


**Image Signal Processing**  
**Professor A. N. Rajagopalan**  
**Department of Electrical Engineering**  
**Indian Institute of Technology, Madras**  
**Lecture - 52**  
**Karhunen-Loeve Transform (KLT) - Concept**

(Refer Slide Time: 00:16)



KLT

$$\underline{R} = U \underline{R} U^H$$

$$E\{r(i,j)} = \frac{\sum_{i=1}^M r(i,i)}{\sum_{i=1}^M r(i,i)}$$


$$\alpha = \frac{\sum_{i=1}^M |r(i,i)|}{\sum_{i=1}^M |r(i,i)|}$$

$$\beta = \sum |r(i,i)|$$
  

$\underline{x}^H \underline{R} \underline{x} \geq 0$   
 $\underline{x}^H \lambda \underline{x} \geq 0$   
 $\lambda \underline{x}^H \underline{x} \geq 0 \quad \lambda \geq 0$

Eigenvalues of  $(\underline{R})$  are real and  $\geq 0$ .

$$\underline{x}^H \underline{R} \underline{x} = \underline{x}^H \underline{E} \underline{E}^H \underline{x} = \underline{E} (\underline{E}^H \underline{x}) (\underline{E}^H \underline{x}) = \underline{E} \underline{z} \underline{z}^H = \underline{E} \underline{z} \underline{z}^* \geq 0$$



Prof. A.N. Rajagopalan  
 Department of Electrical Engineering  
 IIT Madras

( Karhunen-Loeve Transform (KLT) - Concept )

So this how it was and by the way, you all know that for example, if in this case, we are taking, we are taking R to have kind of complex entries also. So given that, what was it that I wanted to say, so your eigenvalues, of course, eigenvalues of R are real and are greater than or equal to 0, because R is anyway a positive semi-definite matrix, it is a kind of a covariance.

So it is like saying that if you do if you take any x which is non-zero, do like Hermitian R x and then given the fact that R itself has the form, some expectation, some let us say y y Hermitian because it is actually a covariance matrix. So when I say R, I mean it is a covariance.

So then y y Hermitian x, so you can push this as x Hermitian y and then y Hermitian x and this is, of course, a scalar and so if you call some z is equal to x Hermitian y, which is a scalar then this is z z star which is like, which is always a number is actually greater than or equal to 0, it will be like magnitude z square and so we know this.

So, and because of, because of which we know that R is PSD, and therefore if you try to do, so given the fact that  $x^H R x$  is actually, is always a number greater than or equal to 0, if you take x to be and since this relation is true for all x, and if you take x to be the eigenvector itself then clearly  $x^H R x$  is some  $\lambda x$ . This is kind of greater than or equal to 0,  $\lambda x^H R x$  is actually a number greater than or equal to 0, and since you have taken x to eigenvector,  $x^H R x$  is non-zero, therefore  $\lambda$  itself is a number which is actually, which is greater than or equal to 0.

These are all things that you know. So when somebody gives you a covariance matrix, know that its eigenvalues are real, its eigenvalues have a value greater than or equal to 0. So all of this kind of makes sense, in the sense that energy and all that that we are talking about, all of that makes sense.

(Refer Slide Time: 03:01)

NPTEL

Random vector  $\underline{u}$

$\underline{v} = \underline{\psi}(\underline{u} - \underline{m})$  (zero-mean)

Columns of  $\underline{\psi}^H$  are the eigenvectors of  $R$

$E \underline{v} \underline{v}^H = E \underline{\psi}(\underline{u} - \underline{m})(\underline{u} - \underline{m})^H \underline{\psi}^H$

$= \underline{\psi} E(\underline{u} - \underline{m})(\underline{u} - \underline{m})^H \underline{\psi}^H$

$= \underline{\psi} R \underline{\psi}^H = \underline{\Lambda}_R$

$\underline{u} - \underline{m} = \underline{\psi}^H \underline{v}$

$\underline{u} = \underline{\psi}^H \underline{v} + \underline{m}$

$\underline{v} = \underline{A} \underline{u}$

$\underline{A}^H = \begin{bmatrix} | & | & | \\ \hline & & \\ \hline \end{bmatrix}$

Prof. A.N. Rajgopal  
Department of Electrical Engineering  
IIT Madras

(Karhunen-Loeve Transform (KLT) - Concept)

Now coming back to what I wrote the other day, so we said that we have, let us say samples from some sort of a random vector  $u$ , so we have this class of examples and this entire class is being represented by  $u$  and let us say this has some covariance  $R$ , this has some mean  $\mu$ ;  $m$ , maybe I will use  $m$ . So the idea is that if you take, if you do a transformation similar to the case which you have done with respect to a data-independent transform, we did some  $v$  is equal to  $A$  times  $u$  there.

