

Machine Learning

Textbook: Machine Learning Refined, by Watt, Borhani, and Katsaggelos, 2nd edition:
MLRefined.com

Previous Lecture

- ♦ Orthogonal projection is the least-squares approximation of a vector by a pt in a chosen subspace.
- ♦ For a unit direction $\mathbf{u}$, the coordinate is $z = \mathbf{u}^T \mathbf{x}$ and the reconstruction is $\hat{\mathbf{x}} = \mathbf{u} \mathbf{u}^T \mathbf{x}$.
- ♦ For an orthonormal basis U , coordinates are $\mathbf{z} = U^T \mathbf{x}$ and the projection matrix is UU^T .
- ♦ Projection residuals are orthogonal to the retained subspace, giving a Pythagorean decomposition of squared length.
- ♦ Centering lets a linear subspace through the origin represent an affine subspace through the data mean.
- ♦ Minimizing reconstruction error is equivalent to maximizing the squared length captured by the projection.



§8.3 - Principal Component Analysis

Last time:

once a projection subspace was chosen,
we knew how to compress and reconstruct the data.

But different directions produced different reconstruction errors.

Today:

let the data choose the best projection direction.

This is **principal component analysis**.

Start With Centered Data

Let

$$\mathbf{x}_1, \dots, \mathbf{x}_P \in \mathbb{R}^N.$$

Mean:

$$\boldsymbol{\mu} = \frac{1}{P} \sum_{p=1}^P \mathbf{x}_p.$$

Centered observations:

$$\mathbf{q}_p = \mathbf{x}_p - \boldsymbol{\mu}.$$

Then

$$\frac{1}{P} \sum_{p=1}^P \mathbf{q}_p = \mathbf{0}.$$

PCA is applied to the centered data.

The mean is added back only when reconstructing in the original coordinates.

Choose One Unit Direction

Let

$$\mathbf{u} \in \mathbb{R}^N$$

satisfy

$$\|\mathbf{u}\| = 1.$$

The scalar coordinate of observation p is

$$z_p = \mathbf{u}^T \mathbf{q}_p.$$

The reconstructed centered observation is

$$\hat{\mathbf{q}}_p = z_p \mathbf{u} = \mathbf{u} \mathbf{u}^T \mathbf{q}_p.$$

Question:

which unit vector

$$\mathbf{u}$$

should we choose?

Criterion 1: Minimize Reconstruction Error

From Lesson 1.22:

$$E(\mathbf{u}) = \frac{1}{P} \sum_{p=1}^P \left\| \mathbf{q}_p - (\mathbf{u}^T \mathbf{q}_p) \mathbf{u} \right\|^2.$$

PCA chooses the direction

with the smallest mean squared reconstruction distance.

But there is an equivalent and more revealing criterion.

Criterion 2: Maximize Captured Squared Length

For each observation:

$$\begin{aligned} & \left\| \mathbf{q}_p \right\|^2 \\ &= (\mathbf{u}^T \mathbf{q}_p)^2 \\ &+ \left\| \mathbf{q}_p - (\mathbf{u}^T \mathbf{q}_p) \mathbf{u} \right\|^2. \end{aligned}$$

Average over all observations:

$$\begin{aligned} & \frac{1}{P} \sum_p \left\| \mathbf{q}_p \right\|^2 \\ &= \frac{1}{P} \sum_p (\mathbf{u}^T \mathbf{q}_p)^2 \\ &+ E(\mathbf{u}). \end{aligned}$$

The left side does not depend on

$$\mathbf{u}.$$

Therefore:

$$\min E(\mathbf{u}) \quad \Leftrightarrow \quad \max \frac{1}{P} \sum_p (\mathbf{u}^T \mathbf{q}_p)^2.$$

Why This Is Variance

The projected coordinates are

$$z_p = \mathbf{u}^T \mathbf{q}_p.$$

Their mean is

$$\begin{aligned} & \frac{1}{P} \sum_p z_p \\ &= \mathbf{u}^T \left(\frac{1}{P} \sum_p \mathbf{q}_p \right) \\ &= 0. \end{aligned}$$

Therefore their variance using denominator P is

$$\text{Var}(z) = \frac{1}{P} \sum_{p=1}^P z_p^2 = \frac{1}{P} \sum_{p=1}^P (\mathbf{u}^T \mathbf{q}_p)^2.$$

So PCA chooses the direction
of greatest projected variance.

