PCA and Autoencoders

CS229: Machine Learning

Sanmi Koyejo
Stanford University, Winter 2024
(Adapted from slides by Matgus Telgarsky and Alexander Schwing)

Goals of this lecture

- Understand Principal Components Analysis (PCA)
- Understand the relationship between PCA and Singular Value Decomposition
- Understand Autoencoders as a non-linear generalization of PCA

Reading material:

- Course Notes, Section 12
- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 12

Lecture notation (same as course notes)

Notation	Usage
$x^{(i)}, x$	Input(s), $\boldsymbol{x} \in \mathcal{X}$
$\Pi.$	Projection operator, e.g., $\Pi_{\mathcal{V}} x$ projects x the subspace $\mathcal{V}$
$\{Q, \Lambda\}$	Eigendecomposition, with eigenvectors $\{q_i\}$, and eigenvalues $\{\lambda_i\}$
$\{U, S, V^\top\}$	Singular value decomposition, with left singular vectors $\{u_i\}$
	right singular vectors $\{v_i\}$ and singular values $\{s_i\}$
$f(\cdot), g(\cdot)$	Encoder and Decoder functions, respectively

Overview.

So far, we have mostly focused on **supervised learning.**
i.e. constructing a mapping $f : \mathcal{X} \to \mathcal{Y}$ given pairs $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$.

Examples:

- Least squares.
- Logistic regression.
- ~~SVM.~~
- Neural networks.

Overview.

So far, we have mostly focused on **supervised learning.**
i.e. constructing a mapping $f : \mathcal{X} \to \mathcal{Y}$ given pairs $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$.

Examples:

- Least squares.
- Logistic regression.
- SVM.
- Neural networks.

- **P.S.** Autoregressive models (e.g., RNNs) are often trained for self-supervised tasks.

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Find structure in **unlabeled** data $\{x^{(i)}\}_{i=1}^{n}$.

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Find structure in **unlabeled** data $\{x^{(i)}\}_{i=1}^{n}$.

- Recover “hidden structure” (e.g., cliques in noisy graphs).

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Find structure in **unlabeled** data $\{x^{(i)}\}_{i=1}^{n}$.

- Recover “hidden structure” (e.g., cliques in noisy graphs).
- Data compression/dimension reduction.

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Find structure in **unlabeled** data $\{x^{(i)}\}_{i=1}^{n}$.

- Recover “hidden structure” (e.g., cliques in noisy graphs).
- Data compression/dimension reduction.
- Interpret / explain data and models.

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Find structure in **unlabeled** data $\{x^{(i)}\}_{i=1}^{n}$.

- Recover “hidden structure” (e.g., cliques in noisy graphs).
- Data compression/dimension reduction.
- Interpret / explain data and models.
- Features for supervised learning (e.g., word embeddings).

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Find structure in **unlabeled** data $\{x^{(i)}\}_{i=1}^{n}$.

- Recover “hidden structure” (e.g., cliques in noisy graphs).
- Data compression/dimension reduction.
- Interpret / explain data and models.
- Features for supervised learning (e.g., word embeddings).

Unsupervised learning.

Next we will study **unsupervised learning.**

What is the **goal** in unsupervised learning?

Find structure in **unlabeled** data $\{x^{(i)}\}_{i=1}^{n}$.

- Recover “hidden structure” (e.g., cliques in noisy graphs).
- Data compression/dimension reduction.
- Interpret / explain data and models.
- Features for supervised learning (e.g., word embeddings).

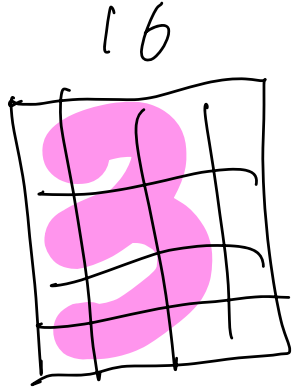
The task in unsupervised learning is less clear-cut.

Examples of unsupervised learning.

Examples of unsupervised learning.

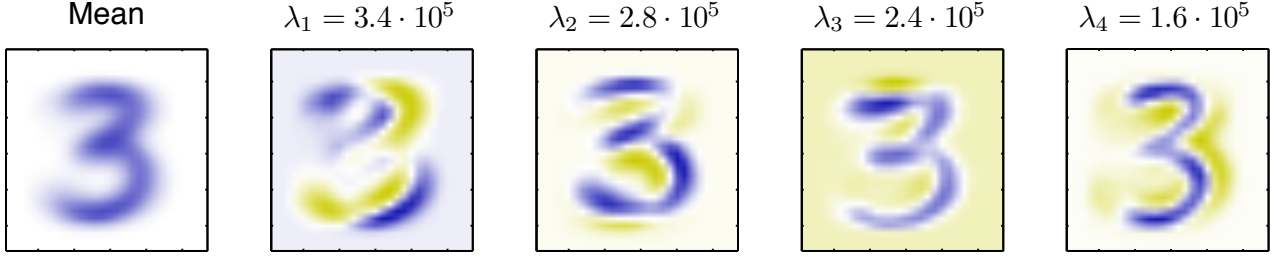
- PCA.
- k-means.
- Gaussian Mixture Models.
- Hidden Markov Models.
- Generative Adversarial Networks.

PCA Application 1: digit data.



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

Mean μ and eigenvectors v_1, v_2, v_3, v_4



Reconstructions:

Only have to store k numbers per image,
along with the mean μ and k eigenvectors ($256(k+1)$ numbers).

Application 2: eigenfaces.

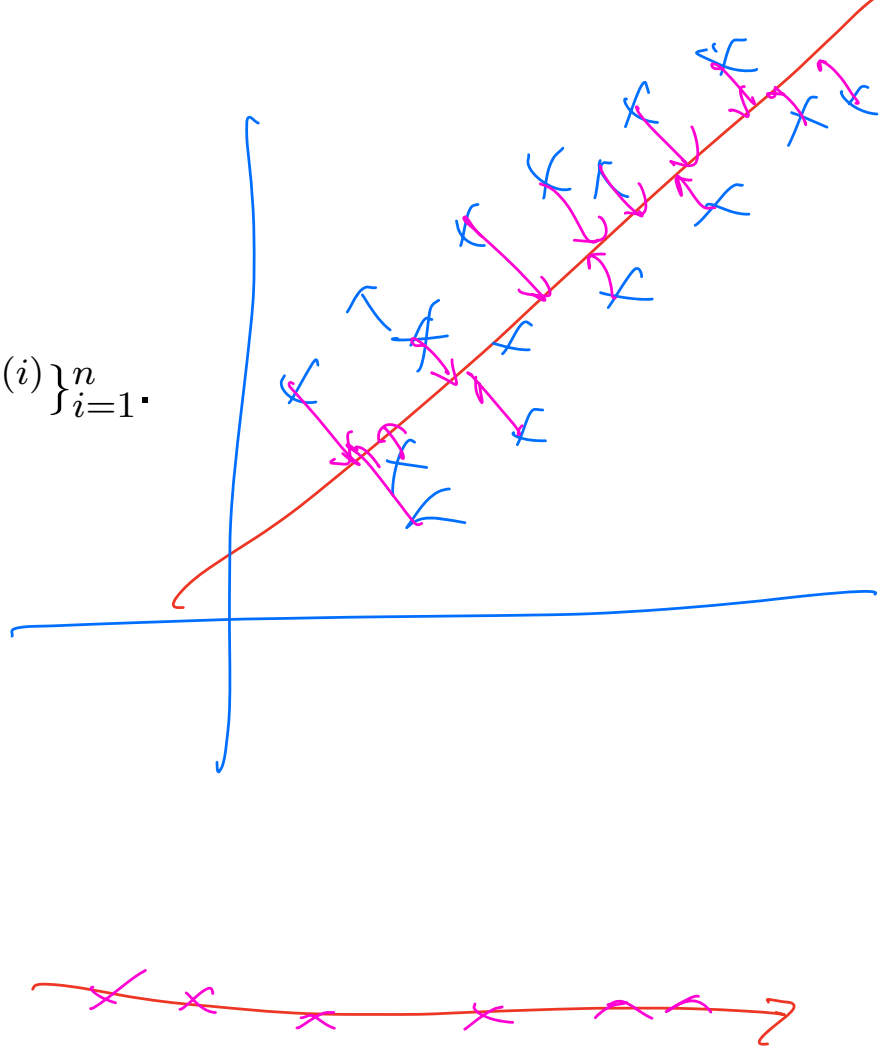
92×112 pixel images of faces (as vectors in $\mathbb{R}^{10304}$)

100 example images

top $k = 48$ eigenvectors

PCA (Principal Component Analysis).

