PCA

CS 446

2020-12-27 (a50ee91)

Plan for today

- **Un**supervised learning?
- PCA.
- PCA example.
- PCA proofs.

Supervised learning

So far, we've done supervised learning:
Given $((\boldsymbol{x}_i, \boldsymbol{y}_i))_{i=1}^n$, find predictor f with $f(\boldsymbol{x}_i) \approx \boldsymbol{y}_i$.

Linear regression, deep networks, k-nn, decision trees, ...

Supervised learning

So far, we've done supervised learning:
Given $((\boldsymbol{x}_i, \boldsymbol{y}_i))_{i=1}^n$, find predictor f with $f(\boldsymbol{x}_i) \approx \boldsymbol{y}_i$.

Linear regression, deep networks, k-nn, decision trees, . . .

Most methods used (regularized) ERM:
minimize $\widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i)$, hope $\mathcal{R}$ is small.

least squares, logistic regression, deep networks, SVM, perceptron, . . .

Unsupervised learning

Now we only receive $(\boldsymbol{x}_i)_{i=1}^n$, and the goal is. . . ?

Unsupervised learning

Now we only receive $(\boldsymbol{x}_i)_{i=1}^n$, and the goal is. . . ?

- Encoding data in some compact representation (and decoding this).
- Data analysis;
  recovering "hidden structure" in data
  (e.g., recovering cliques or clusters).
- Features for supervised learning.
- . . . ?

Unsupervised learning

Now we only receive $(\boldsymbol{x}_i)_{i=1}^n$, and the goal is. . . ?

- Encoding data in some compact representation
  (and decoding this).
- Data analysis;
  recovering "hidden structure" in data
  (e.g., recovering cliques or clusters).
- Features for supervised learning.
- . . . ?

The task is less clear-cut, and lacks a single accepted formalization.

(Side note: can *still* use the "pytorch meta-algorithm".)

Principal Component Analysis (PCA) motivation

Principal Component Analysis (PCA) motivation

Let's formulate a simplistic *linear* unsupervised method.

- Encoding (and decoding) data in some compact representation.
  Let's linearly map data in $\mathbb{R}^d$ to $\mathbb{R}^k$ and back.
- Data analysis;
  recovering "hidden structure" in data.
  Let's find if data mostly lies on a low-dimensional subspace.
- Features for supervised learning.
  Let's feed the $\mathbb{R}^k$-dimensional encoding to supervised methods.

Let $\boldsymbol{M} \in \mathbb{R}^{n \times d}$ be given. $((s_i, \boldsymbol{u}_i, \boldsymbol{v}_i))_{i=1}^r$ is an SVD of $\boldsymbol{M}$ if:

- $\boldsymbol{M}$ has rank r;
- $s_1 \geq s_2 \cdots \geq s_r > 0$;
- $(\boldsymbol{u}_i)_{i=1}^r$ are orthonormal (orthogonal and unit length), and span the column space of $\boldsymbol{M}$;
- $(\boldsymbol{v}_i)_{i=1}^r$ are orthonormal, and span the row space of $\boldsymbol{M}$.
- $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}$.

Let's also collect these into matrices: $\boldsymbol{S} := \operatorname{diag}(s_1, \ldots, s_r) \in \mathbb{R}^{r \times r}$, and

$$\boldsymbol{U} := \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r \\ \downarrow & & \downarrow \end{bmatrix} \in \mathbb{R}^{n \times r}, \qquad \boldsymbol{V} := \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_r \\ \downarrow & & \downarrow \end{bmatrix} \in \mathbb{R}^{d \times r},$$

whereby $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$. Additionally define best rank-k approximations $\boldsymbol{S}_k := \operatorname{diag}(s_1, \ldots, s_k) \in \mathbb{R}^{k \times k}$,

$$\boldsymbol{U}_k := \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_k \\ \downarrow & & \downarrow \end{bmatrix} \in \mathbb{R}^{n \times k}, \qquad \boldsymbol{V}_k := \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_k \\ \downarrow & & \downarrow \end{bmatrix} \in \mathbb{R}^{d \times k},$$

and $\boldsymbol{M}_k := \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k^\mathsf{T}$. (Side note: not unique!)

