

Singular Value Decomposition (SVD)

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$$

Singular Valued Decomposition, SVD, general linear maps

1. **Singular Value Decomposition**
2. General Linear Mappings
3. Linear Regression Revisited

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$$

Singular Valued Decomposition

- Consider an arbitrary $n \times m$ matrix X , and construct the $(n + m) \times (n + m)$ symmetric matrix, B ,

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$$

$\begin{pmatrix} u \\ v \end{pmatrix}$ is an eigenvector of B with eigenvalue s

- We observe that

$$\begin{aligned} Xv &= su \\ X^T Xv &= sX^T u = s^2 v \\ X^T u &= sv \\ XX^T u &= sXv = s^2 u \end{aligned}$$

Eigenvectors

- Note that as $Xv = su$ and $X^T u = sv$ then

$$X(-v) = (-s)u \quad X^T u = (-s)(-v)$$

if $\begin{pmatrix} u \\ v \end{pmatrix}$ is an eigenvector of B with eigenvalue s then so is $\begin{pmatrix} u \\ -v \end{pmatrix}$ with eigenvalue $-s$

- If $n < m$ then $X^T X$ is not full rank so some eigenvalues are zero
- As a consequence $m - n$ vectors exist such that $Xv = 0$
- The eigenvalues and eigenvectors are

$$n \times \left(s_i, \begin{pmatrix} u_i \\ v_i \end{pmatrix} \right) \quad n \times \left(-s_i, \begin{pmatrix} u_i \\ -v_i \end{pmatrix} \right) \quad m - n \times \left(0, \begin{pmatrix} 0 \\ v_k \end{pmatrix} \right)$$

Matrix Decomposition

- Stacking the eigenvectors into a matrix

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} = \begin{pmatrix} U & U & 0 \\ V & -V & V_0 \end{pmatrix} \begin{pmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- Since the vectors $\begin{pmatrix} u_i \\ v_i \end{pmatrix}$ are eigenvectors of a symmetric matrix they form an orthogonal matrix if they are normalised.
- Multiply on the right by the transpose of the orthogonal matrix

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} = \begin{pmatrix} U & U & 0 \\ V & -V & V_0 \end{pmatrix} \begin{pmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} U^T & V^T \\ U^T & -V^T \\ 0 & V_0^T \end{pmatrix}$$

SVD

- Any matrix, X , can be written as $X = USV^T$
 - U, V are orthogonal matrices
 - $S = \text{diag}(s_1, s_2, \dots, s_n)$
- s_i can always be chosen to be positive and are known as **singular values**
- Singular value decomposition applies to both square and non-square matrices—they describe general linear mappings

Normalisation Subtlety

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} = \begin{pmatrix} U & U & 0 \\ V & -V & V_0 \end{pmatrix} \begin{pmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} U^T & V^T \\ U^T & -V^T \\ 0 & V_0^T \end{pmatrix}$$

- Multiplying out we have

$$X = 2USV^T \quad X^T = 2VSU^T$$

- Now the vectors u_i and v_i form an orthogonal set as it satisfy

$$X^T X v = s^2 v \quad X X^T u = s^2 u$$

- But they are not normalised (since $\begin{pmatrix} u_i \\ v_i \end{pmatrix}$ is normalised). If we define $\tilde{U} = \sqrt{2}U$ and $\tilde{V} = \sqrt{2}V$ we find

$$X = \tilde{U} S \tilde{V}^T \quad X^T = \tilde{V} S \tilde{U}^T$$

Finding SVD

- Most libraries will compute the SVD for you
- They can do this by choosing the smaller of two matrices XX^T and $X^T X$ and then compute the eigenvalues
- The singular values are the square root of the eigenvalues (notice that XX^T and $X^T X$ are both positive semi-definite so the eigenvalues will be non-negative)
- It can compute the U matrix or V matrix by multiplying through by X or X^T ($U = XVS^{-1}$ and $V = X^T US^{-1}$)
- In practice to perform PCA most people subtract the mean from their data and then perform SVD

