Principal Component Analysis

Natasha Jaques

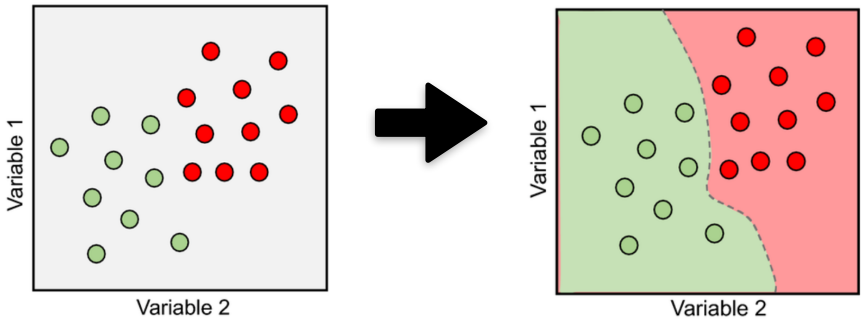
UNIVERSITY of WASHINGTON

Unsupervised vs. supervised learning

Previously: supervised learning

- Each data point x_i has a corresponding label y_i;$\{x_i, y_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. Try to predict the label y for a new test point x

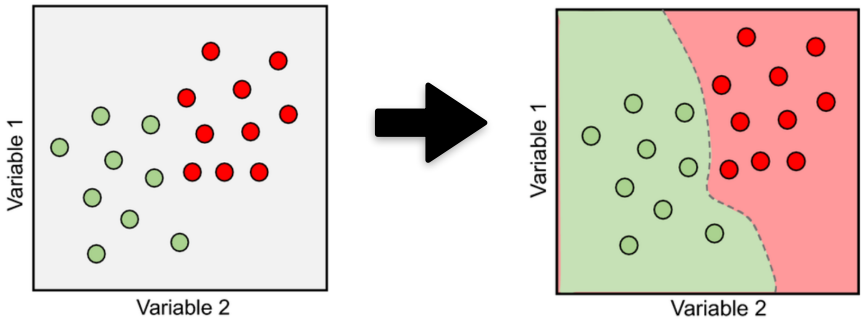
label



Unsupervised vs. supervised learning

Previously: supervised learning

- Each data point x_i has a corresponding label y_i;$\{x_i, y_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. Try to predict the label y for a new test point x



P(price | house)

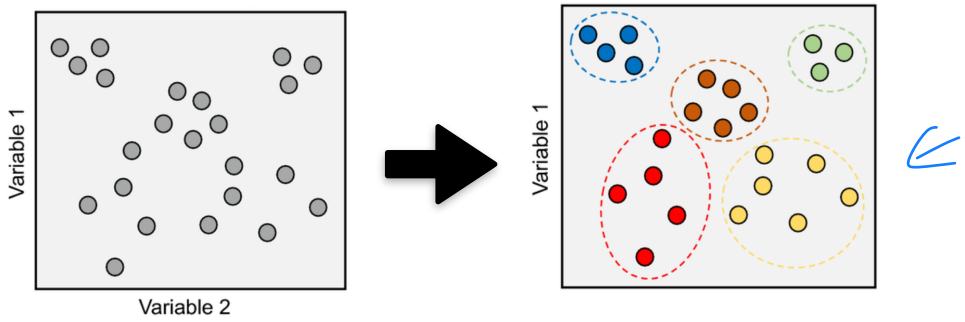
P(features)

P(x1 ... xd)

P(house)

Now: Unsupervised learning

- No labels: data $\{x_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$. Try to model the data distribution $P(X)$, potentially by finding patterns/clusters, or a low-dimensional representation



Motivation: dimensionality reduction

- It takes $n \times d$ memory to store data $\{x_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$
- But many real data have patterns that repeat over samples. Can we find some patterns and use them?

d=32x32pixels per image

n images

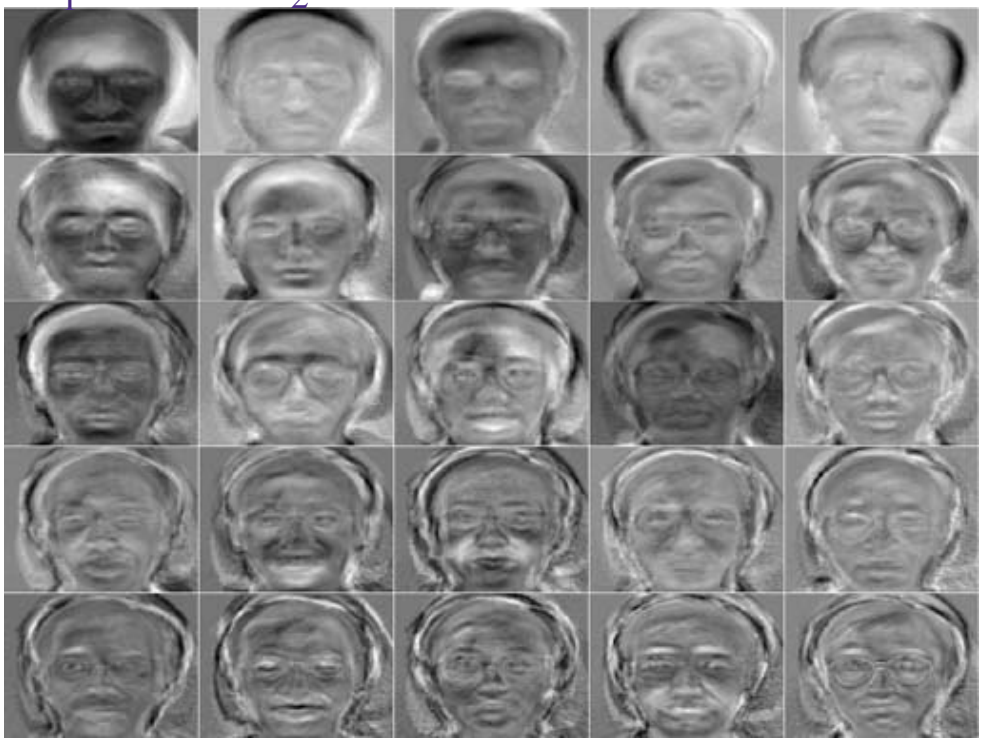
$d \times n$ real values to store the data

Principal component analysis finds a compact linear representation

- patterns that capture the distinct features of the samples is called **principal component** (to be formally defined later)

Principal components:

