PCA

have data set $x_1, \dots, x_n$ $x_i \in \mathbb{R}^d$

$$X = \begin{pmatrix} \text{—} x_1 \text{—} \\ \vdots \\ \text{—} x_n \text{—} \end{pmatrix}$$

assume mean 0 $\left(\sum_{i=1}^{n} x_i = 0\right)$

Examples

- n images, d pixels each
- n measurements, d sensors
- n docs, d words
- n people, d movies
- n customers, d products

Fix k

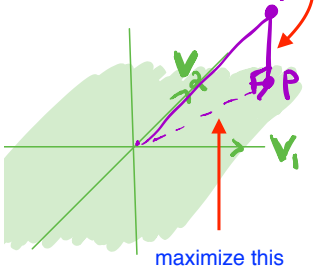
Find k-dimensional subspace (defined by orthonormal vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k$)

so as to minimize $\frac{1}{n}\sum_{i=1}^{n}$ [distance $(x_i \leftrightarrow$ subspace S spanned by $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k)$]^2

$=$ maximizing $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i, v_j)^2$

variance of projected pts

projection of $x_i \to S$: $\sum_{j=1}^{k}(x_i, v_j)\vec{v}_j$



minimze this (i.e. error from points to subspace)

maximize this (component of x in subspace)

find $v_1 \dots v_k$ to max $\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{k}(x_i, v_j)^2\right]$ $\|p\|^2$

$p = a_1 v_1 + a_2 v_2$

(x_i, v_1)

$(x - p, v_1) = 0$

$(x_i, v_1) = (p, v_1) = (a_1 v_1 + a_2 v_2, v_1)$

$= a_1$

$p = (x_i, v_1)v_1 + (x_i, v_2)v_2$

maximizing $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i v_j)^2 = \sum_{j=1}^{k}(Xv_j)^T Xv_j$

$= \frac{1}{n}\sum_{j=1}^{k} v_j^T X^T X v_j$

$X^T X$: empirical covariance matrix

How to find $\vec{v}_1, \ldots, \vec{v}_k$

① Compute eigendecomposition of empirical covariance matrix $X X^T$

$A := X^T X = Q D Q^T$ where $Q = \begin{pmatrix} | & & | \\ q_1 & \cdots & q_d \\ | & & | \end{pmatrix}$ orthonormal eigenvectors

As A is symmetric eigendecomposition is special: Q is orthogonal! orthonormal evectors

$q_1, \ldots, q_n$ orthonormal

$QQ^T = Q^T Q = I$

$\forall w \quad \|Qw\| = \|w\|$

$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_d \end{pmatrix}$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

all eigenvalues ≥ 0

$Aq_i = \lambda_i q_i \quad \forall i$

What happens when we multiply a vector by this matrix A? Its the projection of v onto the basis given by the eigenvectors

QDQ.Tv = QD(q1 * v , q2 * v, …, qn * v).T —> qi * v is the component of v in the direction of qi

② Set $v_1 := q_1$

$v_2 := q_2$

$\vdots$

$v_k := q_k$

New coordinates of x_i in the basis of eigenvectors are $(x_i v_1), (x_i v_2), \ldots, (x_i v_k)$

low dimensional representation of $\vec{x}_i$

1. Find the top component, $\mathbf{v}_1$, using power iteration.
2. Project the data matrix orthogonally to $\mathbf{v}_1$:

$$\begin{bmatrix} — & \mathbf{x}_1 & — \\ — & \mathbf{x}_2 & — \\ & \vdots & \\ — & \mathbf{x}_m & — \end{bmatrix} \mapsto \begin{bmatrix} — & (\mathbf{x}_1 - \langle \mathbf{x}_1, \mathbf{v}_1 \rangle \mathbf{v}_1) & — \\ — & (\mathbf{x}_2 - \langle \mathbf{x}_2, \mathbf{v}_1 \rangle \mathbf{v}_1) & — \\ & \vdots & \\ — & (\mathbf{x}_m - \langle \mathbf{x}_m, \mathbf{v}_1 \rangle \mathbf{v}_1) & — \end{bmatrix}.$$

This corresponds to subtracting out the variance of the data that is already explained by the first principal component $\mathbf{v}_1$.