Economical Forms of SVD

- Often the rows or columns of the orthogonal matrices \mathbf{U} and \mathbf{V} that are not associated with a singular value are ignored

$$\begin{pmatrix} \mathbf{X} \\ \text{---} \end{pmatrix} = \begin{pmatrix} \mathbf{U} \\ \text{---} \end{pmatrix} \begin{pmatrix} \mathbf{S} \\ 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}^T \\ \text{---} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{X} \\ \text{---} \end{pmatrix} = \begin{pmatrix} \mathbf{U} \\ \text{---} \end{pmatrix} \begin{pmatrix} \mathbf{S} \\ 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}^T \\ \text{---} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{X} \\ \text{---} \end{pmatrix} = \begin{pmatrix} \mathbf{U} \\ \text{---} \end{pmatrix} \begin{pmatrix} \mathbf{S} \\ 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}^T \\ \text{---} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{X} \\ \text{---} \end{pmatrix} = \begin{pmatrix} \mathbf{U} \\ \text{---} \end{pmatrix} \begin{pmatrix} \mathbf{S} \\ 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}^T \\ \text{---} \end{pmatrix}$$

- In Matlab these are obtained using

```
>> [U, S, V] = svd(X)
>> [U, S, V] = svd(X, 'econ')
```

General Matrix

- Recall that we can compute the SVD for any matrix, \mathbf{X}
- As matrices describe the most general linear mapping

$$\mathbf{v} \rightarrow \mathcal{T}[\mathbf{v}] = \mathbf{X}\mathbf{v}$$

- We can use SVD to understand any linear mapping
- Thus any linear mapping can be seen as a rotation followed by a squashing or expansion independently in each coordinate followed by another rotation

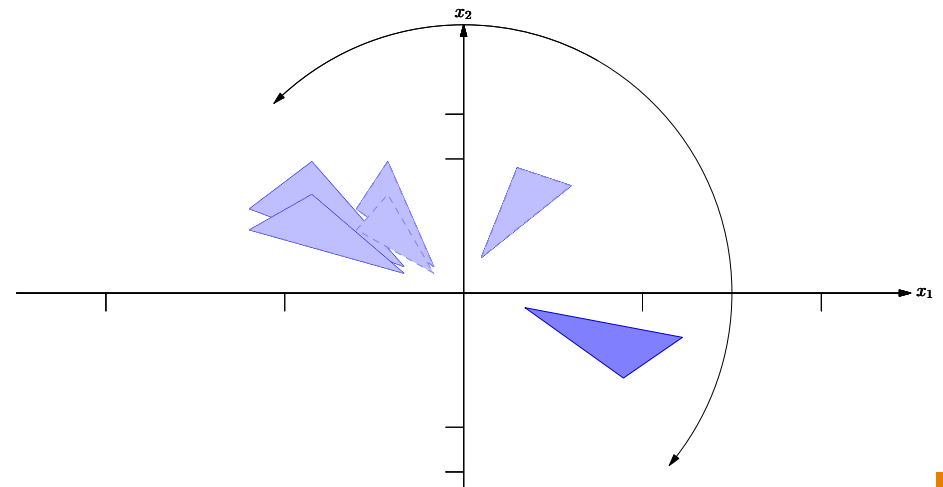
Outline

- Singular Value Decomposition
- General Linear Mappings
- Linear Regression Revisited

$$\begin{pmatrix} 0 & \mathbf{X} \\ \mathbf{X}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = s \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$$

Matrices

$$\mathbf{M} = \begin{pmatrix} -0.45 & 1.9 \\ -0.77 & -0.025 \end{pmatrix} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \begin{pmatrix} \cos(-175) & \sin(-175) \\ -\sin(-175) & \cos(-175) \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 0.75 \end{pmatrix} \begin{pmatrix} \cos(75) & \sin(75) \\ -\sin(75) & \cos(75) \end{pmatrix}$$



Determinants

- The determinant, $|\mathbf{M}|$ of a matrix \mathbf{M} is defined for square matrices
- It describes the change in volume under the mapping
- Now for any two matrices $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
- Thus

$$|\mathbf{M}| = |\mathbf{U}||\mathbf{S}||\mathbf{V}^T|$$

- For an orthogonal matrix $|\mathbf{U}| = \pm 1$
- Thus

$$|\mathbf{M}| = \pm |\mathbf{S}| = \pm \prod_i s_i$$

Non-Square Matrices

- When the matrices are non-square then the matrix of singular value matrix will either
 - ★ Squash some directions to zero
 - ★ Introduce new dimensions orthogonal to the vector

$$\begin{pmatrix} \mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{U} \end{pmatrix} \begin{pmatrix} \mathbf{S} \end{pmatrix} \begin{pmatrix} \mathbf{V}^T \end{pmatrix} \quad \begin{pmatrix} \mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{U} \end{pmatrix} \begin{pmatrix} \mathbf{S} \end{pmatrix} \begin{pmatrix} \mathbf{V}^T \end{pmatrix}$$

