

Unsupervised Dimensionality Reduction via PCA (Part 2)

David I. Inouye

Before

6	6	3	1	5	0
2	-	3	1	0	9
1		0	-	1	2
2		8	-	1	0
0		5	-	0	1
0	1	1	-	0	1

Eigenvalues and
same diagonal

3	0	7	0
0	6	0	6
1	2	7	7
5	4	0	0
0	0	2	5
1	1	1	0
0	0	2	1
0	0	1	4

Eigenvalues
Eigenvectors

Solution

PCA Intuition and Math

Before

6	6	3	1	5	0	
2	-	3	1	0	9	
1		0	-	1	2	7
2		8	-	1	0	0
0		5	-	0	1	0
0	1	1	-	0	1	

Eigenvalues
Eigenvectors

Transformation

After

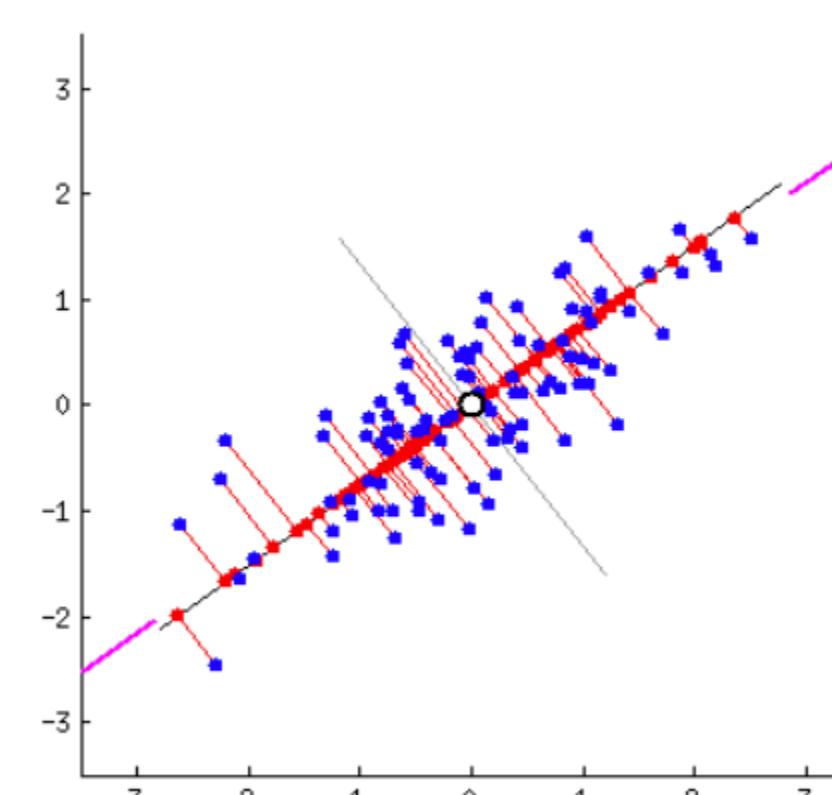
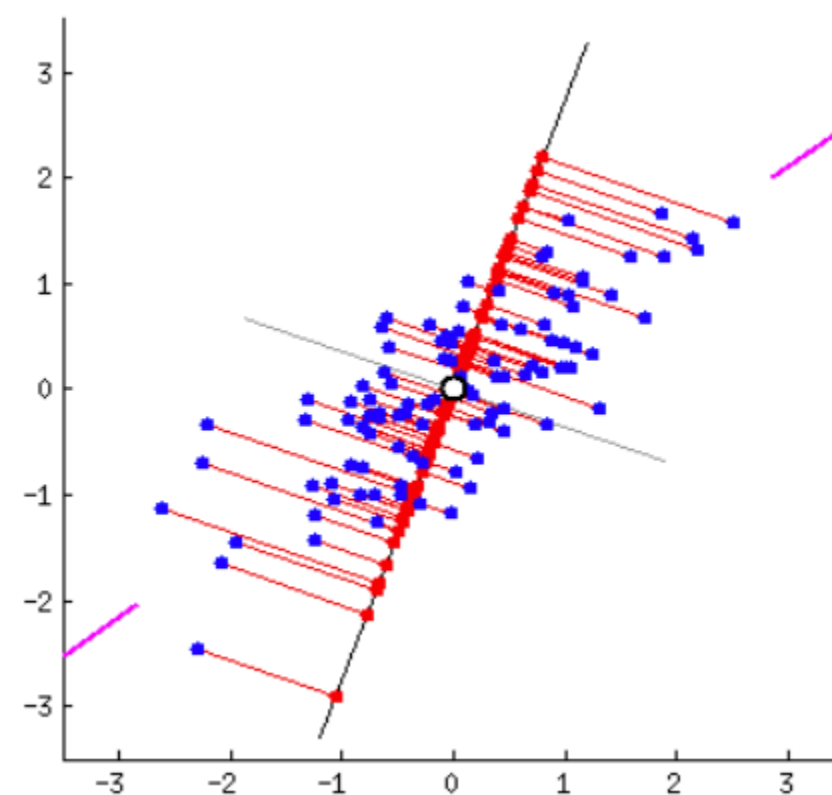
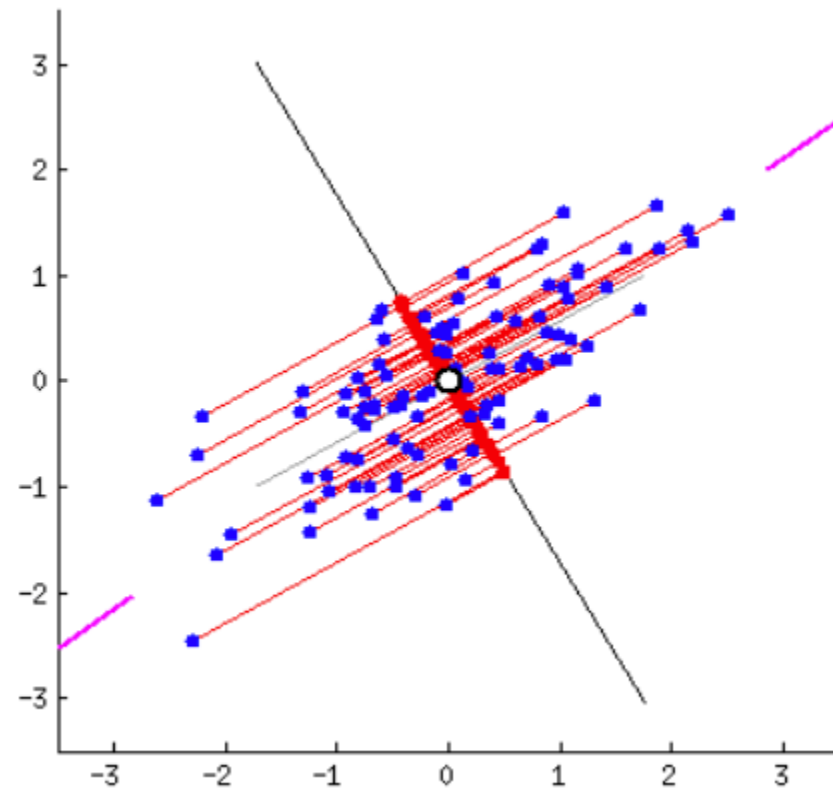
Eigenvalues and
same diquation

