Reconstruction and Principal Component Analysis

David I. Inouye

Dynamic Content · Last Updated: September 11, 2026

A useful reduction should reconstruct the observed data well

Return to the cells described by many measured features.

- Fewer coordinates cannot preserve every possible input exactly.
- The observed data may nevertheless lie near a lower-dimensional subspace.
- We will judge a representation by how well it reconstructs those observations.

> **What reconstruction error are we willing to accept in exchange for fewer features?**

Four steps lead to the PCA formulation

We will turn “reconstruct the data well” into a problem involving one matrix.

1. Define what a good reconstruction means — Next

Choose a representation size and measure reconstruction error.

2. Choose a convenient decoder basis.

An orthonormal basis can represent the same reconstruction subspace.

3. Find the best encoder for that decoder.

The closest point in the subspace determines the coordinates.

4. Formulate the subspace choice.

Combine these results into the PCA optimization problem.

Centering separates the mean from variation

For notation simplicity, we now call the **raw data** $\widetilde{X}$ and the **centered data** X.

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i, \qquad X = \widetilde{X} - \mathbf{1}_n \boldsymbol{\mu}^T.$$

- Each row is $\boldsymbol{x}_i^T = (\widetilde{\boldsymbol{x}}_i - \boldsymbol{\mu})^T$.
- Every feature column of X has mean zero.
- **From here on, X and x_i always denote centered data.**

Add the mean back to reconstruct raw measurements: $\widehat{\widetilde{\boldsymbol{x}}}_i = \boldsymbol{\mu} + \widehat{\boldsymbol{x}}_i$.

A linear encoder and decoder make reconstruction explicit

For one centered observation, use the same operator convention as $\boldsymbol{y} = A\boldsymbol{x}$:

$$\boldsymbol{z}_i = \underbrace{E}_{k\times d}\,\boldsymbol{x}_i, \qquad \widehat{\boldsymbol{x}}_i = \underbrace{D}_{d\times k}\,\boldsymbol{z}_i.$$

Stacking observations as rows gives:

$$\underbrace{Z}_{n\times k} = \underbrace{X}_{n\times d}\,\underbrace{E^T}_{d\times k}, \qquad \underbrace{\widehat{X}}_{n\times d} = \underbrace{Z}_{n\times k}\,\underbrace{D^T}_{k\times d}.$$

The batch reconstruction is $\widehat{X} = XE^TD^T$, with rank at most k.

Reconstruction error formalizes one meaning of preservation

Choose a linear encoder and decoder to minimize squared reconstruction error:

$$\min_{E,D} \|X - XE^T D^T\|_F^2.$$

The **Frobenius norm** measures the error across every entry:

$$\|M\|_F^2 = \sum_{i,j} M_{ij}^2, \qquad \|X - \widehat{X}\|_F^2 = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\|_2^2.$$

Four steps lead to the PCA formulation

We have an error measure. Now consider which reconstructions the decoder can produce.

1. Define what a good reconstruction means — Established

Choose a representation size and measure reconstruction error.

2. Choose a convenient decoder basis — Next

An orthonormal basis can represent the same reconstruction subspace.

3. Find the best encoder for that decoder.

The closest point in the subspace determines the coordinates.

4. Formulate the subspace choice.

Combine these results into the PCA optimization problem.

Matrix–vector multiplication combines the matrix's columns

For a 3×2 decoder, the familiar row dot products can be regrouped by column:

$$\begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \underbrace{\begin{bmatrix} d_{11}z_1 + d_{12}z_2 \\ d_{21}z_1 + d_{22}z_2 \\ d_{31}z_1 + d_{32}z_2 \end{bmatrix}}_{\text{one dot product per row}} = z_1 \underbrace{\begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \end{bmatrix}}_{\boldsymbol{d}_1} + z_2 \underbrace{\begin{bmatrix} d_{12} \\ d_{22} \\ d_{32} \end{bmatrix}}_{\boldsymbol{d}_2}.$$

For any number of columns: $\widehat{\boldsymbol{x}} = D\boldsymbol{z} = \sum_{j=1}^{k} z_j \boldsymbol{d}_j$.

- **Let z be arbitrary for now.** Every reconstruction lies in $\mathcal{S} = \operatorname{span}\{\boldsymbol{d}_1, \dots, \boldsymbol{d}_k\}$.
- First choose how D describes that subspace; then choose the best coordinates.

Scaling a decoder direction only changes its coordinates

A line can use a unit vector $\boldsymbol{u}$ or the scaled vector $3\boldsymbol{u}$.