- The rank of an arbitrary matrix is the number of non-zero singular values (also number of linearly independent rows or columns)

Duality Revisited

- If $\mathbf{X} = \mathbf{USV}^T$ then

$$\begin{aligned} \mathbf{C} &= \mathbf{XX}^T & \mathbf{D} &= \mathbf{X}^T\mathbf{X} \\ &= \mathbf{USV}^T\mathbf{VS}^T\mathbf{U}^T & &= \mathbf{VS}^T\mathbf{U}^T\mathbf{USV}^T \\ &= \mathbf{U}(\mathbf{SS}^T)\mathbf{U}^T & &= \mathbf{V}(\mathbf{S}^T\mathbf{S})\mathbf{V}^T \end{aligned}$$

- If \mathbf{X} is an $p \times m$ matrix then \mathbf{SS}^T is a $p \times p$ diagonal matrix with elements $S_{ii}^2 = s_i^2$
- $\mathbf{S}^T\mathbf{S}$ is an $m \times m$ matrix with elements $S_{ii}^2 = s_i^2$
- \mathbf{U} and \mathbf{V} are matrices of eigenvectors for \mathbf{C} and \mathbf{D}
- The eigenvalues are $\lambda_i = S_{ii}^2 = s_i^2$

\mathbf{SS}^T and $\mathbf{S}^T\mathbf{S}$

$$\mathbf{S} = \begin{pmatrix} s_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_m & 0 & 0 \dots & 0 \end{pmatrix}$$

$$\mathbf{S}^T\mathbf{S} = \begin{pmatrix} s_1^2 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & s_2^2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_m^2 & 0 & 0 \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 \dots & 0 \end{pmatrix}$$

$$\mathbf{SS}^T = \begin{pmatrix} s_1^2 & 0 & \dots & 0 \\ 0 & s_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_m^2 \end{pmatrix}$$

Having A Go

- It's really easy to verify this in MATLAB or OCTAVE

```
>> X = rand(3,2)
>> [U, S, V] = svd(X)
>> U*S*V'
>> U(:,1)'*U(:,2)
>> U'*U
>> U*U'
>> [Ua,L] = eig(X*X')
>> S*S'
```

- Test yourself!

Outline

- Singular Value Decomposition
- General Linear Mappings
- Linear Regression Revisited

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} u \\ v \end{pmatrix}$$

Linear Regression

- Given a set of data $\mathcal{D} = \{(x_i, y_i) | k = 1, 2, \dots, m\}$
- In linear regression we try to fit a linear model

$$f(x|w) = x^T w$$

- Which we fit by minimising the squared error loss

$$L(w) = \sum_{k=1}^m (f(x_i|w) - y_i)^2$$

Matrix Form

- In matrix form we write $L(w) = \|Xw - y\|^2$

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

- Then $\nabla L(w^*) = 0$ implies

$$w^* = (X^T X)^{-1} X^T y = X^+ y$$

- This is known as the pseudo-inverse

Using SVD

- Using $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ then

$$\begin{aligned}\mathbf{X}^+ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ &= (\mathbf{V} \mathbf{S}^T \mathbf{S} \mathbf{V}^T)^{-1} \mathbf{V} \mathbf{S}^T \mathbf{U}^T \\ &= \mathbf{V} (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{V}^T \mathbf{V} \mathbf{S}^T \mathbf{U}^T \\ &= \mathbf{V} (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \mathbf{U}^T = \mathbf{V} \mathbf{S}^+ \mathbf{U}^T\end{aligned}$$

- If $m > p$

$$\mathbf{X}^T = \begin{pmatrix} | & | & | & | & | & | & | & | \\ s_1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & s_2 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & s_3 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_p & 0 & 0 & \dots & 0 \end{pmatrix}, \mathbf{S}^T = \begin{pmatrix} s_1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & s_2 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & s_3 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_p & 0 & 0 & \dots & 0 \end{pmatrix}$$