3	0	7	0		
0	6	0	6	5	1
1	2	7		7	4
5	4	0	0		8
0		0	2	5	
1		1	1	0	
0		0	2	1	
0		0	1	4	

Solution

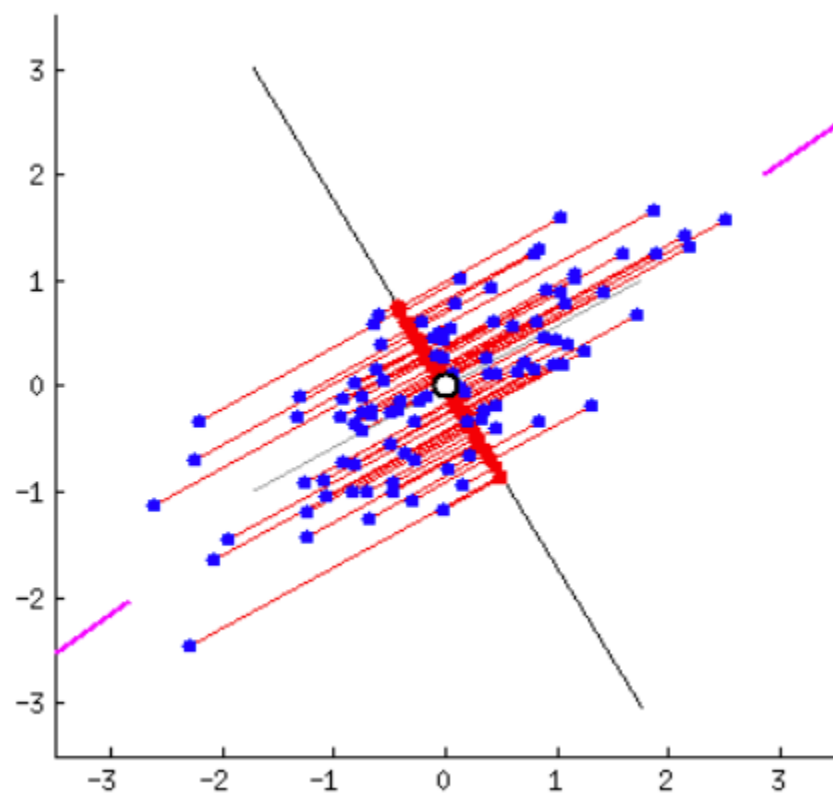
The Goal: Finding the Best Lower-Dimensional Representation

Before diving into equations, let's start with the core visual intuition. What does it mean to find the "best" lower-dimensional projection of our data?