Task (informal):
find best-fitting low-dimensional subspace to $\{x^{(i)}\}_{i=1}^{n}$.



PCA (Principal Component Analysis).

Task (informal):
find best-fitting low-dimensional subspace to $\{x^{(i)}\}_{i=1}^n$.

Task: Given $\{x^{(i)}\}_{i=1}^n$,
find linear subspace $\mathcal{V}$ (with projection operator $\Pi_{\mathcal{V}}$)
which minimizes variance:

$$\min_{\substack{\text{subspaces } \mathcal{V} \subseteq \mathbb{R}^d \\ \dim(\mathcal{V})=k}} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - \Pi_{\mathcal{V}} x^{(i)} \right\|^2.$$

Basis $V \in \mathbb{R}^{d \times k}$

$$V = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ v_1 & v_2 & \cdots & v_k \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$\Pi_{\mathcal{V}}(x) = V V^T x$$

PCA – matrix form.

Original form:

$$\min_{\substack{\text{subspaces } \mathcal{V}\subseteq\mathbb{R}^d \\ \dim(\mathcal{V})=k}} \frac{1}{n}\sum_{i=1}^{n}\left\|x^{(i)} - \Pi_{\mathcal{V}}x^{(i)}\right\|^2.$$

PCA – matrix form.

Original form:

$$\min_{\substack{\text{subspaces } \mathcal{V} \subseteq \mathbb{R}^d \\ \dim(\mathcal{V}) = k}} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - \Pi_{\mathcal{V}} x^{(i)} \right\|^2 .$$

To derive a simpler matrix form:

- Collect $\{x^{(i)}\}_{i=1}^{n}$ as rows of matrix $X \in \mathbb{R}^{n \times d}$.

PCA – matrix form.

Original form:

$$\min_{\substack{\text{subspaces } \mathcal{V} \subseteq \mathbb{R}^d \\ \dim(\mathcal{V}) = k}} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - \Pi_{\mathcal{V}} x^{(i)} \right\|^2.$$

To derive a simpler matrix form:

- Collect $\{x^{(i)}\}_{i=1}^n$ as rows of matrix $X \in \mathbb{R}^{n \times d}$.
- $\mathcal{V}$ is k-dimensional $\iff$ has basis $\{v_1, \ldots, v_k\}$. Collect $\{v_i\}_{i=1}^k$ into $V \in \mathbb{R}^{d \times k}$.

PCA – matrix form.

Original form:

$$\min_{\substack{\text{subspaces } \mathcal{V} \subseteq \mathbb{R}^d \\ \dim(\mathcal{V}) = k}} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - \Pi_{\mathcal{V}} x^{(i)} \right\|^2 .$$

To derive a simpler matrix form:

- Collect $\{x^{(i)}\}_{i=1}^n$ as rows of matrix $X \in \mathbb{R}^{n \times d}$.
- $\mathcal{V}$ is k-dimensional $\iff$ has basis $\{v_1, \ldots, v_k\}$. Collect $\{v_i\}_{i=1}^k$ into $V \in \mathbb{R}^{d \times k}$. What is the orthogonal projection operator onto columns of V?

PCA – matrix form.

Original form:

$$\min_{\substack{\text{subspaces } \mathcal{V} \subseteq \mathbb{R}^d \\ \dim(\mathcal{V}) = k}} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - \Pi_{\mathcal{V}} x^{(i)} \right\|^2 .$$

To derive a simpler matrix form:

- Collect $\{x^{(i)}\}_{i=1}^n$ as rows of matrix $X \in \mathbb{R}^{n \times d}$.
- $\mathcal{V}$ is k-dimensional $\iff$ has basis $\{v_1, \ldots, v_k\}$. Collect $\{v_i\}_{i=1}^k$ into $V \in \mathbb{R}^{d \times k}$. What is the orthogonal projection operator onto columns of V? **Answer**: $VV^\top$

PCA – matrix form.

Original form:

$$\min_{\substack{\text{subspaces } \mathcal{V}\subseteq\mathbb{R}^d \\ \dim(\mathcal{V})=k}} \frac{1}{n}\sum_{i=1}^{n}\left\|x^{(i)} - \Pi_{\mathcal{V}}x^{(i)}\right\|^2.$$

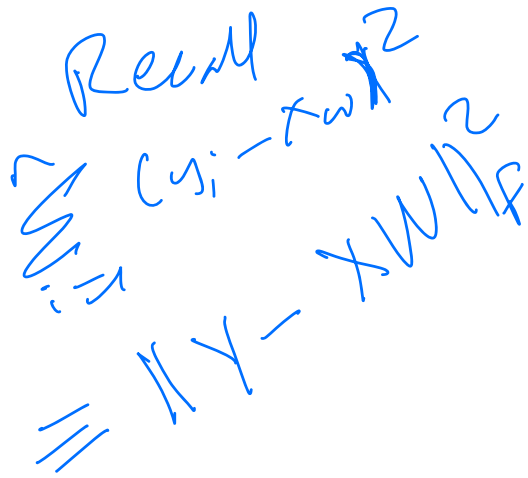
To derive a simpler matrix form:

- Collect $\{x^{(i)}\}_{i=1}^n$ as rows of matrix $X \in \mathbb{R}^{n\times d}$.
- $\mathcal{V}$ is k-dimensional $\iff$ has basis $\{v_1,\ldots,v_k\}$. Collect $\{v_i\}_{i=1}^k$ into $V \in \mathbb{R}^{d\times k}$. What is the orthogonal projection operator onto columns of V? **Answer**: $VV^\top$
- For matrix M, define **Frobenius norm** $\|M\|_{\mathsf{F}}^2 = \sum_{i,j} M_{ij}^2$.

PCA – matrix form.

