COMP 3223/COMP6245: Foundations of Machine Learning
Linear Algebra
Singular Value Decomposition

Srinandan Dasmahapatra

2018/9
Matrices

You should all know the following

- Matrix notation
- Matrix transpose
- Scalar multiplication
- Matrix addition & multiplication
- Matrix inverse
- System of linear equations in matrix form
- Matrix determinant
- Eigenvalues and eigenvectors
In linear regression the model is $\hat{y} = A w$:

$$
w_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + w_1 \begin{pmatrix} \phi_1(x_1) \\ \phi_1(x_2) \\ \vdots \\ \phi_1(x_N) \end{pmatrix} + \cdots + w_p \begin{pmatrix} \phi_p(x_1) \\ \phi_p(x_2) \\ \vdots \\ \phi_p(x_N) \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}
$$

Find best weighted linear combination of columns of design matrix that spans $y$.

The residual must not be in the space spanned by the columns of $A$ – residual orthogonal to each column.
Design matrix has information on patterns in data

- In linear regression the model is $\hat{y} = A w$:

$$
\begin{pmatrix}
1 & \phi_1(x_1) & \cdots & \phi_p(x_1) \\
1 & \phi_1(x_2) & \cdots & \phi_p(x_2) \\
\vdots & \vdots & \ddots & \vdots \\
1 & \phi_1(x_N) & \cdots & \phi_p(x_N)
\end{pmatrix}
$$

- Idea of this lecture: decompose matrix using transformations and data appropriate descriptive bases
Reminder: Solving Linear Equations – Geometrical Picture

- Solve set of equations:

\[
\begin{pmatrix}
a & b \\
c & d \\
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
\end{pmatrix}
=
\begin{pmatrix}
r \\
s \\
\end{pmatrix}
\]

- Geometrically viewed as intersection of linear linear combination of vectors:

\[
\begin{align*}
ax + by &= r \\
\underline{cx + dy} &= s
\end{align*}
\]

\[
\begin{align*}
x \begin{pmatrix} a \\ c \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix} &= \begin{pmatrix} r \\ s \end{pmatrix}
\end{align*}
\]
The Geometrical Picture: An example

- Solve set of equations:
  \[
  \begin{pmatrix}
  -2 & 1 \\
  -1 & 3
  \end{pmatrix}
  \begin{pmatrix}
  x \\
  y
  \end{pmatrix}
  =
  \begin{pmatrix}
  4 \\
  -2
  \end{pmatrix}
  \]

- red vectors are columns of matrix

- Solution of \( y - 2x = 4, \ 3y - x = -2 \), is \((x, y) = (-2.8, -1.6)\).
Fundamental operations on vectors – multiply by scalars and perform addition

Multiply red vectors by numbers (elements of a field) and add vectors together
Systems of equations: underdetermined/overdetermined?

- Instead of solving
  \[
  \begin{pmatrix}
  -2 & 1 \\
  -1 & 3 \\
  \end{pmatrix}
  \begin{pmatrix}
  x \\
  y \\
  \end{pmatrix}
  =
  \begin{pmatrix}
  4 \\
  -2 \\
  \end{pmatrix},
  \text{ with soln. } (-2.8, -1.6)
  \]

- Solve
  \[
  \begin{pmatrix}
  -2 & 1 \\
  \end{pmatrix}
  \begin{pmatrix}
  x \\
  y \\
  \end{pmatrix}
  =
  \begin{pmatrix}
  4 \\
  \end{pmatrix}
  \iff
  \begin{pmatrix}
  -2 & 1 \\
  -2 & 1 \\
  \end{pmatrix}
  \begin{pmatrix}
  x \\
  y \\
  \end{pmatrix}
  =
  \begin{pmatrix}
  4 \\
  4 \\
  \end{pmatrix}
  \]

- What about
  \[
  \begin{pmatrix}
  -2 & 1 \\
  -1 & 3 \\
  0 & 5 \\
  \end{pmatrix}
  \begin{pmatrix}
  x \\
  y \\
  \end{pmatrix}
  =
  \begin{pmatrix}
  4 \\
  -2 \\
  4 \\
  \end{pmatrix}
  \text{? least sq.: } (-0.8, 0.4)
  \]
Thus $A\mathbf{x} = \mathbf{y}$ can be solved if and only if $\mathbf{y}$ is a linear combination of columns of $A$.