$u_1 \in \mathbb{R}^d$ u_2



https://en.wikipedia.org/wiki/Eigenface

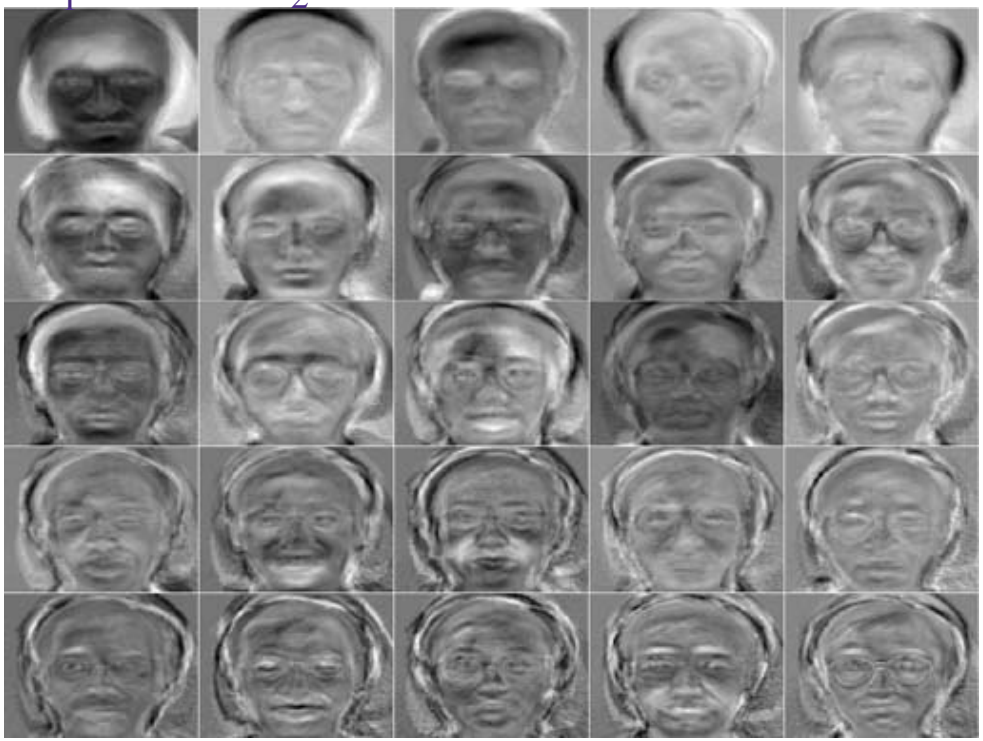
Principal component analysis finds a compact linear representation

- patterns that capture the distinct features of the samples is called **principal component** (to be formally defined later)

- we can represent each sample as a **weighted linear combination** of, say, q=25 principal components, and just store the weights

Principal components:

$u_1 \in \mathbb{R}^d$ u_2



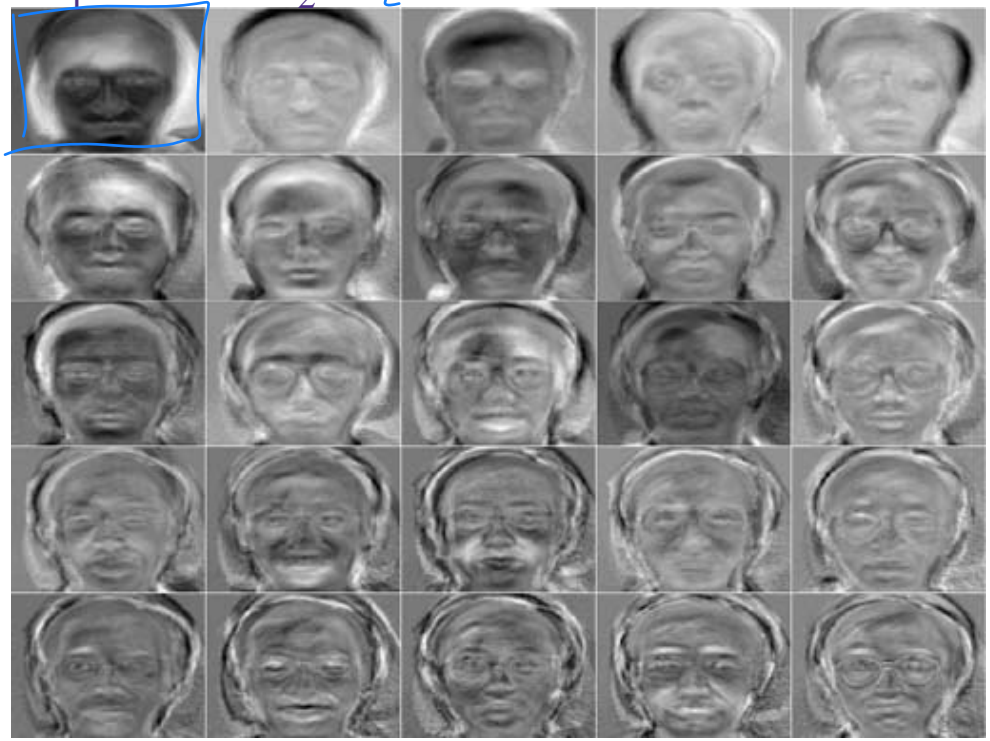
$$\approx z[1]u_1 + z[2]u_2 + \cdots + z[25]u_{25}$$

Principal component analysis finds a compact linear representation

- patterns that capture the distinct features of the samples is called **principal component** (to be formally defined later)

- we can represent each sample as a **weighted linear combination** of, say, q=25 principal components, and just store the weights

Principal components:



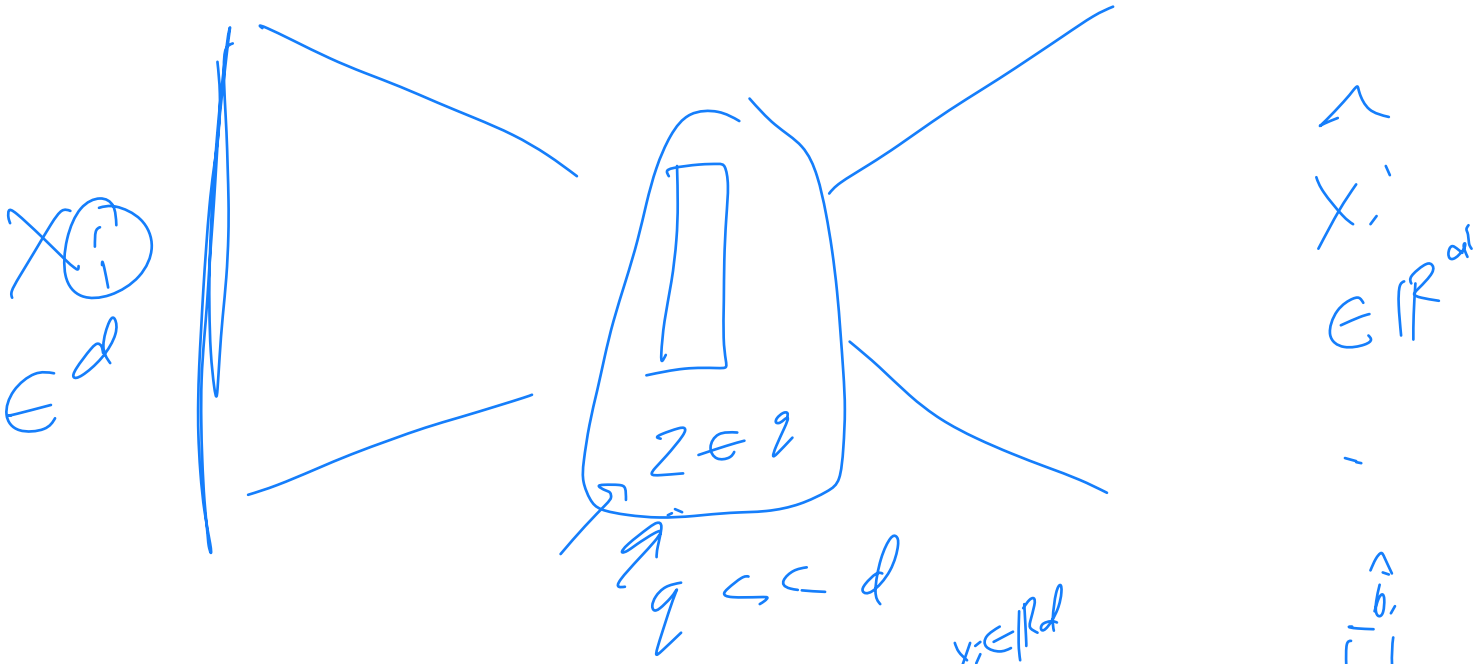
$$\approx z[1]u_1 + z[2]u_2 + \cdots + z[25]u_{25}$$

