Principal Component Analysis

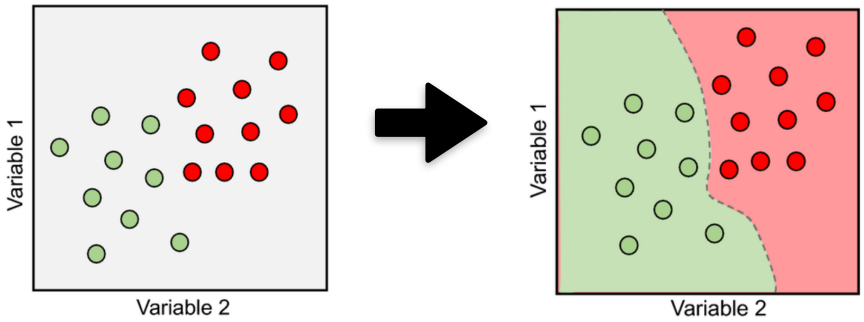
Natasha Jaques

UNIVERSITY of WASHINGTON

Unsupervised vs. supervised learning

Previously: supervised learning

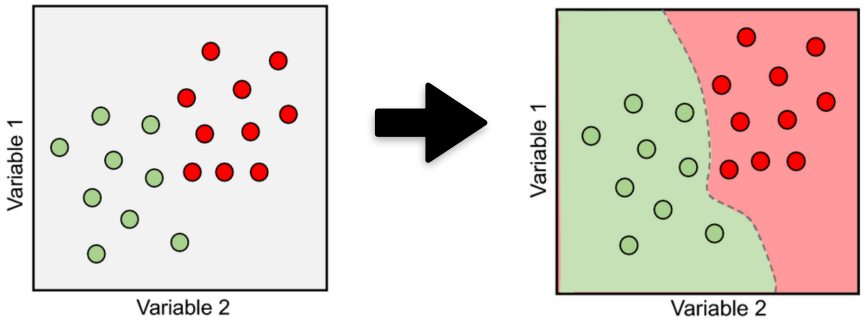
- Each data point x_i has a corresponding label y_i;$\{x_i, y_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. Try to predict the label y for a new test point x



Unsupervised vs. supervised learning

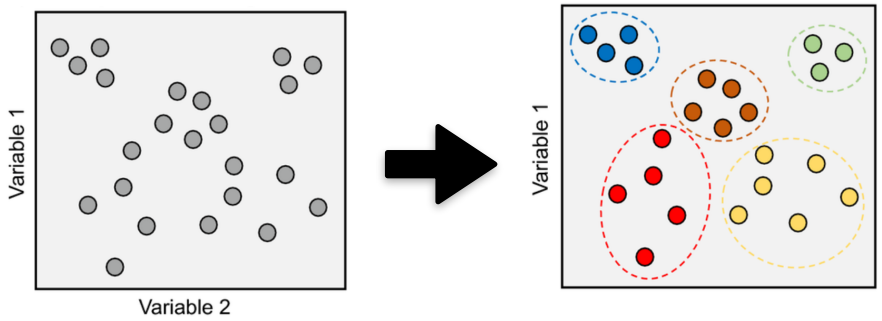
Previously: supervised learning

- 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



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

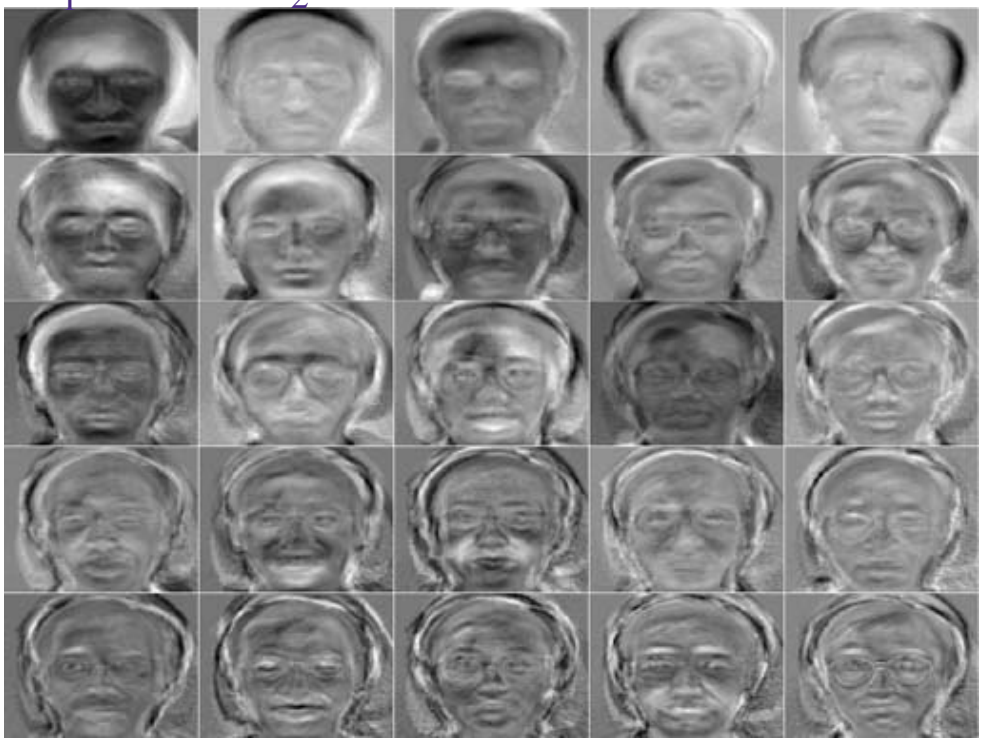
$d \times n$ real values to store the data