Now, a similar transformation we would like to do now on  $u$ , but now we are going to use it on say, all the samples of  $u$  because  $u$  by itself it is not simply one sequence or something.  $u$  represents this entire class now, and therefore what we would like to do is, we would like to do a transformations, something like  $v$ ,  $v$  which is, which will again be a random vector,  $u$  is a random vector with some mean and covariance, mean  $m$  and covariance  $R$ .

So we want to do something like  $\psi$  and then  $u$  minus  $m$ .  $u$  minus  $m$  because we would like to make the data zero-mean, because  $u$  by itself need not be zero-mean data so we like to make it zero-mean. So we make the data zero-mean and then what is this  $\psi$ ? The rows of  $\psi$ ,  $\psi$  contain the or for example, or maybe we will go with the columns of, the columns of  $\psi$  Hermitian are the eigenvectors of  $R$ .

And this comes from, of course, the original theorem that I said, what is called a spectral theorem for normal matrices. So where we said that if this matrix is normal then it is a unitarily diagonalizable, which means that you can write  $R$  as some  $\psi$ , I think this we did that day. So if you do, if you act  $\psi$  on  $R$  from the left and  $\psi$  Hermitian on the right, then this is going to say, diagonalize  $R$ .

So, it is that  $\psi$  that I have taken here, so where we know that the columns of  $\psi$  Hermitian contain eigenvectors of  $R$ . So columns of  $\psi$  Hermitian are the eigenvectors of  $R$ . So  $\psi$  is, of course, a matrix, so  $u$  is a vector and  $m$  is a vector. So therefore clearly if you try to see after you do the transformation what will happen to this random vector  $v$ , clear that mean of  $v$  is 0 because mean of  $u$  is  $m$ , and then if you look at expectation  $v$   $v$  Hermitian that will be expectation  $v$  which is  $\psi u$  minus  $m$ , then  $v$  Hermitian is  $u$  minus  $m$  Hermitian,  $\psi$  Hermitian.

And since expectation is simply a linear operator we can actually send it inside, then it becomes expectation  $u$  minus  $m$ ,  $u$  minus  $m$  Hermitian  $\psi$  Hermitian. Then it means that you have got like, and since you, so this  $R$  is exactly this, expectation  $u$  minus  $m$   $u$  minus  $m$  Hermitian is  $R$ . Therefore this means that this is equal to  $\psi R \psi$  Hermitian, which we know is simply a diagonal form of  $R$ .

So we know, so the idea is that because we started with this result which came from the spectral theorem, we started with that result and then if you use that  $\psi$  here to do the transformation, so the point of what kind of a transformation should we then be doing on  $u$  in order to be able to arrive at this, say diagonal  $R$ . So the transformation that we need to do is this.

But if we do this transformation then we can see that expectation  $v^T v$  Hermitian which is the covariance of  $v$  is simply diagonal, that means you have been able de-correlate the data, so which is exactly what we want. We want to go from the standard basis to a different basis which is also orthonormal but then in that basis we want this diagonalization to happen and that diagonalization does happen and the de-correlation which is what we call as the de-correlation.

And the whole point is we can actually reconstruct now. If you are interested in  $u$ , we do the same thing just as we did there,  $u$  is equal to  $A$  Hermitian  $v$  and similar to that we can do like  $u$  minus  $m$  is equal to  $\psi$  Hermitian, which is the inverse of  $\psi$ , so  $\psi$  Hermitian  $v$  or we can get  $u$  as  $\psi$  Hermitian  $v$  plus  $m$ .

So this mean, of course, we have to add it back because we removed it from  $u$ , from  $v$ . When we did  $v$ , we removed the mean, therefore the mean now gets in and you can actually construct your  $u$ . Now, in this case, of course, it will all be exact because if you are using all the eigenvectors and the idea is that you can actually, you can afford to not use all the eigenvectors.

Now you can ask since, see as opposed to the data-independent case, here all your eigenvectors are coming from  $R$ . This  $\psi$  is kind of special for that  $R$ , correct. So since these eigenvectors have since these eigenvectors are kind of have been derived from this kind of a data, the data samples that you have with you, so in reality what you would do is somebody anyway, so maybe when I show that example I will show you in reality how you do it.

