Multivariate Statistics – PCA (population view)

Let $X \in \mathbb{R}^p$ be a random vector with mean 0 (without limitation of generality) and covariance

$$\mathrm{Cov}(X) = E(XX^t) = \Sigma = V\Lambda V^t$$

for an orthogonal V (using the EVD of Σ).

> Goal of Principal Component Analysis (PCA): find a linear projection P of rank $q < p$ that retains "most of the signal".

We have to make the statement "most of the signal" more concrete and will use a Euclidean metric.

We can always decompose X for any projection P as

$$X = \underbrace{PX}_{\text{projected vector}} + \underbrace{(X - PX)}_{\text{error}}.$$

Then

$$\begin{aligned} E\big(\|X\|_2^2\big) = E\big(\sum_{k=1}^{p} X_k^2\big) &= E\big(\|PX + (X - PX)\|_2^2\big) \\ &= E\big(\|PX\|_2^2\big) + E\big(\|X - PX\|_2^2\big) + 2E\big(X^t P^t (X - PX)\big) \end{aligned}$$

The last term can be written as

$$X^t P^t (X - PX) = X^t (PX - P^2 X) = X^t (PX - PX) \equiv 0,$$

since $P^t = P$ and $P^2 = P$. Hence

$$E\big(\|X\|_2^2\big) = E\big(\|PX\|_2^2\big) + E\big(\|X - PX\|_2^2\big)$$

Maximising variance of projected data $E\big(\|PX\|_2^2\big)$ is thus identical to minimizing error variance $E\big(\|X - PX\|_2^2\big)$.

We now want to maximise the variance of the projected data $E\big(\|PX\|_2^2\big)$. Let us fix a $q < p$. Let $\mathcal{P}(q)$ be the space of all linear projections with rank q. Then our problem writes as

$$\mathrm{argmax}_{P \in \mathcal{P}(q)} E\big(\|PX\|_2^2\big) = \textbf{?}$$

Which projection is achieving the optimum?

Let $w_1, \dots, w_q \in \mathbb{R}^p$ be a basis for the space preserved by the projection P. Let the matrix W be defined by

$$W = \{w_k\}_{k \in \{1,\dots,q\}}$$

by setting the k-th column equal to w_k. Then the projection P can be written as

$$P = W(W^tW)^{-1}W^t$$

Without loss of generality we can write

$$W = VB,$$

where

(i) V is the matrix from the EVD of Σ and contains the k-th eigenvector of Σ in column k for $k = 1, \dots, p$.

(ii) $B \in \mathbb{R}^{p \times q}$ is a matrix of coefficients for $w_1, \dots, w_q$ in basis given by $v_1, \dots, v_p$

Then

$$\begin{aligned} E\big(\|PX\|_2^2\big) &= E\big(\sum_{k=1}^{p}(PX)_k^2\big) \\ &= E\big(\mathrm{trace}(PXX^tP^t)\big) \\ &= \mathrm{trace}\big(P\underbrace{E(XX^t)}_{\Sigma}P^t\big) = \mathrm{trace}\big(P\Sigma P^t\big), \end{aligned}$$

where $\Sigma = E(XX^t) = V\Lambda V^t$ is EVD of covariance of X. Hence, using $P^t = P$ and $P^2 = P$,

$$\begin{aligned} E\big(\|PX\|_2^2\big) &= \mathrm{trace}\big(P\Sigma P^t\big) \\ &= \mathrm{trace}\big(P^tP\Sigma\big) \\ &= \mathrm{trace}\big(P\Sigma\big) \end{aligned}$$

Using $P = W(W^tW)^{-1}W^t$ and using $W = VB$,

$$P = VB(B^t\underbrace{V^tV}_{\mathbf{1}_{p\times p}}B)^{-1}B^tV^t = VB(B^tB)^{-1}B^tV^t,$$

we get

$$\begin{aligned} E\big(\|PX\|_2^2\big) &= \mathrm{trace}\big(P\Sigma\big) \\ &= \mathrm{trace}\big(VB(B^tB)^{-1}B^t\underbrace{V^tV}_{\mathbf{1}_{p\times p}}\Lambda V^t\big) \\ &= \mathrm{trace}\big(\underbrace{V^tV}_{\mathbf{1}_{p\times p}}B(B^tB)^{-1}B^t\Lambda\big) \\ &= \mathrm{trace}\big(P_B\Lambda\big) = \sum_{k=1}^{p}\Lambda_{kk}(P_B)_{kk} = (*), \end{aligned}$$

where $P_B = B(B^tB)^{-1}B^t$. You can think of P_B as the projection if coefficients are expressed in terms of the eigenvector basis of Σ. Note that

(i) $(P_B)_{kk} \leq 1$ for all $k = 1, \ldots, p$ as $\|P_B x\|_2 \leq \|x\|_2$ for all $x \in \mathbb{R}^p$. Taking x as first unit vector (a zero vector with exception of entry 1 at first component) implies that $(P_B)_{11} \leq 1$ and analogously for $k > 1$.

(ii) $\text{trace}(P_B) = \sum_{k=1}^{p}(P_B)_{kk} = \text{trace}(B(B^tB)^{-1}B^t) = \text{trace}(\underbrace{B^tB(B^tB)^{-1}}_{\mathbf{1}_{q\times q}}) = q.$

Under these two constraints the maximal value that can be achieved in (*) is $\sum_{k=1}^{q}\Lambda_{kk}$, the sum of the first q largest eigenvalues. This optimum is achieved by choosing for example a zero matrix with exception of entries 1 along the diagonal:

$$B^* = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ & 0 & \end{pmatrix}.$$

Then $P_{B^*} \in \mathbb{R}^{p\times p}$ is diagonal with the first q entries being 1 and the remaining ones 0:

$$P_{B^*} = B^*((B^*)^tB^*)^{-1}(B^*)^t = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

and this matrix P_{B^*} "keeps" the first q eigenvalues of Σ in $(*)$ and corresponds to projecting into the space spanned by the first q eigenvectors of Σ. A solution that minimizes the approximation error under a linear projection of rank q is hence a projection into the space spanned by the first q eigenvectors of Σ. The solution B that achieves the optimum is not unique, though, as B could also be chosen as a rotation of B^* where the rotation operates only on the first q coordinates. The solution is hence not unique for a fixed value of q. The basis in original space (not eigenvector space) is given by

$$W = VB^* = \{v_k\}_{k\in\{1,\ldots,q\}}$$

the first q eigenvectors of Σ (or any other basis spanning the same space). The projection P_B will, however, be unique for all these choices of B, as long as the eigenvalues of Σ are all distinct.