Module 4

Optimal basis vectors

In the last session we found that minimizing the average squared reconstruction error is equivalent to minimizing the projection of variance of data when projected onto the subspace that we'll ignore in PCA

In this session we will exploit this insight and determine an orthonormal basis of the M dimensional principal subspace using the results from earlier (see pic) we can write (act) our loss function as:

$$J = \sum_{j=M+1}^{D} b_j^T S b_j$$

where S is the data covariance matrix

minimizing this objective, requires us to find the orthonormal basis that spans the subspace that we will ignore

and when we have that basis, we take the orthogonal compliment as the basis of the principal subspace

Remember that the orthogonal compliment of a subspace U, consist of all vectors in the original vector space, that are orthogonal to every vector in U.

Let us start with an example to determine the b vectors.

and let start in 2 Dimensions, where we wish to find a 1 Dim subspace such that the variance of data when projected onto that subspace is maximized.

So we looking at 2 basis vectors b_1 and b_2 in $\mathbb{R}^2$: b_1, b_2

b_1 will be spanning the principal subspace, b_2 is orthogonal complement, i.e the subspace we will ignore.

we also have the constraint that b_1 and b_2 are orthonormal, i.e "$b_i^T b_j$ is Delta(δ)$_{ij}$"

$$b_i^T b_j = \delta_{ij}$$

meaning that, this dot product is 1, if $i=j$, and zero (0) otherwise

"So, ∂L over $\partial \lambda$ is 1 minus b_2 transpose times b_2, and this is 0, if and only if ~~$b_2^T b_2$~~ b_2 transpose b_2 is 1"

$$\frac{\partial L}{\partial \lambda} = 1 - b_2^T b_2 = 0 \iff b_2^T b_2 = 1$$

(so we recover our constraint)

So now let's have a look at the partial derivative of L wrt b_2

We get:

A: from first term

B: from second term

needs to be zero

$$\frac{\partial L}{\partial b_2} = 2 b_2^T S - 2\lambda b_2^T = 0$$

"and zero, if and only if, S times b_2 is λ times b_2"

$$\iff S b_2 = \lambda b_2$$

Here we end up with an Eigenvalue problem

b_2 is an Eigenvector of ~~(vector)~~ data covariance matrix

and the Lagrange multiplier plays the role of the corresponding ~~Eigenvector~~ Eigenvalue

If we now go back to the loss function, we can use this expression:

we can write " of which was b_2 transpose times S times b_2 "

$$J = b_2^T S b_2$$

"But now we know that S times b_2 can be written as λ times b^2, so we get b_2 transpose times b_2 times ~~*~~ λ"

$$= \underbrace{b_2^T b_2}_{\text{orthonormal}} \lambda$$

and because we have an orthonormal basis: Δ
" we end up with $k\lambda$ "

$\Delta = \lambda$

as our loss function.

Therefore the average squared reconstruction error is minimized if λ is the smallest eigenvalue of the data covariance matrix

That means we need to choose b_2 as the corresponding eigen~~value~~ vector, ~~and~~ and that one will span the subspace that we will ignore.

b_1 which spans the principal subspace,
is then the eigen vector that belongs to the largest eigenvalue of the data covariance matrix

Keep in mind that the eigenvectors of the Covariance matrix are already orthogonal to each other because of the symmetry of the Co-variance matrix

If we look at a two Dim example, if there is our data (see pic):

Then the best projection that we can get, that retains most of the information is the one that projects onto the subspace that's spanned by the eigenvector of data covariance matrix which belong to the largest eigenvalue, and that's indicated by this long arrow over here (see pic)

Lets go to the general case:

If we want to find the n dim. principal subspace of a D dimensional dataset, and we solve for the basis vectors " b_j where j equals M+1 to D"

$$b_j \, , \; j = M+1, \ldots, D$$

we do optimize these ones, we end up with the same kind of eigenvalue problems that we had earlier with the simple example we end up with "S times b_j equals λ_j times b_j for j equal M+1 to D"

$$S b_j = \lambda_j b_j \;, \; j = M+1, \ldots, D$$

And the loss function is given by the sum of the corresponding eigenvalues.

we can write:

$$J = \sum_{j=M+1}^{D} \lambda_j$$

Also in the general case the average reconstruction error is minimized, if we choose the basis vectors that spans the ignored subspace to be eigenvectors of the data co-variance matrix that belong to the smallest eigenvalues.

This equivalently means, that the principal subspace is spanned by the eigenvectors belonging to the M largest eigenvalues of the data co-variance matrix

This nicely aligns with properties of the covariance "matrix

① The eigenvectors of the covariance matrix are orthogonal to each other, because of symmetry

② The eigenvector belonging to the largest eigenvalue, points ~~to~~ in the direction of data with largest variance

③ and the variance in that direction is given by the corresponding eigenvalue.

Similarly, the eigenvector belonging to the second largest eigenvalue, points in the direction 2nd largest variance of data and so on....

In this session we identified the orthonormal basis of the principal subspace as the eigenvectors of the data covariance matrix that are associated with the largest eigenvalues

In next session we going to put all pieces together and run through the PCA algorithm in detail.