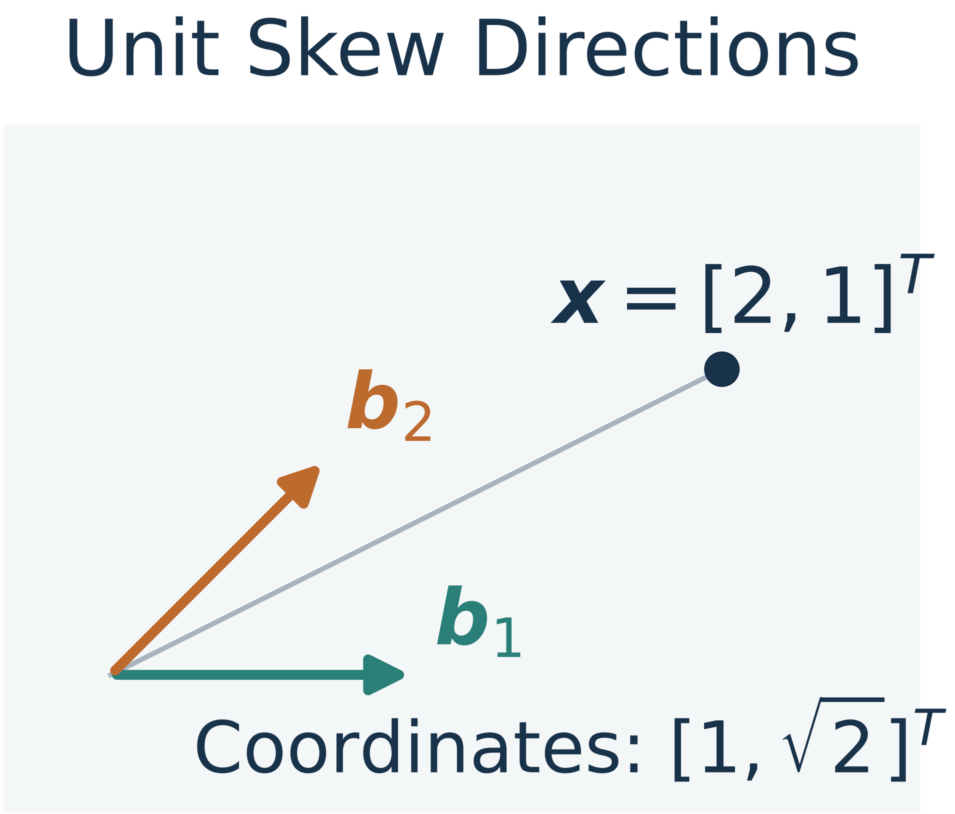
$$\underbrace{\boldsymbol{u}}_{\text{unit basis}}\ \underbrace{t}_{\text{coordinate}} = \underbrace{(3\boldsymbol{u})}_{\text{scaled basis}}\ \underbrace{(t/3)}_{\text{adjusted coordinate}}.$$

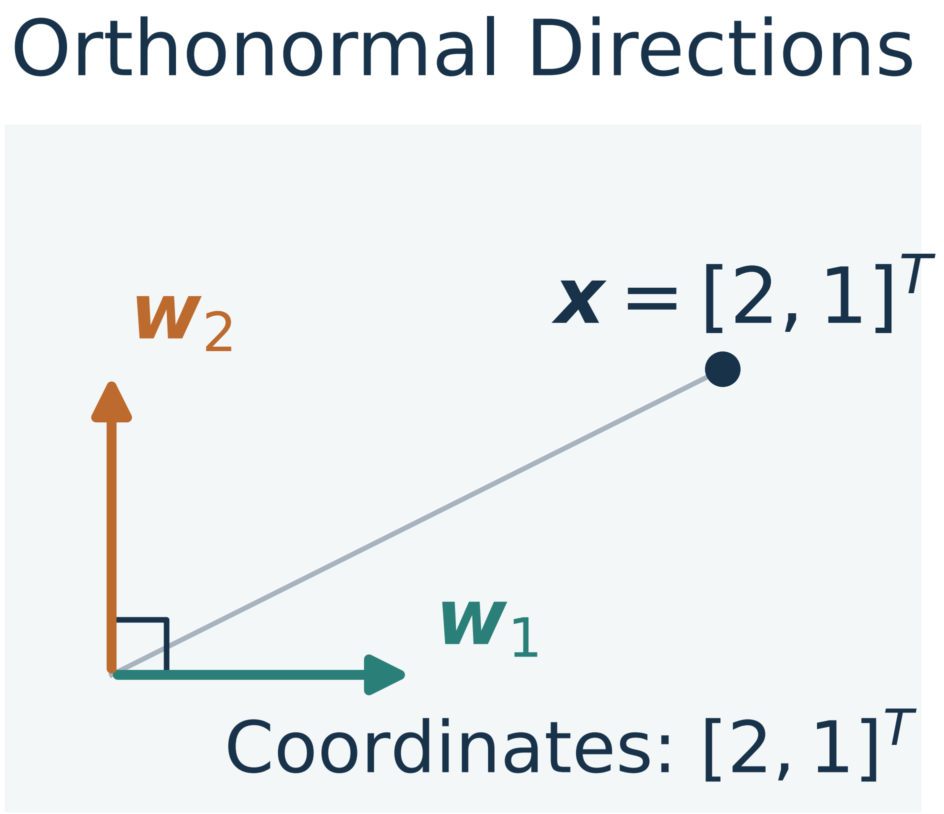
Both describe **the same point on the same line**.

- A nonzero column can be normalized if we rescale its coordinate to compensate.
- Unit columns remove an arbitrary scale choice; no reconstruction is lost.
- Next: can we choose perpendicular columns while keeping the same subspace?

Perpendicular unit directions can describe the same subspace

Both unit-vector pairs span the plane: either represents **every** $\boldsymbol{x} = [x_1, x_2]^T$.





For $\boldsymbol{b}_1 = [1, 0]^T, \boldsymbol{b}_2 = [1, 1]^T/\sqrt{2}$:

$$\boldsymbol{x} = (x_1 - x_2)\boldsymbol{b}_1 + \sqrt{2}\, x_2 \boldsymbol{b}_2.$$

For $\boldsymbol{w}_1 = [1, 0]^T, \boldsymbol{w}_2 = [0, 1]^T$:

$$\boldsymbol{x} = (\boldsymbol{w}_1^T \boldsymbol{x})\boldsymbol{w}_1 + (\boldsymbol{w}_2^T \boldsymbol{x})\boldsymbol{w}_2.$$

Same plane; simpler coordinates: an orthonormal basis gives each coefficient by a dot product.