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

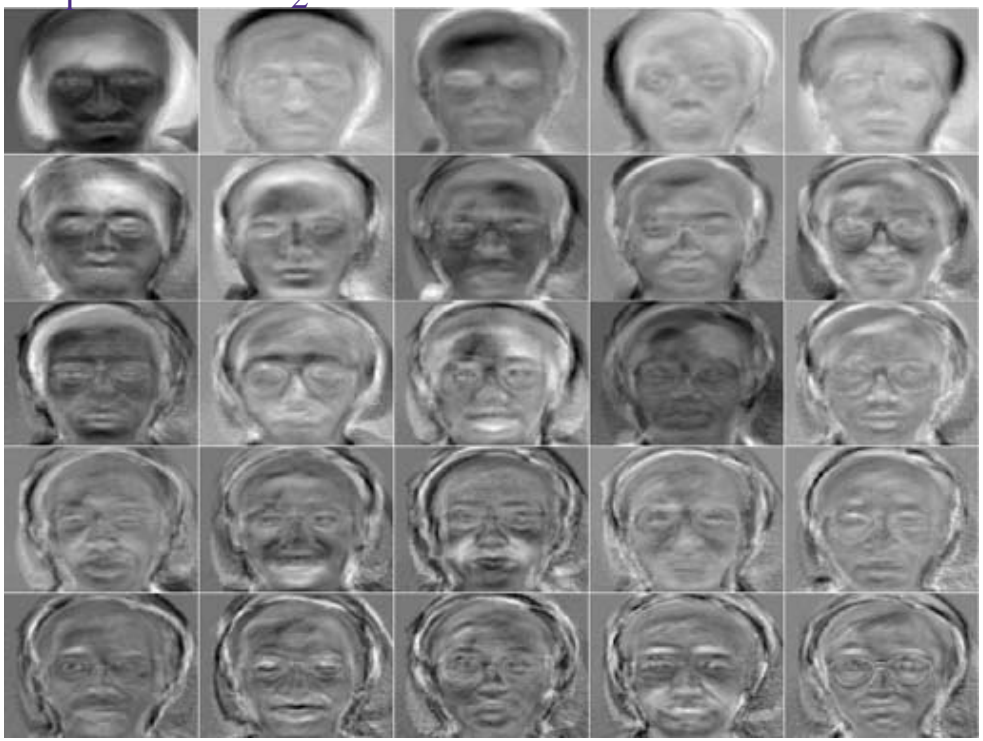
- patterns that capture the distinct features of the samples is called **principal component** (to be formally defined later)

face

- 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



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

each u is a basis face

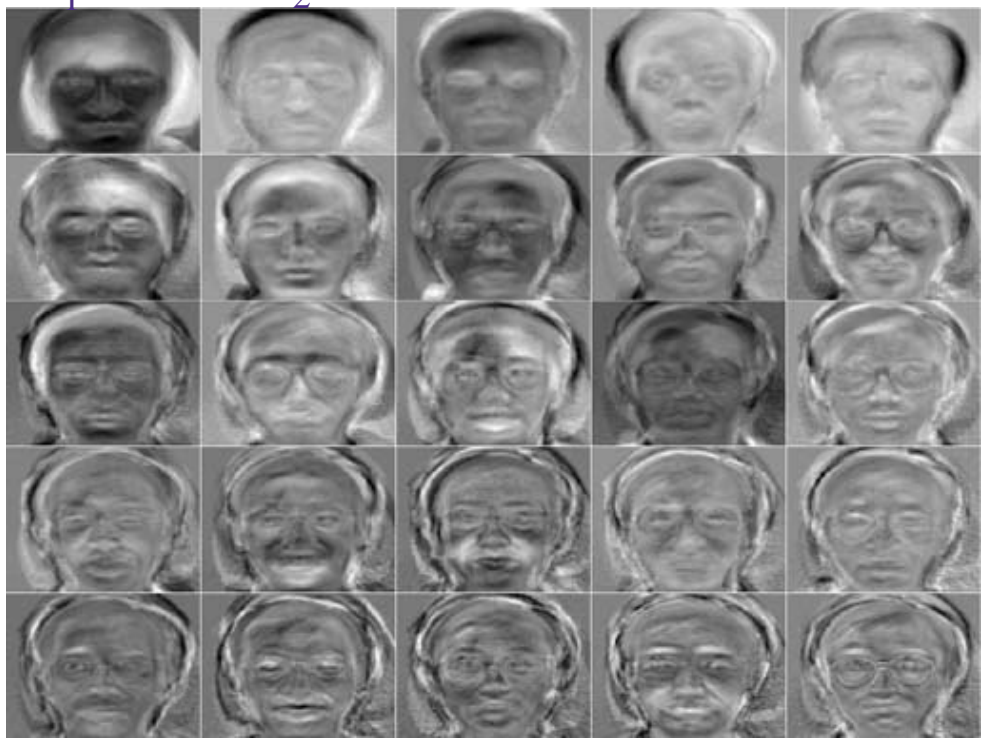
Principal component analysis finds a compact linear representation

- 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



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

- With q=25, to store n images, it requires memory of only $d \times q + q \times n \ll d \times n$

10 principal components give a pretty good reconstruction of a face

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

10 << 1024

reconstruction error decreases as q increases; 0 when?

q = n

Ground truths real face

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 $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

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

z_i = coefficient on basis vectors $z_i \in \mathbb{R}^{q \times 1}$

Orthonormal constraint: what is it?

all columns have unit norm and are mutually orthogonal.

Why?

No redundancy. Unique sol'n.

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

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

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 $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

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

How to solve?

