

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: We introduced the data last time. These were some macroeconomic variables that can be used for forecasting the economy in terms of growth and factors such as inflation or unemployment. The case note goes through analyzing just three of these economic time series-- the unemployment rate, the federal funds rate, and a measure of the CPI, or Consumer Price Index.

When one fits vector autoregression model to this data, it turns out that the roots of the characteristic polynomial are 1.002, then 0.9863. And you recall when our discussion of vector autoregressive models, there's a characteristic equation sort of in matrix form, the determinant is just like the univariate autoregressive case. And in order for the process to be invertible, basically, the roots of the characteristic polynomial need to be less than 1 in magnitude.

In this implementation of the vector autoregression model, the characteristic roots are the inverses of the characteristic roots that we've been discussing. So anyway, this particular fit of the vector autoregression model suggests that the process is non-stationary. And so one should be considering different series to model this as a stationary time series.

But in terms of interpreting the regression model, one can see-- to accommodate the non-stationarity, we can take differences of all the series and fit the vector autoregressive to the difference series. So one way of eliminating any non-stationarity and time series models, basically and eliminate the random walk aspect of the processes is to be modeling first differences.

And so doing that with this series-- let's see. Here is just a graph of the time series properties of the different series. So with our original series, we take differences and

eliminate missing values in this r code. And this autocorrelation function shows us basically the correlations and auto correlations of individual series and the cross correlations across the different series.

So along the diagonals are the autocorrelation function. And one can see that every series is correlation one with itself. But then at the first lag positive for the Fed funds and the CPI measured. And there's also some cross correlations that are strong. And whether or not a correlation is strong or not depends upon how much uncertainty there is in our estimate of the correlation.

And these dashed lines here correspond to plus or minus two standard deviations of the correlation coefficient when the correlation coefficient is equal to 0. So any correlations that sort of go beyond those bounds is statistically significant. The partial autocorrelation function is graphed here.

And let's say our time series problem set goes through some discussion of the partial autoregressive coefficients and the interpretation of those. The partial autoregressive coefficients are the correlation between one variable and the lag of another after explaining for all lower degree lags. So it's like the incremental correlation of a variable with a lag term that exists.

And so if we actually fitting regression models where we include extra lags of a given variable, that partial lot of correlation coefficient is essentially the correlation associated with the addition of the final lagged variable. So here, we can see that each of these series is quite strongly correlated with itself. But there are also some cross correlations with, like, the unemployment rate and the Fed funds rate.

Basically, the Fed funds rate tends to go down when the unemployment rate goes up. And so this data is indicating the association between these macroeconomic variables and the evidence of that behavior. In terms of modeling the actual structural relations between these, we need several, up to about 10 or 12 variables more than these three. And then one can have a better understanding of the drivers of various macroeconomic features.

But this sort of illustrates the use of these methods with this reduced variable case. Let me also go down here and just comment on the unemployment rate or the Fed funds rate. When fitting these vector autoregressive models with the packages that exist in R, they give us output which provides the specification of each of the autoregressive models for the different dependent variables, the different series of the process.

And so here is the case of the regression model for Fed funds as a function of unemployment rate lagged, Fed funds rate lagged, and CPI lagged. These are all on the different scale.

When you're looking at these results, what's important is basically how strong the signal to noise ratio is for estimating these autoregressive parameters, vector autoregressive parameters. And so with the Fed funds, you can look at the t values. And t values that are larger than 2 are certainly quite significant. And you can see that basically when the unemployment rate coefficient is a negative 0.71, so if the unemployment rate goes up, we expect to see the Fed rate going down the next month.

And the Fed funds rate for the lag 1 has a t value of 7.97. So these are now models on the differences. So if the Fed funds rate was increased last month or last quarter, it's likely to be increased again. And that's partly a factor of how slow the economy is in reacting to changes and how the Fed doesn't want to shock the economy with large changes in their policy rates.

Another thing to notice here is that there's actually a negative coefficient on the lag 2 Fed funds term, a negative 0.17. And in interpreting these kinds of models, I think it's helpful just to think of, if you have Fed funds sub t, that's equal to minus 0.71 times the unemployment rate. of t minus 1.