So for the time being, let us assume that somebody gave you  $R$ . So since you know  $R$  now you could actually compute all of this because, in the earlier data-independent and

all we never worried about covariance of the data or anything, you gave me a sequence  $u$ , I would just apply a data-independent transform give you  $v$ , I could not even ask what will be the error that I will make, I mean if let us say if I do not use all of the eigenvectors in  $A$  Hermitian. This is typically  $N$  cross  $N$  if you use  $N$  cross  $1$ .

Now suppose, I chose only  $M$  of those columns, then it means that I have got something like  $N$  cross  $M$ . So now if I try to use only  $M$  number of these columns, which means if I use only  $m$  number of these eigenvectors then what will be the error? We cannot say anything because for each  $u$  it can start to change. But, here with respect to this kind of a KL transform or the PCA, we can ask the same question but then we will have an answer for this.

Because right now, suppose I ask, in general, how well will I do if I sort of knock-out, knock off some of the eigenvectors? Now how do I knock off? I know the eigenvalues with respect to each of these eigenvectors, and therefore I will know the significance in each one of them. Because if you see here, so you can think of this as the linear combination of the eigenvector sitting in  $\psi$  Hermitian.

So  $v$  is like a column you can think of this as  $v_0$  times the first eigenvector plus  $v_1$  times the second eigenvector, all the way, use all of them and then add mean.

(Refer Slide Time: 10:05)

The slide contains the following handwritten content:

- NPTEL** logo in the top left corner.
- Equation 1:** 
$$\frac{v}{m \times 1} = \psi \begin{pmatrix} u - m \\ n \times 1 \\ n \times 1 \end{pmatrix}$$
  - Annotations: "Transform coeff" points to  $\psi$ ; "Ignore less significant eigenvalues" points to the second and third terms in the vector.
  - Underneath: "Dimensionality Reduction".
- Equation 2:** 
$$\frac{u}{n \times 1} = \psi \begin{pmatrix} m \\ n \times m \end{pmatrix} + \frac{m}{n \times 1}$$
  - Annotation: "KLT: Basis given are orthogonal spanning basis not there." points to the  $\psi$  matrix.
  - Annotation: "Meaning:  $C = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ " points to the second term.
  - Annotation: " $R = A \otimes B$ " is written below.
- Equation 3:** 
$$\|u - \hat{u}\|^2 = \sum_{i=m+1}^N \lambda_i \quad \text{MSE}$$
  - Annotation: "KLT: Benchmark" points to the equation.
  - Below it: "DFT" and "DCT" are written and underlined.
- Small photo:** Prof. A.N. Rajagopalan in the bottom left corner.
- Page-Footer:** "Prof. A.N. Rajagopalan, Department of Electrical Engineering, IIT Madras" and "(Karhunen-Loeve Transform (KLT) - Concept)".

But instead of this, suppose I say that, suppose we say that we do not use all of them and instead we do something like  $v$ , suppose I call this as an  $M \times 1$  vector instead of  $N \times 1$ . Ideally, I should have got an  $N \times 1$  as  $v$ , but suppose I chose to do just do  $M \times 1$  which then means that when I actually multiplies  $\psi$ , I said I will do my  $u$  minus, what was that,  $M$ , right?

Now,  $u$  is  $N \times 1$ , of course, this is  $N \times 1$  and therefore  $\psi$ , of course, clearly is an  $M \times N$ . So it is like saying that you have got like  $M$  rows in  $\psi$  and not  $N$  rows. So, when you actually do that you kind of, so such an operation is you can think of it as a transformation, I mean as a transform coefficient, this is what we call. So we call  $v$  as a transform coefficient.

Now, we can even afford to just include the first  $M$  eigenvectors in  $\psi$  Hermitian, not use all of them. So this we arrive by ignoring, ignore insignificant, or ignore less significant eigenvectors based on their eigenvalues.

So from now on, we will assume that everything is ordered. So the eigenvectors are ordered according to their significance, so the guy which is most significant is the first one that means it has a highest eigenvalue, then second and third. So ignore less significant eigenvectors and try to do a reconstruction which would only use the first  $M$  and not all of  $N$ , all the  $N$  eigenvectors.

So if we ignore the less significant eigenvectors and suppose, reconstruct; so this transform coefficient is what we will call it but then this is also what is called a dimensionality reduction. So people, when they talk about PCA one of the things that they like to talk about is doing some kind of dimensionality reduction. So it is like saying that you could, it could, it would ideally allow you to go from a kind of a very high dimensional feature space to or in this case, the image space to a kind of a lower-dimensional feature space.

