Video 2: Principal Component Analysis

wbg231

December 2022

1 introduction

- vedio link
- today we are talking about pca

motivation

- our goal is describe data with multiple features.
- we mode our d-dimensional data set as a random vector

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ .. \\ \tilde{x}_d \end{pmatrix} \in \mathbb{R}^d$$

- our idea is to find the directions in which our random vector has the most variance

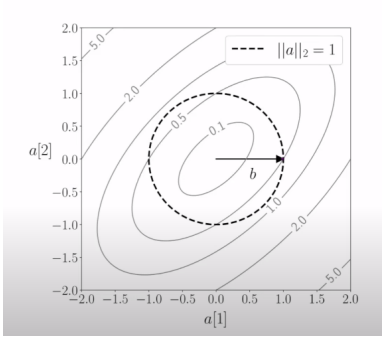
projection in a certain direction

- recall from the last Video what the how we compute the variance of a random vector $\tilde{x}$ in the direction of unit vector $b \in \mathbb{R}^d$ we do this by projecting x onto b that is $P_b(\tilde{x}) = \frac{b^t \tilde{x}}{||b||} = b^t \tilde{x}$
- in general we can think of the random vector $\tilde{x}$ as composed of a section which is colinear to the vector b and a section that is orthogonal to b that is $\tilde{x} = P_b\tilde{x}(b) + (\tilde{x} - (P_b\tilde{x})b) = (b^t\tilde{x})(b) + (\tilde{x} - (b^t\tilde{x})b)$ where b has unit norm
- we write in this way due to the fact that the variance of any linear combination of the random $\tilde{x}$ and a vector $b \in \mathbb{R}d$ can be expressed as

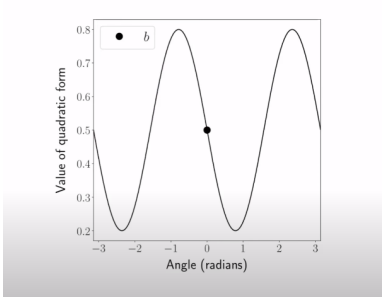
$$var(b^t\tilde{x}) = b^t\Sigma_{\tilde{x}}b$$

where $\Sigma_{\tilde{x}}$ is the covariance matrix of $\tilde{x}$

- this allows us to consider the function $f : A \Rightarrow \mathbb{R} : A := \{a \in \mathbb{R}^d : ||a|| = 1\}$ where $f(a) = var(a^t\tilde{x}) = a^t\Sigma_{\tilde{x}}a$
- plotted over the contour diagram of a certain gaussian random vector that function looks like this



- then plotting this in 2d as a function of radians looks like this



- so the direction of max variance in this case would be any point at the top that wave
- this expands to higher dimensional objects
- so now we want to know how we can find the optima in a more general sense.

the spectral theorem

- given $M \in \mathbb{R}^{d\times d}$ is symmetric then it has an eigen decomposition
- $M = \begin{pmatrix} u_1 & u_2 & ... & u_d \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ .. & ... & ... & .. \\ 0 & 0 & ... & \lambda_d \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ ... \\ u_d \end{pmatrix} = U\Lambda U^t = U\Lambda U^{-1}$
- where the eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$ (that is are sorted)
- and eigenvectors $u_1...u_d$ are orthogonal

- and u_i is the eigenvectors corresponding to λ_i
- and U is an orthogonal matrix
- we are going to go into this more next vedio
- from this we see $\lambda_1 = max_{||a||_2=1} a^t M a$ ie our largest eigenvalue is the maximal variance in any single direction
- and $u_1 = argmax_{||a||_2=1} a^t M a$ and our eigenvector corresponding to the largest eigenvalue is our direction of maximal variance
- then we can find the kth maximal direction of variance and it's associated variance orthogonal to our previous directions as

$$\lambda_k = max_{||a||=1, a \perp u_1 \dots a \perp u_k} a^t M a, \quad \forall k \in [2, d-1]$$