And then we have plus 0.37 times the Fed funds, so t times 1. And this is delta. And then minus 1.8 times the Fed funds. So t minus 2.

In interpreting these coefficients, notice that these two terms correspond to 0.19

times the Fed funds change 1 lag ago plus 0.18 times the change in that rate. So when you see multiple lags coming into play in these models, the interpretation of them can be made by considering different transformations essentially at the underlying variables.

In this form, you can see that OK, the Fed funds tends to change the way it changed the previous month. But it also may change depending on the double change in the previous month. So there's a degree of acceleration in the Fed funds that is being captured here. So the interpretation of these models sometimes requires some care.

This kind of analysis, I find it quite useful. So let's push on to the next topic. So today's topics are going to begin with a discussion of cointegration. Cointegration is a major topic in time series analysis, which is dealing with the analysis of non stationary time series.

And in the previous discussion, we addressed non-stationarity of series by taking first differences to eliminate that non-stationarity. But we may be losing some information with that differencing. And cointegration provides a framework within which we characterize all available information for statistical modeling, in a very systematic way.

So let's introduce the context within which cointegration is relevant. It's relevant when we have a stochastic process, a multivariate stochastic process, which is integrated of some order, d . And to be integrated to order d means that if we take the d 'th difference, then that d 'th difference is stationary.

So and if you look at a time series and you plot that over time, well, OK, a stationary time series we know should be something that basically has a constant mean over time. There's some steady mean which that has. And the variability is also constant.

With some other time series, it might increase linearly over time. And a series that increases linearly over time, well, if you take first differences, that tends to take out that linear trend. If there are higher order differencing is required, then that means

that there's some curvature quadratic, say, that may exist in the data that is being taken out.

So this differencing is required to result in stationarity. If the process does have vector autoregressive representation in spite of its non-stationarity, then it can be represented by a polynomial lag of the x 's is equal to white noise ϵ . And the polynomial ϕ of L going to have a factor term in there of $1 - L$, basically the first difference to the d power.

So if taking these the d 'th order difference reduces it to stationarity, then we can express this vector autoregressive in this way. So the ϕ star of L basically represents the stationary vector autoregressive process on the d 'th different series.

Now, as it says here, each of the component series may be non stationary and integrated, say, of order one. But the process itself may not be jointly integrated. And that it may be that there are linear combinations of our multivariate series, which are stationary. And so these linear combinations basically represent the stationary features of the process. And those features can be apparent without looking at differences.

So in a sense, if you just focused on differences of these non stationary multivariate series, you would be losing out on information of the stationary structure of contemporaneous components of the multivariate series. And so cointegration deals with this situation where some linear combinations of the multivariate series in fact are stationary.

So how do we represent that mathematically? Well, we say that this multivariate time series process is cointegrated if there exists an m vector β such that defining linear weights on the x 's and $\beta' x_t$ is a stationary process. The cointegration vector of β can be scaled arbitrarily.

So it's common practice, if one has an interest, some primary perhaps, in the first component series of process to set that equal to 1. And the expression basically says that our time t value of the first series is related in a stationary way to a linear

combination of the other m minus 1 series.

And this is a long run equilibrium type relationship. How does this arise? Well, it arises in many, many ways in economics and finance. The term structure of interest rates, purchase power parity. In the term structure of interest rates, basically the differences between yields on interest rates over different maturities, those differences might be stationary. The overall level of interest might not be stationary, but the spreads ought to be stationary.

The purchase power parity in foreign exchange, if you look at the value of currencies for different countries, basically different countries ought to be able to purchase the same goods for roughly the same price. And so if there are disparities in currency values, purchase power parity suggests that things will revert back to some norm where everybody is pay on average over time the same amount for different goods. Otherwise, there would be arbitrage.

Money demand, covered interest rate parity, law of one price, spot and futures. Let me show you another example that will be in the case study for this chapter. View, full screen. Let's think about energy futures. In fact, next Tuesday's talk from Morgan Stanley is going to be an expert in commodity futures and options. And that should be very interesting.

Anyway, here, I'm looking at energy futures from the Energy Information Administration. Actually, for this course, trying to get data that's freely available to students is was one of the things we do. So this data is actually available from the Energy Information Administration of the government, which is now open, so I guess that'll be updated over time.

