

OK. Here's lecture sixteen and if you remember I ended up the last lecture with this formula for what I called a projection matrix.

And maybe I could just recap for a minute what is that magic formula doing?

For example, it's supposed to be -- it's supposed to produce a projection, if I multiply by a b , so I take P times b , I'm supposed to project that vector b to the nearest point in the column space.

OK.

Can I just -- one way to recap is to take the two extreme cases.

Suppose a vector b is in the column space?

Then what do I get when I apply the projection P ?

So I'm projecting into the column space but I'm starting with a vector in this case that's already in the column space, so of course when I project it I get B again, right.

And I want to show you how that comes out of this formula.

Let me do the other extreme.

Suppose that vector is perpendicular to the column space.

So imagine this column space as a plane and imagine b as sticking straight up perpendicular to it.

What's the nearest point in the column space to b in that case?

So what's the projection onto the plane, the nearest point in the plane, if the vector b that I'm looking at is -- got no component in the column space, it's sticking completely -- ninety degrees with it, then Pb should be zero, right.

So those are the two extreme cases.

The average vector has a component P in the column space and a component perpendicular to it, and what the projection does is it kills this part and it preserves this part.

OK. Can we just see why that's true?

Just -- that formula ought to work.

So let me start with this one.

What vectors are in the -- are perpendicular to the column

space? How do I see that I really get zero?

I have to think, what does it mean for a vector b to be perpendicular to the column space?

So if it's perpendicular to all the columns, then it's in some other space.

We've got our four spaces so the reason I do this is it's perfectly using what we know about our four spaces.

What vectors are perpendicular to the column space?

Those are the guys in the null space of A transpose,

right? That's the first section of this chapter, that's the key geometry of these spaces.

If I'm perpendicular to the column space, I'm in the null space of A transpose.

OK. So if I'm in the null space of A transpose, and I multiply this big formula times b , so now I'm getting Pb , this is now the projection, Pb , do you see that I get zero?

Of course I get zero.

Right at the end there, A transpose b will give me zero right away.

So that's why that zero's here.

Because if I'm perpendicular to the column space, then I'm in the null space of A transpose and A transpose b is

OK, what about the other possibility. zilch.

How do I see that this formula gives me the right answer if b is in the column space?

So what's a typical vector in the column space?

It's a combination of the columns.

How do I write a combination of the columns?

So tell me, how would I write, you know, your everyday vector that's in the column space?

It would have the form A times some x , right?

That's what's in the column space, A times something.

That makes it a combination of the columns.

So these b 's were in the null space of A transpose.

These guys in the column space, those b 's are Ax -s.

Right?

If b is in the column space then it has the form Ax .

I'm going to stick that on the quiz or the final for sure.

That you have to realize -- because we've said it like a thousand times that the things in the column space are vectors A times x .

OK. And do you see what happens now if we use our formula?

There's an A transpose A .

Gets canceled by its inverse.

We're left with an A times x .

So the result was Ax .

Which was b .

Do you see that it works?

This is that whole business.

Cancel, cancel, leaving Ax .

And Ax was b .

So that turned out to be b , in this case.

OK, so geometrically what we're seeing is we're taking a vector -- we've got the column space and perpendicular

to that is the null space of A transpose.

And our typical vector b is out here.

There's zero, so there's our typical vector b , and what we're doing is we're projecting it to P .

And the -- and of course at the same time we're finding the other part of it which is e .

So the two pieces, the projection piece and the error piece, add up to the original b .

OK.

That's like what our matrix does.

So this is P -- P is -- this P is Ab , is sorry -- is Pb , it's the projection, applied to b , and this one is -- OK, that's a projection too.

That's a projection down onto that space.

What's a good formula for it?

Suppose I ask you for the projection of the projection matrix onto the -- this space, this perpendicular space?

So if this projection was P , what's the projection that gives me e ?

It's the -- what I want is to get the rest of the vector, so it'll be just I minus P times b , that's a projection too.

That's the projection onto the perpendicular space.

OK.

