[SQUEAKING]

[RUSTLING]

[CLICKING]

PROFESSOR: All right, I want to welcome Stefan Andreev as guest lecture today. He has been a significant part of this course over the past, I think, 11 years. Is that right, Stefan?

STEFAN ANDREEV: A long period. And I can't count that far.

PROFESSOR: Partly because his lectures are the most popular in the course, and we really are glad to have him come back again. His background is very interesting. He has a PhD in Chemical Physics from Harvard. And he has worked at the top firms on Wall Street, Morgan Stanley, Citadel, and now to Sigma. So he's a really great person to share his expertise with us. So thank you.

STEFAN ANDREEV: Thank you. Thank you. Thank you, guys. Very nice to meet you all. Big class, big class this year. That's exciting. I'll start with this before everybody gets tired. I wanted to say that like I-- we're going to talk about principal component analysis today. And some of the main things I want you to take away from this are this is a absolutely core data science tool in finance.

That's number one. Number two, it's great to-- this kind of stuff, you only learn by doing it, so actually doing some kind of project where you use this toolkit is essential for you to really learn it. And three, if you ever want to get a job in finance, I think having done a project with this and having some familiarity with what this is about might help you during interviews quite a bit.

We're going to have lectures and slides. And we're going to have some-- the slides are on the web, on the classroom website. In addition to that, there is a notebook, and a Jupyter Notebook and a data set that I've used to prepare these slides, at least the first example we're going to talk about.

Like I think that's a great reference if you ever want to see how this stuff gets computed. And if you want to do like a project, final project in the class, it would be something like that notebook, but a little more elaborate. So it gives you a little bit of a flavor of what the starting point and what the project would be like, probably be a little more extensive than what we're talking about here, but like in the same spirit.

So yeah. Now that we got this over with, let's talk about what is PCA and where does it fit within the world of machine learning algorithms. And it's basically-- it's a method for unsupervised learning. So it's basically a method to cluster, reduce dimensionality to make sense of the data.

It's not a regression. There's no causality implied when you use PCA. Yeah, it's a way to understand data better, especially data that has a lot of dimensions, where many of those dimensions are highly correlated.

Here I've listed a bunch of examples of where PCA can be applied. Generally, it's widely applied in large data sets, multi-dimensional data, like a lot of data. The actual calculations are mostly linear algebra.

So it's efficient, and fast, and scalable. And this is one of the reasons people generally like to use it. I'm not going to go through all of these. Their applications vary widely-- they vary across science. And in finance, it's used very extensively as well.

So first thing I wanted to talk about is go through a visual example of-- some of these things that I want to talk about are like my mental toy models that I always refer to when I think about what is happening when I'm using PCA. In this example-- you guys see well? OK, good.

In this example, I've prepared effectively three data sets. But they're really like one data set. I've just made this up. There's just one. It's like one-dimensional data set.

And all I've done is I've taken these one-dimensional data set, and I've made it two dimensional by just adding an extra coordinate like the blue line. The extra coordinate y, I made 0. And then I took this data set and I just rotated it like 45 degrees and then 90 degrees.

And in this two-dimensional world, it's two sides. Each point has two coordinates. So it looks like it's two dimensional. But I know it's one dimensional because I made it. And the key thing that PCA would do is PCA would immediately tell you, when it looks at this data, that it's actually one dimensional. You download it. You do the calculation. It's going to be one dimensional.

So how would it work? You can think of PCA as a way to rotate the coordinate system. Basically when you apply PCA to this, it will tell you-- if you apply it to the blue data, it will tell you that the PC1, the first principal component-- Peter, you guys have done the math part already of PCA? Yeah.

So the first principal component is going to be the x direction. When you do the red line, it's going to tell you that it's like the 1, 1 direction. And it's going to just discover this direction dimensionality.

The first data set was very, very simple because it's purely one dimensional. This one here is the same thing, but I've added a little bit of noise in the other dimension. And now when I apply PCA to this, it's going to tell me basically the same thing. PC1 is going to be the same, but PC2 is going to have a little bit of variance along PC2, along the perpendicular direction.

So in both of these cases, A and B, it will tell me that PC1 and PC2 eigenvectors are the same. And what's going to be different is there's going to be 0 variance in case of B on the PC2 direction and a little bit of variance in example A, along the PC2 direction.

So it basically tells us-- the two things it tells us, the directions along which the data is concentrated and the variance or how important that direction is on a relative basis. What's nice about this-- like I've done this two-dimensional plots.

But PCA, like when you actually do it, it's just linear algebra. And you can do it on-- you can do it on any number of dimensions. You can do it on-- yeah. So when you have high dimensionality, this is a very robust and powerful tool.

What's really nice about that in the high dimensionality is like it's not a minimization. It's not an optimization. It's just a linear algebra calculation. And it's fast. So you can process very large amounts of data very quickly.

A very common-- like when you use this in practice, like the answers that you get from-- the things you can learn from PCA are going to be-- it tells you a relationship between variables. Like in our first example, if you look at this diagonal line, it's going to tell you that-- like y equals x is kind of the line.

And then you can think, well, I can discover the same thing if I regress y on x. And then it's going to tell me pretty much like the same-- if I look at my diagonal line, it's going to tell me the same thing. And that is true in the one-dimensional example. But I wanted to give an illustration of what is the difference between regression and PCA?

Remember, PCA is-- remember, it's unsupervised learning. So there's no causality implied. So x and y are kind of on the same footing. Each dimension is on the same footing. There's no special dimension that's dependent and one that is independent.

So when you regress, this is like the same data set. And you can see here the regression of y versus x, the regression of x versus y. And the PC1 line, they're similar. But they're not the same. Like it's actually optimizing-- it's minimizing the distance to the line.

This is minimizing the distance in the perpendicular direction. And the others are minimizing distance in the independent variable direction, so when you do a regression. So it's not exactly the same.

I think it's good to keep that in mind. Because sometimes if you really want to understand, given an observation x, how I should model y, maybe PCA is not the right approach. Maybe you should run a regression. But if you want to understand the correlation structure of x and y, of two variables, there is no-- you don't have necessarily one depends on the other. You don't know anything about their relationship. Then PCA is probably a more appropriate tool.

