PROFESSOR: All right. I want to complete the discussion on volatility modeling in the first part of the lecture today. And last time we addressed the definition of ARCH models, which allow for time varying volatility in modeling the returns of a financial time series.

And we were looking last time at modeling the Euro-dollar exchange rate returns. And we went through fitting arch models to those returns, and also looked at fitting the GARCH model to those returns. And to recap, the GARCH model extends upon the ARCH model by adding some extra terms.

So if you look at this expression for the GARCH model, the first two terms for the time varying volatility sigma squared t is a linear combination of the past sort of residual returns squared. That's the ARCH model, p [? of ?] [? those. ?] So the current volatility depends upon what's happened in excess returns over the last p periods.

But then we add an extra term, which is corresponds to queue levels of the previous volatility. And so what we're doing with GARCH models is adding extra parameters to the ARCH, but an advantage of considering these extra parameters which relate basically the current volatility sigma squared t with the previous or lagged value sigma squared t minus j for lags j is that we may be able to have a model with many fewer parameters.

So indeed, if we fit these models to the exchange rate returns, what we found last time-- let me go through and show that-- was basically here are various fits of the three cases of ARCH models. ARCH orders 1, 2, and 10, thinking we maybe need many lags to fit volatility. And then the GARCH model 1,1, where we only have one ARCH term and one GARCH term. And so basically the green line-- or rather the

blue line in this graph, shows the plot of the fitted GARCH 1,1, model as compared with the ARCH models.

Now, in looking at this graph, one can actually see some features of how these models are fitting volatility, which is important to understand. One is that the ARCH models have a hard lower bound on the volatility. Basically there's a constant term in the volatility equation. And because the additional terms are squared, excess returns, it basically-- the volatility does have lower bound of that intercept. So depending on what range you fit the data over, that lower bound is going to be defined by-- or it will be determined by the data you're fitting to.

As you increase the ARCH order, you basically allow for a greater range of-- or a lower lower bound of that. And with the GARCH model you can see that this blue line is actually predicting very different levels of volatility over the entire range of the series. So it really is much more flexible.

Now-- and in these fits, we are assuming Gaussian distributions for the innovations in the return series. We'll soon pursue looking at alternatives to that, but let me talk just a little bit more about the GARCH model going back to lecture notes here. So let me expand this.

OK. So there's the specification. The GARCH 1,1 model. One thing to note is that this GARCH 1,1 model does relate to an ARMA, an Auto Regressive Moving Average, process in the squared residuals.

So if we look at the top line, which is the equation for the GARCH 1,1 model, consider eliminating sigma squared t by using a new innovation term, little ut, which is the difference between the squared residual and the true volatility given by the model. So if you plug in the difference between our squared excess return and the current volatility, that should have mean 0 because sigma squared t, the t volatility, squared is equal to the square-- or is equal to the expectation of the squared excess residual return, epsilon t squared.

So if we plug that in, we basically get an ARMA model for the squared residuals.

And so epsilon t squared is alpha 0 plus alpha 1 plus beta 1 the squared t minus 1 lag plus ut minus beta 1 ut. And so what this implies is an ARMA 1,1 model with white noise that has mean 0 and variance 2 sigma to the fourth. Just plugging things in.

And through our knowledge, understanding, of univariate time series models, ARMA models, we can express this ARMA model for the squared residuals as basically a polynomial lag of the squared residuals is equal to a polynomial lag of the innovations. And so we have this expression for what the innovations are.

And it's required that the roots of this a of l operator, when it thought of it on the complex plane, have roots outside the unit circle, which corresponds to alpha 1 plus beta 1 being less than 1 in magnitude. So in order for these volatility models not to blow up and be stationary, covariant stationary, we have these bounds on the parameters.

OK, let's look at the unconditional volatility or long run variance of the GARCH model. If you take expectations on both sides of the GARCH model equation, you basically have the expectation of sigma squared sub t-- in the long run is sigma star squared-- is alpha 0 plus alpha 1 plus beta 1 sigma star squared.

So that sigma star squared there is the expectation of the t minus 1 volatility squared in the limit. And then you can just solve for this and see the sigma star squared is equal alpha 0 over 1 minus alpha 1 minus beta 1.

And in terms of the stationarity conditions for the process, if the long run variance, in order for that to be finite, you need alpha 1 plus beta 1 to be less than 1 in magnitude. And if you consider the general GARCH p,1 model, then the same argument leads to a long run variance being equal to alpha 0, the sort of intercept term in the GARCH model, divided by 1 minus the sum of all the parameters.

So these GARCH models lead to constraints on the parameters that are important to incorporate when we're doing any estimation of these underlying parameters. And it does complicate things, actually.

So with maximum likelihood estimation, the routine for maximum likelihood estimation is the same for all models. We basically want to determine the likelihood function of our data given the unknown parameters. And the likelihood function is the probability density function of the data conditional on the parameters.