- With q=25, to store n images, it requires memory of only

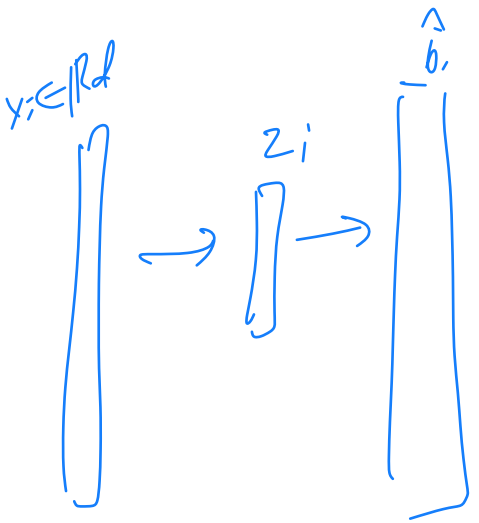
$$d \times q + q \times n \ll d \times n$$

auto encoder

→ learned version of PCA

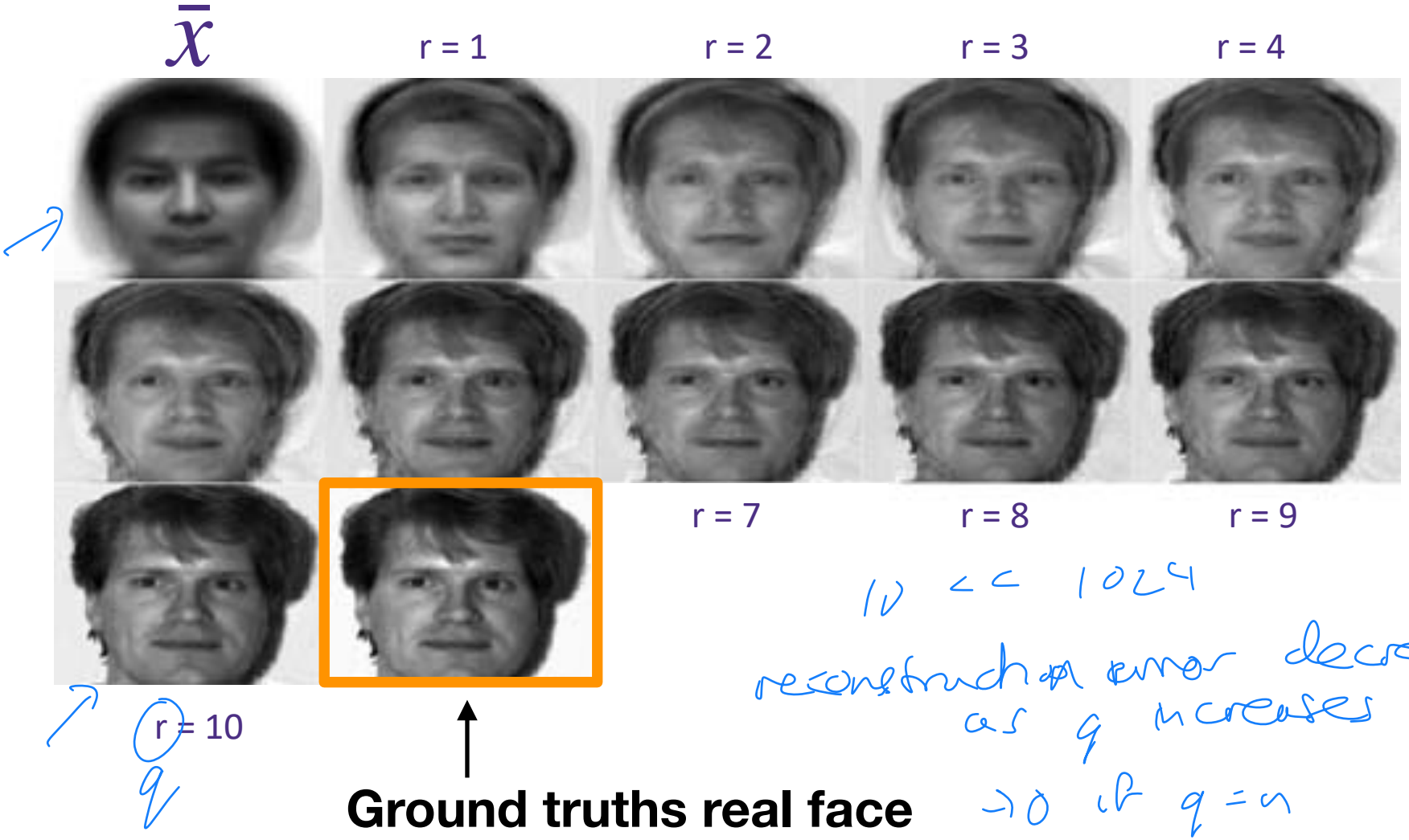


$x_i \to z_i$



10 principal components give a pretty good reconstruction of a face

average face $\bar{x} + a[1]u_1$ $\bar{x} + a[1]u_1 + a[2]u_2$



$N << 1024$

reconstruction error decreases as q increases

$\to 0$ if $q = n$

PCA: a high-fidelity linear projection

Given $x_1, ..., x_n \in \mathbb{R}^d$, find a compressed representation $z_1, ..., z_n \in \mathbb{R}^q$ with $q \ll d$ such that $x_i \approx \bar{x} + \mathbf{V}_q z_i$ and $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

$$\min_{\mathbf{V}_q, \{z_i\}} \sum_{i=1}^{n} ||x_i - \bar{x} - \mathbf{V}_q z_i||_2^2$$

coefficient on basis vectors

basis vectors

$V_q \in \mathbb{R}^{d \times q}$

each column is a basis face

orthonormal

all columns have unit norm and are mutually orthogonal

↓

no redundancy

↓

unique solution

$x_i \approx$ $\bar{x}$ + $\begin{bmatrix} | & & | \\ V_1 & \cdots & V_q \\ | & & | \end{bmatrix}$ $\begin{bmatrix} z_{i,1} \\ \vdots \\ z_{i,q} \end{bmatrix}$ $q \times 1$

$d \times q$

V_q z_i

$= \bar{x} + z_{i,1} V_1 + \cdots z_{i,q} V_q$

PCA: a high-fidelity linear projection

Given $x_1, ..., x_n \in \mathbb{R}^d$, find a compressed representation $z_1, ..., z_n \in \mathbb{R}^q$ with $q \ll d$ such that $x_i \approx \bar{x} + \mathbf{V}_q z_i$ and $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

reconstruction MSE

$$\min_{\mathbf{V}_q, \{z_i\}} \sum_{i=1}^{n} ||x_i - \bar{x} - \mathbf{V}_q z_i||_2^2$$

reconstructed x_i

→ how to solve?