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

Recall: $\left\|y - Xw\right\|_2^2$

$$\hat{w} = \left(X^\top X\right)^{-1} X^\top y$$

So if you know V_q $z_i = \left(V_q^\top V_q\right)^{-1} V_q^\top \left(x_i - \bar{x}\right)$

With orthonormal assumption…

$$z_i = V_q^\top \left(x_i - \bar{x}\right)$$

PCA: a high-fidelity linear projection

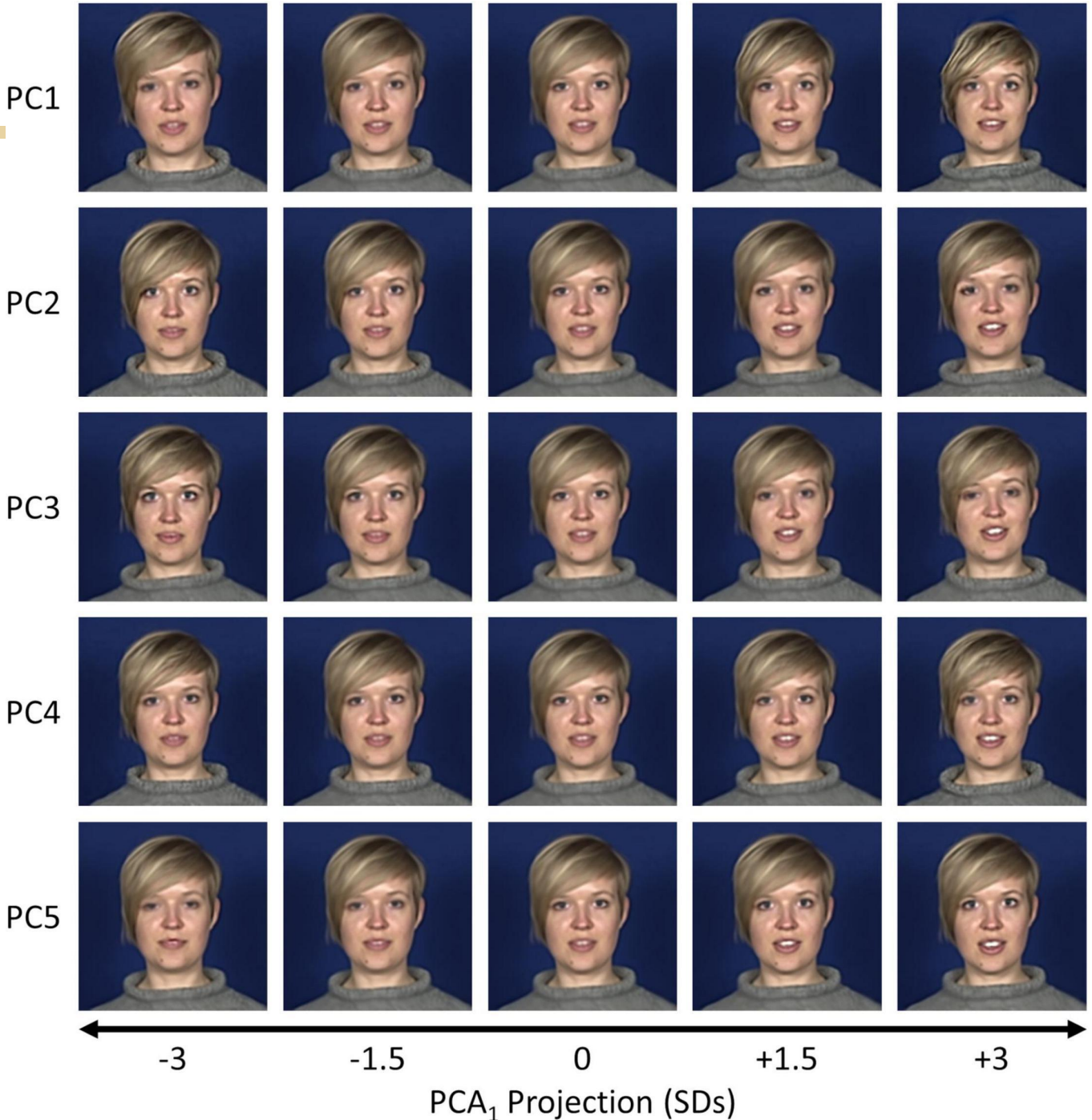
Given $x_1, ..., x_n \in \mathbb{R}^d$, find a compressed representation $z_1, ..., z_n \in \mathbb{R}^q$ with $q \ll d$ such that $x_i \approx \bar{x} + \mathbf{V}_q z_i$ and $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

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

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

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

Think of $V_q V_q^\top$ as a projection matrix, that projects data onto a q-dimensional linear subspace embedded into original higher-dimensional (d-dim) space