So if P 's a projection, I minus P is a projection.

If P is symmetric, I minus P is symmetric.

If P squared equals P , then I minus P squared equals I minus P .

It's just -- the algebra -- is only doing what your -- picture is completely telling you.

But the algebra leads to this expression.

That expression for P given -- given a basis for the subspace, given the matrix A whose columns are a basis for our column space.

OK, that's recap because you -- you need to see that formula more than once.

And now can I pick up on using it?

So now -- and the -- it's like, let me do that again, I'll go right through a problem that I started at the end, which is find a best straight line.

You remember that problem, I -- I picked a particular set of points, they weren't specially brilliant, t equal one, two, three, the heights were one, two, and then two again.

So they were -- heights were that point, that point, which makes it look like I've got a nice forty-five-degree line -- but then the third point didn't lie on the line.

And I wanted to find the best straight line.

So I'm looking for the -- this line, $y=C+Dt$.

And it's not going to go through all three points, because no line goes through all three points.

So I'm going to pick the best line, the -- the best being the one that makes the overall error as small as I can make it.

Now I have to tell you, what is that overall error?

And -- because that determines what's the winning line.

If we don't know -- I mean we have to decide what we mean by the error -- and then we minimize and we find the right -- the best C and D .

So if I went through this -- if I went through that point,

OK. I would solve the equation $C+D=1$.

Because at t equal to one -- I'd have C plus D , and it would come out right.

If it went through this point, I'd have C plus two D equal to

two. Because at t equal to two, I would like to get the answer

two. At the third point, I have C plus three D because t is three, but the -- the answer I'm shooting for is

two again. So those are my three equations.

And they don't have a solution.

But they've got a best solution.

What do I mean by best solution?

So let me take time out to remember what I'm talking about for best solution.

So this is my equation $Ax=b$.

A is this matrix, one, one, one, one, two, three. x is my -- only have two unknowns, C and D, and b is my right-hand side, one, two, three.

OK. No solution.

Three eq- I have a three by two matrix, I do have two independent columns -- so I do have a basis for the column space, those two columns are independent, they're a basis for the column space, but the column space doesn't include that vector.

So best possible in this -- what would best possible mean?

The way that comes out to linear equations is I -- I want to minimize the sum of these -- I'm going to make an error here.

I'm going to make an error here.

I'm going to make an error there.

And I'm going to sum and square and add up those errors.

So it's a sum of squares.

It's a least squares solution I'm looking for.

So if I -- those errors are the difference between Ax and b .

That's what I want to make small.

And the way I'm measuring this -- this is a vector,

right? This is e_1, e_2, e_3 .

The $Ax=b$, this is the e .

The error vector.

And small means its length.

The length of that vector.

That's what I'm going to try to minimize.

And it's convenient to square.

If I make something small, I make -- this is a never negative quantity, right?

The length of that vector.

The length will be zero exactly when the -- when I have the zero vector here.

That's exactly the case when I can solve exactly, b is in the column space, all great.

But I'm not in that case now.

I'm going to have an error vector, e .

What's this error vector in my picture?

I guess what I'm trying to say is there's -- there's two pictures of what's going on.

There's two pictures of what's going on.

One picture is -- in this is the three points and the line.

And in that picture, what are the three errors?

The three errors are what I miss by in this equation.

So it's this -- this little bit here.

That vertical distance up to the line.

There's one -- sorry there's one, and there's C plus D .

And it's that difference.

Here's two and here's $C+2D$.

So vertically it's that distance -- that little error there is e_1 .

This little error here is e_2 .

This little error coming up is e_3 .

e_3 .

And what's my overall error?

Is e_1 square plus e_2 squared plus e_3 squared.

That's what I'm trying to make small.

I -- some statisticians -- this is a big part of statistics, fitting straight lines is a big part of science -- and specifically statistics, where the right word to use would be regression.

I'm doing regression here.

Linear regression.

And I'm using this sum of squares as the measure of error.

Again, some statisticians would be -- they would say, OK, I'll solve that problem because it's the clean problem.

