Discount Rates

John H. Cochrane*

December 27, 2010

Abstract

I argue that characterizing discount rate variation over time and across assets has replaced informational efficiency as the central organizing question of asset pricing research.

I survey the facts: in the last 40 years we have learned that discount rates vary dramatically. Most views of the world changed 100%: we thought 100% of the variation in market dividend yields was due to variation in expected cashflows; now we know 100% is due to variation in discount rates. We thought 100% of the cross-sectional variation in expected returns came from the CAPM, now we think that's about zero and a zoo of new factors describes the cross section. I show how time-series, cross-section, regression, and portfolio approaches are really the same, and think about how the empirical project can achieve a needed unification.

I survey theories. I break discount-rate theories into categories. Frictionless theories include macro, both consumption and investment-focused, and behavioral approaches. Theoreis that focus on frictions include segemented markets, institutional frictions and liquidity.

I survey applications. The simple facts of large variation in risk premiums has and will continue to dramatically change finance applications inluding portfolio formation, accounting, and corporate finance, such as cost of capital, capital structure, and compensation.

*University of Chicago, Booth School of Business, and NBER. 5807 S. Woodlawn Ave. Chicago IL 60637. john.cochrane@chicagobooth.edu. 773 702 3059. http://faculty.chicagobooth.edu/john.cochrane/research/Papers/. I gratefully acknowledge research support from CRSP and outstanding research assistance from Yoshio Nozawa.

Contents

Part I

Introduction

Prices should equal expected discounted value. In 1970, Gene Fama argued that the expected part, "testing market efficiency," provided the framework for organizing finance research. My theme is that finance research today is really all about the "discounted" part: *How* risk premia vary over time and across assets, *why* they do so, and how to *apply* this understanding.

"Efficiency" isn't wrong, it just doesn't describe the focus of what we do. When we see information, it is quickly incorporated into market prices. When we see anomalies – and we see many – informational frictions aren't an interesting alternative. Anomalous discount rate variation is. The "model of market equilibrium," dimly glimpsed in 1970, is now the central question, much better understood, and much more elaborated.

I start with a review of facts and views of the empirical project. I'll attempt a categorization of theory, something analogous to Fama's "weak" "semi-strong" and "strong" forms of efficiency. I will then point to some of the many areas of application that are ripe for change when they reflect discount rate variation.

Part II

Facts

Our empirical view of the world has changed 100% and more since the early 1970s. In the early 1970s, it seemed that expected returns were constant over time, and the CAPM accounted well for their variation across assets. Now we know that expected return variation over time and across assets is much larger than anyone anticipated. Asset valuations move on discount rate news far more than on news of expected cashflows. The CAPM explains nearly none of the cross-sectional variation in stock average returns. Such variation is related to a bewildering variety of new factors instead.

I do not present a congratulatory review of an accomplished agenda. This empirical investigation has only begun. The work of just *documenting* the time-series and cross-sectional pattern of expected returns, its correlation with the covariance structure of returns, and its implications for *prices*, which I emphasize, is only really beginning.

Really, past achievements did not constitute supplying answers. The big achievements have been defining the right questions. The point of my review is to start thinking about the right questions we need to ask to deepen our understanding of discount rate variation.

1 Time Series

1.1 Simple DP regression

Table 1 forecasts returns with dividend yields in the manner of Fama and French (1987). It doesn't seem remarkable. The t statistic is "significant," though contaminated by fishing and small sample biases. The R^2 is low. R^2 is also a measure of potential trading profits. The maximum unconditional sharpe ratio s attainable from market-timing is related to the maxmum sharpe ratio s_0 available from static investments by $s^2 \approx s_0^2 + R^2$ (Cochrane (1999).) Since the market Sharpe ratio is about $s_0 = 0.5$, R^2 of 0.06 or so indicates little benefit from market timing. For these reasons statistically significant return forecastability was long regarded as an economically unimportant nuisance.

	$R^e_{t\to t+k} = a + b\frac{D_t}{P_t} + \varepsilon_{t+k}$				
Horizon k	b	t(b)	R^2	$\sigma\left[E_t(R^e)\right]$	$\frac{\sigma[E_t(R^e)]}{E(R^e)}$
1 month	0.3	(2.9)	0.01	0.45	0.85
1 year	3.8	(2.6)	0.09	5.46	0.76
5 years	20.6	(3.4)	0.28	29.3	0.62

Table 1. Return forecasting regressions using the dividend yield. CRSP value weighted return 1947-2009.

