Principal Component Analysis

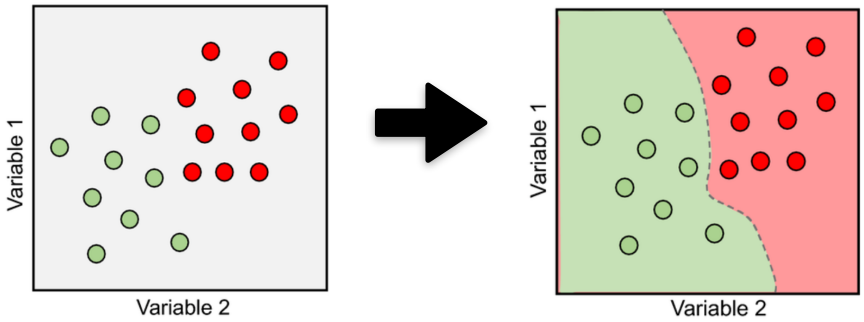
Natasha Jaques

UNIVERSITY of WASHINGTON

Unsupervised vs. supervised learning

Previously: supervised learning

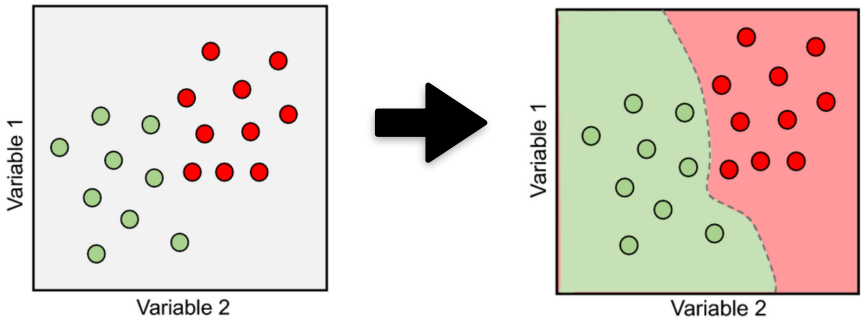
- Each data point x_i has a corresponding label y_i;$\{x_i, y_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. Try to predict the label y for a new test point x



