Video 1: covariance matrix

wbg231

December 2022

1 introduction

- video link
- today we are going to talk about the covariance matrix which captures the variance of multi dimensional features.

motivation

- our goal is to describe data with multiple features
- we are going to be working with a random vector $\tilde{x} \in mathbbR^d$ such that

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ .. \\ \tilde{x}_n \end{pmatrix}$$

that is each element of the random vector $\tilde{x}$ is it's self a random vector

mean of random vector

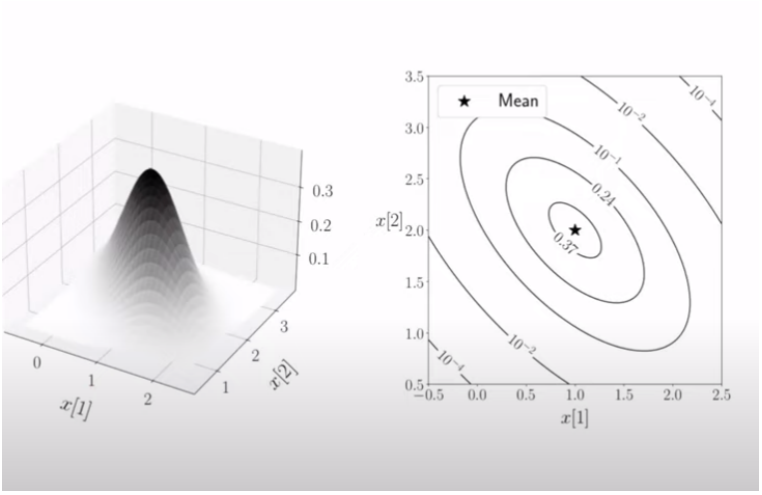
- **the mean of a random vector** $\tilde{x} \in mathbbR^d$ is defined as the mean of each of the random vectors dimensions that is

$$E[\tilde{x}] = \begin{pmatrix} E[\tilde{x}_1] \\ .. \\ E[\tilde{x}_n] \end{pmatrix}$$

gaussian random vector

- a random vector gaussian random vector $x \in \mathbb{R}^d$ has joint pdf

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

- where $\mu \in \mathbb{R}^d$ is the mean parameter and $\Sigma \in \mathbb{R}^{d \times d}$ is the covariance matrix
- note that what we saw in 1-dimension holds and $E[\tilde{x}] = \mu$ as we would expect
- 

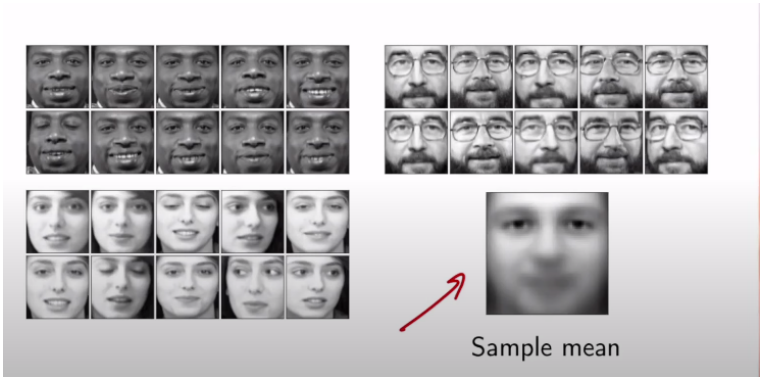
- here are visualizations of the pdf of a gaussian random vector

sample mean

- suppose where have a dataset $\mathcal{D} = \{x_1 ... x_n\} : x_i \in \mathbb{R}^d$
- the sample is given as $m(x) := \frac{1}{n}\Sigma_{i=1}^{n} x_i$ that is we just take the sample mean of each dimension

faces example

- suppose we have a dataset where each observation $x_i \in \mathbb{R}^{4096}$ that is a 64 by 64 image represented as a vector