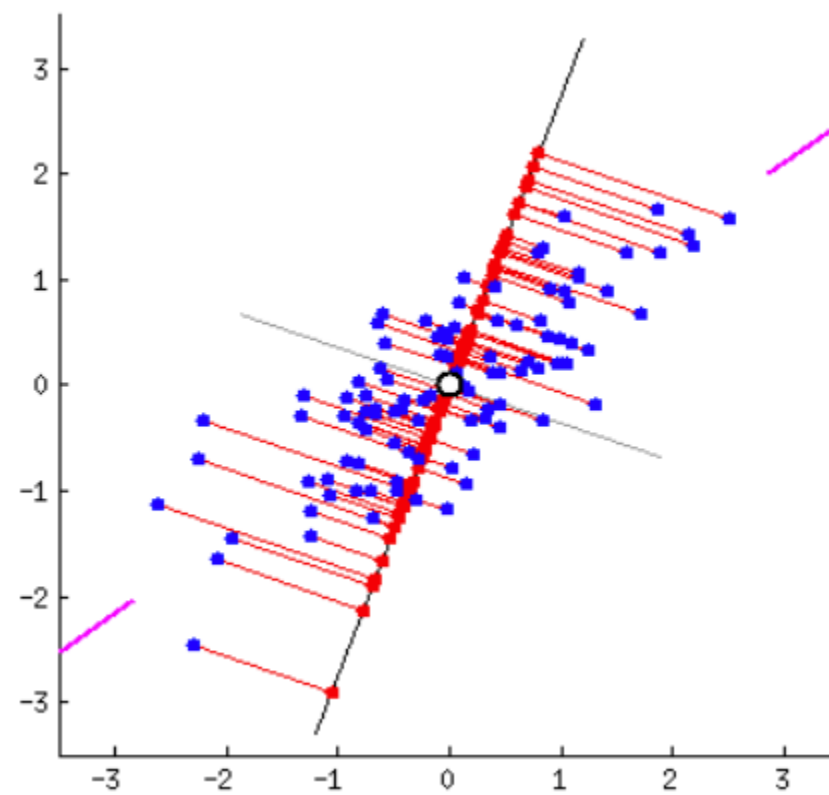


Which is the best projection of the 2D data onto a 1D space (i.e., a line)?

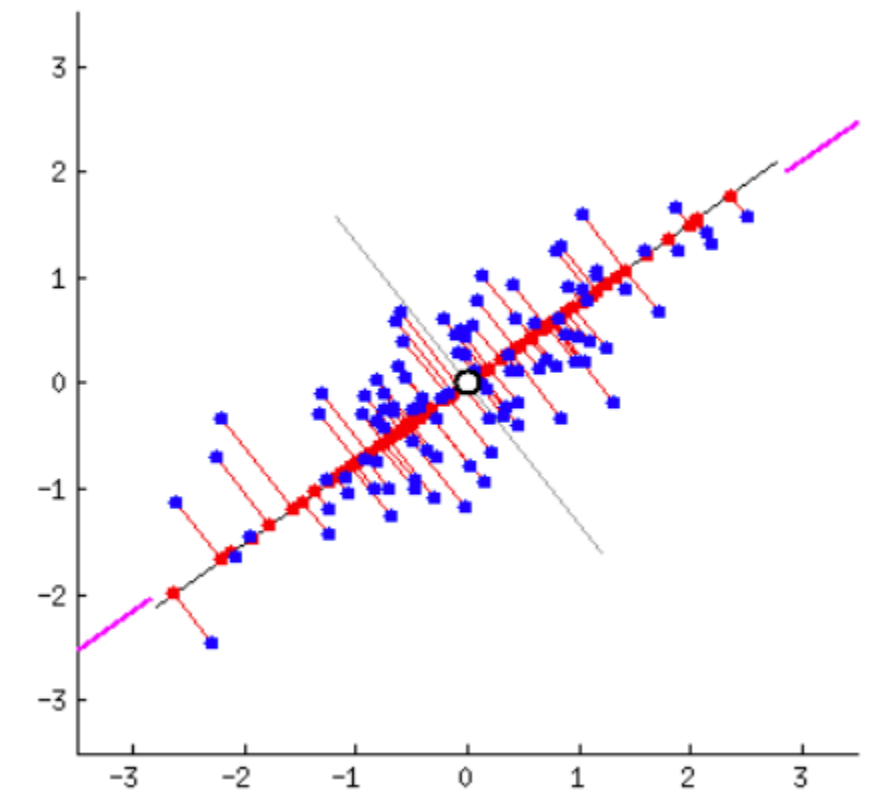
Intuition: Principal Component Analysis Finds The Best Linear Projection Onto a Lower-Dimensional Space



Poor reconstruction



Okay reconstruction



Best reconstruction

- PCA finds the line that has the smallest projection error: $\min_{w: \|w\|_2=1} \|X_c - zw^T\|_F^2$.