Original form:

$$\min_{\substack{\text{subspaces } \mathcal{V} \subseteq \mathbb{R}^d \\ \dim(\mathcal{V}) = k}} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - \Pi_{\mathcal{V}} x^{(i)} \right\|^2.$$



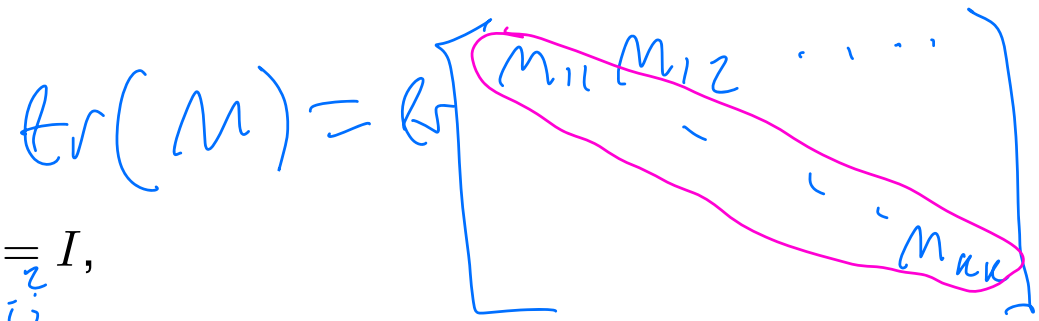
To derive a simpler matrix form:

- Collect $\{x^{(i)}\}_{i=1}^n$ as rows of matrix $X \in \mathbb{R}^{n \times d}$.
- $\mathcal{V}$ is k-dimensional $\iff$ has basis $\{v_1, \ldots, v_k\}$. Collect $\{v_i\}_{i=1}^k$ into $V \in \mathbb{R}^{d \times k}$. What is the orthogonal projection operator onto columns of V? **Answer**: $VV^\top$
- For matrix M, define **Frobenius norm** $\|M\|_{\mathsf{F}}^2 = \sum_{i,j} M_{ij}^2$.

With this notation, obtain alternate matrix form:

$$\min_{\substack{V \in \mathbb{R}^{d \times k} \\ V^\top V = I}} \frac{1}{n} \left\| X^\top - VV^\top X^\top \right\|_{\mathsf{F}}^2.$$

PCA – alternate matrix form.



Given $X \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{d\times k}$ with $V^\top V = I$,
since $\|M\|_{\mathsf{F}}^2 = \mathsf{trace}(M^\top M)$,

$$\left\|X^\top - VV^\top X^\top\right\|_{\mathsf{F}}^2 = \left\|X^\top\right\|_{\mathsf{F}}^2 - 2\mathsf{trace}(XVV^\top X^\top) + \mathsf{trace}(XVV^\top VV^\top X^\top)$$
$$= \|X\|_{\mathsf{F}}^2 - \mathsf{trace}(V^\top X^\top XV) = \|X\|_{\mathsf{F}}^2 - \|XV\|_{\mathsf{F}}^2 .$$

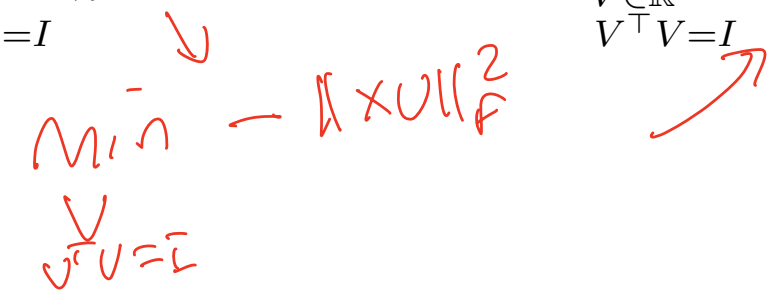
PCA – alternate matrix form.

Given $X \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{d \times k}$ with $V^\top V = I$,
since $\|M\|_{\mathsf{F}}^2 = \mathsf{trace}(M^\top M)$,

$$\begin{aligned}\left\|X^\top - VV^\top X^\top\right\|_{\mathsf{F}}^2 &= \left\|X^\top\right\|_{\mathsf{F}}^2 - 2\mathsf{trace}(XVV^\top X^\top) + \mathsf{trace}(XVV^\top VV^\top X^\top) \\ &= \|X\|_{\mathsf{F}}^2 - \mathsf{trace}(V^\top X^\top XV) = \|X\|_{\mathsf{F}}^2 - \|XV\|_{\mathsf{F}}^2 .\end{aligned}$$

PCA can thus be rewritten:

$$\min_{\substack{V \in \mathbb{R}^{d \times k} \\ V^\top V = I}} \frac{1}{n}\left\|X^\top - VV^\top X^\top\right\|_{\mathsf{F}}^2 = \max_{\substack{V \in \mathbb{R}^{d \times k} \\ V^\top V = I}} \|XV\|_{\mathsf{F}}^2 .$$

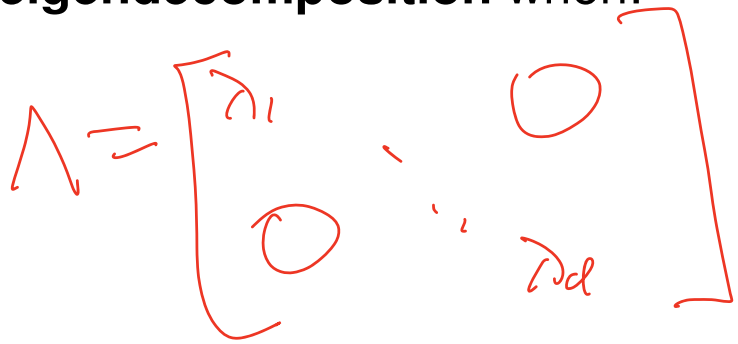


Aside: eigendecompositions.

Aside: eigendecompositions.

Recall: given a matrix M, then $\{Q, \Lambda\}$ are an **eigendecomposition** when:

- Q is orthonormal ($Q^\top Q = I$).
- Λ is diagonal.
- $M = Q\Lambda Q^\top = \sum_{i=1}^{d} \lambda_i q_i q_i^\top$.



Aside: eigendecompositions.

Recall: given a matrix M, then $\{Q, \Lambda\}$ are an **eigendecomposition** when:

- Q is orthonormal ($Q^\top Q = I$).
- Λ is diagonal.
- $M = Q\Lambda Q^\top = \sum_{i=1}^{d} \lambda_i q_i q_i^\top$.

$q_i \in \mathbb{R}^d$

$\lambda_i \in \mathbb{R}$

Notation and details:

- $\{q_1, \ldots, q_d\}$ are **eigenvectors**, $\{\lambda_1, \ldots, \lambda_d\}$ are **eigenvalues**.
- When M is symmetric, its eigendecomposition **exists** and is **real-valued**. Convention: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.
- Eigendecomposition is not, in general, unique! (e.g., zero matrix, . . .)

PCA via eigenvalues.

We've boiled PCA down to

$$\min_{\substack{V \in \mathbb{R}^{d \times k} \\ V^\top V = I}} \frac{1}{n} \left\| X^\top - V V^\top X^\top \right\|_{\mathsf{F}}^2 = \max_{\substack{V \in \mathbb{R}^{d \times k} \\ V^\top V = I}} \mathsf{trace}(V^\top X^\top X V).$$

PCA via eigenvalues.

We've boiled PCA down to

$$\min_{\substack{V \in \mathbb{R}^{d \times k} \\ V^\top V = I}} \frac{1}{n} \left\| X^\top - V V^\top X^\top \right\|_{\mathsf{F}}^2 = \max_{\substack{V \in \mathbb{R}^{d \times k} \\ V^\top V = I}} \text{trace}(V^\top X^\top X V).$$

$X^\top X$ is symmetric, with eigendecomposition $X^\top X = Q \Lambda Q^\top$.
We can also rewrite V in the basis Q, **How?**

$\exists Z$
s.t. $V = QZ$

PCA via eigenvalues.

We've boiled PCA down to

$$\min_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \frac{1}{n}\left\|X^\top - VV^\top X^\top\right\|_{\mathsf{F}}^2 = \max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV).$$

$(AB)^T = B^T A^T$

$X^\top X$ is symmetric, with eigendecomposition $X^\top X = Q\Lambda Q^\top$.
We can also rewrite V in the basis Q, **How?**
thus,

$$\max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV) = \max_{\substack{V=QZ\in\mathbb{R}^{d\times k}\\ (QZ)^\top (QZ)=I}} \text{trace}\left((QZ)^\top Q\Lambda Q^\top (QZ)\right)$$

PCA via eigenvalues.

