

# Unsupervised Dimensionality Reduction via PCA

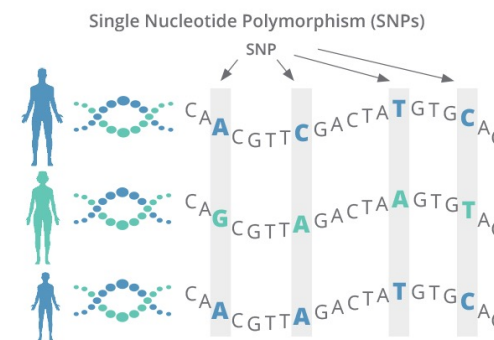
ECE57000: Artificial Intelligence

David I. Inouye

Thursday, September 9, 2021

Very high-dimensional data is becoming ubiquitous

- ▶ Images (1 million pixels)
- ▶ Text (100k unique words)
- ▶ Genetics (4 million SNPs)
- ▶ Business data (12 million products)

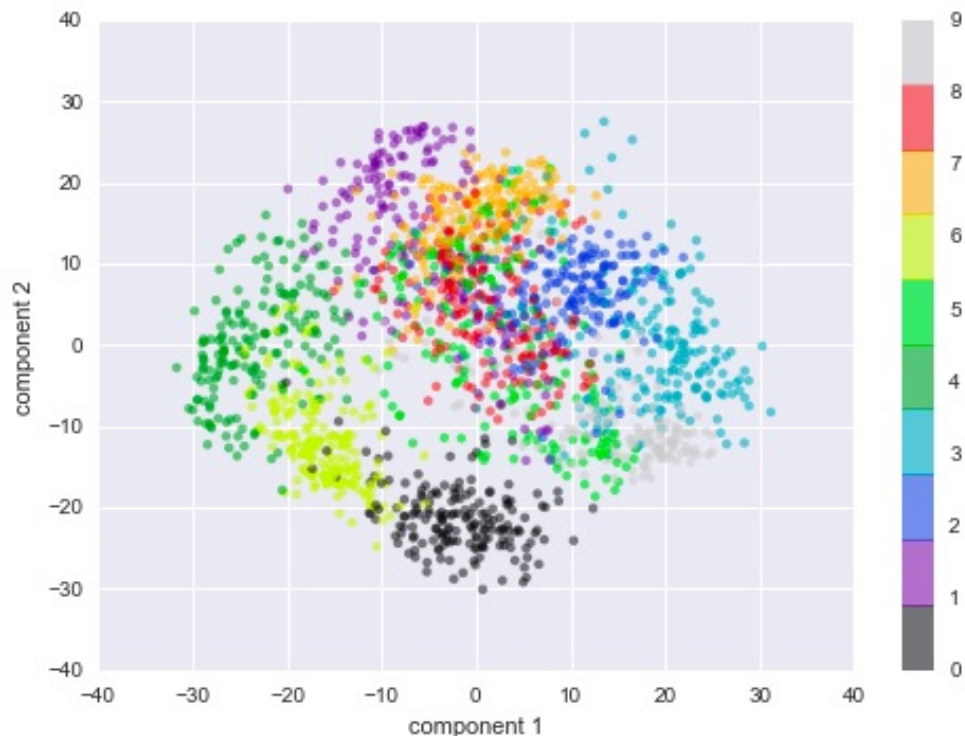




# Why dimensionality reduction?

## Visualization

- ▶ Allows 2D scatterplot visualizations even of high-dimensional data (2D projection of digits)



