Principal Component Analysis

Matt Golub
Hunter Schafer

UNIVERSITY of WASHINGTON

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?

1024

d=32x32pixels per image

n images

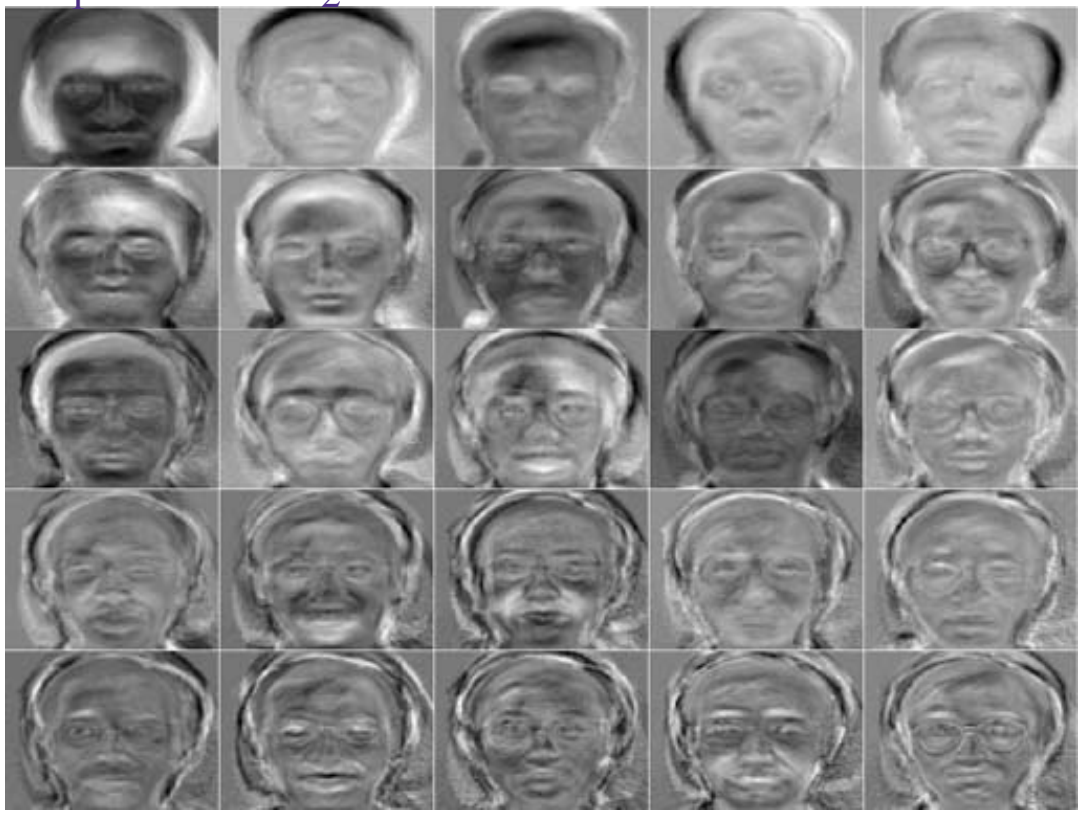
$d \times n$ real values to store the data

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



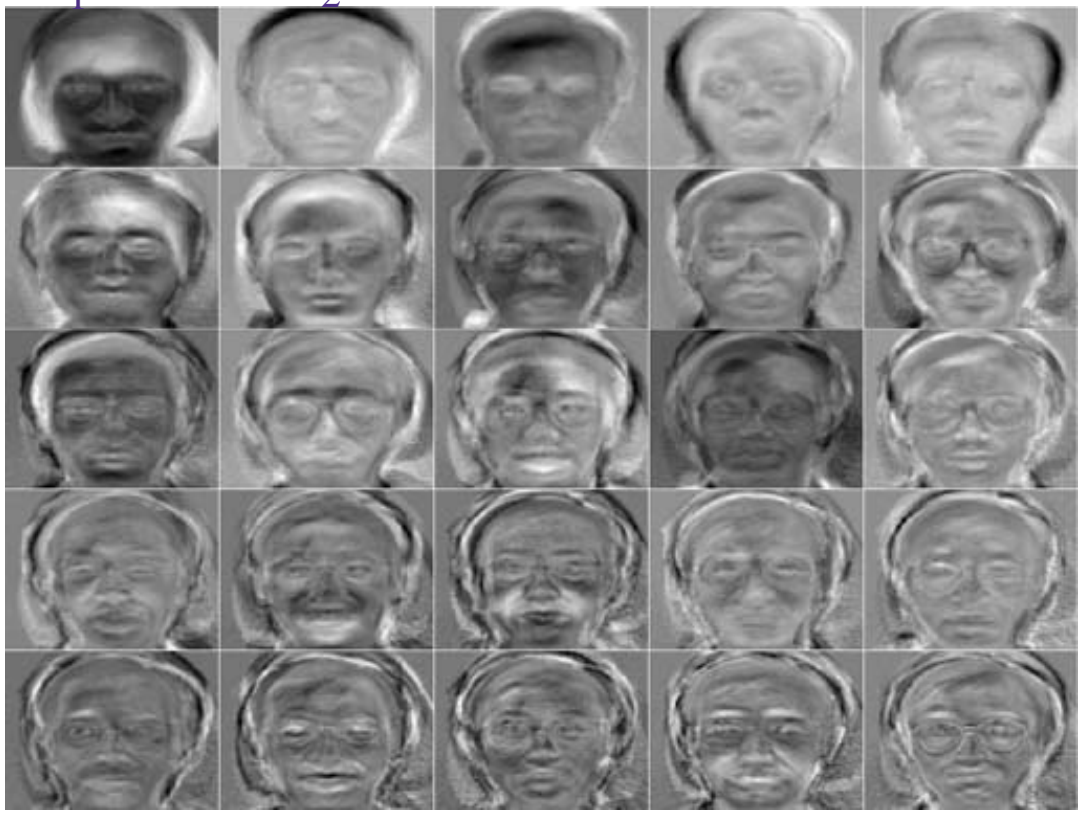
Principal component analysis finds a compact linear representation

- 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



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

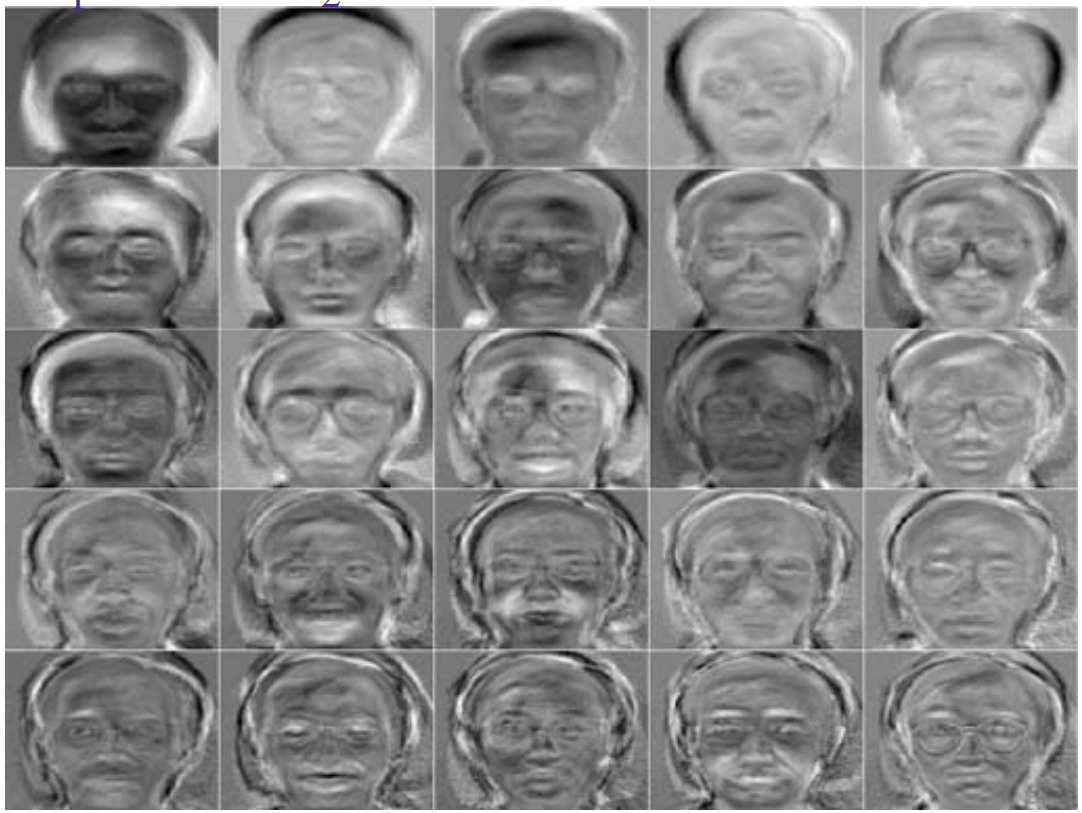
Principal component analysis finds a compact linear representation

- 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



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

- 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

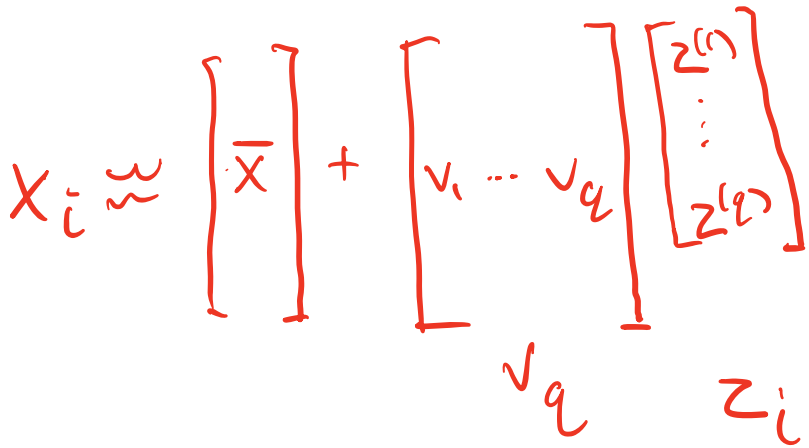
average face $\bar{x} + a[1]u_1$ $\bar{x} + a[1]u_1 + a[2]u_2$

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

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

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

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

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

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

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

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

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

$\mathbf{V}_q$: $d \times q$

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

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

$\mathbf{V}_q \mathbf{V}_q^\top$: $d \times d$

$\mathbf{V}_q \mathbf{V}_q^\top$ projects onto a q-dim linear subspace embedded in $\mathbb{R}^d$.

