1 Definition of covariance matrix

Suppose **X** is a d-dimensional random vector (with d random variables), and $\boldsymbol{X_1}$,...,$\boldsymbol{X_n}$ is n independent copies of **X**.

Write $\boldsymbol{X_i} = (X_i^1, \dots, X_i^d)^T$, the subscript means the i_{th} copy, the superscript means the number of random variable (i.e. scala).

$$\boldsymbol{X} = \begin{pmatrix} X^1 \\ X^2 \\ \dots \\ X^d \end{pmatrix} \tag{1}$$

Then we can know the covariance matrix, which means take two different scalas or coordinates (notice the superscript) from a vector and compute their covariance. For convenience, not use bold X again as before.

$$\begin{aligned} \Sigma &= cov(X^i, X^j) && (2) \\ &= \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X)^T && (3) \\ &= \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T] && (4) \end{aligned}$$

When it comes to empirical data, we use average $\bar{X}$ to replace expectation[1] and use the empirical covariance matrix **S** to replace the Σ),

$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X^1) \\ \vdots \\ \mathbb{E}(X^d) \end{pmatrix} \rightarrow \begin{pmatrix} \frac{\sum}{n} X_i^1 \\ \vdots \\ \frac{\sum}{n} X_i^d \end{pmatrix} \tag{5}$$

$$S = \frac{1}{n} \sum (X_i X_i^T) - \bar{X}\bar{X}^T \tag{6}$$

In order to eliminate the sum character, we multiply a $\mathbb{1}$ to replace the average. $\mathbb{1} = (1, \dots, 1)^T$

$$\bar{X} = \frac{1}{n} \sum X_i \qquad \mathbb{X} = \begin{bmatrix} \vdots & \vdots & \vdots \\ X_1 & X_2 & X_n \\ \vdots & \vdots & \vdots \end{bmatrix} \tag{7}$$

$$\frac{1}{n} \mathbb{X}^T \mathbb{1} = \frac{1}{n} \sum X_i = \bar{X} \tag{8}$$

[1] Here can be a little comfused because in we used subscript before but here we have X_i. This is because in theory, $E(X^1)$ is the expectation of random variable X^1, but empirically we sampled many times and calculate their average

And we can see that

$$M_i = \begin{bmatrix} 0 & \vdots & 0 & 0 \\ 0 & X_i & 0 & 0 \\ 0 & \vdots & 0 & 0 \end{bmatrix} \tag{9}$$

$$\mathbb{X}^T\mathbb{X} = \sum_i^n M_i M_i^T = \sum_i^n X_i X_i^T \tag{10}$$

$$\mathbb{X}^T = M_1 + M_2 + \cdots + M_n \tag{11}$$

Then in Eq.6 can be transformed into

$$\begin{aligned} S =& \frac{1}{n}\mathbb{X}^T\mathbb{X} - \frac{1}{n^2}\mathbb{X}^T(\mathbb{1}\mathbb{1}^T)\mathbb{X} && (12) \\ =& \frac{1}{n}\mathbb{X}^{\mathbb{T}}(I_d - \frac{1}{n}\mathbb{1}\mathbb{1}^T)\mathbb{X} && (13) \\ =& \frac{1}{n}\mathbb{X}^T H \mathbb{X} && (14) \end{aligned}$$

So, obviously matrix H is a prthogonal projector (you can proof by calculate $H^T H$), what's the subspace this projector project a vector to?

$$\begin{aligned} H &= (I_d - \frac{1}{n}\mathbb{1}\mathbb{1}^T) && (15) \\ &= \begin{bmatrix} 1 - \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & 1 - \frac{1}{n} \end{bmatrix} && (16) \end{aligned}$$

so for any vector $\boldsymbol{v}$, we have

$$\begin{aligned} H\boldsymbol{v} &= \boldsymbol{v} - \frac{1}{n}(\boldsymbol{v}^T\mathbb{1})\mathbb{1} && (17) \\ &= \boldsymbol{v} - \bar{\boldsymbol{v}}\mathbb{1} && (18) \end{aligned}$$

which means a vector minus its means by all elements. And it's clear that

$$avg(H\boldsymbol{v}) = 0 \tag{19}$$

means H projects vector $\boldsymbol{v}$ to the subspace that has the mean of 0. Or in another words, Hv $\perp$ span of $\mathbb{1}$ because $(Hv)^T\mathbb{1} = 0$.