But basically these energy futures are traded on the Chicago Mercantile Exchange. And basically CL is crude, West Texas intermediate crude, light crude, which we have here, a time series from 2006 to basically yesterday. And you can see how at the start of the period around \$60 and then went up to close to \$140, and then it dropped down to around \$40. And it's been hovering around \$100 lately.

The second series here is gasoline, RBOB gasoline. Always have to look this up. This is that reformulated blend stock for oxygenated blending gasoline. Anyway, futures on this product are traded at the CME as well. And then heating oil. And what's happening with these data is that we have basically a refinery which processes crude oil as an input. And it basically refines it, distills it, and generates outputs, which include heating oil, gasoline, and various other things like jet fuel and others.

So if we're looking at the prices, the futures prices of, say, gasoline and heating oil, relating those to crude oil, well, certainly, the cost of producing these products should depend on the cost of the input. So I've got in the next plot, a translation of these futures contracts into their price per barrel. Turns out crude is quoted in dollars per barrel. And the gasoline heating oil are in cents per gallon.

So one multiplies. There are 42 gallons in a barrel. So you multiply those previous years by 42. And this shows the plot of the prices of the futures where we're looking at essentially the same units of output relative to input. And what's evident here is that while the futures for gasoline, the blue, is consistently above the green, the input, and same for heating oil.

And those vary depending on which is greater. So if we look at the difference between, say, the price of the heating oil future and the crude oil future, what does that represent? That's the spread in value of the output minus the input. Ray?

AUDIENCE: [INAUDIBLE] cost of running the refinery?

PROFESSOR: So cost of refining. So let's look at, say, heating oil minus CL and, say, this RBOB minus CL. So it's cost of refining. What else could be a factor here?

AUDIENCE: Supply and demand characteristics [INAUDIBLE].

PROFESSOR: Definitely. Supply and demand. If one product is demanded a lot more than another. Supply and demand. Anything else?

AUDIENCE: Maybe for the outputs, if you were to find the difference between the outputs, it

would be something cyclical. For example, in the winter, heating oil is going to get far more valuable as gasoline, because people drive less and people demand more for heating homes.

PROFESSOR: Absolutely. That's a very significant factor with these. There are seasonal effects that drive supply and demand. And so we can put seasonal effects in there as affecting supply and demand. But certainly, you might expect to see seasonal structure here. Anything else?

Put on your traders hat. Profit, yes. The refinery needs to make some profit. So there has to be some level of profit that's acceptable and appropriate.

So we have all these things driving basically these differences. Let's just take a look at those differences. These are actually called the crack spreads. Cracking in the business of refining is basically the breaking down of oil into components, products. And on the top is the gasoline crack spread. And the bottom is the heating oil crack spread.

And one can see that as time series, these actually look stationary. There certainly doesn't appear to be a linear trend up. But there are, of course, many factors that could affect this. So with that as motivation, how would we model such a series? So let's go back to our lecture here.

All right, View, full size. This is going to be a very technical discussion, but it's, at the end of the day, I think fairly straightforward. And the objective actually of this lecture is to provide an introduction to the notation here, which should make it seem like it's a very straightforward derivation process of these models.

So let's begin with just a recap of the vector autoregressive model of order p . This is the extension of the univariate case where we have a vector C of constants, m constants, and matrices ϕ_1 to ϕ_p corresponding to basically how the autoregression of one series depends on all the other series. And then there's multivariate white noise η_t , which has mean 0 and some covariate structure in it.

And the stationarity-- if this series were stationary, then the determinant of this

matrix polynomial would have roots outside the unit circle for complexity. And if it's not stationary, then some of those roots will be on the unit circle or beyond. So let's actually go to that non stationary case and suppose that the process is integrated of order one.

So if we were to take first differences, we would have stationarity. Well, the derivation of the model proceeds by converting the original vector autoregressive equation into an equation that's mostly relating to differences but with also some extra terms. So let's begin the process by just subtracting the lagged value of the multi-variant vector from the original series.

