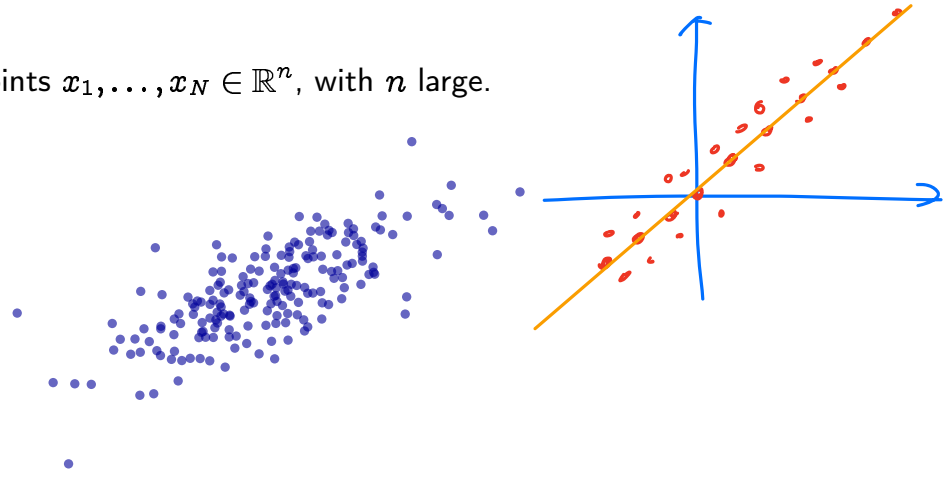
Principal Component Analysis

Stephen Boyd and Sanjay Lall

EE263
Stanford University

The problem

We are given a cloud of N data points $x_1, \ldots, x_N \in \mathbb{R}^n$, with n large.

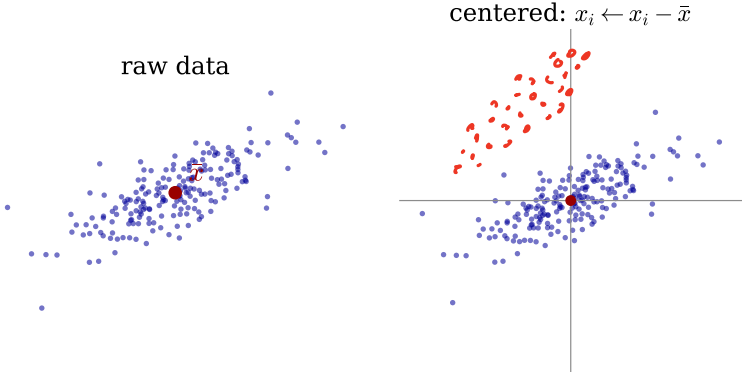


We want to *summarize* each point with just a few numbers, throwing away as little as possible:

- is there a single direction along which the data mostly lives?
- a low-dimensional subspace the data stays close to?

These turn out to be *the same question*, and its answer is an eigenvalue problem.

First step: center the data



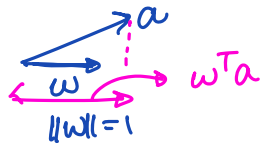
PCA is about *variation about the center*, so we subtract the mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad x_i \leftarrow x_i - \bar{x}.$$

From now on the x_i are *centered*, and we collect them as the rows of the *data matrix*

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} \in \mathbb{R}^{N\times n}.$$

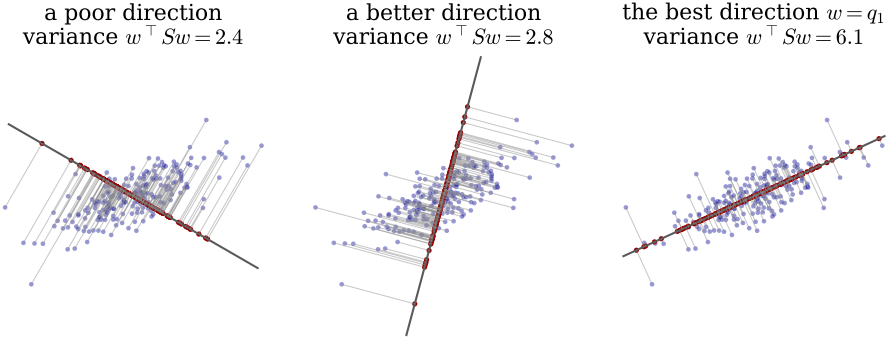
How spread out is the data along a direction?