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

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

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

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

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

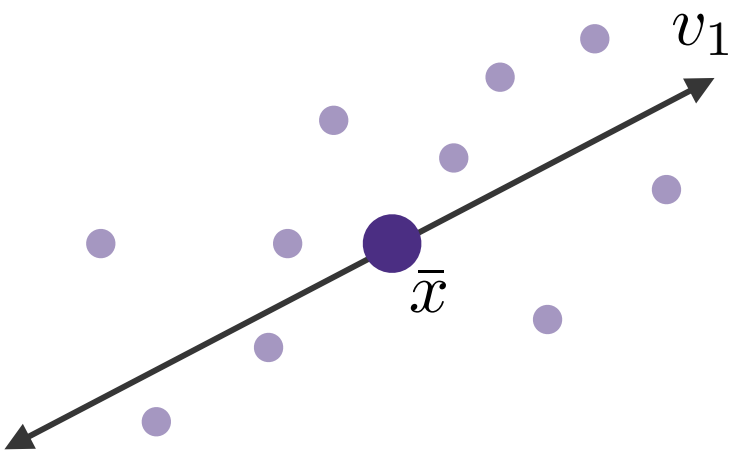
Case when $q = 1$

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

a b

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

$$\|x_i - \bar{x}\| - 2(x_i - \bar{x})^\top v v^\top (x_i - \bar{x})$$
$$+ (x_i - \bar{x})^\top v \underbrace{v^\top v}_{=1} v^\top (x_i - \bar{x})$$



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

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

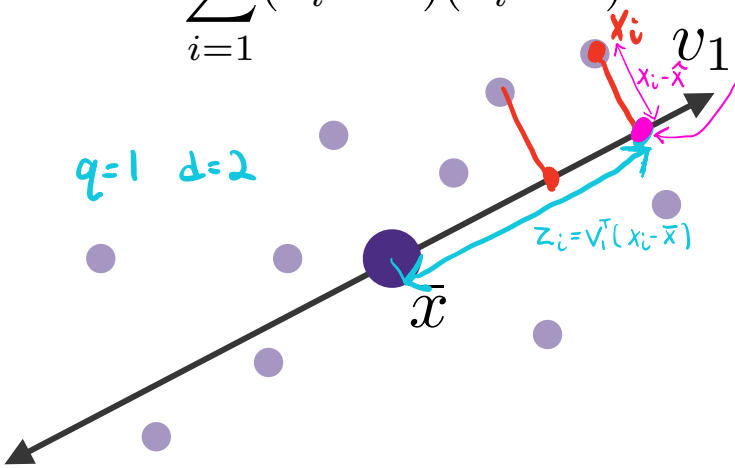
Case when $q = 1$

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

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

$$+ (x_i - \bar{x})^\top vv^\top vv^\top (x_i - \bar{x})$$

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



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

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

scalar scalar

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

Maximize the projected variance along direction v.

$$= \arg\max_v \sum_{i=1}^{N} \left[v^\top (x_i - \bar{x})(x_i - \bar{x})^\top v \right]$$

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

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

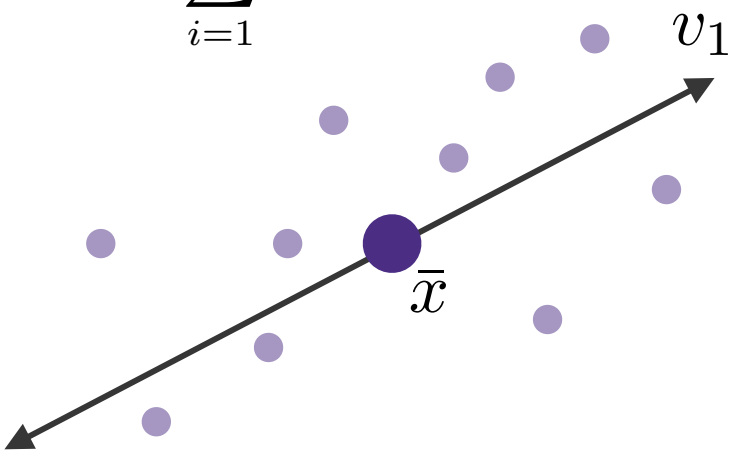
$$\widehat{x}_i := \bar{x} + \mathbf{V}_q \mathbf{V}_q^\top (x_i - \bar{x}) = \bar{x} + \sum_{j=1}^{q} v_j \langle v_j, x_i - \bar{x} \rangle$$

Case when $q = 1$

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

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

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



(v, λ) are an (eigenvector, eigenvalue) pair of (square) matrix A

if [by definition]: $Av = \lambda v$

We have $A = \Sigma$. $\Sigma v = \lambda v$

$\Rightarrow v^T \Sigma v = \lambda v^T v = \lambda$

$\Rightarrow$ Eigenvalue λ_i is the variance of the data projected onto direction v_i.

Choose v to be the eigenvector corresponding to the leading eigenvalue, λ_{max}.

To prove,
- Write out lagrangian, take derivatives, set to 0.

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

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

scalar = Tr(scalar)

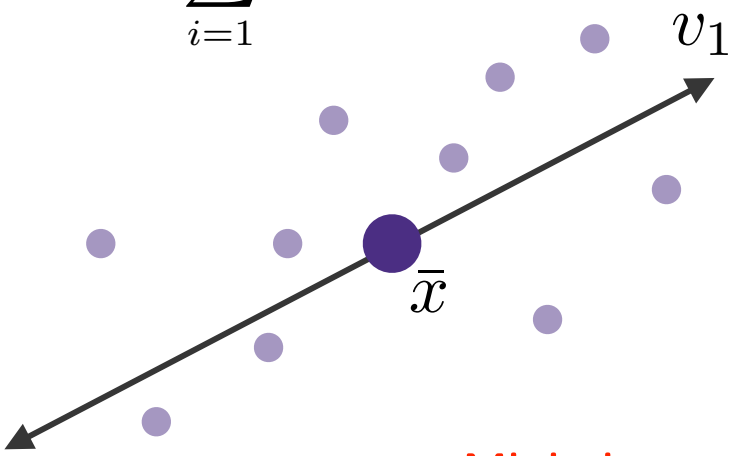
Tr(ABC) = Tr(CAB)

= Tr(BCA)

General $q \geq 1$

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

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



np.linalg.eig(Sigma)

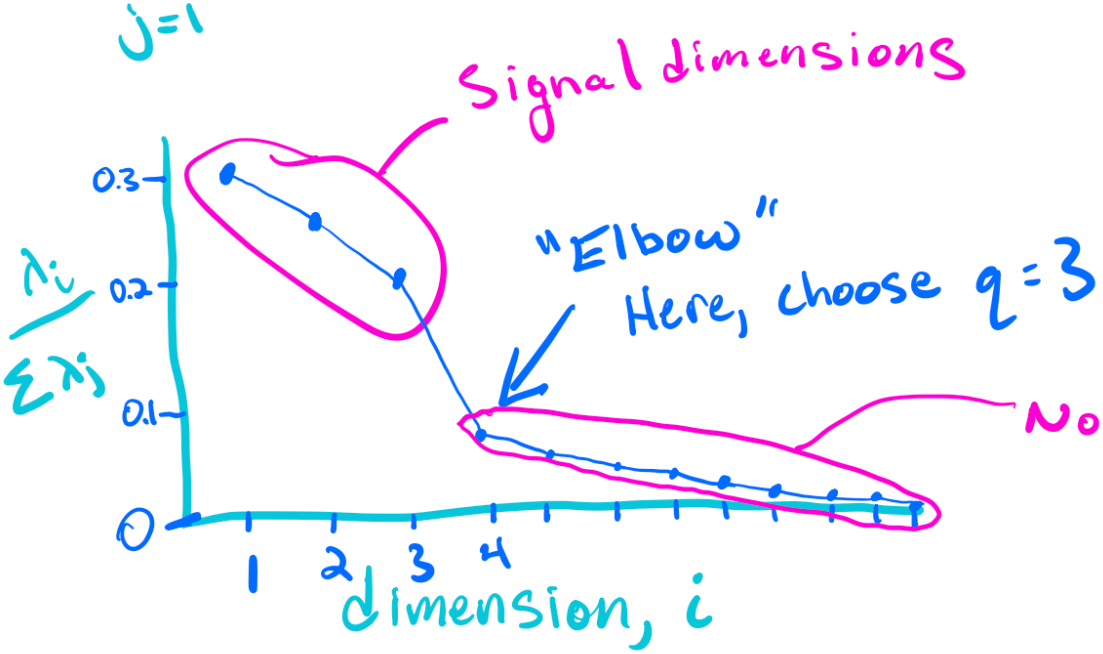
$\mathbf{V}_q$ are the first q eigenvectors of Σ

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)

For more principled approach, define probabilistic model. Covered in CSE 599N.

