Business 35905 **John H. Cochrane**

Problem Set 4

The point of this probem is to think through agents that have more information than we do, and the point and nature of state-space models.

1. To warm up, consider the state space model,

$$\begin{aligned} x_{t+1} &= \phi x_t + \varepsilon^x_{t+1} \\ r_{t+1} &= x_t + \varepsilon^r_{t+1} \\ \sigma^2_x &\equiv \sigma^2(\varepsilon^x_{t+1}); \sigma^2_r \equiv \sigma^2(\varepsilon^r_{t+1}); \sigma_{xr} \equiv cov(\varepsilon^x_{t+1}, \varepsilon^r_{t+1}) \end{aligned} \tag{1}$$

We allow arbitrary correlation between ε^x and ε^r.

a) By matching autocorrelations, show that the univariate (Wold representation) return process is an ARMA(1,1) $(1-\phi L)r_{t+1} = (1-\theta L)v^r_{t+1}$ where $v^r_{t+1} \equiv r_{t+1} - E(r_{t+1}|I_t)$; $I_t = \{r_t, r_{t-1}, r_{t-2}, ...\}$. Find an equation which you can solve for θ. (You can leave your answer in the form $\theta/(1+\theta^2) = ...$ The solution is a nasty quadratic which we will only do numerically.) You do not need to find v^r_{t+1}, though you should see how to do it if you had to. (There will be two equations in the two unknowns $\theta, \sigma^2(v^r)$)

b) Find the autoregressive Wold representation of r_t, i.e. what function of $\{r_t, r_{t-1}, r_{t-2}...\}$ you would optimally use to forecast r_t if you could not see x_t. You're looking for $r_{t+1} = \sum_{j=0}^{\infty}(terms_j)r_{t-j} + v^r_{t+1}$. In contrast to the simple state-space representation, in which you would forecast using $r_{t+1} = x_t + \varepsilon^r_{t+1}$ using only current x_t, you should see that long averages or r_t will be useful. However, those long lags have a particular structure, so that estimating them by an unconstrained autoregression would not be a good idea.

c) Let's compete the circle of ideas. Suppose we had started with a univariate $ARMA(1,1)$ representation

$$(1-\phi L)r_{t+1} = (1-\theta L)v^r_{t+1}. \tag{2}$$

Now show you could construct an equivalent "state space" representation of this process. Construct

$$\hat{x}_t = E(r_{t+1}|I_t)$$

(don't forget, $I_t \equiv \{r_t, r_{t-1}, r_{t-2}, ...\}$). Show that you can represent (2) as

$$\begin{aligned} \hat{x}_{t+1} &= \phi\hat{x}_t + k \times v^r_{t+1} \\ r_{t+1} &= \hat{x}_t + v^r_{t+1} \end{aligned} \tag{3}$$

You will also have to find the value of k.

Equations (1) and (3) look similar but they are very different! Of course the errors are perfectly correlated in (3). They are also different errors, innovations relative to a much smaller information set. For example, recall in lecture that we found our "standard" VAR with $\phi = \rho$ implied a pure iid return process, i.e. $k = 0$.

d) Check everything is ok by starting with your answer to c, and deriving the univariate representation that it implies. Do this directly, by taking $(1-\phi L)r_{t+1}$, and by using the result in a to verifying that you get the same θ that you started with.

d)We're here to learn how to get back and forth from "state space" to Wold representations. Our beginning and ending points were very nice: We started with

$$\begin{aligned} x_{t+1} &= \phi x_t + \varepsilon^x_{t+1} \\ r_{t+1} &= x_t + \varepsilon^r_{t+1} \end{aligned} \tag{4}$$

We defined $\hat{x}_t \equiv E(r_{t+1}|I_t),\ v^r_{t+1} \equiv r_{t+1} - E(r_{t+1}|I_t)$ and ended up with

$$\begin{aligned} \hat{x}_{t+1} &= \phi\hat{x}_t + k \times v^r_{t+1} \\ r_{t+1} &= \hat{x}_t + v^r_{t+1}. \end{aligned} \tag{5}$$

However, the path was torturous – We matched autocorrelation functions, created an ARMA(1,1) representation, noticed it had a nice AR representation, then defined $\hat{x}_t$ from the right side of that AR representation. It sure looks like we can get from (4) to (5) directly. We can, and this will make our multivariate task easier.