So our likelihood function as a function of the unknown parameters c, alpha, and beta is the value of the probability density, the joint density of all the data conditional on those parameters. And that joint density function can be expressed as the product of successive conditional expectations of the time series.

And those conditional densities are normal random variables. So we can just plug in what we know to be the probability densities of normals for the t-th innovation epsilon t. And we just optimize that function.

Now, the challenge with estimating these GARCH models in part is the constraints on the underlying parameters. Those need to be enforced. So we have to have that the alpha i are greater than 0. Also, the beta j are greater than 0. And the sum of all of them is between 0 and 1.

Who in this class has had courses in numerical analysis and done some optimization of functions? Non-linear functions? Anybody?

OK. Well, in addressing this kind of problem, which will come up with any complex model that you need to estimate, say via maximum likelihood, the optimization methods do really well if you're optimizing a convex function, finding the minimum of a convex function. And it's always nice to do minimization over sort of an unconstrained range of underlying parameters.

And one of the tracks in solving these problems is to transform the parameters to a scale where they're unlimited in range, basically. So if you have a positive random variable, you might use to log of that variable as the thing to be optimizing over. If the variable's between 0 and 1, then you might use that variable divided by 1 minus that variable and then take the log of that. And that's unconstrained. So there are tricks for how you do this optimization, which come into play.

Anyway, that's the likelihood with the normal distribution. And we have computer programs that will solve that directly so we don't have to worry about this particular case.

Once we fit this model, we want to evaluate how good it is and the evaluation is based upon looking at the residuals from the model. So what we have are these innovations, epsilon hat t, which should be distributed with variance or volatility sigma hat t. Those should be uncorrelated with themselves or at least to the extent that they can be.

And the squared standardized residuals should also be uncorrelated. What we're trying to do with these models is to capture the dependence, actually, in this squared residuals, which is measuring the magnitude of the excess in returns. So though should be uncorrelated.

There are various test for normality. I've listed some of those that are the most popular here. And then there's issues of model selection for deciding sort of which GARCH model to apply.

I wanted to go through an example of this analysis with the Euro-dollar exchange rate. So let me go to this case study note. So let's see. There's a package in r called rugarch for univariate GARCH models, which fits various GARCH models with different-- and fits them by maximum likelihood.

So with this packet-- with this particular library in r, I fit the GARCH model after actually fitting the mean process for the exchange rate returns. Now, when we looked at things last time, we basically looked at modeling the squared returns. In fact, there may be an underlying mean process that needs to be specified as well.

So in this section of the case note, I initially fit an auto regressive process using the Akaike information criterion to choose the order of the auto regressive process and then fit a GARCH model with normal GARCH terms. And this is a plot of the normal q, q plot of the auto regressive residuals.

And what you can see is that the points lie along a straight line sort of in the middle of the range. But on the extremes, they depart from that straight line. This basically is a measure of standardized quantiles. So in terms of standard units away from the mean for the residuals, we tend to get many more high values and many more low values with the Gaussian distribution. So that really isn't fitting very well.

If we proceed and fit-- OK, actually that plot was just the simple ARCH model with no GARCH terms. And then this is the graph of the q, q plot with the Gaussian assumption. So here we can see that the residuals from this model are suggesting that it may do a pretty good job when things are only a few standard deviations away from the mean. Less than two 2, 2.5.

But when we get to more extreme values, this isn't modeling things well. So one alternative is to consider a heavier tailed distribution than the normal, namely the t distribution. And consider identifying what t distribution best fits the data.

So let's just look at what ends up being the maximum likelihood estimate for the degrees of freedom parameter, which is 10 degrees of freedom. This shows the q, q plot when you have a non-Gaussian distribution that's t with 10 degrees of freedom. It basically is explaining these residuals quite well, so that's accommodating the heavier tailed distribution of these values.

With this GARCH model, let's see-- if you compare sort of estimate of volatility under the GARCH and ARCH models-- the GARCH models with the t distribution-- sorry t distribution versus Gaussian. Here's just a graph showing time series plots of the estimated volatility over time, which actually look quite close. But when you look at the differences, there really are differences.

And so it turns out that the volatility function or the volatility estimate GARCH models with Gaussian versus GARCH with t distributions are really very, very similar. And the heavier tailed distribution of the t distribution means that the distribution of actual volatility is greater. But in terms of estimating the volatility, you have quite similar estimates of the volatility coming out.

And this display-- which you'll be able to see more clearly in the case notes that I'll post up-- but show that these are really quite similar in magnitude. And the value at risk concept that was just-- by Ken couple weeks ago in his lecture from Morgan Stanley-- concerns the issue of estimating what is the likelihood of returns exceeding some threshold. And if we use the t distribution for measuring variability of the excess returns, then the computations in the notes indicate how you would compute these value at risk limits.

If you compare the t distribution with a Gaussian distribution at these nominal levels for value at risk of like 2.5% or 5%, surprisingly you won't get too much difference. It's really in looking at sort of the extreme tails of the distribution that things come into play.

