Principal Component Analysis and Least Squares

Distance from a Point to a Line in \( \mathbb{R}^n \)

Let \( Z \in \mathbb{R}^n \) be a point and \( \mathcal{L} \) the straight line

\[
\mathcal{L} = \{X \in \mathbb{R}^n \mid X = X_0 + tV\},
\]

where \( X_0 \in \mathbb{R}^n \) is a specified point, \( V \in \mathbb{R}^n \) is a unit vector, and \( t \in \mathbb{R} \) (think of \( X(t) \) as the position of a particle at time \( t \)).

Compute the (Euclidean) distance from \( Z \) to the line.

**Solution:** We minimize

\[
\varphi(t) = \|Z - (X_0 + tV)\|^2.
\]

At a minimum,

\[
0 = \varphi'(t) = 2\langle Z - (X_0 + tV), -V \rangle.
\]

Because \( \|V\| = 1 \) then

\[
t = \langle Z - X_0, V \rangle
\]

and hence

\[
\text{Distance}^2(Z, \mathcal{L}) = \|Z - X_0 - \langle Z - X_0, V \rangle V\|^2 = \|Z - X_0\|^2 - \langle Z - X_0, V \rangle^2.
\]

Because \( \langle Z - X_0, V \rangle V \) is just the orthogonal projection of \( Z - X_0 \) into \( \mathcal{L} \), this formula is also a geometrically obvious consequence of the Pythagorean Theorem.

Fit Data to a Straight Line

Let \( Z_1, \ldots, Z_N \) be \( N \) data points in \( \mathbb{R}^n \) and \( \mathcal{L} \) be the straight line

\[
\mathcal{L} = \{X \in \mathbb{R}^n \mid X = X_0 + tV\},
\]

where \( X_0 \in \mathbb{R}^n \) is a specified point, \( V \in \mathbb{R}^n \) is a unit vector, and \( t \in \mathbb{R} \) (think of \( X(t) \) as the position of a particle at time \( t \)).

Find the line \( \mathcal{L} \) that best fits the data in the sense that it minimizes the error (see the previous problem)

\[
E(X_0, V) = \sum_{j=1}^{N} \text{Distance}^2(Z_j, \mathcal{L}) = \sum_{j=1}^{N} \left( \|Z_j - X_0\|^2 - \langle Z_j - X_0, V \rangle^2 \right).
\]

**Solution:** With hindsight it is helpful to introduce the mean of the data, \( \bar{Z} := \frac{1}{N} \sum_{j=1}^{N} Z_j \).

Let \( W_j := Z_j - \bar{Z} \) and \( \bar{X}_0 := X_0 - \bar{Z} \). Then \( Z_j - X_0 = W_j - \bar{X}_0 \). Note that \( \sum_j W_j = 0 \).

We want to minimize

\[
E(X_0, V) = \sum_{j=1}^{N} \left( \|W_j - \bar{X}_0\|^2 - \langle W_j - \bar{X}_0, V \rangle^2 \right).
\]
Since $\sum W_j = 0$

$$\sum_{j=1}^{N} \left( \|W_j - \hat{X}_0\|^2 \right) = \sum_j \|W_j\|^2 - 2 \sum_j \langle W_j, \hat{X}_0 \rangle + \sum_j \|\hat{X}_0\|^2$$

$$= \sum_j \|W_j\|^2 + N\|\hat{X}_0\|^2$$

and

$$\sum_j (W_j - \hat{X}_0, V)^2 = \sum_j \left( \langle W_j, V \rangle - \langle \hat{X}_0, V \rangle \right)^2$$

$$= \sum_j \langle W_j, V \rangle^2 - 2 \sum_j \langle W_j, V \rangle \langle \hat{X}_0, V \rangle + \sum_j \langle \hat{X}_0, V \rangle^2$$

$$= \sum_j \langle W_j, V \rangle^2 + N\langle \hat{X}_0, V \rangle^2.$$

Therefore

$$E(X_0, V) = \sum_j \|W_j\|^2 + N(\|\hat{X}_0\|^2 - \langle \hat{X}_0, V \rangle^2) - \sum_j \langle W_j, V \rangle^2$$ (1)

Since $\|V\| = 1$, the Cauchy-Schwarz inequality shows that $|\langle \hat{X}_0, V \rangle| \leq \|\hat{X}_0\|$. Thus the middle term of $E(X_0, V)$ is minimized by choosing $\hat{X}_0 = 0$, that is $X_0 = Z$ so the best straight line $\mathcal{L}$ should contain the center of mass of the data points.

Next we find the unit vector $V$. Because after setting $\hat{X}_0 = 0$ only the last term in equation (??) depends on $V$, we should pick a unit vector that maximizes

$$Q(V) := \sum_j \langle W_j, V \rangle^2.$$

It is time to look more closely at the data matrix $W$ for this problem.

$$W := \begin{pmatrix}
\cdots & W_1 & \cdots \\
\cdots & W_2 & \cdots \\
\cdots & W_3 & \cdots \\
\vdots & \vdots & \vdots \\
\cdots & W_N & \cdots 
\end{pmatrix},$$

where $W_j$ is a row vector with the data for the $j^{th}$ data point. Then, with $V$ as a column vector, the product vector $WV$ is the column vector

$$WV = \begin{pmatrix}
\langle W_1, V \rangle \\
\langle W_2, V \rangle \\
\vdots \\
\langle W_N, V \rangle
\end{pmatrix}.$$
Thus

\[ Q(V) = \|WV\|^2 = \langle WV, WV \rangle = \langle V, W^*WV \rangle \]

The matrix \( W^*W \) is a positive semidefinite symmetric matrix, so its eigenvalues, \( \sigma_j \) (which are referred to as the singular values of \( W \)) are either positive or zero

\[ \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0. \]

Call the corresponding (orthonormal) eigenvectors \( e_1, e_2, \ldots, e_n \). Then \( W^*e_j = \sigma_j e_j \)

\( j = 1, \ldots, n \). These eigenvectors are a basis for \( \mathbb{R}^n \) so we can write \( V \) in this basis:

\[ V = v_1 e_1 + v_2 e_2 + \cdots + v_n e_n. \]

This gives

\[ W^*WV = \sigma_1 v_1 e_1 + \sigma_2 v_2 e_2 + \cdots + \sigma_n v_n e_n \]

and

\[ Q(V) = \langle V, W^*WV \rangle = \sigma_1 v_1^2 + \sigma_2 v_2^2 + \cdots + \sigma_n v_n^2. \]

Consequently \textit{the unit vector} \( V \text{ that maximizes } Q(V) \text{ is the eigenvector } e_1 \text{ of } W^*W \text{ corresponding to largest eigenvalue of } W^*W \).

\textbf{Example.} Say the data are points \((a_1, b_1), \ldots, (a_N, b_N)\) in the plane \( \mathbb{R}^2 \) so the data matrix is

\[
Z = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_N & b_N \end{pmatrix}.
\]

We assume the data has been normalized of the sum of each column is zero. Then

\[
Z^*Z = \begin{pmatrix} a_1 & a_2 & \cdots & a_N \\ b_1 & b_2 & \cdots & b_N \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_N & b_N \end{pmatrix} = \begin{pmatrix} \sum_j a_j^2 & \sum_j a_j b_j \\ \sum_j a_j b_j & \sum_j b_j^2 \end{pmatrix}
\]

which is easy to use.