So I will, so this dimensionality part, I will talk about it later. But for the time being, we will simply look upon it as some kind of a transform coefficient that we derived. But

earlier, we never did something like this, we were not truncating, we were always looking at exact reconstruction.

But in this case, because we can say something about the whole class, we can actually try doing this, and if you do this then when you reconstruct you will have your  $\hat{u}$ , let us say because I no longer can get my  $u$  back, so I will write it as  $\hat{u}$  which is  $N \times 1$  and then I will have a  $\psi$  Hermitian but now let me call this  $A$ , so what this is like  $N \times M$ ,  $\psi$  Hermitian will be  $N \times M$  and then this multiply is your  $v$  which is  $M \times 1$  plus  $m$  which is  $N \times N$ ,  $N \times 1$ .

Student: (12:56).

Professor: Correct. By class, I mean that, for example, all the samples that are sitting in that classroom which you derived your  $R$  and mean. So it is like saying suppose I had face, human faces, and suppose I collected lots and lots of human faces, I computed the covariance  $R$ , I computed the mean  $\mu$ , so I will have something like an average human face, it is like the mean image.

And then I will have what kind of spread it has and then after I do the, let us say eigenvector, eigenvalue decomposition then suppose I do not choose, so what might happen is some of the variations that are happening across us is because of maybe not, see think of the most significant eigenvectors trying to capture the real structure of the face and maybe there are lots and lots of eigenvectors which do not even mean anything, I mean so those variations are such things that you can ignore.

Ignore in the sense that I mean, if you ignore it you would like to know what will happen if we start ignoring. And because it is like saying that not, let us say every eigenvector sitting there is significant because the variations that happen across faces you may still be able to reconstruct that is something reasonable with which you may be able to work by not using all the Eigenfaces or the eigenvectors.

So in that sense, this is like an ensemble property. So over the entire class, we can ask suppose, but all this assumes that you have a good estimate of  $R$  which means that it

assumes that you have lots and lots of, you have collected lots and lots of examples, you have been able to estimate robustly some covariance.

Now you can ask if I reconstruct with let us say, a fewer not with all of them but then, fewer in the sense that sometimes it can be drastic, by fewer we are not talking 10, 20 lists. For example, if you look at an image let us say it is 64 square, minimum that is the kind of face you would look at; so that is like what 4096.

Now if you look at it, people have done this covariance analysis and then they say we will take the top 100 eigenvectors. So it is like 4096 to 100, somebody may even take 50. Again, what they do is they look at the significance of the eigenvalues, typically you say that I go down till maybe I reach a fraction of the maximum eigenvalue; you may say, you may cap it at whatever, 40 percent, 20 percent, 5 percent.

It depends on the application, but then what has been found is most of the time not just with faces and all, anytime there is a correlation that is high, you might be able to throw away lots of those eigenvalues because they do not really, they are not really the ones that are carrying the information, the ones that are the most significant ones are the ones that are carrying information.

What I wanted to tell was, so if you do this kind of reconstruction which is not exact anymore because you can only get an approximation of  $u$  now, but if you ask what will be the norm of  $u$  minus  $\hat{u}$ ,  $\hat{u}$  square in a sort of a Frobenius sense; okay, in this case, it is simply a Euclidean norm because  $u$  is a vector,  $\hat{u}$  is a vector now.

Now if you ask, now in this case, until now when we did the data-independent thing we could never tell anything. Suppose, I just take the, in a kind of a DFT, suppose I do not use all the, you look at  $\phi$  Hermitian, suppose I choose only a few of them what kind of an error will I get? You cannot say anything, if I take a  $u$  of a certain type it will get something, you change  $u$  you will get something else.

But here, you can actually, we can show that this is equal to, when we do SVD I will show an alternative way of doing this. So, I do not want to prove this but this is like



summation  $\lambda_i$ , where let us say, where  $i$  is equal to  $M + 1$  to  $N$ , that means all the eigenvalues that you actually left out, which is a very kind of a strong result.

It is like saying that, it does not mean that every time you take an image from that class and if you do a reconstruction will kind of hit this value, it means that if you really take a lot if you take lots and lots of images and you compute the average error, so this is like a mean square error. This is in a sort of a mean square sense because this over the whole class.

So over the whole class on the average, you will get, I mean you can expect that at least you have some sort of a ballpark kind of a figure, I mean you are not entirely blind to what will happen. So you know that on the average I will do only so bad or if you really, if you really ignore the good eigenvalues then it mean that you really did badly or else you will know that even if I ignore then I may not be, it may not be such a bad thing at all.