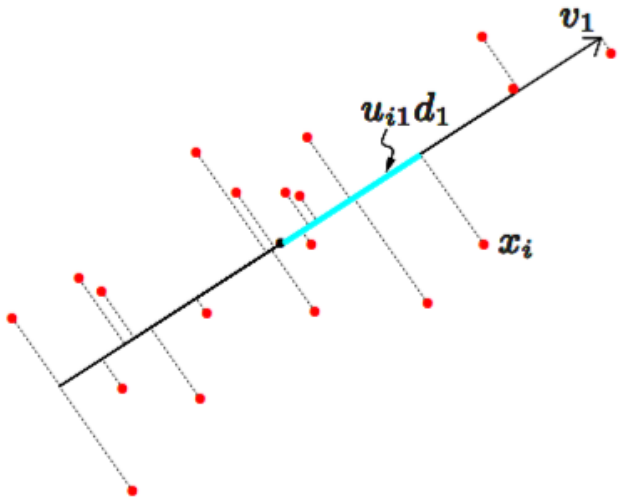
PCA: a high-fidelity linear projection

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)

Theorem (SVD): Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{USV}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = I$, $\mathbf{V}^T\mathbf{V} = I$.

"singular values"

$$\mathbf{A}^T\mathbf{A}v_i = (USV^T)^T(USV^T)v_i = VS^TU^TUSV^Tv_i$$

$$= VS^TSV^Tv_i = VS^2V^Tv_i$$

$$= VS^2e_i$$

$$= v_i s_{ii}^2$$

$$\mathbf{A}\mathbf{A}^T u_i = u_i s_{ii}^2$$

Singular Value Decomposition (SVD)

np.linalg.svd(X)

Theorem (SVD): Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = I$, $\mathbf{V}^T\mathbf{V} = I$.

$$\mathbf{A}^T\mathbf{A}v_i = \mathbf{S}_{i,i}^2 v_i$$

$$\mathbf{A}\mathbf{A}^T u_i = \mathbf{S}_{i,i}^2 u_i$$

$\mathbf{V}$ are the first r eigenvectors of $\mathbf{A}^T\mathbf{A}$ with eigenvalues $\text{diag}(\mathbf{S}^2)$
$\mathbf{U}$ are the first r eigenvectors of $\mathbf{A}\mathbf{A}^T$ with eigenvalues $\text{diag}(\mathbf{S}^2)$

Suppose $X = \begin{bmatrix} (x_1 - \bar{x})^T \\ \vdots \\ (x_n - \bar{x})^T \end{bmatrix} \in \mathbb{R}^{n \times d}$ $\quad X = USV^T$

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

$$= \frac{1}{n}(VS^TU^T)(USV^T)$$

$$= \frac{1}{n}VS^2V^T$$

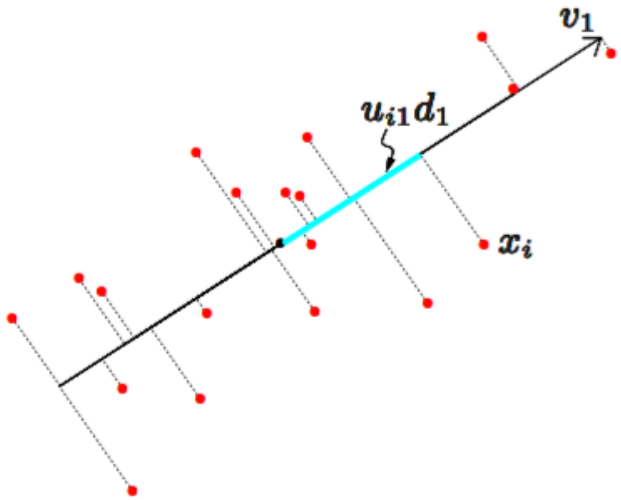
This is the diagonalization (eigen-decomposition) of Σ, where the columns of V are the eigenvectors of Σ and $\lambda_i = \frac{s_{ii}^2}{n}$ are the eigenvalues of Σ.
Or, alternatively, $s_{ii} = \sqrt{n\lambda_i}$.

SVD on a de-meaned data matrix X is equivalent to running PCA on the data.

Linear projections

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 $\mathbf{V}_q$

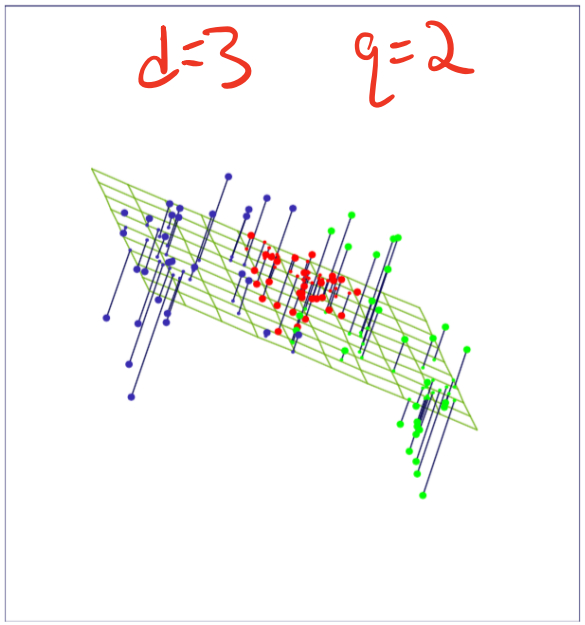
$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U}_q \mathrm{diag}(d_1, \ldots, d_q) \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

Singular Value Decomposition defined as

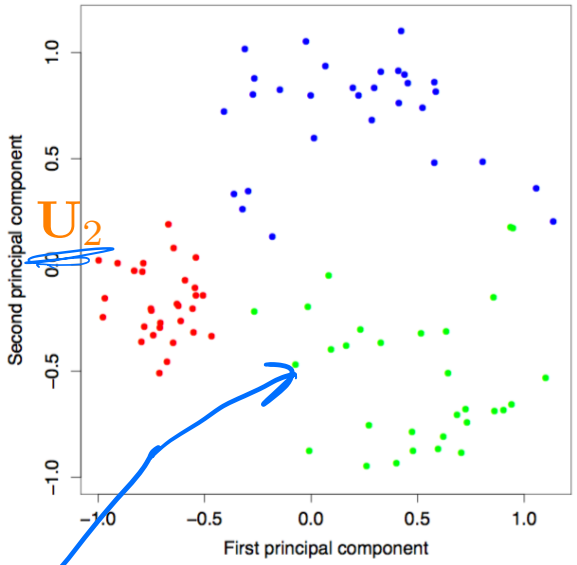
$$\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T \approx \sum_{i=1}^{d} u_i v_i^T s_{ii}$$

Dimensionality reduction

$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T$



$\mathbf{X} - \mathbf{1}\bar{x}^T$



The PCA projection

$(X - 1\bar{x}^T)V_q = USV^TV_q$

$n \times d \quad d \times q \quad = US[e_1 \cdots e_q]$

$= [u_1 \cdots u_q][s_{11} \cdots s_{qq}] = U_q S_q$

$n \times q \quad q \times q$

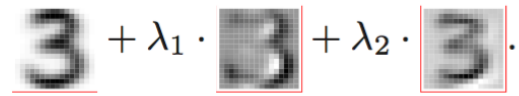
Each row of $U_q S_q$ is a projected datapoint (z_i^T). Here plotting for $q=2$.

Dimensionality reduction

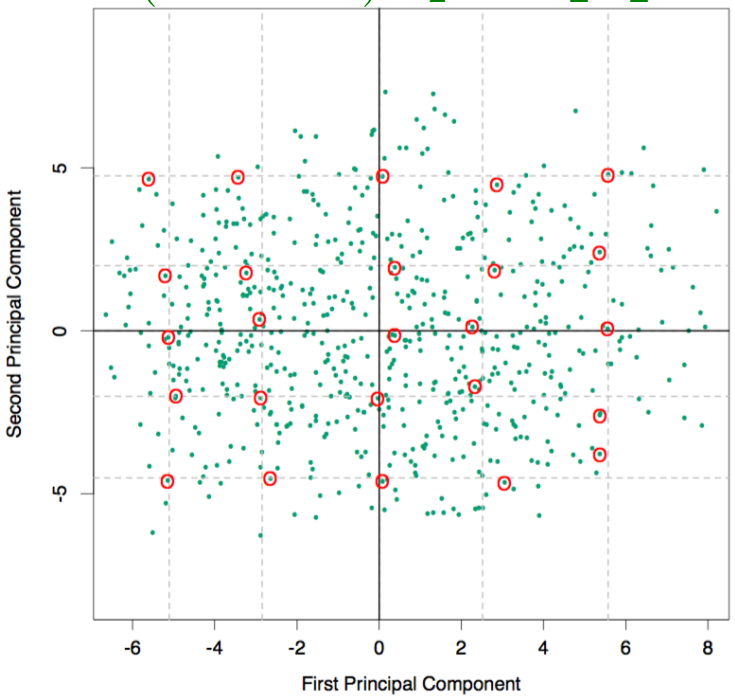
$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$

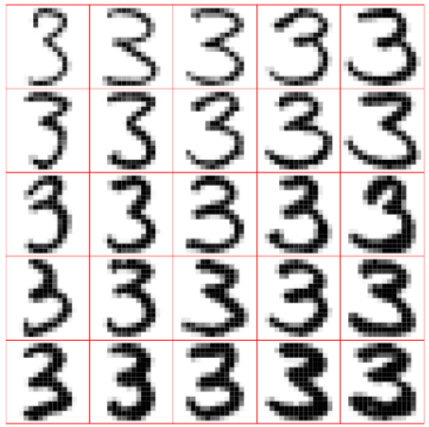
Handwritten 3's, 16x16 pixel image so that $x_i \in \mathbb{R}^{256}$

