

AAIL100: Introduction to Data Science

Ragesh Jaiswal, IITD

Best Fit Subspaces and Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD)

Applications of SVD

- The line **through the origin** that minimizes the sum of squared distances is defined by the first singular vector.
- Question: What if we drop the constraint that the line should pass through the origin? This general case is more relevant in data analysis.

Singular Value Decomposition (SVD)

Applications of SVD

- The line **through the origin** that minimizes the sum of squared distances is defined by the first singular vector.
- Question: What if we drop the constraint that the line should pass through the origin? This general case is more relevant in data analysis.

Theorem

The best fit line of a set of data points must pass through the centroid of the points.

Singular Value Decomposition (SVD)

Applications of SVD

Theorem

The best fit line of a set of data points must pass through the centroid of the points.

- We can generalize for higher dimensional objects.

Definition (Affine subspace)

An affine subspace is a subspace translated by a vector. So, it is a set of the form: $\{\mathbf{v}_0 + \sum_{i=1}^k c_i \mathbf{v}_i \mid c_1, \dots, c_k \in \mathbb{R}\}$. Here $\mathbf{v}_0$ is the translation and $\mathbf{v}_1, \dots, \mathbf{v}_k$ form an orthonormal basis of the subspace.

Theorem

The k -dimensional affine space that minimizes the sum of squared perpendicular distances to the data points must pass through the centroid of the points.

Singular Value Decomposition (SVD)

Applications of SVD: PCA

Theorem

The best fit line of a set of data points must pass through the centroid of the points.

Theorem

The k -dimensional affine space that minimizes the sum of squared perpendicular distances to the data points must pass through the centroid of the points.

- Given the above, the standard first step in data analysis is to **center the data** which means subtracting the centroid from all the data points (so that the origin is the centroid).

Singular Value Decomposition (SVD)

Applications of SVD: PCA

- Given the above, the standard first step in data analysis is to **center the data** which means subtracting the centroid from all the data points (so that the origin is the centroid).
- Usually, a next natural data analysis investigation on the centered data is to find the **principle components** of the data. That is, the directions along which the data has the maximum variation.
- Our discussion of Singular values and Singular vectors are already in tune with this idea. Note that the Singular vectors give the direction along which the data has the maximum variance.
- If we consider the projection Y of the data points X onto V_k (i.e., the subspace defined by the first k Singular vectors), then Y may have much smaller dimension than X and yet has most of the “content” of the original data X .
- This is a standard **dimension reduction** technique and is popularly known as **Principal Component Analysis (PCA)**.

Singular Value Decomposition (SVD)

Applications of SVD: Linear Regression

- Linear regression: Given $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^d$ and $y_1, \dots, y_m \in \mathbb{R}$ we want to find a vector $\mathbf{w}$ that minimizes:

$$\sum_{i=1}^n (\mathbf{x}_i^T \mathbf{w} - y_i)^2 = \|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2.$$

Where $\mathbf{A}$ is a matrix with the i^{th} row as $\mathbf{x}_i^T$.

- The following theorem gives a very nice application of SVD.

Theorem

The vector $\mathbf{w}$ that minimizes $\|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2$ is $\mathbf{w} = \mathbf{A}^\dagger \mathbf{y} = \mathbf{V}\mathbf{D}^\dagger \mathbf{U}^T \mathbf{y}$ for $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and $D_{ii}^\dagger = 1/D_{ii}$ for all i .

End