Now this kind of an, this kind of a mean square error number we do not even have for the data-independent cases. We cannot say anything in general as to what will happen, but in this case, as far as KLT is concerned we can actually have this figure. But all of this assumes that we have been able to compute  $R$  and mean and all pretty well, this all banks on these statistics, if you do badly there then all this will get affected.

If there are no doubts then what I thought I will do is, now one of the thing is you might say that all of this is so nice that why do not we then use it, right? The problem is KLT, there is no fast algorithm because first of all the basis images are, the basis the eigenvectors which are the basis images are data-dependent.

So, for example, if I change the class I will get another set of eigenvectors, if I go from face to something else, let us say I go from face to leaves, I will get something else; if we go from leaves to something like tools, I will get something else. So there is nothing like fixed this one, so as you keep hopping from one class to another, you have to tell the other person what is the basis image that you are using.

Then this notion of separability, as you go to a higher dimension also is generally not there, separability notion, but some people just assume it, so separability notion not there. In general, it is not there but sometimes what will happen is see, for example, if you think about the Markovian process that we had for 1D which was like  $\rho^m$  whatever  $m$  minus  $m'$  mod, this is what we had.

Now you can, so some people will now when they go to a higher dimension they will kind of assume that the  $R$  that you will get, let me put this as some script  $R$  it is simply the  $\rho$ , the  $R$  that you get for 1D which is a Kronecker  $R$ , which is like saying that your  $R$ , if you think about this, is  $R_{m, n}$ , and maybe  $m'$  comma  $n'$ , then they will assume that this follows the model into  $\rho_{n, n'}$  mod.

This is something that they will impose. One does not know whether it will actually happen, but if it does happen when you go from a lower dimension to a higher dimension, then all of this we can do. For example, if you want to go from, similar to the case that we did from 1D to sort of a 2D, then you can actually, you can kind of reuse the basis but then it is not always true that such a thing will happen.

So the separability notion is not automatic, fast algorithms just do not exist, and therefore it ends up being, so if you ask then why would I even bother about it then like I said right at the beginning, so the KLT is really, really a benchmark, so it is used as a benchmark.

So anything else that can come close to it will be nice because, on the one hand, you will have fast algorithms like we said, at DFT if  $R$  is circulant so it would actually mean that DFT will be the KLT so which will then mean that it will have all these kind of nice things that we said, it will have maximum energy packing efficiency, it will have data de-correlation that is 100 percent, then this kind of a norm error on the average, the mean square error notion you will know.

You will know all of that and on top of that because it is a data-independent transform you also know that a DFT can be run very fast, you have fast algorithms for the separability notion, there are all of that.

Similarly, DCT if you do it for, let us say 1D this one, a Markov process then we know that, a first-order Markov process then we know that this guy will come close to being the KLT, and therefore so in that sense, this is still, as far as the coding part is concerned where you want to analyze from that sort of a perspective, then you only use this as some something that you put way up there and then try to see which other transform even tends to come close.

See when I say that we do not talk about data, I said data-independent for a class because data-independent means, itself means that I do not even want to know from where, from what class it is coming. The moment you say data, I mean, otherwise it will not be data-independent, if I make it dependent on the class it will no longer be data-independent.

The very notion of data-independent is that I do not care from where it is coming, I do not care whether it is a human face or whether it is some other things, whether it is simply a natural scene, I do not want to really worry about from where it is coming. The moment you start worrying about examples from that and all then it means, then it means you are fine-tuning yourself to a particular class, in which case you will end up with this, I mean if you had enough data then this is what you would end up doing.

The whole idea is that data-independent, nice thing is that those have these, those are those are fast and separable all those things are there, and on top of that, you do not need to worry about what should be my this one, basis. But if you can show that something that is data-independent can be an optimal transform that is very good.

So that is where this equivalence to KLT makes sense because if you can somehow show that something else that is data-independent comes very close to a KLT then it means that you will rather pick that then pick something else. So this is like a, this helps you in order to know what transform to pick if you knew that your data was following a certain characteristic and then if you knew that a certain sort of independent transform would diagonally a covariance corresponding to that then you can talk about it.

So, yeah. This is all about ensemble characteristics, this is not about one particular example or something this is kind of an ensemble characteristic. Now the, but then that by itself, it does not mean that we limit the KLT or PCA just for this.

As the transform part is concerned because we have started with image transform, we wanted to talk about this because this and the SVD are like two things which are, which depend on the, on your image but whereas the others and all were totally independent. Now one more, but then that is as far as you look at unitary transforms and all.