3. Recurse by finding the top $k-1$ principal components of the new data matrix.

$X - X v v^T$

Applications

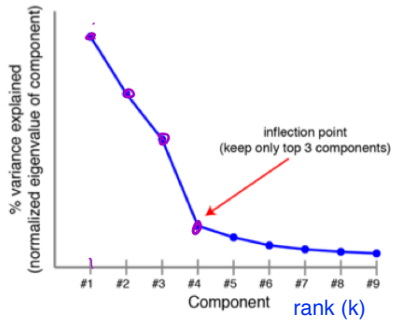
① Visualization ② Compression ③ Learning

How to choose k?

How do we choose what low rank approximation to use? What rank?

- For data visualization: a few
- compression: Look at eigenvalues. As soon as small enough, happy.

sum of eigenvalues / total sum



rank (k)

Scree plot. Principal components are ranked by the amount of variance they capture in the original dataset, a scree plot can provide some sense of how many components are needed.

Singular Value Decomposition (SVD)

gives us best way to approximate a matrix with a "low rank" matrix

$$\begin{bmatrix} 7 & ? & ? \\ ? & 8 & ? \\ ? & 12 & 6 \\ ? & ? & 2 \\ 21 & 6 & ? \end{bmatrix}.$$

Motivation: can we reconstruct missing entries?

construct the missing entries using as small as rank as possible (that matches the given entry)

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 7 & & \\ & 8 & \\ & 12 & 6 \\ & & 2 \\ 21 & 6 & \end{bmatrix} \end{array}$$

Rank 0

all zeros matrix

Rank 1

rank 1 matrix; every rows is a multiple of v.T; every column is a multiple of u

$$\mathbf{A} = \mathbf{u}\mathbf{v}^\top = \begin{bmatrix} — & u_1\mathbf{v}^\top & — \\ — & u_2\mathbf{v}^\top & — \\ & \vdots & \\ — & u_m\mathbf{v}^\top & — \end{bmatrix} = \begin{bmatrix} | & | & & | \\ v_1\mathbf{u} & v_2\mathbf{u} & \cdots & v_n\mathbf{u} \\ | & | & & | \end{bmatrix}$$

Rank 2

$$\mathbf{A} = \mathbf{u}\mathbf{v}^\top + \mathbf{w}\mathbf{z}^\top = \begin{bmatrix} — & u_1\mathbf{v}^\top + w_1\mathbf{z}^\top & — \\ — & u_2\mathbf{v}^\top + w_2\mathbf{z}^\top & — \\ & \vdots & \\ — & u_m\mathbf{v}^\top + w_m\mathbf{z}^\top & — \end{bmatrix} = \begin{bmatrix} | & | \\ \mathbf{u} & \mathbf{w} \\ | & | \end{bmatrix} \cdot \begin{bmatrix} — & \mathbf{v}^\top & — \\ — & \mathbf{z}^\top & — \end{bmatrix}$$

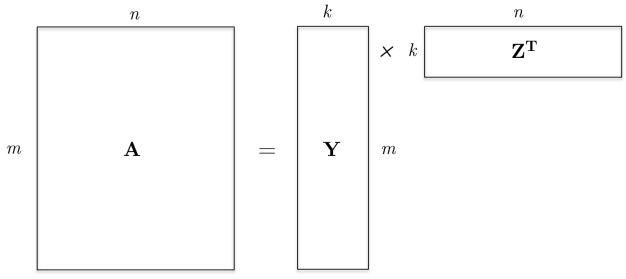


Figure 1: Any matrix $\mathbf{A}$ of rank k can be decomposed into a long and skinny matrix times a short and long one.