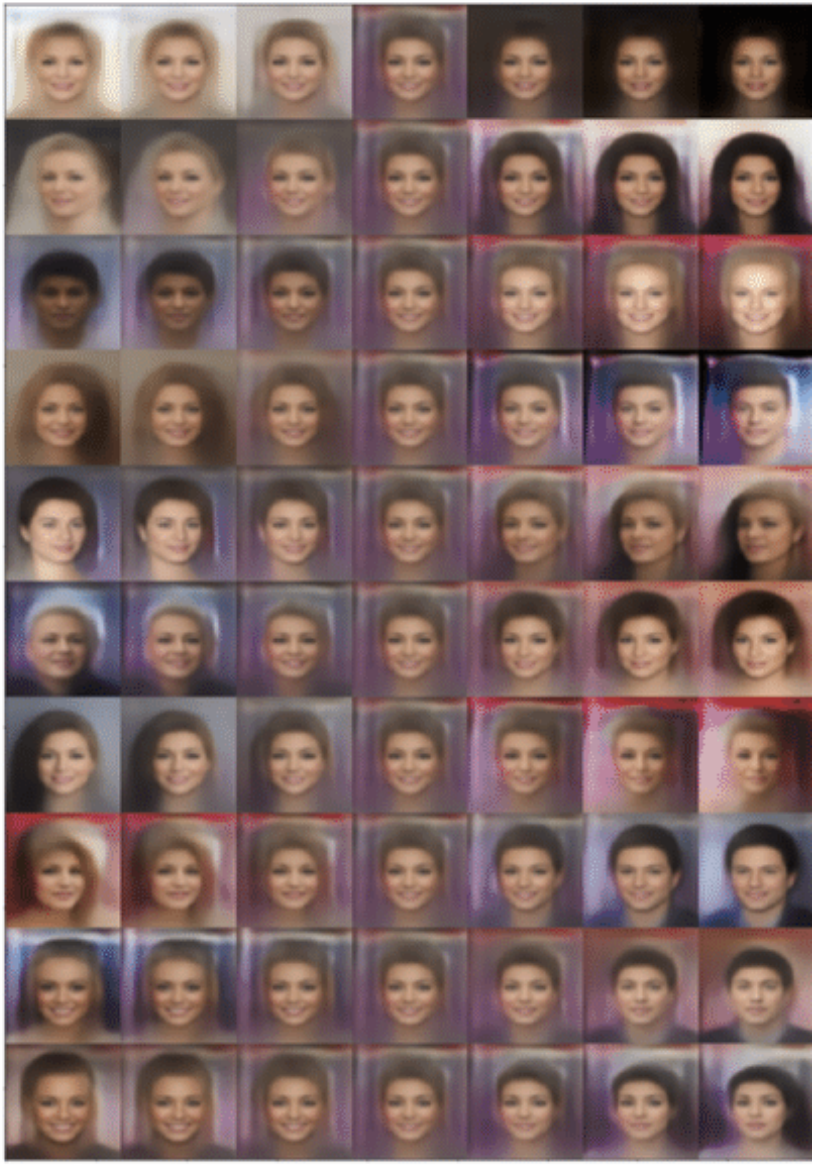


A PCA-Based Active Appearance Model for Characterising Modes of Spatiotemporal Variation in Dynamic Facial Behaviours
https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.880548/full

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

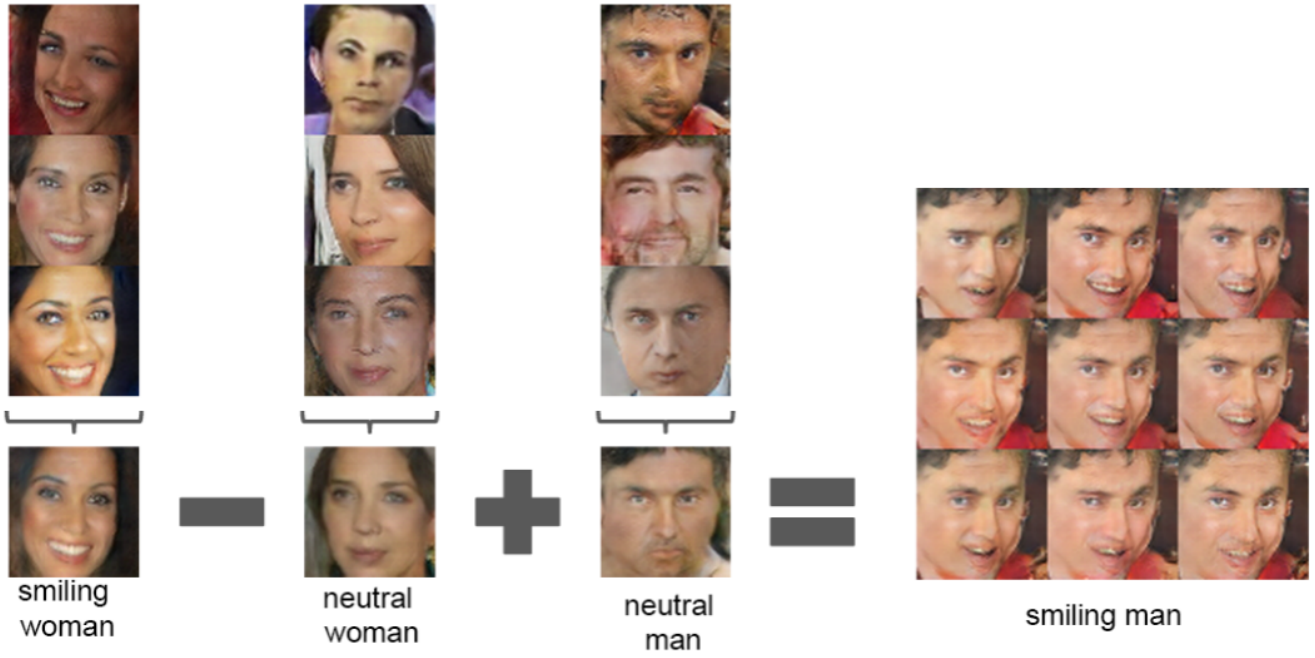


$-3\sigma \quad -2\sigma \quad -\sigma \quad 0 \quad \sigma \quad 2\sigma \quad 3\sigma$