So we subtract x_{t-1} from both sides, and we get Δx_t is equal to C plus ϕ_1 minus ϕ_1 times x_{t-1} plus the rest. So that's a very simple step. We're just subtracting the lagged multivariate series from both sides.

Now, what we want to do is convert the second term in the middle line into a difference term. So what do we do? Well, we can subtract and add ϕ_1 minus ϕ_1 times x_{t-2} . If we do that, subtract and add that, we then get the Δx_t is C plus a multiple of Δx_{t-1} plus this multiple of x_{t-2} .

So we basically reduced the equations to differences in the first two terms or in the current series and the lagged. But then we have the original series for lags $t-2$. We can continue this process with the third. And then at the end of the day, we end up getting this equation for the difference of the series is equal to a constant plus a matrix multiple of the first difference multivariate series, plus another matrix times the second difference, all the way down to the p 'th difference, or the p minus first difference.

But at the end, we're left with terms at p lags that have no differences in them. So we've been able to represent this series as an autoregressive function of differences. But there's also a term on the undifferenced series at the end that's left over.

And or this argument can actually proceed by eliminating differences in the reverse

way, starting with the p 'th lag and going up. And one then can represent this as Δx_t of C plus some matrix times just the lagged series plus various matrices times the differences going back p minus 1 lags.

And so at the end of the day, this model basically for Δx_t is a constant plus a matrix times the previous lagged series or the first lag of the multivariate time series, plus various autoregressive lags of the differenced series. So these notes give you the formulas for those, and they're very easy to verify if you go through them one by one.

And when we look at this expression for the model, this expresses the stochastic process model for the difference series. This difference series is stationary. We've eliminated the non-stationarity in the process. So that means the right hand side has to be stationary as well.

And so while the terms which are matrix multiples of lags of the differenced series, those are going to be stationary because we're just taking lags of the stationary multivariate time series, the difference series. But this πx_t term has to be stationary as well. So this πx_t contains the cointegrating terms.

And fitting a sort of cointegrated vector autoregression model involves identifying this term, πx_t . And given that the original series had unit roots, it has to be the case that π , the matrix, is singular. So it's basically a transformation of the data that eliminates that unit root in the overall series.

So the matrix π is of reduced rank, and it's either rank zero, in which case there's no cointegrating relationships, or its rank is less than m . And the matrix π does define the cointegrating relationships. Now, these cointegrating relationships are the relationships in the process that are stationary.

And so basically there's a lot of information in that multivariate series with contemporaneous values of the series. There is stationary structure at every single time point, which can be the target of the modeling. So this matrix π is of rank r less than m . And so it can be expressed as basically $\alpha \beta'$, where these

matrices of rank r α and β .

And the columns of β define linearly independent vectors which cointegrate x . And the decomposition of π isn't unique. You can basically, for any invertible r by r matrix, g define another set of cointegrating relationships. So in the linear algebra structure of these problems, there's basically an r dimensional space where the process is stationary, and how you define the coordinate system in that space is up to you or subject to some choice.

So how do we estimate these models? Well, rather nice result of Sims, Stock, and Watson. Actually, Sims, Christopher Sims, he got the Nobel Prize a few years ago for his work in econometrics. And so this is a rather significant work that he did.

Anyway, he, together with Stock and Watson, prove that if you're estimating a vector autoregression model, then the least squares estimator of the original model is basically sufficient to do an analysis of this cointegrated vector autoregression process.

The parameter estimates from just fitting the vector autoregression are consistent for the underlying parameters. And they have asymptotic distributions that are identical to those of maximum likelihood estimators. And so what ends up happening is the least squares estimates of the vector autoregression parameters lead to an estimation of the π matrix. And the constraints on the π matrix which are basically π is of reduced rank, those will hold asymptotically.

So let's just go back to the equation before, to see if that looks familiar here. So what that work says is that if we basically fit the linear regression model regressing the difference series on the lag of the series plus lags of differences, the least squares estimates of these underlying parameters will give us asymptotically efficient estimates of this overall process.

So we don't need to use any new tools to specify these models. There's an advanced literature on estimation methods for these models. Johansen does describe maximum likelihood estimation when the innovation terms are normally

distributed. And that methodology applies reduced rank regression methodology and yields tests for what the rank is of the cointegrating relationship.

