GILBERT STRANG: So this is a pretty key lecture. This lecture is about principal component analysis, PCA-- which is a major tool in understanding a matrix of data. So what is PCA about? Well first of all, let me remember what was the whole point of last-- yesterday's lecture-- the singular value decomposition, that any matrix A could be broken into r rank 1 pieces-- r being the rank of the matrix.

And each piece has a U times a V transpose. And the good-- special thing is, the U's are orthonormal, and also, the V's are orthonormal. OK. So that's the whole matrix. But we have a big matrix, and we want to get the important information out of it-- not all the information. And people say, in machine learning, if you've learned all the training data, you haven't learned anything, really. You've just copied it all in. The whole point of neural nets and the process of machine learning is to learn important facts about the data.

And now, here we're at the most basic stage of that. And I claim that the important facts about the matrix are in its largest k singular values-- the largest k pieces. We can take-- k equal 1 would tell us the largest single piece. But maybe we have space and computing power to handle a hundred pieces. So I would take k equal 100. The matrix might have ranked thousands. So I claim that Ak is the best. Now here's the one theorem for today, that Ak-- using the first k pieces of the SVD-- is the best approximation to A of rank k.

So I'll write that down. So that really says why the SVD is perfect. OK. So that statement says, that if B-- another matrix-- has rank k, then the distance from A to B-- the error you're making in just using B-- that error is greater than or equal to the error you make for the best guy. Now that's a pretty straightforward, beautiful fact. And it goes back to people who discovered the SVD in the first place.

But then a couple of psychologists gave a proof in a later paper-- and it's often called the Eckart-Young Theorem. There is the theorem. Isn't that straightforward? And the hypothesis is straightforward. That's pretty nice. But of course, we have to think, why is it true? Why is it true? And to give meaning to the theorem, we have to say what these double bars are. Do you know the right name for this? So that double bar around a matrix is called the-- the norm of the matrix, the norm. So I have to say something about matrix norms.

How big is-- that's a measure of how big it is. And what I have to say is, there are many different measures of a matrix-- how large that matrix is. Let me tell you, for today, three possible measures of a matrix. So different ways to measure-- I'll call the matrix just A, maybe. But then I'm going to apply the measure to A minus B, and to A minus AK, and show that that is smaller. OK.

So I want to tell you about the norm of A-- about some possible norms of A. And actually, the norms I'm going to take today will be-- will have the special feature that they can be found-- computed by their singular values. So let me mention the L2 norm. That is the largest singular value. So that's an important measure of the-- sort of the size of a matrix. I'm talking here about a general m by n matrix A. Sigma 1 is an important norm-- often called the L2 norm. And that's where that index 2 goes.

Oh. I should really start with vectors-- norms of vectors-- and then build to the norms of matrices. Let me do norms of vectors over on this side. The L2 norm of a vector-- do we know what that is? That's the regular length of the vector that we all expect-- the square root of v1 squared up to vn squared. The hypotenuse-- the length of the hypotenuse in n dimensional space. That's the L2 norm, because of that 2. The L1 norm of a vector is just add up those pieces without squaring and square rooting them. Just add them.

That's the L1 norm. And you might say, why do we want two norms? Or there are more norms. Let me just tell you one more. The infinity norm-- and there is a reason for the 1 and the 2 and the infinity-- is the largest of the v's. OK. Have you met norms before? I don't know. These are vector norms, but maybe you have met. Then we're going to have matrix norms, that maybe will be new. So this is the norm that we usually think of.

But this one has become really, really important, and let me tell you just why. And then we'll-- later section of the notes and a later lecture in this course will develop that-- develop this. This is the L1 norm. So this is L2, L1, and L infinity-- [INAUDIBLE]. So what's special about this one? Well, it just turned out-- and it was only discovered in our lifetimes-- that when you minimize some function using the L1 norm, you minimize some, let's say, signal the noise, or whatever you minimize-- some function.