Unsupervised vs. supervised learning

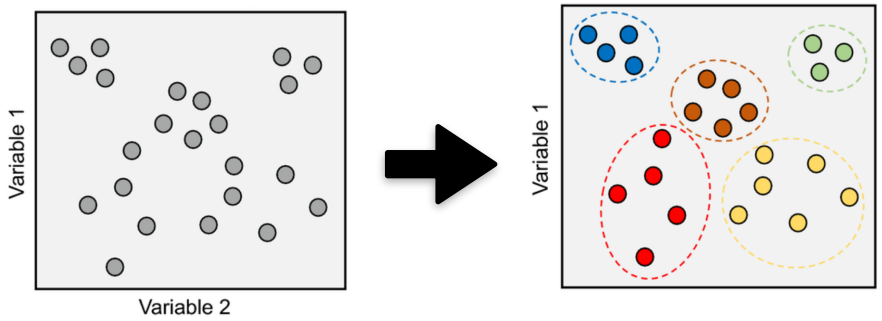
Previously: supervised learning

- 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

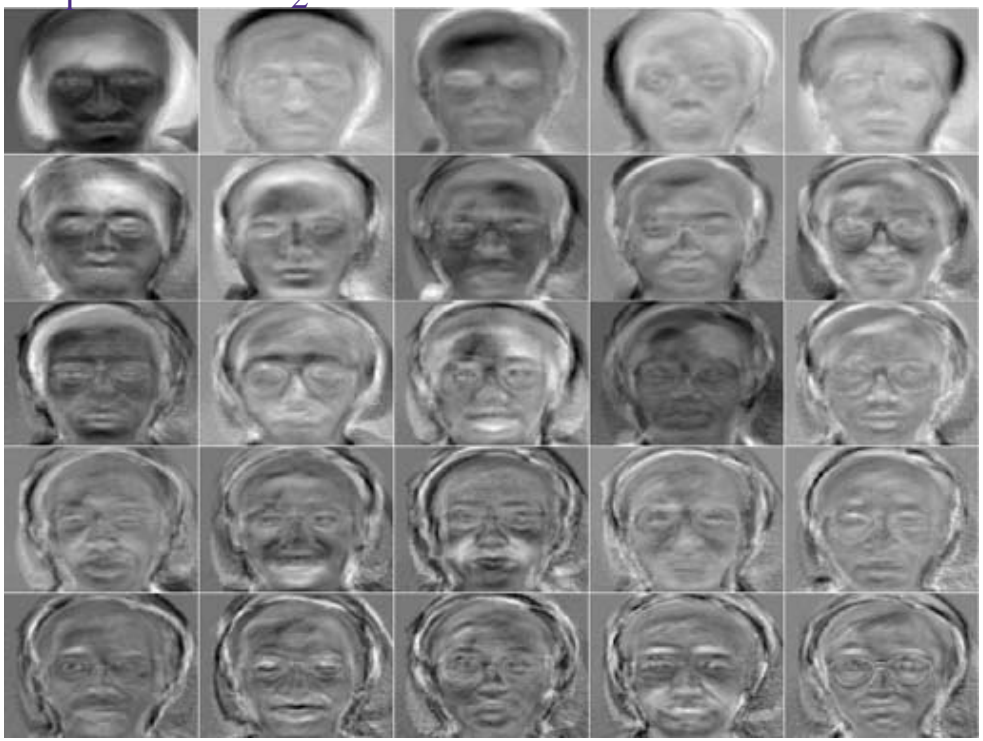
$d \times n$ real values to store the data

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

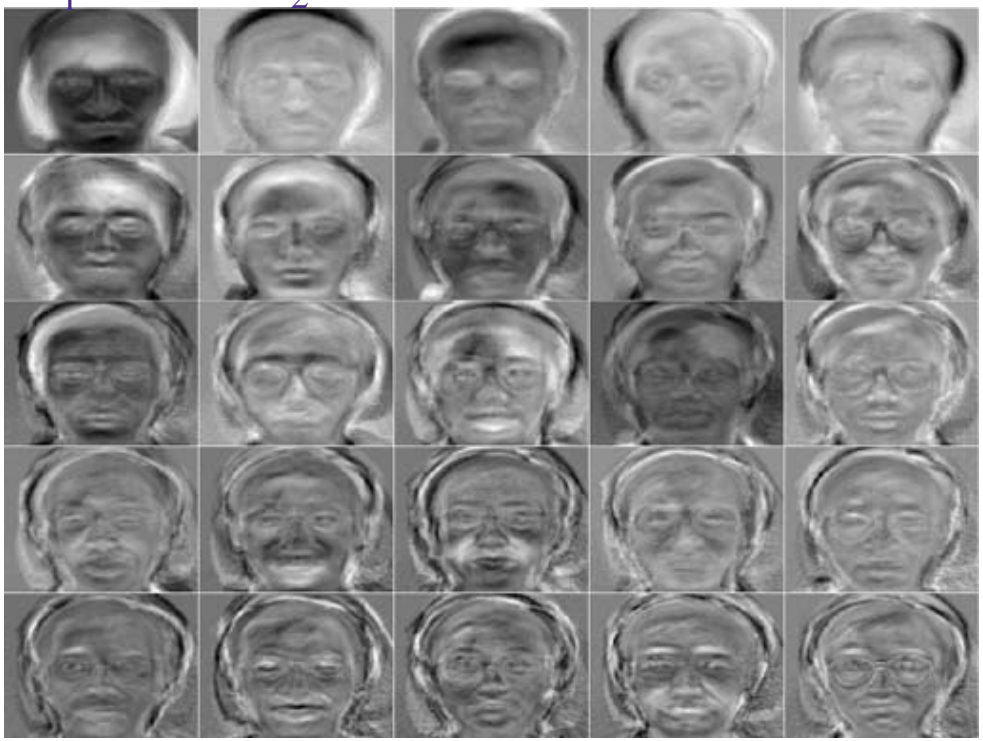
Principal component analysis finds a compact linear representation