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$



$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_2 = \mathbf{U}_2\mathbf{S}_2 \in \mathbb{R}^{n\times 2}$$





$\mathrm{diag}(\mathbf{S})$

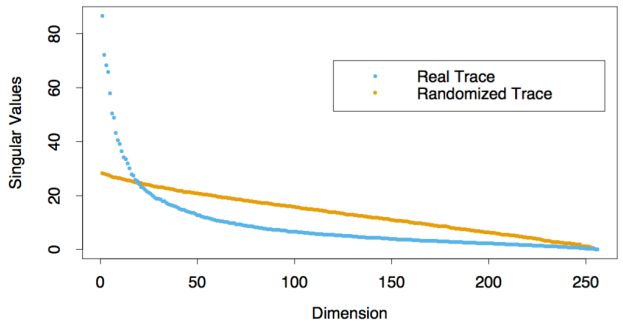
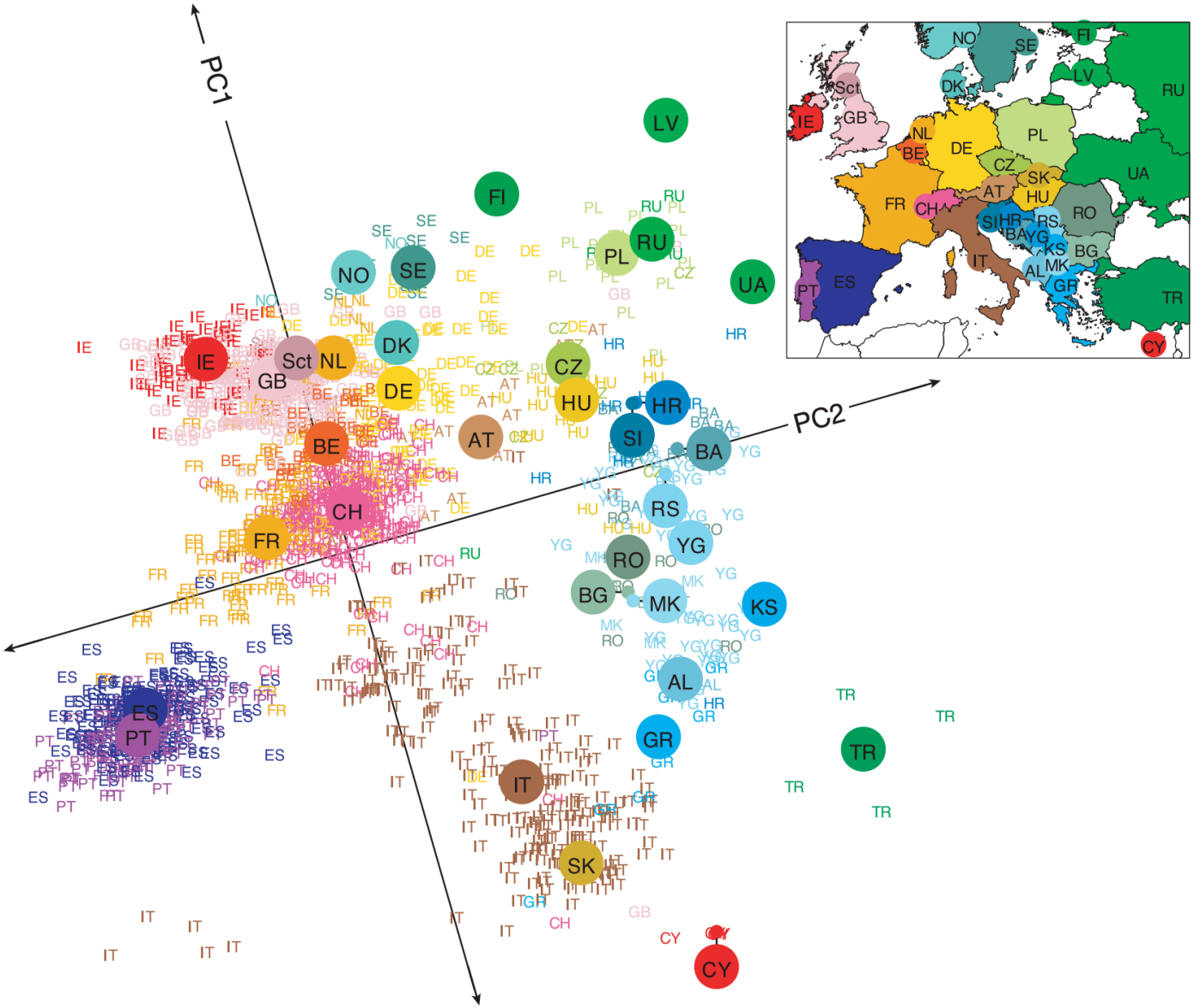


FIGURE 14.24. *The 256 singular values for the digitized threes, compared to those for a randomized version of the data (each column of* $\mathbf{X}$ *was scrambled).*

Dimensionality reduction



Novembre, et al, “Genes mirror geography within Europe” Nature 2008.

Kernel PCA

$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U_q}\mathbf{S_q} \in \mathbb{R}^{n \times q}$$

$$\mathbf{J}\mathbf{X} = \mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T \qquad \mathbf{J} = I - \mathbf{1}\mathbf{1}^T/n$$

$$(\mathbf{J}\mathbf{X})(\mathbf{J}\mathbf{X})^T =$$

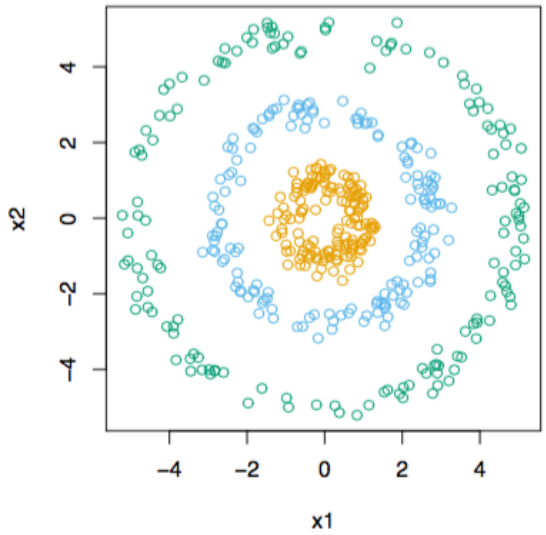
Kernel PCA

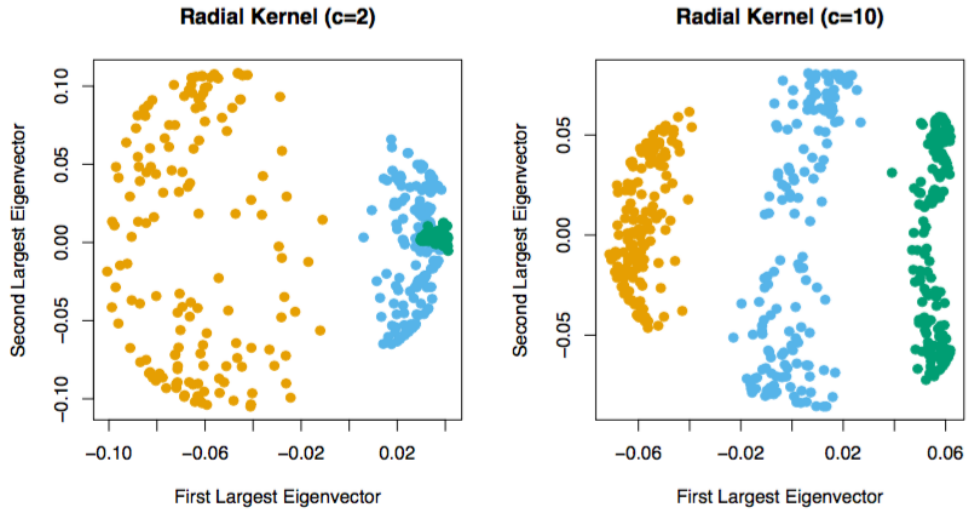
$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U_q}\mathbf{S_q} \in \mathbb{R}^{n \times q}$$

$$\mathbf{J}\mathbf{X} = \mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T \qquad \mathbf{J} = I - \mathbf{1}\mathbf{1}^T/n$$

$$(\mathbf{J}\mathbf{X})(\mathbf{J}\mathbf{X})^T = \mathbf{U}\mathbf{S}^2\mathbf{U}^T$$





Matrix completion

Given historical data on how users rated movies in past:

NETFLIX

17,700 movies, 480,189 users, 99,072,112 ratings (Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for $1 million prize)

Matrix completion

n movies, d users, $|S|$ ratings

$$\underset{\tilde{U}\in\mathbb{R}^{n\times q},\tilde{V}\in\mathbb{R}^{d\times q}}{\arg\min} \sum_{(i,j)\in S} \left([\tilde{U}\tilde{V}^\top]_{ij} - X_{ij}\right)^2$$

How do we solve it? With full information?