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

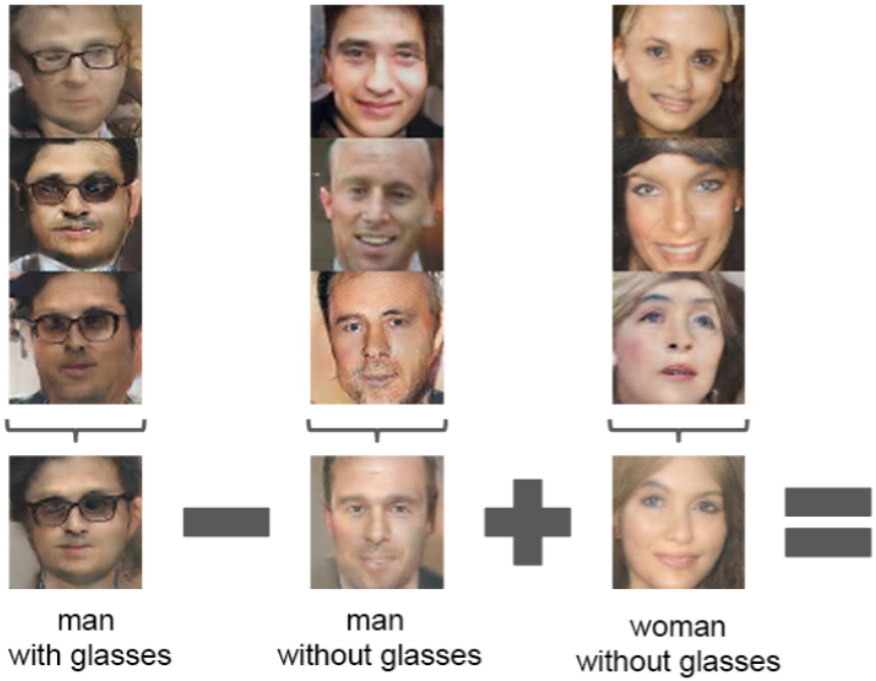
Representations of images are meaningful



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.

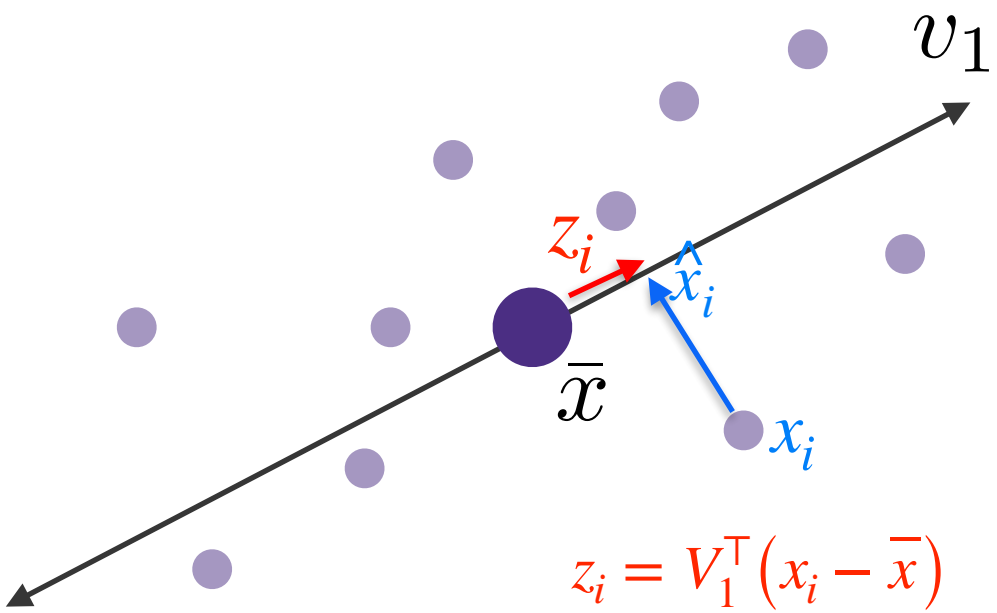
Representations of images are meaningful



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



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

Assume d=2, q=1

v_1 is the principal component / basis vector

Project each point onto the principal component

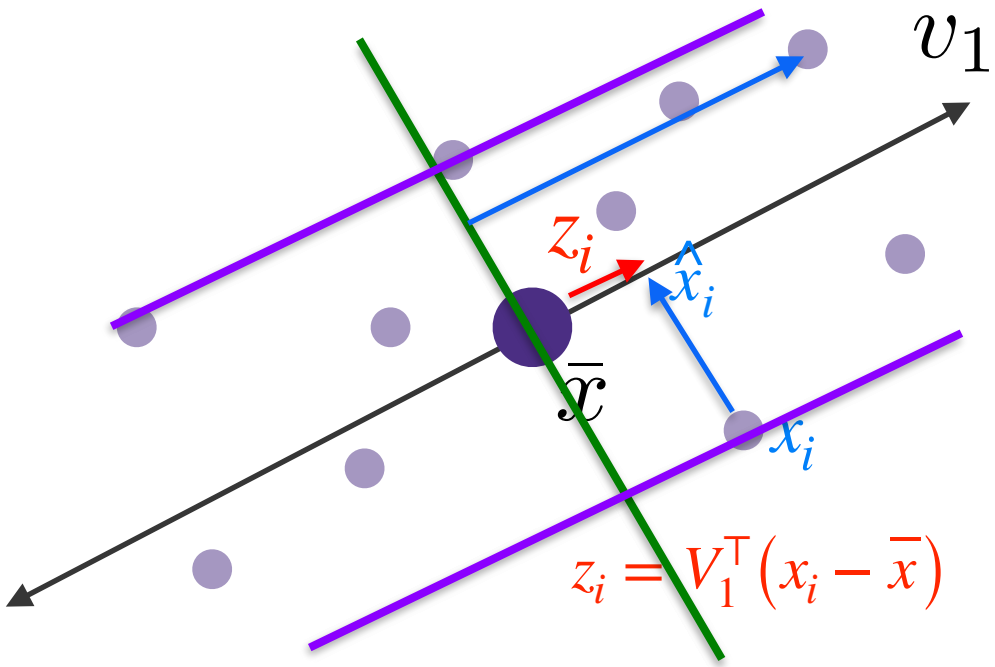
$$\hat{x}_i = \bar{x} + V_1 z_i$$

$$\hat{x}_i = \bar{x} + V_1 V_1^\top \left(x_i - \bar{x} \right)$$

Reconstruction error is the distance between the original point and the projected version