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

- 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



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

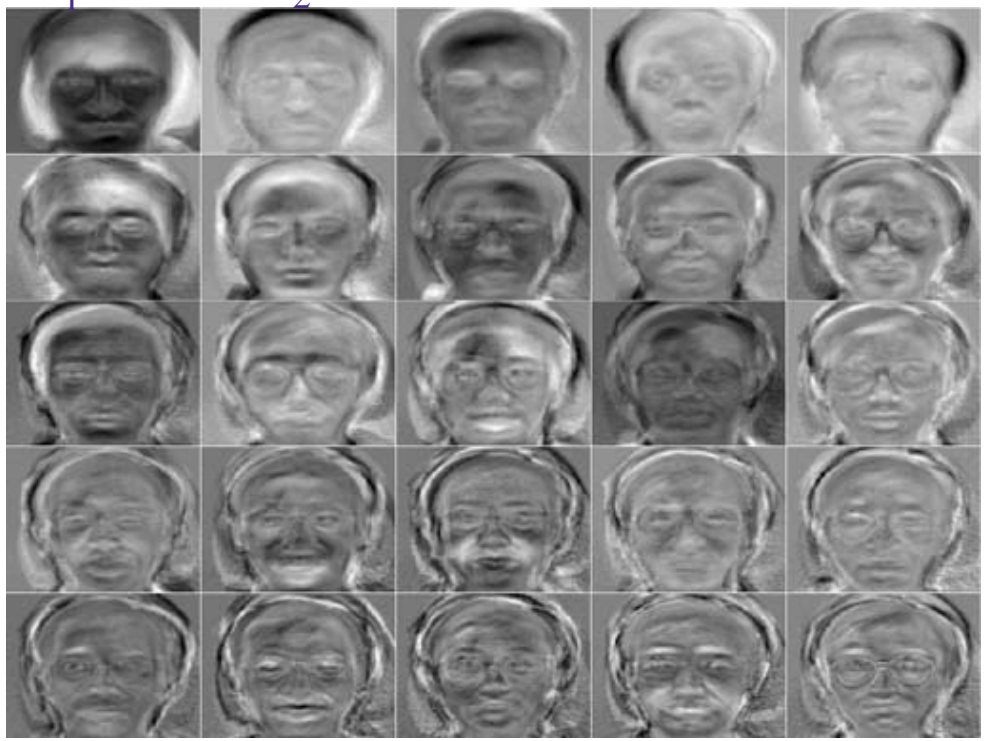
Principal component analysis finds a compact linear representation

- 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:

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

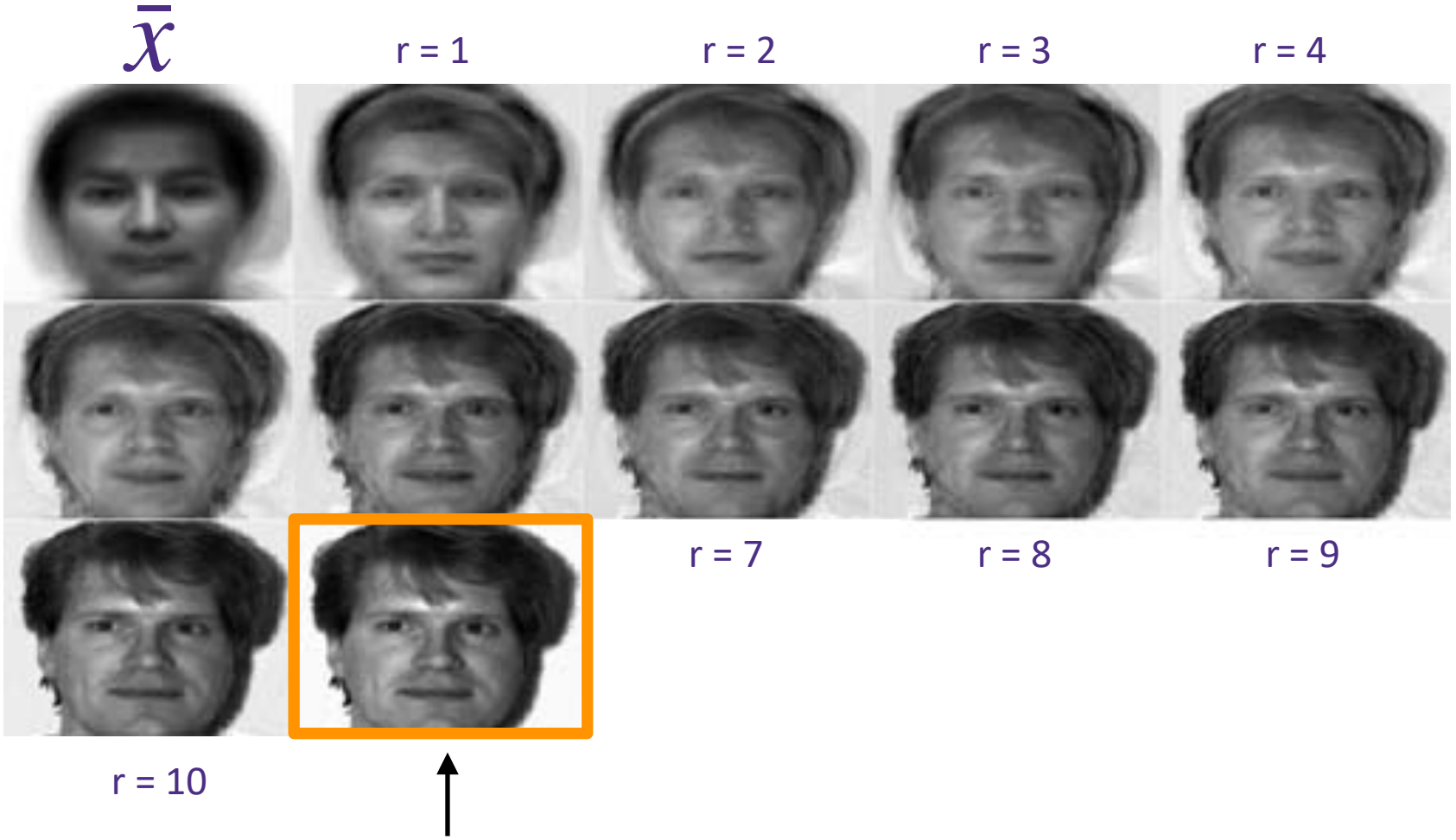


$$\approx z[1]u_1 + z[2]u_2 + \cdots + z[25]u_{25}$$

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



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

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

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 $\{z_i\}$:

PCA: a high-fidelity linear projection

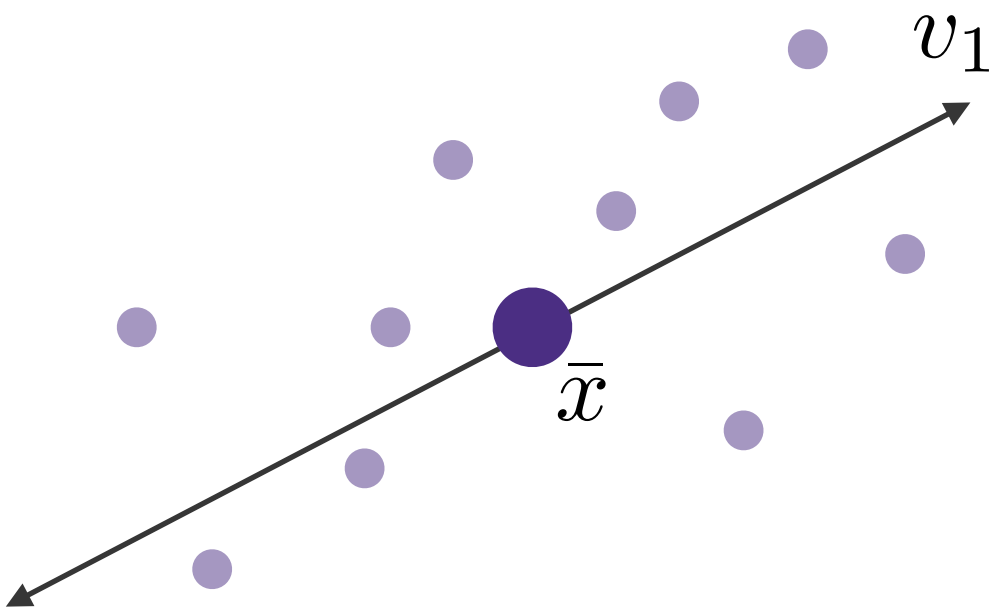
Given $x_1, ..., x_n \in \mathbb{R}^d$, find a compressed representation $z_1, ..., z_n \in \mathbb{R}^q$ with $q \ll d$ such that $x_i \approx \bar{x} + \mathbf{V}_q z_i$ and $\mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}$.

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

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

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

PCA: the geometrical interpretation



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

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

Case when $q = 1$

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

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 $q = 1$

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

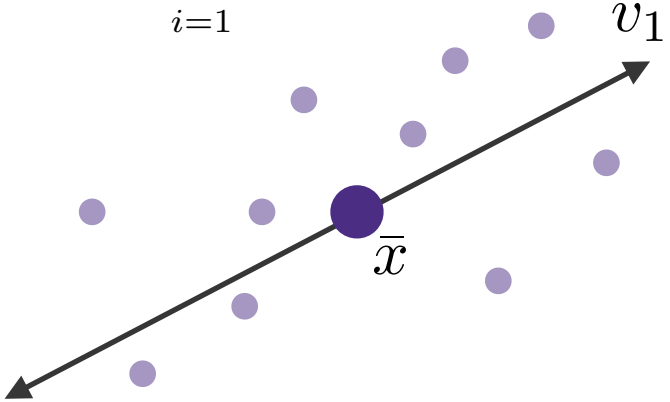
$$= \arg \min_{v:\|v\|_2=1} \sum_{i=1}^{N} \|x_i - \bar{x}\|_2^2 - 2(x_i - \bar{x})^\top vv^\top (x_i - \bar{x}) + (x_i - \bar{x})^\top vv^\top vv^\top (x_i - \bar{x})$$

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

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

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

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



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

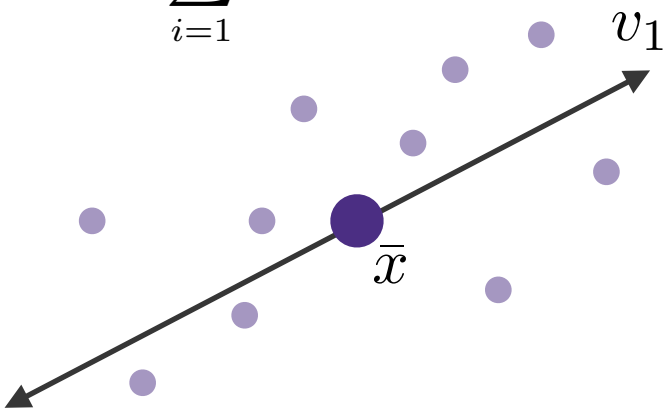
$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j \langle v_j, x_i - \bar{x} \rangle$$