Pseudo-Inverse of S

$$\mathbf{S}^T \mathbf{S} = \begin{pmatrix} s_1^2 & 0 & \dots & 0 \\ 0 & s_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_p^2 \end{pmatrix}, (\mathbf{S}^T \mathbf{S})^{-1} = \begin{pmatrix} s_1^{-2} & 0 & \dots & 0 \\ 0 & s_2^{-2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_p^{-2} \end{pmatrix}$$

$$\mathbf{S}^+ = (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T = \begin{pmatrix} s_1^{-1} & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & s_2^{-1} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & s_3^{-1} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_p^{-1} & 0 & 0 & \dots & 0 \end{pmatrix}$$

III-Conditioned Data Matrix

- Recall that

$$\mathbf{w}^* = \mathbf{X}^+ \mathbf{y} = \mathbf{V} \mathbf{S}^+ \mathbf{U}^T \mathbf{y}$$

- If any of the singular values of \mathbf{X} are small then \mathbf{S}^+ will magnify components in that direction
- Any errors in the target \mathbf{y} will be magnified
- This leads to poor weights

Regularisation

- Consider linear regression with a regulariser

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \eta \|\mathbf{w}\|^2 \\ &= \mathbf{w}^T (\mathbf{X}^T \mathbf{X} + \eta \mathbf{I}) \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}\end{aligned}$$

- Thus

$$\nabla \mathcal{L}(\mathbf{w}) = 2(\mathbf{X}^T \mathbf{X} + \eta \mathbf{I}) \mathbf{w} - 2\mathbf{X}^T \mathbf{y}$$

- and $\nabla \mathcal{L}(\mathbf{w}^*) = 0$ gives

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X} + \eta \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Regularisation Continued

- Using $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$

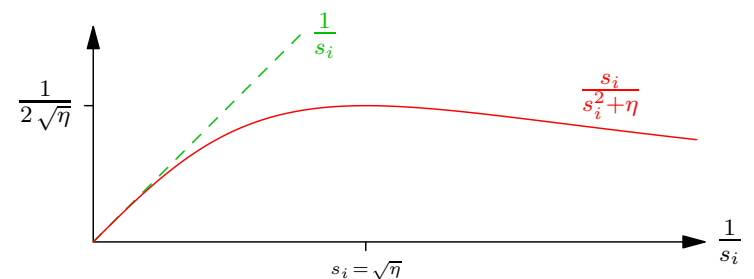
$$\begin{aligned} \mathbf{w}^* &= (\mathbf{X}^T\mathbf{X} + \eta\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{V}(\mathbf{S}^T\mathbf{S} + \eta\mathbf{I})^{-1}\mathbf{S}^T\mathbf{U}^T\mathbf{y} \end{aligned}$$

- where

$$(\mathbf{S}^T\mathbf{S} + \eta\mathbf{I})^{-1}\mathbf{S}^T = \begin{pmatrix} \frac{s_1}{s_1^2 + \eta} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{s_2}{s_2^2 + \eta} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{s_3}{s_3^2 + \eta} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{s_p}{s_p^2 + \eta} & 0 & 0 & \cdots & 0 \end{pmatrix}$$

Effect of Regularisation

- Without regularisation if $s_i = 0$ the problem would be ill-posed (even \mathbf{S}^+ does not exist since s_i^{-1} would be ill defined) and if s_i is small then \mathbf{S}^+ is ill conditioned
- Using $\hat{\mathbf{S}}^+ = (\mathbf{S}^T\mathbf{S} + \eta\mathbf{I})^{-1}\mathbf{S}^T$ instead of \mathbf{S}^+ then



- Regularisation makes the machine much more stable (reduces the variance)

Summary

- Any matrix can be decomposed as $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where
 - ★ \mathbf{U} and \mathbf{V} are orthogonal (rotation matrices)
 - ★ $\mathbf{S} = \text{diag}(s_1, \dots, s_n)$ is a diagonal matrix of positive singular values
- This describes the most general linear transform
- The transform exploits the duality between $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$
- In linear regression the pseudo-inverse involves the reciprocal of the singular values, which can lead to poor generalisation
- Regularisation improves the conditioning of the “inverse” matrix