And so I wanted to show you how that plays out by showing you another graph here. Those of you who have had a statistics course before have heard that sort of a t distribution can be a good approximation to a normal-- or it can be approximated well by a normal if the degrees of freedom for the t are at some level. And who wants to suggest a degrees of freedom that you might have before you're comfortable approximating a t with a normal? Danny?

AUDIENCE: 30 or 40.

PROFESSOR: 30 or 40. Sometimes people say even 25. Above 25, you can almost expect the t distribution to be a good approximation to the normal.

Well, this is a graph the PDF for a standard normal versus a standard t with 30 degrees of freedom. And you can see that the density functions are very, very close. The standard-- the CDFs, the Cumulative Distribution Functions, which is the likelihood of being less than or equal to the horizontal value ranges between 0 and 1-- is almost indistinguishable.

But if you look at the tails of the distribution, here I've computed the log of the CDF function. You basically have to move much more than two standard deviations away from the mean before there's really a difference in the t distribution with 30 degrees

of freedom.

Now I'm going to page up by reducing the degrees of freedom. Let's see. If we could do a page down here. Page down. Oh, page up. OK.

So here is 20 degrees of freedom. Here's 10 degrees of freedom, in our case, which turns out to be sort of the best fit of the t distribution. And what you can see is that, in terms of standard deviation units, up to about two standard deviations below the mean, we're basically getting virtually the same probability mass at the extreme below.

But as we go to four or six standard deviations, then we get heavier mass with the t distribution. In discussion of results in finance when you sort of fit models, people talk about, oh, there was six standard deviation move or-- which is just virtually impossible to occur. Well, with t distributions a six standard deviation move occurs about 1 in 10,000 times according to this fit.

And so it actually is a common [? idiomatic. ?] And so it's important to know that these t distributions are benefiting us by giving us a much better gauge of what the tail distribution is like. And we call these distributions leptokurtic, meaning they're heavier tailed than a normal distribution. Actually, lepto means slender, I believe, if you're Greek or have the Greek origin of the word. And you can see that the blue curve, which is the t distribution, is sort of a bit more slender in the center of the distribution, which allows it to have heavier tails.

All right. So t distributions are very useful. Let's go back to this case note here which discusses-- this case note goes through, actually, fitting the t distribution-- identifying the degrees of freedom for this t model. And so with ru GARCH package, we can get the log likelihood of the data fit under the t distribution assumption.

And here's a graph of the negative log likelihood versus the degrees of freedom in the t model. So with maximum likelihood we identify the value, which minimizes the negative log likelihood. And that comes out as that 10 value.

All right. Let's go back to these notes and see what else we want to talk about. All

right.

OK, with these GARCH models we actually are able to model volatility clustering. And volatility clustering is where, over time, you expect volatility to be high during some periods and to be low during other periods. And the GARCH model can accommodate that. So large volatilities tend to be followed by large, small volatilities tend to be followed by small ones.

OK. The returns have heavier tails than Gaussian distributions. Actually, even if we have Gaussian errors in the GARCH model, it's still heavier tailed than a Gaussian. The homework goes into that a little bit. And the-- well, actually one of the original papers by Engle with Bollerslev, who introduced the GARCH model, discusses these features and how useful they are for modeling financial time series.

Now, a property of these models that may be obvious, perhaps, but it is-- OK, these are models that are appropriate for modeling covariant stationary time series. So the volatility measure, which is a measure of the squared excess return, is basically a covariant stationary process.

So what does that mean? That means that's going to have a long term mean. So with these are GARCH models that are covariant stationary, there's going to be a long term mean of the GARCH process. And this discussion here details how this GARCH process is essentially a mean reversion of the volatility to that value.

So basically, the sort of excess volatility of the squared residuals relative to their long term average is some multiple of the previous period's excess volatility. So if we build forecasting models of volatility with GARCH models, what's going to happen?

Basically, in the long run we predict that any volatility value is going to revert to this long run average. And in the short run, it's going to move incrementally to that value. So these GARCH models are very good for describing volatility relative to the long term average.

In terms of their usefulness for prediction, well, they really predict that volatility is

going to revert back to the mean at some rate. And the rate at which the volatility reverts back is given by alpha 1 plus beta 1. So that number, which is less than 1 for covariant stationarity, is sort of measuring, basically, how quickly you are reverting back to the mean. And that sum is actually called a persistence parameter in GARCH models as well.

So is volatility persisting or not? Well, the larger alpha 1 plus beta 1 is. The more persistent volatility is, meaning it's reverting back to that long run average very, very slowly.

In the implementation of volatility estimates with the risk metrics methodology, they actually don't assume that there is a long run volatility. And so that basically you'll have alpha 1 be equal to 0 and beta 1 equal to, say, 0.95. So or rather the alpha 0 is 0 and the alpha 1 and beta 1 will actually sum to 1. And so you actually are tracking a potentially non-stationary volatility, which allows you to be estimating the volatility without presuming a long run average is consistent with the past.