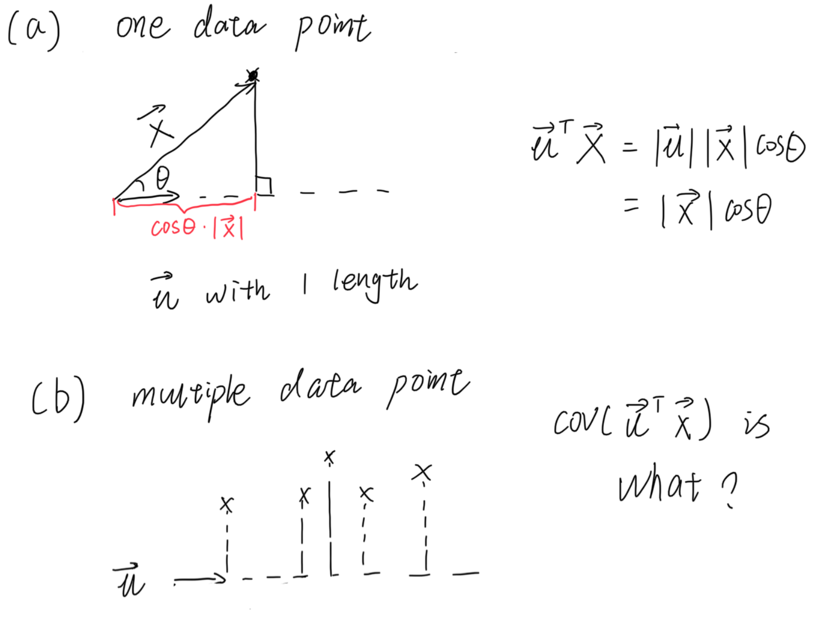
2 Core: $u^T \Sigma u$

Take a vector $\boldsymbol{u} \in \mathbb{R}^d$ (column vector), then

$$
\begin{aligned}
u^T \Sigma u &= u^T [E(XX^T) - E(X)E(X)^T]u && (20)\\
&= E[(u^T X)(X^T u)] - E(u^T X)E(X^T u) && (21)\\
&= E[(u^T X)^2] - [E(u^T X)]^2 && (22)\\
&= var(u^T X) && (23)
\end{aligned}
$$

The transition to Eq.22 is because $u^T X = X^T =$ a number. So this is the magic now, the covariance matrix is equal to the variance of $u^T X$. What's is $u^T X$?

$u^T X$ is the the inner product betwen u and X. Look at my handnote,in geometric, it means the length of red line. So with multiple points, the variance means *the degree of dispersion along the vector u.*



Therefore, we need to find the vector $\boldsymbol{u}$ to maxmize our variance, because we reduce the dimension but don't want to lose too much information (image a 3D olive, we cut it and wanna get the cross section with as long and wide as possible).

3 Spectral decomposition/Eigendecomposition

3.1 Variance is eigenvalue

Since Σ and S are symmetric, we can decompose it into this form:

$$\Sigma = PDP^T \ (or PDP^{-1}) \tag{24}$$

We know that matrix P consists of all eigenvectors of Σ, and

$$\Sigma v_1 = PDP^T v_1 = \lambda_1 v_1 \tag{25}$$

$$v_1^T \Sigma v_1 = \lambda_1 v_1^T v_1 = \lambda_1 \tag{26}$$

Therefore, the variance along eigenvectors(here v_1 means the first and largest eigenvector) is simply the eigevalue λ.

Assume $\bar{X} = 0$ to ensure $\bar{X}\bar{X}^T = 0$ and make calculation easier, the Equation 6 becomes

$$S = \Sigma X_i X_i^T \tag{27}$$

3.2 Another way to proof

Suppose $y_i = P^T X_i$ (which is the projected vector). Then

$$\bar{y_i} = \overline{P^T X_i} = P^T \bar{X}_i = 0 \tag{28}$$

$$S' = \frac{1}{n} \sum y_i y_i^T \tag{29}$$

$$= \frac{1}{n} \sum (P^T X_i)(P^T X_i)^T \tag{30}$$

$$= \frac{1}{n} \sum (P^T X_i X_i^T P) \tag{31}$$

$$= \frac{1}{n} \sum (P^T S P) \tag{32}$$

And because $S = PDP^T$, we have

$$S' = P^T (PDP^T) P \tag{33}$$

$$= D \tag{34}$$

We know D is a diagonal matrix made up of eigevalue λ_i. So $cov(y^i, y^j) = 0$ when $i \neq j$. In other words, $lambda_i = var(P^T X_i)$.

3.3 Why eigenvector is best?

Here we need to proof why eigenvectors are the ones make variance largest, because there're so many choices.

Suppose $b = P^T u$ and u is unit vector.

$$u^T S u = b^T D b = \sum_{j=1}^{d} \lambda_j b_j^2 \leq \sum_{j=1}^{d} \lambda_1 b_j^2 \tag{35}$$

λ_1 here still means the largest eigenvalue. So for any vector u, we can know that λ_1 is the largest variance and the Nth largest eigenvectors are called (N_{th}) Principal Components.

In extrme cases, if $n >> d$(much more data samples than dimension), then the empirical data converge to a consistent estimator (which means perfect). Otherwise, if $d >> n$, the angle between eigenvectors of Σ and S will be very large (which means very bad estimator). And we need sparse PCA (I don't konw this either).