- The column space of a matrix $A$ (denoted $\text{col } A$) is the subspace spanned by all linear combinations of the columns of $A$.
- This is also the range of the linear map: $\text{range}(A) = AV = \{ \mathbf{w} \in \mathbf{W} : \mathbf{w} = A\mathbf{v} \text{ for some } \mathbf{v} \in \mathbf{V} \}$
Examples illustrating linear dependence and nullspace

Let \( B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \). For vector \( \mathbf{v} \) in direction \( \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \), \( B\mathbf{v} = 0. \mathbf{v} \) in nullspace or kernel of \( B \).

For \( A = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \), \( \text{col}(1) + \text{col}(2) = \text{col}(3) \), so

\[
\begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \ker(A) = c \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}
\]

Show \( \ker(A^T) = c \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix} \).
Kernel or Null space of a matrix

- In the previous example $A = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, there are 3 variables $v$ in $Av = y$ but only two independent equations.
- If $Av = y$ and $x \in \ker(A)$ then $A(v + x) = y$. Either there are no solutions or there are (infinitely) many solutions.
- The kernel of a map (or matrix) $\ker(A) = \text{nullspace } A = \{v \in V : Av = 0\}$.
- Let $A$ be a $3 \times q$ matrix.

$$A = \begin{pmatrix} - & u & - \\ - & v & - \\ - & w & - \end{pmatrix},$$

where $u$, $v$ and $w$ are $q$-dim row vectors. Then, $x \in \ker(A) \iff Ax = 0$. This means $x \perp \{u, v, w\}$. 

Rank of a matrix = number of independent equations

- The rank (column rank) of $A$ is the dimension of the column space of $A$.
- A vector space is partitioned into its range and null spaces:

$$\dim V = \dim \ker(A) + \dim \text{range}(A).$$

- We can do the same for the transpose: $\text{col}(A^T)$ and $\ker(A^T)$.
- 4 fundamental subspaces: $\text{col}(A)$, $\ker(A^T)$, $\text{col}(A^T)$ and $\ker(A)$.
Four fundamental subspaces of a matrix

Linear regression with fixed functions of data

- Linear combinations of fixed functions $\phi_j(x_n)$:

$$w_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + w_1 \begin{pmatrix} \phi_1(x_1) \\ \phi_1(x_2) \\ \vdots \\ \phi_1(x_N) \end{pmatrix} + \cdots + w_p \begin{pmatrix} \phi_p(x_1) \\ \phi_p(x_2) \\ \vdots \\ \phi_p(x_N) \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

- Sets of functions constitute vector spaces

- Approximate outputs/targets $\mathbf{y}$ by element of column space of the design matrix.
Functions constitute vector spaces

- $\mathbb{R}[x]$, the space of polynomials $\sum_m a_m x^m$, where $a_m \in \mathbb{R}$ forms a vector space:

\[
(a_0 + a_1 x + a_2 x^2) + (b_0 + b_1 x) = (a_0 + b_0) + (a_1 + b_1) x + a_2 x^2
\]

Monomials as basis elements $a + b = c$:

\[
(a_0, a_1, a_2) + (b_0, b_1, 0) = (a_0 + b_0, a_1 + b_1, a_2) = (c_0, c_1, c_2)
\]

- Similarly, the set $\mathbb{R}[x_1, \ldots, x_k]$ of polynomials in $k$ variables forms a vector space.
- Set of functions of the form $\sum_{|n|<N} a_n e^{in\theta}$ (Fourier series).
- Extension – replace sums (where the summation index is from a discrete set) by integrals (where the index being summed over is now continuous)
Matrices form a vector space: multiply $n \times m$ matrices $A$ with entries $a_{ij} \in \mathbb{R}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ by scalars and add any two such matrices together:

$$3 \begin{pmatrix} -2 & 1 \\ -1 & 4 \end{pmatrix} - 2 \begin{pmatrix} 2 & 2 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} -10 & -1 \\ -1 & 0 \end{pmatrix}.$$
 Reminder: Linear combination and dependence