There are many extensions of the GARCH models. And there's wide literature on that. For this course, I think it's important to understand the fundamentals of these models in terms of how they're specified under Gaussian and t assumptions. Extending them can be very interesting. And there are many papers to look at for that.

OK. let's pause for a minute and get the next topic.

All right. The next topic is time series, multivariate time series. In two lectures ago of mine, we talked about univariate time series and basic methodologies there. We're now going to be extending that to multivariate time series.

Turns out there's a multivariate Wold representation theorem, extension of the univariate one. There are auto regressive processes for multivariate cases, which are vector auto regressive processes. Least squares estimation comes into play. And then we'll see where our regression analysis understanding allows us to specify these vector auto regressive processes nicely.

There's an optimality properties of ordinary least squares estimates component wise, which we'll highlight in about a half an hour. And go through the maximum likelihood estimation model selection methods, which are just very straightforward extensions of the same concepts for univariate time series and univariate regressions.

So let's talk-- let's introduce the notation for multivariate time series. We have a stochastic process, which now is multivariate. So we have bold x of t is some m dimensional valued random variable. And it's stochastic process that varies over time t. And we can think of this as m different time series corresponding to the m components of the given process.

So, say, with exchange rates we could be modeling m different exchange rate values and want to model those jointly as a time series. Or we could have collections of stocks that we're modeling.

So each of the components individually can be treated as univariate series with univariate methods. With the multivariate case, we extend the definition of covariance stationarity to correspond to finite bounded first and second order moments. So we need to talk about the first order moment of the multivariate time series.

Mu now is an m vector, which is the vector of expected values of the individual components, which we can denote by mu 1 through mu m. So we basically have m vectors for our mean.

Then for the variance/covariance matrix, let's define gamma 0 to be the variance/covariance matrix of the t-th observation of our multivariate process. So that's equal to the expected value of xt minus mu xt minus mu prime. So when we write that down, we have xt minus mu. This is basically an m by 1 vector and then xt minus mu prime is a 1 by m vector.

And so the product of that is an m by m quantity. So the 1, 1 element of that product is the variance of x1t. And the diagonal entries are the variances of the components

series. And the off diagonal values are the covariance between the i-th row series and the j-th column series, as given by the i-th row of x and the j-th column of x transpose.

So we're just collecting together all the variances covariances together. And the notation is very straightforward and simple with the matrix notation given here.

Now, the correlation matrix, r0, is obtained by pre and post multiplying this covariance matrix gamma 0 by a diagonal matrix with the square roots of the diagonal of this matrix.

Now what's a correlation? Correlation is the correlation between two random variables where we've standardize the variables to have mean 0 and variance 1. So what we want to do is basically divide through all of these variables by their standard deviation and compute the covariance matrix on that new scaling.

That's equivalent to just pre and post multiplying by that diagonal of the inverse of the standard deviations. So with matrix algebra, that formula is-- I think it's very clear. And this is-- now with-- the previous discussion was just looking at the sort of contemporaneous covariance matrix of the time series values at the given time t with itself. We want to look at, also, the cross covariance matrices.

So how are the current values of the multivariate time series xt-- how do they covary with the k-th lag of those values? So gamma k is looking at how the current period is vector values as covaried with the k-th lag of those values. So this covariance matrix has covariance elements given in this display. And we can define the cross correlation matrix by similarly pre and post multiplying by the inverse of the standard deviations.

The diagonal of gamma 0 is the covariance-- or is the matrix of diagonal entries of variances. Now, properties of these matrices is-- OK, gamma 0 is a symmetric matrix that we had before. But gamma k where k is greater than 1 or less than-- or greater or equal to 1 or less than-- basically different from 0. This is not symmetric.

Basically, you may have lags of some variables that are positively correlated with

others and not vice versa. So the off diagonal entries here aren't necessarily even of the same sign, let alone equal and symmetric.

So with these covariance matrices, one can look at how things covary and whether they are-- whether there is, basically, a dependence between them. And you can define-- it's basically the j star series-- the j star component of the multivariate time series may lead the j-th one if the covariance of the k-th lag of j star is different from 0-- or the covariance of j star k lags ago is non-zero covaries with the j-th lag. Sorry. The current lag. So xt j star will lead xtj.

Basically, there's information in the lagged values of j star for the component j. So if we're trying to build models-- linear regression models, even, where we're trying to look at how-- trying to predict values, then if there's a non-zero covariance, then we can use those variables' information to actually project what the one variable is given the other.

Now, it can be the case that you have non-zero covariance in both directions. And so that suggests that there can be sort of feedback between these variables. It's not just that one variable causes another, but there can actually be feedback.

In economics and finance, there's a notion of Granger causality. And basically that-- well, Granger and Engle got the Nobel Prize number of years ago based on their work. And that work deals with identifying, in part, judgments of causality between-- or Granger causality between variables economic time series. And so Granger causality basically is sort of positive or non-zero correlation between variables where lags of one variable will cause another or cause changes in another.