If you use L1, the winning vector-- the minimizing vector-- turns out to be sparse. And what does sparse mean? Sparse means mostly zero components. Somehow, when I minimize in L2-- which historically goes back to Gauss, the greatest mathematician of all time. When you minimize something in L2, you do the least squares. And you find that the guy that gives you the minimum has a lot of little numbers-- lot of little components. Because when you're square those little ones, they don't hurt much.

But Gauss-- so Gauss didn't do least L1 norm. That has different names-- basis pursuit. And it comes into signal processing and sensing. Right. And then it was discovered that if you minimize-- as we'll see in that norm-- you amazingly get-- the winning vector has-- is mostly zeros. And the advantage of that is that you can understand what its components are. The one with many small components, you have no interpretation for that answer. But for an answer that just has a few non-zero components, you really see what's happening.

And then this is a important one, too. OK. Now I'm going to turn just to-- so what's the property of a norm? Well, you can see that the norm of C times a vector is-- just multiplying by 6, or 11, or minus pi, or whatever-- is the size of C. Norms have that nice property. They're homogeneous, or whatever word. If you double the vector, you should double the norm-- double the length. That makes sense. And then the important property is that-- is the famous triangle in equality-- that if v and w are two sides of a triangle, and you take the norm of v and add to the norm of w-- the two sides-- you get more than the straight norm along the hypotenuse.

Yeah. So those are properties that we require, and the fact that the norm is positive, which is-- I won't write down. But it's important too. OK. So those are norms, and those will apply also to matrix norms. So if I double the matrix, I want to double its norm. And of course, that works for that 2 norm. And actually, probably-- so the triangle in equality for this norm is saying that the largest singular value of A plus B-- two matrices-- is less or equal to the larger the singular value of A plus the largest singular value of B.

And that's-- we won't take class time to check minor, straightforward things like that. So now I'm going to continue with the three norms that I want to tell you about. That's a very important one. Then there is another norm that's named-- has an F. And it's named after Frobenius. Sorry about that. And what is that norm? That norm looks at all the entries in the matrix-- just like it was a long vector-- and squares them all, and adds them up.

So in a way, it's like the 2 norm for a vector. It's-- so the squared-- or shall I put square root? Maybe I should. It's the square root of all the little people in the matrix. So a1, n squared, plus the next a2, 1 squared, and so on. You finally get to a-m-n squared. You just treat the matrix like a long vector. And take this square root just like so. That's the Frobenius norm. And then finally, not so well known, is something that's more like L1. It's called the nuclear norm.

And not all the faculty would know about this nuclear norm. So it is the sum of the sigma of the singular values. I guess there are r of them. So that's where we would stop. Oh, OK. So those are three norms. Now why do I pick on those three norms? And here's the point-- that for those three norms, this statement is true. I could cook up other matrix norms for which this wouldn't work. But for these three highly important norms, this Eckart-Young statement, that the closest rank k approximation is found from the first k pieces.

You see, that's a good thing, because this is what we compute from the SVD. So now we've solved an approximation problem. We found the best B is Ak. And the point is, it could use all the-- any of those norms. So there would be a-- well, somebody finally came up with a proof that does all three norms at once. In the notes, I do that one separately from Frobenius. And actually, I found-- in an MIT thesis-- I was just reading a course 6 PhD thesis-- and the author-- who is speaking tomorrow, or Friday in IDSS-- Dr. [? Cerebro ?] found a nice new proof of Frobenius. And it's in the notes, as well as an older proof. OK.

You know, as I talk here, I'm not too sure whether it is essential for me to go through the proof, either in the L2 norm-- which takes half a page in then notes-- or in the Frobenius norm, which takes more. I'd rather you see the point. The point is that, in these norms-- and now, what is special about these norms of a matrix? These depend only on the sigmas-- only on the-- oh. Oh. I'll finish that sentence, because it was true.