Matrix completion

n movies, d users, $|S|$ ratings

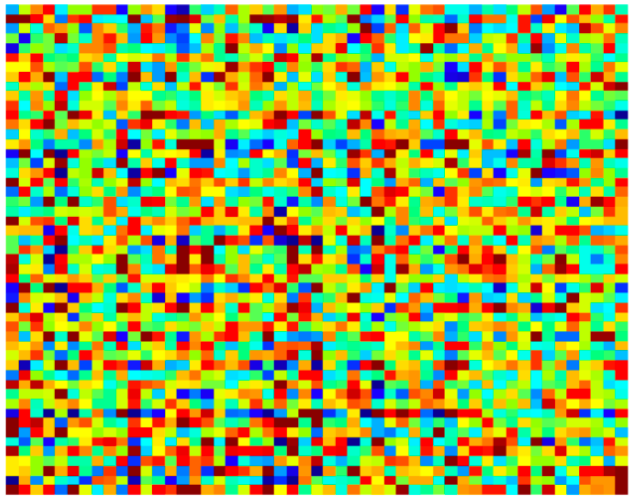
$$\underset{\tilde{U}\in\mathbb{R}^{n\times q},\tilde{V}\in\mathbb{R}^{d\times q}}{\arg\min} \sum_{(i,j)\in S} \left([\tilde{U}\tilde{V}^\top]_{ij} - X_{ij}\right)^2$$

What about the general case, with (many!) missing entries?

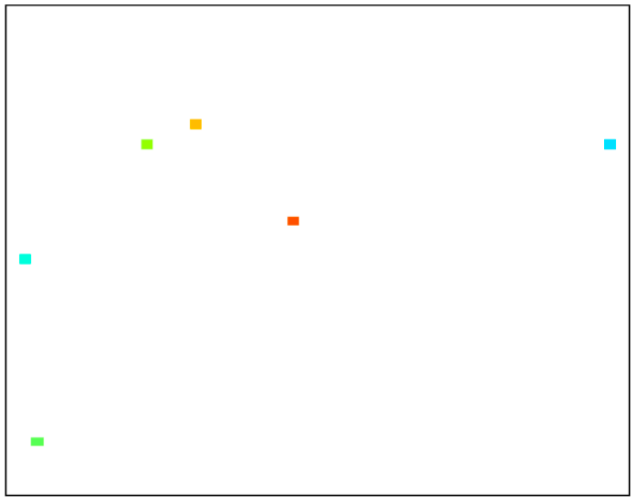
Example: 2000×2000 rank-8 random matrix

For illustration, we zoom in to a 50x50 submatrix

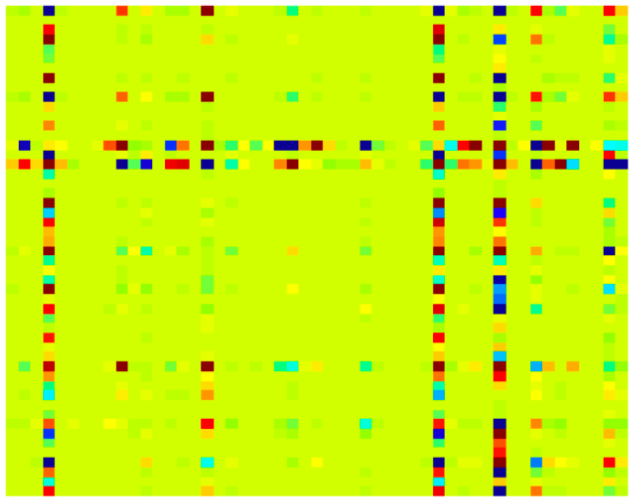
low-rank matrix $\mathbf{X}$



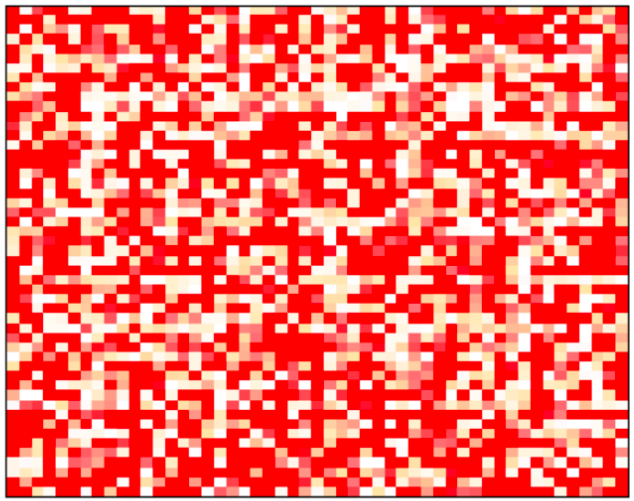
sampled matrix



Gradient descent output $\mathbf{UA}$



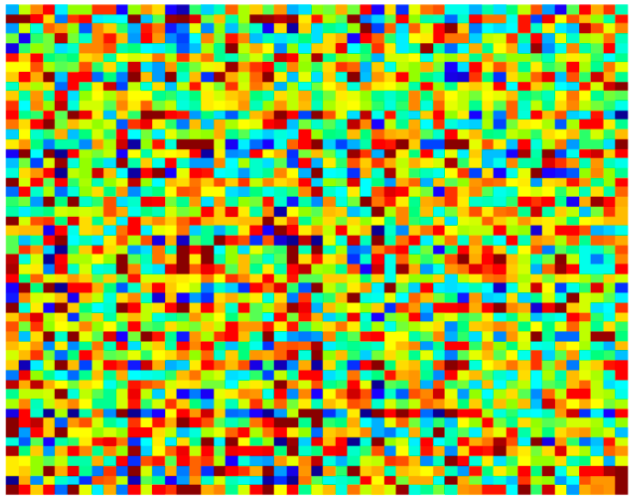
squared error $(\mathbf{X}_{ji} - (\mathbf{UA})_{ji})^2$



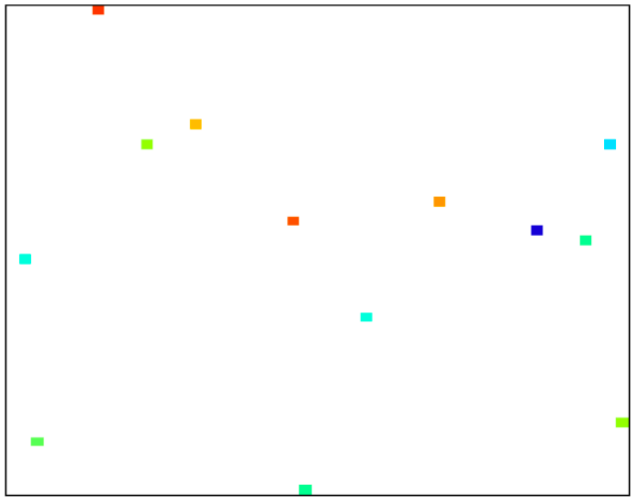
0.25% sampled

Example: 2000×2000 rank-8 random matrix

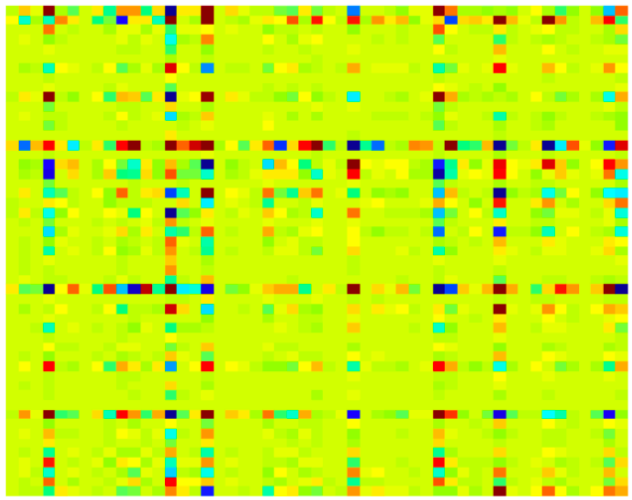
low-rank matrix $\mathbf{X}$



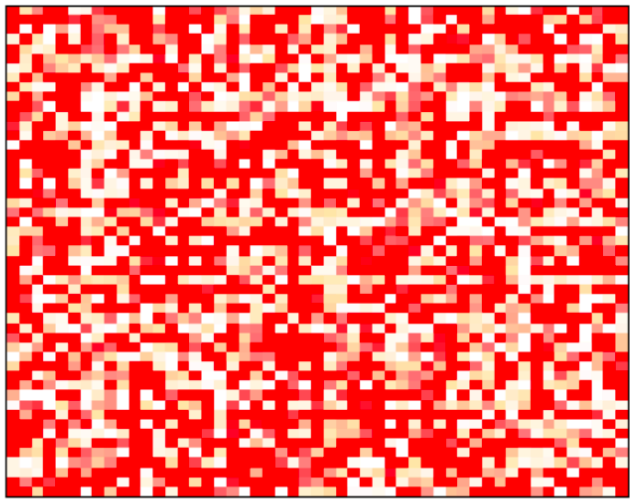
sampled matrix



Gradient descent output $\mathbf{UA}$



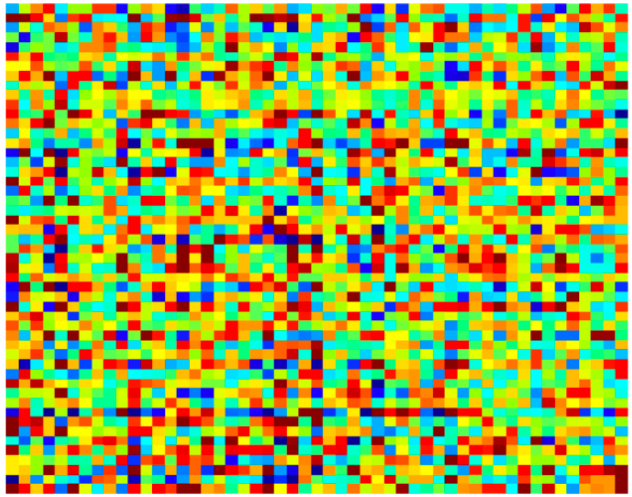
squared error $(\mathbf{X}_{ji} - (\mathbf{UA})_{ji})^2$