All right. I want to just alert you to the existence of this Wold decomposition theorem. This is an advanced theorem, but it's a useful theorem to know exists. And this extends-- the univariate Wold decomposition theorem, which concerns the-- whenever we have a covariant stationary process, there exists a representation of that process, which is the sum of a deterministic process and s moving average process of a white noise.

So if you're modeling a time series and you're going to be specifying a covariant stationary process for that, there does exist a Wold decomposition representation of that. You can basically determine-- identify the deterministic process that the process might follow. It might be a linear trend over time or an exponential trend.

And if you remove that sort of deterministic process vt, then what remains is a process that can be modeled with a moving average of white noise. These. Now here, everything is changed from univariate case to multivariate case, so we have matrices in place of constants from before.

So these-- new concepts here are we have a multivariate white noise process. That's going to be a process a to t which is m dimensional which has mean 0. And the variance matrix of this m vector is going to be sigma, which is now m by m variance/covariance matrix of the components. And that must be a positive semi-definite definite.

And for white noise, we have covariances between, say, the current t innovation and a lag of its value are 0. So these are uncorrelated multivariate white noise processes. And so they're uncorrelated with each other at various lags. And the innovation a to t has a covariance of 0 with the deterministic process. Actually, that's pretty much a given if we have a deterministic process.

Now, the term psi k-- basically we have this vector xt is equal to the sum m vectored process vt plus this weighted average of innovations. What's required is that the sum of this-- basically each term psi k and its transpose converges.

Now, if you were to take that xt process and say let me compute the variance/covariance matrix of that representation, then you would basically get terms in the covariance matrix which includes this sum of terms. So that sum has to be finite in order for this to be covariant stationary.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Yes?

AUDIENCE: Could you define what you mean by innovation?

PROFESSOR: Oh, OK. Well, the innovation is-- let's see. With-- let me go back up here. OK.

The innovation process-- innovation process. OK, if we have, as in this case, we have sort of our xt stochastic process. And we have sort of, say, f sub t minus 1 equal to the information on x2 minus 1 x2 minus 2. Basically consisting of the information set available before time t.

Then we can model xt to be the expected value of xt given ft minus 1 plus an innovation. And so our objective in these models is to be thinking of how is that process evolving where we can model the process as well as possible using information up to time before t.

And then there's some disturbance about that model. There's something new that's happened at time t that wasn't available before. And that's this innovation process.

So this representation with the Wold decomposition is converting the-- or representing, basically, the bits of information that are affecting the process that are occurring at time t and wasn't available prior to that. All right.

Well, let's move on to vector auto regressive processes. OK, this representation for a vector auto regressive process is an extension of the univariate auto regressive process to p dimensions. Sorry, to m dimensions.

And so our xt is an m vector. That's going to be equal to some constant vector c plus a matrix phi 1 times lag of xt first order, xt minus 1. Plus another matrix, phi 2 times the second lag of xt, xt minus 2. Up to the p-th term, which is a phi p m by m matrix times x2 minus p plus this innovation term.

So this is essentially-- this is basically how a univariate auto regressive process extends to an m variate case. And what this allows one to do is model how a given component of the multivariate series-- like how one exchange rate varies depending on how other exchange rates might vary. Exchange rates tend to co-move together in that example.

So if we look at what this represents in terms of basically a component series, we can consider fixing j, a component of the multivariate process. It could be the first, the last, or the j-th, somewhere in the middle. And that component time series-- like a fixed exchange rate series or time series, whatever we're focused on in our modeling-- is a generalization of the auto regressive model where we have the auto regressive terms of the j-th series on lags of the j-th series up to order p.

So we have the univariate auto regressive model, but we also add to that terms corresponding to the relationship between xj and xj star. So how does xj, the j-th component, depend on other variables, other components of the multivariate series. And those are given here. So it's a convenient way to allow for interdependence among the components and model that.

OK. This slide deals with representing a p-th order process as a first order process with vector auto regressions. Now the concept here is really a very powerful concept that's applied in time series methods, which is when you are modeling dependence that goes back, say, a number of lags like p lags, the structure can actually be re-expressed as simply a first order dependence only.

And so it's much easier sort of to deal with just a lag one dependence then to consider p lag dependence and the complications involved with that. So-- and this technique is one where, in the early days of fitting, like auto regressive moving average processes and various smoothing methods, the model-- basically accommodating p lags complicated the analysis enormously. But one can actually re-express it just as a first order lag problem.

So in this case, what one does is one considers for a vector auto regressive process of order of p, simply stacking the values of the process. So let me just highlight what's going on there.

So if we have basically-- OK, so if we have x1, x2, xn, which are all m by 1 values, m vectors of the stochastic process. Then consider defining zt to be equal to xt transpose xt minus 1 transpose up to x t minus p minus 1 transpose. Or this is t minus p minus 1. So there are p terms.

And then if we consider the lagged value of that, that's x2 minus 1, x2 minus 2, x2 minus p transpose. So what we've done is we're considering zt. This is going to be m times p. It's actually 1 by m times p in this notation.