Why might a matrix be approximately low rank?

Example: movie ratings

Suppose each movie characterized by relatively small # of attributes

e.g. romance, violence, comedy, ...

and each person characterized by their preferences on each of these

Singular Value Decomposition (SVD)

If the matrix was square and symmetric this would be the eigendecomposition?

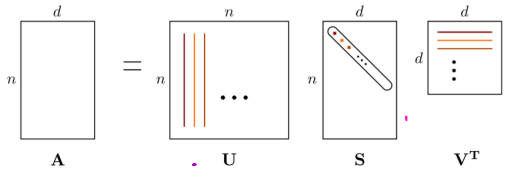


Figure 2: The singular value decomposition (SVD). Each singular value in $\mathbf{S}$ has an associated left singular vector in $\mathbf{U}$, and right singular vector in $\mathbf{V}$.

Running time O(min(n^2d), O(d^2n))

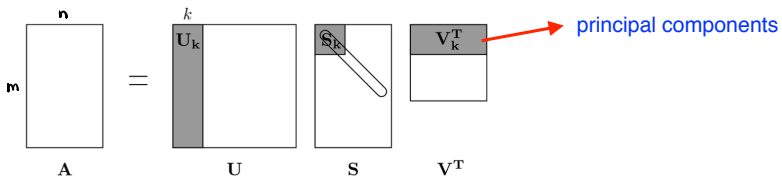


Figure 3: Low rank approximation via SVD. Recall that $\mathbf{S}$ is non-zero only on its diagonal, and the diagonal entries of S are sorted from high to low. Our low rank approximation is $\mathbf{A}_k = \mathbf{U}_k\mathbf{S}_k\mathbf{V}_k^\top$.

ll A - A_k ll_F ^ 2 <= ll A - B ll_F ^ 2 for any k rank matrix B

i.e. in the least squared sense, the SVD is the best rank k approximation possible.

exercise: Knowing that v_1 is the first principal component, prove this theorem for k = 1 (i.e. via the definition of PCA

Relationship between SVD and PCA

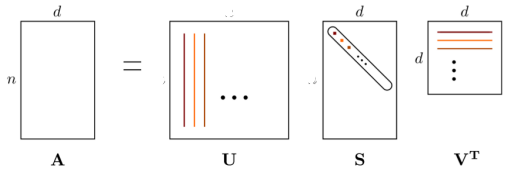


Figure 2: The singular value decomposition (SVD). Each singular value in $\mathbf{S}$ has an associated left singular vector in $\mathbf{U}$, and right singular vector in $\mathbf{V}$.

PCA

X^TX = QDQ^T

SVD

X = USV^T

X^TX = VS^TU^TUSV^T = VS^2V^T

Thus Q = V and D = S^2 so the eigenvalues are just the squares of the singular values from SVD!

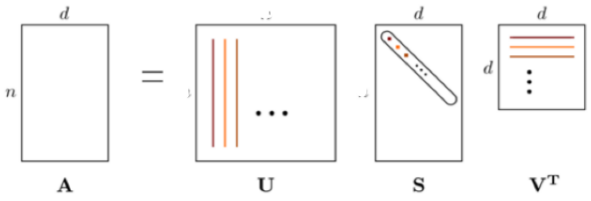


Figure 2: The singular value decomposition (SVD). Each singular value in $\mathbf{S}$ has an associated left singular vector in $\mathbf{U}$, and right singular vector in $\mathbf{V}$.

