# Linear Autoencoder → PCA

A clean derivation from reconstruction error to eigenvectors

**Notation note.** The handwritten pages mix row- and column-vector notation in one step. Here the same derivation is written consistently using *column vectors*. Thus

$$z = U^\top x \qquad \Longleftrightarrow \qquad z^\top = x^\top U.$$

No conceptual step is changed.

## 1. Low-dimensional representation

Suppose the data matrix is

$$X \in \mathbb{R}^{D\times N},$$

with $N$ samples in $D$ dimensions. We approximate it by

$$\boxed{X \approx UZ}$$

where

$$U \in \mathbb{R}^{D\times d}, \qquad Z \in \mathbb{R}^{d\times N}, \qquad d < D.$$

For one sample,

$$x \in \mathbb{R}^{D}, \qquad z \in \mathbb{R}^{d},$$

and the reconstruction is $Uz$.

We constrain the columns of $U$ to be orthonormal:

$$\boxed{U^\top U = I_d.}$$

The reconstruction objective for one sample is

$$\min_z \|x - Uz\|^2 .$$

## 2. First solve for the code $z$

Expand the squared error:

$$\begin{aligned}\|x - Uz\|^2 &= (x - Uz)^\top (x - Uz)\\ &= x^\top x - x^\top Uz - z^\top U^\top x + z^\top U^\top Uz.\end{aligned}$$

Because $U^\top U = I_d$,

$$z^\top U^\top Uz = z^\top z = \|z\|^2 .$$

Also $x^\top Uz$ is a scalar, so

$$z^\top U^\top x = (x^\top Uz)^\top = x^\top Uz.$$

Therefore

$$\|x - Uz\|^2 = \|x\|^2 + \|z\|^2 - 2x^\top Uz.$$

Differentiate with respect to $z$:

$$\frac{\partial}{\partial z}\left(\|x\|^2 + \|z\|^2 - 2x^\top Uz\right) = 2z - 2U^\top x.$$

At the minimum,

$$2z - 2U^\top x = 0,$$

so

$$\boxed{z = U^\top x.}$$

**Why differentiate with respect to $z$ first?** For a fixed basis $U$, the best code $z$ has a simple closed-form solution. Once that is substituted back, the only remaining unknown is $U$.

## 3. Encoder, decoder, and projection

Substituting the optimal code gives

$$\hat{x} = Uz = UU^\top x.$$

So the model is a linear autoencoder with tied encoder/decoder weights:

$$\boxed{x} \xrightarrow[\text{encode}]{U^\top} \boxed{z} \xrightarrow[\text{decode}]{U} \boxed{\hat{x}}$$

For all samples,

$$Z = U^\top X, \qquad \hat{X} = UU^\top X.$$

Thus $UU^\top$ is the orthogonal projector onto the $d$-dimensional subspace spanned by the columns of $U$.

The average reconstruction problem is

$$\boxed{\min_U \frac{1}{N} \sum_{i=1}^{N} \left\| x_i - UU^\top x_i \right\|^2} \qquad \text{subject to } U^\top U = I_d.$$

## 4. Reconstruction minimization becomes projection maximization

For an orthogonal projection,

$$\hat{x} = UU^\top x$$

lies in the chosen subspace, while

$$x - \hat{x}$$

is perpendicular to that subspace. Hence Pythagoras gives

$$\|x\|^2 = \left\| UU^\top x \right\|^2 + \left\| x - UU^\top x \right\|^2.$$

For fixed data, $\|x\|^2$ is constant. Therefore minimizing the reconstruction error is equivalent to maximizing the length of the projected component:

$$\min_U \left\| x - UU^\top x \right\|^2 \iff \max_U \left\| UU^\top x \right\|^2.$$

Because the columns of $U$ are orthonormal, $U$ preserves the norm of any vector in $\mathbb{R}^d$. For any $y \in \mathbb{R}^d$,

$$\|Uy\|^2 = (Uy)^\top (Uy) = y^\top U^\top U y = y^\top y = \|y\|^2.$$

Taking $y = U^\top x$,

$$\boxed{\left\| UU^\top x \right\|^2 = \left\| U^\top x \right\|^2.}$$

So the objective becomes

$$\boxed{\max_U \frac{1}{N} \sum_{i=1}^{N} \left\| U^\top x_i \right\|^2} \qquad \text{subject to } U^\top U = I_d.$$

**Interpretation.** Choose an orthonormal $d$-dimensional subspace that captures as much of the data's squared length as possible. This is the PCA objective.