- 

- here the mean is actually meaningfully this is primally because the images are aligned

mean of a random matrix

- if we have a random matrix $\tilde{M} = \begin{pmatrix} \tilde{x}_1 & ... & \tilde{x}_{d_2} \end{pmatrix} \in \mathbb{R}^{d_1 \times d_2} : \tilde{x}_i \in \mathbb{R}^{d_1}$ that is a matrix where all columns (or rows) are random vectors and thus all entries are random variables
- thus we can see that $E[\tilde{M}] = \begin{pmatrix} E[\tilde{x}_1] & ... & E[\tilde{x}_{d_2}] \end{pmatrix} = \begin{pmatrix} E[\tilde{x}_{1,1}] & ... & E[\tilde{x}]_{1,d_2} \\ ... & ... & ... \\ E[\tilde{x}_{d_1,1}] & ... & E[\tilde{x}]_{d_1,d_2} \end{pmatrix}$

liniarty of expecations

- for any random vector $\tilde{x} \in \mathbb{R}^d$ deterministic matrix A $\in \mathbb{R}^{n \times d}$ and vector b$\in \mathbb{R}^n$
- what is $E[A\tilde{x} + b]$?
- lets look at one entry $E[A\tilde{x} + b][i]$ we know that $(A\tilde{x} + b) \in \mathbb{R}^n$ thus by definition of random vector expectations $E[A\tilde{x} + b][i] = E[\quad [A\tilde{x} + b][i] \quad]$
- then we know that $A\tilde{x}[i]$ is a matrix times a vector and thus can be written as $A\tilde{x}[i] = \Sigma_{j=1}^{d} A[i,j]\tilde{x}[j]$ and that will be added onto by the ith entry of b
- thus we have $E[\quad [A\tilde{x}+b][i] \quad] = E[\Sigma_{j=1}^{d} A[i,j]\tilde{x}[j]+b[i]]$ then we can see that this is an affine combination of d random varible and a constant thus by linearity of expecations $= \Sigma_{j=1}^{d} A[i,j]E[\tilde{x}_j]+b[i]$ and as A is deterministic we can again apply linearity of expectations to get $= (AE[\tilde{x}] + b)[i]$
- this all more or less goes to show that linearity of expectations holds for random vectors

- the same thing also holds for random matrix $\tilde{M} \in \mathbb{R}^{n \times d}$ and deterministic matrices $A \in \mathbb{R}^{x \times n}, A \in \mathbb{R}^{x \times d}$ we have

$$E[A\tilde{X} + b] = AE[\tilde{X}] + B$$

- this will help us with the covariance matrix

2 variance

- the variance characterizes the average variation of a random variable
- for a random vector we can think of the variance as comprised of the linear combination of all entries of the random vector

variance of a linear combination

- for any deterministic vector a
- $var(< a, \tilde{x} >) = var(a^t\tilde{x}) = E[(a^t\tilde{x} - E[a^t\tilde{x}])] =$
  $= E[(a^t\tilde{x} - a^tE[\tilde{x}])^2] = E[a^t(\tilde{x} - E[\tilde{x}])^2]$
- here we are going to take a second and define $c_t(\tilde{x}) = \tilde{x} - E[\tilde{x}]$ that is the centered version of the random vector $\tilde{x}$
- $var(< a, \tilde{x} >) = var(a^t\tilde{x}) = E[(a^t\tilde{x} - E[a^t\tilde{x}])] =$
  $= E[(a^t\tilde{x} - a^tE[\tilde{x}])^2] = E[a^t(\tilde{x} - E[\tilde{x}])^2] = E[(a^tc_t(\tilde{x}))^2] = E[(a^tc_t(\tilde{x}))(a^tc_t(\tilde{x}))^t] =$
  $E[(a^tc_t(\tilde{x}))(c_t(\tilde{x})^ta] = a^tE[ct(\tilde{x})(ct(\tilde{x})^t)]a$
- this is the covariance matrix as we will show
- lets take a look at the entries of this matrix $E[ct(\tilde{x})(ct(\tilde{x})^t)[i,j]]$
- first consider diagonal entries $E[ct(\tilde{x})(ct(\tilde{x})^t)[i,i]] = E[ct(\tilde{x}_{[}i])^2] = E[(\tilde{x}[i] -$
  $E[\tilde{x}[i]])^2] = var(\tilde{x}[i])$
- consider the off diagonal now $E[ct(\tilde{x})(ct(\tilde{x})^t)[i,j]] = E[ct(\tilde{x}[i])ct(\tilde{x})[j]] =$
  $E[(\tilde{x}[i] - E[\tilde{x}[i]]) * (\tilde{x}[j] - E[\tilde{x}[j]])] = cov(\tilde{x}[i], \tilde{x}[j])$
- so with this in mind we are able to define **the covariance matrix** of a random vector $\tilde{x}$ as

$$\Sigma_{\tilde{x}} = E[ct(\tilde{x})ct(\tilde{x})^t] = \begin{pmatrix} var(\tilde{x}[1]) & cov(\tilde{x}[1], \tilde{x}[2]) & ... & ...cov(\tilde{x}[1], \tilde{x}[d]) \\ cov(\tilde{x}[1], \tilde{x}[2]) & var(\tilde{x}[2]) & ... & cov(\tilde{x}[2]\tilde{x}[d]) \\ cov(\tilde{x}[i], \tilde{x}[d]) & ... & ... & var(\tilde{x}[d]) \end{pmatrix}$$