QR lets us choose an orthonormal decoder without changing reconstruction

Start with any encoder $E_0 \in \mathbb{R}^{k \times d}$ and decoder $D_0 \in \mathbb{R}^{d \times k}$, where $k \leq d$. Suppose D_0 does not have orthonormal columns.

The **thin QR decomposition** writes $D_0 = WR$: $W^T W = I_k$ and R is upper triangular. Here $W \in \mathbb{R}^{d \times k}$ and $R \in \mathbb{R}^{k \times k}$.

Can we choose $D = W$ and a new E to keep every reconstruction the same?

$$
\begin{aligned}
D_0 E_0 \boldsymbol{x} &= (WR) E_0 \boldsymbol{x} && \text{(QR of the original decoder)} \\
&= W (R E_0) \boldsymbol{x} && \text{(Regroup the factors)} \\
&= D E \boldsymbol{x} && \text{(Choose } D = W, E = R E_0\text{)}
\end{aligned}
$$

Same reconstruction for every input: an orthonormal decoder is without loss of generality. Next, fix $D = W$ and find its best encoder.

Four steps lead to the PCA formulation

We may choose an orthonormal decoder without changing what we can reconstruct.

1. Define what a good reconstruction means — Established

Choose a representation size and measure reconstruction error.

2. Choose a convenient decoder basis — Established

An orthonormal basis can represent the same reconstruction subspace.

3. Find the best encoder for that decoder — Next

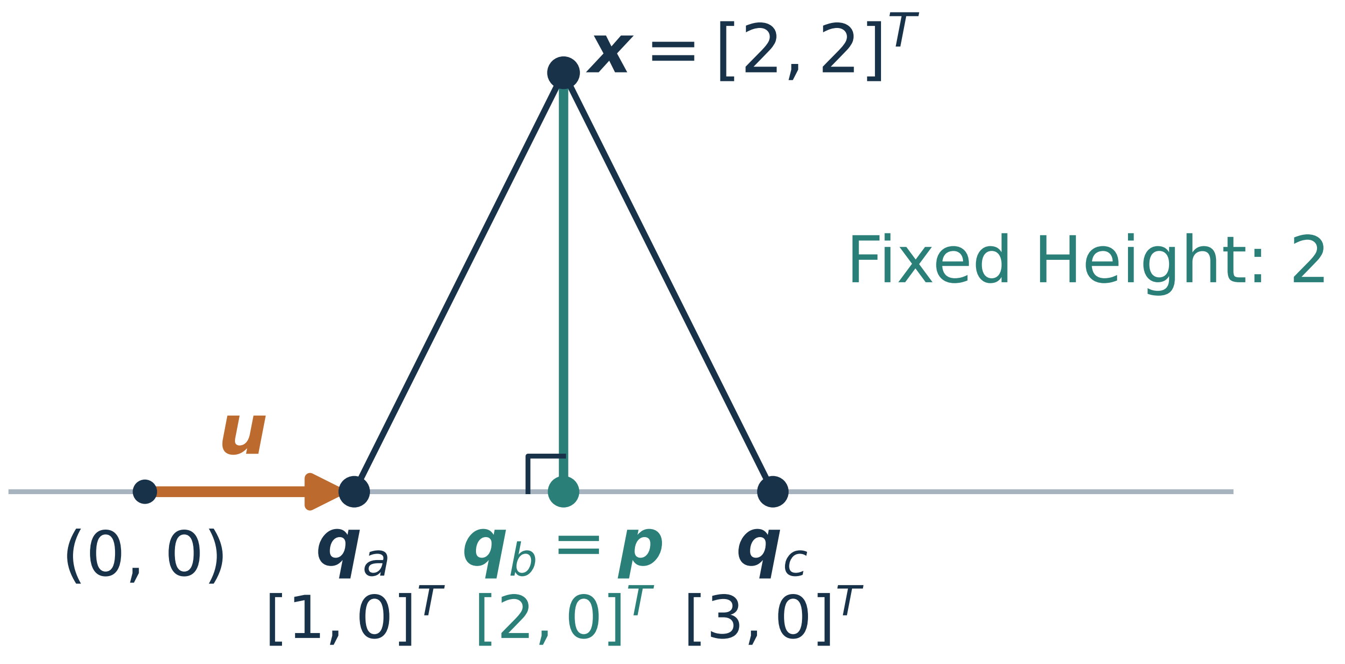
The closest point in the subspace determines the coordinates.

4. Formulate the subspace choice.

Combine these results into the PCA optimization problem.

In 2D, the perpendicular foot is the best reconstruction

Any unit u works. Illustrate with $\boldsymbol{u} = [1, 0]^T$, so $\boldsymbol{q} = z\boldsymbol{u} = [z, 0]^T$.