Yeah, so I was just saying, if there is some data that is in some-- if data is in a particular line, PC1, which is the red direction here, you will always discover where that direction is. And if you actually map the data into PC coordinates, it's just going to look like a straight line like that. And a couple of things that I like about this example is you can see how rotating the data set doesn't change the PC coordinate output. So it finds the direction appropriately.

The second thing that's very important is robustness. In the real world, when you do-- it could be in science or in finance. You're not going to have perfect data. Everything is very noisy. And noise is a huge problem. And robustness to data noise is very important. So this illustrates how robust things are.

So if I take one point and I just move it around, you can see that the red line, which is the direction of my PC1, doesn't-- it's quite robust. It's quite robust to noise, to one particular. If everything else is aligned, just moving around one data, it doesn't really change things too much. If I make them a little bit more spread out, it's a little less robust, but it's still pretty robust.

And the thing that I wanted to say-- what happened? People are just going to-- very often, you just take some data and you just shove it into PCA. And you don't actually care too much before you even know what the data is. And you just put it into PCA. And PCA always gives you something.

But suppose the data doesn't really have a dominant direction. Then the PCA algorithm becomes not robust. So you can see now. Look, I'm moving one data. And look what's happening to the coordinate.

One data point is-- I can just move that a little bit, and it's just completely giving me-- the output is super sensitive to any noise in the data. So it's worthless. So it's very important when we're looking at the data if we don't really know what this data looks like.

Sometimes it's high-dimensional. You can't even really visualize it. You have to think, how do I understand whether there's actual structure in the data that I'm trying to discover?

This will help me discover-- quantify the structure. But it's also good to understand qualitatively how much the structure is. One thing that PCA also-- part of the PCA analysis at the output-- yes?

AUDIENCE: Yeah, how many is [INAUDIBLE] making it around the parts to see how robust it is, to see if the data is actually close to the line or not?

STEFAN ANDREEV: Well, it's-- no, but what-- I don't have-- you can write tools to do that. But I think what you want-- one thing that you can do that is very important when you do your PCA analysis is to look at the eigenvalues. Because very often, people look at the eigenvectors. And they're like, oh, these are my directions. And I'm going to project things, and I'm going to assume everything is factorized nicely.

And we don't really look as much at the eigenvalues, but it's very important to look at the eigenvalues. And the reason is that in this example, when there's no dominant direction, the eigenvalues, which measure the variance along each direction, would be very similar. So when you look at-- when you're using a principal component like the first principal component, you want to make sure that the eigenvalue of that component you're using is quite distinct from the rest of the eigenvalues.

In finance, we'll see later we're dealing with the time series of PCA. Every day, you construct a PCA model. So anything that's a number here, the eigenvalue is actually a time series of eigenvalue.

So the thing I look at a lot when I'm doing analysis is I look at eigenvalue plots in log space. And I want to see that each distinct component that I use-- ideally, you want to see that it's an order of magnitude different than the next one in variance, basically, which if that's true, that's great. And that doesn't always happen.

But at least you want it to be-- visually, you want it to be distinct. You want it to be different. You don't want-- if you have a time series of things, you don't want the first and the second one to cross. And if that's happening, then the first and second principal component-- if the eigenvalues are small changes or causing them to switch places, it basically means there isn't a lot of structure. And the PCA output is noisy, and it's actually going to obfuscate rather than clarify the thing you're looking at.

I've put here a little bit of math notation. You guys have done your math lecture on PCA, so we can be relatively quick here. But I do want to go through it. The main thing is you have your data. Data matrix is generally-- there's n observations in p dimensions. And we think of that as a matrix.

And then if this matrix is x, then every day's observation are vectors of size p. And then you have a covariance matrix, which is XT times X. And then there is-- I've here defined the wk, or the orthonormal vectors representing the-- these are the eigenvectors. Wk. And then I've written down the math of how you do the projection of data onto principal components and how you compute the variance.

How do you get these vectors? Well, there is-- you could do-- you could take your data, compute covariance matrix, feed it into an eigenvalue calculator, which is basically-- in my work, our data is not really-- the number of dimensions is not actually that high. So that's pretty-- that's OK. It does scale, as that thing is potentially maybe a little bit slow, but it's not actually that slow.

There is also a more efficient method called SVD, Singular Value Decomposition, where you can just compute your projections without calculating the covariance matrix. I like to compute the covariance matrix, because we generally need to compute the covariance matrix for other things in any case. So if you have it already, then it's-- you might as well use it.

So there's some-- I've listed here some properties of this eigenvector matrix that are important, that it's orthonormal, which means that the transpose of the eigenvector matrix is also an inverse. The covariance matrices can be written as the-- in this form, where lambda is the eigenvalue matrix. That's the SVD decomposition.

There's total variance quantity you can think about, which is the variance of x, which is just the trace of lambda. So the product of eigenvalues is the variance of your data set. And there is something called percent variance explained, which is just defined as how much variance is along a particular PC direction divided by the total variance. So that adds up to 100% when you're-- but that's very useful because if you remember how I was saying want to look at the eigenvalues to understand how important each direction is, variance explained tells you that as a number.

You want to look at-- you want to-- when you're trying to model your data set with as few dimensions as possible, you want to add more PC directions until you get to a relatively high variance explained. If you're able to explain-- if you have a 10-dimensional data set, and you're able to explain 80% of it using three dimensions, that's a pretty big win. And I've also written out this out-of-sample variance explained. Or very often, what we do is we calculate a PCA. We calibrate it on historical data, and then we apply it to future data that we're trying to model what will happen in the future. Especially in finance we do that.

So when the data you're calibrating your PCA on is different than the other data, then this-- you hope that the structure of the two data sets are the same. So if the structure is the same, you'll get the same variance explained structure. But to the extent that it's not true, you may not.

So you may see that your dominance of certain PC1 factors becomes-- it starts to explain less variance as your out-of-sample data becomes more different from the data, which you calibrate. So this is a little bit of the modeling part of it, where going to be using some data to calibrate the model. And then you're going to use this to predict the variance, the correlation structure of what happens in the future, like future market moves or future returns.