$||y - Xw||_2^2$

$\hat{w} = (X^T X)^{-1} X^T y$

If you know V_q:

$z_i := (V_q^T V_q)^{-1} V_q (x_i - \bar{x})$

$= I$

$z_i = V_q^T (x_i - \bar{x})$

Fix $\mathbf{V}_q$ and solve for $\{z_i\}$:

PCA: a high-fidelity linear projection

Given $x_1, ..., x_n \in \mathbb{R}^d$, find a compressed representation $z_1, ..., z_n \in \mathbb{R}^q$ with $q \ll d$ such that $x_i \approx \bar{x} + \mathbf{V}_q z_i$ and $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

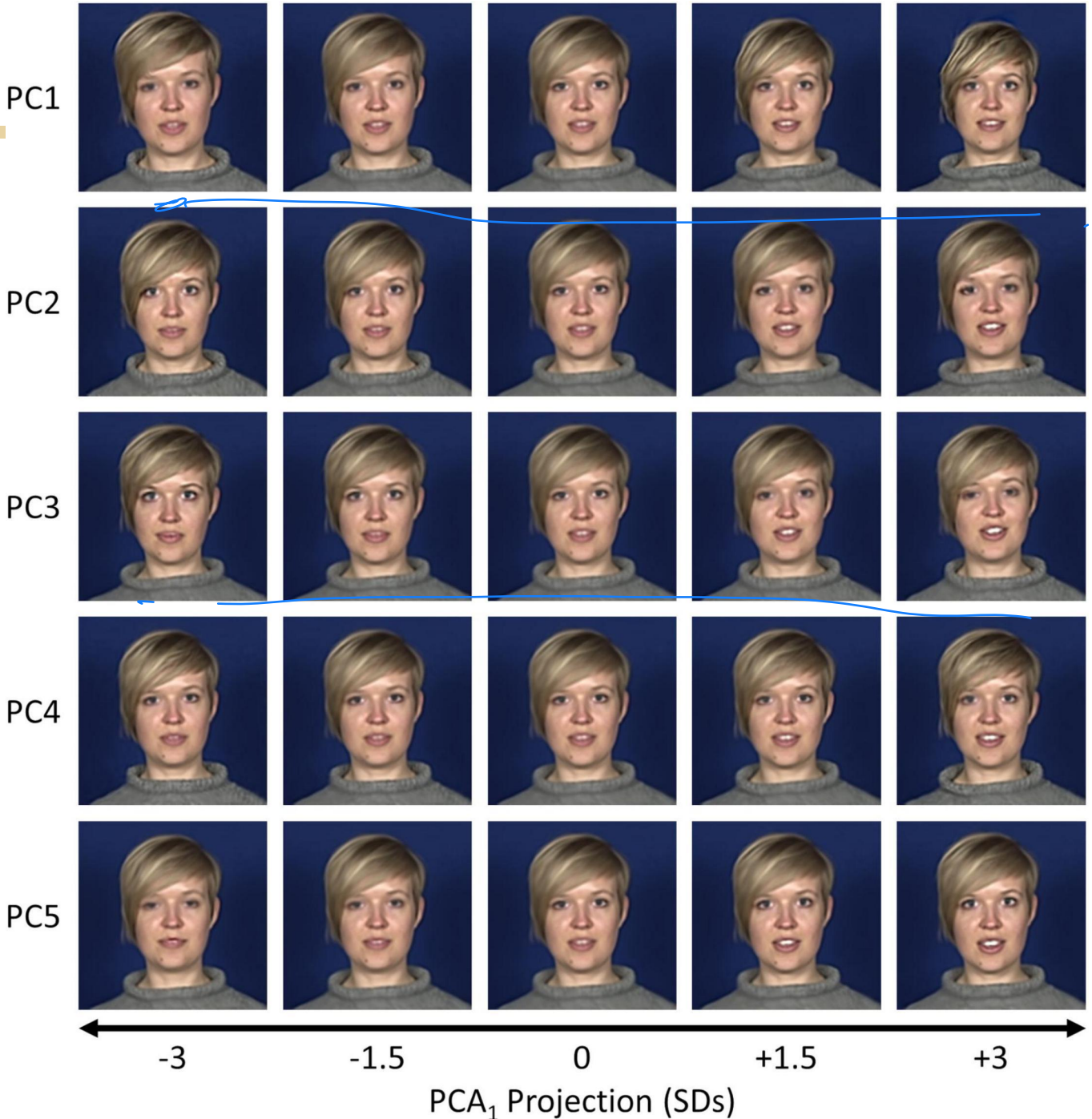
$$\min_{\mathbf{V}_q, \{z_i\}} \sum_{i=1}^{n} ||x_i - \bar{x} - \mathbf{V}_q z_i||_2^2$$

Fix $\mathbf{V}_q$ and solve for $\{z_i\}$: $z_i = \mathbf{V}_q^\top (x_i - \bar{x})$

→ projection matrix → projects onto q-dimensional linear subspace embedded into original high D space ($\mathbb{R}^d$)

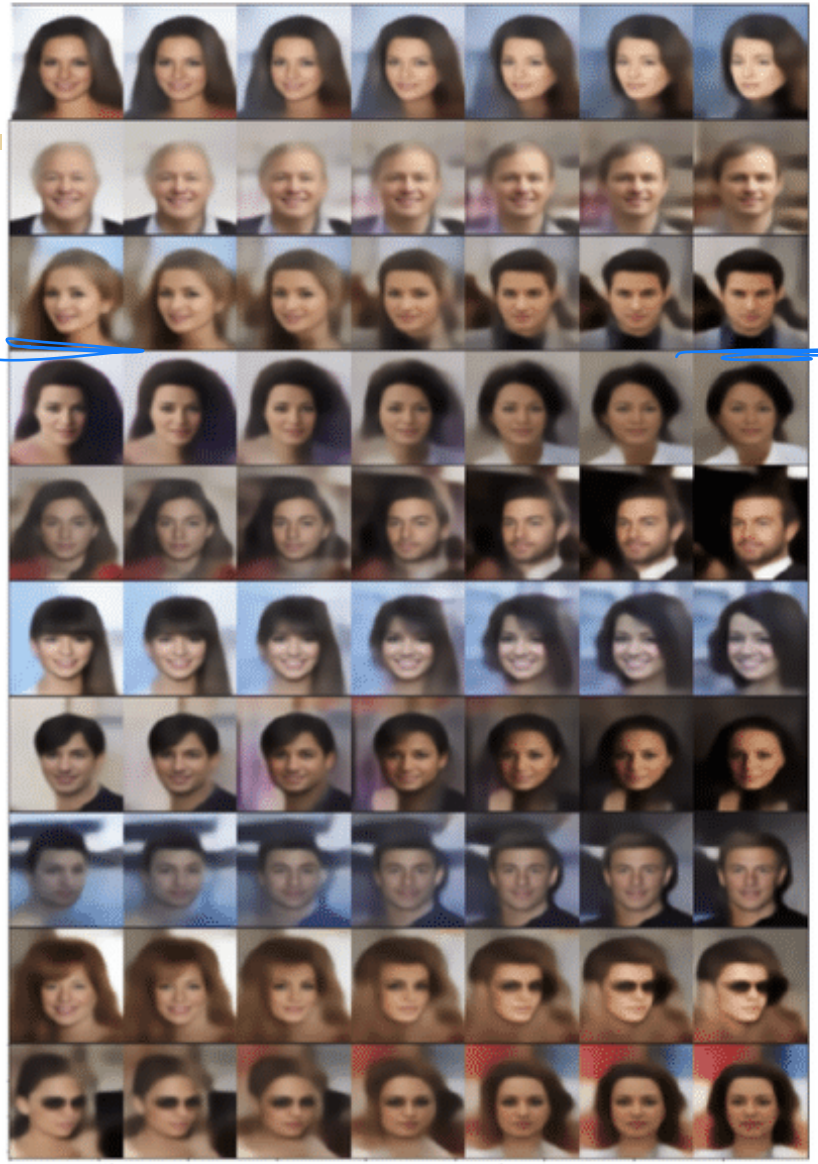
$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j v_j^\top (x_i - \bar{x})$$