Linear combination of vectors: \( \mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i, \ \alpha_i \in \mathbb{F}, \mathbf{v}_i \in V. \)

- The vectors in the figure are linear combinations of \( \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \) and \( \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \). They are in the span of \( \{ \mathbf{e}_1, \mathbf{e}_2 \} \).
- \( \mathbf{v} = a_1 \mathbf{e}_1 + a_2 \mathbf{e}_2 = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \) can be zero iff \( a_1 = 0 = a_2 \).
A set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are called **linearly independent** if none of them can be represented as a linear combination of the others:

$$\mathbf{v}_k \neq \sum_{i \neq k} c_i \mathbf{v}_i, \quad \text{for any } c_i \in \mathbb{F}$$

Equivalently, condition for a set of vectors $\{\mathbf{v}_i\}_i$ to be linearly independent:

If $\sum_{i=1}^{n} \alpha_i \mathbf{v}_i = 0$, then $\alpha_i = 0$ for all $i$.

A **basis** for $V$ is a set $B \subset V$ which is both spanning and independent. A finite dimensional vector space has a finite basis, and its dimension $\dim V$ is the number of elements in $B$. 
Linear dependence & Linear Regression

- In linear regression the model is $\hat{y} = A\mathbf{w}$:

\[
\begin{pmatrix}
1 \\
1 \\
\vdots \\
1
\end{pmatrix}
+ w_1
\begin{pmatrix}
\phi_1(x_1) \\
\phi_1(x_2) \\
\vdots \\
\phi_1(x_N)
\end{pmatrix}
+ \cdots + w_p
\begin{pmatrix}
\phi_p(x_1) \\
\phi_p(x_2) \\
\vdots \\
\phi_p(x_N)
\end{pmatrix}
= \begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_N
\end{pmatrix}
\]

- Find best weighted linear combination of columns of design matrix that spans $\mathbf{y}$.

- The residual must not be in the space spanned by the columns of $A$ – orthogonal to it.
Dot Products, Orthogonality and Norms

- We can associate, with two vectors $\mathbf{v}$ and $\mathbf{w}$ an element of $\mathbb{R}$ called their **scalar** (or **dot**) **product**:

  \[ \text{dot : } V \times V \rightarrow \mathbb{R} \]

  \[ \text{dot}(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w} \rightarrow a \]

- Two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are called **orthogonal** if their dot product is zero, i.e. $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$. If $k$ vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are mutually orthogonal, i.e. $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $i \neq j$, they are called an **orthogonal set**.

- Euclidean norm: for $\mathbf{v} \in V$, dim $V = N$,

  \[ \|\mathbf{v}\|_2 := \sqrt{\mathbf{v}^T \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_N^2} = |\mathbf{v}| \]

- If all vectors are of unit length $|v_i| = 1$, the set is called **orthonormal**.
Using dot products to introduce projections

- **Project** a vector \( y \) on a direction given by a vector \( u \)

[Diagram showing vector projection]

- The **projection** is given by the value \( w \) (length \( \overrightarrow{0w} \)). Note, \( w \) could be negative if \( \alpha \) is bigger than 90\(^\circ\). From the figure, we see that

\[
w = |y| \cos \alpha = \frac{y \cdot u}{|u|},
\]

because the **dot** product is \( y \cdot u = |u||y| \cos \alpha \).
Example: expand vector in orthogonal basis

- Let \( e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \), \( e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \). Expand \( \mathbf{v} = \begin{pmatrix} -5 \\ 3 \end{pmatrix} \) as a linear combination of the set \( \{e_i\} \), i.e. find numbers \( \alpha_1, \alpha_2 \) such that

\[
\mathbf{v} = \sum_{i=1}^{n} \alpha_i e_i
\]

- Solution: Multiply \( \mathbf{v} \) by \( e_j \), use orthogonality \( (e_1 \cdot e_2 = 0) \):
  \( e_1 \cdot \begin{pmatrix} -5 \\ 3 \end{pmatrix} = -5 \), \( e_2 \cdot \begin{pmatrix} -5 \\ 3 \end{pmatrix} = 3 \).

\[
\begin{pmatrix} -5 \\ 3 \end{pmatrix} = (-5)\begin{pmatrix} 1 \\ 0 \end{pmatrix} + (3)\begin{pmatrix} 0 \\ 1 \end{pmatrix}
\]
Expanding a vector in a set of orthogonal vectors

Let \( \{ \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \} \) be a set of orthogonal \( n \times 1 \) column vectors and \( \mathbf{v} \) an arbitrary \( n \times 1 \) column vector.

