Example: Handwritten digits

Handwritten digit data, but with no labels

What can we do?

• Suppose we know that there are 10 groupings, can we find the groups?

• What if we don’t know there are 10 groups?

• How can we discover/explore other structure in such data?
Dimensionality Reduction

Data: \( \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d \)

Goal: find a ‘useful’ transformation \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^k \) that helps in the downstream prediction task.

Some previously seen useful transformations:

• z-scoring \( (x_1, \ldots, x_d) \mapsto \left( \frac{x_1 - \mu_1}{\sigma_1}, \ldots, \frac{x_d - \mu_d}{\sigma_d} \right) \) \( \text{Keeps same dimensionality but with better scaling} \)

• Kernel transformations. \( \text{Higher dimensionality, making data linearly separable} \)

What are other desirable feature transformations? \( \text{How about lower dimensionality while keeping the relevant information?} \)
Principal Components Analysis (PCA)

Data: \( \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n \in \mathbb{R}^d \)

Goal: find the best **linear** transformation \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^k \) that best maintains reconstruction accuracy.

Equivalently, minimize aggregate residual error

Define: \( \Pi_k^k : \mathbb{R}^d \rightarrow \mathbb{R}^d \) \( k \)-dimensional orthogonal linear projector

\[
\text{minimize } \frac{1}{n} \sum_{i=1}^{n} \left\| \bar{x}_i - \Pi_k^k(\bar{x}_i) \right\|^2
\]

How do we optimize this?
Dimensionality Reduction via Projections

A $k$ dimensional subspace can be represented by $\vec{q}_1, \ldots, \vec{q}_k \in \mathbb{R}^d$ orthonormal vectors.

The projection of any $\vec{x} \in \mathbb{R}^d$ in the $\text{span}(\vec{q}_1, \ldots, \vec{q}_k)$ is given by

$$\sum_{i=1}^{k} (\vec{q}_i \cdot \vec{x}) \vec{q}_i = \left( \sum_{i=1}^{k} \vec{q}_i \vec{q}_i^T \right) \vec{x}$$

To represent it in $\mathbb{R}^k$ (using basis $\vec{q}_1, \ldots, \vec{q}_k$) the coefficients simply are: $(\vec{q}_1 \cdot \vec{x}), \ldots, (\vec{q}_k \cdot \vec{x})$
PCA: $k = 1$ case

If projection dimension $k = 1$, then looking for a $q$ such that

$$\underset{\|q\|=1}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{x}_i - (\bar{q} \bar{q}^T) x_i \right\|^2$$

$$= \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i^T \tilde{x}_i \right) - \bar{q}^T \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T \right) \bar{q}$$

$$\propto - \bar{q}^T \left( \frac{1}{n} XX^T \right) \bar{q}$$

Equivalent formulation:

$$\underset{\|q\|=1}{\text{maximize}} \quad \bar{q}^T \left( \frac{1}{n} XX^T \right) \bar{q}$$

How to solve?
Recall for any matrix $M$, the $(\lambda, v)$ pairs of the fixed point equation

$$Mv = \lambda v$$

are the eigenvalue and the eigenvectors of $M$. ($v \neq 0$)

$$v^T M v = \lambda v^T v$$

$$\lambda = \frac{v^T M v}{v^T v} = \bar{v}^T M \bar{v}$$

where $\bar{v} = \frac{v}{||v||}$ (ie, unit length)

So,

$$\text{maximize}_{||q||=1} \quad q^T \left( \frac{1}{n} XX^T \right) q$$

Basically is the top eigenvector of matrix $(1/n) XX^T$!
PCA: $k = 1$ case

\[
\text{maximize}_{\|q\| = 1} \quad q^T \left( \frac{1}{n} X X^T \right) q
\]

Covariance of data (if mean = 0)

For any $q$ the quadratic form $q^T \left( \frac{1}{n} X X^T \right) q$ is the empirical variance of data in the direction $q$, ie, of data $q^T \vec{x}_1, \ldots, q^T \vec{x}_n$

Therefore, the top eigenvector solution implies that the direction of maximum variance minimizes the residual error!

What about general $k$?
PCA: general $k$ case

\[
\arg\min_{Q \in \mathbb{R}^{d \times k}} \min_{Q^\top Q = I} \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{x}_i - QQ^\top \tilde{x}_i \right\|^2 = \arg\max_{Q \in \mathbb{R}^{d \times k}} \min_{Q^\top Q = I} \text{tr} \left( Q^\top \left( \frac{1}{n} XX^\top \right) Q \right)
\]

Solution: Basically is the top $k$ eigenvectors of the matrix $XX^\top$!

\[
\text{tr} \left( Q^\top \left( \frac{1}{n} XX^\top \right) Q \right) = \sum_{i=1}^{k} \text{empirical variance of } q_i^\top x
\]

$k$-dimensional subspace preserving maximum amount of variance
**PCA: Example Handwritten Digits**

*Images of handwritten 3s in $\mathbb{R}^{784}$*

Any example: $x = q_1 + w_1 + q_2 + w_2 + \ldots$

Data Reconstruction:

- $k = 1$: Project $x$ onto $q_1$.
- $k = 10$: Project $x$ onto $q_1$ and $q_2$.
- $k = 50$: Project $x$ onto $q_1$, $q_2$, and $q_3$.
- $k = 200$: Project $x$ onto all $q_i$.

*We can compress each datapoint to just $k$ numbers!*
Other Popular Dimension Reduction Methods

Multi-dimensional Scaling

Independent Component Analysis (ICA) (for blind source separation)

Non-negative matrix factorization (to create additive models)

Dictionary Learning

Random Projections

... All of them are linear methods
Consider non-linear data

Linear embedding

non-linear embedding
Basic optimization criterion:

Find an embedding that:

- Keeps neighboring points close
- Keeps far-off points far

**Example variation 1:**

*Distort neighboring distances by at most \((1 \pm \varepsilon)\) factor, while maximizing non-neighbor distances.*

**Example variation 2:**

*Compute geodesic (local hop) distances, and find an embedding that best preserves geodesics.*
Non-linear embedding: Example
Popular Non-Linear Methods

Locally Linear Embedding (LLE)

Isometric Mapping (IsoMap)

Laplacian Eigenmaps (LE)

Local Tangent Space Alignment (LTSA)

Maximum Variance Unfolding (MVU)

...
What We Learned...

• Dimensionality Reduction
  Linear vs non-linear Dimensionality Reduction

• Principal Component Analysis
Questions?