Fix a *unit* direction w ($||w|| = 1$). The coordinate of point i along w is $w^\top x_i$, and the spread of these coordinates is their variance

$\frac{1}{N-1}\sum (y_i - \bar{y})^2$

$$\sigma_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}(w^\top x_i)^2.$$



goal: find the direction along which the projected data is most spread out (it retains the most information). We fix $||w|| = 1$ because only the *direction* matters: scaling w just rescales the coordinate.

The spread is a quadratic form

$$\sigma_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}(w^\top x_i)^2 = \frac{1}{N-1}\sum_{i=1}^{N} w^\top x_i x_i^\top w = w^\top \underbrace{\left(\frac{1}{N-1}\sum_i x_i x_i^\top\right)}_{S} w$$

$$\sigma_w^2 = w^\top S w, \qquad S = \frac{1}{N-1}X^\top X \quad \text{(the } \textit{covariance matrix}\text{)}$$

- $S = S^\top$ and $S \geq 0$ (it is $\frac{1}{N-1}X^\top X$)
- S_{jj} is the variance of coordinate j; S_{jk} its covariance with coordinate k

$x^\top A x \leq \lambda_1 x^\top x$ for $\|x\| = 1$

So "find the most spread-out direction" becomes:

maximize $w^\top S w$ subject to $\|w\| = 1$.

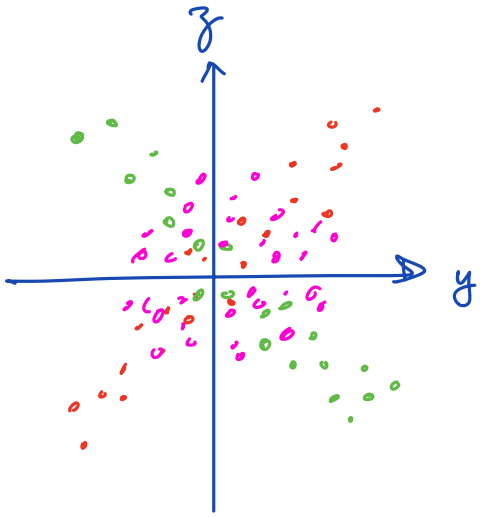
$z_1, \cdots, z_N$

$y_1, \cdots, y_N$

$t_1, \cdots$

$u_1, \cdots$

- mean: $\bar{z} = \mathbb{E}[z] \rightarrow \frac{1}{N}(z_1 + \cdots + z_N)$
- variance: $\sigma_z^2 = var(z) \rightarrow \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2$
- standard deviation: $\sigma_z = \sqrt{var(z)}$
- covariance of y and z: $\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})$



$\bar{y} = \bar{z} = 0$

$Cov(y,z) = \frac{1}{N}\sum_{i=0}^{N} y_i z_i$

$\simeq 10$

$\simeq 0$

$\simeq -10$

$cov(z,z) = var(z)$

$X^{(1)}: \; X_1^{(1)} \; \cdots \; X_N^{(1)}$

$X^{(2)} \quad \vdots$

$\vdots$

$X^{(n)}: \; X_1^{(n)} \; \cdots \; X_N^{(n)}$

$\rightarrow cov(X^{(1)} \cdots X^{(n)}) =$

$Cov(X^{(i)}, X^{(j)})$

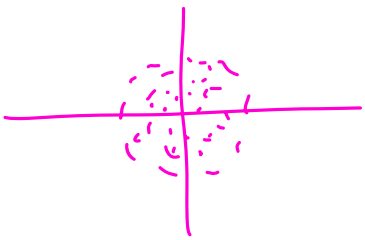
ij

ii

$n \times n$

$Var(X^{(i)})$

$\frac{1}{N} X^T X$



$Cov(y,z) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \rightarrow \sigma^2 I$

$\rightarrow \begin{bmatrix} 3\sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \rightarrow \lambda = 3\sigma^2, \sigma$

We have already solved this problem

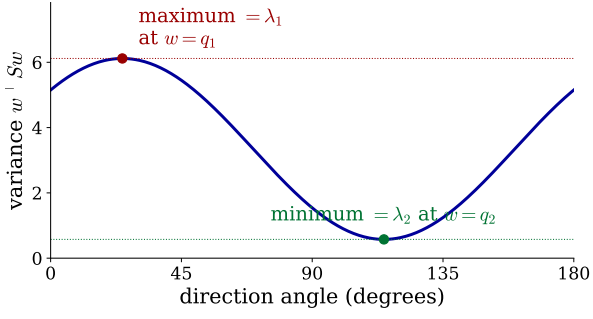
Recall (symmetric matrices lecture)