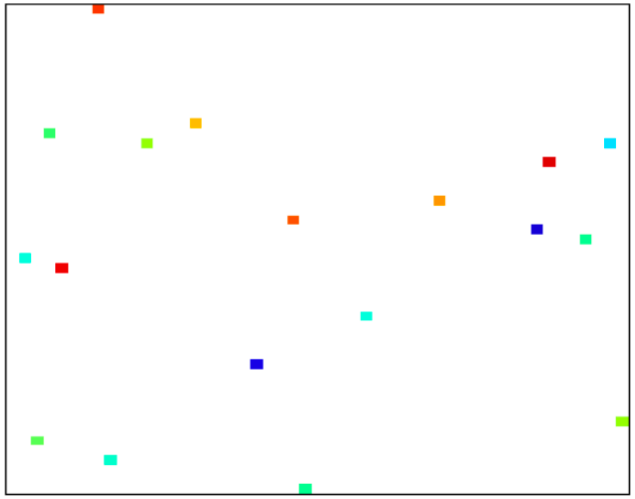
0.50% sampled

Example: 2000×2000 rank-8 random matrix

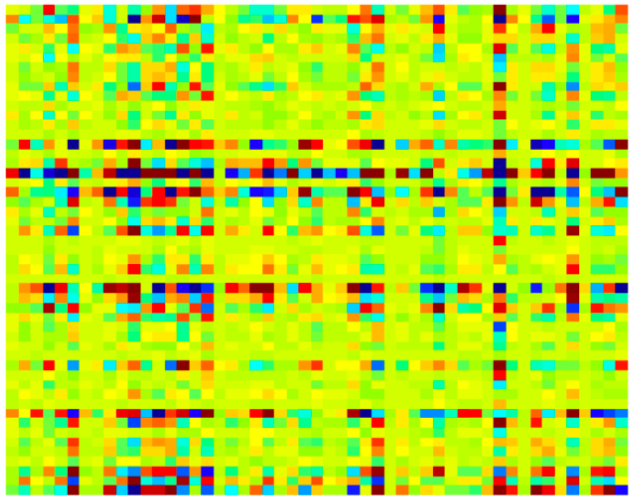
low-rank matrix $\mathbf{X}$



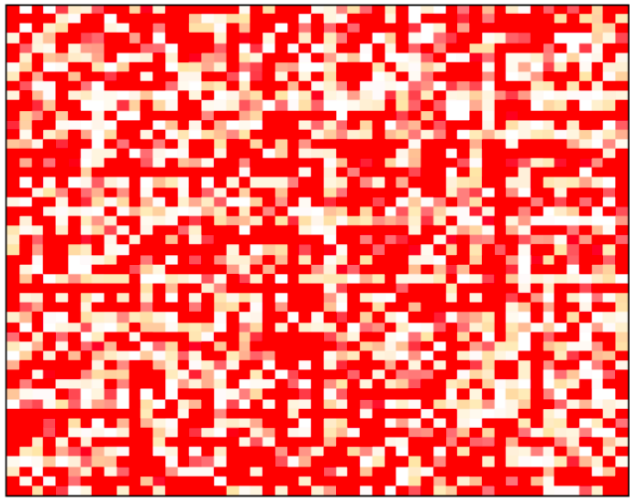
sampled matrix



Gradient descent output $\mathbf{UA}$



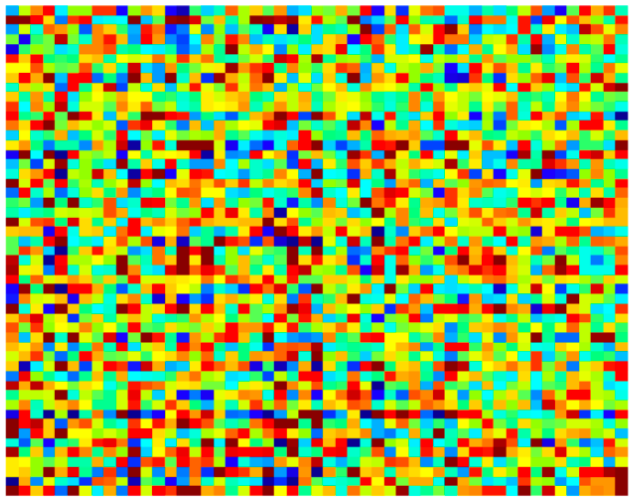
squared error $(\mathbf{X}_{ji} - (\mathbf{UA})_{ji})^2$



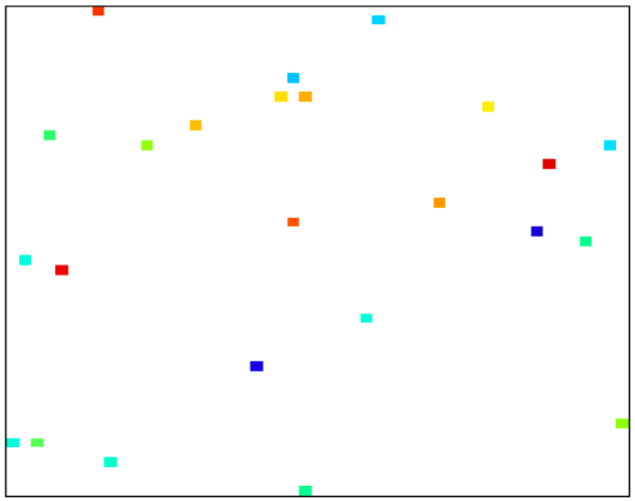
0.75% sampled

Example: 2000×2000 rank-8 random matrix

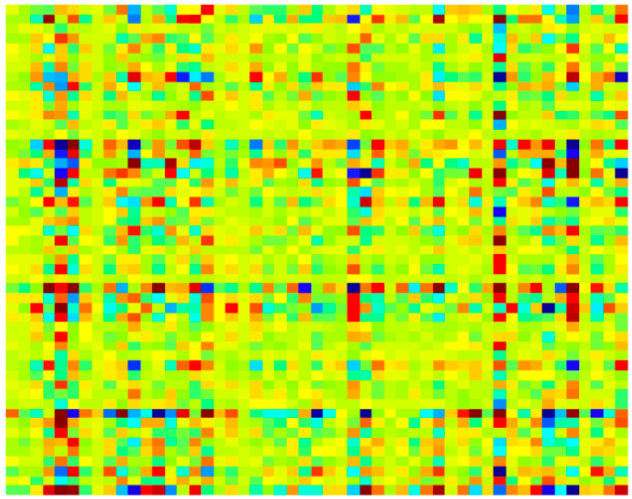
low-rank matrix $\mathbf{X}$



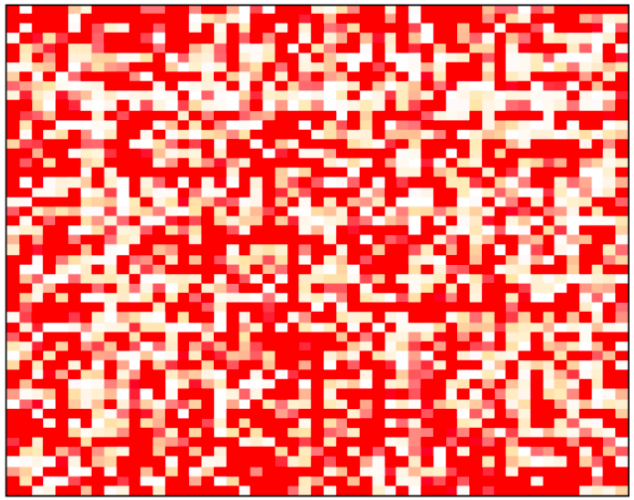
sampled matrix



Gradient descent output $\mathbf{UA}$



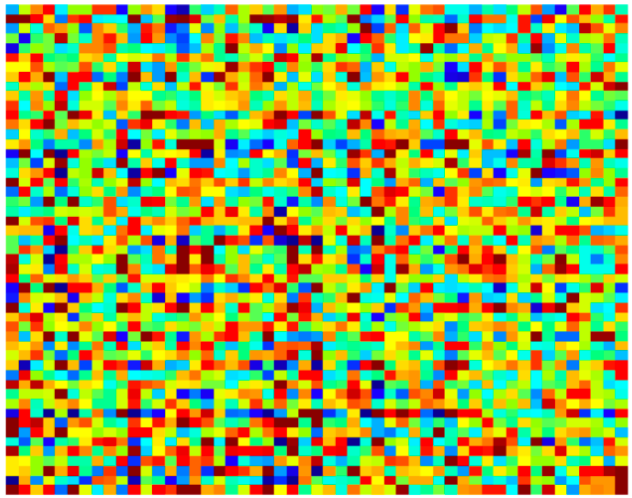
squared error $(\mathbf{X}_{ji} - (\mathbf{UA})_{ji})^2$