$$\left\| x_i - \hat{x}_i \right\|_2^2$$

PCA: the geometrical interpretation



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

Maximize the variance captured in the low-d representation

Assume d=2, q=1

v_1 is the principal component / basis vector

Project each point onto the principal component

$$\hat{x}_i = \bar{x} + V_1 z_i$$

$$\hat{x}_i = \bar{x} + V_1 V_1^\top \left(x_i - \bar{x} \right)$$

If v1 were oriented this way, not only would reconstruction errors be higher…

The spread of the points of the points when projected on the basis vector would be smaller

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

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

$V_q^\top V_q = I$

v1 = 1 principal component

Case when $q = 1$

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

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

No v

$$v_1 = \arg \min_{v:\|v\|_2=1} \cancel{\sum_{i=1}^{n} \|x_i - \bar{x}\|_2^2} - \underbrace{2(x_i - \bar{x})^\top vv^\top (x_i - \bar{x})}_{\text{Same}} + \underbrace{(x_i - \bar{x})^\top vv^\top vv^\top (x_i - \bar{x})}_{\text{Same}}$$

= I

$$= \arg\min \; - \sum_{i=1}^{n} (x_i - \bar{x})^\top vv^\top (x_i - \bar{x})$$

$$= \arg\max \sum_{i=1}^{n} (x_i - \bar{x})^\top vv^\top (x_i - \bar{x}) = \sum_{i=1}^{n} z_i z_i^\top$$

$$z_i = \mathbf{V}_q^\top (x_i - \bar{x})$$

Maximizing $z_i z_i^\top$ → Maximizing variance

$$= \arg\max \sum_{i=1}^{n} (x_i - \bar{x})^\top v v^\top \left(x_i - \bar{x} \right) = \sum_{i=1}^{n} z_i z_i^\top$$

The mean of all z_i's is 0 (Recall z_i is distance from mean of x)

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i \quad = \frac{1}{n} \sum_{i=1}^{n} v^\top \left(x_i - \bar{x} \right) \quad = v^\top \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)$$

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

So we can freely add it to our prev equation:

$$\underset{v}{\operatorname{argmax}} \sum_{i=1}^{n} z_i z_i^\top = \sum_{i=1}^{n} \left(z_i - \bar{z} \right) (z_i - \bar{z})^\top$$

This is equivalent to maximizing empirical variance!

Minimizing reconstruction error is equivalent to maximizing the variance captured in the low-d representation

So how do we actually solve for v?

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

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

$$\underset{v}{\text{argmax}}\ v^\top \underbrace{\left(\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top \right)}_{\text{d x d}} v$$

$$= \underset{v}{\text{argmax}}\ v^\top \Sigma v$$

$z_i \in \mathbb{R}^1 \rightarrow$ scalar

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

Take v outside sum (similar to ridge regression derivation)

Define this term as the *covariance matrix:* $\Sigma \in \mathbb{R}^{d \times d}$

 # Find the direction that maximizes the projected variance

Does this look familiar to any linear algebra whizzes in the crowd?

So how do we actually solve for v?

$$= \underset{v}{\operatorname{argmax}}\ v^\top \Sigma v$$

Solution: choose v to be the leading eigenvectors of Σ (those with the highest eigenvalues λ

Defn: Eigenvector and Eigenvalue

$$Av = \lambda v$$

Amount of variance captured

Direction capturing the most variance

Eigenvalues and Eigenvectors

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

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

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

$$= V \Lambda V^\top$$

Λ is a diagonal matrix where elements are eigenvalues λ

If A is PSD (Positive Semi-Definite), all eigenvalues are non-negative

$$Av_i = \lambda_i v_i$$

$$v_i^\top A v_i = \lambda_i v_i^\top v_i$$

$$v_i^\top A v_i = \lambda_i$$

By orthonormal constraint

Bringing this back to PCA

So, if A is a covariance matrix Σ, then it is symmetric and PSD.

And we assume. $\|v_i\|_2 = 1$ ….

Then $v_i^\top A v_i = \lambda_i$ is the variance along direction v_i (which is λ_i). Variance is real and non-negative.

So to compute PCA in practice:

- Compute the eigendecomposition of the covariance matrix Σ
- The top q eigenvectors (with the largest eigenvalues) are the first q principal components
- Each eigenvalue λ_i tells you how much variance is captured along its direction

Takeaway: find important 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 $q = 1$

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

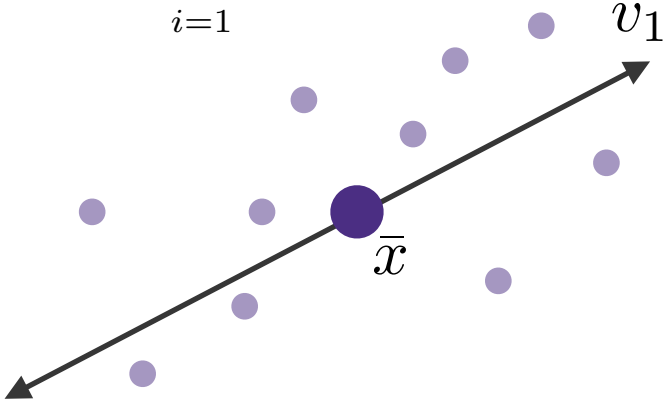
$$= \arg \min_{v:\|v\|_2=1} \sum_{i=1}^{N} \|x_i - \bar{x}\|_2^2 - 2(x_i - \bar{x})^\top vv^\top (x_i - \bar{x}) + (x_i - \bar{x})^\top vv^\top vv^\top (x_i - \bar{x})$$