We've boiled PCA down to

$$\min_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \frac{1}{n}\left\|X^\top - VV^\top X^\top\right\|_{\mathsf{F}}^2 = \max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV).$$

$X^\top X$ is symmetric, with eigendecomposition $X^\top X = Q\Lambda Q^\top$.
We can also rewrite V in the basis Q, **How?**
thus,

$$\max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV) = \max_{\substack{V=QZ\in\mathbb{R}^{d\times k}\\ (QZ)^\top (QZ)=I}} \text{trace}\left((QZ)^\top Q\Lambda Q^\top (QZ)\right)$$

PCA via eigenvalues.

We've boiled PCA down to

$$\min_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \frac{1}{n}\left\|X^\top - VV^\top X^\top\right\|_{\mathsf{F}}^2 = \max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV).$$

$X^\top X$ is symmetric, with eigendecomposition $X^\top X = Q\Lambda Q^\top$.
We can also rewrite V in the basis Q, **How?**
thus,

$$\max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV) = \max_{\substack{V=QZ\in\mathbb{R}^{d\times k}\\ (QZ)^\top(QZ)=I}} \text{trace}\left((QZ)^\top Q\Lambda Q^\top (QZ)\right)$$

$$= \max_{\substack{QZ\in\mathbb{R}^{d\times k}\\ Z^\top Z=I}} \text{trace}\left(Z^\top \Lambda Z\right) = \lambda_1 + \cdots + \lambda_k.$$

PCA via eigenvalues.

We've boiled PCA down to

$$\min_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \frac{1}{n}\left\|X^\top - VV^\top X^\top\right\|_{\mathsf{F}}^2 = \max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV).$$

$X^\top X$ is symmetric, with eigendecomposition $X^\top X = Q\Lambda Q^\top$.
We can also rewrite V in the basis Q, **How?**
thus,

$$\max_{\substack{V\in\mathbb{R}^{d\times k}\\ V^\top V=I}} \text{trace}(V^\top X^\top XV) = \max_{\substack{V=QZ\in\mathbb{R}^{d\times k}\\ (QZ)^\top(QZ)=I}} \text{trace}\left((QZ)^\top Q\Lambda Q^\top (QZ)\right)$$

$$= \max_{\substack{QZ\in\mathbb{R}^{d\times k}\\ Z^\top Z=I}} \text{trace}\left(Z^\top \Lambda Z\right) = \lambda_1 + \cdots + \lambda_k.$$

- The solution to PCA is the top k eigenvectors of $X^\top X$.
- The sum of eigenvalues is the maximum *value* of the variance.

PCA summary.

We are given data $\{x^{(i)}\}_{i=1}^n$;
We want subspace $\mathcal{V}$, $\dim(\mathcal{V}) = k$, minimizing $\sum_{i=1}^n \|x^{(i)} - \Pi_{\mathcal{V}} x^{(i)}\|^2$.

PCA summary.

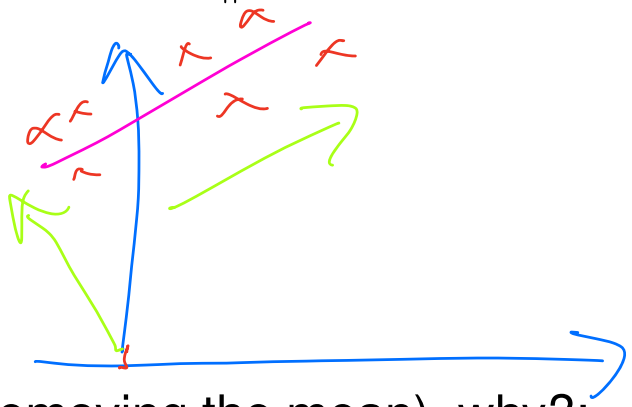
We are given data $\{x^{(i)}\}_{i=1}^n$;
We want subspace $\mathcal{V}$, $\dim(\mathcal{V}) = k$, minimizing $\sum_{i=1}^n \|x^{(i)} - \Pi_{\mathcal{V}} x^{(i)}\|^2$.

- Form matrix $X \in \mathbb{R}^{n \times d}$ with $x^{(i)}$ as row i.
- Compute top eigenvectors $\{v_1, \ldots, v_k\}$ of $X^\top X$. $\in \mathbb{R}^{d \times d}$
- Collect $\{v_1, \ldots, v_k\}$ as columns of $V \in \mathbb{R}^{d \times k}$.
- Output V; note $\Pi_{\mathcal{V}} = VV^\top$.

PCA summary.

We are given data $\{x^{(i)}\}_{i=1}^n$;
We want subspace $\mathcal{V}$, $\dim(\mathcal{V}) = k$, minimizing $\sum_{i=1}^n \|x^{(i)} - \Pi_{\mathcal{V}} x^{(i)}\|^2$.

- Form matrix $X \in \mathbb{R}^{n \times d}$ with $x^{(i)}$ as row i.
- Compute top eigenvectors $\{v_1, \ldots, v_k\}$ of $X^\top X$.
- Collect $\{v_1, \ldots, v_k\}$ as columns of $V \in \mathbb{R}^{d \times k}$.
- Output V; note $\Pi_{\mathcal{V}} = VV^\top$.



Remark. Often we want **PCA with centering** (i.e., first removing the mean), why?:

PCA summary.

We are given data $\{x^{(i)}\}_{i=1}^n$;
We want subspace $\mathcal{V}$, $\dim(\mathcal{V}) = k$, minimizing $\sum_{i=1}^n \|x^{(i)} - \Pi_{\mathcal{V}} x^{(i)}\|^2$.

- Form matrix $X \in \mathbb{R}^{n \times d}$ with $x^{(i)}$ as row i.
- Compute top eigenvectors $\{v_1, \ldots, v_k\}$ of $X^\top X$.
- Collect $\{v_1, \ldots, v_k\}$ as columns of $V \in \mathbb{R}^{d \times k}$.
- Output V; note $\Pi_{\mathcal{V}} = VV^\top$.

Remark. Often we want **PCA with centering** (i.e., first removing the mean), why?:
Find the mean $\mu = \frac{1}{n} \sum_{i=1}^n x^{(i)}$,
Form $X \in \mathbb{R}^{n \times d}$ where row i has $x^{(i)} - \mu$.
Associate $x^{(i)}$ with $\mu + \Pi_{\mathcal{V}}(x^{(i)} - \mu)$.

SVD (Singular Value Decomposition).

Every matrix $M \in \mathbb{R}^{n \times d}$ has an SVD $\{U, S, V^\top\}$.

- $U \in \mathbb{R}^{n \times r}$ with $U^\top U = I$ and $r := \text{rank}(M)$.
  Columns of U are **left singular vectors** $\{u_1, \ldots, u_r\}$.
- $S = \text{diag}(s_1, \ldots, s_r)$; these are the **singular values** $s_1 \geq \cdots \geq s_r$.
- $V \in \mathbb{R}^{d \times r}$ with $V^\top V = I$.
  Columns of V are **right singular vectors** $\{v_1, \ldots, v_r\}$.
- $M = USV^\top = \sum_{i=1}^{r} s_i u_i v_i^\top$.

SVD (Singular Value Decomposition).

Every matrix $M \in \mathbb{R}^{n \times d}$ has an SVD $\{U, S, V^\top\}$.