So a typical-- what do we generally do when we apply PCA? We're going to take our data set. We're going to demean it. And then maybe we're going to normalize the variance to 1. When I say "maybe"-- and the reason this "maybe" is-- sometimes you want to do that. Sometimes you don't. Very often, whether you want to do that or not depends on how exactly are you going to use your PC factors and your dimensions.

Generally, I would say that most of the time, you want to normalize it in a certain-- you want to normalize it in a way that you want the expected-- at least the expected variance along different-- along your dimensions to be unit, to be similar. Then you're going to compute your covariance matrix and calculate the eigenvalues. And then you're going to order them by decreasing eigenvalue.

And then you're going to have a certain cutoff, where you say, only the top-- I'm going to use only the top x factors to model the of my data set. Yeah?

AUDIENCE: So you said maybe you would normalize it. What would be a concrete example of where it would be better to not normalize it?

STEFAN ANDREEV: Well, here's an example. So actually-- yes, I will start with this. So this is an example of-- it's a little bit of a toy example. And I'll talk about a real example too.

So in this toy example, there's the blue and the red data set. And the red one looks like it's one-dimensional. And the blue looks like it's just a blob. There's no direct dimensionality.

So if I tell you, should we use PCA? Is there a way-- is PCA going to work on the blue data set? Is it going to reduce dimensionality? And the answer would be no, because there is no dominant direction. And in the red one, it definitely will. It will pick the right direction.

Now, what's fun about these two examples is that these two are the same data, except in the blue data set, I've normalized the variance along x and y. And in the red one, I haven't. So if these two dimensions were temperature versus-- temperature in Seattle versus pressure in Atlanta or something, that are very distinct and generally different things, where you're only interested in correlation and the relative variance is not very meaningful, then you would want to normalize it. Because otherwise, the units are not going to be-- you want to have somewhat related units.

So the units will be standard deviations if you normalize it. If you don't, the units will be arbitrary. And they don't-- it wouldn't really make sense.

On the other hand, if this was-- if x-axis was your grades on quiz 1 and y-axis was your grades on quiz 2, these two are the same units now. And you might not want to normalize. You want to actually-- these grades are-- these numbers are meaningful relative to each other. So there's not necessarily-- you want to actually model the actual grades, including the variance of quiz 1 and quiz 2. So there is no necessarily need to normalize it.

AUDIENCE: That makes sense. Thank you.

STEFAN ANDREEV: Yeah, but that part is-- to be honest, that part is very much the art of things. We have a lot of discussion at work. Are we normalizing it? How are we normalizing it?

What are we using to normalize it? That's like-- surprisingly, this is-- sometimes cutting-edge research is to actually understand how you normalize it. These are the things that actually make a difference.

The other thing I mentioned is that we need to demean the data set. Why is that important? Because remember, the principal component analysis is a rotation. So it always rotates around (0, 0). So what 0 is, it matters a lot.

And this is two examples of the same data set. These two plots are the same data set. But one is not demeaned, and the other is demeaned.

So this one here is centered at 0. And PCA finds that there is a direction, a dominant direction. It finds it. It works like it's supposed to.

This one here is not centered at 0. For example, let's say you do a bunch of stock prices. You just put prices into PCA. Stock prices are not centered at 0. They're always positive.

So it's going to have some direction, but it's going to be somewhere far away from 0. When you run PCA on it, it's going to be-- it's going to try to rotate around 0, but it's not. It can't find anything because it's not-- you're not-- you can see why.

You can't rotate this thing. You can't. There's nothing to find if you don't center it at 0. You cannot rotate around. It's rotating always around the origin. So mean near 0 or close to 0 or exactly 0 is very important.

And surprisingly, it's surprising how often it comes up. Somebody just does PCA on some numbers that are just large, positive numbers. And nothing comes out. And they're like, it doesn't work. But you just have to normalize to 0.

How do we actually do it? Well, this is just the examples of various places that have implementations of these eigenvalue calculations. I think when I started doing this 11-- I don't know how many years ago, this actually was more relevant. Now, it's like everybody uses Python. Everybody knows Jupyter Notebooks. And you guys-- now it's standard toolkit.

Now, what's special about finance when it comes to PCA? Mostly, what is special about finance-- and this is not exclusive, by the way. But most of the time, we are going to be doing PCA on time series. In quantitative finance, in the Quant world that I live in and I work in, I would say, the general job is to discover patterns that have happened in the past and try to use that information to model the future and to somehow use that as part of your toolkit to make predictions and to place bets in the markets.

So every day, every trading session, every observation is an opportunity to bet and to make predictions. And you can understand the thing about time series is that the history now and the history on the next opportunities to bet is not that different. It has one extra data point. Later, it has an additional data point, but everything else is the same.

So there is a slow-changing time series of information. And how we handle that-- there is a whole time series data science class that one can take. But especially with regard to PCA, it's important to understand what impact does that time series aspect have on PCA. A couple of things, I think, are important.

We generally are going to be interested in modeling the changes of things, of variables, not the levels. Part of that is the thing I was just talking about, which is demeaning. And part of it is that when you make bets in the market, you make money based on how things change after your bet, not the price, the absolute level of where the price is at. So returns are what matters, not the price of something, the absolute price.

Then the next thing is, so when we do-- when we model-- when we have our data in finance, generally, the data is the historical returns or history of something. History of prices. History of returns. History of anything, of whatever you're modeling. It doesn't have to be just prices. It could be anything that-- any technical indicator.

And that means that every day, there is a history that is similar, but a little different. So the history changes. Your information about what has happened changes slowly.

And that means that when you're looking at your PCA, you have to now define, how am I going to define my data matrix, which I'm using to calculate my eigenvectors and eigenvalues? And one of the main choices are how long of a history I want to use and how much-- how do I weight history, more recent history versus further away history? And these are some of the main hyperparameters of principal component analysis in finance. How many things I'm including in my data set. How far back am I looking? How am I weighting those observations as I pass through time?

And then the last thing is the robustness of the data. Very often, there is a trade-off between including more data points. But each marginal addition of data-- each marginal dimension of data that I'm looking at could be less robust. And we'll talk about in the examples later on.