# Why dimensionality reduction?

Noise reduction via reconstruction



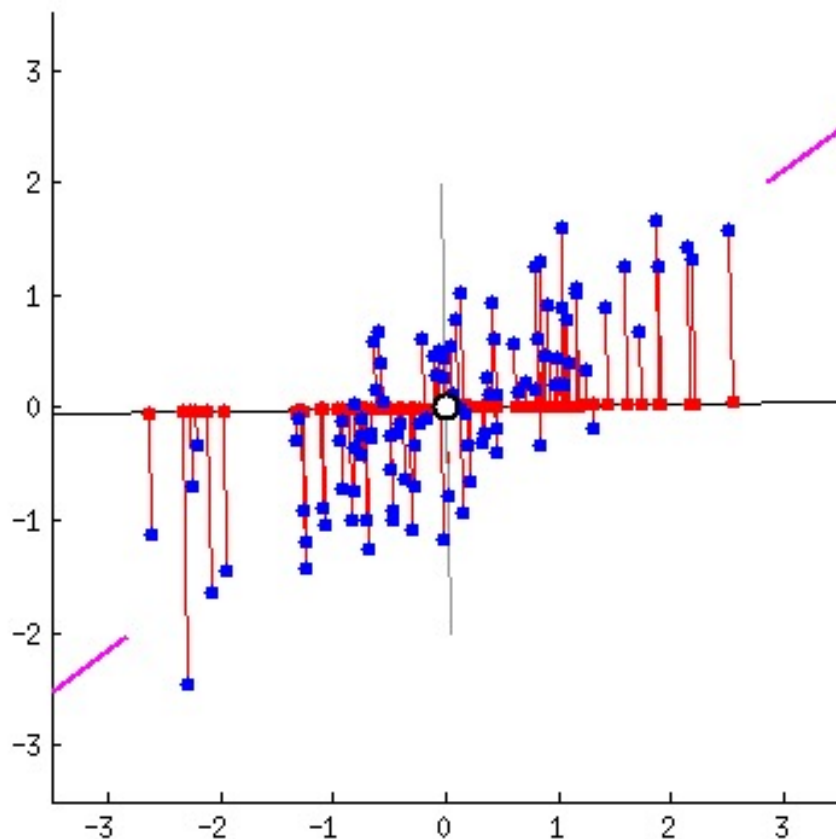
# Outline of Principal Components Analysis (PCA)

1. Motivation for dimensionality reduction
2. Intuition of PCA
3. Formal PCA problem: Min reconstruction
4. Derive PCA formulation for 1D
  - ▶ Least error 1D projection is orthogonal
  - ▶ Sum over all data points
5. Solution is based on truncated SVD
6. Alternative viewpoint: Max variance
  - ▶ Derive equivalence
  - ▶ Derive equivalent solutions

# Review of linear algebra and introduction to numpy Python library

- ▶ See Jupyter notebook, which can be opened and run in Google Colab

Intuition: Principal component analysis finds the best linear projection onto a lower-dimensional space



2D to 1D projection: Red lines show the projection error onto 1D lines. PCA finds the line that has the smallest projection error (in this example, when it aligns with the purple).

<https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues>



Math: Principal Component Analysis (PCA) can be formalized as minimizing the linear reconstruction error of the data using only  $k \leq d$  dimensions

- ▶ PCA can be formalized as

$$\min_{Z, W} \|X_c - ZW^T\|_F^2$$

- ▶ where

$X_c = X - \mathbf{1}_n \mu_x^T \in \mathbb{R}^{n \times d}$  (centered input data)

$Z \in \mathbb{R}^{n \times k}$  (latent representation or “scores” )

$W^T \in \mathbb{R}^{k \times d}$  (principal components)

$w_s^T w_t = 0, w_s^T w_s = \|w_s\|_2 = 1, \forall s, t$   
(orthogonal constraint)

Math: Principal Component Analysis (PCA) can be formalized as minimizing the linear reconstruction error of the data using only  $k \leq d$  dimensions