- $U \in \mathbb{R}^{n \times r}$ with $U^\top U = I$ and $r := \text{rank}(M)$.
  Columns of U are **left singular vectors** $\{u_1, \ldots, u_r\}$.
- $S = \text{diag}(s_1, \ldots, s_r)$; these are the **singular values** $s_1 \geq \cdots \geq s_r$.
- $V \in \mathbb{R}^{d \times r}$ with $V^\top V = I$.
  Columns of V are **right singular vectors** $\{v_1, \ldots, v_r\}$.
- $M = USV^\top = \sum_{i=1}^{r} s_i u_i v_i^\top$.

Remarks.

- Commonly known as the **thin SVD** or **truncated SVD** (e.g., Murphy book).
- $\sum_i s_i u_i v_i^\top$ is convenient representation (especially when r is small).
- *Again* in general not unique (consider $s_1 = s_2$).

More on the SVD.

Every matrix $M \in \mathbb{R}^{n \times d}$ has SVD $M = USV^\top$
with $U^\top U = I \in R^{r \times r}$, $V^\top V \in \mathbb{R}^{r \times r}$, $S = \text{diag}(s_1, \ldots, s_r)$.

- $M^\top M$ is symmetric and positive semi-definite
  PSD since $x^\top M^\top M x = |Mx|^2 \geq 0$.
  Note $M^T M = VS^2V^\top$.

$$M^\top M = (USV^\top)^\top (USV^\top)$$
$$= VSU^\top USV = VSSV^\top$$
$$= VS^2V^\top$$

More on the SVD.

Every matrix $M \in \mathbb{R}^{n \times d}$ has SVD $M = USV^\top$
with $U^\top U = I \in R^{r \times r}$, $V^\top V \in \mathbb{R}^{r \times r}$, $S = \text{diag}(s_1, \ldots, s_r)$.

- $M^\top M$ is symmetric and positive semi-definite
  PSD since $x^\top M^\top M x = |Mx|^2 \geq 0$.
  Note $M^T M = V S^2 V^\top$.
- $MM^\top$ is symmetric and positive semi-definite; also $MM^\top = US^2U^\top$.

More on the SVD.

Every matrix $M \in \mathbb{R}^{n \times d}$ has SVD $M = USV^\top$
with $U^\top U = I \in R^{r \times r}$, $V^\top V \in \mathbb{R}^{r \times r}$, $S = \text{diag}(s_1, \ldots, s_r)$.

- $M^\top M$ is symmetric and positive semi-definite
  PSD since $x^\top M^\top M x = |Mx|^2 \geq 0$.
  Note $M^T M = V S^2 V^\top$.
- $MM^\top$ is symmetric and positive semi-definite; also $MM^\top = US^2U^\top$.
- Eigenvalues of $MM^\top$ and $M^\top M$ coincide;
  agree with $\{s_1^2, \ldots, s_r^2, 0, \ldots 0\}$.

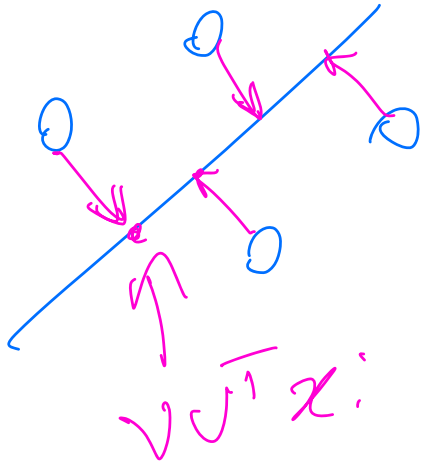
More on the SVD.

Every matrix $M \in \mathbb{R}^{n \times d}$ has SVD $M = USV^\top$
with $U^\top U = I \in R^{r \times r}$, $V^\top V \in \mathbb{R}^{r \times r}$, $S = \text{diag}(s_1, \ldots, s_r)$.

- $M^\top M$ is symmetric and positive semi-definite
  PSD since $x^\top M^\top M x = |Mx|^2 \geq 0$.
  Note $M^T M = VS^2V^\top$.
- $MM^\top$ is symmetric and positive semi-definite; also $MM^\top = US^2U^\top$.
- Eigenvalues of $MM^\top$ and $M^\top M$ coincide;
  agree with $\{s_1^2, \ldots, s_r^2, 0, \ldots 0\}$.
- Eigenvectors of $M^\top M$ are **right singular vectors**;
  Eigenvectors of $MM^\top$ are **left singular vectors**.

SVD and PCA.

Given data $\{x^{(i)}\}_{i=1}^{n}$ collected as rows of $X \in \mathbb{R}^{n \times d}$,
PCA solution was top k eigenvectors of $X^\top X$,
the projected points are $VV^\top X^\top$ with eigenvectors V.



SVD and PCA.

Given data $\{x^{(i)}\}_{i=1}^{n}$ collected as rows of $X \in \mathbb{R}^{n \times d}$,
PCA solution was top k eigenvectors of $X^\top X$,
the projected points are $VV^\top X^\top$ with eigenvectors V.

- Eigenvectors of $X^\top X$ are right singular vectors V in $X = USV^\top$.
- PCA solution is V_k (first k columns of V).
- Projected data is $V_k V_k^\top X^\top = V_k V_k^\top VSU^\top = V_k S_k U_k^\top$.
  Reduced dimension description is $S_k U_k^\top$.

PCA summary so far.

- Goal in PCA: find linear subspace $\mathcal{V}$ close to data, $\dim(\mathcal{V}) = k$.
- Objective function:

$$\min_{\substack{\text{subspaces } \mathcal{V} \subseteq \mathbb{R}^d \\ \dim(\mathcal{V}) = k}} \sum_{i=1}^{n} \left\| x^{(i)} - \Pi_{\mathcal{V}} x^{(i)} \right\|^2 .$$

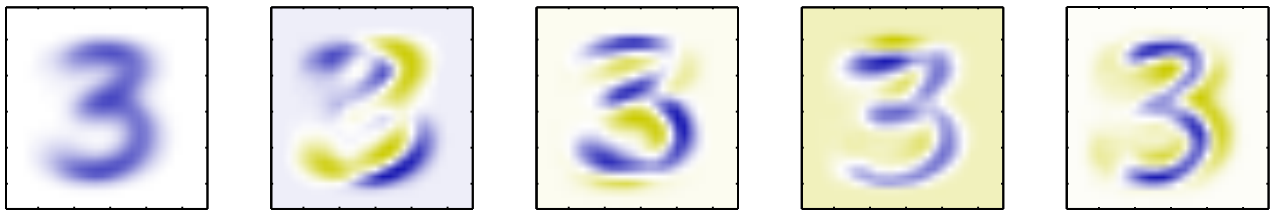
- Solution 1: top k eigenvectors of $X^\top X$.
- Solution 2: top k right singular vectors of X.

PCA Application 1: digit data.

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

Mean μ and eigenvectors v_1, v_2, v_3, 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 μ and k eigenvectors ($256(k+1)$ numbers).

PCA Application 1: digit data.