And these methods are implemented in our packages. Let's see. Let me just go back now to the-- so let's see. The case study on the crack spread data actually goes through sort of testing for non-stationarity in these underlying series. And actually, why don't I just show you that? Let's go back here.

If you can see this, for the crack spread data, looking at the crude oil futures, basically the crude oil future can be evaluated to see if it's non stationary. And there's this augmented Dickey-Fuller test for non-stationarity. And it basically has a null hypothesis that the model or the series is non stationary, or it has a unit root versus the alternative that it doesn't.

And so testing that null hypothesis that it's non stationary yields a p value of 0.164 for CLC 1, the first nearest contract, near month contract of the futures like crude. And so the data suggests that crude has a distribution that's non stationary integrated order 1.

And the HOC1 also basically has a test for p value for non-stationarity of 0.3265. So we can't reject non-stationarity or unit root in those series with these test statistics. In analyzing the data, this suggests that we basically need to accommodate that non-stationarity when we specify the models. Let me just see if there's some results here.

For this series, actually the case notes will go through actually conducting this Johansen procedure for testing for the rank of the cointegrated process. And that test basically has different test statistic for testing whether the rank is 0, 1, less than or equal to 1, less than or equal to 2. And one can see that there's marginal-- the test statistic is almost significant at the 10% level for the overall series.

It's not significant for the rank being less than or equal to 1. And so these results, it doesn't suggest there's strong non-stationarity. But certainly with that non-stationarity is no more than rank one for the series. And the eigenvector

corresponding to the stationarity relationship is given by these coefficients of 1 on the crude oil future, 1.3 on the RBOB and minus 1.7 on the heating oil.

So what this suggests is that there's considerable variability in these energy futures contracts. What appears to be stationary is some linear combination of crude plus gasoline minus heating oil. And in terms of why does it combine that way, well, there are all kinds of factors that we went through-- cost of refining, supply and demand, seasonality, which affect things.

And so when analyzed, sort of ignoring seasonality, these would be the linear combinations that appear to be stationary over time. Yeah?

AUDIENCE: Why did you choose to use the futures prices as opposed to the spot? And how did you combine the data with actual [INAUDIBLE]?

PROFESSOR: I chose this because if refiners are wanting to hedge their risks, then they will go to the futures market to hedge those. And so working with these data, one can then consider problems of hedging refinery production risks. And so that's why.

AUDIENCE: [INAUDIBLE]

PROFESSOR: OK, well, the Energy Information Administration provides historical data which gives the first month, the second month, the third month available for each of these contracts. And so I chose the first month contract for each of these features. Those 10 are the most liquid. Depending on what one is hedging, one would use perhaps longer periods for those.

There's some very nice finance problems dealing with hedging, hedging these kinds of risks, and as well as trading these kinds of risk. Traders can try to exploit short term movements in these. Anyway, I'll let you look through these, the case note later. And it does provide some detail on the coefficient estimates. And one can basically get a handle on how these things are being specified.

So let's go back. The next topic I want to cover is linear state space models. It turns out that many of these time series models appropriate in economics and finance

can be expressed as a linear state space model.

I'm going to introduce the general notation first and then provide illustrations of this general notation with a number of different examples. So the formulation is we have basically an observation vector at time t , y_t . This is our multivariate time series that we're modeling.

Now, I've chosen it to be k dimensional for the observations. There's an underlying state vector that's of m dimensions, which basically characterizes the state of the process at time t . There's an observation error vector at time t , ϵ_t So it's k by 1 as well, corresponding to y . And there's a state transition innovation error vector, which is n by 1 , which actually can be different from m , the dimension of the state vector.

So we have in the state space specification, we're going to specify two equations, one for how the states evolve over time and another for how the observations or measurements evolve, depending on the underlying states. So let's first focus on a state equation which describes how the state progresses from the state at time t to the state at time $t + 1$.

Because this is a linear state space model, basically the state at $t + 1$ is going to be some linear function of the states at time t plus some noise. And that noise is given by η_t being independent [η_t identically η_t] distributed white noise or normally distributed with some covariance matrix Q_t positive definite. And A_t is some linear transformation of those, which characterize the uncertainty in the particular states.