Viewpoint 1: Minimizing Reconstruction Error

Now, let's formalize this idea of "smallest projection error." This is the first of two ways we will define PCA.

Before

6	6	3	1	5	0
2	-	3	1	0	9
1		0	-	1	2
2		8	-	1	0
0		5	-	0	1
0	1	1	-	0	1

Eigenvalues
Eigenvectors

Eigenvalues and
same diquation

3	0	7	0
0	6	0	6
1	2	7	7
5	4	0	0
0	0	2	5
1	1	1	0
0	0	2	1
0	0	1	4

Solution

Math: PCA Minimizes The Linear Reconstruction Error 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)
- $w_s^T w_t = 0, w_s^T w_s = \|w_s\|_2^2 = 1, \forall s, t \neq s$ (orthogonal constraint)

Eigenvalues and same diquation

3	0	7	0
0	6	0	6
1	2	7	7
5	4	0	0
0	0	2	5
1	1	1	0
0	0	2	1
0	0	1	4

Solution

Math: PCA Minimizes The Linear Reconstruction Error Using Only $k \leq d$ Dimensions

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

- 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?
- $\|X_c - ZW^T\|_F^2 = \sum_{i=1}^n \|x_i^T - z_i^T W^T\|_2^2 = \sum_{i=1}^n \|x_i - W z_i\|_2^2$
- For analysis, let's simplify to a single dimension (i.e., $k = 1$):
 - $\sum_{i=1}^n \|x_i - z_i w\|_2^2$ where z_i is a scalar

Eigenvalues and same diquation

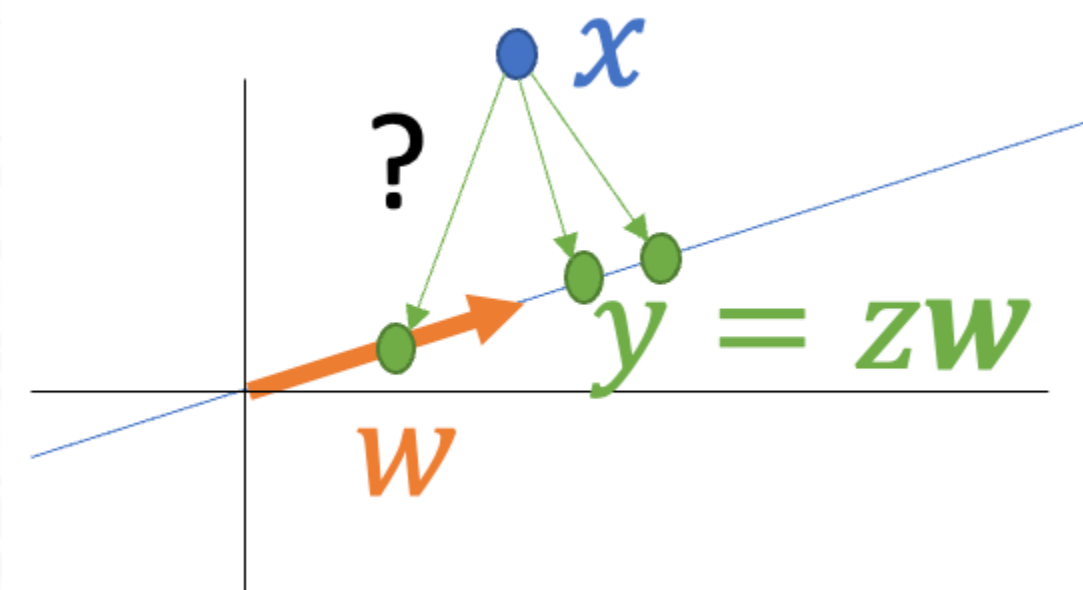
3	0	7	0
0	6	0	6
1	2	7	7
5	4	0	0
0	0	2	5
1	1	1	0
0	0	2	1
0	0	1	4

Solution

What is The Best Projection Given a Fixed Subspace (Line in 1D Case)?

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

$$\min_z \|x - zw\|_2^2$$



- The orthogonal projection!

- $z = x^T w = \|x\| \|w\| \cos \theta = \|x\| \cos \theta$

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

- zw is a scaled vector along the line defined by w .