These norms depend only on the singular values. Right? That one, at least, depends only on the singular value. It's the largest one. This one is the sum of them all. This one comes into the Netflix competition, by the way. This was the right norm to win a zillion dollars in the Netflix competition. So what did Netflix put-- it did a math competition. It had movie preferences from many, many Netflix subscribers. They gave their ranking to a bunch of movies.

But of course, they hadn't seen-- none of them had seen all the movies. So the matrix of rankings-- where you had the ranker and the matrix-- is a very big matrix. But it's got missing entries. If the ranker didn't see the movie, he isn't-- he or she isn't ranking it. So what's the idea about Netflix? So they offered like a million dollar prize. And a lot of math and computer science people fought for that prize. And over the years, they got like higher 92, 93, 94% right. But it turned out that this was-- well, you had to-- in the end, you had to use a little psychology of how people voted.

So it was partly about human psychology. But it was also a very large matrix problem with an incomplete matrix-- an incomplete matrix. And so it had to be completed. You had to figure out what would the ranker have said about the post if he hadn't seen it, but had ranked several other movies, like All the President's Men, or whatever-- given a ranking to those? You have to-- and that's a recommender system, of course. That's how you get recommendations from Amazon.

They've got a big matrix calculation here. And if you've bought a couple of math books, they're going to tell you about more math books-- more than you want to know. Right. OK. So anyway, it just turned out that this norm was the right one to minimize. I can't give you all the details of the Netflix competition, but this turned out to be the right norm to do a minimum problem, a best not least squares. These squares would look at some other norm, but a best nuclear norm completion of the matrix.

And that-- and now it's-- so now it's being put to much more serious uses for MRI-- magnetic resonance stuff, when you go in and get-- it's a noisy system, but you get-- it gives a excellent picture of what's going on. So I'll just write Netflix here. So it gets in the-- and then MRIs. So what's the point about MRIs? So if you don't-- if you stay in long enough, you get all the numbers. There isn't missing data. But if you-- as with a child-- you might want to just have the child in for a few minutes, then that's not enough to get a complete picture.

And you have, again, missing data in your matrix in the image from the MRI. So then, of course, you've got to complete that matrix. You have to fill in, what would the MRI have seen in those positions where it didn't look long enough? And again, a nuclear norm is a good one for that. OK. So there will be a whole section on norms, maybe just about-- in stellar by now. OK. So I'm not going to-- let me just say, what does this say? What does this tell us?

I'll just give an example. Maybe I'll take-- start with the example that's in the notes. Suppose k is 2. So I'm looking among all rank 2 matrices. And suppose my matrix is 4, 3, 2, 1, and all the rest 0's. Diagonal. And it's rank 4 matrix. I can see its singular values. They're sitting there. Those would be the singular values, and the eigenvalues, and everything, of course. Now, what would be A2? What would be the best approximation of rank 2 to that matrix, in this sense to be completed? What would A2 do?

Yeah. It would be 4 and 3. It would pick the two largest. So I'm looking at Ak. This is k to the 2, so it has to have rank 2. This has got rank 4. The biggest pieces are those. OK.

So this thing says that if I had any other matrix B, it would be further away from A than this guy. It says that this is the closest. And I just-- could you think of a matrix that could possibly be closer, and be rank 2? Rank two 2 the tricky thing.

The matrices of rank 2 form a kind of crazy set. If I add a rank 2 matrix to a rank 2 matrix, probably the rank is up to 4. So the rank 2 matrices are all kind of floating around in their own little corners. This looks like the best one. But in the notes I suggest, well, you could get a rank 2-- well, what about B? What about this B?

For this guy, I could get closer-- maybe not exact-- but closer, maybe by taking 3.5, 3.5. But I only want to use rank-- I've only got two rank 2 to play with. So I better make this into a rank-- I have to make this into a rank 1 piece, and then the 2 and the 1.