z_i



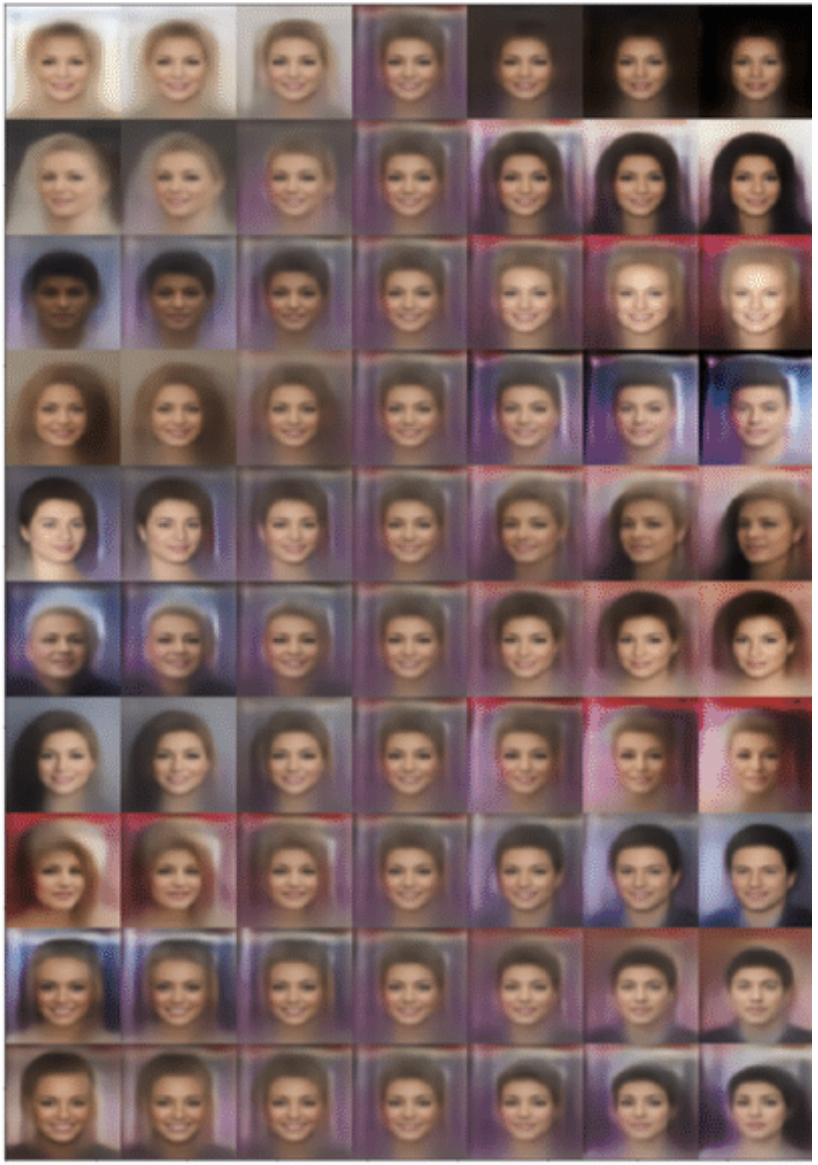
A PCA-Based Active Appearance Model for Characterising Modes of Spatiotemporal Variation in Dynamic Facial Behaviours
https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.880548/full

Sample linear interpolations in the encoding space



$\psi(\alpha\phi(x_1) + (1-\alpha)\phi(x_2)), \quad \alpha \in [0,1]$

PCA modes calculated in the encoding space

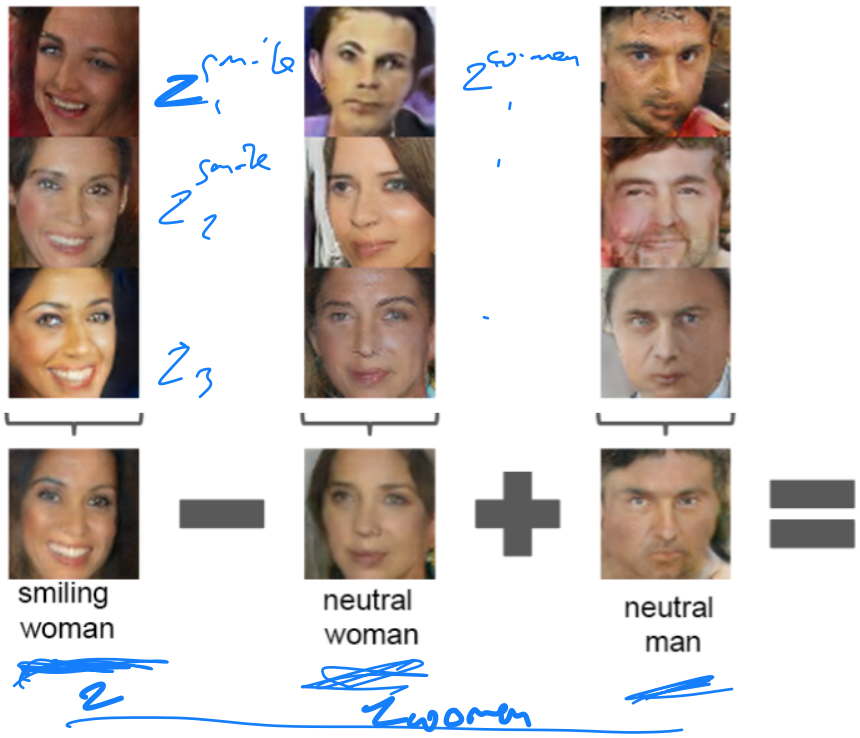


$-3\sigma \quad -2\sigma \quad -\sigma \quad 0 \quad \sigma \quad 2\sigma \quad 3\sigma$

Sliced Wasserstein Autoencoder: An Embarrassingly Simple Generative Model

https://www.researchgate.net/figure/The-results-of-SWAE-on-the-CelebA-face-dataset-with-a-128-dimensional-uniform_fig5_324246144