PCA (Principal component analysis)

Input: Data as rows of $\mathbb{R}^{n \times d} \ni \boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$, integer k.
Output: Encoder $\boldsymbol{V}_k$, decoder $\boldsymbol{V}_k^{\mathsf{T}}$, encoded data $\boldsymbol{X}\boldsymbol{V}_k = \boldsymbol{U}_k\boldsymbol{S}_k$.

PCA (Principal component analysis)

Input: Data as rows of $\mathbb{R}^{n\times d} \ni \boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$, integer k.
Output: Encoder $\boldsymbol{V}_k$, decoder $\boldsymbol{V}_k^\mathsf{T}$, encoded data $\boldsymbol{X}\boldsymbol{V}_k = \boldsymbol{U}_k\boldsymbol{S}_k$.

The goal in unsupervised learning is unclear.
We'll try to define this as "best encoding/decoding in Frobenius sense":

$$\min_{\substack{\boldsymbol{D}\in\mathbb{R}^{k\times d}\\ \boldsymbol{E}\in\mathbb{R}^{d\times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_\mathsf{F}^2 = \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \left\|\boldsymbol{X} - \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \|\boldsymbol{X} - \boldsymbol{X}_k\|_\mathsf{F}^2 .$$

PCA (Principal component analysis)

Input: Data as rows of $\mathbb{R}^{n\times d} \ni \boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$, integer k.
Output: Encoder $\boldsymbol{V}_k$, decoder $\boldsymbol{V}_k^\mathsf{T}$, encoded data $\boldsymbol{X}\boldsymbol{V}_k = \boldsymbol{U}_k\boldsymbol{S}_k$.

The goal in unsupervised learning is unclear.
We'll try to define this as "best encoding/decoding in Frobenius sense":

$$\min_{\substack{\boldsymbol{D}\in\mathbb{R}^{k\times d}\\ \boldsymbol{E}\in\mathbb{R}^{d\times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_\mathsf{F}^2 = \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \left\|\boldsymbol{X} - \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \|\boldsymbol{X} - \boldsymbol{X}_k\|_\mathsf{F}^2 .$$

Note $\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}$ performs orthogonal projection onto subspace spanned by $\boldsymbol{V}_k$;
thus we are finding "best k-dimensional projection of the data".

PCA properties

Let's try to capture "best low-rank approximation" and "best linear encoder/decoder".

PCA properties

Let's try to capture "best low-rank approximation" and "best linear encoder/decoder".

Theorem. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$ and integer $k \leq r$ be given.

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{k \times d} \\ \boldsymbol{E} \in \mathbb{R}^{d \times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_\mathsf{F}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2$$

$$= \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \sum_{i=k+1}^{r} s_i^2.$$

Additionally,

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2$$

$$= \|\boldsymbol{X}\|_\mathsf{F}^2 - \|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \sum_{i=1}^{k} s_i^2.$$

Remark 1. SVD is not unique, but $\sum_{i=1}^{r} s_i^2$ is identical across SVD choices.

PCA properties

Let's try to capture "best low-rank approximation" and "best linear encoder/decoder".

Theorem. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$ and integer $k \leq r$ be given.

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{k \times d} \\ \boldsymbol{E} \in \mathbb{R}^{d \times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_\mathsf{F}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2$$

$$= \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \sum_{i=k+1}^{r} s_i^2.$$

Additionally,

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2$$

$$= \|\boldsymbol{X}\|_\mathsf{F}^2 - \|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \sum_{i=1}^{k} s_i^2.$$

Remark 1. SVD is not unique, but $\sum_{i=1}^{r} s_i^2$ is identical across SVD choices.

Remark 2. As written, this is not a convex optimization problem!

PCA properties

Let's try to capture "best low-rank approximation" and "best linear encoder/decoder".

Theorem. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$ and integer $k \leq r$ be given.

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{k \times d} \\ \boldsymbol{E} \in \mathbb{R}^{d \times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_\mathsf{F}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2$$

$$= \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \sum_{i=k+1}^{r} s_i^2.$$

Additionally,

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2$$

$$= \|\boldsymbol{X}\|_\mathsf{F}^2 - \|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \sum_{i=1}^{k} s_i^2.$$

Remark 1. SVD is not unique, but $\sum_{i=1}^{r} s_i^2$ is identical across SVD choices.

Remark 2. As written, this is not a convex optimization problem!

Remark 3. The second form is interesting...

PCA properties

Let's try to capture "best low-rank approximation" and "best linear encoder/decoder".