## 5. Write the objective using the trace

For one sample,

$$\begin{aligned}\left\|U^\top x\right\|^2 &= (U^\top x)^\top (U^\top x) \\ &= x^\top U U^\top x.\end{aligned}$$

A scalar equals its own trace, so

$$x^\top U U^\top x = \operatorname{tr}\left(x^\top U U^\top x\right).$$

Using the cyclic property of the trace,

$$\operatorname{tr}(ABCD) = \operatorname{tr}(DABC) = \operatorname{tr}(CDAB),$$

we obtain

$$\operatorname{tr}\left(x^\top U U^\top x\right) = \operatorname{tr}\left(U^\top x x^\top U\right).$$

Averaging across samples gives

$$\max_U \operatorname{tr}\left(U^\top C U\right),$$

where

$$\boxed{C = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top = \frac{1}{N} X X^\top.}$$

Thus

$$\boxed{\max_U \operatorname{tr}(U^\top C U) \quad \text{subject to } U^\top U = I_d.}$$

## 6. Insert the eigendecomposition

Write the eigendecomposition of $C$ as

$$C = Q \Lambda Q^\top,$$

where

$$Q \in \mathbb{R}^{D\times D}, \qquad Q^\top Q = I_D,$$

and

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_D).$$

Then

$$\max_U \operatorname{tr}\left(U^\top Q \Lambda Q^\top U\right).$$

Define

$$\boxed{W = Q^\top U} \qquad (W \in \mathbb{R}^{D\times d}).$$

Since both $Q$ and the columns of $U$ are orthonormal,

$$W^\top W = U^\top Q Q^\top U = U^\top U = I_d.$$

The objective is now

$$\boxed{\max_W \operatorname{tr}(W^\top \Lambda W), \qquad W^\top W = I_d.}$$

## 7. Why the largest eigenvectors are selected

First consider only one direction, so $d = 1$. Then $W$ is a vector $w \in \mathbb{R}^D$, and

$$w^\top w = 1.$$

The objective becomes

$$w^\top \Lambda w = \lambda_1 w_1^2 + \lambda_2 w_2^2 + \cdots + \lambda_D w_D^2.$$

Because

$$w_1^2 + w_2^2 + \cdots + w_D^2 = 1,$$

the maximum is achieved by placing all the weight on the largest diagonal entry of $\Lambda$. If

$$\lambda_1 = \max_i \lambda_i,$$

then

$$w = e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^\top.$$

Therefore

$$Q^\top u = e_1 \quad \Longrightarrow \quad u = Qe_1 = q_1,$$

so the optimal direction is the eigenvector $q_1$ associated with the largest eigenvalue.

For $d > 1$, the columns of $W$ must be orthonormal. The same argument selects the coordinates corresponding to the $d$ largest eigenvalues. Hence

$$\boxed{U = [q_1\ q_2\ \cdots\ q_d]},$$

where $q_1, \ldots, q_d$ are the top $d$ eigenvectors of $C$.

**Final result.**

$$\boxed{\text{Linear autoencoder} + \text{squared reconstruction loss} + \text{orthonormal } U \implies \text{PCA subspace.}}$$

The encoder computes $z = U^\top x$, the decoder reconstructs $\hat{x} = Uz = UU^\top x$, and the optimal columns of $U$ are the principal eigenvectors of the data second-moment matrix $C = XX^\top / N$.

## 8. Dimension check

| **Quantity** | **Shape** | **Role** |
|---|---|---|
| $X$ | $D \times N$ | data matrix |
| $x$ | $D \times 1$ | one sample |
| $U$ | $D \times d$ | orthonormal basis / decoder |
| $z$ | $d \times 1$ | low-dimensional code |
| $Z$ | $d \times N$ | codes for all samples |
| $Q$ | $D \times D$ | eigenvector matrix of $C$ |
| $\Lambda$ | $D \times D$ | diagonal eigenvalue matrix |
| $W = Q^\top U$ | $D \times d$ | coordinates of the chosen subspace in the eigenbasis |

Prepared as a cleaned, typed version of the uploaded handwritten notes. The mathematical structure and sequence of the original notes have been preserved; notation was standardized for readability.