Thus, We Can Simplify to Only Minimizing Over w

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

Now we can return to the Frobenius norm:

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

- What is zw^T ? Have we seen something like this before?
- This is the best low-rank approximation to X_c , which is given by the SVD!
- $w = v_1$ and $z = \sigma_1 u_1$, where σ_1 , u_1 , 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

Before

6	6	3	1	5	0	
2	-	3	1	0	9	
1		0	-	1	2	7
2		8	-	1	0	0
0		5	-	0	1	0
0	1	1	-	0	1	

Eigenvalues
Eigenvectors

Eigenvalues and
same diquation

3	0	7	0		
0	6	0	6	5	1
1	2	7		7	4
5	4	0	0		8
0		0	2	5	
1		1	1	0	
0		0	2	1	
0		0	1	4	

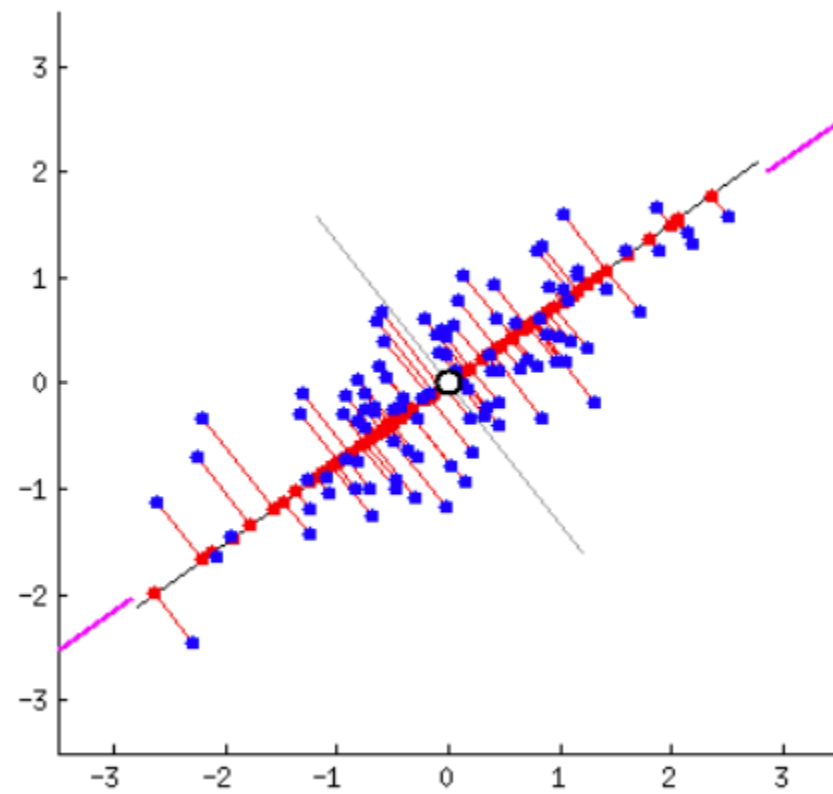
Solution

Viewpoint 2: Maximizing Projection Variance

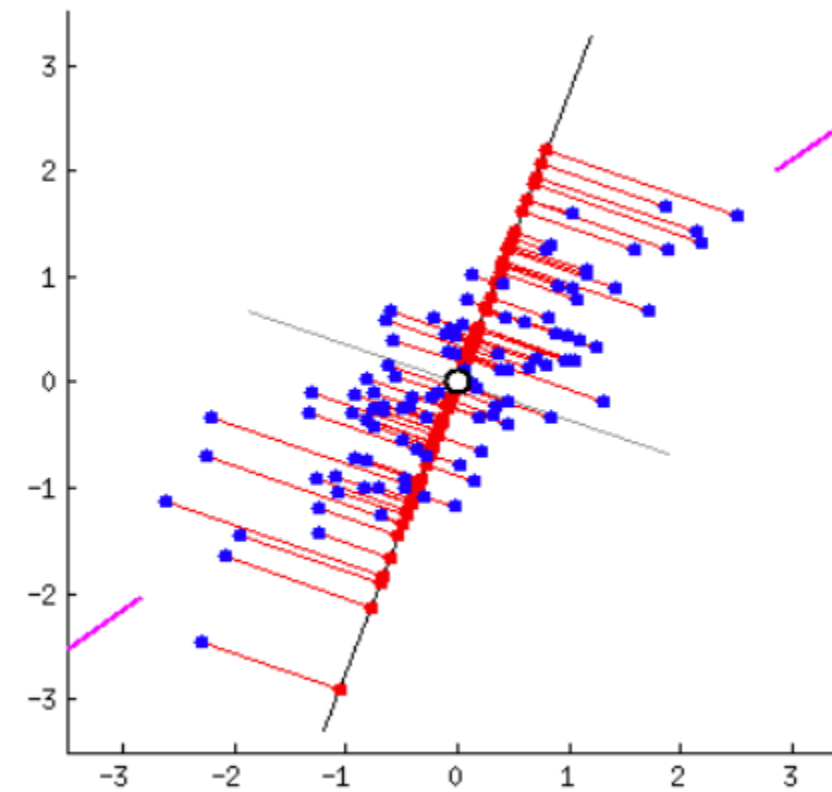
It turns out there's a completely different way to think about the "best" projection that leads to the same answer. Instead of minimizing error, we can try to maximize the information, or variance, of the projected data.