Theorem. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$ and integer $k \leq r$ be given.

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{k \times d} \\ \boldsymbol{E} \in \mathbb{R}^{d \times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_\mathsf{F}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2$$

$$= \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 = \sum_{i=k+1}^{r} s_i^2.$$

Additionally,

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2$$

$$= \|\boldsymbol{X}\|_\mathsf{F}^2 - \|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \sum_{i=1}^{k} s_i^2.$$

Remark 1. SVD is not unique, but $\sum_{i=1}^{r} s_i^2$ is identical across SVD choices.

Remark 2. As written, this is not a convex optimization problem!

Remark 3. The second form is interesting...

Remark 4. Easy to show "best rank-k approximation" from here.

Centered PCA

Some treatments replace $\boldsymbol{X}$ with $\boldsymbol{X} - \mathbf{1}\boldsymbol{\mu}^\mathsf{T}$,
with mean $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1} \boldsymbol{x}_i$.

Centered PCA

Some treatments replace $\boldsymbol{X}$ with $\boldsymbol{X} - \mathbf{1}\boldsymbol{\mu}^\mathsf{T}$,
with mean $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1} \boldsymbol{x}_i$.

$\frac{1}{n}\boldsymbol{X}^\mathsf{T}\boldsymbol{X} \in \mathbb{R}^{d\times d}$ is data covariance;

Centered PCA

Some treatments replace $\boldsymbol{X}$ with $\boldsymbol{X} - \mathbf{1}\boldsymbol{\mu}^\mathsf{T}$,
with mean $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1} \boldsymbol{x}_i$.

$\frac{1}{n}\boldsymbol{X}^\mathsf{T}\boldsymbol{X} \in \mathbb{R}^{d\times d}$ is data covariance;
$\frac{1}{n}(\boldsymbol{X}\boldsymbol{D})^\mathsf{T}(\boldsymbol{X}\boldsymbol{D})$ is data covariance after projection;

Centered PCA

Some treatments replace $\boldsymbol{X}$ with $\boldsymbol{X} - \mathbf{1}\boldsymbol{\mu}^\mathsf{T}$,
with mean $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1} \boldsymbol{x}_i$.

$\frac{1}{n}\boldsymbol{X}^\mathsf{T}\boldsymbol{X} \in \mathbb{R}^{d\times d}$ is data covariance;
$\frac{1}{n}(\boldsymbol{X}\boldsymbol{D})^\mathsf{T}(\boldsymbol{X}\boldsymbol{D})$ is data covariance after projection;
lastly

$$\frac{1}{n}\|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2 = \frac{1}{n}\operatorname{tr}\Big((\boldsymbol{X}\boldsymbol{D})^\mathsf{T}(\boldsymbol{X}\boldsymbol{D})\Big) = \frac{1}{n}\sum_{i=1}^{k}(\boldsymbol{X}\boldsymbol{D}\boldsymbol{e}_i)^\mathsf{T}(\boldsymbol{X}\boldsymbol{D}\boldsymbol{e}_i).$$

Since PCA is also solving $\max_{\substack{\boldsymbol{D}\in\mathbb{R}^{d\times k}\\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D}=\boldsymbol{I}}}\|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2$, therefore centered PCA is maximizing the resulting per-coordinate variances!

PCA example

- Image data; e.g., “eigenfaces”.

PCA example

- Image data; e.g., "eigenfaces".
  Weirdness: negative faces?
  This motivates *non-negative matrix factorization*.

PCA example

- Image data; e.g., "eigenfaces".
  Weirdness: negative faces?
  This motivates *non-negative matrix factorization*.
- *LSI (Latent Semantic Indexing)*:
  collect many documents into $\boldsymbol{X} \in \mathbb{R}^{n \times d}$,
  where $\boldsymbol{x}_i$ is a normalize bag-of-words vector (plus nonlinear mappings).
  Can interpret new representation as weighting over "topics".