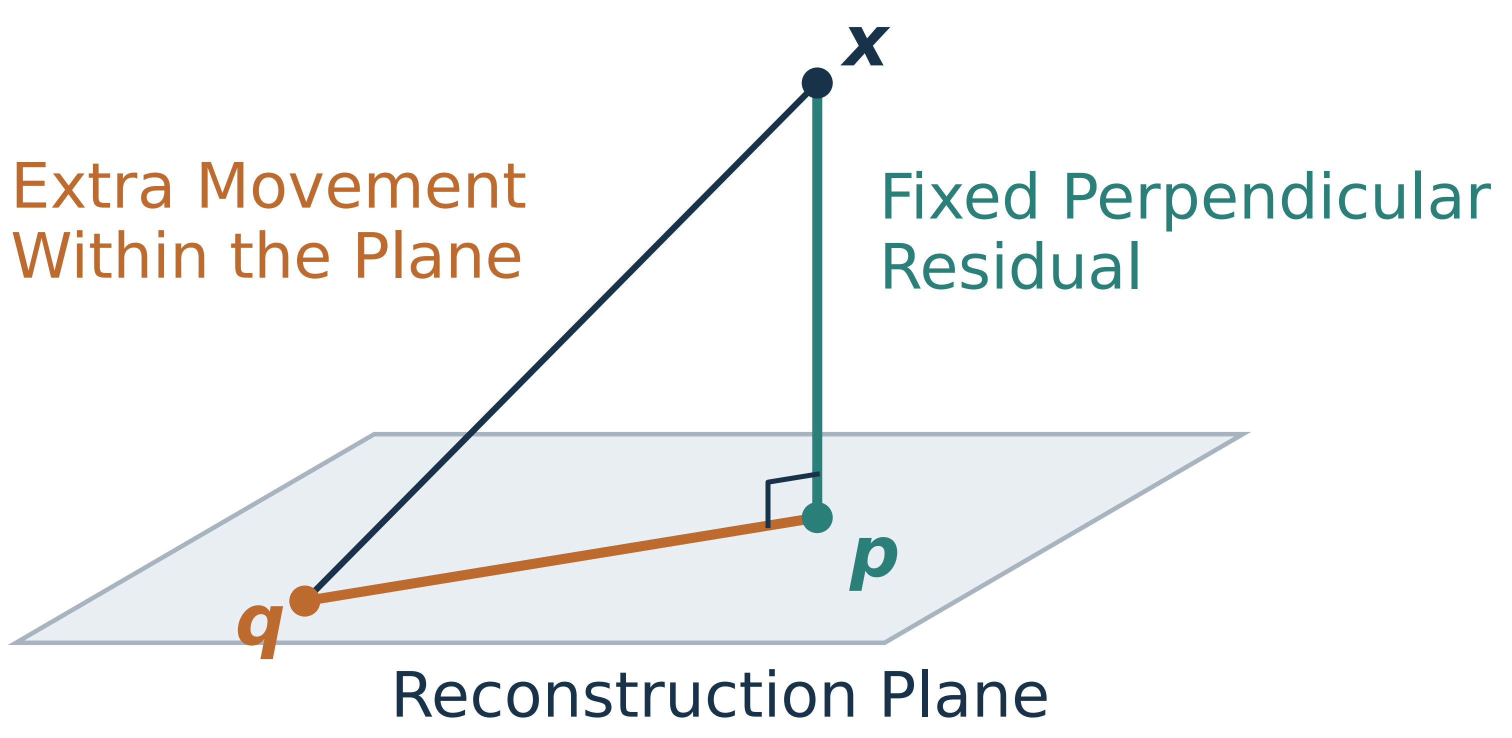


$$
\begin{aligned}
\|\boldsymbol{x} - \boldsymbol{q}\|_2^2 &= (x_1 - q_1)^2 + (x_2 - q_2)^2 && \text{(Squared distance)} \\
&= (2 - z)^2 + (2 - 0)^2 && \text{(Substitute coordinates)} \\
&= (2 - z)^2 + 4 && \text{(Simplify)}
\end{aligned}
$$

Choose $z = 2$: **the horizontal error is zero; the height remains.**

In 3D, moving within the plane cannot remove the height

Fix $D = W$ with two orthonormal columns spanning this plane. **Which point should we reconstruct?**

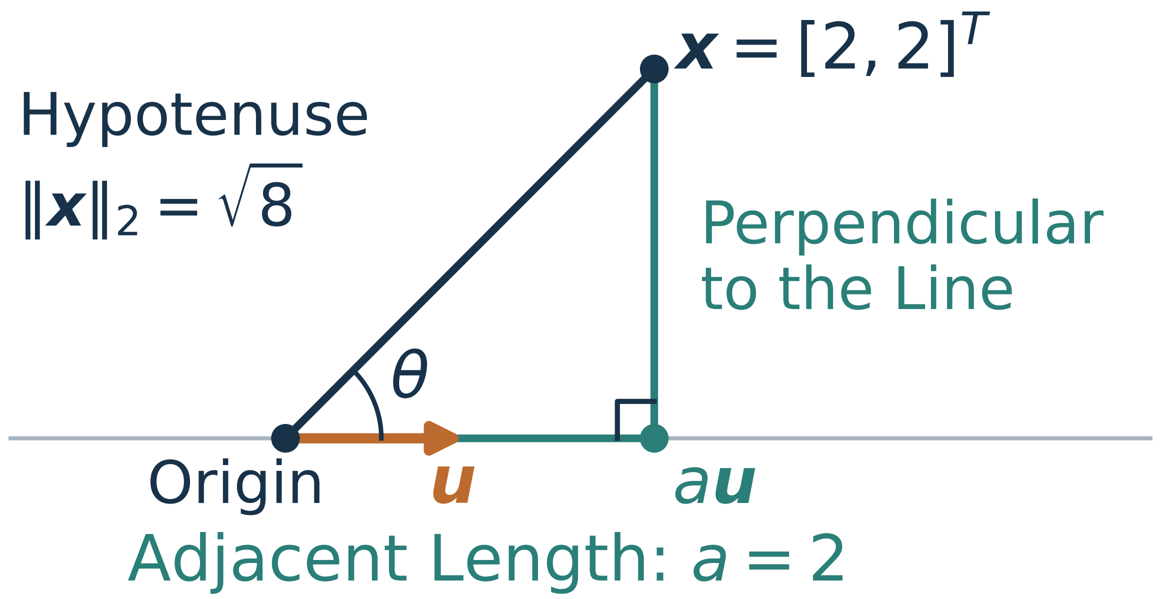


$$\|\boldsymbol{x} - \boldsymbol{q}\|_2^2 = \underbrace{\|\boldsymbol{x} - \boldsymbol{p}\|_2^2}_{\text{fixed perpendicular error}} + \underbrace{\|\boldsymbol{p} - \boldsymbol{q}\|_2^2}_{\text{extra error within the plane}}.$$