Task: Expand \( \mathbf{v} \) as linear combination of set \( \{ \mathbf{v}_i \} \), ie. find numbers \( \alpha_1, \alpha_2, \ldots, \alpha_n \) s.t. \( \mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i \).

Solution: Multiply \( \mathbf{v}_j \cdot (\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i) \) use \( \mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij} \| \mathbf{v}_i \|^2 \) (orthogonality) to get \( \mathbf{v}_j \cdot \mathbf{v} = \sum_{i=1}^{n} \alpha_i \delta_{ij} = \alpha_j \). Explicitly,

\[
\begin{align*}
\mathbf{v}_j \cdot \mathbf{v} & = \alpha_1 \mathbf{v}_j \cdot \mathbf{v}_1 + \cdots + \alpha_j \mathbf{v}_j \cdot \mathbf{v}_j + \cdots + \alpha_n \mathbf{v}_j \cdot \mathbf{v}_n \\
& = \alpha_j \mathbf{v}_j \cdot \mathbf{v}_j
\end{align*}
\]

Hence

\[
\alpha_j = \frac{\mathbf{v}_j \cdot \mathbf{v}}{\mathbf{v}_j \cdot \mathbf{v}_j} = \frac{\mathbf{v}_j \cdot \mathbf{v}}{\| \mathbf{v}_j \|^2}
\]

Seek to characterise design matrix in terms of some orthonormal bases
Real symmetric matrices appear a lot in ML

- A matrix $A$ is called *real symmetric* if all matrix elements are real numbers and $A^T = A$, where $A^T$ is the transpose of $A$.
- For a matrix $A$ with elements $(A)_{ij} = a_{ij}$, $A^T$ is defined as $(A^T)_{ij} = a_{ji}$.
- *The eigenvalues of a real symmetric matrix are real.*
- Examples of symmetric matrices: covariance matrices and kernel/gram matrices; generated from data sets, used extensively in ML algorithms.
Eigenvectors of real symmetric matrices are orthogonal

- **Let** \( A \) **be a real symmetric matrix. Then eigenvectors associated with distinct eigenvalues are orthogonal.**

- **Proof:** Let \( u \) and \( v \) be two eigenvectors with distinct eigenvalues \( \lambda \) and \( \mu \) respectively, i.e \( Au = \lambda u \) and \( Av = \mu v \). We shall prove \( u^T v = 0 \).

  Since \( A \) is symmetric, \( (Au)^T = u^T A \). Therefore \( Au = \lambda u \iff u^T A^T = \lambda u^T \). Multiply on the right with \( v \):

  \[
  \lambda u^T v = u^T Av = \mu u^T v.
  \]

  Hence \( (\lambda - \mu)u^T v = 0 \) and, since \( \mu \neq \lambda \), \( u^T v = 0 \).

- Can find orthogonal eigenvectors for real symmetric matrix, even with repeated (degenerate) eigenvalues.
Design matrix is not square

- The domain and range of matrix have different dimensions
- Need descriptive basis for each: action of matrix on vector space pieced together from its action on orthonormal basis
- For instance, each feature corresponds to a direction represented by a unit vector:

\[
Aw = w_0A \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + w_1A \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + w_pA \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}
\]

- Generalise notion of eigenvalue/eigenvector pair
- Introduce **singular value decomposition (SVD)**
Singular Value Decomposition (SVD) of a Matrix

- The action of an arbitrary matrix on a vector space can be pieced together from its action on an orthonormal basis in that vector space.
- SVD measures how a circle is mapped into an ellipse; how an $n$-dimensional hyper-sphere is mapped into an $n$-dimensional hyper-ellipse.