Application: digit data

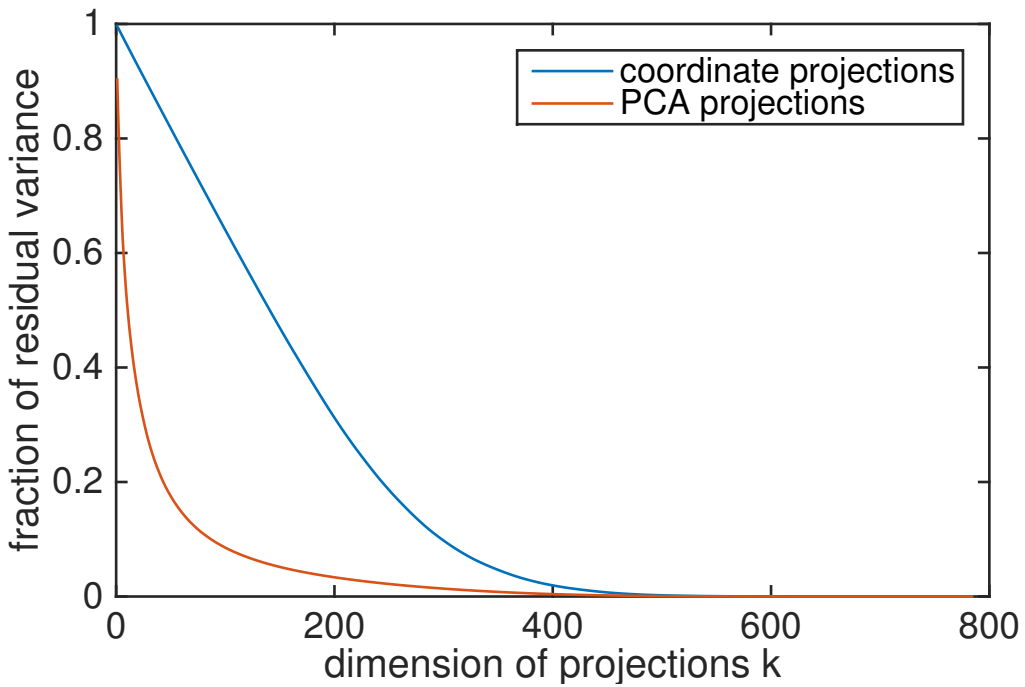
Data $(\boldsymbol{x}_i)_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^{784}$.

- Residual variance left by rank-k PCA projection:

$$1 - \frac{\sum_{j=1}^k \text{variance in direction } \boldsymbol{v}_j}{\text{total variance}} = 1 - \frac{\|\boldsymbol{X}\boldsymbol{V}_k\|_{\mathsf{F}}^2}{\|\boldsymbol{X}\|_{\mathsf{F}}^2}.$$

- Residual variance left by best k *coordinate* projections:

$$1 - \frac{\sum_{j=1}^k \text{variance in direction } \mathbf{e}_j}{\text{total variance}} = 1 - \frac{\sum_{j=1}^k (\boldsymbol{X}\mathbf{e}_j)^{\mathsf{T}}(\boldsymbol{X}\mathbf{e}_j)}{\|\boldsymbol{X}\|_{\mathsf{F}}^2}.$$