For symmetric $S = Q\Lambda Q^{\mathsf{T}}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and orthonormal eigenvectors $q_1, \ldots, q_n$,

$$\lambda_n w^{\mathsf{T}} w \;\leq\; w^{\mathsf{T}} S w \;\leq\; \lambda_1 w^{\mathsf{T}} w$$

with equality $w^{\mathsf{T}} S w = \lambda_1$ at $w = q_1$ (and $= \lambda_n$ at $w = q_n$).

So over all unit vectors, the maximum spread is λ_1, achieved along q_1:

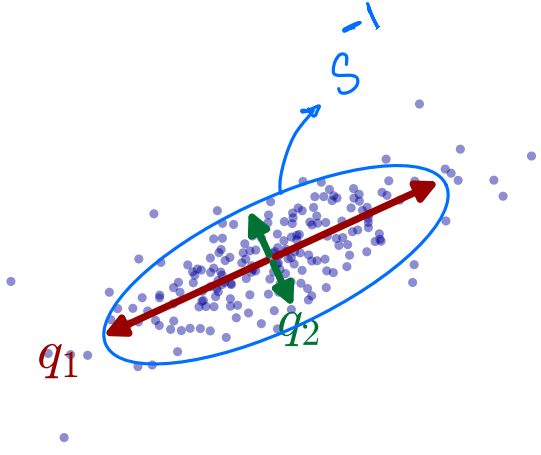


The first principal component

$\sigma^2 = w^T S w$

$w = q \rightarrow q^T S q = \lambda \|q\|^2 = \lambda$

$Sq = \lambda q$, $\|q\|^2 = 1$



The direction of maximum spread is the top eigenvector of the covariance:

$$q_1 = \text{first principal component}, \quad \sigma^2_{q_1} = \lambda_1$$

The next component is the most spread-out direction *orthogonal* to q_1: that is q_2, with variance λ_2; and so on.

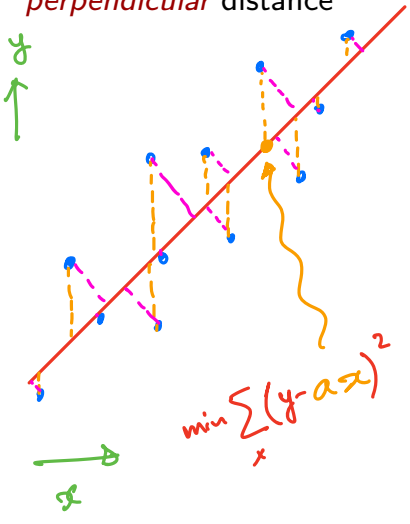
The principal components $q_1, \ldots, q_n$ are the orthonormal eigenvectors of S, with variances $\lambda_1 \geq \cdots \geq \lambda_n$.

Total variance $= \sum_i \sigma^2_{q_i} = \sum_i \lambda_i = \operatorname{tr} S$: the eigenvalues split the total spread among orthogonal directions.

A second, different-looking goal

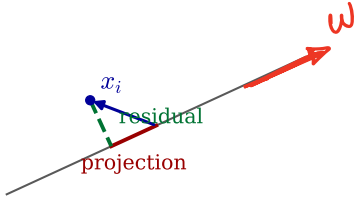
LS: $\min \sum_{i=1}^{m} \| y_i - a^T x_i \|$

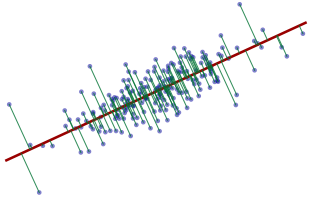
Instead of *most spread*, ask for the line (through the origin) the data is *closest* to: minimize the total squared *perpendicular* distance



$$\min_{||w||=1} \sum_{i=1}^{N} ||x_i - (w^T x_i)\, w||^2.$$

$\|x_i\|^2 = \|\text{proj}\|^2 + \|\text{resid}\|^2$ all residuals to the line





$(w^T x_i)\, w$ is the projection of x_i onto the line; the leftover $x_i - (w^T x_i)w$ is the perpendicular residual.