So there's a great deal of flexibility here in how things depend on each other. And right now, it will appear just like a lot of notation. But as we see it in different cases, you'll see how these terms come into play. And they're very straightforward.

So we're considering simple linear transformations of the states plus noise. And then the observation equation or measurement equation is a linear transformation of the underlying states plus noise. So the matrix z_t is the observation coefficients matrix. And the noise or innovations ϵ_t are, we'll assume, independent [ϵ_t

identically ?] distributed, normal multivariate normal, [INAUDIBLE] some covariance matrix, Σ_t .

To be fully general, the subscript t means the covariance can depend on time t . It doesn't have to, but it can. These two equations can be written together in a joint equation where we see that the underlying state at time t gets transformed with T_{t+1} to the state at $t+1$ plus residual innovation term. And the observation equation y_t is $z_t' x_t$ plus that.

So we're representing how the states evolve over time and how the observations depend on the underlying states in this joint equation. And the structure of basically this sort of linear function of states plus error, the error term u_t here is normally distributed with covariance matrix Σ , which has this structure. It's a block diagonal.

We have the covariance of the epsilons as the Σ . And the covariance of r_t is $r_t \Sigma r_t'$. So you may recall when we take a covariance matrix of linear function of random variables given by a matrix, then it's that linear function, r , times the covariance matrix times the transpose.

So that term comes into play. So let's see how a capital asset pricing model with time varying betas can be represented as a linear state space model. You'll recall, we discussed this model a few lectures ago, where we have the excess return of a given stock, r_t , is a linear function of the excess return of the market portfolio, r_{mt} , plus error.

What we're going to do now is extend that previous model by adding time dependence, t , to the regression parameters. The alpha is not a constant. It is going to vary by time. And the beta is also going to vary by time.

And how will they vary by time? Well, we're going to assume that the alpha t is a Gaussian random walk. And the beta is also a Gaussian random walk.

And with that set up, we have the following expression for the state equation. OK, the state equation, which is just the unknown parameters-- it's the alpha and the

beta at given time t . The state at time t gets adjusted to the state at time $t + 1$ by just adding these random walk terms to it. So it's a very simple process. We have the identity times the previous state plus the identity times this vector of these innovations.

So s_{t+1} is equal to t_t , s_t plus $r_t \eta_t$, where this matrix, t sub t and r sub t are trivial. They're just the identity. And η_t has a covariance matrix which is just given by $q_t \sigma^2$, $\sigma^2 \epsilon$. This is a complex way, perhaps, of representing this model. But it puts this simple model into that linear states based framework.

Now, the observation equation is given by this expression defining the z_t matrix as the unit element in R_{mt} . So it's basically a row vector or a row matrix, one-row matrix. And ϵ_t is the white noise process.

Now, putting these equations together, we basically have the equation for the state transition and the observation equation together. We have this form for that.

So now, let's consider a second case of linear regression models where we have a time varying beta. In a way, this case we just looked at is a simple case of that. But let's look at a more general case where we have p independent variables, which could be time varying. So we have a regression model almost as we've considered it previously. y_t is equal to x_t transpose beta t plus ϵ_t .

The difference now is our regression coefficients, beta, are allowed to change over time. How do they change over time? Well, we're going to assume that those also follow independent random walks with variances of the random walks that may depend on the component.

So the joint state space equation here is given by the identity times s_t plus η_t . That's basically the random walk process for the underlying regression parameters. And y_t is equal to x_t transpose times the same regression parameters plus the observation error. I guess needless to say, if we consider the special case where the random walk process is degenerate and they're basically steps of size zero,

then we get the normal linear regression model coming out of this.

If we were to be specifying the linear state space implementation of this model and consider successive estimates of the model parameters over time, then these equations would give us recursive estimates for updating regressions as we add additional values to the data, additional observations to the data. Let's look at autoregressive models of order p .

The autoregressive model of order p for a univariate time series has the setup given here. It's a polynomial lag of the response variable y_t is equal to the innovation ϵ_t . And we can define the state vector to be equal to the vector of p values, p successive values of the process.