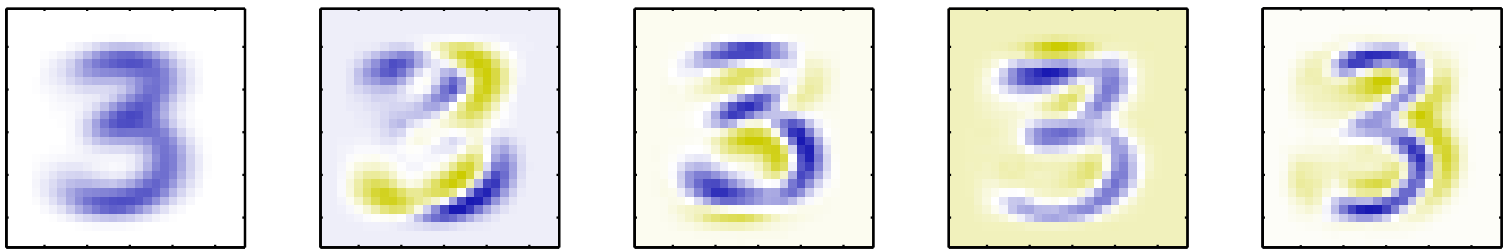


Application: digit data

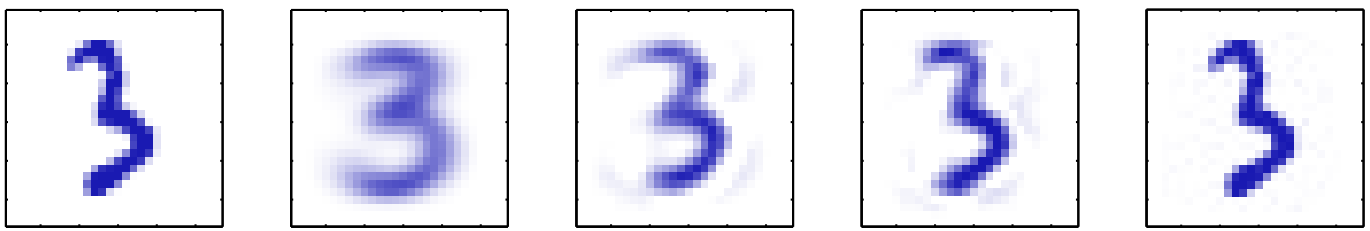
16×16 pixel images of handwritten 3s (as vectors in $\mathbb{R}^{256}$)

Mean μ and right singular vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4$

Mean | $\lambda_1 = 3.4 \cdot 10^5$ | $\lambda_2 = 2.8 \cdot 10^5$ | $\lambda_3 = 2.4 \cdot 10^5$ | $\lambda_4 = 1.6 \cdot 10^5$



Reconstructions:



x | $k = 1$ | $k = 10$ | $k = 50$ | $k = 200$

Only have to store k numbers per image,
along with the mean $\boldsymbol{\mu}$ and k eigenvectors ($256(k+1)$ numbers).

Proof of PCA theorem.

Fact. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $k \leq r$ be given.

$$\min_{\substack{\boldsymbol{M} \in \mathbb{R}^{d \times d} \\ \operatorname{rank}(\boldsymbol{M}) = k}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{M}\|_{\mathsf{F}}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{k \times d} \\ \boldsymbol{E} \in \mathbb{R}^{d \times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_{\mathsf{F}}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}}\right\|_{\mathsf{F}}^2.$$

Proof of PCA theorem.

Fact. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $k \leq r$ be given.

$$\min_{\substack{\boldsymbol{M} \in \mathbb{R}^{d \times d} \\ \operatorname{rank}(\boldsymbol{M}) = k}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{M}\|_{\mathsf{F}}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{k \times d} \\ \boldsymbol{E} \in \mathbb{R}^{d \times k}}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{E}\boldsymbol{D}\|_{\mathsf{F}}^2 = \min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}}\right\|_{\mathsf{F}}^2.$$

Proof. Since

$$\begin{aligned}\left\{\boldsymbol{M} \in \mathbb{R}^{d \times d} : \operatorname{rank}(\boldsymbol{M}) = k\right\} &\supseteq \left\{\boldsymbol{D}\boldsymbol{E} : \boldsymbol{D} \in \mathbb{R}^{k \times d}, \boldsymbol{E} \in \mathbb{R}^{d \times k}\right\} \\ &\supseteq \left\{\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}} : \boldsymbol{D} \in \mathbb{R}^{d \times k}, \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D} = \boldsymbol{I}\right\},\end{aligned}$$

it suffices to show

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D} = \boldsymbol{I}}} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}}\right\|_{\mathsf{F}}^2 \leq \min_{\substack{\boldsymbol{M} \in \mathbb{R}^{d \times d} \\ \operatorname{rank}(\boldsymbol{M}) = k}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{M}\|_{\mathsf{F}}^2.$$

Proof (continued). For any $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T} \in \mathbb{R}^{d\times d}$ with $\operatorname{rank}(\boldsymbol{M}) \leq k$ (whereby $\boldsymbol{M} = \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^\mathsf{T}$),

$$\begin{aligned}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{M}\|_\mathsf{F}^2 &= \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T} + \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T} - \boldsymbol{X}\boldsymbol{M}\right\|_\mathsf{F}^2 \\ &= \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 + 2\operatorname{tr}\left(\left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right)^\mathsf{T}\left(\boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T} - \boldsymbol{X}\boldsymbol{M}\right)\right) \\ &\quad + \left\|\boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T} - \boldsymbol{X}\boldsymbol{M}\right\|_\mathsf{F}^2.\end{aligned}$$

We'll show the middle (trace) term is 0, and therefore

$$\|\boldsymbol{X} - \boldsymbol{M}\|_\mathsf{F}^2 = \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2 + \left\|\boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T} - \boldsymbol{X}\boldsymbol{M}\right\|_\mathsf{F}^2 \geq \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\right\|_\mathsf{F}^2.$$