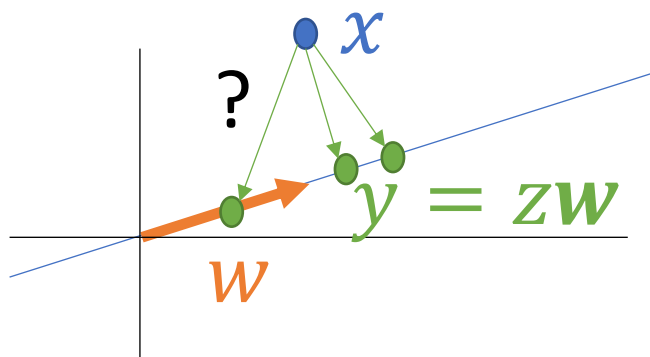
$$\min_{Z \in \mathbb{R}^{n \times k}, W \in \mathbb{R}^{d \times k}} \|X_C - ZW^T\|_F^2 \quad \text{s.t. } W^T W = I_k$$

- ▶ Let's stare at this equation some more 😊
- ▶ Why is this dimensionality reduction?
- ▶ What does the orthogonal constraint mean?
- ▶ Why minimize the squared Frobenius norm?
- ▶  $\|X_C - ZW^T\|_F^2 = \sum_{i=1}^n \|\mathbf{x}_i^T - \mathbf{z}_i^T W^T\|_2^2 = \sum_{i=1}^n \|\mathbf{x}_i - W \mathbf{z}_i\|_2^2$
- ▶ For analysis, let's simplify to a single dimension (i.e.,  $k = 1$ )
  - ▶  $\sum_{i=1}^n \|\mathbf{x}_i - z_i \mathbf{w}\|_2^2$  where  $z_i$  is a scalar

What is the best projection given a fixed subspace (line in 1D case)?

- ▶ If we are given  $\mathbf{w}$ , what is the best  $z$  (i.e. minimum reconstruction error) for a given  $\mathbf{x}$ ?

- ▶  $\min_z \|\mathbf{x} - z\mathbf{w}\|_2^2$



- ▶ The orthogonal projection!

- ▶  $z = \mathbf{x}^T \mathbf{w} = \|\mathbf{x}\| \|\mathbf{w}\| \cos \theta = \|\mathbf{x}\| \cos \theta$

- ▶  $z = \|\mathbf{x}\| \cos \theta = \text{hyp} \cdot \frac{\text{adj}}{\text{hyp}} = \text{adj}$

- ▶  $z\mathbf{w}$  is a scaled vector along the line defined by  $\mathbf{w}$

Thus, we can simplify to only minimizing over  $W$

$$\min_{\mathbf{z}, \mathbf{w}: \|\mathbf{w}\|_2=1} \sum_{i=1}^n \|\mathbf{x}_i - z_i \mathbf{w}\|_2^2 = \min_{\mathbf{w}: \|\mathbf{w}\|_2=1} \sum_{i=1}^n \|\mathbf{x}_i - (\mathbf{x}_i^T \mathbf{w}) \mathbf{w}\|_2^2$$

- ▶ Now we can return to the Frobenius norm:

$$\min_{\mathbf{w}: \|\mathbf{w}\|_2=1} \|X_C - \mathbf{z} \mathbf{w}^T\|_F^2 \quad \text{where } \mathbf{z} = X_C \mathbf{w}$$