Choose $q = p$. **Moving in either plane direction only adds error.**

A dot product gives the length along a unit direction

Recall $\boldsymbol{u}^T\boldsymbol{x} = \|\boldsymbol{u}\|_2\|\boldsymbol{x}\|_2 \cos\theta$. Here $\|\boldsymbol{u}\|_2 = 1$.



$$
\begin{aligned}
a &= \text{hypotenuse}\,\frac{\text{adjacent}}{\text{hypotenuse}} && (a = \text{adjacent length}) \\
&= \|\boldsymbol{x}\|_2 \cos\theta && (\text{Definition of cosine}) \\
&= \boldsymbol{u}^T\boldsymbol{x} && (\text{Dot product; } \|\boldsymbol{u}\|_2 = 1)
\end{aligned}
$$

The foot is $(\boldsymbol{u}^T\boldsymbol{x})\boldsymbol{u}$. **This works in any dimension.**

Coordinates along the basis directions locate the point in the plane

For a plane with orthonormal basis $\boldsymbol{w}_1, \boldsymbol{w}_2$, read the two signed lengths:

$$a_1 = \boldsymbol{w}_1^T \boldsymbol{x}, \qquad a_2 = \boldsymbol{w}_2^T \boldsymbol{x}.$$

Combine the corresponding moves within the plane:

$$\boldsymbol{p} = a_1 \boldsymbol{w}_1 + a_2 \boldsymbol{w}_2 = \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = W \boldsymbol{a}.$$

- Perpendicular basis directions let us read each coordinate independently.
- For k orthonormal directions, the same construction gives $\boldsymbol{a} = W^T \boldsymbol{x}$ and $\boldsymbol{p} = W \boldsymbol{a}$.

Next, verify the geometry algebraically: is $\boldsymbol{x} - \boldsymbol{p}$ perpendicular to every decoder direction?

The residual is perpendicular to every decoder direction

Fix $D = W$, with $W^T W = I_k$. For one centered input, define:

$$\boldsymbol{a} = W^T \boldsymbol{x}, \qquad \boldsymbol{p} = W\boldsymbol{a}, \qquad \boldsymbol{r} = \boldsymbol{x} - \boldsymbol{p}.$$

The point $\boldsymbol{p}$ lies in the decoder's subspace. Check its residual:

$$\begin{aligned} W^T \boldsymbol{r} &= W^T(\boldsymbol{x} - W\boldsymbol{a}) && \text{(Substitute the residual)} \\ &= W^T \boldsymbol{x} - W^T W \boldsymbol{a} && \text{(Distribute)} \\ &= \boldsymbol{a} - \boldsymbol{a} = \boldsymbol{0} && (\boldsymbol{a} = W^T\boldsymbol{x} \text{ and } W^T W = I_k) \end{aligned}$$

Each entry is $\boldsymbol{w}_j^T \boldsymbol{r} = 0$: the residual is perpendicular to **every** basis direction.

Every reconstruction error has two perpendicular parts

For any coordinates $\boldsymbol{z}$, the decoder returns $W\boldsymbol{z}$. Insert the perpendicular foot $\boldsymbol{p} = W\boldsymbol{a}$:

$$\begin{aligned} \boldsymbol{x} - W\boldsymbol{z} &= (\boldsymbol{x} - \boldsymbol{p}) + (\boldsymbol{p} - W\boldsymbol{z}) && \text{(Add and subtract the same point)} \\ &= \boldsymbol{r} + (W\boldsymbol{a} - W\boldsymbol{z}) && \text{(Substitute the two parts)} \\ &= \boldsymbol{r} + W(\boldsymbol{a} - \boldsymbol{z}) && \text{(Factor out the decoder)} \end{aligned}$$

- $\boldsymbol{r}$ is the **fixed perpendicular part**, outside the subspace.
- $W(\boldsymbol{a} - \boldsymbol{z})$ is the **adjustable part**, inside the subspace.

Changing $\boldsymbol{z}$ moves only within the subspace, so it **cannot cancel** $\boldsymbol{r}$.

The best coordinates remove all adjustable error

Orthonormal columns preserve coordinate lengths: $\|W\boldsymbol{b}\|_2^2 = \boldsymbol{b}^T W^T W \boldsymbol{b} = \|\boldsymbol{b}\|_2^2$.

$$
\begin{aligned}
\|\boldsymbol{x} - W\boldsymbol{z}\|_2^2 &= \|\boldsymbol{r} + W(\boldsymbol{a} - \boldsymbol{z})\|_2^2 && \text{(Use the error split)} \\
&= \|\boldsymbol{r}\|_2^2 + \|W(\boldsymbol{a} - \boldsymbol{z})\|_2^2 && \text{(Pythagoras: the parts are perpendicular)} \\
&= \|\boldsymbol{r}\|_2^2 + \|\boldsymbol{a} - \boldsymbol{z}\|_2^2 && \text{(Orthonormal columns)}
\end{aligned}
$$

The first term is fixed. The second is nonnegative and becomes zero **only at** $\boldsymbol{z} = \boldsymbol{a} = W^T\boldsymbol{x}$.