Proof (continued). Note

$$\begin{aligned}
&\operatorname{tr}\left(\left(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}\right)^{\mathsf{T}}\left(\boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}-\boldsymbol{X}\boldsymbol{M}\right)\right)\\
&=\operatorname{tr}\left(\left(\boldsymbol{I}-\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}\right)^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\left(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathsf{T}}\right)\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}\right)\\
&=\operatorname{tr}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\left(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^{\mathsf{T}}\right)\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}\left(\boldsymbol{I}-\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}\right)^{\mathsf{T}}\right),
\end{aligned}$$

and

$$\begin{aligned}
(\boldsymbol{I}-\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}})^{\mathsf{T}}\left(\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}\right) &= \left(\sum_{i=1}^{d}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathsf{T}}-\sum_{i=1}^{k}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathsf{T}}\right)^{\mathsf{T}}\sum_{j=1}^{k}\boldsymbol{v}_j\boldsymbol{v}_j^{\mathsf{T}}\\
&= \left(\sum_{i=k+1}^{d}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathsf{T}}\right)\sum_{j=1}^{k}\boldsymbol{v}_j\boldsymbol{v}_j^{\mathsf{T}}=0.
\end{aligned}$$

Therefore

$$\operatorname{tr}\left(\left(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}\right)^{\mathsf{T}}\left(\boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}-\boldsymbol{X}\boldsymbol{M}\right)\right)=0.$$

□

Fact. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be given along with $\boldsymbol{D} \in \mathbb{R}^{d \times k}$ with $\boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}$. Then $\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2$, and

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T} \right\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2,$$

$$\arg\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T} \right\|_\mathsf{F}^2 = \arg\max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2.$$

Fact. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be given along with $\boldsymbol{D} \in \mathbb{R}^{d \times k}$ with $\boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}$. Then $\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2$, and

$$\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T} \right\|_\mathsf{F}^2 = \|\boldsymbol{X}\|_\mathsf{F}^2 - \max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2,$$

$$\arg\min_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T} \right\|_\mathsf{F}^2 = \arg\max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2.$$

Proof. Since

$$\begin{aligned} \|\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\|_\mathsf{F}^2 &= \operatorname{tr}\left((\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T})^\mathsf{T}(\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T})\right) = \operatorname{tr}\left((\boldsymbol{X}\boldsymbol{D})^\mathsf{T}(\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\boldsymbol{D})\right) \\ &= \operatorname{tr}\left((\boldsymbol{X}\boldsymbol{D})^\mathsf{T}(\boldsymbol{X}\boldsymbol{D})\right) = \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2, \end{aligned}$$

therefore

$$\begin{aligned} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T} \right\|_\mathsf{F}^2 &= \|\boldsymbol{X}\|_\mathsf{F}^2 - 2\operatorname{tr}\left((\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T})^\mathsf{T}\boldsymbol{X}\right) + \|\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\|_\mathsf{F}^2 \\ &= \|\boldsymbol{X}\|_\mathsf{F}^2 - \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2. \end{aligned}$$

□

Fact. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be given with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$. Then

$$\max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_{\mathsf{F}}^2 = \|\boldsymbol{X}\boldsymbol{V}_k\|_{\mathsf{F}}^2 = \sum_{i=1}^{k} s_k^2.$$

Fact. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be given with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$. Then

$$\max_{\substack{\boldsymbol{D} \in \mathbb{R}^{d \times k} \\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}}} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2 = \|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F}^2 = \sum_{i=1}^{k} s_k^2.$$

Proof. Define

$$S_1 := \{\boldsymbol{D} \in \mathbb{R}^{d \times k} : \boldsymbol{D}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}\}, \qquad S_2 := \{\boldsymbol{V}\boldsymbol{D} : \boldsymbol{D} \in S_1\}.$$

Note $S_1 = S_2$:

- $S_1 \subseteq S_2$, since $\boldsymbol{D} \in S_1$ implies $(\boldsymbol{V}^\mathsf{T}\boldsymbol{D})^\mathsf{T}\boldsymbol{V}^\mathsf{T}\boldsymbol{D} = \boldsymbol{I}$, thus $\boldsymbol{D} = \boldsymbol{V}(\boldsymbol{V}^\mathsf{T}\boldsymbol{D}) \in S_2$.
- $S_2 \subseteq S_1$, since $\boldsymbol{V}\boldsymbol{D} \in S_2$ implies $(\boldsymbol{V}\boldsymbol{D})^\mathsf{T}(\boldsymbol{V}\boldsymbol{D}) = \boldsymbol{I}$ thus $\boldsymbol{V}\boldsymbol{D} \in S_1$.