But there are some-- we have to think carefully about how robust the data we're putting in is. Because if it's too noisy, if it's not representative of where our bets are going to be, then it becomes less relevant to our modeling down the line. On the other hand, if we don't include enough, we're also going to be missing opportunities to understand the structure of the information. So again, that's part of the art of-- actually, of the practitioner, is to actually have some feeling, have some intuition about what I should include and what I should not include.

And we'll talk about an example here. The first example, and the main example I want to go through, is applying principal component analysis to the United States bond market.

Yes?

AUDIENCE: Well, earlier you mentioned that your time [INAUDIBLE], in this case, this time series was really, like, historical [INAUDIBLE]. Is there any play with the granularity of the time series, the different, I guess, window sizes, I would say, of how granular the data is going to be, whether it's second--

STEFAN ANDREEV: Yeah, that's a whole other dimension of how granular it is, absolutely right. It's not just about how long of a history, but how granular it is. Well, the examples I'm going to show you right now are based on daily observations. But why did I decide to use daily observations? The short answer is every-- by the way, very often, in my work, we use daily observations. The reason is that daily observations are just more available. Generally, more high-quality data history is available for longer periods of time for daily observations.

Second, yes, intraday observations are-- they represent additional data. However, the quality of the imagery that data, meaning how much noise there is, can very often, in practice, be much higher for intraday data than it is for end of day data. That may not always be true. There are situations where that's not true. And there's entire fields in finance that deal with high-frequency trading in those fields end of day data is just like their whole trading, buying, holding, and selling happens 10 times for the entire day. So day start and end of day with no positions. And for them, daily is not relevant. It's too slow. They need to understand how the market moves in a very fast timescale.

So how often how you sample time is a whole other story. And basically, how we sample time is going to impact the units of volatility because volatility is measured in units of square root of time, basically. So it's very important to know that and to understand, to keep that in mind, that whatever the frequency of sampling is, that's going to give you effectively the unit of your volatility. And then if you want to annualize it, you need to apply appropriate factors. And comparing volatilities of different units is not going to be meaningful. So that's important.

Going back to this example, the US bond market-- so why am I-- why is this an interesting example? I work in fixed income, which is basically where we trade bonds and various derivatives of bonds. And those are the biggest markets in the world. There's a huge amounts of money being transferred. They're not necessarily the most volatile. Equities are much more volatile. But they are extremely deep, extremely liquid. And they also allow people who trade in them to take large amounts of leverage.

So even though they're not by themselves that volatile, if once you take leverage, you can increase volatility by taking leverage. So you can actually make-- you can construct portfolios that are actually quite volatile and deliver substantial market volatility and returns. So there's a lot of huge industry of money managers, hedge funds that trade fixed income. And the US bond market is the prototypical example of fixed income. It's the most liquid in the world by far, the most well-structured. So it's a great example to play with.

Another reason I'm using it is because there is pricing data from the Federal Reserve that is freely available. And that's the data that we're going to use right now to do our analysis, this data is like, it's actually quite high-quality data. I think that if you want to do a project like this, would be a great data set to play with because it's real and it goes on for a very long time, and you can apply you can apply our PCA techniques to it.

So the first question to ask is-- remember, we have our data set now. What does it consist of? Well, in the United States, the US Treasury issues a bunch of T-bills, short maturity bills, that they raise money. That's how they raise money for the debt. But the part we're going to concern with is going to be the US Treasury notes and bonds which are kind of the longer-dated instruments that US Treasury issues.

They issue on a very regular basis-- every month, there is a schedule of auctions that is very well known and repetitive. And they always issue a 2-year bond, a 3-year bond, 5-year bond, 7-year bond, 10-year bond, and a 30-year bond. And there's actually now a 20-year bond that started about three or four years ago. But we're not going to look at that right now because it's very new and it's not as liquid. So we're just going to focus on these.

And the other thing about these bonds is that they're all debts of the US Treasury. So in a sense, they're very similar. If I give you money for six months and I also give you a separate loan for three years, I've given you two loans. The first thing that these two things are-- basically, you still owe me money. One is longer in maturity than the other, but you're still owing me money. So in terms of the way the price of these securities evolves, they all depend on the interest rates, the path of short-term interest rates of the Federal Reserve, and the expectations of that path. So they're all derivatives of that.

The implications of that is that the returns or the changes of bond yields are very highly correlated. So here, on the bottom right, I've shown the correlation matrix of changes. And you can see that it's like these numbers are very, very high. So this is a data set where everything is very, very highly correlated prime candidate for PCA. Whenever you see a correlation matrix, it has high concentrations of high correlations, the prime candidate of PCA.

So what are these numbers I'm looking at? On the left chart, I've shown the US yield curve. So on any given day, we can observe the prices of 2-year bonds, 3-year bonds, 5-year bonds, 7-year bonds, 10-year bonds, 30-year bonds. What is the price of that? And I don't know if-- Peter, have you guys done bond math yet?

PETER: A little bit bond math, a little.

STEFAN ANDREEV: A little. So we know the prices where these things buy and sell. They're very transparent prices. And we can convert prices to yields. So basically, a yield is like the implied rate of return of the bond, effectively, given its price over the course of its maturity. And then, when we plot these yields, on this plot here, I've plotted the yields. The yield is on the y-axis versus the maturity of the bond. And I've sampled five days from history. And this is what's known as the yield curve. That is the most important input of macroeconomics in the world is the US-- the shape of the dollar yield curve.

Most of the time, the yield curve is upward sloping, so people require a bigger return to lend money for longer. And the US Treasury generally requires that. But sometimes, that is not true. Sometimes, you can see this top blue line. Sometimes, the yields actually-- it starts out downward sloping and then flattens out. So these yields can have a variety of shapes. And this is just the level of yields.

What we're betting on is the changes. So we're betting on how this curve evolves day to day. And you can see from these yield curves that you can imagine how, if I were to show you a movie of the yield curve, it's going to flop around. It's going to be like little strings that flop around. It's going to be going up and down. It's going to slope, like this. It's going to have a whole set of dynamics. And that those dynamics is what we're trying to model with PCA here.

Bonus question is do you guys know what is the shape of the yield curve now? What does it look like? Are short-term interest rates higher than long-term interest rates? Has anybody-- we had a very prominent Federal Reserve meeting recently, and do you remember what they did?