The Covariance Matrix

Define

$$C = \frac{1}{P} \sum_{p=1}^P \mathbf{q}_p \mathbf{q}_p^T.$$

If centered observations are stored in the rows of

$$\begin{aligned} & X_c: \\ & C = \frac{1}{P} X_c^T X_c. \end{aligned}$$

This is an

$$N \times N$$

symmetric matrix.

Its diagonal entries measure feature variances.

Its off-diagonal entries measure pairwise covariation.

Projected Variance as a Quadratic Form

Start with

$$\frac{1}{P} \sum_p (\mathbf{u}^T \mathbf{q}_p)^2.$$

Since each term is a scalar:

$$\begin{aligned} & (\mathbf{u}^T \mathbf{q}_p)^2 \\ &= \mathbf{u}^T \mathbf{q}_p \mathbf{q}_p^T \mathbf{u}. \end{aligned}$$

Therefore

$$\begin{aligned} & \frac{1}{P} \sum_p (\mathbf{u}^T \mathbf{q}_p)^2 \\ &= \mathbf{u}^T \left[\frac{1}{P} \sum_p \mathbf{q}_p \mathbf{q}_p^T \right] \mathbf{u}. \end{aligned}$$

Thus

$$\text{Var}(z) = \mathbf{u}^T C \mathbf{u}.$$

PCA solves:

$$\begin{aligned} & \underset{\|\mathbf{u}\|=1}{\text{maximize}} \quad \mathbf{u}^T C \mathbf{u}. \end{aligned}$$

The Unit-Length Constraint Is Essential

Without

$$\|\mathbf{u}\| = 1:$$

if one direction gives positive projected variance,

then replacing

$$\mathbf{u}$$

by

$$c\mathbf{u}$$

multiplies

$$\mathbf{u}^T C \mathbf{u}$$

by

$$c^2.$$

The objective could grow without bound.

The unit-length condition means:

choose a direction,
not an arbitrary scale.

Use a Lagrange Multiplier

Maximize

$$\mathbf{u}^T C \mathbf{u}$$

subject to

$$\mathbf{u}^T \mathbf{u} = 1.$$

Define

$$\begin{aligned} L(\mathbf{u}, \lambda) \\ = \mathbf{u}^T C \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1). \end{aligned}$$

Because

$$\begin{aligned} C &= C^T: \\ \nabla_{\mathbf{u}}(\mathbf{u}^T C \mathbf{u}) \\ &= 2C\mathbf{u}. \end{aligned}$$

Also

$$\begin{aligned} \nabla_{\mathbf{u}}(\mathbf{u}^T \mathbf{u}) \\ = 2\mathbf{u}. \end{aligned}$$

Stationarity gives

$$2C\mathbf{u} - 2\lambda\mathbf{u} = \mathbf{0}.$$

Therefore

$$C\mathbf{u} = \lambda\mathbf{u}.$$

The PCA direction must be an eigenvector of the covariance matrix.

What Does the Eigenvalue Mean?

If

$$C\mathbf{u} = \lambda\mathbf{u}$$

and

$$\|\mathbf{u}\| = 1,$$

then

$$\begin{aligned} &\mathbf{u}^T C \mathbf{u} \\ &= \mathbf{u}^T (\lambda \mathbf{u}) \\ &= \lambda \mathbf{u}^T \mathbf{u} \\ &= \lambda. \end{aligned}$$

So:

eigenvalue = variance captured along its eigenvector.

Therefore the best direction is the eigenvector
with the largest eigenvalue.

The First Principal Component Direction

Order the covariance eigenvalues:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N \geq 0.$$

Let corresponding orthonormal eigenvectors be

$$\mathbf{u}_1, \dots, \mathbf{u}_N.$$

Then

$\mathbf{u}_1$ = first principal component direction.

It solves

$$\underset{\|\mathbf{u}\|=1}{\text{maximize}} \quad \mathbf{u}^T C \mathbf{u}.$$

The variance of the first principal-component coordinates is

$$\lambda_1.$$

Why Are the Eigenvalues Nonnegative?

For any vector

$$\begin{aligned} &\mathbf{v}: \\ &\mathbf{v}^T C \mathbf{v} \\ &= \frac{1}{p} \mathbf{v}^T X_c^T X_c \mathbf{v} \\ &= \frac{1}{p} \|X_c \mathbf{v}\|^2 \\ &\geq 0. \end{aligned}$$

Therefore

$$C \succeq 0.$$

The covariance matrix is positive semidefinite.

Hence all of its eigenvalues satisfy

$$\lambda_j \geq 0.$$

Variance cannot be negative.