Therefore

$$\begin{aligned}
\max_{\boldsymbol{D} \in S_1} \|\boldsymbol{X}\boldsymbol{D}\|_\mathsf{F}^2 = \max_{\boldsymbol{M} \in S_2} \|\boldsymbol{X}\boldsymbol{M}\|_\mathsf{F}^2 = \max_{\boldsymbol{D} \in S_1} \|\boldsymbol{X}\boldsymbol{V}\boldsymbol{D}\|_\mathsf{F}^2 = \max_{\boldsymbol{D} \in S_1} \left\|\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}\boldsymbol{D}\right\|_\mathsf{F}^2 \\
= \max_{\boldsymbol{D} \in S_1} \operatorname{tr}\left((\boldsymbol{U}\boldsymbol{S}\boldsymbol{D})^\mathsf{T}(\boldsymbol{U}\boldsymbol{S}\boldsymbol{D})\right) = \max_{\boldsymbol{D} \in S_1} \operatorname{tr}\left(\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\right) \\
= \max_{\boldsymbol{D} \in S_1} \sum_{j=1}^{r} s_j^2 \sum_{i=1}^{k} \boldsymbol{D}_{ij}^2.
\end{aligned}$$

Proof (continued). We've the reduced the proof to showing

$$\max_{\substack{\boldsymbol{D}\in\mathbb{R}^{d\times k}\\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D}=\boldsymbol{I}}} \sum_{j=1}^{r} s_j^2 \sum_{i=1}^{k} \boldsymbol{D}_{ij}^2 = \|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F}^2,$$

and note moreover $\|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F} = \operatorname{tr}\left((\boldsymbol{U}_k\boldsymbol{S}_k)^\mathsf{T}(\boldsymbol{U}_k\boldsymbol{S}_k)\right) = \sum_{i=1}^{k} s_i^2$. Lastly:

- Since $\boldsymbol{V}_k \in \mathbb{R}^{d\times k}$ and $\boldsymbol{V}_k^\mathsf{T}\boldsymbol{V}_k = \boldsymbol{I}$,

$$\max_{\substack{\boldsymbol{D}\in\mathbb{R}^{d\times k}\\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D}=\boldsymbol{I}}} \sum_{j=1}^{r} s_j^2 \sum_{i=1}^{k} \boldsymbol{D}_{ij}^2 \geq \|\boldsymbol{X}\boldsymbol{V}_k\|_\mathsf{F}^2.$$

- For any feasible $\boldsymbol{D} \in \mathbb{R}^{d\times k}$, extend it to orthonormal $\boldsymbol{M} \in \mathbb{R}^{d\times d}$; since $\boldsymbol{M}^\mathsf{T}\boldsymbol{M} = \boldsymbol{I} = \boldsymbol{M}\boldsymbol{M}^\mathsf{T}$, then $\boldsymbol{M}^\mathsf{T}$ is orthonormal as well, and $\sum_{i=1}^{k} \boldsymbol{D}_{ij}^2 \leq \sum_{i=1}^{d} \boldsymbol{M}_{ij}^2 = 1$. Moreover, $\sum_{i,j} \boldsymbol{D}_{ij}^2 \leq k$, so

$$\max_{\substack{\boldsymbol{D}\in\mathbb{R}^{d\times k}\\ \boldsymbol{D}^\mathsf{T}\boldsymbol{D}=\boldsymbol{I}}} \sum_{j=1}^{r} S_j^2 \left(\sum_{i=1}^{k} \boldsymbol{D}_{ij}^2\right) \leq \max_{\substack{\boldsymbol{w}\in[0,1]^d\\ \sum_i w_i \leq k}} \sum_{j=1}^{r} S_j^2 w_j^2 \leq \sum_{j=1}^{k} S_j^2.$$

□

Summary for today

- **Un**supervised learning?
- PCA.
- PCA example.
- PCA proofs.