Start from (4).By definition,

$$E\left(x_{t+1}|I_{t+1}\right) = E\left(x_{t+1}|I_t\right) + \left\{E\left(x_{t+1}|I_{t+1}\right) - E\left(x_{t+1}|I_t\right)\right\}$$

Substitute from (4) on the right side to conclude that $\hat{x}_{t+1} = \phi\hat{x}_t + k \times v^r_{t+1}$ for some k. (Note it is not true that $k \times v^r_{t+1} = E(\varepsilon^x_{t+1}|I_{t+1})$.) Showing $r_{t+1} = \hat{x}_t + v^r_{t+1}$ should be easy. You've shown that *You can write the Wold representation in a state-space form that looks just like the structural representation, except for the errors.*

This argument won't get you the value of k, since that depends on the covariance structure of the ε. I can't think of a better way to do that than match autocorrelations between (4) and (5). Since you already have don't that and know θ, you can quickly deduce the value of k in (5) needed to produce the right θ. If you can think of a better way to find k, let me know!

2. Now, the real question. Last week I hope I persuaded you to ignore anything past an AR(1), at least with only lags of dp, r, and Δd. This week, I want to persuade you that we might have been wrong. At least for the purposes of forecasting one year returns and dividend growth, we might well be able to do better!

We know that agents certainly have more information about dividend growth and returns than we see in our simple VAR. I want to use a simple parameterization of agents' information to motivate more terms of the VAR, and suggest a structure to higher terms of the VAR. (This is a simplified version of Ralph Koijen and Jules Van Binsbergen's "Predictive Regressions," and my "State Space vs. VAR models for stock returns." Denis Chaves came up with this special case of correlated errors that makes the calculations easy.)

Suppose that expected returns $\mu_t = E(r_{t+1}|\text{investor information}_t)$ and expected dividend growth $g_t = E(\Delta d_{t+1}|\text{investor information}_t)$ both follow $AR(1)$ processes,

$$\begin{aligned} \mu_t &= \phi\mu_{t-1} + \sigma_\mu v_t \\ g_t &= \theta g_{t-1} + \sigma_g v_t. \end{aligned}$$

This is an obvious generalization to two variables of the expected-return latent variable models of the notes.

I want to think of θ as being substantially less than ϕ, maybe 0.4. My guess about the world is that we can forecast dividend growth for a year or two, but not much beyond that. But we'll let the data speak.

I specify the same shock v_t in each equation. If you allow different shocks, then you cannot recover μ_t and g_t from the Wold representation of $\{dp, r, \Delta d\}$, so you have to do a fun job of projecting down on that space to figure out what VAR to run. As you'll see, by setting the shocks equal to each other in this way, you can still recover v_t from dp,as we could in the simple state-space models with only one variable, and then we can derive the implied VAR representation. A positive σ_g means expected returns

and dividend growth move in the same direction, which is my intuition; they offset so dp doesn't move; both expected returns and expected dividend growth rise in recessions. However, negative values could be allowed here.

You recognize these processes as models that imply ARMA(1,1) for r_{t+1} and Δd_{t+1}, which suggests using some long moving averages of r_t and Δd_t to improve forecasts, i.e. that the VAR could be improved by adding such long moving averages or restricted coefficients. However, we'll see that moving averages of dp_t are even better.

a) Using the present value identity, find dp_t in terms of μ_t and g_t.