The two goals are the same

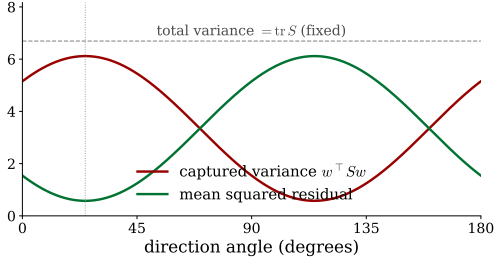
The projection and the residual are orthogonal, so by Pythagoras

$$||x_i||^2 = \underbrace{(w^\top x_i)^2}_{\text{captured}} + \underbrace{||x_i - (w^\top x_i)w||^2}_{\text{residual}}.$$

Summing over i,

$$\underbrace{\sum_i ||x_i||^2}_{\text{fixed total, } = (N-1)\,\mathrm{tr}\,S} = \underbrace{\sum_i (w^\top x_i)^2}_{(N-1)\,w^\top S w} + \sum_i ||\text{residual}_i||^2.$$

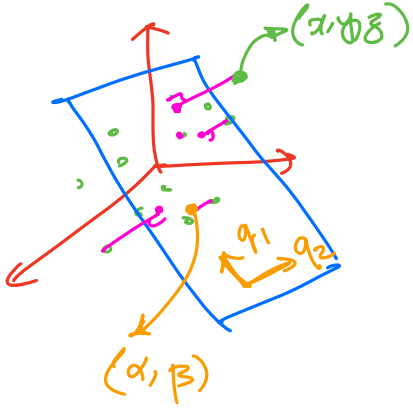
The total is fixed, so *maximizing captured variance* = *minimizing residual*



So the max-variance direction q_1 is also the closest line: *one answer, two questions*.

The best-fit k-dimensional subspace

The same argument in k directions at once: the k-dimensional subspace the data is closest to is spanned by the top k principal components $q_1, \ldots, q_k$ (maximizing captured variance $\sum_{j \le k} q_j^\top S q_j$ over orthonormal q_j gives the top k eigenvectors).



Each point gets k *reduced coordinates*

$$z_i = \begin{bmatrix} q_1^\top x_i \\ \vdots \\ q_k^\top x_i \end{bmatrix} = W_k^\top x_i, \qquad W_k = \begin{bmatrix} q_1 & \cdots & q_k \end{bmatrix}$$

and is reconstructed as $\bar{x} + W_k z_i$ (the best rank-k description).

The leftover (lost) variance is exactly $\sum_{i>k} \lambda_i$.

(The data has at most $\mathbf{min}(n, N-1)$ nonzero components, so PCA is meaningful only with many points: a single point is described exactly by one component.)

PCA is the SVD of the centered data

Recall (SVD lecture)

If $X = U\Sigma V^\top$ then $X^\top X = V\Sigma^2 V^\top$, so the eigenvectors of $X^\top X$ are the right singular vectors v_i, with eigenvalues σ_i^2.

Since $S = \frac{1}{N-1} X^\top X$, this says

$$q_i = v_i \quad \text{(principal components)}, \qquad \lambda_i = \frac{\sigma_i^2}{N-1}$$

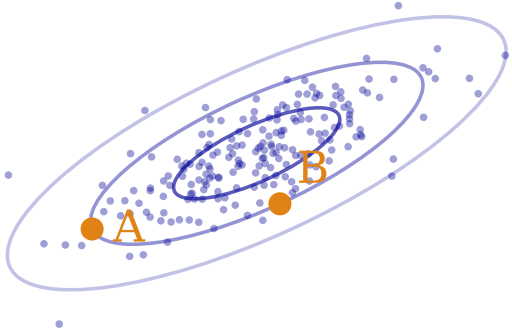
- *to do PCA, take the SVD of the centered data matrix X*
- the reduced coordinates are the columns of $\Sigma V^\top$ (that is, $U\Sigma$ up to scaling)

Recall (SVD lecture: low-rank approximation)

Keeping the top k components is *exactly* the best rank-k approximation $X_k = \sum_{i \le k} \sigma_i u_i v_i^\top$ of the data matrix (Eckart–Young). So *PCA is low-rank approximation, applied to the centered data matrix X*.

The covariance ellipsoid: the idea

To judge how far a point is from the center, count *standard deviations*, not raw distance: a point is unusual only relative to how much the data varies in that direction.