It leads to a beautiful linear system.

But they would be a little careful about these squares, for -- in this case.

If one of these points was way off.

Suppose I had a measurement at t equal zero that was way off.

Well, would the straight line, would the best line be the same if I had this fourth point?

Suppose I have this fourth data point.

No, certainly the line would -- it wouldn't be the -- that wouldn't be the best line.

Because that line would have a giant error -- and when I squared it it would be like way out of sight compared to the others.

So this would be called by statisticians an outlier, and they would not be happy to see the whole problem turned topsy-turvy by this one outlier, which could be a mistake, after all.

So they wouldn't -- so they wouldn't like maybe squaring, if there were outliers, they would want to identify them.

OK. I'm not going to -- I don't want to suggest that least squares isn't used, it's the most used, but it's not exclusively used because it's a little -- overcompensates for outliers.

Because of that squaring.

OK. So suppose we don't have this guy, we just have these three equations.

And I want to make -- minimize this error.

OK. Now, what I said is there's two pictures to look at.

One picture is this one.

The three points, the best line.

And the errors.

Now, on this picture, what are these points on the line, the points that are really on the line?

So they're -- points, let me call them P_1 , P_2 , and P_3 , those are three numbers, so this -- this height is P_1 , this height is P_2 , this height is P_3 , and what are those guys?

Suppose those were the three values instead of -- there's b_1 , ev- everybody's seen all these -- sorry, my art is as usual not the greatest, but there's the given b_1 , the given b_2 , and the given b_3 .

I promise not to put a single letter more on that picture.

OK. There's b_1 , P_1 is the one on the line, and e_1 is the distance

between. And same at points two and same at points three.

OK, so what's up?

What's up with those P s?

P1, P2, P3, what are they?

They're the components, they lie on the line,

right? They're the points which if instead of one, two, two, which were the b's, suppose I put P1, P2, P3 in here.

I'll figure out in a minute what those numbers are.

But I just want to get the picture of what I'm doing.

If I put P1, P2, P3 in those three equations, what would be good about the three equations?

I could solve them.

A line goes through the Ps.

So the P1, P2, P3 vector, that's in the column

space. That is a combination of these columns.

It's the closest combination.

It's this picture.

See, I've got the two pictures like here's the picture that shows the points, this is a picture in a blackboard plane, here's a picture that's showing the vectors.

The vector b, which is in this case, in this example is the vector one, two, two.

The column space is in this case spanned by the -- well, you see A there.

The column space of the matrix one, one, one, one, two, three.

And this picture shows the nearest point.

There's the -- that point P1, P2, P3, which I'm going to compute before the end of this hour, is the closest point in the column space.

OK. Let me -- t I don't dare leave it any longer -- can I just compute it now.

So I want to compute -- find P.

All right.

Find P.

Find x , which is CD, find P and P.

OK.

And I really should put these little hats on to remind myself that they're the estimated the best line, not the perfect line.

OK. OK.

How do I proceed?

Let's just run through the mechanics.

What's the equation for x ?

The -- or \hat{x} .

The equation for that is $A^T \hat{x} = A^T b$.

The most -- I'm -- will venture to call that the most important equation in statistics.

And in estimation.

And whatever you're -- wherever you've got error and noise this is the estimate that you use first.

OK. Whenever you're fitting things by a few parameters, that's the equation to use.

OK, let's solve it.

What is $A^T A$?

So I have to figure out what these matrices are.

One, one, one, one, two, three and one, one, one, one, two, three, that gives me some matrix, that gives me a matrix, what do I get out of that, three, six, six, and one and four and nine, fourteen.

OK.

And what do I expect to see in that matrix and I do see it, just before I keep going with the calculation?

I expect that matrix to be symmetric.

I expect it to be invertible.

And near the end of the course I'm going to say I expect it to be positive definite, but that's a future fact about this crucial matrix, $A^T A$.

OK.

And now let me figure $A^T b$.

So let me -- can I tack on b as an extra column here, one, two, two?

And tack on the extra $A^T b$ is -- looks like five and one and four and six, eleven.