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

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

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

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



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

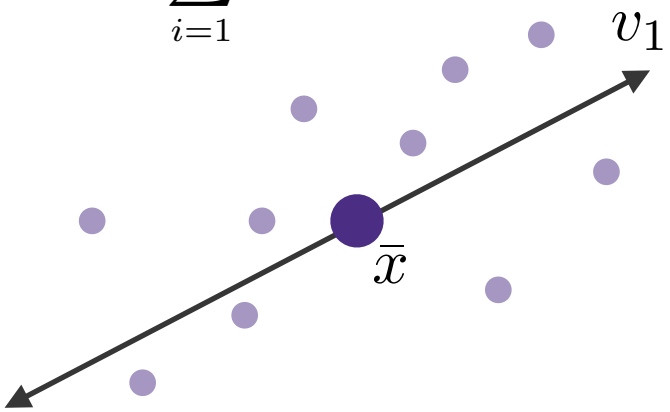
$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j \langle v_j, x_i - \bar{x} \rangle$$

Case when $q = 1$

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

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

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



PCA: a high-fidelity linear projection

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

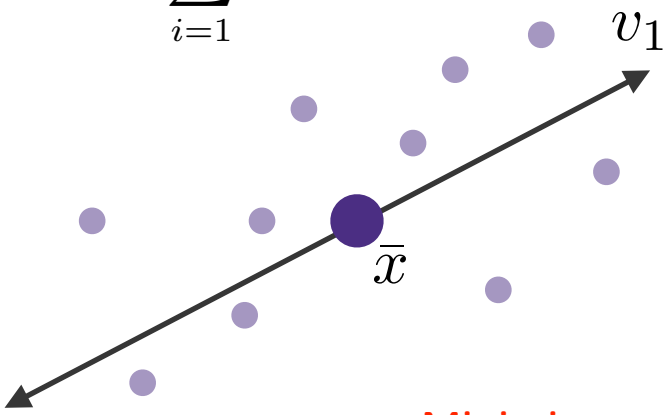
$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

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

General $q \geq 1$

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

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



For the general case, use the trace trick where we previously used the scalar transpose trick

Tr(ABC) = Tr(CAB) = Tr(BCA)

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

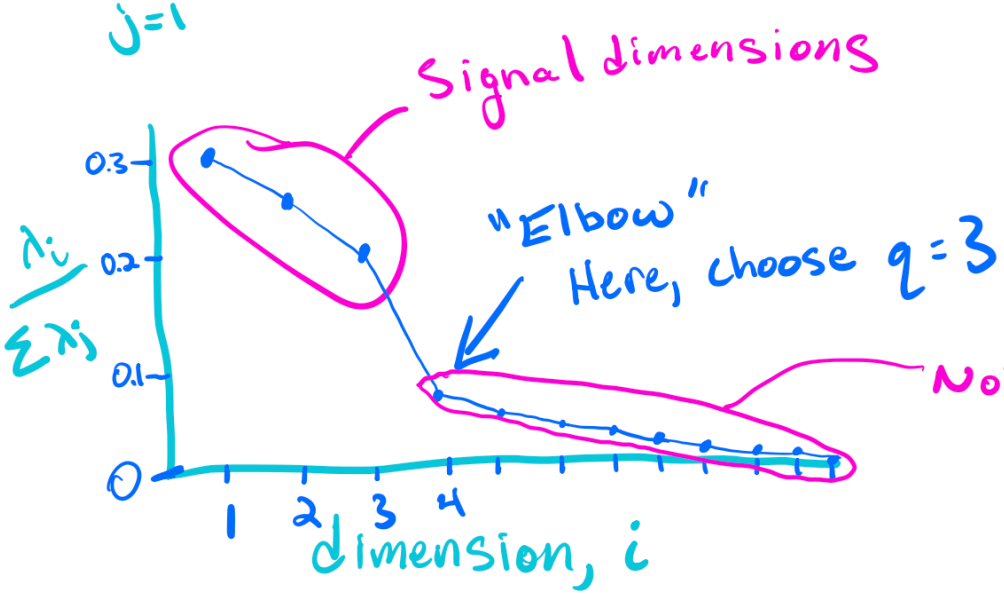
Minimize reconstruction error = capture the most variance in your data.

How to choose the dimensionality, *q*

q is like your model complexity

HOW TO CHOOSE q

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



CROSS VALIDATION DOESN'T WORK

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

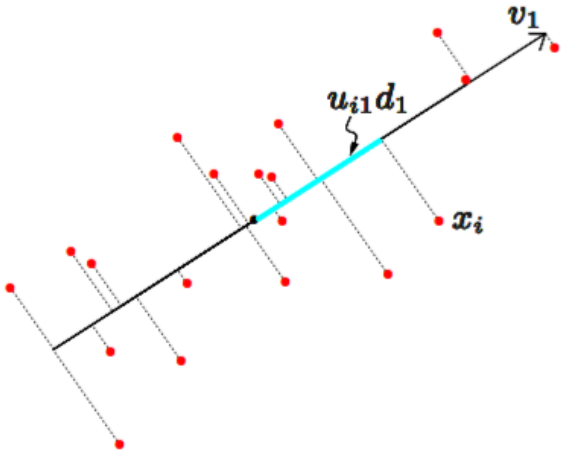
- Ad-hoc approach: # dims needed to explain 95% of variance.
- Leave-one-feature-out cross-validation. (LOFO-CV)

here dims are sorted by largest eigenvalues

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 $\mathbf{V}_q$

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

Singular Value Decomposition (SVD)