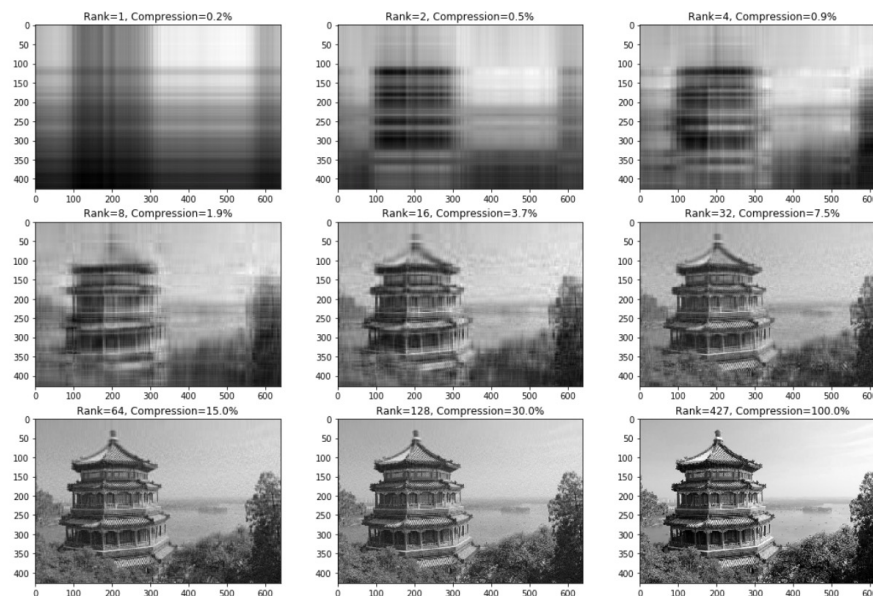
- ▶ What is  $\mathbf{z} \mathbf{w}^T$ ? Have we seen something like this before?
- ▶ This is the best low-rank approximation to  $X_C$ , which is given by the SVD!
  - ▶  $\mathbf{w} = \mathbf{v}_1$  and  $\mathbf{z} = \sigma_1 \mathbf{u}_1$ , where  $\sigma_1, \mathbf{u}_1, \mathbf{v}_1$  are the first singular value, left singular vector and right singular vector respectively.

For  $k \geq 1$ , the PCA solution is the top  $k$  right singular vectors

- ▶ If  $X_c = USV^T$ , then the general solution is

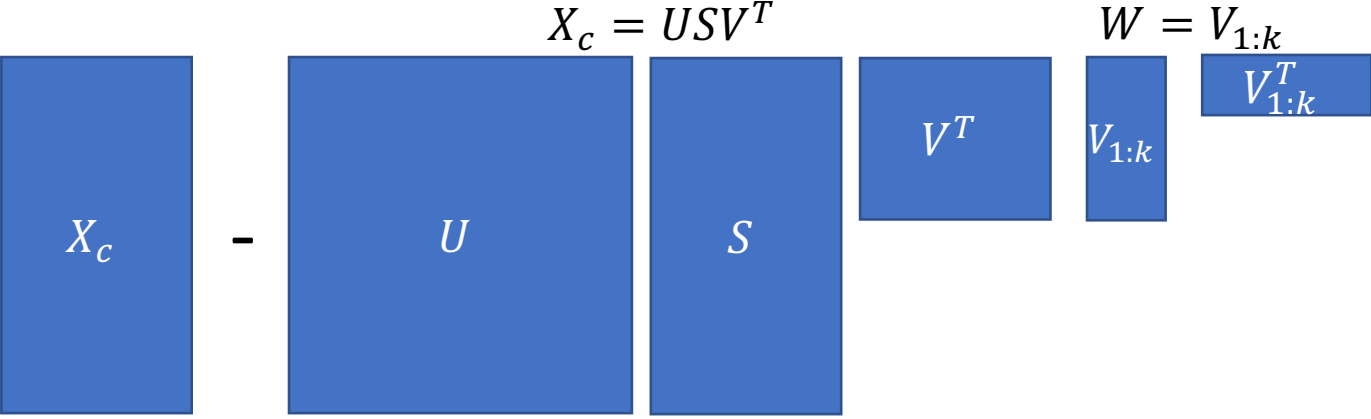
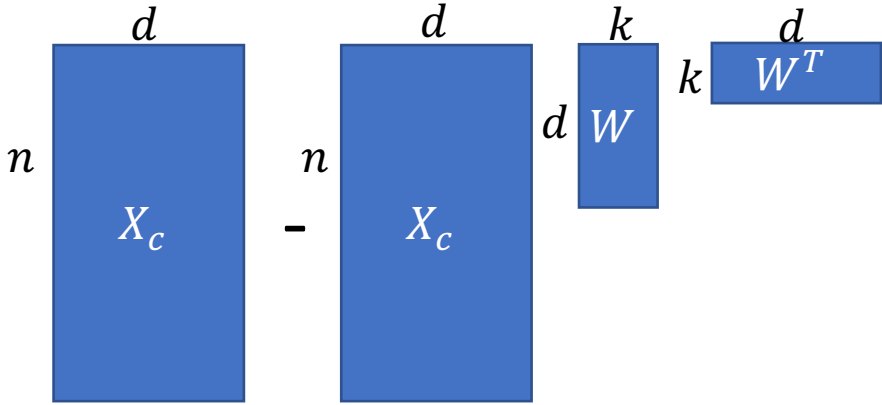
$$W^* = V_{1:k}$$