Well, actually I guess I should put transpose here. So m minus p by 1. OK, in the lecture notes it actually is primed there to indicate the transpose.

Well, if you define zt and zt minus 1 this way, then zt is equal to d plus a of zt minus 1 plus f, where this is d. Basically the constant term has the c entering and then 0's everywhere else. And the a matrix is phi 1 phi 2 up to phi p. And so basically the zt vector transforms the zt-- or is the transpose-- this linear transformation of the zt minus 1.

And we have sort of a very simple form for the constant term and a very simple form for the f vector. And this is-- renders the model into a sort of a first order time series model with a larger multivariate series, basically mp by 1.

Now, with this representation we basically have-- we can demonstrate that the process is going to be stationary if all eigenvalues of the companion matrix a have modulus less than 1. And let's see-- if we go back to the expression.

OK, if the eigenvalues of this matrix A are less than 1, then we won't get sort of an explosive behavior of the process when this basically increments over time with every previous value getting multiplied by the A matrix and scaling the process over time by the A-th power. So that is required. All eigenvalues of A have to be less than 1.

And equivalently, all roots of this equation need to be outside the unit circle. You remember there was a constraint of-- or a condition for univariate auto regressive models to be stationary, that the roots of the characteristic equation are all outside the unit circle. And the class notes go through and went through the derivation of that. This is the extension of that to the multivariate case.

And so basically one needs to solve for roots of a polynomial in Z and determine

whether those are outside the unit circle. Who can tell me what the order of the polynomial is here for this sort of determinant equation?

AUDIENCE: [INAUDIBLE] mp.

PROFESSOR: mp. Yes. It's basically of power mp. So in a determinant you basically are taking products of the m components in the matrix, various linear combinations of those. So that's going to be an mp dimensional polynomial. All right.

Well, the mean of the stationary VAR process can be computed rather easily by taking expectations of this on both sides. So if we take the expectation of xt and take expectations across both sides, we get that mu is the c vector plus the product of the phi case times mu plus 0.

So mu, the unconditional mean of the process, actually has this formula just solving for mu in the top-- in the second line to the third line. So here we can see that basically this expression 1 minus phi 1 through phi p, that inverse has to exist.

And actually, if we then plug in the value of c in terms of the unconditional mean, we get this expression for the original process. So the unconditional mean c, if we demeaned the process, there's busy know mean term. There's 0. And so basically the mean adjusted process x follows this multivariate vector auto regression with no mean, which is actually used when this is specified.

Now, this vector auto regression model can be expressed as a system of regression equations. And so what we have with the multivariate series, if we have multivariate data, we'll have n sample observations, xt which is basically the m vector of the multivariate process observed for n time points. And for the computations here, we're going to assume that we have p sort of-- we have pre-sample observations available to us. So we're essentially going to be considering models where we condition on the first p time points in order to facilitate the estimation methodology.

Then we can set up m regression models corresponding to each component of the m variate series. And so what we have is our original-- we have our collection of data values, which is x1 transpose, x2 transpose, down to xn transpose, which is an

n by m matrix.

OK, this is our multivariate time series where we were just-- the first row corresponds to the first time values, nth row to the nth time values. And we can set up m regression models where we're going to consider modeling the j-th column of this matrix. So we're just picking out the univariate time series corresponding to the j-th component. That's yj.

And we're going to model that has z beta j plus epsilon j where z is given by the vector of lagged values of the multivariate process where there's, for the t-th-- t minus first value we have that current value-- or the t minus first, t minus second, up to t minus p. So we have basically p m vectors here.

And so this j-th time series has elements that follow a linear regression model on the lags of the entire multivariate series up to p lags with their progression parameter given by beta j. And basically the beta j regression parameters corresponds to the various elements of the phi matrices. So now there's a one-to-one one correspondence between those.

All right. So I'm using now a notation where superscript j corresponds to the j-th component of the series, of the multivariate stochastic process. So we have an mp plus 1 vector progression parameters for each series j, and we have an epsilon j for an n-vector innovation errors for each series.

And so basically if this, the j-th column, is yj, we're modeling that to be equal to the simple matrix Z times beta j plus epsilon j, where this is n by 1. This is n by np plus 1. And this beta j is the mp plus 1 progression parameter. OK.

One might think, OK, one can consider each of these regressions for each of the component series, you could consider them separately. But to consider them all together, we can define the multivariate regression model, which has the following form. We basically have the n vectors for the first component, and then the second component up to nth component. So an n by p matrix of dependent variables, where each column corresponds to a different component series, follows a linear

regression model with the same Z matrix with different regression coefficient parameters, beta 1 through beta m corresponding to the different components of the multivariate series.

And we have epsilon 1, epsilon 2, up to epsilon m. So we're thinking of taking-- so basically the y1 y2 up to ym is essentially this original matrix of our multivariate time series because it's the first component in the first column and the nth component in the nth column.