And so we basically get a combination here of the observation equation and state equation joining where basically one of the states is actually equal to the observation. And basically, with this definition for a state of the vector at the next time point, t , that is equal to this linear transformation of the lag state vector plus that innovation term. I dropped the mic.

So the notation here shows the structure for how this linear state space model is evolving. Basically, the observation equation is the linear combination of the five multiples of lags of the values plus the residual. And the previous lags of the states are just simply the identities times those values, shifted.

So it's a very simple structure for the autoregressive process as a linear state space model. We have, as I was just saying, for the transition matrix, T_{t-1} , this matrix and the observation equation is essentially picking out the first element of the state vector, which has no measurement error. So that simplifies that.

The moving average model of order q could also be expressed as a linear state space model. Remember, the moving average model is one where our response variable, y , is simply some linear combination of innovations, q past innovations. And this state vector, if we consider the state vector just being basically q lags of the innovations, then the transition of those underlying states is given by this expression

here.

And we have a state equation, an observation equation, which has these forms for these various transition matrices and for how the innovation terms are related. Let me just finish up with example showing with the autoregressive moving average model. And many years ago, it was actually very difficult to specify the estimation methods for autoregressive moving average models.

But the implementation of these models as linear state space models facilitated that greatly. And with the ARMA model, the set up basically is a combination of the autoregressive moving average processes. We have an autoregressive of the y 's is equal to a moving average of the residuals or the innovations. And it's convenient in the setup for linear state space models to define the dimension, m , which is the maximum of p and q plus 1, and think of having basically a possibly m order polynomial lag for each of those two series.

And we can basically constrain those values to be 0 if m is greater than p or m is greater than q . And Harvey, in a very important work in '93, actually defined a particular state space representation for this process. And I guess it's important to know that with these linear state space models, we're dealing with characterizing structure in m dimensional space. There's often some choice in how you represent your underlying states.

You can basically re-parametrize the models by considering invertible linear transformations of the underlying states. So let me go back here. In expressing the state equation generally is T sub t st -- sorry ta to t . This matrix, T sub t and st , the st can be replaced by a linear transformation of st , so long as we multiply the T sub t by the inverse of that transformation.

So there's flexibility in the choice of our linear state space specification. And so there really are many different equivalent linear state space models for a given process depending on exactly how you define the states and the underlying transformation matrix, t . And the beauty of Harvey's work was coming up with a nice representation for the states, where we had very simple forms for the various

matrices.

And the lecture notes here go through the derivation of that for the ARMA process. And this derivation is-- I just want to go through the first case just to highlight how the argument goes. We basically have this equation, which is the original equation for an ARMA pq process.

And Harvey says, well, define the first or the state at time t to be equal to the observation at time t . If we do that, then how does this equation relate to the basically-- this is the state at the next time point, $T + 1$, is equal to ϕ_1 times the state at time t , plus a second state at time t and a residual innovation, η_t .

So by choosing the first state to be the observation value at that time, we can then solve for the second state, which is given by this expression just by rewriting our model equation in terms of s_{1t} , s_{2t} , and η_t . So this s_{2t} is this function of the observations and η_t .

So it's a very simple specification of the second state. Just what is that second state element given this definition of the first one? And one can do this process iteratively getting rid of the observations and replacing them by underlying states. And at the end of the day, you end up with this very simple form for the transition matrix, t .

Basically, the t has the autoregressive components as the first column of the t matrix. And this r matrix has this vector of the moving average components. So it's a very nice way to represent the model. Coming up with it was something very clever that he did.

But what one can see is that this basic model where you have the states transitioning according to a linear transformation of the previous state plus error, and the observation being some function of the current states, plus error or not, depending on the formulation, is the representation. Now, with all of these models, a reason why linear states space modeling is in fact effective is that they're specification is fully specified with the Kalman filter.

So with this formulation of linear state space models, the Kalman filter as a

methodology is the recursive computation of the probability density functions for the underlying states at basically $t + 1$ given information up to time t , as well as the joint density of the future state and the future observation at $t + 1$, given information up to time t .