The *covariance ellipsoids* are the shells of points equally many standard deviations from the center (nested ellipsoids are the cloud's contour lines).

A and B lie on the same shell: both 2 *standard deviations* out. Yet A is physically far (a wide direction) and B near (a narrow one): the same "surprise" reaches farther where the cloud is spread out.

How many standard deviations? (why S^{-1})

Along a principal direction q_i the data has standard deviation $\sqrt{\lambda_i}$, so a point $t\,q_i$ is $t/\sqrt{\lambda_i}$ standard deviations out.

For a general x, *whiten* the cloud: with $S^{-1/2} = Q\Lambda^{-1/2}Q^{\mathsf{T}}$, set $z = S^{-1/2}x$. Its sample covariance is

$$\tfrac{1}{N-1}\textstyle\sum_i z_i z_i^{\mathsf{T}} = S^{-1/2}\Big(\tfrac{1}{N-1}\textstyle\sum_i x_i x_i^{\mathsf{T}}\Big)S^{-1/2} = S^{-1/2}\,S\,S^{-1/2} = I,$$

so in z *every* direction has unit standard deviation. The number of standard deviations of x is then the ordinary length

$$d(x) = ||z||, \qquad d(x)^2 = z^{\mathsf{T}}z = x^{\mathsf{T}}S^{-1}x.$$

The *inverse* appears because whitening divides by the variances; the ellipsoids $\{x \mid x^{\mathsf{T}}S^{-1}x \le c^2\}$ are exactly "within c standard deviations."

The covariance ellipsoid: geometry

$S \to \lambda_i$

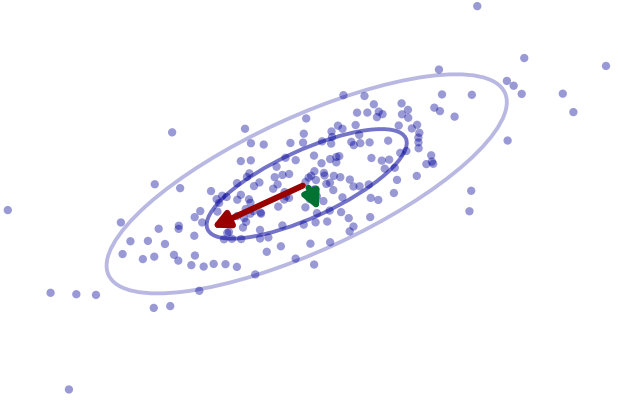
$S^{-1} \to 1/\lambda_i$

$\{x : x^\top S^{-1} x \le c\}$

$\frac{1}{\sqrt{\lambda_i(S^{-1})}} = \sqrt{\lambda_i(S)}$

$\mathcal{E} = \{x \mid x^T A x \le 1\}$

$1/\sqrt{\lambda_i}$

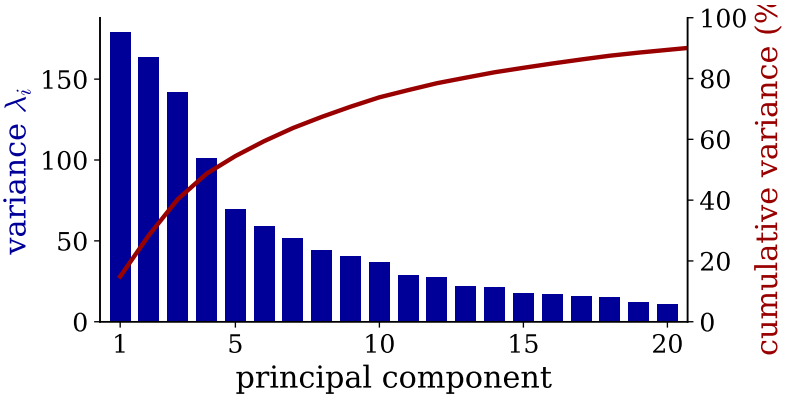


Recall (ellipsoids lecture)
$\{x \mid x^\top M x \le 1\}$ has semi-axes $\mu_i^{-1/2} q_i$, from the eigen-pairs (μ_i, q_i) of M.

Here $M = S^{-1}$: same eigenvectors q_i as S, eigenvalues $1/\lambda_i$. So the semi-axes are $\sqrt{\lambda_i}\, q_i$, *longest along the large-variance directions*.