b) Now, you're typically stuck here; you can't infer μ_t and g_t separately from $\{dp_t, r_t, g_t\}$, so you have a fun projection job to figure out $E_t(r_{t+1}|dp_t, r_t, \Delta d_t, dp_{t-1}, ...)$ and so forth. However, since the shocks are the same in the μ and g equations, you can express dp_t in terms of v_t, which become dp_t regression errors. Show that dp_t follows an ARMA(2,1) in terms of the v_t shocks. ($\phi < 1$ and $\theta < 1$ with $\sigma_\mu \geq 0$ and $\sigma_g \geq 0$ doesn't always mean that the MA coefficient is less than one in absolute value, but convince yourself that it is for reasonable numbers. Note $\frac{\sigma_\mu}{1-\rho\phi}$ is the contribution of expected return shocks and $\frac{\sigma_g}{1-\rho\theta}$ the contribution of expected dividend shocks to dp changes.) Denote the MA root by ξ, i.e show that dp_t has the Wold representation

$$(1-\phi L)(1-\theta L)dp_t = k\,(1-\xi L)\,v_t$$

where k and ξ are expressions you will find in terms of the original parameters.

c) Now, since you can recover v_t from the history of $\{dp_t\}$, you can recover μ_t and g_t from the history of dp_t! Find these expressions, and hence find what the return and dividend growth regressions should look like. You're looking for expressions of the form

$$\begin{aligned} r_{t+1} &= (\text{expression})\frac{1-aL}{1-\xi L}dp_t + \varepsilon^r_{t+1} \\ \Delta d_{t+1} &= (\text{expression})\frac{1-bL}{1-\xi L}dp_t + \varepsilon^d_{t+1} \end{aligned}$$

(I gave you a hint with the lag polynomials, but you have to get there, and find a and b!) We're only looking for the lag polynomials. You can express the return forecast with our old fried $(1-\rho\phi)$ as a leading term but then new things on the right hand side. (We can also find ε^d and ε^r in terms of v, but I didn't do that.)

d) Now have a prediction for a VAR, which is an ARMA(2,1) for DP and regression coefficients for r_{t+1} and Δd_{t+1} on dp_t that have the ARMA(1,1) pattern.

I don't want to get too fancy in estimating ARMA processes for a problem set. Since

$$\frac{1-aL}{1-\xi L} = 1 - \left(\frac{a-\xi}{1-\xi L}\right)L$$

we can summarize our predictions for the lag polynomials that *a moving average of* $dp_{t-1} + \xi dp_{t-1} + \xi^2 dp_{t-2} + ...$ *should help* dp_t *to forecast returns* r_{t+1} *and dividend growth* Δd_{t+1}.

And this is the big point. A "simple" AR(1) in the state-space implies patterns in the VAR representation. We looked at extra lags of dp last week and concluded there isn't much there. But we were looking at each one individually, not looking for patterns, for coefficients that might be *jointly* helpful.

Ok, to evaluate this write a program that takes a guess for ξ, forms a moving average $dp_{t-1} + \xi dp_{t-1} + \xi^2 dp_{t-2} + .$ and adds that to the right hand side, i.e.

$$\begin{aligned} r_{t+1} &= a + b_r dp_t + c_r\left(dp_{t-1} + \xi dp_{t-2} + \xi^2 dp_{t-3} + ..\right) + \varepsilon^r_{t+1} \\ \Delta d_{t+1} &= a + b_d dp_t + c_d\left(dp_{t-1} + \xi dp_{t-2} + \xi^2 dp_{t-3} + ..\right) + \varepsilon^d_{t+1} \end{aligned}$$

I used real (not implied) dividend growth here so I could make sure I was really forecasting something.

i) Report (at least) coefficients, t statistics, and R^2. Search a bit for a ξ that improves the R^2 the most (this is called "manual nonlinear least squares without t statistics for ξ"), and pay attention to the Δd results especially (if it belongs in the Δd equation, it also belongs in the r equation, by an identity!) I found a quite low value of ξ did the trick. I used the sample r_{t+1} starting in $t+1 = 1947$, and I know the results look decent there. Dividends are much less predictable before WWII, and any quick look at the data shows a big change in dividend behavior there. Include a plot of actual and predicted r_{t+1} and actual and predicted Δd_{t+1}.