Return to the Four-Point Dataset

From Lesson 1.22:

p	x_1	x_2
1	0	1
2	1	2
3	2	2
4	3	3

Mean:

$$\boldsymbol{\mu} = (1.5, 2)^T.$$

Centered rows:

$(-1.5, -1),$
 $(-0.5, 0),$
 $(0.5, 0),$
 $(1.5, 1).$

Now compute the covariance matrix.

Covariance Matrix for the Example

We have

$$X_c^T X_c = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}.$$

With

$$P = 4:$$

$$C = \frac{1}{4} \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 5/4 & 3/4 \\ 3/4 & 1/2 \end{bmatrix}.$$

Trace:

$$\begin{aligned} \text{tr}(C) &= \frac{5}{4} + \frac{1}{2} \\ &= \frac{7}{4}. \end{aligned}$$

Determinant:

$$\begin{aligned} \det C &= \frac{5}{8} - \frac{9}{16} \\ &= \frac{1}{16}. \end{aligned}$$

Find the Eigenvalues

Characteristic equation:

$$\lambda^2 - \frac{7}{4}\lambda + \frac{1}{16} = 0.$$

Therefore

$$\lambda_{1,2} = \frac{7 \pm 3\sqrt{5}}{8}.$$

Numerically:

$$\lambda_1 \approx 1.7135,$$

and

$$\lambda_2 \approx 0.0365.$$

Almost all centered variation lies in one direction.

Find the Principal Direction

For

$$\lambda_1 = (7 + 3\sqrt{5})/8,$$

an eigenvector is proportional to

$$\begin{bmatrix} 1 \\ (\sqrt{5} - 1)/2 \end{bmatrix}.$$

After normalization:

$$\mathbf{u}_1 \approx \begin{bmatrix} 0.8507 \\ 0.5257 \end{bmatrix}.$$

The sign is arbitrary.

Thus

$$-\mathbf{u}_1$$

is the same principal direction.

Projection onto a line does not distinguish
between the two orientations of that line.

Compare With the Direction Chosen Last Time

Lesson 1.22 used

$$\mathbf{u} = (1, 1)^T / \sqrt{2}$$

as a plausible diagonal direction.

Its mean squared reconstruction error was
0.125.

PCA chooses

$$\mathbf{u}_1 \approx (0.8507, 0.5257)^T.$$

For the optimal one-dimensional projection:

mean squared reconstruction error

$$= \lambda_2$$
$$\approx 0.0365.$$

PCA improves the fit
because it optimizes over all unit directions.

Why Is the Reconstruction Error λ_2 ?

Total mean squared centered length:

$$\frac{1}{P} \sum_p \|\mathbf{q}_p\|^2$$
$$= \text{tr}(C)$$
$$= \lambda_1 + \lambda_2.$$

The first principal component captures
 λ_1 .

Therefore discarded mean squared length is

$$(\lambda_1 + \lambda_2) - \lambda_1$$
$$= \lambda_2.$$

More generally:
PCA reconstruction error equals
the sum of the omitted eigenvalues.

Principal-Component Coordinates

For the first component:

$$z_{p1} = \mathbf{u}_1^T \mathbf{q}_p.$$

Using

$$\mathbf{u}_1 \approx (0.8507, 0.5257)^T,$$

the four coordinates are approximately
 $(-1.8017, -0.4253, 0.4253, 1.8017)$.

These one-dimensional numbers
preserve nearly all variation in the original two-dimensional cloud.

The reconstructed points lie on the line
through the mean
in direction
 $\mathbf{u}_1$.

How Much Variance Is Explained?

Total variance:

$$\lambda_1 + \cdots + \lambda_N$$

$$= \text{tr}(C).$$

The explained-variance ratio of component j is

$$\frac{\lambda_j}{\sum_{r=1}^N \lambda_r}.$$

For our first component:

$$\begin{aligned} & \frac{\lambda_1}{\lambda_1 + \lambda_2} \\ &= \frac{1}{2} + \frac{3\sqrt{5}}{14} \\ &\approx 0.9792. \end{aligned}$$

So one dimension explains about
97.9%
of the centered variance.

The Second Principal Component

The covariance matrix is symmetric.

Therefore eigenvectors corresponding to distinct eigenvalues
are orthogonal.