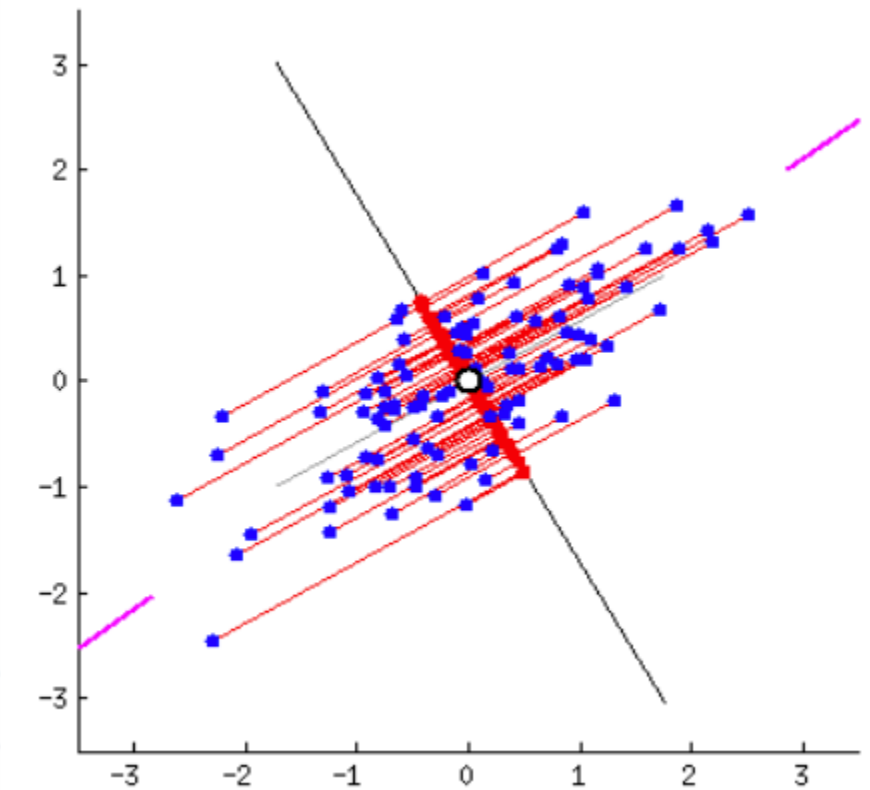
Minimizing Reconstruction Error (Red Lines) is Equivalent to Maximizing The Variance of Projection (Spread of Red Points)



Min reconstruction error, Max variance



Middle error and variance



Max reconstruction error, Min variance

Eigenvalues
Eigenvectors

Solution

Proving the Equivalence in 1D (Part 1)

This visual equivalence isn't just a coincidence. Let's formally prove it for the 1D case. The first step is to rewrite the squared reconstruction error for a single point x_i in a more useful form.

$$\begin{aligned}\|x_i - (x_i^T w)w\|_2^2 &= (x_i - (x_i^T w)w)^T (x_i - (x_i^T w)w) \\ &= x_i^T x_i - 2(x_i^T w)w^T x_i + (x_i^T w)^2 w^T w \\ &= \|x_i\|^2 - 2(x_i^T w)^2 + (x_i^T w)^2 \|w\|^2 \\ &= \|x_i\|^2 - (x_i^T w)^2\end{aligned}$$

Proving the Equivalence in 1D (Part 2)

Now we plug that result back into the total error minimization problem. Notice that minimizing the error is equivalent to maximizing the term we are subtracting.

$$\begin{aligned}\arg \min_w \sum_i \|x_i - (x_i^T w)w\|_2^2 &= \arg \min_w \sum_i (\|x_i\|^2 - (x_i^T w)^2) \\ &= \arg \min_w \sum_i -(x_i^T w)^2 \\ &= \arg \max_w \frac{1}{n} \sum_i (x_i^T w)^2 \\ &= \arg \max_w \frac{1}{n} \sum_i z_i^2 \\ &= \arg \max_w \sigma_z^2\end{aligned}$$

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

Reframing the Problem: Maximizing Variance

We've shown that minimizing reconstruction error is the same as maximizing the variance of the projected points. Let's formalize this new maximization problem. The term σ_z^2 can be rewritten using the data covariance matrix $\hat{\Sigma}$.