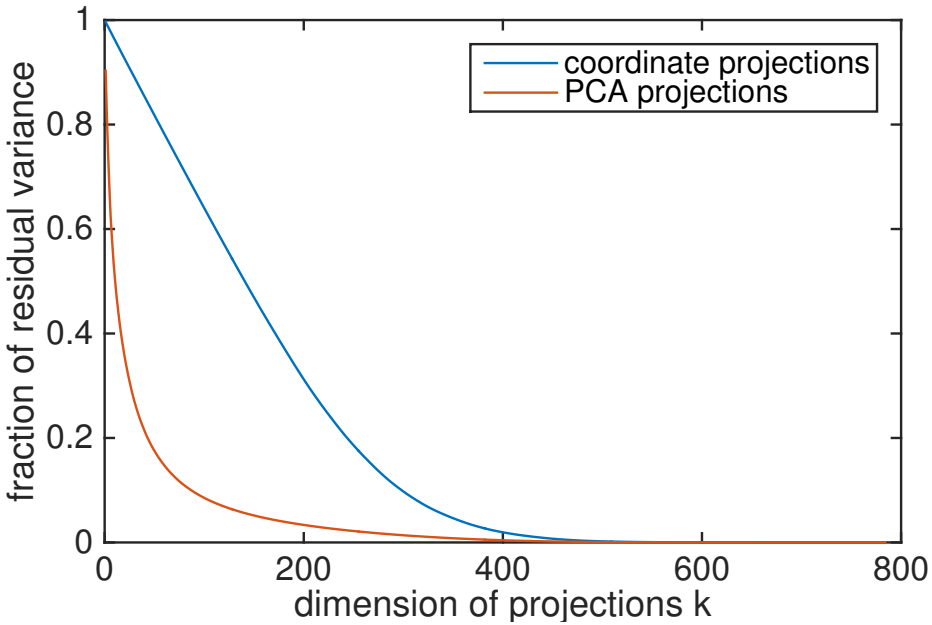
Data $\{x^{(i)}\}_{i=1}^n$ with $x^{(i)} \in \mathbb{R}^{784}$.

- Residual variance left by rank-k PCA projection:

$$1 - \frac{\sum_{j=1}^k \text{variance in direction } v_j}{\text{total variance}}.$$

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

$$1 - \frac{\sum_{j=1}^k \text{variance in direction } \mathbf{e}_j}{\text{total variance}}.$$



Application 2: topic modeling.

- Let $\{x^{(i)}\}_{i=1}^{n}$ denote *text documents*:
  each $x^{(i)} \in \mathbb{R}^d$ contains normalized word counts
  (d possible words).
- With SVD/PCA, replace $x^{(i)}$ with $VV^\top x = Vy$;
  now $y \in \mathbb{R}^k$ (e.g., $k = 100 \ll 30,000 = d$).
- **Problem:** negative values!
  Common solution: Look up non-negative matrix factorization (NMF)
- **Further reading (beyond this class):** *LSA (latent semantic analysis)* and *LSI (latent semantic indexing)*.

Algorithms.

- We reduced PCA to eigenvectors of $X^\top X$.
- An easy solver here is the **power method**.

Algorithms.

- We reduced PCA to eigenvectors of $X^\top X$.
- An easy solver here is the **power method**.
- Basic observation: given $M = Q\Lambda Q^\top$, then

$$M^k = Q\Lambda^k Q^\top = \sum_{i=1}^{d} \lambda_i^k q_i q_i^\top.$$

- i.e., M^k has clearer "eigenvalue structure" then M.
  How do we leverage this algorithmically?

Power method.

For top eigenvector, $M^k x / \|M^k x\| \approx q_1$ (details in appendix), iterate as follows.

- Randomly initialize x_0 with $\|x_0\| = 1$.
- Iterate $x_{t+1} := \frac{Mx_t}{\|Mx_t\|}$.

Power method.

For top eigenvector, $M^k x / \|M^k x\| \approx q_1$ (details in appendix), iterate as follows.

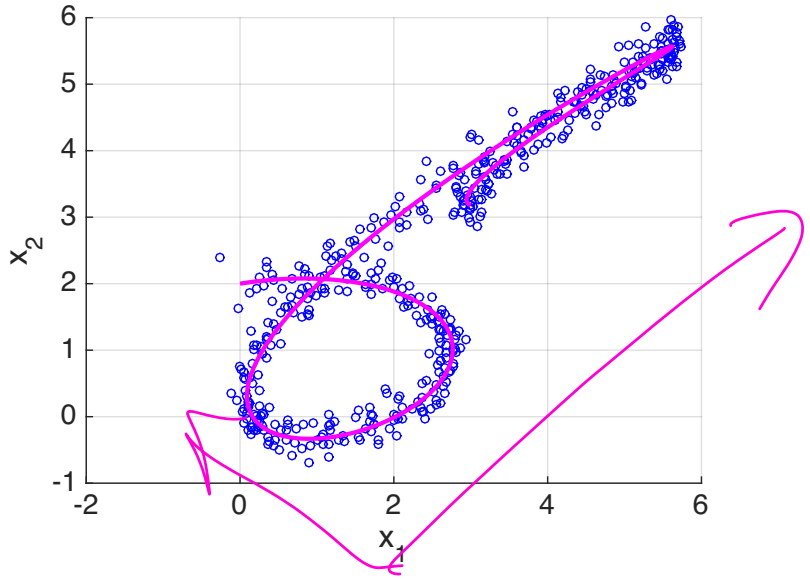
- Randomly initialize x_0 with $\|x_0\| = 1$.
- Iterate $x_{t+1} := \frac{Mx_t}{\|Mx_t\|}$.

Remarks.

- After $\ln(1/\epsilon)$ steps, power method achieves ϵ-approx solution!
- For left and right singular vectors: replace M with $MM^\top$ and $M^\top M$ respectively.
- For multiple eigenvectors, the most common approach is deflation (project the matrix to the orthogonal basis of the estimated eigenvector) before estimating subsequent eigenvectors.

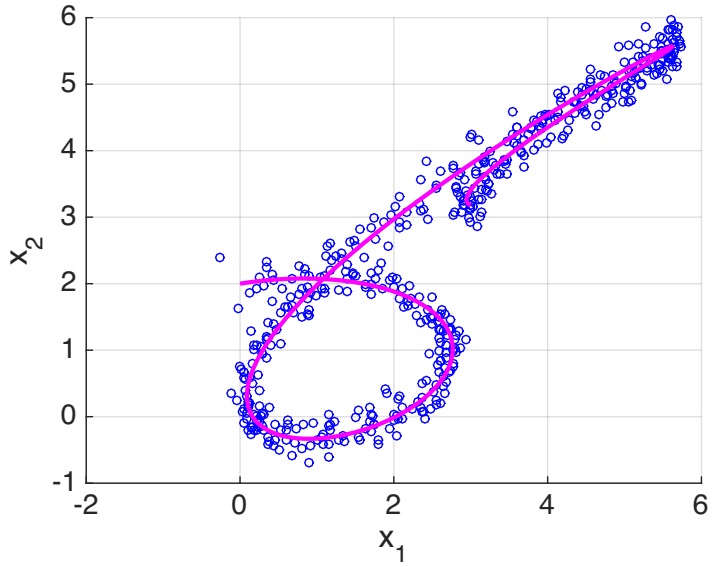
Nonlinear Embeddings and the Manifold Hypothesis.

- **New assumption:** data is concentrated around a low dimensional (non-linear) manifold.
- What will PCA do for such data?



Nonlinear Embeddings and the Manifold Hypothesis.

