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

$$\mathcal{L} = \{X \in \mathbb{R}^n \,|\, X = X_0 + tV\},$$

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). Since adding a multiple of V to X_0 does not change the line, we may assume that X_0 is orthogonal to V, so $\langle X_0, V\rangle = 0$. Then for fixed V and various X_0 define parallel lines whose distance from the origin is $\|X_0\|$

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 $t = \langle Z - X_0, V\rangle = \langle Z, V\rangle$ and hence

$$\begin{aligned}\text{Distance}^2(Z, \mathcal{L}) =& \|Z - X_0 - \langle Z, V\rangle V\|^2 \\ =& \|Z - X_0\|^2 - \langle Z, V\rangle^2.\end{aligned}$$

Because $\langle Z, V\rangle V$ is just the orthogonal projection of Z 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, as above, $\mathcal{L}$ be the straight line

$$\mathcal{L} = \{X \in \mathbb{R}^n \,|\, X = X_0 + tV\},$$

where $V \in \mathbb{R}^n$ is a unit vector and $X_0 \perp V$ is a specified point on the line.
Find the line $\mathcal{L}$ that best fits the data in the sense that it minimizes the error (see the previous problem)

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

SOLUTION: We use the mean of the data $\bar{Z} := \frac{1}{N}\sum_1^N Z_j$ to simplify (1).

$$\begin{aligned}\sum_{j=1}^{N}\|Z_j - X_0\|^2 &= \sum_{j=1}^{N}\left(\|Z_j\|^2 - 2\langle Z_j,\, X_0\rangle + \|X_0\|^2\right)\\ &= \sum_{j=1}^{N}\|Z_j\|^2 - 2N\langle \bar{Z},\, X_0\rangle + N\|X_0\|^2\\ &= \sum_{j=1}^{N}\|Z_j\|^2 + N\left[\|X_0 - \bar{Z}\|^2 - \|\bar{Z}\|^2\right]\end{aligned}$$

Thus

$$E(X_0, V) = \sum_{j=1}^{N}\|Z_j\|^2 + N\left[\|X_0 - \bar{Z}\|^2 - \|\bar{Z}\|^2\right] - \sum\langle Z_j,\, V\rangle^2$$

This is clearly minimized by letting $X_0 = \bar{Z}$, so *the best straight line $\mathcal{L}$ should contain the center of mass of the data points* and choosing the unit nit vector V *maximizes*

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

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

$$Z := \begin{pmatrix} \cdots & Z_1 & \cdots \\ \cdots & Z_2 & \cdots \\ \cdots & Z_3 & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & Z_N & \cdots \end{pmatrix},$$

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

$$ZV = \begin{pmatrix} \langle Z_1,\, V\rangle \\ \langle Z_2,\, V\rangle \\ \vdots \\ \langle Z_N,\, V\rangle \end{pmatrix}.$$

Thus

$$Q(V) = \|ZV\|^2 = \langle ZV,\, ZV\rangle = \langle V,\, Z^*ZV\rangle$$

The matrix Z^*Z is a positive semidefinite symmetric matrix, so its eigenvalues, σ_j (which are refered to as the *singular values of Z* 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 $Z^*Ze_j = \sigma_j e_j$, $j = 1, \ldots, n$. These eigenvectors are a basis for $\mathbb{R}^n$ so we can write V in this basis:

$$V = v_1e_1 + v_2e_2 + \cdots + v_ne_n.$$

This gives

$$Z^*ZV = \sigma_1v_1e_1 + \sigma_2v_2e_2 + \cdots + \sigma_nv_ne_n$$

and

$$Q(V) = \langle V, Z^*ZV\rangle = \sigma_1v_1^2 + \sigma_2v_2^2 + \cdots + \sigma_nv_n^2.$$

Consequently *the unit vector V that maximizes $Q(V)$ is the eigenvector e_1 of Z^*Z corresponding to largest eigenvalue* of Z^*Z.

In statistical applications Z^*Z is called the *covariance matrix of Z*.

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

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

which is easy to use.