$$u_k = argmax_{||a||=1, a \perp u_1 \dots a \perp u_k} a^t M a, \quad \forall k \in [2, d-1]$$

showing covariance matrix is symmetric

- let $\tilde{a}$ be an arbitrary random vector
- we can define the centered version of $\tilde{a}$ as $\tilde{x} := \tilde{a} - E[\tilde{a}]$
- so we can see that $E[\tilde{x}] = E[\tilde{a} - E[\tilde{a}]] = E[\tilde{a}] - E[\tilde{a}] = 0$
- and $var(\tilde{x}) = var(\tilde{a} - E[\tilde{a}]) = E[(\tilde{a} - E[\tilde{a}] - E[\tilde{a} - E[\tilde{a}]])^2] = E[(\tilde{a} - E[\tilde{a}])^2] = var(\tilde{a}) = \Sigma_{\tilde{x}} = \Sigma_{\tilde{a}}$
- so this just is saying can take an arbitrary random vector and center it to have mean 0 and the same variance
- now we want to show $\Sigma_{\tilde{x}} = \Sigma_{\tilde{x}}^T$
- we know that $\Sigma_{\tilde{x}} = var(\tilde{x}) = E[(\tilde{x} - E[\tilde{x}])^2] = E[(\tilde{x})^2] = E[\tilde{x}\tilde{x}^t]$
- so thus $\Sigma_{\tilde{x}}^t = (E[\tilde{x}\tilde{x}^t])^t = E[(\tilde{x}\tilde{x}^t)^t] = E[(\tilde{x}\tilde{x}^t)^t = E[\tilde{x}\tilde{x}^t] = \Sigma_{\tilde{x}}$
- so yeah the covariance matrix is symmetric, and thus we can use the spectral theorem

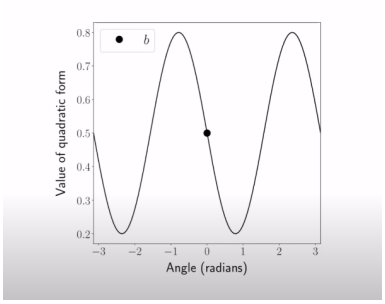
Principal directions

- let $u_1..u_d$ be the eigenvectors of $\Sigma_{\tilde{x}}$ and $\lambda_1 ... \lambda_d$ be there corresponding eigenvalues
- from this we see $\lambda_1 = max_{||a||_2=1} a^t M a = max_{||a||_2=1} var(a^t M)$ ie our largest eigenvalue is the maximal variance in any single direction

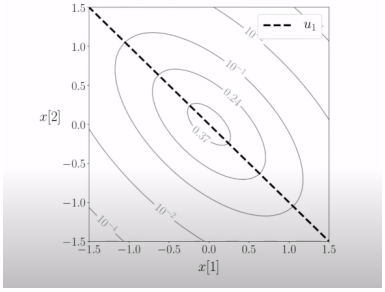
- and $u_1 = argmax_{||a||_2=1} a^t M a = argmax_{||a||_2=1} var(a^t M)$ and our eigenvector corresponding to the largest eigenvalue is our direction of maximal variance
- then we can find the kth maximal direction of variance and it's associated variance orthogonal to our previous directions as

  $\lambda_k = max_{||a||=1, a \perp u_1 \ldots a \perp u_k} a^t M a = max_{||a||=1, a \perp u_1 \ldots a \perp u_k} var(a^t M), \quad \forall k \in [2, d-1]$

  $u_k = argmax_{||a||=1, a \perp u_1 \ldots a \perp u_k} a^t M a = argmax_{||a||=1, a \perp u_1 \ldots a \perp u_k} var(a^t M), \quad \forall k \in [2, d-1]$
- we can call u_i the ith Principal direction of $\tilde{x}$
- so back to this graph



- we know that our function will be maximized at the first Principal direction (the first eigenvector of the covariance matrix of $\tilde{x}$)
- looking at the joint pdf of our random vector we can indeed see that the first Principal direction does capture the most variance graphically