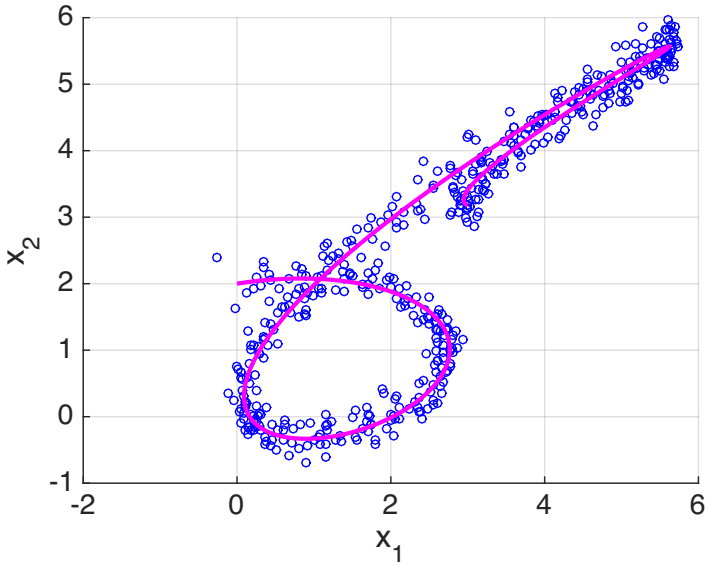
- **New assumption:** data is concentrated around a low dimensional (non-linear) manifold.
- What will PCA do for such data?



- **How can we fix this?**

Nonlinear Embeddings and the Manifold Hypothesis.

- **New assumption:** data is concentrated around a low dimensional (non-linear) manifold.
- What will PCA do for such data?



- **How can we fix this?**
- Idea: Replace linear mapping with non-linear mapping!

Autoencoders.

Task (informal): find best-fitting non-linear representation of $\{x^{(i)}\}_{i=1}^{n}$.

Autoencoders.

Task (informal): find best-fitting non-linear representation of $\{x^{(i)}\}_{i=1}^n$.

Task: Given $\{x^{(i)}\}_{i=1}^n$, find an "encoder" $f(\cdot)$ and a "decoder" $g(\cdot)$ which minimizes reconstruction error:

$$\min_{f,g} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - f\left(g\left(x^{(i)}\right)\right) \right\|^2 .$$

Swap f and g

$$\min_{f,g} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - g\left(f\left(x^{(i)}\right)\right) \right\|^2$$

(note we removed the orthogonality constraint)

Autoencoders.

Task (informal): find best-fitting non-linear representation of $\{x^{(i)}\}_{i=1}^n$.

Task: Given $\{x^{(i)}\}_{i=1}^n$, find an "encoder" $f(\cdot)$ and a "decoder" $g(\cdot)$ which minimizes reconstruction error:

$$\min_{f,g} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - f\left(g\left(x^{(i)}\right)\right) \right\|^2 .$$

Swap f and g

(note we removed the orthogonality constraint)

- The encoder f maps each input $x^{(i)}$ to a low-dimensional "code" as $f(x^{(i)}) = z^{(i)} \in \mathbb{R}^m$, where $m < d$.

Autoencoders.

Task (informal): find best-fitting non-linear representation of $\{x^{(i)}\}_{i=1}^n$.

Task: Given $\{x^{(i)}\}_{i=1}^n$, find an "encoder" $f(\cdot)$ and a "decoder" $g(\cdot)$ which minimizes reconstruction error:

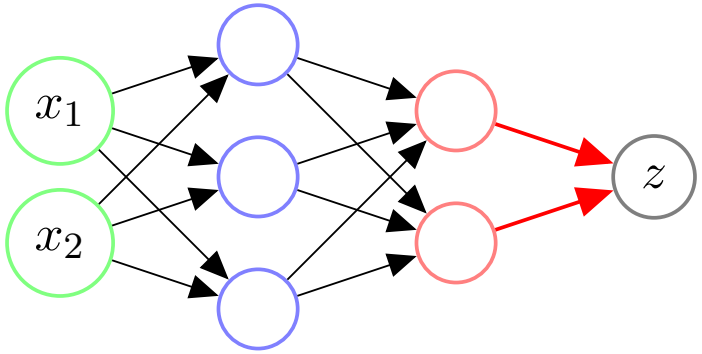
$$\min_{f,g} \frac{1}{n} \sum_{i=1}^{n} \left\| x^{(i)} - f\left(g\left(x^{(i)}\right)\right) \right\|^2 .$$

Swap f and g

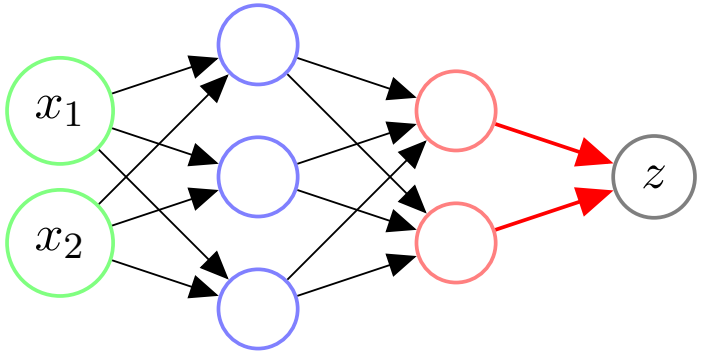
(note we removed the orthogonality constraint)

- The encoder f maps each input $x^{(i)}$ to a low-dimensional "code" as $f(x^{(i)}) = z^{(i)} \in \mathbb{R}^m$, where $m < d$.
- The decoder g maps each code $z^{(i)}$ back to the high-dimensional space as $g(z^{(i)}) = \hat{x}^{(i)} \in \mathbb{R}^d$.

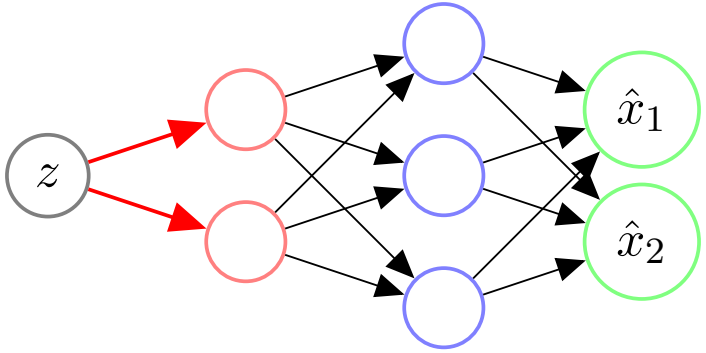
Common architecture: We chose the encoder f as a neural network



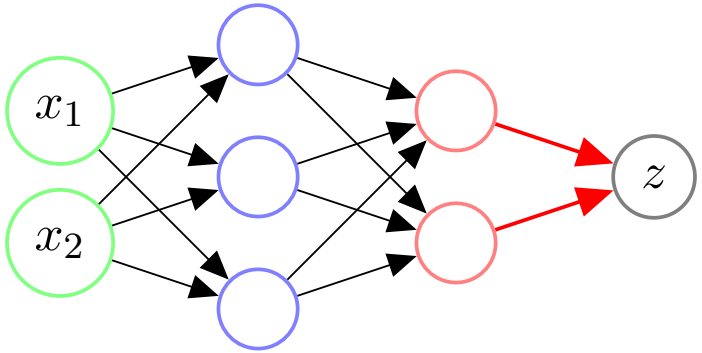
Common architecture: We chose the encoder f as a neural network



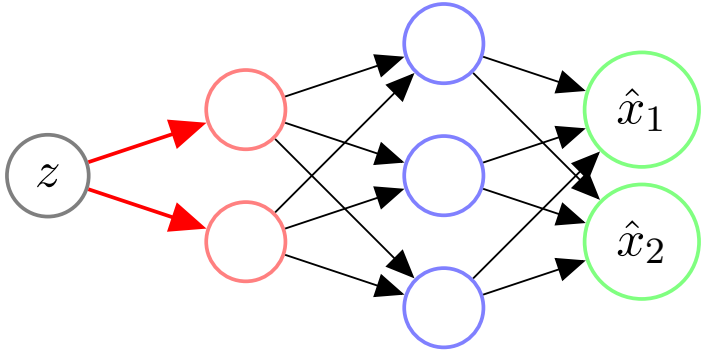
Common architecture: Similarly, chose the decoder g as a neural network (usually mirroring f)