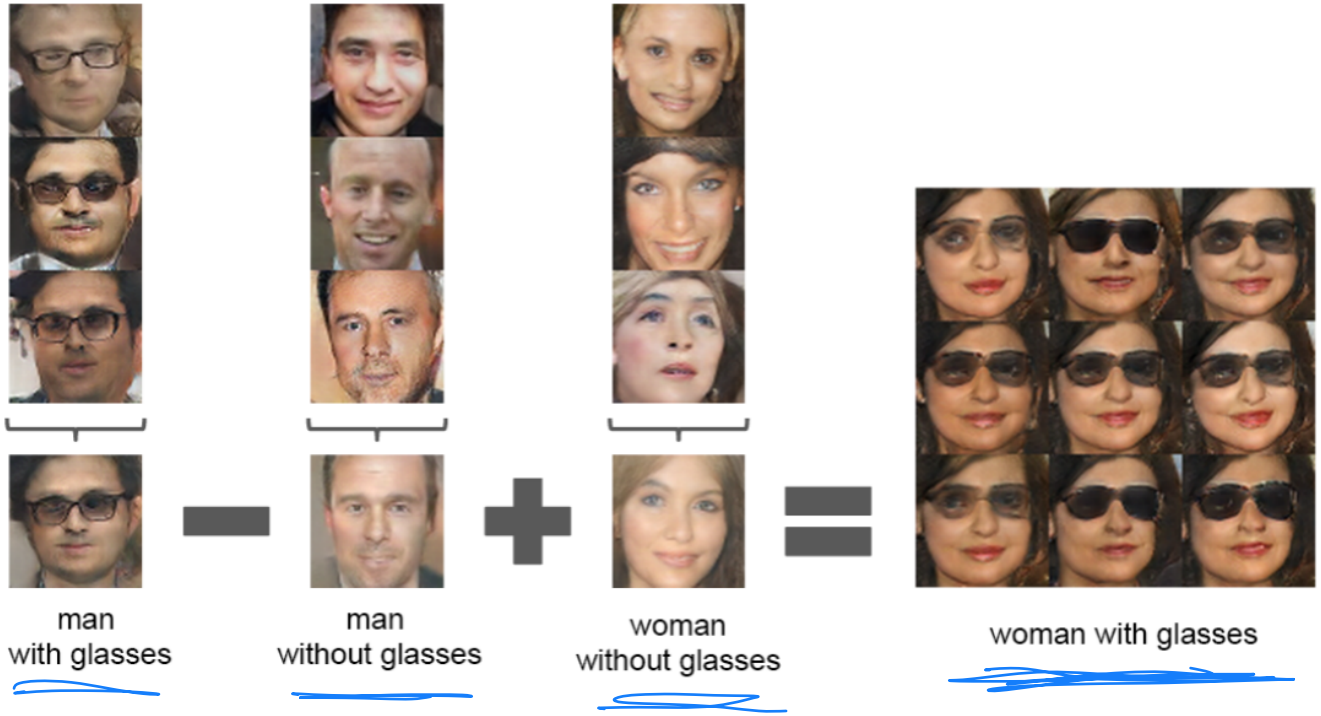
Representations of images are meaningful



Using same math on pixels only

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.

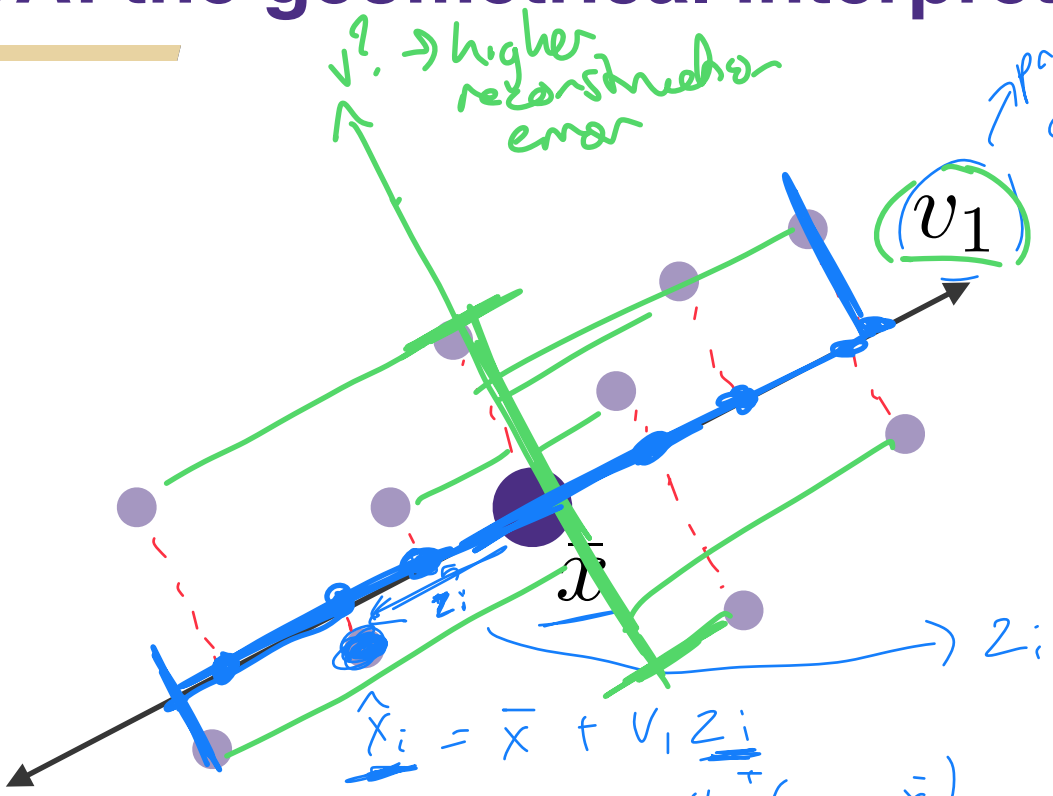
Representations of images are meaningful



Using same math on pixels only

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.

PCA: the geometrical interpretation



? → higher reconstruction error

principal component / basis vector

$d = 2$

$q = 1$

reconstruction error is the distance b/w the original point & projected version

$z_i = v_1^T (x_i - \bar{x})$

$\hat{x}_i = \bar{x} + v_1 z_i$

$= \bar{x} + v_1 v_1^T (x_i - \bar{x})$

Goal: orient the direction of v_1 to minimize squared reconstruction error

→ maximize the variance captured in the low-d representation

PCA: a high-fidelity linear projection

Given $x_1, ..., x_n \in \mathbb{R}^d$, find a compressed representation $z_1, ..., z_n \in \mathbb{R}^q$ with $q \ll d$ such that $x_i \approx \bar{x} + \mathbf{V}_q z_i$ and $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

$$\min_{\mathbf{V}_q, \{z_i\}} \sum_{i=1}^{n} ||x_i - \bar{x} - \mathbf{V}_q z_i||_2^2$$

Fix $\mathbf{V}_q$ and solve for $\{z_i\}$: $\quad z_i = \mathbf{V}_q^\top (x_i - \bar{x})$

$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j v_j^\top (x_i - \bar{x})$$

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \left|\left|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x})\right|\right|_2^2$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

PCA: a high-fidelity linear projection

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \left|\left|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x})\right|\right|_2^2$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j v_j^\top (x_i - \bar{x})$$

$V_q^T V_q = I$ orthonormal

Case when $q = 1$ → principal component

$$v_1 = \arg \min_{v: \|v\|_2 = 1} \sum_{i=1}^{N} \|(x_i - \bar{x}) - vv^\top (x_i - \bar{x})\|_2^2$$

a b

$\|a-b\|_2^2 = (a-b)^T(a-b)$

$= a^Ta - 2a^Tb + b^Tb$

$$v_1 = \arg\min_{v:\|v\|_2=1} \sum_{i=1}^{n} \|x_i - \bar{x}\|_2^2 - 2(x_i - \bar{x})^T v v^T (x_i - \bar{x}) + (x_i - \bar{x})^T v \underbrace{v^T v}_{I} v^T (x_i - \bar{x})$$

does not depend on v — same — same

$$= \arg\min - \sum_{i=1}^{n} (x_i - \bar{x})^T v v^T (x_i - \bar{x})$$

$$= \arg\max \sum_{i=1}^{n} \underbrace{(x_i - \bar{x})^T v}_{z_i} \underbrace{v^T (x_i - \bar{x})}_{z_i^T} = \sum_{i=1}^{n} z_i z_i^T$$

Show max $\sum z_i z_i^T$ ends up maximizing the variance!