The action the Fed took was-- the action the Fed took was to cut interest rates by 50 basis points. So what the Fed is doing-- it's cutting the interest rate that is the leftmost point on this yield curve that they're setting that point. Everything else, the rest of the curve, is market expectations of what that point will do in the future. So they moved this point down by 50 basis points. 50 basis points means half a percent. So it went down by half a percent.

Right now, this blue curve, even though it's from 2023, that's pretty recent. The curve is still downward sloping. It looks a little bit more like this right now. For a very long time, it was actually really inverted. It was going downward sloping, like that. Now the back is starting to flatten out, but it's still quite high in the front end and downward sloping. So unrelated to what we're talking about, but kind of a fun fact.

So now, when we look at your changes, when we look at the-- we take our data, we compute yield changes, so the changes of yields on day to day. We should demean that data set. But generally, demeaning doesn't really help that much because the mean is going to be very close to 0 because changes are sometimes going to go up. Sometimes, they're going to go down. Over the history of 20 or 30 years that we're talking about here, you're not going to have a large mean. Still, it's always good practice to demean it. It won't change, much, the result.

But in this example, I didn't actually normalize the variance. I think we just left it in there. So we did PCA on covariance matrix, and the reason is that the units, the yield changes are comparable units. And looking at the PCA without normalizing the variance is actually-- it's a very meaningful output. We want to understand how volatile. It's going to tell us something about the relative volatilities of points.

So when we do that and then we look at-- what I'm plotting here on the top are the eigenvectors, so the principal component factors, the shapes of the principal component factors on various days. So this plot can be confusing. And I want to make sure that we're all on the same page what it is. This is the yields. And we take these yields. We look at the changes of yields. And we put them through PCA. And remember, the output of PCA is eigenvectors and eigenvalues. Eigenvectors means every point, like 2-year, 3-year, 5-year, 10-year, 30-year. Everything has a number now, loading. We call that, the number, the eigenvector coefficient. I call that loading. That's a term that we use.

That loading is, again, there's a time series of those. And I'm plotting here select five different dates in history of what that eigenvector looks like-- what the first principal component looks like, what the second principal component looks like, what the third principal component looks like. These are snapshots, like five snapshots in time. And I'm plotting maturity versus loading. And there's actually three dimensions here. There's maturity, loading, and time.

So I can show you a three-dimensional plot, but that would just be like confusing. So I'm showing here cross-section. So these are various cross-sections on time. And you can see that there is a clear and persistent pattern in what the eigenvectors look like.

The first one, PC1, is called the level. That basically means that-- all the loadings are positive. So it means when the yield curves move, generally, they move up and down together in the same direction. So if 10-year rate is going up, probably 2-year rate is going up as well, 5-year rate is going up as well. And hence the term rates are going up. Rates are going up means, generally speaking, everything is going up or down. And that is the move that explains most of the variance of the yield curve.

But, on top of that, now there is a second move, which is we call this PC2. And when we look at the shape of it, we can say, oh it starts out negative on the front end of the curve and it's positive on the back end the curve. That's the slope. So that's kind of a steepening of the curve versus flattening of the curve. That is the second PC factor.

So this first PC factor, by the way, explains around like 90% of the variance, maybe 85 or something. And the second PC factor explains between 5% and 10% of the variance-- very huge difference. And the third factor is like the curvature, so how curvy-- what is the curvature of the-- it's positive on the front end, positive at the third year, and negative at the belly of the curve, 10-year point.

And that has the smallest amount of variance. So what does that mean? If you think mentally of the movie of what the yield curve is doing, it's mostly going up and down, mostly going up and down like that. But it's also slope, a little bit of slope. So it's mostly up and down, but a little bit of slope, and a little bit of curvature, like that. So that's what this thing tells us is the dynamics of the curve.

Now some of these things are-- the structure of this thing actually tells you a lot. You can see how the PC1 level, the maturity at which it-- it starts out steep, upward sloping, and then it levels out. The maturity in years at which it levels out-- I've plotted here what the maturity is, at which it kind of levels out, where the short-term rate decoupled from long-term rates.

And if you plot that in time series, you can see that it was very short. Monetary policy was basically only governing the short end of the curve up to two years up until right after the post-financial crisis. And then the Fed became very, very powerful post-financial crisis. The Fed said, we're going to keep rates low for a very long time. And all of a sudden, this thing exploded and became very, very, very high. So basically, the short end rates were governing, effectively, the rates up to four or five years.

And then, up to COVID, they started to decrease up until COVID. COVID happened, Fed came in and said, we're going to bring the bazooka. So everybody's like, OK, they're going to keep rates low for a long time as well. It exploded again, and now it's back down again. So that's already something that is kind of like a regime indicator. You can see how the market behavior is reflected in the shape of these PC factors. That tells you so much about what's going on in the market when you look in the details of these factors.

Something like where is the bottom of PC3? Where is the belly? You can see that, sometimes, it's at 10 years. Sometimes, it becomes close to 5. Sometimes, it's close to 15. That's also a regime indicator. PC loading stability. So here I'm taking this graph PC1 graph. But instead of plotting loading versus maturity, I'm plotting loading versus time. And each maturity point is line. So it tells me the stability of my loadings. Generally, you want to look at that because you want factors that are stable. And PC1 loadings are pretty stable.

But where the changes in PC1 loadings, very often, they're quite meaningful. In 2008 crisis, 2-year rate-- this blue line is a 2-year rate. The short end becomes very volatile. Nobody knows what the Fed is going to do. There's a lot of volatility. It then collapses post-2012. The Fed is not doing anything, so the 2-year point is not moving. Then it goes up. Maybe the Fed is going to move. Then, after COVID, it also collapses. And now it's back up again. The Fed is very active right now. This ends in 2023, this data set. If I make it a little bit longer, you see this line actually went very high in the recent history.

So these things are-- they all come from this data, but they tell you a lot about the behavior, the regime, and the dynamics of interest rate market.

This is the variance explained. I think this is a way-- remember how I was telling you how we look at eigenvalues and we want to make sure that they're well-segregated? Well, this is a way to look at it. If you look at this chart, and the first factor explains so much variance, it's like, yeah, that's a good factor. That's a good factor. That's a solid factor. And you know that this thing is persistent. This thing is likely to continue. It's likely to be stable. And it's a good part of your model for the future.