Common architecture: We chose the encoder f as a neural network

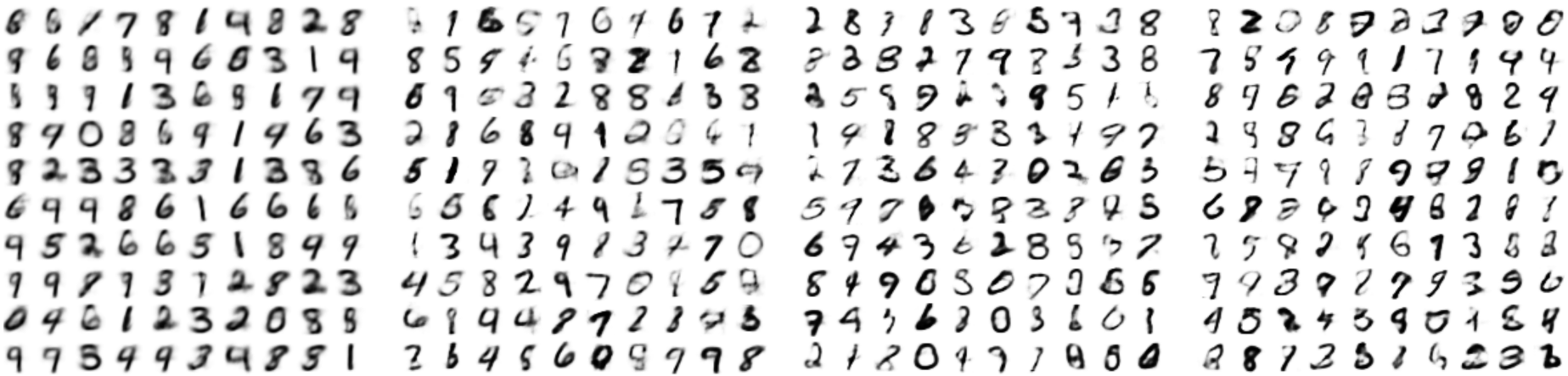


Common architecture: Similarly, chose the decoder g as a neural network (usually mirroring f)



All the parameters (for f and g) can be trained using SGD (backpropagation).

Example: Auto-encoder for MNIST digits.



(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Quiz / Summary:

- What is the goal of PCA?

Quiz / Summary:

- What is the goal of PCA?
- What are some applications of PCA?
- How do you solve PCA using eigendecomposition?

Quiz / Summary:

- What is the goal of PCA?
- What are some applications of PCA?
- How do you solve PCA using eigendecomposition?
- How do you solve PCA using the SVD?
- How is PCA related to Autoencoders?

Important topics of this lecture

- PCA
- SVD
- Autoencoders

Up next:

- Clustering

Extra Slides (not covered in lecture).

Power method background.

- From $M = Q\Lambda Q^\top$, have $M^k = Q\Lambda^k Q^\top = \sum_i \lambda_i^k q_i q_i^\top$.
- Pick any unit vector x; write it as Qy for unit vector y.
- Therefore $M^k x = \sum_i \lambda_i^k q_i q_i^\top x = \sum_i \lambda_i^k y_i q_i$.
  Seems to “amplify” top eigenvalue!

Power method background.

- From $M = Q\Lambda Q^\top$, have $M^k = Q\Lambda^k Q^\top = \sum_i \lambda_i^k q_i q_i^\top$.
- Pick any unit vector x; write it as Qy for unit vector y.
- Therefore $M^k x = \sum_i \lambda_i^k q_i q_i^\top x = \sum_i \lambda_i^k y_i q_i$.
  Seems to "amplify" top eigenvalue!
- Indeed, setting $\Delta := \max_{j \geq 2} \frac{\lambda_j}{\lambda_1} \left(\frac{y_j}{y_1}\right)^{\frac{1}{k}}$,

$$\frac{(q_1^\top M^k x)^2}{\|M^k x\|^2} = \frac{\lambda_1^{2k} y_1^2}{\sum_i \lambda_i^{2k} y_i^2} = \frac{1}{1 + \sum_{i \geq 2} \left(\frac{\lambda_i}{\lambda_1}\right)^{2k} \left(\frac{y_i}{y_1}\right)^2} = \frac{1}{1 + \sum_{i \geq 2} \left(\frac{\lambda_i}{\lambda_1} \left(\frac{y_i}{y_1}\right)^{\frac{1}{k}}\right)^{2k}}$$

Power method background.

- From $M = Q\Lambda Q^\top$, have $M^k = Q\Lambda^k Q^\top = \sum_i \lambda_i^k q_i q_i^\top$.
- Pick any unit vector x; write it as Qy for unit vector y.
- Therefore $M^k x = \sum_i \lambda_i^k q_i q_i^\top x = \sum_i \lambda_i^k y_i q_i$.
  Seems to "amplify" top eigenvalue!
- Indeed, setting $\Delta := \max_{j \geq 2} \frac{\lambda_j}{\lambda_1} \left(\frac{y_j}{y_1}\right)^{\frac{1}{k}}$,

$$\frac{(q_1^\top M^k x)^2}{\|M^k x\|^2} = \frac{\lambda_1^{2k} y_1^2}{\sum_i \lambda_i^{2k} y_i^2} = \frac{1}{1 + \sum_{i \geq 2} \left(\frac{\lambda_i}{\lambda_1}\right)^{2k} \left(\frac{y_i}{y_1}\right)^2} = \frac{1}{1 + \sum_{i \geq 2} \left(\frac{\lambda_i}{\lambda_1} \left(\frac{y_i}{y_1}\right)^{\frac{1}{k}}\right)^{2k}}$$

$$\geq \frac{1}{1 + k\Delta^{2k}} = 1 - \frac{k\Delta^{2k}}{1 + k\Delta^{2k}} \geq 1 - k\Delta^{2k}.$$

Power method background.

- From $M = Q\Lambda Q^\top$, have $M^k = Q\Lambda^k Q^\top = \sum_i \lambda_i^k q_i q_i^\top$.
- Pick any unit vector x; write it as Qy for unit vector y.
- Therefore $M^k x = \sum_i \lambda_i^k q_i q_i^\top x = \sum_i \lambda_i^k y_i q_i$.
  Seems to "amplify" top eigenvalue!
- Indeed, setting $\Delta := \max_{j \geq 2} \frac{\lambda_j}{\lambda_1} \left(\frac{y_j}{y_1}\right)^{\frac{1}{k}}$,

$$\frac{(q_1^\top M^k x)^2}{\|M^k x\|^2} = \frac{\lambda_1^{2k} y_1^2}{\sum_i \lambda_i^{2k} y_i^2} = \frac{1}{1 + \sum_{i \geq 2} \left(\frac{\lambda_i}{\lambda_1}\right)^{2k} \left(\frac{y_i}{y_1}\right)^2} = \frac{1}{1 + \sum_{i \geq 2} \left(\frac{\lambda_i}{\lambda_1} \left(\frac{y_i}{y_1}\right)^{\frac{1}{k}}\right)^{2k}}$$

$$\geq \frac{1}{1 + k\Delta^{2k}} = 1 - \frac{k\Delta^{2k}}{1 + k\Delta^{2k}} \geq 1 - k\Delta^{2k}.$$

- **Thus:** if gap λ_1/λ_2 large and y_1 not too small, $1 - k\Delta^{2k} \approx 1$, then $M^k x / \|M^k x\| \approx q_1$.