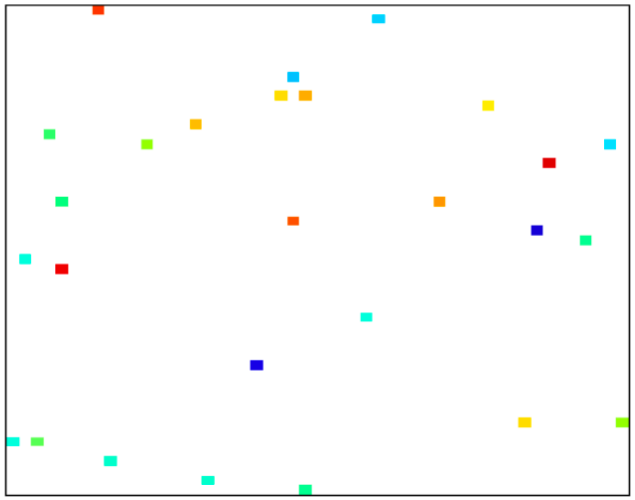
1.00% sampled

Example: 2000×2000 rank-8 random matrix

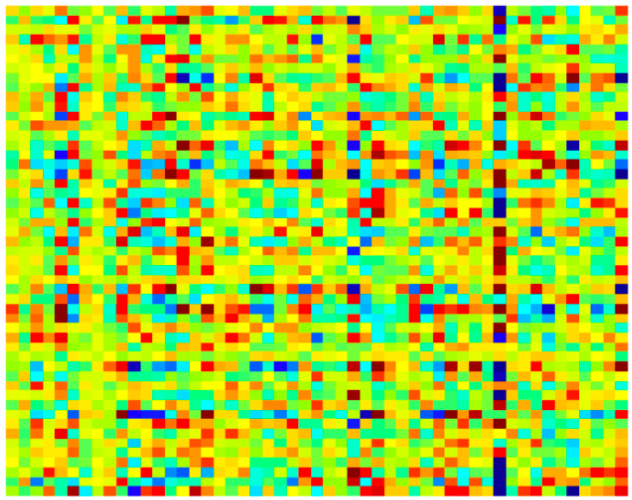
low-rank matrix $\mathbf{X}$



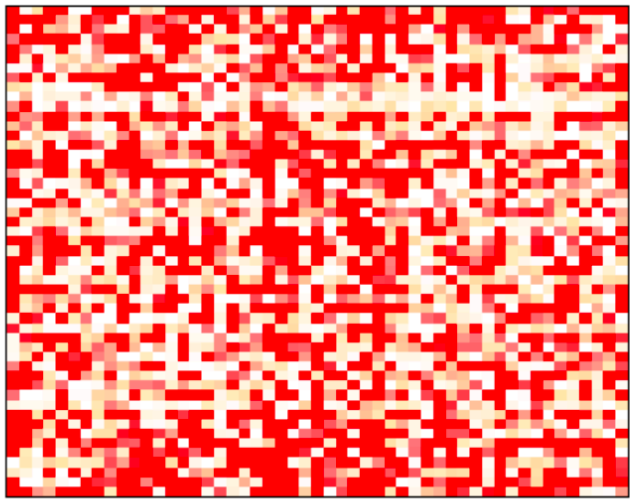
sampled matrix



Gradient descent output $\mathbf{UA}$



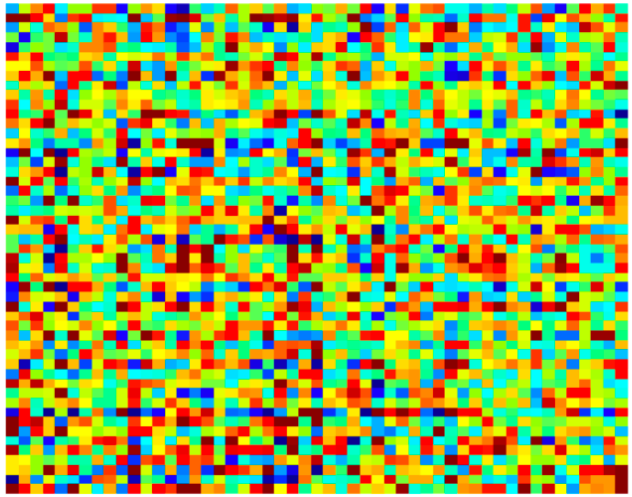
squared error $(\mathbf{X}_{ji} - (\mathbf{UA})_{ji})^2$



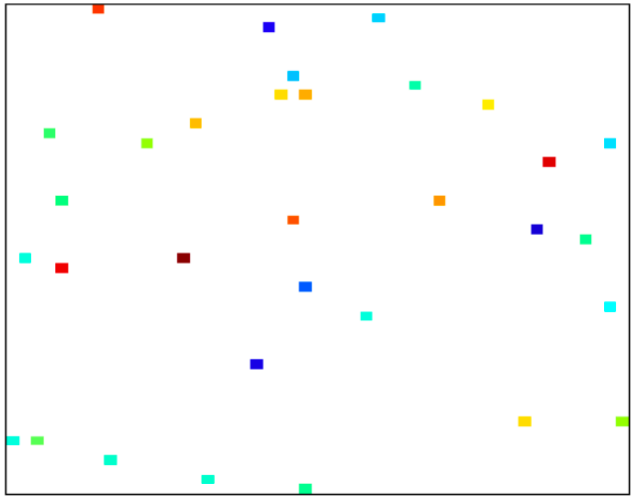
1.25% sampled

Example: 2000×2000 rank-8 random matrix

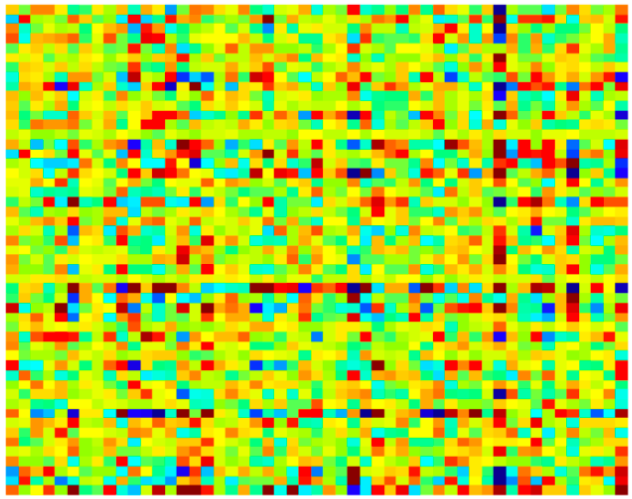
low-rank matrix $\mathbf{X}$



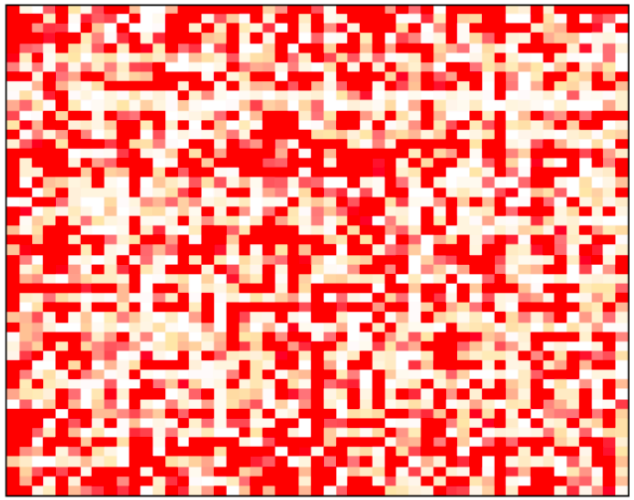
sampled matrix



Gradient descent output $\mathbf{UA}$



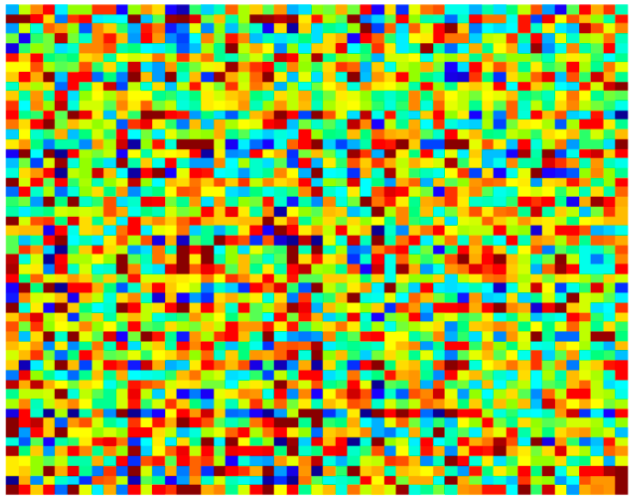
squared error $(\mathbf{X}_{ji} - (\mathbf{UA})_{ji})^2$