- we can naturally find the minimum variance direction by just looking for the eigenvector associated with the minimum eigenvalue

Principal Components

- let $ct(\tilde{x}) = \tilde{x} - E[\tilde{x}]$ be our centered random vector

- the **Principal Component** corresponding to each **Principal direction** is defined as

$$\tilde{w_i} = u_i^t ct(\tilde{x}) = P_{u_i} ct(\tilde{x}) \quad \forall i \in [1, d]$$

that is the inner product between our centered random vector $ct(\tilde{x})$ and the Principal direction $\tilde{u_i}$ or in other words our random vector projected onto that Principal Component

variance of the ith Principal Component

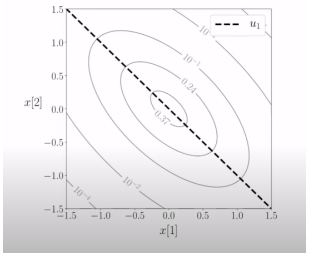
- $var(\tilde{w_i}) = var(u_i ct(\tilde{x})) = u_i^t \Sigma_{ct(\tilde{x})} u_i = u_i^t \Sigma_{\tilde{x}} u_i = \lambda_i u_i^t u_i = \lambda_i ||u_1|| = \lambda_i$ since our eigenvector are unit norm
- recall that by spectral theorem these variances correspond to the variance in the ith Principal direction that is

$$\lambda i = max_{||a||_2=1, a \perp u_1, \ldots, a \perp u_{i-1}} var(a^t \tilde{x}_i) = var(\tilde{w_i}) = var(u_i ct(\tilde{x})), \quad \forall i \in [2, d]$$

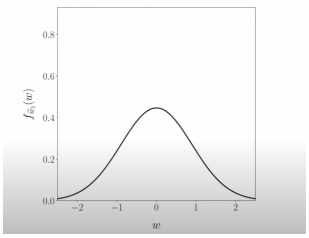
- so this goes to show that our definition of Principal Components result in Components that have maximal variance

example

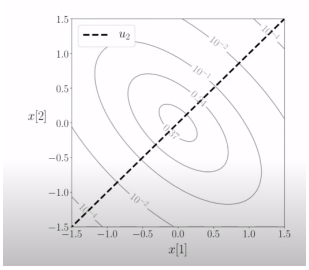
- so for a certain gaussian random vector $\tilde{x} \in \mathbb{R}^2$
- we say can plot it's first Principal direction as



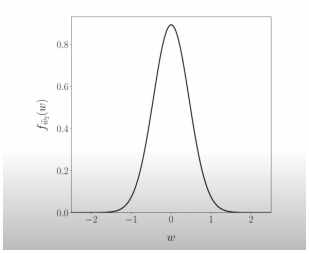
- then computing the corresponding first Principal Component as $\tilde{w_i} = u_i^t ct(\tilde{x})$ and plotting its pdf



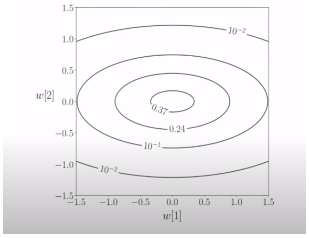
- this tells us that the first principal component of our gaussian random vector is a gaussian random varibale is a gaussian with fairly high variance
- as this is only a 2 dimensional random vector we know it's other principal direction will be the direction of minimal variance graphing this we see



- and plotting the pdf of the second principal Component we get



- which is good as we can see the second Principal Component of our random vector is a gaussian random variable with lower variance
- now we can look at the joint pdf of our two Principal Components $\tilde{w}_1, \tilde{w}_2$