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

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

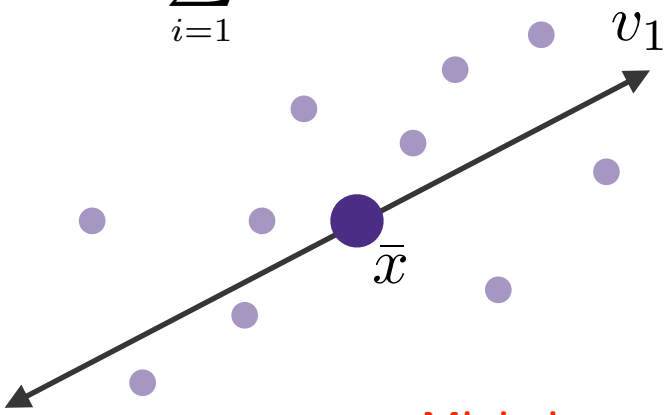
$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

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

General $q \geq 1$

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

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



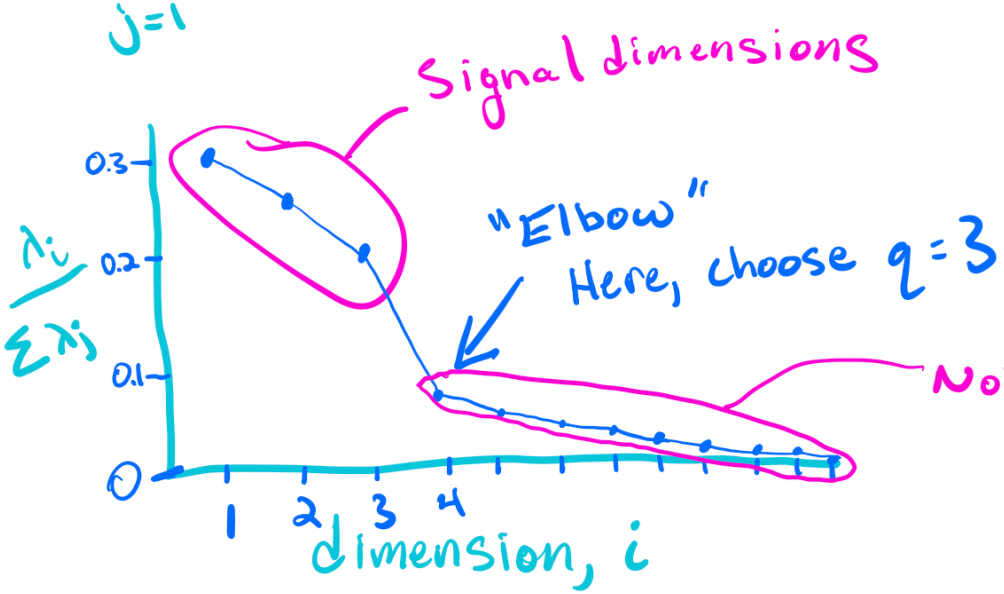
$\mathbf{V}_q$ are the first q eigenvectors of Σ

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

How to choose the dimensionality, *q*

HOW TO CHOOSE q

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



CROSS VALIDATION DOESN'T WORK

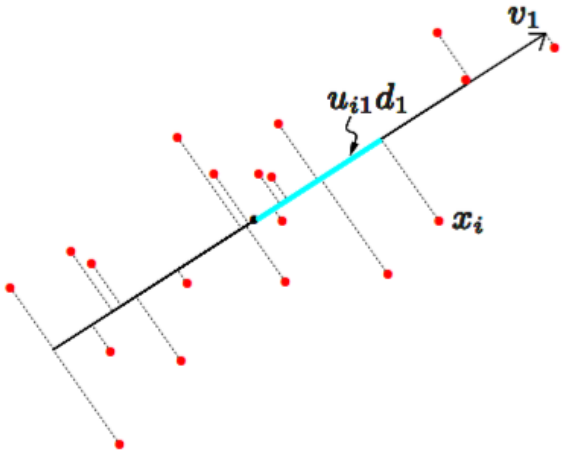
- More dimensions always increases projected variance (decreases reconstruction error), INCLUDING ON VAL DATA.

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

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 $\mathbf{V}_q$

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