So you see what I-- what I thought of? I thought, man, maybe that's better-- like on the diagonal, I'm coming closer. Well, I'm not getting it exactly here. But then I've got one left to play with. And I'll put, maybe, 1.5 down here. OK. So that's a rank 2 matrix-- two little rank 1s. And on the diagonal, it's better. 3.5s-- I'm only missing by a half. 1.5s-- I'm missing by half. So I'm only missing by a half on the diagonal where this guy was missing by 2. So maybe I've found something better. But I had to pay a price of these things off the diagonal to keep the rank low. And they kill me.

So that B will be further away from A. The error, if I computed A minus B, and computed its norm, I would see bigger than A minus A2. Yeah. So, you see the point of the theorem? That's really what I'm trying to say, that it's not obvious. You may feel, well, it's totally obvious. Pick 4 and 3. What else could do it? But it depends on the norm and so on. So it's not-- Eckart-Young had to think of a proof, and other people, too. OK. So that's-- now, but you could say-- also say-- object that I started with a diagonal matrix here. That's so special.

But what I want to say is the diagonal matrix is not that special, because I could take A-- so let me now just call this diagonal matrix D-- or let me call it sigma to give it another sort of appropriate name. So if I thought of matrices, what I want to say is, this could be the sigma matrix. And there could be a U on the left of it, and a sigma on the right of it. So A is U sigma V transpose. So this is my sigma. And this is like any orthogonal matrix U. And this is like any V transpose. Right?

I'm just saying, here's a whole lot more matrices. There is just one matrix. But now, I have all these matrices with Us multiplying on the left, and V transpose ones on the right. And I ask you this question, what are the singular values of that matrix, A? Here the singular values were clear-- 4, 3, 2, and 1. What are the singular values of this matrix A, when I've multiplied by a orthogonal guy on both sides? That's a key question. What are the singular values of that one?

AUDIENCE: 4, 3, 2, 1.

GILBERT STRANG: 4, 3, 2, 1. Didn't change. Why is that? Because the singular values are the-- because this has a SVD form-- orthogonal times diagonal times orthogonal. And that diagonal contains the singular values. What I'm saying is, that my-- and our-- trivial little example here, actually was all 4 by 4's that have these singular values. I could-- my whole problem is orthogonally invariant, a math guy would say. When I multiply by U or a V transpose, or both-- the problem doesn't change. Norms don't change.

Yeah, that's a point. Yeah. I realize it now. This is the point. If I multiply the matrix A by an orthogonal matrix U, it has all the same norms-- doesn't change the norm. Actually, that was true way back for vectors with this length-- with this length. What's the deal about vectors? Suppose I have a vector V, and I've computed its hypotenuse and the norm. And now I look at Q times V in that same 2 norm. What's special about that? So I took any vector V and I know what its length is-- hypotenuse. Now I multiply by Q.

What happens to the length? Doesn't change. Doesn't change. Orthogonal matrix-- you could think of it as just like rotating the triangle in space. The hypotenuse doesn't change. And we've checked that, because we could-- the check is to square it. And then you're doing QV, transpose QV. And you simplify it the usual way. And then you have Q transpose Q equal the identity. And you're golden. Yeah. So the result is you get the same answer as V.

So let me put it in a sentence now, pause. Multiplying that norm is not changed by orthogonal matrix. And these norms are not changed by orthogonal matrices, because if I multiply the A here by an orthogonal matrix, I have-- this is my A. If i multiply by a Q, then I have QU sigma V transpose. And what is really the underlying point? That QU is an orthogonal matrix just as good as U. So if I-- let me put this down.

QA would be QU sigma V transpose. And now I'm asking you, what's the singular value decomposition for QA? And I hope I may actually-- seeing it. What's the singular value decomposition of QA? What are the singular values? What's the diagonal matrix? Just look there for it. The diagram matrix is sigma. What goes on the right of it? The V transpose. And what goes on the left of it is QU. Plus, that's orthogonal times orthogonal. Everybody in this room has to know that if I multiply two orthogonal matrices, the result is, again, orthogonal.