The second direction is approximately

$$\mathbf{u}_2 \approx \begin{bmatrix} -0.5257 \\ 0.8507 \end{bmatrix}.$$

It captures variance

$$\lambda_2 \approx 0.0365.$$

Thus the two coordinates

$$z_{p1} = \mathbf{u}_1^T \mathbf{q}_p$$

and

$$z_{p2} = \mathbf{u}_2^T \mathbf{q}_p$$

form a rotated orthonormal coordinate system.

PCA orders those coordinates

from greatest variance to least variance.

More Than One Principal Component

Suppose we want

$$d < N$$

dimensions.

Collect the first d principal directions:

$$U_d = \begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_d \\ | & & | \end{bmatrix}.$$

Then

$$U_d^T U_d = I_d.$$

Low-dimensional coordinates:

$$\mathbf{z}_p = U_d^T \mathbf{q}_p.$$

Reconstruction:

$$\hat{\mathbf{x}}_p = \boldsymbol{\mu} + U_d U_d^T (\mathbf{x}_p - \boldsymbol{\mu}).$$

Why the Top d Eigenvectors?

For orthonormal columns

$$U,$$

captured mean squared length is

$$\frac{1}{P} \sum_p \|U^T \mathbf{q}_p\|^2.$$

This equals

$$\text{tr}(U^T C U).$$

PCA solves

$$\underset{U^T U = I_d}{\text{maximize}} \quad \text{tr}(U^T C U).$$

The maximum is obtained by choosing
the eigenvectors with the d largest eigenvalues.

The maximum captured variance is

$$\lambda_1 + \cdots + \lambda_d.$$

Reconstruction Error With d Components

Total centered variance:

$$\sum_{j=1}^N \lambda_j.$$

Variance retained by the first d components:

$$\sum_{j=1}^d \lambda_j.$$

Therefore mean squared reconstruction error is

$$\sum_{j=d+1}^N \lambda_j.$$

For the unnormalized total squared reconstruction error:
multiply by
 P .

This gives a direct quantitative rule
for how much information is lost
by dimension reduction.

Cumulative Explained Variance

The fraction of total variance retained by the first d components is

$$R_d = \frac{\lambda_1 + \dots + \lambda_d}{\lambda_1 + \dots + \lambda_N}.$$

This is the **cumulative explained-variance ratio**.

Possible dimension choice:

choose the smallest d

such that

$$R_d$$

exceeds a desired level,

such as

$$0.90$$

or

$$0.95.$$

But:

the appropriate threshold is application-dependent.

PCA Produces Uncorrelated Coordinates

Let

$$U = [\mathbf{u}_1, \dots, \mathbf{u}_N].$$

Define transformed coordinates

$$\mathbf{z} = U^T \mathbf{q}.$$

The covariance matrix of the transformed data is

$$U^T C U.$$

Because the columns of U are covariance eigenvectors:

$$U^T C U = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_N \end{bmatrix}.$$

Thus principal-component coordinates are uncorrelated.

Their variances are the eigenvalues.

PCA Is a Rotation Followed by Truncation

Using all

$$N$$

orthonormal eigenvectors:

$$U^T U = U U^T = I.$$

Then

$$\mathbf{z} = U^T \mathbf{q}$$

is just a rotation / reflection of coordinates.

No information is lost.

Dimension reduction occurs when we keep only

$$d < N$$

coordinates.

So PCA can be viewed as:

1. center;
2. rotate to covariance eigenvector coordinates;
3. order coordinates by variance;
4. discard the low-variance coordinates.

Covariance Denominator Convention

We used

$$C = \frac{1}{p} X_c^T X_c.$$

Classical sample covariance often uses

$$\frac{1}{p-1} X_c^T X_c.$$

These matrices differ by a positive scalar factor.

Therefore:

they have the same eigenvectors.

So the PCA directions are unchanged.

Only the eigenvalues

and reported variance magnitudes
are rescaled.

Always note which convention is being used.

PCA Is Sensitive to Feature Scale

Suppose one feature is measured in
dollars,

and another in
meters.

A feature with large numerical scale
can dominate the covariance matrix.

Centering removes feature means.

It does **not** equalize feature scales.

Sometimes we standardize each feature:

$$x_{pj}^{\text{std}} = \frac{x_{pj} - \mu_j}{s_j}.$$

But standardization changes the geometry
and therefore changes the PCA problem.

Use it only when equalizing feature scales
is substantively appropriate.

PCA Directions Are Not Automatically “Important Features”

A principal component is generally
a linear combination of original features.

Example:
 $z_1 = 0.8507q_1 + 0.5257q_2.$

The coefficients are sometimes called
loadings.