SVD is a general technique for finding right singular values of any matrix

$\tilde{X} \in \mathbb{R}^{n \times d}$ as $\tilde{X} = USV^{\top}$

Doesn't have to be square

Works for any matrix, always exists

- U,V are orthogonal (rotations); S is diagonal with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$
- Squared singular values σ_i^2 = eigenvalues of $\tilde{X}^{\top}\tilde{X}$
- Right singular vectors (columns of V) are eigenvectors of $\tilde{X}^{\top}\tilde{X}$

In practice, use SVD for PCA, since you can do it directly with $\tilde{X}^{\top}\tilde{X}$ rather than having to compute Σ

More efficient since smaller than full Σ

Better numerical stability

PCA is solved using SVD

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

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

Minimizing the reconstruction error for a compressed representation of the data

Maximum eigen-space of Σ

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

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

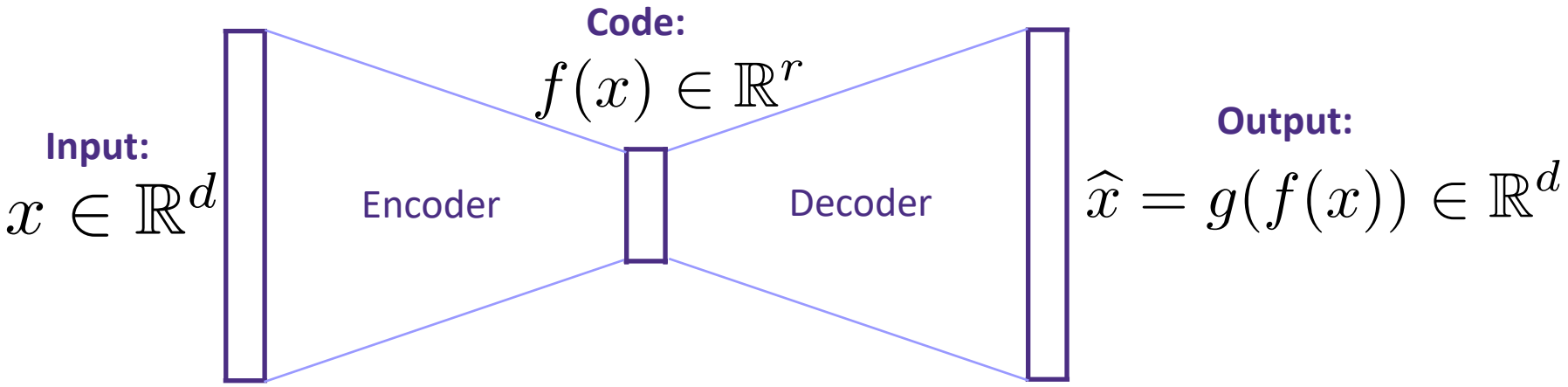
$$\max_{V_q} \quad V_q^T \Sigma V_q$$

$$\text{subject to} \quad V_q^T V_q = I$$

We can use SVD on the matrix $\tilde{X}$, and take the first r eigen pairs to find PCA.

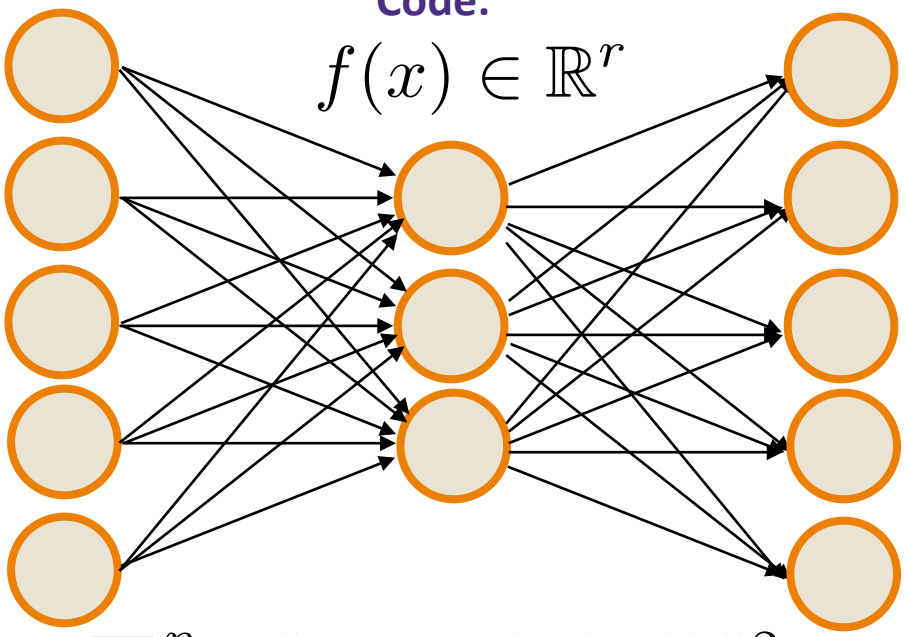
Autoencoders

Find a low dimensional representation for your data by predicting your data



$$\underset{f,g}{\text{minimize}} \sum_{i=1}^{n} \|x_i - g(f(x_i))\|_2^2$$

Autoencoders



$$\underset{f,g}{\text{minimize}} \sum_{i=1}^{n} \|x_i - g(f(x_i))\|_2^2$$

What if $f(X) = Ax$ and $g(y) = By$?

Just PCA