- the covariance matrix captures the variance of any linear combinations of the random vector

- so for any deterministic vector $a \in \mathbb{R}^d$

$$var[a^t\tilde{x}] = a^t E[cv(\tilde{x})cv(\tilde{x})^t] = a^t\Sigma_{\tilde{x}}a$$

- this is nice as this captures the variance of all possible linear combinations

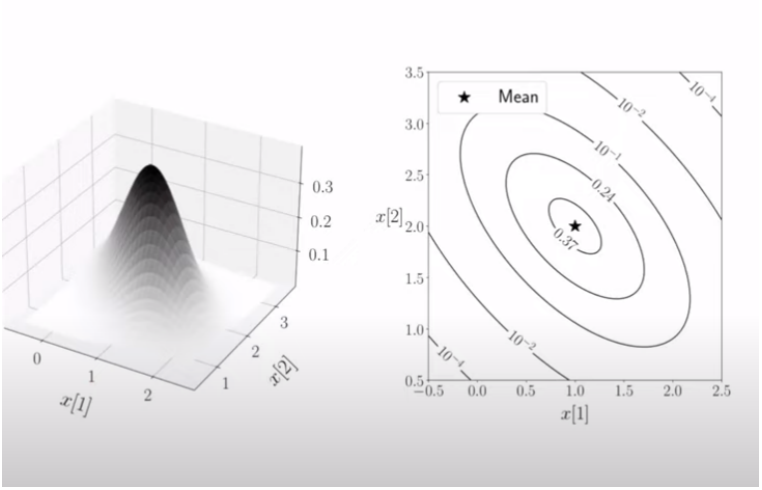
deli example

- suppose we have a deli with 3 ingredients bread local cheese and imported cheese
- suppose that the random vector $\tilde{x} = \begin{pmatrix} \text{price of bread} \\ \text{price of local cheese} \\ \text{price of imported cheese} \end{pmatrix}$
- and has covariance matrix $\Sigma_{\tilde{x}} = \begin{pmatrix} 1 & .8 & 0 \\ .8 & 1 & 0 \\ 0 & 0 & 1.2 \end{pmatrix}$
- this tells us that the price of the bread and local cheese are highly correlated and the price of the imported cheese is uncorrelated
- there are two recipes we are considering
- 1. 100 g bread, 50 g local cheese, 50 g imported cheese can be written as vector $a_1\tilde{x} = \begin{pmatrix} 100 \\ 50 \\ 50 \end{pmatrix} \tilde{x}$
- 2. 100 g bread, 100 g local cheese, 0 g imported cheese can be written as vector $a_2\tilde{x} = \begin{pmatrix} 100 \\ 100 \\ 0 \end{pmatrix} \tilde{x}$
- what is the variance in the price of each recipe?
- as said before the variance of a linear combination of our random vector can be expressed as $var(a_1\tilde{x}) = a_1^t\Sigma_{\tilde{x}}a_1$
- computing this out we see a_1 has lower variance. this makes sense as we are getting a lot of local cheese and bread which are highly correlated so those variances feed off of one another.

gaussian random vector

- a random vector gaussian random vector $x \in \mathbb{R}^d$ has joint pdf

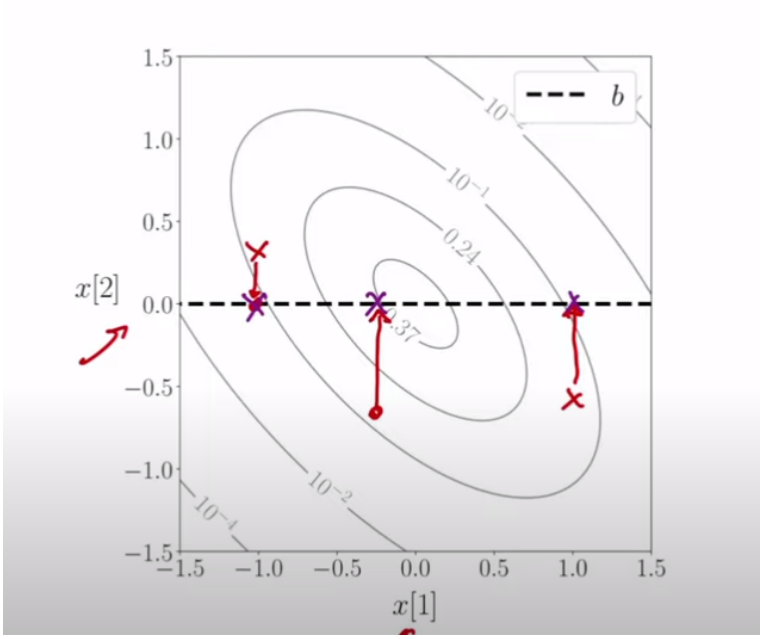
$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)}$$