Action of $\begin{pmatrix} 1.0 & 2.0 \\ 0.5 & 2.5 \end{pmatrix}$ on unit vectors $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. The lengths of the semi-major axes of the hyper-ellipse are properties of the map.
Even when the vectors in the domain and range of the map change, their locus displays the geometrical character of the transformation enacted by the matrix.

While the displayed pairs of vectors in the domain (red) are orthogonal by construction, the pairs they map to (blue) are not.
Example of SVD

The action of an arbitrary matrix on a vector space can be pieced together from its action on an orthonormal basis \( \{ \mathbf{v}_1, \mathbf{v}_2 \} \) in that vector space, here 2-dimensional. So, \( \mathbf{w} = (\mathbf{v}_1^T \mathbf{w}) \mathbf{v}_1 + (\mathbf{v}_2^T \mathbf{w}) \mathbf{v}_2 \).

The range of \( \mathbf{A} \) is spanned by \( \mathbf{u}_1, \mathbf{u}_2 \) with \( \mathbf{A} \mathbf{v}_j = \sigma_j \mathbf{u}_j \) and \( \sigma_j \) scalars for \( j = 1, 2 \).

Action of \( \mathbf{A} \) on vector \( \mathbf{w} \)

\[
\mathbf{A} \mathbf{w} = (\mathbf{v}_1^T \mathbf{w}) \mathbf{A} \mathbf{v}_1 + (\mathbf{v}_2^T \mathbf{w}) \mathbf{A} \mathbf{v}_2 \\
= (\mathbf{v}_1^T \mathbf{w}) \sigma_1 \mathbf{u}_1 + (\mathbf{v}_2^T \mathbf{w}) \sigma_2 \mathbf{u}_2 \\
= (\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T) \mathbf{w} \\
\Rightarrow \mathbf{A} = \mathbf{v}_1^T \sigma_1 \mathbf{u}_1 + \mathbf{v}_2^T \sigma_2 \mathbf{u}_2
\]

Express that as \( \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \), with \( \mathbf{U} \) containing the columns of \( \mathbf{u}_i \), \( \mathbf{V} \) the columns of \( \mathbf{v}_i \), and \( \Sigma \) a diagonal matrix with \( \sigma_i \) along the diagonal.
The full SVD describes both the domain and range of a matrix by orthonormal bases

- For an arbitrary matrix $A \in \mathbb{R}^{m \times n}$ we have an $n \times n$ matrix $V$ and a $m \times m$ matrix $U$ that are both orthonormal, and a $m \times n$ matrix $\Sigma$ whose non-zero entries $\sigma_i = \Sigma_{ii}$ are along the diagonal:

$$A = UV^T$$

- The columns of $V$ and $U$ are the right and left singular vectors, and the diagonal entries of $\Sigma$ are the singular values of $A$. 
Geometry of SVD: choice of basis vectors lying on circle and map

♡ Choose the pre-image of the orthogonal pair in the range of the map.
Singular vectors describe spheres and ellipsoids by semi-major axes

- There is one choice of vector pairs (basis) in the domain that gets mapped into an orthogonal pair along the major axes of the ellipse.
- These pairs are the **singular vectors** of the matrix. The lengths of the semi-major axes of the ellipse are the **singular values**.
- There will be left and right singular vectors
How is the SVD made useful in machine learning?

- Distance minimisation: matrix generalisation of the following
- To find a vector $\mathbf{y}$ from a set $\mathcal{Y}$ closest to $\mathbf{x}$ we perform
  \[ \mathbf{y} = \arg \min_{\mathbf{v} \in \mathcal{Y}} \| \mathbf{x} - \mathbf{v} \|. \]

- SVD helps find a matrix $\tilde{\mathbf{X}}$ from a set $\mathcal{M}$ closest to given matrix $\mathbf{X}$:
  \[ \tilde{\mathbf{X}} = \arg \min_{\mathbf{Y} \in \mathcal{M}} \| \mathbf{X} - \mathbf{Y} \|_2. \]
**SVD gives low-rank approximation of matrices**