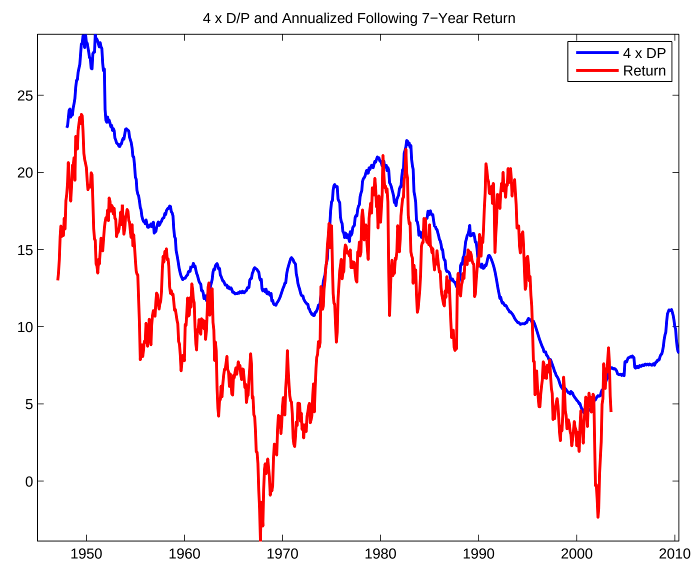

Figure 1: Dividend yield (multiplied by 4) and following 7 year return. CRSP VW market index.

In fact, the point estimates of this regression have huge economic significance.

First, the magnitude of the coefficient is huge. $b = 3.8$ means that a one percentage point rise in dividend yield gives a *four* percentage point increase in return. With no change in prices, a one percentage point change in dividend yield means one percentage point more return. Apparently, price move in the "wrong" direction.

Second, the $\sigma\left(b \times D/P_t\right) = \sigma\left[E_t(R_{t+1})\right] = 5.46$ percentage point variation in expected returns is huge! A 6% equity premium was already a "puzzle." Now expected returns *vary* by as much as this *level*.

The trouble here is that R^2 is not really the right measure of "economic" significance. Ex-post return variance is a poor standard for measuring the variation in expected returns. There will always be news, and R^2 can be low even with unthinkable amounts of expected return variation. Thinking about this regression in terms of "how well can you forecast returns" simply misses the point. In the last column I divide variation in expected returns by the level of expected returns as a way to capture this idea.

Third, the slope coefficients and R^2 rise with horizon. I dramatize this fact in Figure 1 which plots each year's dividend yield along with the *subsequent* 7 years of returns. High prices, relative to dividends, have reliably led to 7 years of poor returns, and low prices have led to high returns. This graph, I think, really captures the basic fact.

1.1.1 Present values, volatility, bubbles, and long-horizons

Long horizons are particularly interesting because they tie predictability to volatility, "bubbles," and the nature of price movements. To address these questions, in Table 2 I regress long-horizon returns on dividend yields, both directly and as implied by a very simplified VAR.

	$\sum_{j=1}^{k}\rho^{j-1}r_{t+j}$	$\sum_{j=1}^{k}\rho^{j-1}\Delta d_{t+j}$	$\rho^k dp_{t+k}$
Direct, $k=15$	1.01	-0.11	-0.11
VAR, $k=15$	1.05	0.27	0.22
VAR, $k=\infty$	1.35	0.35	0.00

Table 2. Left panel: Long-horizon return regression coefficients.

$$y_{t+k} = a + b^{(k)} dp_t + \varepsilon_{t+k}$$

Annual data 1947-2009. VAR calculations use the coefficients from Table 3.

	dp_t	t	R^2	$\sigma(E_t(y_{t+1}))\%$	$\frac{\sigma(E_t(y_{t+1}))}{E(y_{t+1})}$
r_{t+1}	0.13	(2.61)	010	5.54	0.54
Δd_{t+1}	0.04	(0.92)	0.02	1.83	0.33
dp_{t+1}	0.94	(23.8)	0.91		

Table 3. VAR for long-run calculations. Annual data 1947-2009.

To make the link between returns and prices, I use the Campbell-Shiller (1988) linearized present value identity,

$$dp_t \approx \sum_{j=1}^{k} \rho^{j-1}\left(r_{t+j} - \Delta d_{t+j}\right) + \rho^k \left(d_{t+k} - p_{t+k}\right). \tag{1}$$

This identity is simply a forward iteration of a loglinearization of one period log returns,

$$\begin{aligned} R_{t+1} &= \frac{\left(1+\frac{P_{t+1}}{D_{t+1}}\right)}{\frac{P_t}{D_t}} \frac{D_{t+1}}{D_t} \\ r_{t+1} &\approx -\rho dp_{t+1} + dp_t + \Delta d_{t+1}. \end{aligned} \tag{2}$$

Small letters are logs of big letters, and I ignore constants. See the appendix for details.

As a result of (1), the long-run regression coefficients from Table 2 must add up to one,

$$1 \approx b_r^{(k)} - b_d^{(k)} + \rho^k \phi^{(k)}. \tag{3}$$

(Regress both sides of (1) on dp_t. Again, see the appendix for details.)