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} v^T (x_i - \bar{x}) \quad \to 0$$

$$= 0$$

$$\arg\max_v \sum_{i=1}^{n} z_i z_i^T = \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^T$$

n * empirical variance of z_i

Minimizing reconstruction error is equivalent to maximizing the variance of the projected data points

① $\underset{v}{\operatorname{argmax}} \sum_{i=1}^{n} \underbrace{\left((x_i - \bar{x})^T v\right)^T \left(v^T (x_i - \bar{x})\right)^T}_{z_i}$

$z_i \in \mathbb{R}^1$ → scalar

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

// take v outside sum

see ridge regression

$\underset{v}{\operatorname{argmax}} \sum_{i=1}^{n} v^T (x_i - \bar{x})(x_i - \bar{x})^T v$

$\underset{v}{\operatorname{argmax}}\ v^T \underbrace{\left(\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T\right)}_{d \times d} v$ → define to be $\Sigma \in \mathbb{R}^{d \times d}$ covariance matrix

$= \underset{v}{\operatorname{argmax}}\ v^T \Sigma v$ → find the direction that maximizes the project variance

consider $v = e_j = [0\ 0\ 0\ \underset{j}{1}\ 0\ 0\ 0]$

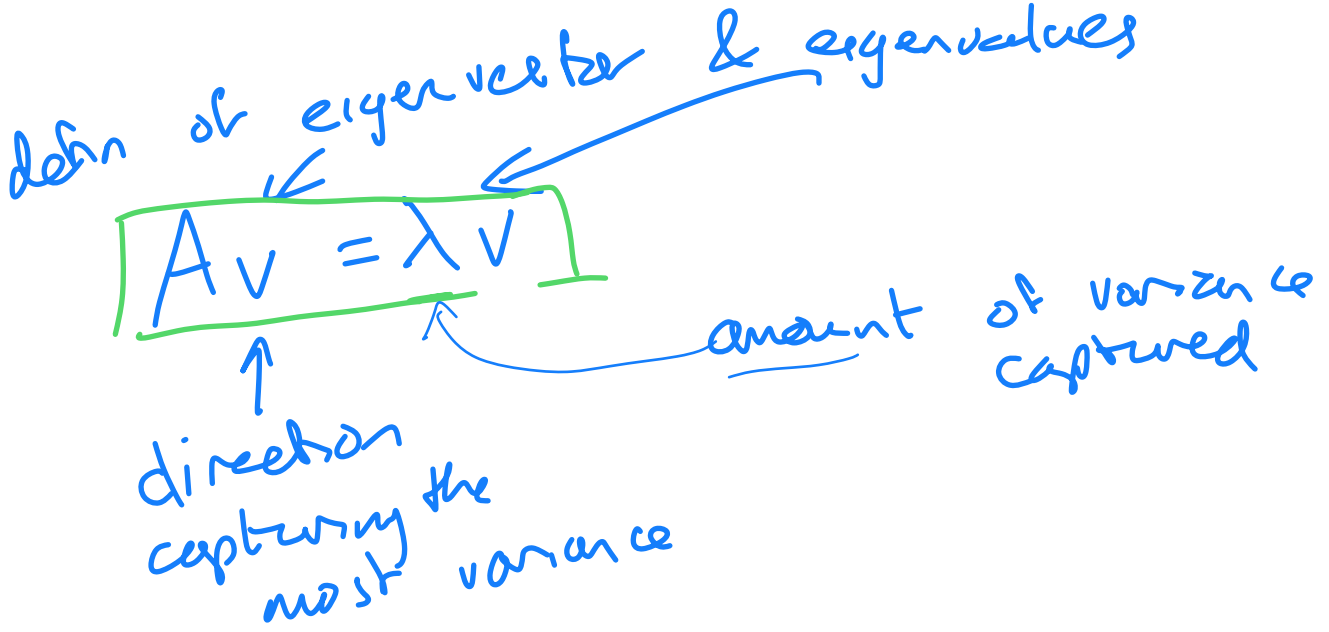
$e_j^T \Sigma e_j = ?$

sample variance of the j^{th} feature

i.e. maximizing the variance in direction e_j

→ not in practice a good idea

☆ Solution / Punchline:
choose v to be the leading eigenvectors of Σ
↳ with highest eigenvalues λ



For a matrix $A \in \mathbb{R}^{d \times d}$ we say (λ, v) is (eigenvalue, eigenvector) pair if $\boxed{Av = \lambda v}$

Fact: if A is symmetric, then all of its eigenvalues are real and

$$A = \sum_{i=1}^{d} \lambda_i v_i v_i^T \quad \text{where } v_i^T v_j = 0 \ \forall i \neq j$$

$$\boxed{= V \Lambda V^T}$$

↳ diagonal matrix where elements are eigenvalues λ

if A is PSD all eigenvalues are non-negative.

$$Av_i = \lambda_i v_i$$

$$v_i^T A v_i = \lambda_i v_i^T v_i \rightarrow I \text{ by orthonormal constraint}$$

$$\boxed{v_i^T A v_i} = \lambda_i$$

if A is a covariance matrix Σ
then it's symmetric & PSD

and if $\|v_i\|_2 = 1$

then $v_i^T A v_i$ is the variance along direction v_i (which is λ_i)

variance should be real & non-negative.

Find structure in your data with basic linear algebra

PCA: a high-fidelity linear projection

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \left|\left|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x})\right|\right|_2^2$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j \langle v_j, x_i - \bar{x} \rangle$$

Case when $q = 1$

$$v_1 = \arg \min_{v:\|v\|_2=1} \sum_{i=1}^{N} \|(x_i - \bar{x}) - vv^\top (x_i - \bar{x})\|_2^2$$

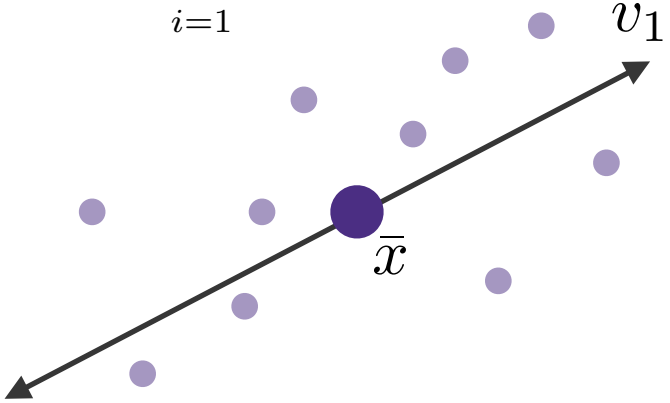
$$= \arg \min_{v:\|v\|_2=1} \sum_{i=1}^{N} \|x_i - \bar{x}\|_2^2 - 2(x_i - \bar{x})^\top vv^\top (x_i - \bar{x}) + (x_i - \bar{x})^\top vv^\top vv^\top (x_i - \bar{x})$$

$$= \arg \min_{v:\|v\|_2=1} \sum_{i=1}^{N} \|x_i - \bar{x}\|_2^2 - \sum_{i=1}^{N} (x_i - \bar{x})^\top vv^\top (x_i - \bar{x})$$

$$= \arg \max_{v:\|v\|_2=1} \sum_{i=1}^{N} (x_i - \bar{x})^\top vv^\top (x_i - \bar{x})$$

$$= \arg \max_{v:\|v\|_2=1} v^\top \Sigma v$$

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$



PCA: a high-fidelity linear projection

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \left|\left|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x})\right|\right|_2^2$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