If you look at variance explained by a factor, this is a log plot. You see how the first and second and third factor are separated by order of magnitude, so they never cross? So that means that these factors are robust and they're very useful factors for modeling. You want to see this kind of picture. What you don't want to see is like lines that are all tangled up with each other. You don't want to see that. You want to see something like this in general, when you're looking at PCA. Again, these are eigenvalues I'm looking at-- variance explained.

And this out-of-sample covariance is when I'm-- this is basically when I calibrate my PCA, remember, if I take a single data set, however I choose it, and I do run PCA, the PC1 projections and the PC2 projections are going to be uncorrelated with each other by construction. But if I take the same loadings now and I look at the future returns, there is nothing that says that if I project my future data, future returns onto the same loadings, those might actually become correlated.

So here, I'm computing kind of future returns projected onto PC factors using data calibrated on past history. And you see that they're not fully orthogonal. However, they're quite orthogonal. So it means that the dynamics are mostly preserved. The PC factors are mostly stable. This is something you would expect, you would hope to see. You don't want to see these numbers, off-diagonal numbers, to become 30%-50%, that means that your PCA model is quite unstable. But if your out-of-sample correlation is pretty small, that's a good thing.

Now what do we do with this PCA? Well, one of the key things is-- I want to go back to this slide for one reason. Suppose you say, well, I want to bet on curve-- I want to bet on the second factor. I have a model that I want to bet on curves, depending on relative value of 10-year bonds versus 5-year bonds or 30-year bonds versus 10-year bonds.

I can use PCA to isolate effects that are idiosyncratic to particular points on the curve to particular factors. And I can basically say, well, if I hedge out the vast majority of the variance by hedging out PC1, and I look at the residual PC1 residuals, which is PC2 and above, what can I say about those returns? How can I model those?

And the key thing-- you can construct these things. You can say, oh, I want to model these things. But the key is, once you have a good model, how do you trade them? So being able to construct-- to understand how PC loadings translate to portfolios is very important because it tells you how I can use this methodology to construct specific portfolios that are uncorrelated to each other and give me a new kind of target to model.

If I didn't do that, I still have, like, eight things I can trade, or six things-- 2-year, 3-year, 5-year, 10-year, and 30-year bonds. But they're so correlated, where I'm trading 10-year bonds or 30-year bond, if I just trade 10-year bond outright or 30-year bond outright, I'm mostly getting the same return. So to be able to increase the diversity of the target I'm capturing, I'm using PCA to construct more portfolios that are going to be uncorrelated and present me with more opportunities to model, to make bets.

That's a key use case for PCA in finance. We take these things that are very highly correlated. And now I'm going to create, out of these six highly-correlated portfolios, I'm going to create three or four portfolios that are not highly correlated. These portfolios are going to be defined by my PCA factors. So remember the factors on any given day, they give me a loading on every tenure, on every maturity. Did you have a question? OK.

Yeah, they give me a loading on every maturity. So it tells me-- and I can think of this loading as telling me how much to own of this particular bond on that maturity. So PC1 means buy every bond in some equal amount. That's PC1. PC2 is like, oh, you should sell some to your bonds, buy some 30-year bonds. Basically, the portfolio is going to be-- the portfolio is going to be defined. You're going to basically buy bonds according to the loading on every point on PC2.

Why is this portfolio interesting? Well, it's interesting because PCA tells us that if I construct this portfolio like this, it would have no sensitivity to PC1. But PC1 is, like, 90% of my variance. So that means that if I construct the PC2 portfolio, I have effectively hedged out 90% of the variance of the individual bonds. And I've constructed something that's new, less volatile, but uncorrelated to my original outright tenure bond.

Same thing with PC3. I can construct these PC3 portfolios that are-- and by the way, these things actually trade in the market. They're called curves. PC2 things are-- they're curves. You buy 2-year, you sell 10-year, or vice versa. And then there's 5's, which are like 2's, 10's, and 30's, where you sell 10-year and you buy 2-year and 30-year. Yes?

AUDIENCE: Is there any application of the higher-order PCs or are they just [INAUDIBLE] up to the variance, you don't really bother with?

STEFAN ANDREEV: They are the most interesting thing because the PC1 stuff is easy to bet on. You just buy 10-year bond, buy and sell. If you can model that and predict it, that's awesome. But first, that's hard. But also, it's only one thing you can predict. What are you doing 10-year or 30-years? You're predicting only one thing, so you can make a model for that.

Being able to construct these portfolios, PC2 and 3, you're able to create, out of these bonds that are highly correlated, you're able to create something that is entirely different from the PC1-- uncorrelated. So now we have something else that you can model. And because of the market is-- so the market is liquid. It needs to be very liquid because, remember this, it doesn't explain a lot of variance. So if you construct this thing, it's not going to move much. But all that means is I just need to buy more of it. If I buy enough of it, even though a single unit might not move much, when I buy enough of it, it's still going to move enough. And that's the point of leverage.

So if I sell a bunch of 2-year bond, buy a bunch of 30-year bond, now if I buy the same notion of-- buy and sell the same notionals as the PC1, it's not going to, like you said, it's not going to move very much. Maybe it's going to move only one third as much as my PC1 portfolio.

But if I scale it up by a factor of 3, now my new portfolio, PC2, has the same daily standard deviation as my PC1 portfolio. And it's orthogonal. So you just need, basically-- so what did I need to do that I need to be able to buy three times as much? So it needs the market needs to be liquid. I need to have access to leverage.

You can do the same with PC3. And yes, in the big liquid markets, like US Treasury, there is enough access to leverage where you can actually do that. You will see a lot of times in the press-- it's a very hot topic now. The CFTC is trying to regulate how much leverage exists in the financial markets, especially in interest rates. Because of this, because of the fact that PC1 is so explains so much, people use massive amounts of leverage to gain access to, basically, enough volatility to make these bets on PC2 and 3.

And that can create-- it means you own large amount. You own long and short, massive amounts of certain bonds. That can be dangerous if it's systemic in a certain way, and it presents risk. But it's definitely doable because there is so much debt. There's so much debt in the world that there is enough out there where you can bet-- you can gain this leverage, basically. And it's very accessible to professional investors.