Let's rewrite this last term:

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

- Thus, our problem can be formulated as:

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

- The solution is the eigenvector q_1 of $\hat{\Sigma} = Q\Lambda Q^T$ corresponding to the largest eigenvalue λ_1 . $w^* = q_1$

Generalizing the Max Variance View to k Dimensions

This max-variance formulation generalizes neatly from one dimension to k dimensions. But what does the new objective function mean?

Minimize reconstruction error

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

Alternative problem

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

$$\text{Tr}(W^T X_c^T X_c W) = \text{Tr}((X_c W)^T (X_c W))$$

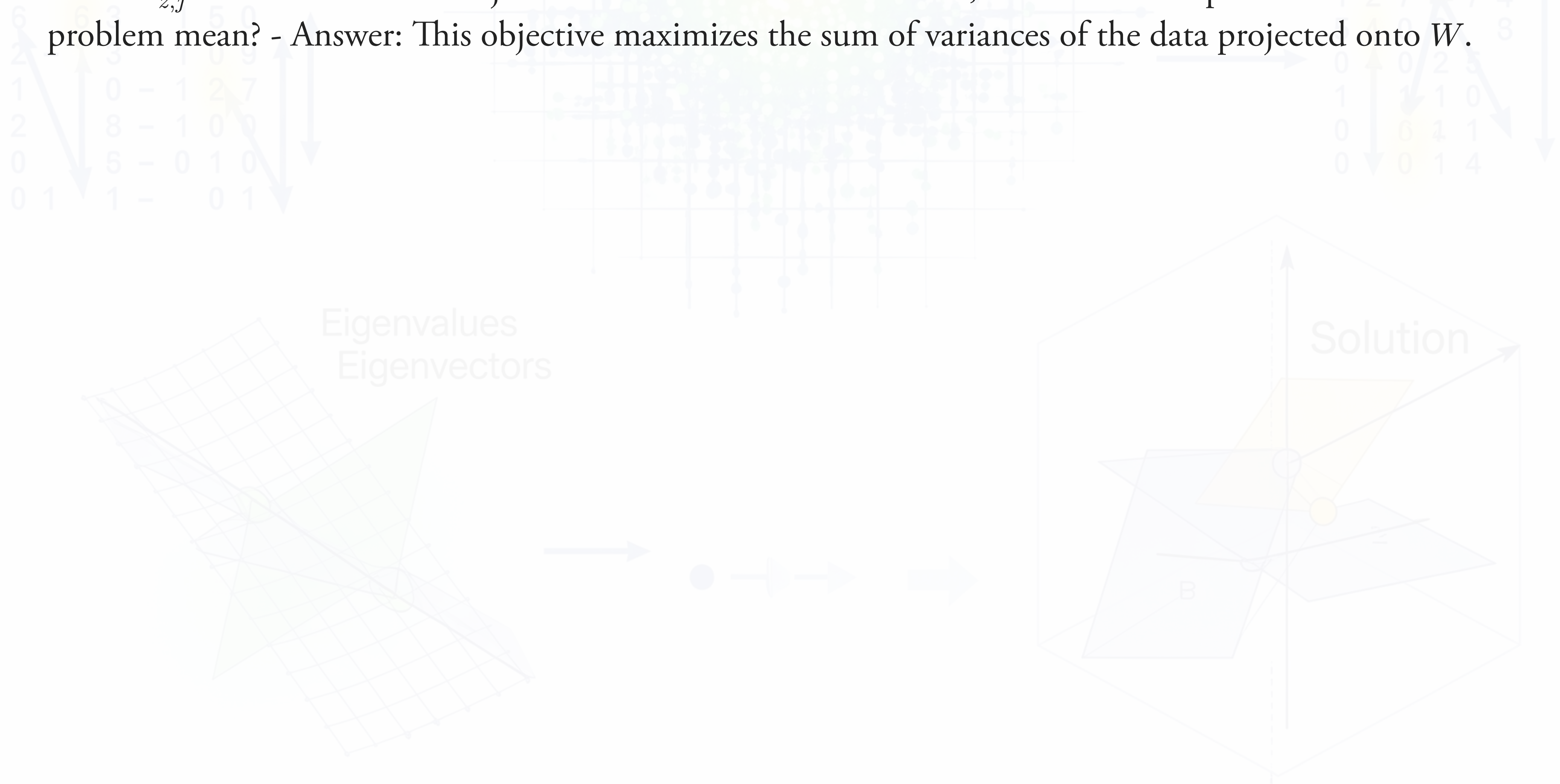
$$= \text{Tr}(Z^T Z)$$

$$= \sum_{j=1}^k z_j^T z_j$$

$$= n \sum_{j=1}^k \frac{1}{n} \sum_{i=1}^n z_{i,j}^2$$

$$= n \sum_{j=1}^k \sigma_{z,j}^2$$

where $\sigma_{z,j}^2$ is the variance of the j -th latent dimension. - Given this, what does the optimization problem mean? - Answer: This objective maximizes the sum of variances of the data projected onto W .