Application 1: Denoising

Suppose a matrix A is a noisy version of a rank k matrix

A = C + N where C is the true rank k structure and N is noise

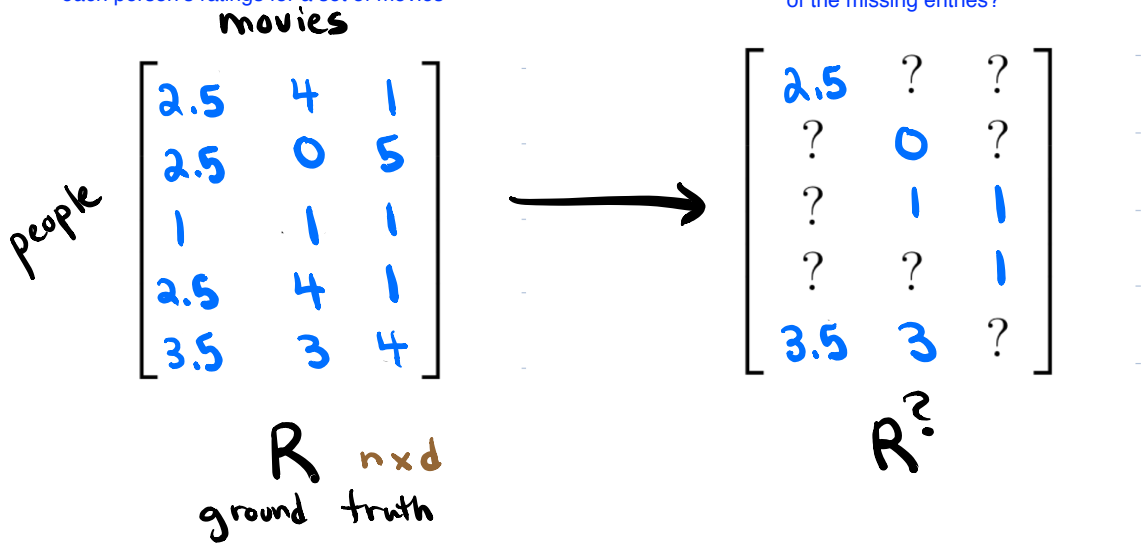
We can “reconstruct” C by considering a rank k approximation of A

Collaborative Filtering

recomendations: which movies to see, which products to buy

Model:

each person's ratings for a set of movies



Assumptions:

R is low rank

e.g. rank k

example: humor, violence, romance

The matrix completion problem:

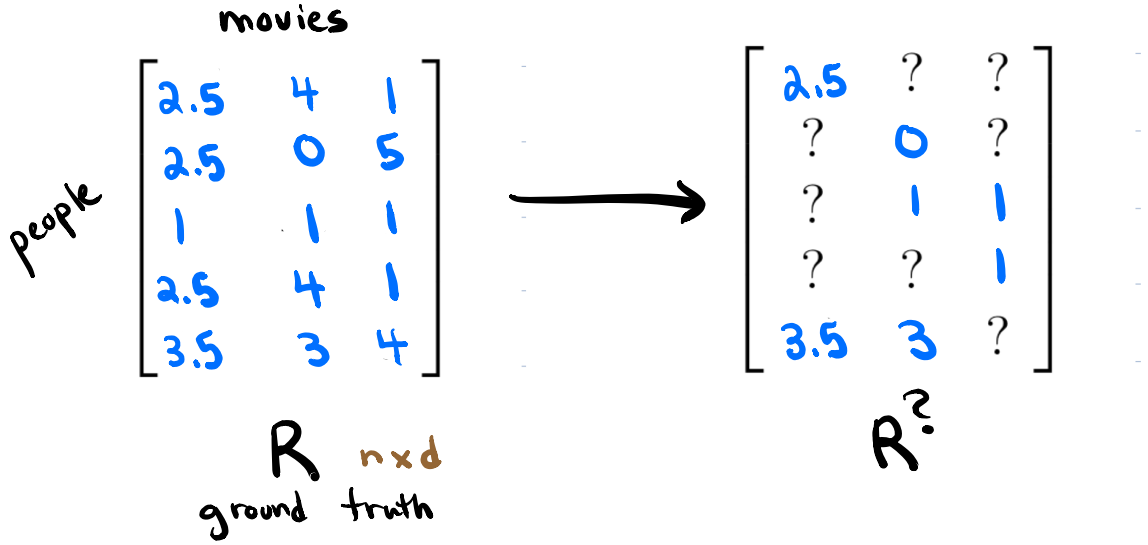
$$R = h \left(\begin{matrix} \\ -\ a_i\ - \end{matrix} \right)^{k} \; k \left(\begin{matrix} d \\ | \\ bj \\ | \end{matrix} \right)$$