Therefore $\widehat{\boldsymbol{x}} = WW^T\boldsymbol{x}$ is the unique closest reconstruction. Write $P = WW^T$.

One orthonormal basis supplies both encoder and decoder

We chose the decoder $D = W$ without loss of generality; minimizing error gives $E = W^T$.

For the batch of centered observations:

$$Z = XW, \qquad \widehat{X} = ZW^T = XWW^T.$$

Role	Choice	Property
Encoder	$E = W^T \in \mathbb{R}^{k \times d}$	Orthonormal rows: $EE^T = I_k$.
Decoder	$D = W \in \mathbb{R}^{d \times k}$	Orthonormal columns: $D^T D = I_k$.

Together, these choices attain the best reconstruction error allowed by general E, D.

What remains to choose is the subspace: which k directions should W contain?

Four steps lead to the PCA formulation

For a fixed orthonormal decoder, its transpose gives the best encoder. Which subspace should we choose?

1. Define what a good reconstruction means — Established

Choose a representation size and measure reconstruction error.

2. Choose a convenient decoder basis — Established

An orthonormal basis can represent the same reconstruction subspace.

3. Find the best encoder for that decoder — Established

The closest point in the subspace determines the coordinates.

4. Formulate the subspace choice — Next

Combine these results into the PCA optimization problem.

Build the objective from requirements and choices

> **We want k linear features that reconstruct centered measurements well. Which parts are requirements, and which are convenient parameter choices?**
> **Explain the roles of k, squared error, and $W^TW = I_k$. Would low reconstruction error alone establish biological usefulness?**

k limits representation size; squared error defines "well." An orthonormal basis is sufficient for an optimum, but biological usefulness needs more evidence.

PCA chooses the best linear reconstruction subspace

For centered data X and a chosen dimension $1 \leq k < d$, **PCA** solves:

$$\min_{W \in \mathbb{R}^{d \times k}} \|X - XWW^T\|_F^2 \quad \text{subject to} \quad W^T W = I_k.$$

- **Objective:** choose the subspace that minimizes total squared reconstruction error.
- **Constraint:** $W^T W = I_k$ means the columns of W are orthonormal.
- **One matrix to choose:** we may use decoder $D = W$ without loss of generality; its optimal encoder is $E = W^T$.

$$\underbrace{W^T W = I_k}_{\text{identity in coordinate space}} \quad , \quad \underbrace{WW^T \neq I_d}_{\text{projection in the original space, since } k<d} \quad .$$

Three steps connect the PCA problem to its solution

We have formulated PCA using one orthonormal basis W. Now we will solve that problem.

1. Rewrite the objective — Next

Least reconstruction error means most retained squared length.

2. Find the optimal directions

Use the SVD to prove that the leading right singular vectors are optimal.

3. Interpret the solution

Read reconstruction error and coordinate variance from the singular values.

Minimizing reconstruction error means keeping the most squared length

For any orthonormal W, the reconstruction and residual are perpendicular.

$$\begin{aligned}
\|X\|_F^2 &= \|XW W^T\|_F^2 + \|X - XWW^T\|_F^2 \\
&= \|XW\|_F^2 + \|X - XWW^T\|_F^2
\end{aligned}$$

(Pythagoras, summed over observations)

(W preserves coordinate lengths)

The data X are fixed, so

$$\underbrace{\min_{W^TW=I_k} \|X - XWW^T\|_F^2}_{\text{least reconstruction error}} \iff \underbrace{\max_{W^TW=I_k} \|XW\|_F^2}_{\text{most retained squared length}} .$$

Three steps connect the PCA problem to its solution

The total squared length is fixed. We only need to maximize what the coordinates retain.

1. Rewrite the objective — Established

Least reconstruction error means most retained squared length.

2. Find the optimal directions — Next

Use the SVD to prove that the leading right singular vectors are optimal.

3. Interpret the solution

Read reconstruction error and coordinate variance from the singular values.

The SVD expresses a data matrix as a sum of simple patterns

Now decompose the centered data itself: $X = U\Sigma V^T$.

$$X = \sum_{j=1}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T, \qquad r = \operatorname{rank}(X).$$

Part	Meaning for the Data Matrix
$\boldsymbol{v}_j \in \mathbb{R}^d$	One pattern across measured features.
$\sigma_j \boldsymbol{u}_j \in \mathbb{R}^n$	How strongly that pattern appears in each observation.
$\sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T$	A matrix of rank one: one shared pattern with observation-specific weights.

Truncation keeps some patterns and drops the rest

For $1 \le k \le r$, keeping the leading k terms gives:

$$X_k = \sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T = U_k \Sigma_k V_k^T.$$

- U_k: first k columns of U.
- Σ_k: the $k \times k$ diagonal matrix of retained singular values.
- V_k: first k columns of V.

Does this choice retain the most squared length among all k-dimensional subspaces?

The SVD measures how much of each direction we keep

Write $W = [\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_k]$ and $X = U\Sigma V^T$, with $\sigma_j = 0$ for $j > \mathrm{rank}(X)$.

$$\begin{aligned}
\|XW\|_F^2 &= \sum_{\ell=1}^{k} \|X\boldsymbol{w}_\ell\|_2^2 && \text{(Sum squared column lengths)} \\
&= \sum_{\ell=1}^{k}\sum_{j=1}^{d} \sigma_j^2 (\boldsymbol{v}_j^T \boldsymbol{w}_\ell)^2 && \text{(V^T reads coordinates; Σ scales them; U preserves length)} \\
&= \sum_{j=1}^{d} \sigma_j^2 \alpha_j && \text{($\alpha_j = \textstyle\sum_{\ell=1}^{k} (\boldsymbol{v}_j^T \boldsymbol{w}_\ell)^2$)}
\end{aligned}$$

- $\alpha_j = \|W^T \boldsymbol{v}_j\|_2^2$: how much of unit direction $\boldsymbol{v}_j$ we keep, so $0 \le \alpha_j \le 1$.
- $\sum_j \alpha_j = \sum_{\ell=1}^{k} \|\boldsymbol{w}_\ell\|_2^2 = k$: k **units in total**.