I think my equations are three C plus six D equals five, and six D plus four-six C plus fourteen D is eleven.

Can I just for safety see if I did that right?

One, one, one times one, two, two is five.

One, two, three, that's one, four and six, eleven.

Looks good.

These are my equations.

That's my -- they're called the normal equations.

I'll just write that word down because it -- so I solve them.

I solve that for C and D .

I would like to -- before I solve them could I do one thing that's on the -- that's just above here?

I would like to -- I'd like to find these equations from calculus.

I'd like to find them from this minimizing thing.

So what's the first error?

The first error is what I missed by in the first equation.

C plus D minus one squared.

And the second error is what I miss in the second equation.

C plus two D minus two squared.

And the third error squared is C plus three D minus two squared.

That's my -- overall squared error that I'm trying to minimize.

OK. So how would you minimize that?

OK, linear algebra has given us the equations for the minimum.

But we could use calculus too.

That's a function of two variables, C and D, and we're looking for the minimum.

So how do we find it?

Directly from calculus, we take partial derivatives, right, we've got two variables, C and D, so take the partial derivative with respect to C and set it to zero, and you'll get that equation.

Take the partial derivative with respect -- I'm not going to write it all out, just -- you will.

The partial derivative with respect to D, it -- you know, it's going to be linear, that's the beauty of these squares, that if I have the square of something and I take its derivative I get something

And this is what I get. linear.

So this is the derivative of the error with respect to C being zero, and this is the derivative of the error with respect to D being zero.

Wherever you look, these equations keep coming.

So now I guess I'm going to solve it, what will I do, I'll subtract, I'll do elimination of course, because that's the only thing I know how to do.

Two of these away from this would give me -- let's see, six, so would that be two Ds equals one?

Ha. So it wasn't -- I was afraid these numbers were going to come out awful.

But if I take two of those away from that, the equation I get left is two D equals one, so I think D is a half and C is whatever back substitution gives, six D is three, so three C plus three is five, I'm doing back substitution now, right, three, can I do it in light letters, three C plus that six D is three equals five, so three C is two, so I think C is two-thirds.

One-half and two-thirds.

So the best line, the best line is the constant two-thirds plus one-half t.

And I -- is my picture more or less right?

Let me write, let me copy that best line down again, two-thirds and a half.

Let me -- I'll put in the two-thirds and the half.

OK.

So what's this P1, that's the value at t equal to one.

At t equal to one, I have two-thirds plus a half, which is -- what's that, four-sixths and three-sixths, so P1, oh, I promised not to write another thing on this -- I'll erase P1 and I'll put seven-sixths.

OK.

And yeah, it's above one, and e1 is one-sixth,

right. You see it all.

Right? What's P2?

OK.

At point t equal to two, where's my line here?

At t equal to two, it's two-thirds plus one, right?

That's five-thirds.

Two-thirds and t is two, so that's two-thirds and one make five-thirds.

And that's -- sure enough, that's smaller than the exact two.

And then final P3, when t is three, oh, what's two-thirds plus three-halves?

It's the same as three-halves plus two-thirds.

It's -- so maybe four-sixths and nine-sixths, maybe thirteen-sixths.

OK, and again, look, oh, look at this, OK.

You have to admire the beauty of this answer.

What's this first error?

So here are the errors. e_1 , e_2 and e_3 .

OK, what was that first error, e_1 ?

Well, if we decide the errors counting up, then it's one-sixth.

And the last error, thirteen-sixths minus the correct two is one-sixth again.

And what's this error in the middle?

Let's see, the correct answer was two, two.

And we got five-thirds and it's the other direction, minus one-third, minus two-sixths.

That's our error vector.

In our picture, in our other picture, here it is.

We just found P and e. e is this vector, one-sixth, minus two-sixths, one-sixth, and P is this guy.

Well, maybe I have the signs of e wrong, I think I have, let me fix it.

Because I would like this one-sixth -- I would like this plus the P to give the original b.

I want P plus e to match b.