- where $\mu \in \mathbb{R}^d$ is the mean parameter and $\Sigma \in \mathbb{R}^{d \times d}$ is the covariance matrix
- note that what we saw in 1-dimension holds and $E[\tilde{x}] = \mu$ as we would expect
- we also have $\Sigma_{\tilde{x}} = \Sigma$
- 

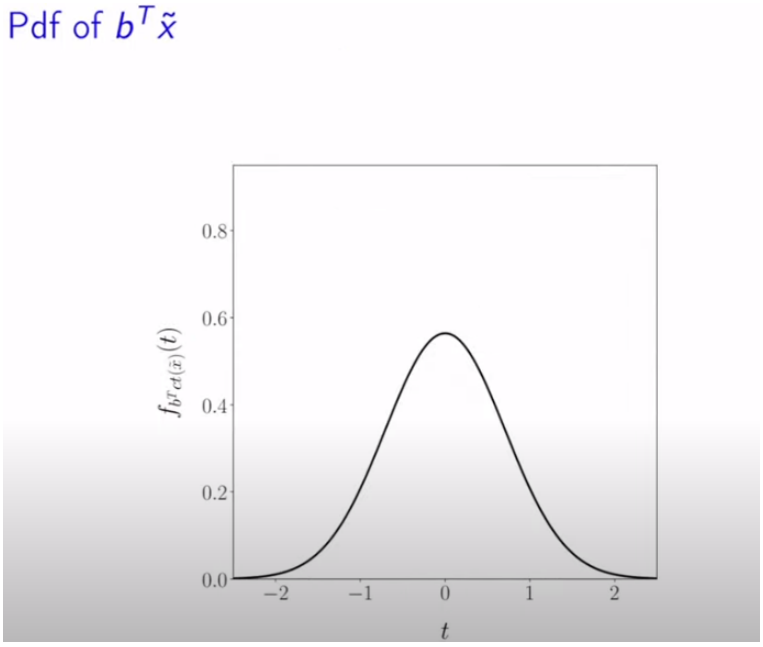
- here are visualizations of the pdf of a gaussian random vector
- basically the shape of the ellipsoids for the contour lines are determined by the covariance matrix