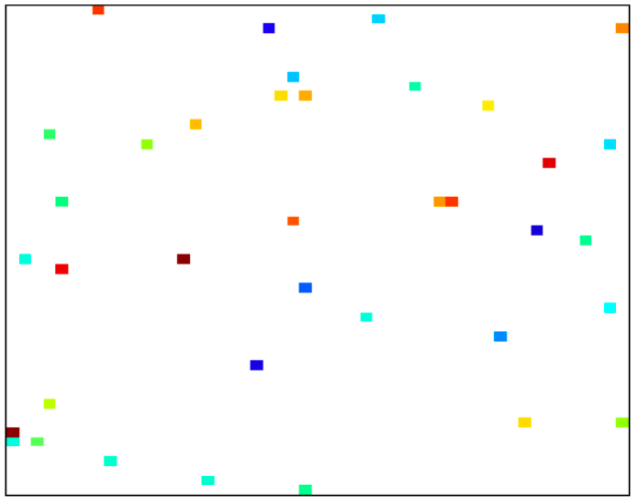
1.50% sampled

Example: 2000×2000 rank-8 random matrix

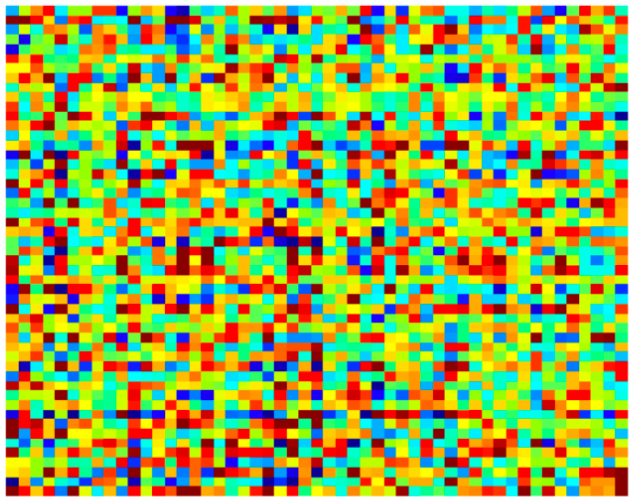
low-rank matrix $\mathbf{X}$



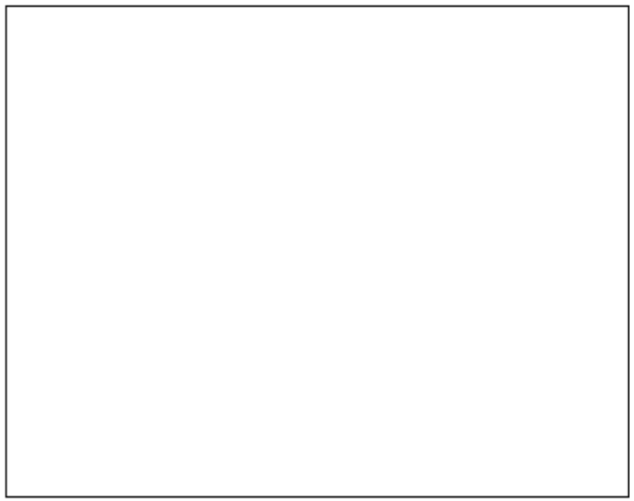
sampled matrix



Gradient descent output $\mathbf{UA}$

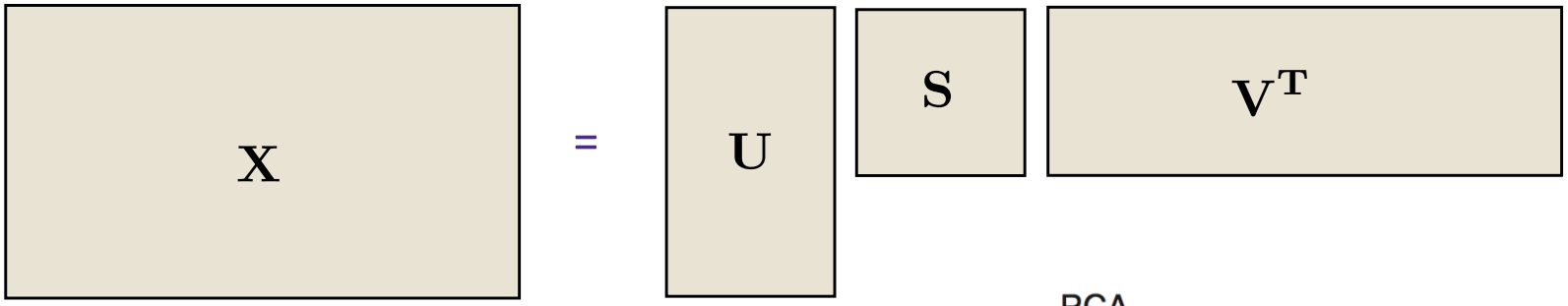


squared error $(\mathbf{X}_{ji} - (\mathbf{UA})_{ji})^2$



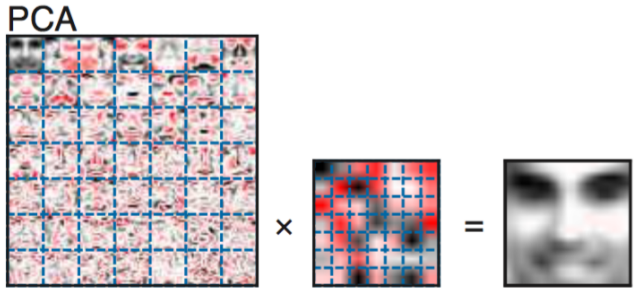
1.75% sampled

Other matrix factorizations



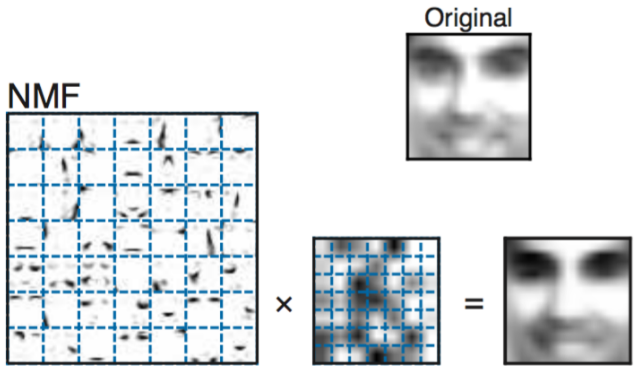
Singular value decomposition

Elements of $\mathbf{U}, \mathbf{S}, \mathbf{V}$ in $\mathbb{R}$



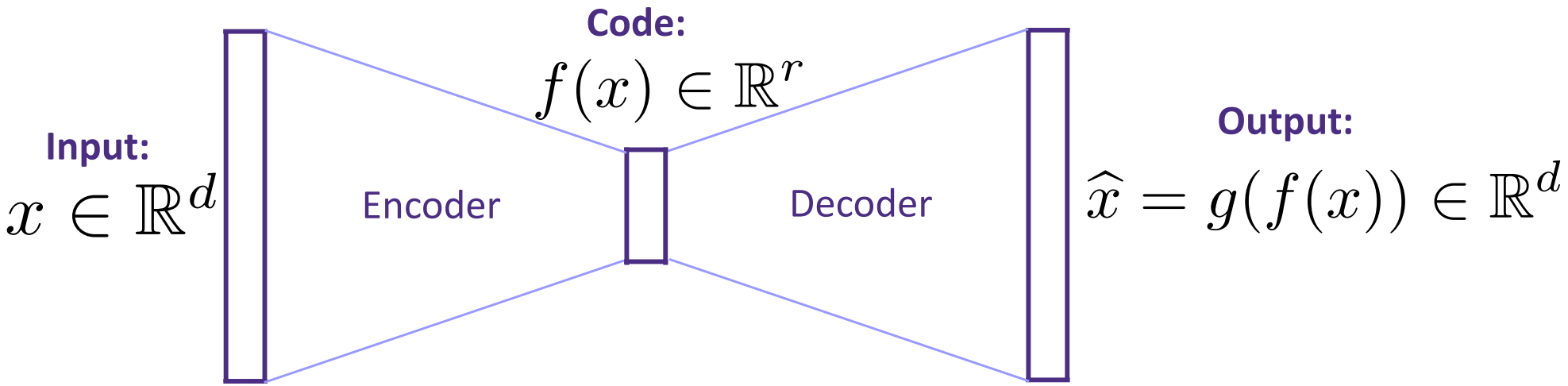
Nonnegative matrix factorization (NMF)

Elements of $\mathbf{U}, \mathbf{S}, \mathbf{V}$ in $\mathbb{R}_+$



Autoencoders

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



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

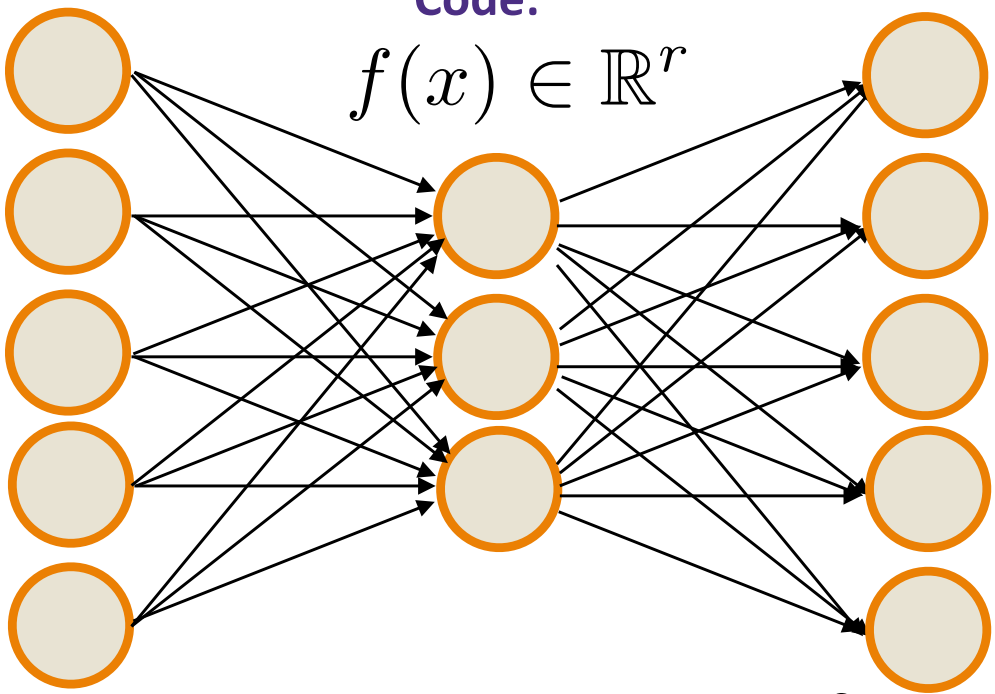
Autoencoders

Code:

$f(x) \in \mathbb{R}^r$

Input:

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



Output:

$\widehat{x} = g(f(x)) \in \mathbb{R}^d$

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

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