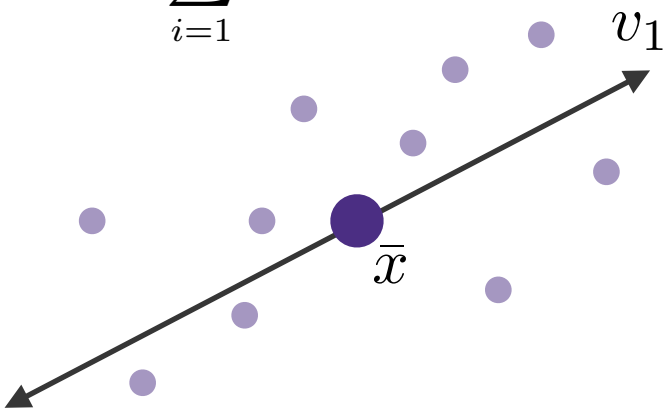
$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j \langle v_j, x_i - \bar{x} \rangle$$

Case when $q = 1$

$$v_1 = \arg \min_{v:\|v\|_2=1} \sum_{i=1}^{N} \|(x_i - \bar{x}) - vv^\top (x_i - \bar{x})\|_2^2$$

$$= \arg \max_{v:\|v\|_2=1} v^\top \Sigma v$$

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$



PCA: a high-fidelity linear projection

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \left|\left|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x})\right|\right|_2^2$$

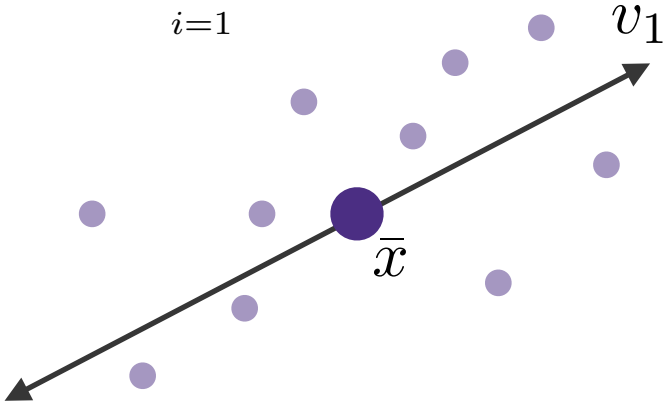
$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j \langle v_j, x_i - \bar{x} \rangle$$

General $q \geq 1$

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2 = \min_{\mathbf{V}_q} Tr(\Sigma) - Tr(\mathbf{V}_q^T \Sigma \mathbf{V}_q)$$

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$



$Tr(ABC) = Tr(CAB)$
$= Tr(BCA)$
by cyclic where you swapped order of 2 scalars

$\mathbf{V}_q$ are the first q eigenvectors of Σ

Minimize reconstruction error = capture the most variance in your data.

How to choose the dimensionality, q

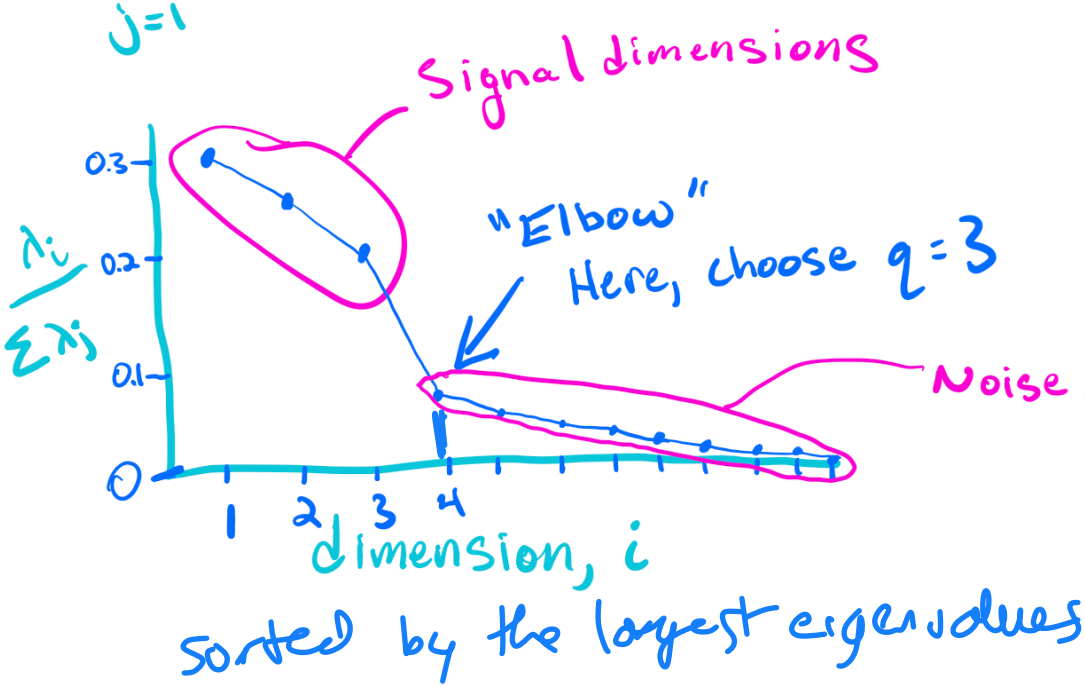
HOW TO CHOOSE q

model complexity

CROSS VALIDATION DOESN'T WORK

- More dimensions always increases projected variance (decreases reconstruction error), INCLUDING ON VAL DATA.

$$\frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} = \frac{\text{variance along } v_i}{\text{total variance}}$$



- Ad-hoc approach:
  → # dims needed to explain 95% of variance.
- Leave-one-feature-out cross-validation. (LOFO-CV)

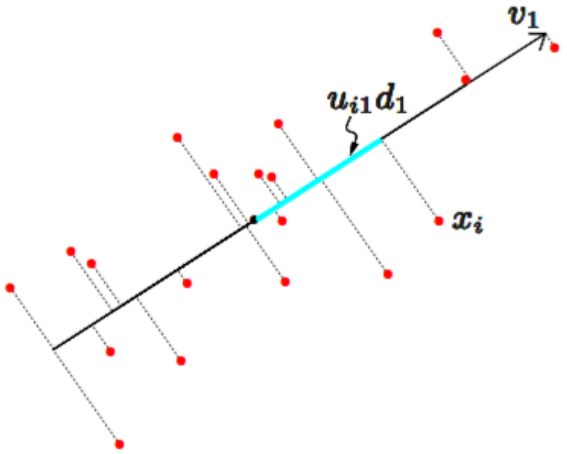
PCA: a high-fidelity linear projection

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$

where $\mathbf{V}_q = [v_1, v_2, \ldots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$



$\mathbf{V}_q$ are the first q eigenvectors of Σ

$\mathbf{V}_q$ are the first q *principal components*

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

Principal Component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x}^T)$ down onto $\mathbf{V}_q$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U}_q \mathrm{diag}(d_1, \ldots, d_q) \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$