So I want minus a sixth, plus seven-sixths to give the correct b equal one.

OK.

Now -- I'm going to take a deep breath here, and ask what do we know about this error vector e?

You've seen now this whole problem worked completely through, and I even think the numbers are right.

So there's P , so let me -- I'll write -- if I can put it down here, B is P plus e .

b I believe was one, two, two.

The nearest point had seven-sixths, what were the others?

Five-thirds and thirteen-sixths.

And the e vector was minus a sixth, two-sixths, one-third in other words, and minus a sixth.

OK. Tell me some stuff about these two vectors.

Tell me something about those two vectors, well, they add to b , right, great.

OK.

What else?

What else about those two vectors, the P , the projection vector P , and the error vector e .

What else do you know about them?

They're perpendicular, right.

Do we dare verify that?

Can you take the dot product of those vectors?

I'm like getting like minus seven over thirty-six, can I change that to ten-sixths?

Oh, God, come out right here.

Minus seven over thirty-six, plus twenty over thirty-six, minus thirteen over thirty-six.

Thank you, God.

OK.

And what else should we know about that vector?

Actually we know -- I've got to say we know even a little more.

This vector, e , is perpendicular to P , but it's perpendicular to other stuff too.

It's perpendicular not just to this guy in the column space, this is in the column space for sure.

This is perpendicular to the column space.

So like give me another vector it's perpendicular to.

Another because it's perpendicular to the whole column space, not just to this -- this particular projection that's -- that is in the column space, but it's perpendicular to other stuff, whatever's in the column space, so tell me another vector in the -- oh, well, I've written down the matrix, so tell me another vector in the column space.

Pick a nice one.

One, one, one.

That's what everybody's thinking.

OK, one, one, one is in the column space.

And this guy is supposed to be perpendicular to one,

one, one. Is it?

Sure. If I take the dot product with one, one, one I get minus a sixth, plus two-sixths, minus a sixth, zero.

And it's perpendicular to one, two, three.

Because if I take the dot product with one, two, three I get minus one, plus four, minus three, zero again.

OK, do you see the -- I hope you see the two pictures.

The picture here for vectors and, the picture here for the best line, and it's the same picture, just -- this one's in the plane and it's showing the line, this one never did show the line, this -- in this picture, C and D never showed up.

In this picture, C and D were -- you know, they determined that line.

But the two are exactly the same.

C and D is the combination of the two columns that gives P.

OK. So that's these squares.

And the special but most important example of fitting by straight line, so the homework that's coming then Wednesday asks you to fit by straight lines.

So you're just going to end up solving the key equation.

You're going to end up solving that key equation and then P will be Ax hat.

That's it.

OK. Now, can I put in a little piece of linear algebra that I mentioned earlier, mentioned again, but I never did write?

And I've -- I should do it right.

It's about this matrix A transpose A .

There.

I was sure that that matrix would be invertible.

And of course I wanted to be sure it was invertible, because I planned to solve this system with with that matrix.

So and I announced like before -- as the chapter was just starting, I announced that it would be invertible.

But now I -- can I come back to that?

OK. So what I said was -- that if A has independent columns, then A transpose A is invertible.

And I would like to -- first to repeat that important fact, that that's the requirement that makes everything go here.

It's this independent columns of A that guarantees everything goes through.

And think why.

Why does this matrix A transpose A , why is it invertible if the columns of A are independent?

OK, there's -- so if it wasn't invertible, I'm -- so I want to prove that.

If it isn't invertible, then what?

I want to reach -- I want to follow that -- follow that line -- of thinking and see what I come to.

Suppose, so proof.

Suppose $A^T x$ is zero.

I'm trying to prove this.

This is now to prove.

I don't like hammer away at too many proofs in this course.

But this is like the central fact and it brings in all the stuff we know.

OK.

So I'll start the proof.

Suppose $A^T x$ is zero.

What -- and I'm aiming to prove $A^T A$ is invertible.

So what do I want to prove now?

So I'm aiming to prove this fact.