For $k > 1$, We Maximize The Sum of Variances for Each Latent Dimension

More generally we can formulate this as:

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

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

Eigenvalues and same diquation

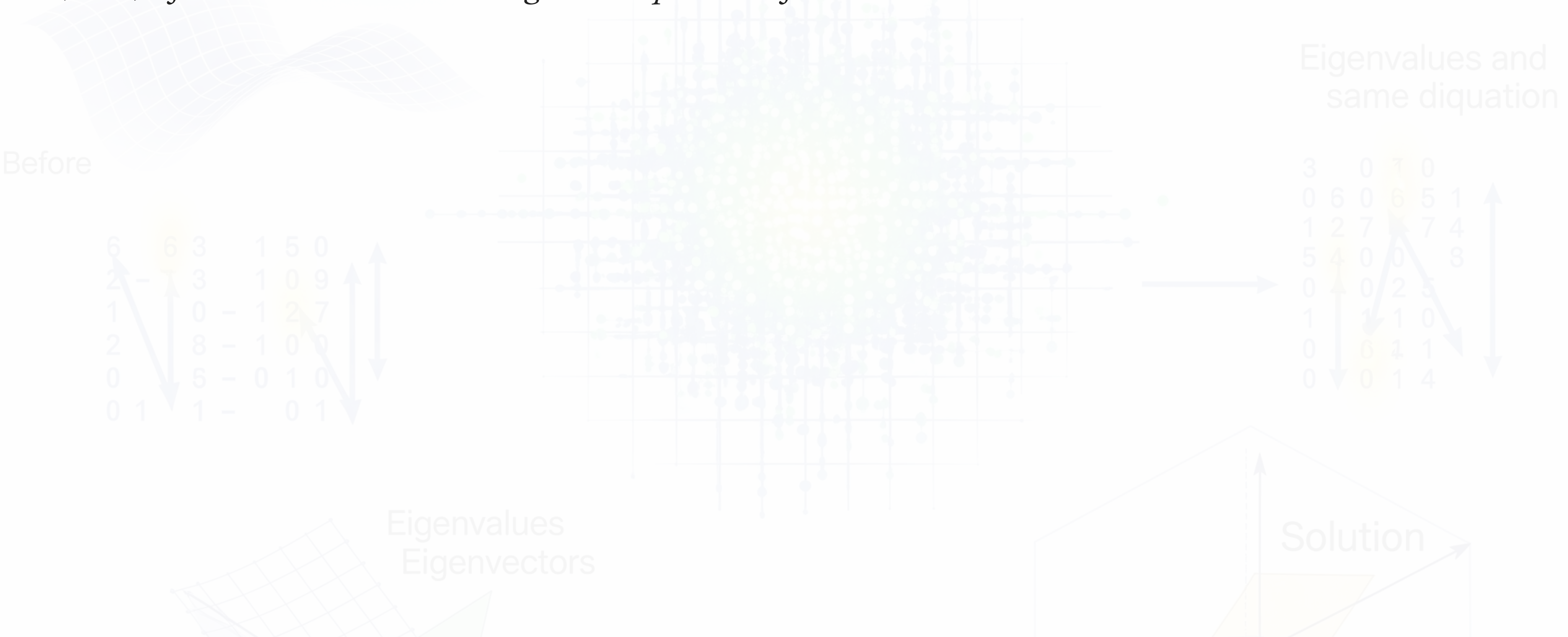
3	0	7	0
0	6	0	6
1	2	7	7
5	4	0	0
0	0	2	5
1	1	1	0
0	0	2	1
0	0	1	4

Eigenvalues
Eigenvectors

Solution

The Unifying Solution

We've now defined PCA in two different ways—minimizing reconstruction error and maximizing projected variance. Remarkably, they both have the same solution. This connects the Singular Value Decomposition (SVD) of the data matrix to the Eigendecomposition of its covariance matrix.



Equivalent Solutions: The Solution to Both Problems is The Top k Right Singular Vectors of X_c

Minimize reconstruction error

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

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

**Maximize variance of latent projection
(equivalent solution)**

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

where $\hat{\Sigma} := \frac{1}{n} X_c^T X_c$ is the covariance matrix

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

Eigenvalues
Eigenvectors

Solution

Questions?

Transformation

After

Before

6	6	3	1	5	0
2	-	3	1	0	9
1		0	-	1	2
2		8	-	1	0
0		5	-	0	1
0	1	1	-	0	1

Eigenvalues
Eigenvectors

Eigenvalues and
same diquation

3	0	7	0
0	6	0	6
1	2	7	7
5	4	0	0
0	0	2	5
1	1	1	0
0	0	2	1
0	0	1	4

Solution