- ▶ Remember: SVD is best  $k$  dim. approximation



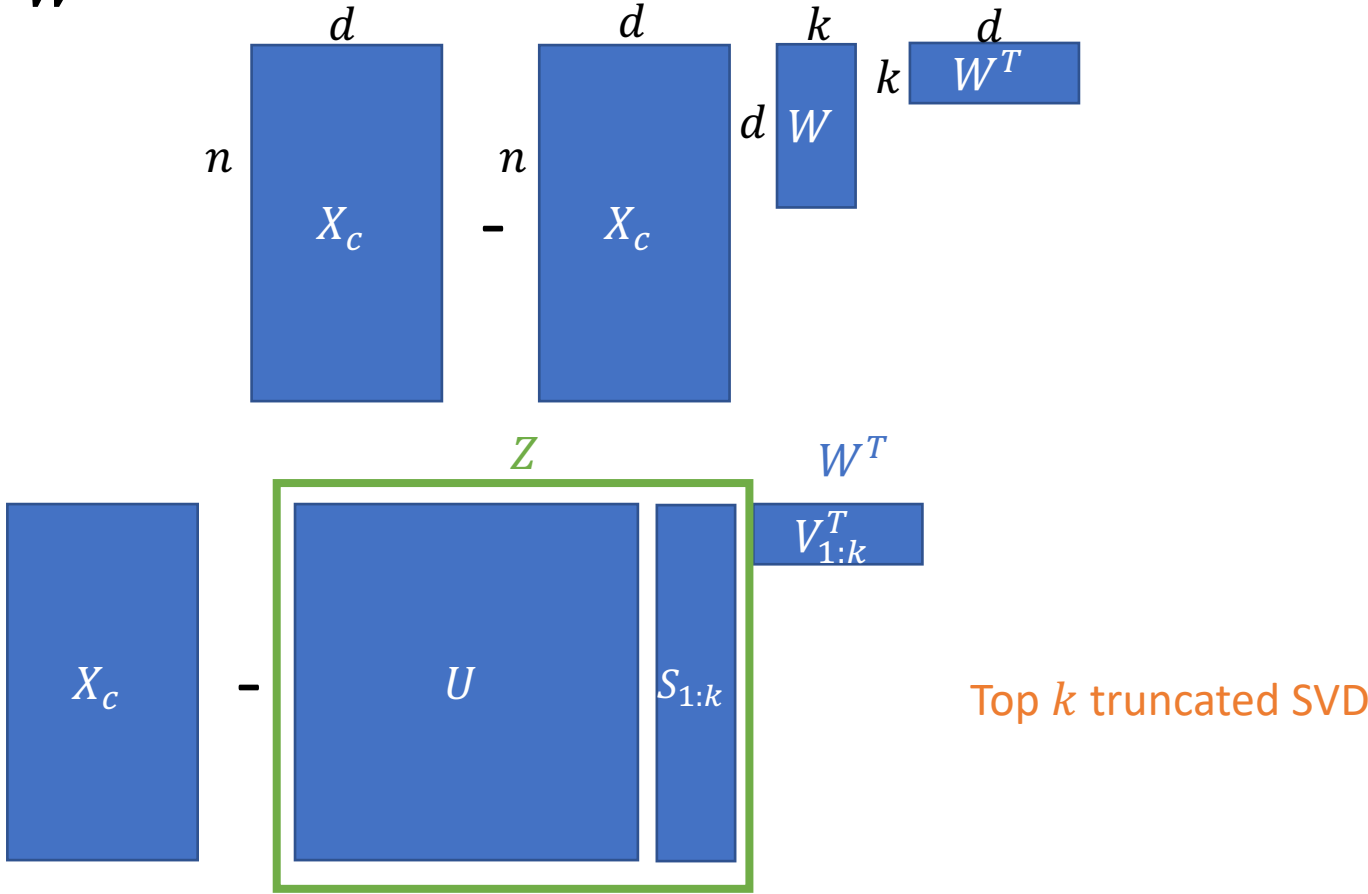
Check: The solution reveals the truncated SVD as best approximation