AUDIENCE: Thank you.

STEFAN ANDREEV: So yeah, actually, far from it being irrelevant, that is actually where, a lot of times, the money is made. But it is true that as you go, why can't we go to PC4 and 5? The amount of leverage you need to make this work becomes so high it becomes really expensive to trade. Yes?

AUDIENCE: This is usually [INAUDIBLE], as you said before, that this allows you where you can scale up to get the volatility that you need is only useful for the fixed income market because it's more liquid than if you're applying the same thing in equities, for example? It's not as good, but you could get-- if you were to buy that many of whatever security costs, too much of shift, or stuff like that?

STEFAN ANDREEV: Well, it turns out that it's not really true what you say, and the reason is the following. Yeah, the volumes and notional volumes of equity market may be a little bit smaller. They're still very large. And the variance explained of the first principal component equities is not nearly as high as it is in rates. So you don't need as much leverage to get-- the second principal component has got a lot of volatility, a lot more relative to the first, than in this example here.

Later in the slides, which I think we might not get to, there is also an example with equity indices in Europe-- the same concept there. And by the way, a lot of equity-- there is a whole field of hedge funds equities that are equity long, short, neutral. They are basically betting on equities without taking directional risk. So everybody is trying-- all these guys are hedging out PC1. They're taking leverage. It exists. It's just a little more expensive to trade, but it gives you opportunities that you wouldn't be able to access otherwise. Yes?

AUDIENCE: I'm curious for some products, especially ones that are more illiquid. They might trade one, rarely, and two, around specific events. The prices that you see are necessarily indicative of how the prices would have traded a day or two before that? And how do you think about how to, I guess, clean data, in a way, or understand that there's a lot of idiosyncratic factors that go on with fixing the price, especially ones that are more liquid?

STEFAN ANDREEV: Yeah, it's a great question. There's a lot of art that goes-- a lot of domain knowledge that goes into that. In this example with bonds, yeah, there are certainly events where the market is very volatile. And when we look at daily returns, if I look at 10 years of daily returns and I'm going to estimate a standard deviation, and I'm going to say, well, on average, yields move by, I don't know, 5 to 6 basis points per day, something which is-- that's roughly how much they move.

And then if I'm sitting there and tomorrow is like-- there was this example in August that there was a huge-- there was a huge number that comes out, a huge print that was related to the inflation. And on that day, if you look at the price of the day before versus the price after the announcement, the volatility that people expected on this day was probably twice as high. So people expected the standard deviation of daily move that was priced in was probably 10 to 12 basis points per day.

Does that mean that all this is irrelevant? No, it's not. It just means that when you're thinking about the risk you're taking, you have to understand that if you take a certain amount of-- for the same number of bonds or the same number of the same investment that you hold, if you hold it on that kind of a day, you would just have a lot more risk. So when you're doing your risk management, you have to account for that. Do you want to do that? Do you want to hold it? Do you want to maybe shrink your position? Or if you have a view, maybe you don't want to shrink your position.

But you have to just be aware of it. The correlations are very likely to continue to hold. Sometimes, things happen. COVID happens. For example, COVID hit. These loadings, the dynamics of the curve, changed overnight. The PC loadings, the way the curve started to behave-- the central bank came, and there was an emergency cut, and things changed super quickly.

Being able to reflect that in your model is very important because now, if you're trying to take a PC2 bet, but your PC loadings are not indicative of how the current dynamics are, you're going to be ending up taking massive leverage. But you're not really PC1 hedged, so all this leverage is going to produce huge amounts of returns and noise in your P&L. So these are the practical questions that you face when you're actually doing this thing.

And remember, you need history to train your model. So you're sitting there. COVID happened. It was a big cut, emergency cut from the central bank. They cut, like, 1% in one day. And you know it's a new world. But there is no data on how the new world is going to behave because it just happened two days ago. You don't have 10 years of history to calibrate your PCA model. So what are you going to do? And there's a lot of answers to that. You can look at implied markets. There's a lot of answers to that. But how you answer that question, that's critical. That's critical, yeah.

AUDIENCE: Can I follow up to that?

STEFAN ANDREEV: Yeah.

AUDIENCE: Do you think it's worth, in some sense, almost slicing your data into pieces where you think things will behave in similar fashions, and then, like, you say something, a similar model, to better understand how they behave in those specific regimes or even times?

STEFAN ANDREEV: Yeah, that's certainly a thing that that is a thing that you can do, like regime-specific-- you can group together data that are specific to this regime. Oh, I'm going to look at, right now, the Fed is cutting rates. I'm going to model my curve behavior not by looking at the last five years of history, but I'm going to look at joined together historical periods where the Fed was cutting rates and just look at that. It's an approach, and it's a sensible approach.

Now, when you do that, you're introducing other-- yes, the Fed is cutting rates, but there are other things that were not the same. So there's no perfect answer in that because, again, you're still trying to use the past to predict the future. And there is no perfect answer. But there are many valid answers.

This is a description of how you construct these PC portfolios and PC residuals. We have some of the pros and cons. I already talked about some of these. And I kind of wanted to talk about the output of this. So here, I've created a little trading strategy based on the returns of-- basically, a momentum mean reversion strategy for PC1 and looked at the performance. It doesn't really work great, but that's not the point. The point is that it's an illustration of how this works.

All the code of how this was created and how you could calculate these things is part of that Jupyter Notebook that is accompanying the slides that you can download. And the data is there in a CSV file, which you can pull from the Fed-- an updated version of.

So if you are interested in doing a project, there's a lot of standard strategies that you can try, for example, to write and to test out. Again, the point is not to actually go and trade it, but to actually practice a little bit, applying PCA to real data. And there's no substitute to actually doing it because you see that, in every single step, there's choices to make. There's little details.

And these details matter and being able to understand the pros and cons of each of these details. One lecture is not enough to go over everything. I've wanted to give you a little taste of what it's like to deal with this stuff, but it's a very powerful tool. It's used all around the street for modeling of things like stocks and bonds. And it's a great thing to have in your toolkit as somebody who is looking for a job in finance, if you decide to do that.