A large loading means a feature contributes strongly
to that principal direction.

But PCA is not supervised.

It finds high-variance structure,
not necessarily structure useful for predicting a future target.

High variance does not automatically mean
high predictive value.

PCA in Python: Center and Form the Covariance Matrix

Use the same dataset:

```
import numpy as np

X = np.array([[0., 1.],
              [1., 2.],
              [2., 2.],
              [3., 3.]])

mu = np.mean(X, axis=0)
Xc = X - mu

C = Xc.T @ Xc / len(X)

print(C)
```

Output:
[[1.25 0.75]]

```
[0.75 0.50]]
```

The formula exactly matches

$$C = (1/P)X_c^T X_c.$$

Compute Eigenvalues and Eigenvectors

Because

C

is symmetric:

```
eigenvalues, eigenvectors = np.linalg.eigh(C)
```

New Python idea:

`np`

is designed for symmetric / Hermitian matrices.

It returns eigenvalues in ascending order.

So reverse them:

```
order = np.argsort(eigenvalues)[::-1]
```

```
eigenvalues = eigenvalues[order]  
eigenvectors = eigenvectors[:, order]
```

New Python idea:

`np`

returns indices that sort an array.

```
[::-1]
```

reverses their order.

Inspect the Principal Components

```
print(eigenvalues)  
print(eigenvectors)
```

One possible output:

```
[1.71352549 0.03647451]  
  
[[-0.85065081  0.52573111]  
 [-0.52573111 -0.85065081]]
```

Your signs may differ.

For example:

$(0.8507, 0.5257)^T$

and

$(-0.8507, -0.5257)^T$

represent the same principal axis.

Eigenvector signs are arbitrary.

Project Onto the First Principal Component

Take

```
u1 = eigenvectors[:, 0]
```

Coordinates:

```
z1 = Xc @ u1
```

Reconstruction:

```
Xc_hat = np.outer(z1, u1)
```

```
X_hat = Xc_hat + mu
```

Mean squared reconstruction distance:

```
mse = np.mean(  
    np.sum((Xc - Xc_hat)**2, axis=1)  
)
```

Output:

approximately
0.03647.

This equals the omitted eigenvalue

λ_2 .

Explained Variance in Python

```
explained_ratio = (  
    eigenvalues / np.sum(eigenvalues)  
)  
  
print(explained_ratio)
```

Output:

```
[0.97915742  0.02084258]
```

The first component explains about

97.9%

of the centered variance.

The second explains about

2.1%.

Keep the First d Components

For general

d :

```
Ud = eigenvectors[:, :d]
```

```
Z = Xc @ Ud
```

```
Xc_hat = Z @ Ud.T
```

```
X_hat = Xc_hat + mu
```

Dimensions:

X_c :
 $P \times N$;

U_d :
 $N \times d$;

Z :
 $P \times d$.

Each observation is now represented by
only
 d
principal-component coordinates.

Choosing d From Explained Variance

Compute cumulative explained variance:
`cumulative = np.cumsum(explained_ratio)`

New Python idea:
`np`
computes cumulative sums.

Then one may choose the smallest
 d
for which
`cumulative[d-1]`
exceeds a desired threshold.

Example:
retain at least
95%
of variance.

But the threshold should be justified
by the application,
not chosen mechanically.

What PCA Has Accomplished

PCA found a coordinate system
determined entirely by the data.

The first coordinate:
greatest variance.

The second:
greatest remaining variance
subject to orthogonality.

And so on.

Keeping only the first few coordinates
gives the best orthogonal linear reconstruction
of that dimension
in squared-error terms.

Next:
connect this eigenvector view
directly to the singular value decomposition.

What did we learn?

- ◆ PCA chooses projection directions from centered data rather than specifying them in advance.
- ◆ The first principal component maximizes projected variance $\mathbf{u}^T C \mathbf{u}$ over unit vectors.
- ◆ Lagrange multipliers show that principal directions are eigenvectors of the covariance matrix.
- ◆ The eigenvalue associated with a principal direction is the variance captured along that direction.
- ◆ The first d principal components are the eigenvectors associated with the d largest covariance eigenvalues.
- ◆ Mean squared reconstruction error after retaining d components equals the sum of the omitted eigenvalues.

Next

- PCA can be described through covariance eigenvectors.
- The same principal directions also appear directly in the singular value decomposition of the centered data matrix.
- This gives a numerically important and conceptually useful alternative view.
- **1.24 - PCA, SVD, and the Linear Autoencoder**