$$\min_W \|X_c - (X_c W)W^T\|_F^2 \quad \text{s. t.} \quad W^T W = I$$



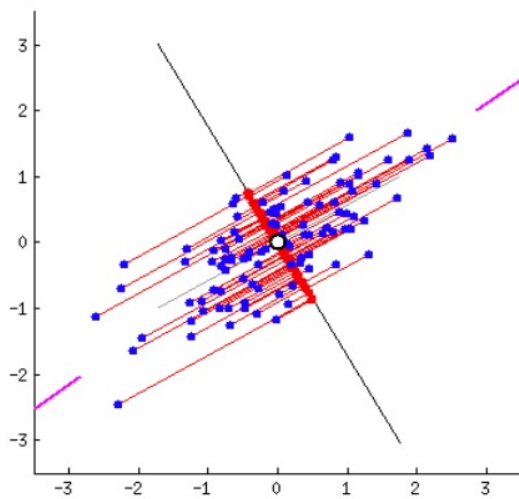
Check: The solution reveals the truncated SVD as best approximation

$$\min_W \|X_c - (X_c W)W^T\|_F^2 \quad \text{s.t.} \quad W^T W = I$$

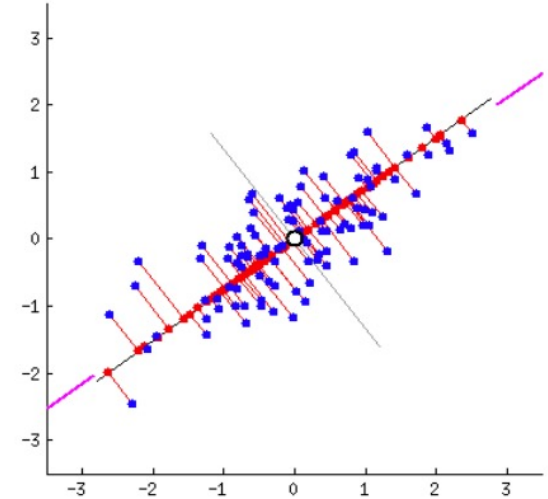
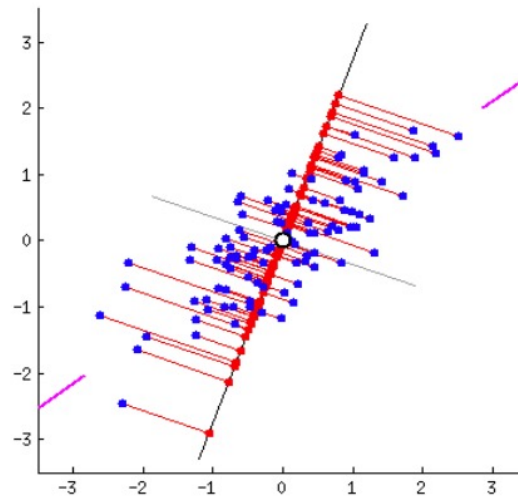


Claim: **Minimizing reconstruction error** (red lines) is equivalent to **maximizing the variance of projection** (spread of red points)