And the-- this regression parameter or this explanatory variables matrix X, Z in this case corresponds to lags of the whole process up to p lags. So we're having lags of all the m variate process up to p lags. So that's mp and then plus 1 for our constant.

So this is the set up for a multivariate regression model. In terms of how one specifies this, well, actually, in economic theory this is also related to seemingly unrelated regressions, which you'll find in econometrics. If we want to specify this multivariate model, well, what we could do is we could actually specify each of the component models separately because we basically have sort of-- can think of the univariate regression model for each component series.

And this slide indicates basically what the formulas are for that. So if we don't know anything about multivariate regression we can say, well, let's start by just doing the univariate regression of each component series on the lags. And so we get our beta hat j's least squares estimates given by the usual formula where the independent variable is matrix Z goes Z transpose Z inverse Z transpose Y of the residual.

So these are familiar formulas. And if we did this for each of the component series j, then we would actually get sample estimates of the innovation process, eta 1. Basically the whole eta series. And we could actually define from these estimates of the innovations our covariance matrix for the innovations as the sample covariance matrix of these etas.

So all of these formulas are-- you're basically applying very straightforward estimation methods for the parameters of a linear regression and then estimating

variances/covariances of these innovation terms. So from this, we actually have estimates of this process in terms of the sigma and the beta hats. But it's made assuming that we can treat each of these component regressions separately.

A rather remarkable result is that these component-wise regressions are actually the optimal estimates for the multivariate regression as well. And as mathematicians, this kind of result is, I think, rather neat and elegant. And maybe some of you will think this is very obvious, but it actually-- it isn't quite obvious. That said, this component-wise estimation should be optimal as well.

And the next section of the lecture notes goes through this argument. And I'm going to, in the interest of time, go through this-- just sort of highlight what the results are. The details are in these notes that you can go through. And I will be happy to go into more detail about them during office hours.

But if we're fitting a vector auto regression model where there are no constraints on the coefficient matrices, phi 1 through phi p, then these component-wise estimates, accounting for arbitrary covariance matrix sigma for the innovations, those basically are equal to the generalized least squares estimates of these underlying parameters.

You'll recall we talked about the Gauss Markov theorem where we were able to extend the assumption of sort equal variances across observations to unequal variances and covariances. Well, it turns out to these component-wise OLS estimates are, in fact, the generalized least squared estimates. And under the assumption of Gaussian distributions for the innovations, they, in fact, are maximum likelihood estimates.

And this theory applies Kronecker products. We're not going to have any homework with Kronecker products. These notes really are for those who have some more extensive background in linear algebra. But it's a very nice use of these Kronecker product operators.

Basically, this notation-- or no, x circle-- I'll call it Kronecker-- is one where you take

a matrix A and a matrix B and you consider the matrix which takes each element of A times the whole matrix B. So we start with an m by n matrix A and end up with an mp by qn matrix by taking each element of A times the whole matrix B. So it's, they say, has this block structure.

So this is very simple definition. If you look at properties of transposition of matrices, you can prove these results. These are properties of the Kronecker product. And there's a vec operator which takes a matrix and simply stacks the columns together. And in the talk last Tuesday of Ivan's, talking about modeling the volatility surface, he basically, he was modeling a two dimensional surface-- or a surface in three dimensions, but there was two dimensions explaining it.

You basically can stack columns of the matrix and be modeling a vector instead of a matrix of values. So the vectorizing operator allows us to manipulate terms into a more convenient form.

And this multivariate regression model is one where it's set up as sort of a n by m matrix Y, having that structure. It can be expressed in terms of the linear regression form as y star equaling the vector, the vec of y. So we basically have y1 y2 down to ym all lined up. So this is pm by 1.

That's going to be equal to some matrix plus the epsilon 1 epsilon 2 down to epsilon n. And then there's going to be a matrix and a regression coefficient matrix beta 1 beta 2 down to beta p. So we consider vectorizing the beta matrix, vectorizing epsilon, and vectorizing y.

And then in order to define this sort of simple linear regression model, univariate regression model, well, we need to have a Z in the first column here corresponding to beta 1 for Y1 and 0's everywhere else. In the second block we want to have a Z in the second off diagonal with 0's everywhere else and so forth. So this is just re-expressing everything in this notation.

But the notation is very nice because, at the end of the day we basically have a regression model like we had when we were doing our regression analysis. So all

the theory we have for specifying these models plays through with univariate regression. And one can go through this technical argument to show that the generalized least squares estimate is, in fact, the equivalent to the component-wise values. And that's very, very good.

Maximum likelihood estimation with these models. Well, we actually use this vectorized notation to define the likelihood function. And if these assumptions are made about the linear regression model, we basically have an n times m vector of dependent variable values, whereas your multivariate normal with mean given by x star beta star and then a covariance matrix epsilon. The covariance matrix of epsilon star is sigma star. Well, sigma star is In Kronecker product sigma.

So if you go through the math of this, everything matches up in terms of what the assumptions are. And the conditional probability density function of this data is the usual functions of log normal or of a normal sample. So we have unknown parameters beta star sigma, which are equal to the joint density of this normal linear regression model.