And also just the marginal distribution of the next observation given the information up to time t . So what I want to do is just go through with you how the Kalman filter is implemented and defined. And the implementation of the Kalman filter requires us to have some notation that's a bit involved, but we'll hopefully explain it so it's very straightforward.

There are basically conditional means of the states. s_t given t is the mean value of the state at time t given the information up to time t . If we condition on $t - 1$, then it's the expectation of the state at time t given the information up to $t - 1$. And then y_{t-1} is the expectation of the observation given information up to $t - 1$.

There's also conditional covariances and mean squared errors. All these covariances are determined by omegas. The subscript corresponds to the states, s , or observation, y . And basically, the conditioning set is either information up to time t , $t - 1$ or $t - 1$ in the second case. And we want to compute basically the covariance matrix of the states given whatever the information is, information up to time t , $t - 1$.

So these covariance matrices are the expectation of the state minus their expectation under the conditioning times the state minus the expectation transpose. That's the definition of that covariance matrix. So the different definitions here correspond to just whether we're conditioning on information.

And then the observation innovations or residuals are the difference between an observation, y_t , and its estimate given information up to $t - 1$. So the residuals in this process are the innovation residuals, one period ahead. And the Kalman filter consists of four steps. We basically want to, first, predict the state vector one step ahead.

So given our estimate of the state vector at time $t - 1$, we want to predict this state vector at time t . And we also want to predict the observation at time t given our estimate at state vector time $t - 1$. And so at time $t - 1$, we can estimate these quantities. [INAUDIBLE]

At $t - 1$, we can basically predict what the state is going to and predict what the observation is going to be. And we can estimate how much error there's going to be in those estimates, by these covariance matrices. The second step is updating these predictions to get our estimate of the state given the observation at time t and to update our uncertainty about that state given this new observation.

So basically, our estimate of the state at time t is an adjustment to our estimate given information up to $t - 1$, plus a function of the difference between what we observed and what we predicted. And this gain function matrix is called the filter gain matrix. And basically, it characterizes how do we adjust our prediction of the underlying state depending on what happened.

So that's the filter gain matrix. So we actually do gain information with each observation about what the new value of the process is. And that information is characterized by filter gain matrix. You'll notice that the uncertainty in the state at time t , this Σ_t given t , that's equal to the covariance matrix given $t - 1$. So it's our beginning level of uncertainty adjusted by a term that tells us how much information did we get from that new information.

So notice that there's a minus sign there. We're basically reducing our uncertainty about the state given the information in the innovation that we now have observed. Then, there's a forecasting step which is used to forecast the state one period forward, is simply given by this linear transformation of the previous state. And we can also update our covariance matrix for future states given the previous state by applying this formula which is a recursive formula for estimating covariances.

So we have forecasting algorithms that are simple linear functions of these estimates. And then finally, there's a smoothing step which is characterizing the

conditional expectation of underlying states, given information in the whole time series.

And so ordinarily with Kalman filters, Kalman filters are applied sequentially over time where one basically is predicting ahead one step, updating that prediction, predicting ahead another step, updating the information on the states. And that overall process is the process of actually computing the likelihood function for these linear state space models.

And so the Kalman filter is basically ultimately applied for successive forecasting of the process but also for helping us identify what the underlying model parameters are using maximum likelihood methods. And so the likelihood function for the linear state space model is basically the-- or the log likelihood is the log likelihood of the entire data series, given the unknown parameters. But that can be expressed as the product of the conditional distributions of each successive observation, given the history.

And so basically, the likelihood of θ is the likelihood of the first observation times the density of the second observation given the first times and so forth for the whole series. And so the likelihood function is basically a function of all these terms that we were computing with the Kalman filter. And with the Kalman filter, it basically provides all the terms necessary for this estimation.

If the error terms are normally distributed, then the means and variances of these estimates are in fact characterizing the exact distributions of the process. Basically, we're taking, if the innovation series are all normal random variables, then the linear state space model, all it's doing is taking linear combinations of normals for the underlying states and for the actual observations.

And normal distributions are fully characterized by their mean vectors and covariance matrices. And the Kalman filter provides a way to update these distributions for all these features of a model, the underlying states as well as the distributions of the observations. So that's a brief introduction the Kalman filter. Let's finish there. Thank you.