So I can multiply by Q, and it only affects the U part, not the sigma part. And so it doesn't change any of those norms. OK. So that's fine. That's what I wanted to say about the Eckart-Young Theorem-- not proving it, but hopefully giving you an example there of what it means-- that this is the best rank to approximate that one. OK. So that's the key math behind PCA. So now I have to-- want to, not just have to-- but want to tell you about PCA. So what's that about?

So we have a bunch of data, and we want to see-- so let me take a bunch of data-- bunch of data points-- say, points in the plane. So I have a bunch of data points in the plane. So here's my data vector. First, vector x1-- well, x. Is at a good-- maybe v1. These are just two component guys. v2. They're just columns with two components. So I'm just measuring height and age, and I want to find the relationship between height and age. So the first row is meant-- is the height of my data. And the second row is the ages.

So these are-- so I've got say a lot of people, and these are the heights and these are the ages. And I've got n points in 2D. And I want to make sense out of that. I want to look for the relationship between height and age. I'm actually going to look for a linear row relation between height and age. So first of all, these are all over the place. So the first step that a statistician does, is to get mean 0. Get the average to be 0.

So what is-- so all these points are all over the place. From row 1, the height, I subtract the average height. So this is A-- the matrix I'm really going to work on is my matrix A-- minus the average height-- well, in all components. So this is a, a, a, a-- I'm subtracting the mean, so average height and average age. Oh, that was a brilliant notation, a sub a can't be a sub a. You see what the matrix has done-- this matrix 2 means?

It's just made each row of A. Now adds to row. Now add to what? If I have a bunch of things, and I've subtracted off their mean-- so the mean, or the average is now 0-- then those things add up to--

AUDIENCE: Zero.

GILBERT STRANG: Zero. Right. I've just brought these points into something like here. This is age, and this is height. And let's see. And by subtracting, it no longer is unreasonable to have negative age and negative height, because-- so, right. The little kids, when I subtract it off the average age, they ended up with a negative age. The older ones ended up still positive. And somehow, I've got a whole lot of points, but hopefully, their mean is now zero. Do you see that I've centered the data at 0, 0? And I'm looking for-- what am I looking for here? I'm looking for the best line.

That's what I want to find. And that would be a problem in PCA. What's the best linear relation? Because PCA is limited. PCA isn't all of deep learning by any means. The whole success of deep learning was the final realization, after a bunch of years, that they had to have a nonlinear function in there to get to model serious data. But here's PCA as a linear business. And I'm looking for the best line. And you will say, wait a minute.

I know how to find the best line, just use least squares. Gauss did it. Can't be all bad. But PCA-- and I was giving a talk in New York when I was just learning about it. And somebody said, what you're doing with PCA has to be the same as least squares-- it's finding the best line. And I knew it wasn't, but I didn't know how to answer that question best. And now, at least, I know better. So the best line in least squares-- can I remind you about least squares? Because this is not least squares.

The best line of least squares-- so I have some data points. And I have a best line that goes through them. And least squares, I don't always center the data to mean zero, but I could. But what do you minimize in least squares-- least squares? If you remember the picture in linear algebra books of least squares, you measure the errors-- the three errors. And it's how much you're wrong at those three points. Those are the three errors.

A-- difference between Ax and B-- the B minus Ax that you square. And you add up those three errors. And what's different over here? I mean, there's more points, but that's not the point. That's not the difference. The difference is, in PCA, you're measuring perpendicular to the line. You're adding up all these little guys, squaring them. So you're adding up their squares and minimizing. So the points-- you see it's a different problem? And therefore it has a different answer.

And this answer turns out to involve the SVD, the sigmas. Where this answer, you remember from ordinary linear algebra, just when you minimize that, you got to an equation that leads to what equation for the best x? So do you remember?