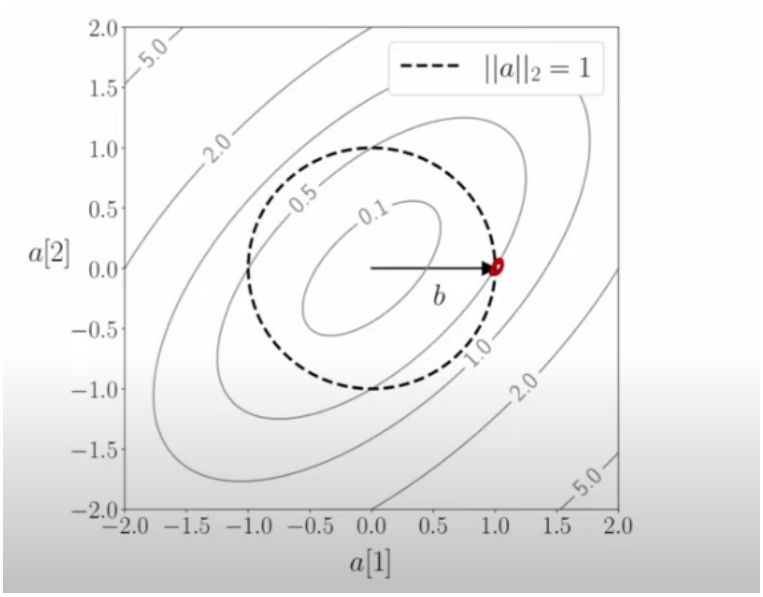
3 variance in a certain direction

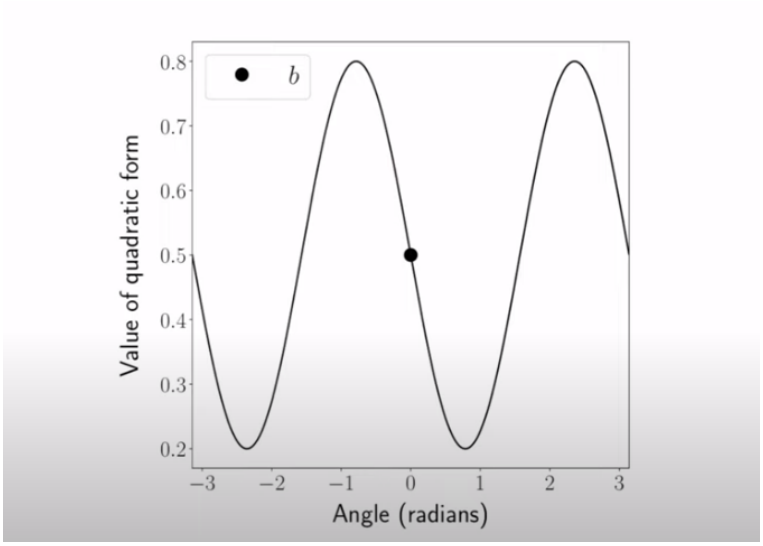
- if we want to know the variance of a random vector $\tilde{x} \in \mathbb{R}^d$ in a certain direction $b \in \mathbb{R}^d$
- we can think of this as the variance of the projection of our random variable onto that direction
- so if we assume b is unit norm than $||b|| = 1$ and $P_b(\tilde{x}) = \frac{b^t \tilde{x}}{||b||} = b^t \tilde{x}$
- we are using unit norm directions to make our math easier, and because you can just rescale any vector to be unit norm
- further note that $\tilde{x} = ((b^t \tilde{x}))b + (\tilde{x} - ((b^t \tilde{x}))b))$ to see this basically notice that $b^t \tilde{x}$ is the projection of x on to b, and thus $(b^t \tilde{x})b$ will be collinear (parallel) with b and thus if we subtract that value from our original vector x we are taking away the orthogonal portion of x
- here we also assume that we have centered x by subtracting it's mean.

- 

- here we can see a gaussian rv x, centered at 0
- and a direction b
- then we want to understand how x varies along that line by looking at how much the projection of x ($b^t\tilde{x}$) varies

- 

- here is the pdf of $b^t \tilde{x}$
- so what is the gaussian from this? we can use our equation $var[b^t \tilde{X}] = b^t \Sigma_{\tilde{x}} b = .5$
- naturally we can do this for any direction
- 

- here is our variance in every direction along the unit circle (ie all vectors with length 1 in $\mathbb{R}^2$)
- 

- we can look at this in 1 dimension, and see our variance in a certain direction as a function of the angle (note it is sinusoidal)
- this is called a quadratic form

covariance matrix on a data set

- suppose we have dataset $\mathbb{D} = \{x_1 ... x_n\} \in \mathbb{R}^{d \times n}$
- we can call the jth feature of our data set $x[j] = \{x_1[j] .. x_n[j]\} \in \mathbb{R}^n$
- the sample variance of a feature is $v(X[j])$
- and the sample covariance of two features is $c(X[j], X[k])$
- so here we can deffine the sample covariance matrix as $\tilde{x}$ as

$$\Sigma_{\tilde{x}} = \begin{pmatrix} v(x[1]) & c(x[1], \tilde{x}[2]) & ... & ...c(x[1], x[d]) \\ c(x[1], x[2]) & v(\tilde{x}[2]) & ... & c(\tilde{x}[2]\tilde{x}[d]) \\ c(\tilde{x}[i], \tilde{x}[d]) & ... & ... & v(\tilde{x}[d]) \end{pmatrix} = \frac{1}{n-1} \Sigma_{i=1}^{b} ct(x_i) ct(x_i)^t$$

where $ct(x_i) = x_i - m(x)$

sample variance in any direction

- $\mathcal{D} = \{x_1..x_n\}$
- for any vector a $\mathcal{D}_a = \{a1^t x_1 ... a^t x_n\}$
- then we can find the variance as $var(\mathcal{D}_a) = \frac{1}{n-1}\Sigma_{i=1}^{n}(a^t x_i - \frac{1}{n}\Sigma_{j=1}^{n} a^t x_j)^2$
- then by doing more or less the same computations as in the random variable case we can see that $var(\mathcal{D}_a) = a^t \Sigma_x a$
- so the sample covariance matrix also exactly captures the sample variance of any linear combination of the data
- so we can do a similar thing as we did in the random case and try to do this along all unit norm vectors to get an idea of the sample variance in all directions as a function of angle