Note: I "orthogonalized" my right hand variables, running instead

$$\begin{aligned} r_{t+1} &= a + b_r dp_t + c_r \left[dp_t - \frac{1}{1+\xi+\xi^2+..} \left(dp_{t-1} + \xi dp_{t-2} + \xi^2 dp_{t-3} + .. \right) \right] + \varepsilon^r_{t+1} \\ \Delta d_{t+1} &= a + b_d dp_t + c_d \left[dp_t - \frac{1}{1+\xi+\xi^2+..} \left(dp_{t-1} + \xi dp_{t-2} + \xi^2 dp_{t-3} + .. \right) \right] + \varepsilon^d_{t+1} \end{aligned}$$

This has the advantage that the b_r and b_d coefficients are almost the same as before. It has no effect on the magnitude or significance of c_r, c_d. It means the extra variable can be nicely interpreted as "*recent increase in dp.*" Rather than just look at the overall level of valuation (dp), we also ask whether there are any big recent *changes* in valuation. Recent changes help us to isolate the faster-moving component, i.e. isolate μ changes and g changes. You saw last week that $dp_t - dp_{t-1}$ helped to forecast, and this whole structure gives a way to interpret that fact

Make sure your coefficients c_r, c_d, c_{dp} obey the correct approximate identity – trace any improvement in forecast ability across the other two variables.

ii) Include plots of actual and forecast returns and actual and forecast dividend growth over time, contrasting the bivariate results with the original that only uses dp_t. Is the improvement in performance consistent and believable over the sample, or just one data point? Do you see the improvement in dividend growth forecast mirrored in improvement in return forecast?

e) Are we really doing all that much? Compare your forecasts with a simple lag of dp, $r_{t+1} = a + b dp_t + c(dp_t - dp_{t-1}) + \varepsilon^r_{t+1}$. Is the moving average capturing a lot of information you can't see with just one lag?

3. Let's replicate Goyal and Welch, and ask how the dividend yield forecast works "out of sample." Let's also look into the GW criticism a little deeper.

Start at the 20th year of the sample. Run a regression

$$r_t = a + b(d-p)_{t-1} + \varepsilon_t \; t = 1, 2, ..20$$

Use this regression to forecast the 21st year,

$$\hat{r}_{21} = \hat{a} + \hat{b}(d-p)_{20}.$$

Compute the forecast error in the 21th year as the difference,

$$e_{21} = r_{21} - \hat{r}_{21}.$$

Now, do the same thing at year 21. Run the regression from $t = 1$ to $t = 21$

$$r_t = a + b(d-p)_{t-1} + \varepsilon_t \; t = 1, 2, ..21$$

Use this regression to forecast the 22st year,

$$\hat{r}_{22} = \hat{a} + \hat{b}(d-p)_{21}$$

Compute the forecast error in the 22nd year as the difference,

$$e_{22} = r_{22} - \hat{r}_{22}.$$

Keep going this way through the end of the sample. At each date you are only using data up to that date to forecast the return. Compute the root mean squared error

$$rmse = \sqrt{\frac{1}{T-20}\sum_{t=21}^{T} e_t^2}$$

of this strategy.

Now, suppose you just took the sample mean as your forecast. Starting at year 20, compute

$$\bar{r}_{21} = \frac{1}{20}\sum_{t=1}^{20} r_t$$

and compute the error

$$e = r_{21} - \bar{r}_{21}$$

Keep going; compute

$$\bar{r}_{22} = \frac{1}{21}\sum_{t=1}^{21} r_t$$

and compute the error

$$e = r_{22} - \bar{r}_{22}$$

Compute *this* root mean squared error as well.

a) Make a plot of the root mean squared errors over time, i.e. plot

$$rmse_t = \sqrt{\frac{1}{t-20}\sum_{j=21}^{t} e_j^2}$$

for each technique as a function of t. This lets you see if there is ever a time in which one does better than the other. GW plot the difference between the two errors. Plot that too, but I find the levels interesting. The levels give you a good sense of "is the difference significant?"

b) Plot the forecast $E_t R_{t+1}$ and the coefficients, $\hat{\mu}_t$, and $\hat{a}_t, \hat{b}_t$. This will let you see how dp forecasts go wrong. (Is uncertainty about a or about b the bigger issue?)

Bottom line? I agree with their result. As in the dog that didn't bark, we expect this from the null!