- this graph makes sense as we are centering our data (then rotating it in the direction of our Principal Components) so that we have the ellipse that are most spread out in the first dimension and least spread out in the second
- this also leaves us with uncorrelated Component

corelation

$E[\tilde{w}_i\tilde{w}_j^t] = E[(u_i^t ct(\tilde{x}))(u_j^t ct(\tilde{x}))^t] = E[u_i^t ct(\tilde{x})ct(\tilde{w})u_j] = u_i^t E[ct(\tilde{x})ct(\tilde{x})]u_j = u_i^t \Sigma_{\tilde{x}} u_j = \lambda_j u_i^t u_j = 0$

- so from this we get $var(\tilde{w}_i + \tilde{w}_j) = var(\tilde{w}_i) + var(\tilde{w}_j) - 2cov(\tilde{w}_i\tilde{w}_j)$ this shows that these two Components are orthogonal
- $cov(\tilde{w}_i^t\tilde{w}_j) = E[\tilde{w}_i^t\tilde{w}_j] - E[\tilde{w}_i]^t E[\tilde{w}_j]$
- $E[\tilde{w}_j] = E[u_i^t ct(\tilde{x})] = u_i^t E[ct(\tilde{x})]$
- $cov(\tilde{w}_i^t\tilde{w}_j) = E[\tilde{w}_i^t\tilde{w}_j] - E[\tilde{w}_i]^t E[\tilde{w}_j] = 0 - (u_i^t E[ct(\tilde{x})])^t u_j^t E[ct(\tilde{x})] = E[ct(\tilde{x})]^t u_i u_j^t E[ct(\tilde{x})] = 0$
- going to show that $var(\tilde{w}_i + \tilde{w}_j) = var(\tilde{w}_i) + var(\tilde{w}_j) - 2cov(\tilde{w}_i\tilde{w}_j) = var(\tilde{w}_i) + var(\tilde{w}_j)$ and thus $\tilde{w}_i, \tilde{w}_j$ are uncorrelated

gaussian example

- the eigenvalues of a gaussian random vector's covariance matrix are the axis of it's contour lines (which are ellipses)

pca from data

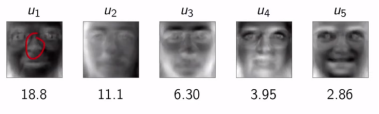
- given a dataset $\mathcal{D} = (x_1...x_n) : x_i \in \mathbb{R}^d$
- our steps of pca are
  1. compute teh sample covariance matrix Σ_d
  2. take the eigen decomposition of Σ_X which yields principal direction $u_1...u_d$
  3. center the data and compte the Principal components as
     $$w_j[i] = uj^t ct(x), \quad \forall i \in [1,d], j \in [1,d]$$
     where $ct(x_i) = x_i - M(x)$ so we are projecting the data onto the Principal Components
- this is optimal with respect to sample variance

example

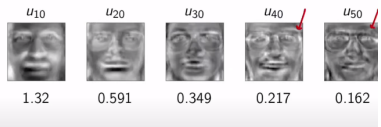
- data set with latitude and longitude of cities
- have data set $x = \{x_1...x_n\}$
- projecting our data onto direction a is given by $X_a := \{a^t x_1...a^t x_n\}$
- we know that the sample variance of X_a is given by $v(X_a) = a^t\Sigma_X a$ we showed in last Video this gives us exactly the sample variance
- we can again show that the Principal Components are the exact directions of maximal sample variance for our dataset by the exact same logic as what we did in the random vector case.

other example

- here we are looking at a dataset of face immages where each image $x_i \in \mathbb{R}^{64\times64}$ then we are going to flatten those to vector such that $x_i \in \mathbb{R}^{4096}$
- now we can build a covariance matrix $\Sigma_x \in \mathbb{R}^{4096\times4096}$ and taking the eigenvectors of that we get the principal direction which we can then project the data onto to get the Principal Component $u_1...u_{4096}$
- here are the first few Principal components



- here are some lower Principal Components few Principal components



- we can see that the higher principal Components capture some of the course features of our data like nose or glasses
- the lower Principal Components are more fine and often give us less interpretable features