The leading right singular vectors solve the PCA objective

Order $\sigma_1^2 \geq \cdots \geq \sigma_d^2$. We have k units to allocate, with at most one per direction.

$$\sum_{j=1}^{d} \sigma_j^2 \alpha_j \leq \sum_{j=1}^{k} \sigma_j^2 \alpha_j + \sigma_k^2 \left(k - \sum_{j=1}^{k} \alpha_j \right) \qquad \text{(For } j > k, \sigma_j^2 \leq \sigma_k^2\text{)}$$

$$= k\sigma_k^2 + \sum_{j=1}^{k} (\sigma_j^2 - \sigma_k^2)\alpha_j \qquad \text{(Regroup)}$$

$$\leq \sum_{j=1}^{k} \sigma_j^2 \qquad \text{(}\alpha_j \leq 1\text{; each coefficient is nonnegative)}$$

$W = V_k$ **attains the bound:** $\alpha_1 = \cdots = \alpha_k = 1$, all others zero.

$$Z = XV_k = U_k \Sigma_k, \qquad \widehat{X} = XV_kV_k^T = X_k.$$

Three steps connect the PCA problem to its solution

The leading right singular vectors attain the optimum. Now read what this solution keeps and loses.

1. Rewrite the objective — Established

Least reconstruction error means most retained squared length.

2. Find the optimal directions — Established

Use the SVD to prove that the leading right singular vectors are optimal.

3. Interpret the solution — Next

Read reconstruction error and coordinate variance from the singular values.

Discarded singular values quantify what reconstruction loses

Subtract the optimal retained squared length from the fixed total:

$$\|X - X_k\|_F^2 = \underbrace{\sum_{j=1}^{r} \sigma_j^2}_{\text{total}} - \underbrace{\sum_{j=1}^{k} \sigma_j^2}_{\text{retained}} = \sum_{j>k} \sigma_j^2 .$$

For singular values $6, 3, 1$:

Retained Features	Squared Reconstruction Error
$k = 1$	$3^2 + 1^2 = 10$
$k = 2$	$1^2 = 1$
$k = 3$	0

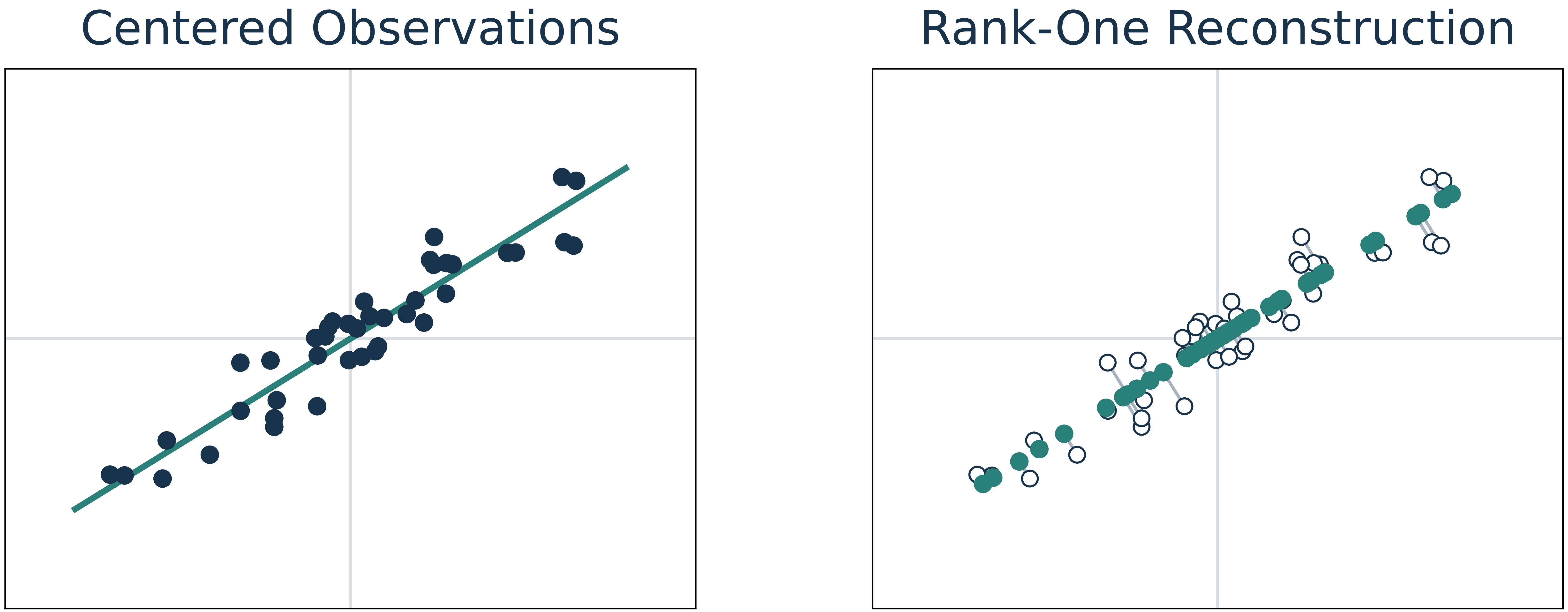
Squared singular values give the variance of each PCA coordinate