So this corresponds to what we had before in our regression analysis. We just had this more complicated definition of the independent variables matrix X star. And a more complicated definition of our variance/covariance matrix sigma star. But the log likelihood function ends up being equal to a term proportional to the log of the determinant of our sigma matrix and minus one half q of beta sigma, where q of beta sigma is the least squares criterion for each of the component models summed up.

So the component-wise maximum likelihood estimation is-- for the underlying parameters, is the same as the large one. And in terms of estimating the covariance matrix, there's a notion called the concentrated log likelihood, which comes into play in models with many parameters. In this model, we have unknown parameters-- our regression parameters beta and our covariance matrix for the innovations sigma.

It turns out that our estimate of the regression parameter beta is independent, doesn't depend-- not statistically independent-- but does not depend on the value of

the covariance matrix sigma. So whatever sigma is, we have the same maximum likelihood estimate for the betas. So we can consider the log likelihood setting the beta parameter equal to its maximum likelihood estimate. And then we have a function that just depends on the data and the unknown parameter sigma.

So that's a concentrated likelihood function that needs to be maximized. And the maximization of the log of a determinant of a matrix minus n over 2, the trace of that matrix times an estimate of it, that has been solved. It's a bit involved. But if you're interested in the mathematics for how that's actually solved and how you take derivatives of determinants and so forth, there's a paper by Anderson and Olkin that goes through all the details of that that you can Google on the web.

Finally, let's see. There's-- well, not finally. There's model selection criteria that can be applied. These have been applied before for regression models for univariate time series model, the Akaike Information Criterion, the Bayes Information Criterion, Hannan-Quinn Criterion.

These definitions are all consistent with the other definitions. They basically take the likelihood function and you try to maximize that plus a penalty for the number of unknown parameters. And that's given here.

OK, then the last section goes through an asymptotic distribution of least squares estimates. And I'll let you read that on your own.

Let's see. For this lecture I put together an example of fitting vector auto regressions with some macroeconomic variables. And I just wanted to point that out to you. So let me go to this document here. What have we got here?

All right. Well, OK. Modeling macroeconomic time series is an important topic. It's what sort of central bankers do. They want to understand what factors are affecting the economy in terms of growth, inflation, unemployment. And what's the impact of interest rate policies.

There are some really important papers by Robert Lederman and Christopher Sims dealing with fitting vector auto regression models to a macroeconomic time series.

And actually, the framework within which they specified these models was a Bayesian framework, which is an extension of the maximum likelihood method where you'll incorporate reasonable sort prior assumptions about what the parameters ought to be.

But in this note, I sort of basically go through collecting various macroeconomic variables directly off the web using the package r. All this stuff is-- these are data that you can get your hands on. Here's the unemployment rate from January 1946 up through this past month. Anyone can see how that's varied between much less than 4% to over 10%, as it was recently.

And there's also the Fed funds rate, which is one of the key variables that the Federal Reserve Open Market Committee controls, or I should say controlled in the past, to try and affect the economy. Now that value of that rate is set almost at zero and other means are applied to have an impact on economic growth the economic situation of the market-- of the economy, rather.

Let's see. There's also-- anyway, a bunch of other variables. CPI, which is a measure of inflation. What this note goes through is the specification of vector auto regression models for these series. And I use just a small set of cases. I look at unemployment rate, federal funds, and the CPI, which is a measure of inflation.

And there's-- if one goes through, there are multivariate versions of the autocorrelation function, as given on the top right panel here, between these variables.

And one can also do the partial autocorrelation function. You'll recall that autocorrelation functions and partial autocorrelation functions are related to what kind of-- or help us understand what kind of order ARMA processes might be appropriate for univariate series. For multivariate series, then there are basically cross lags between variables that are important, and these can call be captured with vector auto regression models.

So this goes through and shows how these things are correlated with themselves.

And let's see. At the end of this note, there are some impulse response functions graphed, which are looking at what is the impact of an innovation in one of the components of the multivariate time series. So like if Fed funds were to be increased by a certain value, what would the likely impact be on the unemployment rate? Or on GNP? Basically, the production level of the economy.

And this looks at-- let's see. Well, actually here we're looking at the impulse function. You can look at the impulse function of innovations on any of the component variables on all the others. And in this case, on the left panel here is-- it shows what happens when unemployment has a spike up, or unit spike. A unit impulse up.

Well, this second panel shows what's likely to happen to the Fed funds rate. It turns out that's likely to go down. And that sort of is indicating-- it's sort of reflecting what, historically, was the policy of the Fed to basically reduce interest rates if unemployment was rising.

And then-- so anyway, these impulse response functions correspond to essentially those innovation terms on the Wold decomposition. And why are these important? Well, this indicates a connection, basically, between that sort of moving average representation and these time series models. And the way these graphs are generated is by essentially finding the Wold decomposition and then incorporating that into these values.

So-- OK, we'll finish there for today.