AUDIENCE: [INAUDIBLE]

GILBERT STRANG: Yeah. What is it now? Everybody should know. And we will actually see it in this course, because we're doing the heart of linear algebra here. We haven't done it yet, though. And tell me again, what equation do I solve for that problem?

AUDIENCE: A transpose A.

GILBERT STRANG: A transpose A x hat equal A transpose b. Called the normal equations. It's sort part of-- it's this regression in statistics language. That's a regression problem. This is a different problem. OK. Just so now you see the answer. So that involves-- well, they both involve A transpose A. That's sort of interesting, because you have a rectangular matrix A, and then sooner or later, A transpose A is coming. But this involves solving a linear system of equations. So it's fast.

And we will do it. And it's very important. It's probably the most important application in 18.06. But it's not the same as this one. So this is now in 18.06, maybe the last day is PCA. So I didn't put those letters-- Principal Component Analysis-- PCA. Which statisticians have been doing for a long time. We're not doing something brand new here. But the result is that we-- so how does a statistician think about this problem, or that data matrix? What-- if you have a matrix of data-- 2 by 2 rows and many columns-- so many, many samples-- what-- and we've made the mean zero.

So that's a first step a statistician takes to check on the mean. What's the next step? What else does a statistician do with data to measure how-- its size? There's another number. There's a number that goes with the mean, and it's the variance-- the mean and the variance. So somehow we're going to do variances. And it will really be involved, because we have two sets of data-- heights and ages. We're really going to have a covariance-- covariance matrix-- and it will be 2 by 2.

Because it will tell us not only the variance in the heights-- that's the first thing a statistician would think about-- some small people, some big people-- and variation in ages-- but also the link between them. How are the height, age pairs-- does more height-- does more age go with more height? And of course, it does. That's the whole point here. So it's this covariance matrix.

And that covariance matrix-- or the sample covariance matrix, to give it its full name-- what's the-- so just touching on statistics for a moment here.

What's the-- when we see that word sample in the name, what is that telling us? It's telling us that this matrix is computed from the samples, not from a theoretical probability distribution. We might have a proposed distribution that the height follows the age-- height follows the age by some formula. And that would give us theoretical variances. We're doing sample variances, also called empirical covariance made. Empirical says-- empirical-- that word means, from the information, from the data.

So that's what we do. And it is exactly-- it's AA transpose. You have to normalize it by the number of data points, N. And then, for some reason-- best known to statisticians-- it's N minus 1. And of course, they've got to be right. They've been around a long time and it should be N minus 1, because somehow 1 degree of freedom was accounted for when we took away-- when we made the mean 0. So we-- anyway, no problem.

But the N minus 1 is not going to affect our computation here. This is the matrix that tells us that's what we've got to work with. That's what we've got to work with-- the matrix AA transpose. And then the-- so we have this problem. So we have a-- yeah. I guess we really have a minimum problem. We want to find-- yeah. What problem are we solving? And it's-- yeah. So our problem was not least squares-- not the same as least squares. Similar, but not the same.

We want to minimize. So we're looking for that best line where age equals some number, c, times the height, times the-- yeah-- or height. Maybe it would have been better to age here and height here. No. No, because there are two unknowns.

So I'm looking for c. I'm looking for the number c-- looking for the number c. And with just two minutes in class left, what is that number c going to be, when I finally get the problem stated properly, and then solve it? I'm going to learn that the best ratio of age to height is sigma 1. Sigma 1.

That's the one that tells us how those two are connected, and the orthogonal-- and what will be the best-- yeah. No. Maybe I didn't answer that the right-- maybe I didn't get that right. Because I'm looking for-- I'm looking for the vector that points in the right way. Yeah. I'm sorry. I think the answer is, it's got to be there in the SVD. I think it's the vector you want. It's the principal component you want. Let's do that properly on Friday. I hope you see-- because this was a first step away from the highlights of linear algebra to problem solve by linear algebra, and practical problems, and my point is that the SVD solves these.