The ellipsoid lines up with the cloud, and *PCA finds its axes*.

How many components to keep?



Keep the top k components; the fraction of variance retained is

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_n}.$$

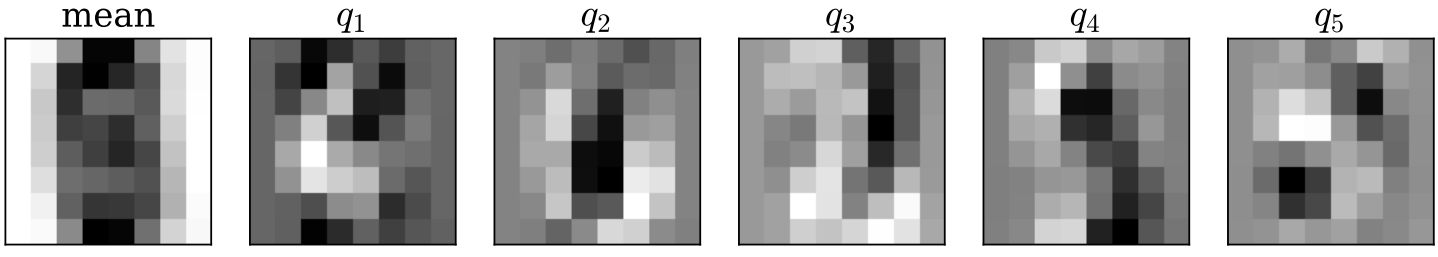
Plot the eigenvalues (the *scree plot*) and look for the elbow: a few components often capture almost all the variance.

(Here: the handwritten-digit data of the next slide.)

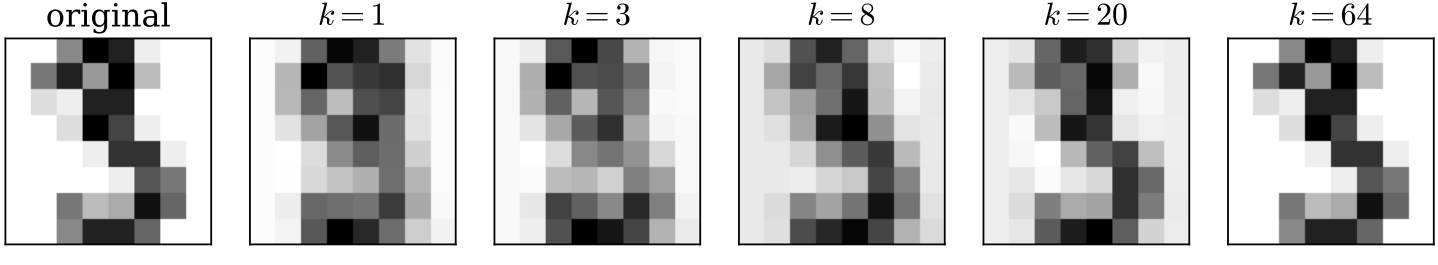
Example: handwritten digits

$X \in \mathbb{R}^{1797 \times 64}$

We have a *dataset* of 1797 handwritten digits, each an 8×8 image (a point in $\mathbb{R}^{64}$). PCA of the whole collection gives a mean and eigenvectors ("eigen-digits") that are themselves images:

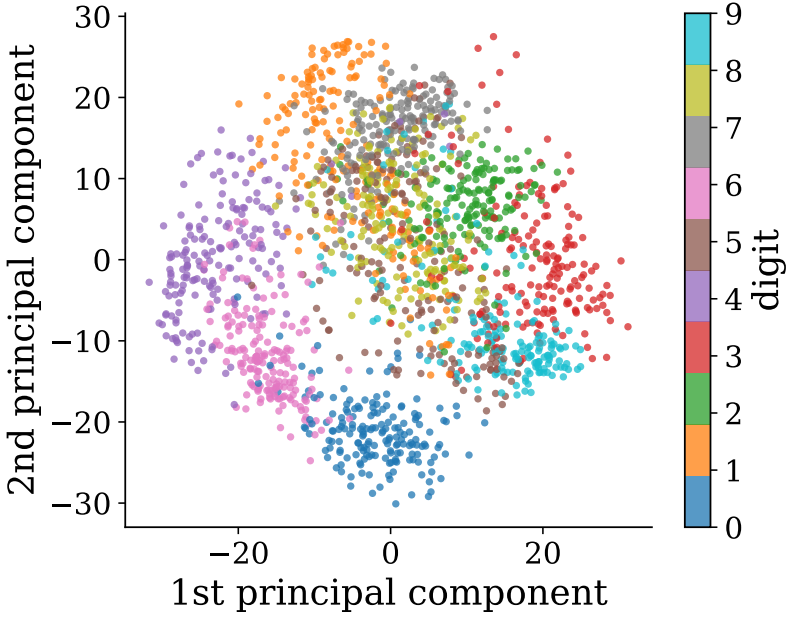


Reconstructing one digit from its top-k coordinates in this *shared* basis, $\bar{x} + W_k z$:



A handful of the 64 components already give a recognizable digit.

Example: digits in two dimensions



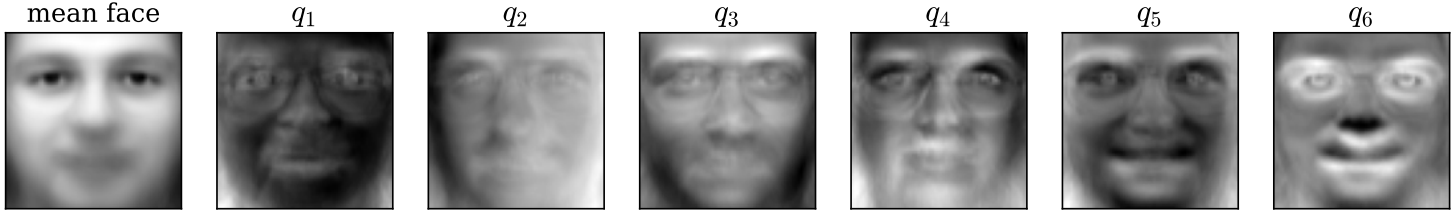
Project all 1797 digits from $\mathbb{R}^{64}$ onto their first two principal components $z_i = W_2^\top x_i$.

Two numbers per image (down from 64), yet the digits already separate into clusters.

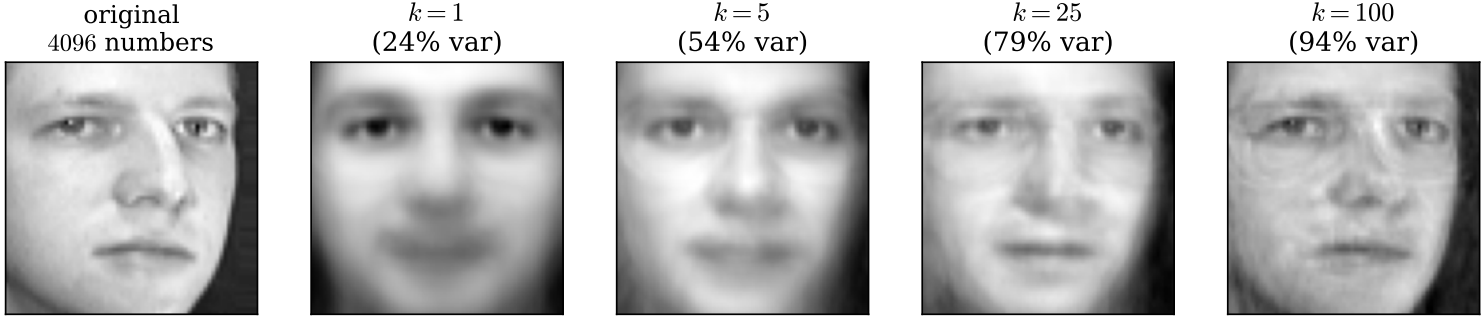
This is the workhorse use of PCA: *visualization and compression* of high-dimensional data.

A larger example: eigenfaces

A *dataset* of 400 face images, each 64×64 (a point in $\mathbb{R}^{4096}$). PCA gives a mean face and eigenvectors ("eigenfaces"):



Reconstructing one face from its top-k coordinates in the shared basis:



About 25 of the 4096 components already give a recognizable face.

Summary

PCA is one object seen from several sides:

- the *directions of maximum variance* of the data
- the *best-fit subspace* (minimum perpendicular distance)
- the *eigenvectors of the covariance* S, with variances the eigenvalues
- the *right singular vectors* of the centered data X (PCA $=$ SVD of X)
- the *axes of the covariance ellipsoid*

Center the data, take the SVD; the top singular vectors are what matters.