Find ai with i = 1, …, n and bj with j = 1, …, d
such that R captures all the known entries as good as possible

Collaborative Filtering

recomendations: which movies to see, which products to buy

Model:



R $n \times d$ ground truth

$R^?$

Assumptions:

① R is low rank, e.g. rank k

example: humor, violence, romance

The matrix completion problem:

fill in the missing entries.

each entry is a random variable; all we see is some sample

Theorem $\hat{A}$ $n \times d$ matrix of indep. r.v.s, each with variance bounded by σ^2

If $A = E(\hat{A})$ is rank k

then with high probability

$$\|A - \hat{A}_k\|_F^2 \text{ is } O(k\sigma^2(n+d))$$

k rank approximation of A_hat is quite close to A

average per element error

avg per elt error $O\left(\frac{k\sigma^2(n+d)}{nd}\right) = o(1)$

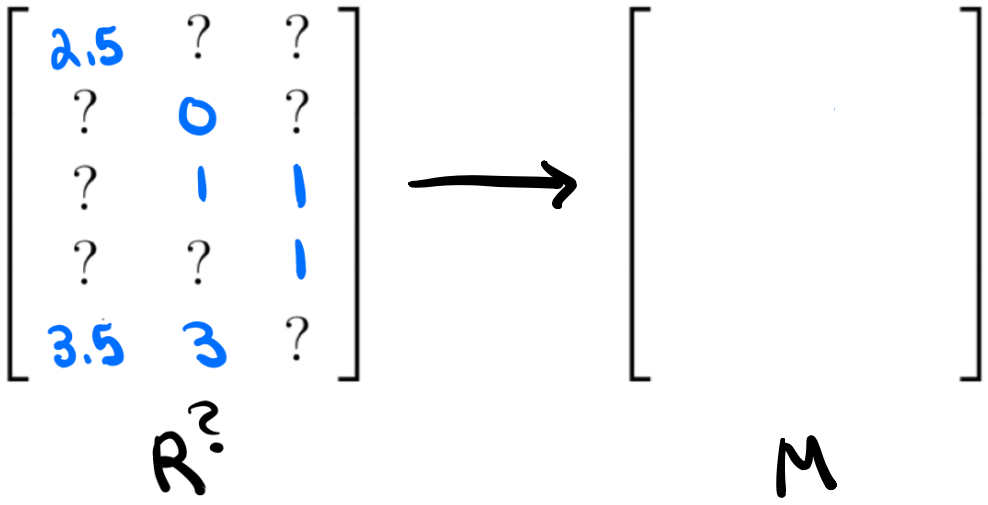
average error per element is constant?

Suppose we have a matrix A that is rank k and that entry (i,j) is deleted with probability 1-p_ij (i.e. goes to a ? in example below ; retained with probability p_ij)

All we see are the nondeleted entries according to this random process (R? below) Then I can essentially reconstruct A

Let \hat{A_ij} = { A_ij / p_ij if entry is present and 0 otherwise

E[\hat{A}_ij] = A_ij the true value! Thus \hat{A} is an unbiased estimator of our true matrix A. A = E[\hat{A}]. By the theorem above then, taking the low-rank approximation (of rank k) of \hat{A} we get a good approximation of A!


2.5 ? ?
? 0 ?
? 1 1
? ? 1
3.5 3 ?
R?
M