- We seek \( \tilde{\mathbf{X}} = \arg\min_{\mathbf{Y} \in \mathcal{M}} \| \mathbf{X} - \mathbf{Y} \|_2 \).
- By partitioning the numerically ordered diagonal entries of \( \Sigma \) into the first \( k \) and the rest, we have (from the SVD)
  \[
  \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T + \mathbf{U}_\perp \Sigma_\perp \mathbf{V}_\perp^T \\
  = (\mathbf{U}_k \mathbf{U}_\perp) \begin{pmatrix} \Sigma_k \\ \Sigma_\perp \end{pmatrix} \begin{pmatrix} \mathbf{V}_k \\ \mathbf{V}_\perp \end{pmatrix} \\
  \approx \mathbf{U}_k \Sigma_k \mathbf{V}_k^T \equiv \tilde{\mathbf{A}}_k
  \]

- \( \mathbf{A} \) is replaced by the rank \( k \) matrix \( \tilde{\mathbf{A}}_k \). Of all possible rank-\( k \) matrices \( \mathbf{B} \in \mathcal{M}_k \), \( \tilde{\mathbf{A}}_k \) constructed via the SVD gives the best approximation to \( \mathbf{A} \) in the sense that it minimises the \( L_2 \)-norm:
  \[
  \tilde{\mathbf{A}}_k = \arg\min_{\mathbf{B} \in \mathcal{M}_k} \| \mathbf{A} - \mathbf{B} \|_2.
  \]
Linear regression using SVD: find $w$ for smallest $\|Aw - y\|_2$

- A vector $w$ that is closest to target vector $y$ along direction $u$ is $w = x^* u$. Proof: all vectors in direction $u$ of the form $x u$,

$$x^* = \arg\min_{x \in \mathbb{R}} \|y - x u\|_2 = \frac{y \cdot u}{u \cdot u} = y \cdot u \text{ (projection)} u$$

- Use SVD to find singular vectors $u_i$ and find projections of $y$ along each.
Linear regression by SVD: express weights and targets in terms of singular vectors

- \( \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T = \sum_{k=1}^{r} \mathbf{u}_k \sigma_k \mathbf{v}_k^T \) or, equivalently, \( \mathbf{A} \mathbf{v}_i = \sigma_i \mathbf{u}_i \).
- Weight space spanned by \( \mathbf{v}_k \), output space spanned by \( \mathbf{u}_k \):
  \[
  \mathbf{w} = \sum_i \alpha_i \mathbf{v}_i \quad \text{and} \quad \mathbf{y} = \sum_k \beta_k \mathbf{u}_k .
  \]
- Nearest vector to \( \mathbf{y} \) along direction \( \mathbf{u}_k \) is \( (\mathbf{u}_k \cdot \mathbf{y}) \mathbf{u}_k \). Let \( \beta_k = (\mathbf{u}_k^T \mathbf{y}) \).
- The model prediction \( \mathbf{A} \mathbf{w} \) combines weighted features
  \[
  \mathbf{A} \mathbf{w} = \mathbf{A} \left( \sum_k \alpha_k \mathbf{v}_k \right) = \sum_k \alpha_k (\mathbf{A} \mathbf{v}_k) = \sum_k \alpha_k \sigma_k \mathbf{u}_k .
  \]
- The best fit vector to \( \mathbf{y} \) along each \( \mathbf{u}_k \) is \( \beta_k \mathbf{u}_k \). The vector in the column space of \( \mathbf{A} \) along direction \( \mathbf{u}_k \) is \( \alpha_k \sigma_k \mathbf{u}_k \).
- Equating coefficients along \( \mathbf{u}_i \),
  \[
  \sum_k \alpha_k \sigma_k \mathbf{u}_k = \mathbf{y} .
  \]
Linear regression by SVD: small singular values are unwelcome

- The best fit weight vector is

\[ w = \sum_i \left( \frac{u_i^T y}{\sigma_i} \right) v_i. \]

- Very small (zero) singular values cause problems. The large (infinite) components of the weight vectors track noise in the targets, not useful signals.

- This requires regularisation: add positive constant to denominator from minimising

\[ \| y - Aw \|^2 + \lambda \| w \|^2 \]
Summary: linear algebra provides conceptual tools

- Column spaces and kernels
- Bases and orthogonality
- Decomposition induced by patterns in data
- Independence of influences