I'll use this, and I'm aiming to prove that this matrix is invertible, OK, so if I suppose $A^T x$ is zero, then what conclusion do I want to reach?

I'd like to know that x must be zero.

I want to show x must be zero.

To show now -- to prove x must be the zero vector.

Is that right, that's what we worked in the previous chapter to understand, that a matrix was invertible when its null space is only the zero vector.

So that's what I want to show.

How come if $A^T x$ is zero, how come x must be zero?

What's going to be the reason?

Actually I have two ways to do it.

Let me show you one way.

This is -- here, trick.

Take the dot product of both sides with x .

So I'll multiply both sides by x transpose.

x transpose A transpose Ax equals zero.

I shouldn't have written trick.

That makes it sound like just a dumb idea.

Brilliant idea, I should have put.

OK.

I'll just put idea.

OK.

Now, I got to that equation, x transpose A transpose $Ax=0$, and I'm hoping you can see the right way to -- to look at that equation.

What can I conclude from that equation, that if I have x transpose A -- well, what is x transpose A transpose Ax ?

Does that -- what it's giving you?

It's again going to be putting in parentheses, I'm looking at Ax and what I seeing here?

Its transpose.

So I'm seeing here this is Ax transpose Ax .

Equaling zero.

Now if Ax transpose Ax , so like let's call it y or something, if y transpose y is zero, what does that tell me?

That the vector has to be zero, right?

This is the length squared, that's the length of the vector Ax squared, that's Ax times Ax .

So I conclude that Ax has to be zero.

Well, I'm getting somewhere.

Now that I know Ax is zero, now I'm going to use my little hypothesis.

Somewhere every mathematician has to use the hypothesis.

Right?

Now, if A has independent columns and we've -- we're at the point where Ax is zero, what does that tell us?

I could -- I mean that could be like a fill-in question on the final exam.

If A has independent columns and if Ax equals zero then what?

Please say it. x is zero, right.

Which was just what we wanted to prove.

That -- do you see why that is?

If Ax equals zero, now we're using -- here we used this was the square of something, so I'll put in little parentheses the observation we made, that was a square which is zero, so the thing has to be zero.

Now we're using the hypothesis of independent columns at the A has independent columns.

If A has independent columns, this is telling me x is in its null space, and the only thing in the null space of such a matrix is the zero vector.

OK.

So that's the argument and you see how it really used our understanding of the -- of the null space.

OK.

That's great.

All right.

So where are we then?

That board is like the backup theory that tells me that this matrix had to be invertible because these columns were independent.

OK. there's one case of independent -- there's one case where the geometry gets even better.

When the -- there's one case when columns are sure to be independent.

And let me put that -- let me write that down and that'll be the subject for next time.

Columns are sure -- are certainly independent, definitely independent, if they're perpendicular.

Oh, I've got to rule out the zero column, let me give them all length one, so they can't be zero if they are perpendicular unit vectors.

Like the vectors one, zero, zero, zero, one, zero and zero, zero, one.

Those vectors are unit vectors, they're perpendicular, and they certainly are independent.

And what's more, suppose they're -- oh, that's so nice, I mean what is $A^T A$ for that matrix?

For the matrix with these three columns?

It's the identity.

So here's the key to the lecture that's coming.

If we're dealing with perpendicular unit vectors and the word for that will be -- see I could have said orthogonal, but I said perpendicular -- and this unit vectors gets put in as the word normal.

Orthonormal vectors.

Those are the best columns you could ask for.

Matrices with -- whose columns are orthonormal, they're perpendicular to each other, and they're unit vectors, well, they don't have to be those three, let me do a final example over here, how about one at an angle like that and one at ninety degrees, that vector would be $\cos \theta$, $\sin \theta$, a unit vector, and this vector would be $-\sin \theta$, $\cos \theta$.

That is our absolute favorite pair of orthonormal vectors.

They're both unit vectors and they're perpendicular.

That angle is ninety degrees.

So like our job next time is first to see why orthonormal vectors are great, and then to make vectors orthonormal by picking the right basis.

OK, see you.

Thanks.