So I'm going to stop here. Down here, there is another example, which is European stock market indices. Same thing-- tells you about the correlation, what the various principal components are, what they mean, and tells you about the strategy. So there's two examples in the slides. We don't need to go through both of them.

I know that, in five minutes, you have to leave. So I wanted to give a little bit of time for questions and for discussion. Thank you so much for being so active and interactive, rather. You guys have. You guys have asked more questions in this class than-- it's like definitely one of the top question-asking classes in my experience doing this class. So I love that. That shows that the stuff is interesting, so I'm very happy to hear that.

So yeah, I definitely wanted to ask if you have any questions. And in particular, if this is clear, I really want to make sure that this stuff is clear-- these top three plots, what they represent. Yeah?

AUDIENCE: Is there any chance you could explain the bottom left one?

STEFAN ANDREEV: Yeah. So the bottom left one, basically. So if there's two PC1 shapes, this is date 1, and there is one that is date 2. I kind of said, well, what I'm looking for is basically the-- I'm looking for the maturity point where the loadings cross a certain threshold because, at the end, both of these are going to be kind of relatively flat towards the back end. So basically, 10-year bonds and 30-year bonds, when the rates move up and down, they're going to move the same amount.

But on the very front end, that's not really true. And why is that? Well, on the very front end, the rates are governed by the central bank. So central bank is saying, we're going to set the rates here. Expectations can vary for what happens 10 years down the line. But expectations are what's going to happen over the next couple of months are not going to vary very much because that is basically-- the Fed sets that. The Fed tells you what is going to happen. So the loading of this PC1-- that front point is much less volatile in general.

That can change. It could also be that in the recent-- in the last couple of months, the actual shape of the PC1 factor was probably like that. And the reason is that the front end was so uncertain. Nobody knew. Are they going to cut? Are they going to hike? We don't really know what's going to happen. But generally speaking, the front end is going to be less volatile than the back end because of the fact that the Federal Reserve-- they control the front end part of the curve.

So what I'm plotting is, effectively, at what point does the loadings cross a threshold? And if the Fed controls the curve, if people believe the Fed is going to have-- if there's clarity on the Fed's policy over a longer period of time, we're going to have lower loadings on the front of the curve for longer periods of time. And if there is less clarity, that period is going to shrink, meaning, yeah, they're always going to control the front end. But how quickly do we decouple from that?

And that's what I'm plotting here. On every day, I'm taking this point. This is on day one. It's going to be here. On some other day, it's going to be here. And I'm plotting a time series of that point. At what time in maturity years, at which maturity do we cross the threshold?

AUDIENCE: Can you explain the connection between the curvature and the macroeconomics situation there?

STEFAN ANDREEV: The curvature and what?

AUDIENCE: The macroeconomics, like.

STEFAN ANDREEV: Ah, yeah. So the PC3 shape, remember, was like-- this is 0. And it's positive, negative, positive. So the short end and the back end are-- basically, once you hedge out, the curvature and the levels, what remains is-- once you hedge out the slope and levels, what remains is the curvature.

And the question is, at that point, how does the curve move? And I can give you two examples. In dollars, you can see that most of the time, the bottom of this graph is at around 10-year maturity. So that means that once you hedge out slope and curvature, 10-year bonds are going to be anticorrelated with 2-year and 30-year bonds. So it's going to move around like that, with where the belly of the curve is 10-year point.

Now if I look at another country that is more like an emerging markets country, for example, like Mexico, it's going to look like this, where this is going to be around five years. What is the significance of this? Well, it tells you-- you can think of this as the curve has distinct parts. Different things drive it.

The very front end, we say, is the central bank. Then there's a period after that is really mostly driven by inflation expectations. And then the very long end is a lot of it is about growth expectations. And the point where this bottom of the curvature is, it really tells you a lot about what part of the curve is-- where does this crossover go from inflation being dominant to growth being dominant? Different players invest in different parts of the curve.

The back end of the curve, there's a lot of pensions, insurance companies investing in the very long end. They have long-term liabilities. They invest here. Very front end is a lot of 2-year rates, a lot of speculators, a lot of people who need short-term financing. 10-year in the US, between 5 and 10 years is where most corporate bonds most corporations they issue, they raise debt in the 5 to 10-year maturity. So different players participate in the different parts of the curve.

And that means the dynamics are different. Once you hedge out those two components, you start to discover those-- these dynamics are small, relatively speaking, on the scale of rates going up and down. But once you remove that, you start to discover these features. They're kind of hidden under this big noise of rates going up and down. And you start to see these things start to pop out.

This being longer term, it tells you, again, something about-- developed markets tend to have this belly be further out because people look at a longer-term horizon, essentially, for investing. The pension funds invest for longer. They buy longer-term bonds. And generally, people invest in longer term than in, say, emerging market countries.

So when you compare these factors across countries, you can actually make sense of some of these things. And you can quantify those things, and they can become features for modeling.

AUDIENCE: One more question, perhaps?

STEFAN ANDREEV: Yeah.

AUDIENCE: For a lot of these markets, you're not going to get prices equally for something like US Treasury. Is there a way in which by-- trying to ask for transactional prices, do a synthetic [INAUDIBLE] with data [INAUDIBLE] would have?

STEFAN ANDREEV: Yeah, getting access to good data is step one in every quantitative finance project. I would say that that may be true, but there's ways to get it. First of all, there's bond futures, which are effectively futures on bonds. Those trade on exchanges. The data also tends to be pretty high quality and readily available. So there's ways to get it. But yeah, for every particular market you look at, there's going to be certain data quality issues you have to overcome. US market is the easiest for that.

Now it's part of, for example, that's why for example, if you're going to do your project, you're not going to look at South African rates because it's just hard to get good data. But if you're in a hedge fund, maybe, oh, it's worth getting the data and putting some effort because maybe you have less competition there. You can maybe make some money. Or maybe you have some ideas where you could model it, where you can actually have some predictive power in South African rates. So you make that investment of time.

But it's all part of that. The harder it is to get the data usually means that there's probably going to be more opportunities for quantitative modeling because it's hard for everybody. If it's easy, then everybody can do it. So you always have to be on the edge of the wave to make money.

PETER: Great. Thank you so much.

STEFAN ANDREEV: You're welcome.