The jth coordinate across observations is $Z_{:j} = X\boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j$. Since X is centered, this coordinate also has mean zero.

$$\begin{aligned}
\mathrm{Var}_{\text{sample}}(Z_{:j}) &= \frac{1}{n-1}\sum_{i=1}^{n} Z_{ij}^2 && \text{(Variance of a centered coordinate; } n > 1\text{)} \\
&= \frac{\sigma_j^2}{n-1}\|\boldsymbol{u}_j\|_2^2 && \text{(Substitute } Z_{:j} = \sigma_j \boldsymbol{u}_j\text{)} \\
&= \frac{\sigma_j^2}{n-1} && (\boldsymbol{u}_j \text{ is a unit vector})
\end{aligned}$$

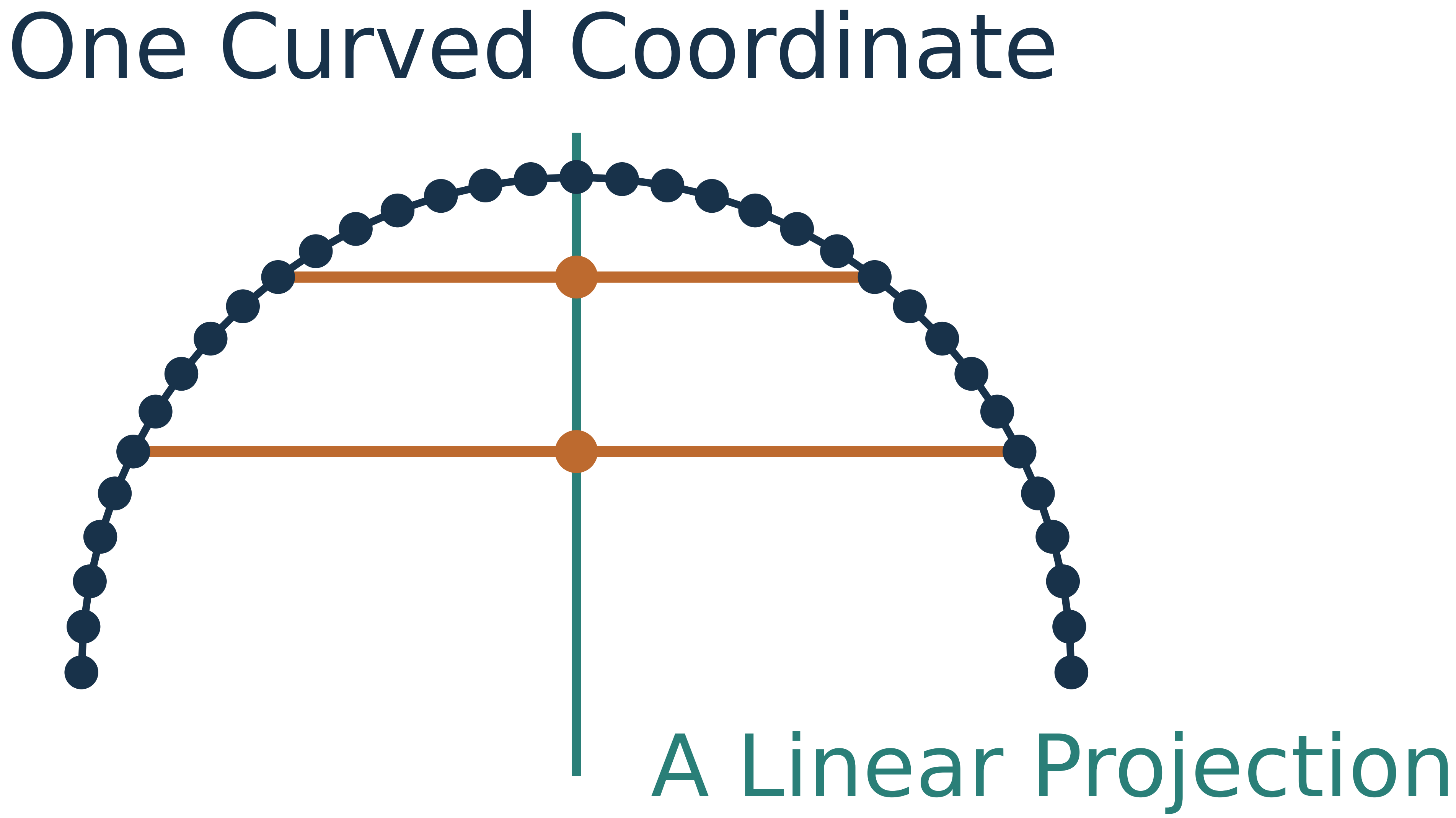
The leading PCA coordinates therefore have the **largest variances**. This interprets the solution we already proved.

One retained direction can reconstruct a narrow data cloud



Original-scale reconstruction: $\widehat{\widetilde{\boldsymbol{x}}}_i = \boldsymbol{\mu} + W\boldsymbol{z}_i$.

A curved structure can be low-dimensional without being linear



Would one linear feature reconstruct every point exactly? Could a nonlinear coordinate do so? What does “one-dimensional” mean in each claim?

PCA solves the stated problem; usefulness still needs evidence

- **Chosen objective:** squared error in the selected measurements and units.
- **Mathematical solution:** leading right singular vectors of centered data.
- **Remaining questions:** which features and preprocessing, how many components, and whether the retained variation matters for the intended use.

Next: translate this mathematical contract into arrays and reliable numerical computation—then ask what the observed data justify claiming about new cases.