbol{b}_1 z_{1n})^\top \boldsymbol{b}_1 = -\frac{2}{N}(\boldsymbol{b}_1^\top \boldsymbol{x}_n - \boldsymbol{b}_1^\top \boldsymbol{b}_1 z_{1n}) = 0$$

- Since $\boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1$, the equation yields

$$z_{1n} = \boldsymbol{b}_1^\top \boldsymbol{x}_n$$

- Consequently,
  - The optimal linear projection $\tilde{\boldsymbol{x}}_n$ of $\boldsymbol{x}_n$ is an orthogonal projection.
  - **Optimal** coordinates z_{1n} of the projection $\tilde{\boldsymbol{x}}_n$ are the coordinates of the **orthogonal projection** of the original data point $\boldsymbol{x}_n$.