Max reconstruction error  
Min variance



Min reconstruction error  
Max variance





# Derivation of min error equivalent to max variance

► First step: Simplify squared distance

$$\|x_i - (x_i^T w)w\|_2^2$$

$$= (x_i - (x_i^T w)w)^T (x_i - (x_i^T w)w)$$

$$= x_i^T x_i - 2(x_i^T w)w^T x_i + (x_i^T w)^2 w^T w$$

$$= \|x_i\|^2 - 2(x_i^T w)^2 + (x_i^T w)^2 \|w\|^2$$

$$= \|x_i\|^2 - (x_i^T w)^2$$

# Derivation of min error equivalent to max variance

- ▶ Equivalence of optimization in 1D

- ▶  $\arg \min_w \sum_i \|x_i - (x_i^T w)w\|_2^2$

- ▶  $= \arg \min_w \sum_i \|x_i\|^2 - (x_i^T w)^2 = \arg \min_w \sum_i -(x_i^T w)^2$

- ▶  $= \arg \max_w \frac{1}{n} \sum_i (x_i^T w)^2 = \arg \max_w \frac{1}{n} \sum_i z_i^2$

- ▶  $= \arg \max_w \sigma_z^2$

Note  $z$  is already centered so  
mean of squares is variance

Therefore, we can reformulate the problem as maximizing the variance

▶ Let's rewrite this last term

$$\begin{aligned}\text{▶ } \sigma_z^2 &= \frac{1}{n} \sum_i (z_i)^2 = \frac{1}{n} \sum_i (x_i^T w)^2 = \frac{1}{n} (X_c w)^T (X_c w) \\ &= w^T \left( \frac{1}{n} X_c^T X_c \right) w = w^T \hat{\Sigma} w\end{aligned}$$

▶ Thus, our problem can be formulated as:

$$\text{▶ } \max_{w: \|w\|=1} w^T \hat{\Sigma} w$$

▶ The solution is the eigenvector  $q_1$  of  $\hat{\Sigma} = Q\Lambda Q^T$  corresponding to the largest eigenvalue  $\lambda_1$

$$w^* = q_1$$

For  $k > 1$ , we maximize the sum of variances for each latent dimension

► More generally we can formulate this as:

$$\begin{aligned} & \max_{W:W^TW=I_k} \sum_{j=1}^k \sigma_{z_j}^2 \\ & = \max_{W:W^TW=I_k} \sum_{j=1}^k w_j^T \hat{\Sigma} w_j \\ & = \max_{W:W^TW=I_k} \text{Tr}(W^T \hat{\Sigma} W) \end{aligned}$$

► The solution is the top  $k$  eigenvectors of  $\hat{\Sigma}$   
 $= Q\Lambda Q^T$   
►  $W^* = Q_{1:k}$

Equivalent solutions: The solution to both problems is the top  $k$  right singular vectors of  $X_c$

- ▶ Minimize reconstruction error

$$\min_{W:W^T W=I_k} \|X_c - (X_c W)W^T\|_F^2$$

- ▶ Singular value decomposition (SVD) of  $X_c = USV^T$
- ▶ Solution:  $W^* = V_{1:k}$

- ▶ Maximize variance of latent projection

$$\max_{W:W^T W=I_k} \text{Tr}(W^T \hat{\Sigma} W)$$

- ▶ Equivalent solution
- ▶  $n\hat{\Sigma} = X_c^T X_c = (USV^T)^T (USV^T) = (VSU^T)(USV^T) = VS(U^T U)SV^T = VS^2V^T = Q\Lambda Q^T$
- ▶ Solution:  $W^* = Q_{1:k} \equiv V_{1:k}$ !

# Recap: Principal Components Analysis (PCA)

1. Motivation for dimensionality reduction
2. Intuition of PCA
3. Formal PCA problem: Min reconstruction
4. Derive PCA formulation for 1D
  - ▶ Least error 1D projection is orthogonal
  - ▶ Sum over all data points
5. Solution is based on truncated SVD
6. Alternative viewpoint: Max variance
  - ▶ Derive equivalence
  - ▶ Derive equivalent solutions