This equation says that high prices must forecast low discount rates, high expected cashflows, or a "rational bubble" of ever higher prices. Equivalently, multiplying both sides of (3) by $var(dp_t)$, we obtain

$$var\left(dp_t\right) \approx cov\left[dp_t, \sum_{j=1}^{k} \rho^{j-1} r_{t+j}\right] - cov\left[dp_t, \sum_{j=1}^{k} \rho^{j-1} \Delta d_{t+j}\right] + \rho^k cov\left(dp_t, dp_{t+k}\right) \tag{4}$$

If the world is i. i. d. , then price-dividend ratios must be constant; variation in price dividend ratios must then come from discount rates, dividend growth, or "rational bubbles." The regression coefficients can be read as the fractions of dividend yield variation attributed to each source.

The empirical question is, which is it? The long-run return coefficients are all slightly larger than 1.0. The dividend-growth forecasts are insignificant, and the point estimates go the "wrong" way – high prices (low dividend yields) signal *low* future dividend growth.

Thus, Table 2 shows us that *all price-dividend ratio volatility corresponds to variation in expected returns. None corresponds to variation in expected dividend growth, and none to "rational bubbles" in which prices are high just on the expectation of even higher future prices with constant expected returns.*

Figure 2 gives another view of the same facts (this is an interpretive view of the impulse-response function to a dp shock with no change in current dividends). Suppose you see a sharp decline in prices relative to dividends, as illustrated. What would you forecast? In the classic view, you would forecast no change in future prices, but a decline in future dividends. Our current view is exactly the reverse.

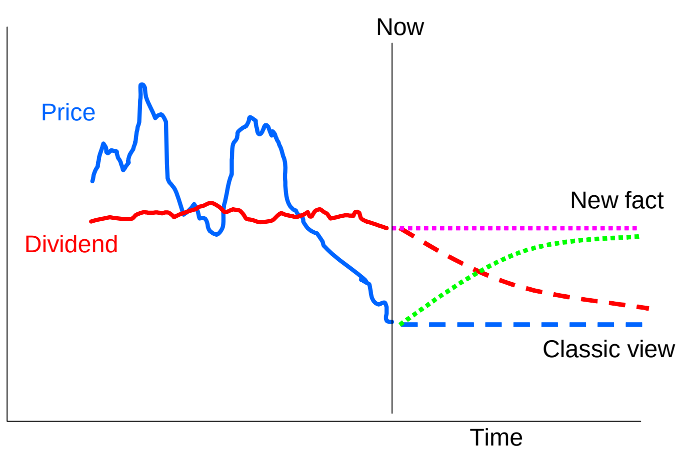

Figure 2: Classic and new view of price-divided ratios.

This is the true meaning of return forecastability, and measure of "how big" it is. Return forecastability is "just enough" to explain the volatility of prices.

And they are a revolution in our understanding of the facts. In the 1970s, we would have guessed exactly the opposite pattern. The (point estimate of the) facts changed form 100% dividend growth, to 100% returns. This really is a big fact!

In this way, the real point of the return-forecasting regression is to analyze the information in prices, not to characterize the conditional mean of returns. Fama and Bliss (1987) was titled "the information in forward rates" not "how to forecast bond returns." Ideally, we'd run the regression with *dp* on the left, and expected return or dividend growth on the right, something like

$$dp_t = (\text{expected return})_t + (\text{expected dividend growth})_t + \text{error}.$$

In these regressions, we are essentially using actual returns as a proxy for expected returns. We run the regression "backwards" because the resulting "measurement error" is a forecast error and orthogonal to dp_t, and the right hand variables of regressions should be orthogonal to error terms. But that fact does *not* mean we assume right hand variables "cause" left hand variables, or that the main interest of the regression is in "explaining" the left hand variable.

Agents have information we will never observe, and prices reflect that information. When something unobservable to us drives down expected returns, it drives up prices, and then we can see it. The point of these regressions is to see one slice of agents' information.

Be careful about the fact. In this setup, expected returns account for almost all variation in *price-dividend ratios*. That does not mean they account for almost all variation in *prices* or *returns.* From (2)

$$r_{t+1} \approx -\rho dp_{t+1} + dp_t + \Delta d_{t+1}.$$

and we can also write

$$\Delta p_{t+1} = -dp_{t+1} + dp_t + \Delta d_{t+1}.$$

Now dividend growth has a substantial, roughly 11% per year volatility. Thus, even if *dividend yield* variation all corresponds to expected return variation, and even if dividend growth is perfectly unforecastable, as much as half of actual *returns* and *price-growth* variation will correspond to dividend growth. This verbal slippage has caused a lot of confusion.

1.2 A pervasive phenomenon

This pattern of predictability is pervasive across markets. Price variation that "should" forecast one thing (dividend growth) does not do so. Instead, prices forecast expected returns. A 1/0 prior became 0/1.

- *Bonds/term structure.* A rising yield curve signals better one-year return for long-term bonds, not higher future interest rates. A one percentage point rise in the n year forward rate over the spot rate signals a one percentage point risk in the expected one year return on n year bonds. Forward-spot spreads do not forecast one-year changes in spot rates (Fama and Bliss (1987); Campbell and Shiller (1991)).

  Macroeconomists have not digested this lesson. The Federal Reserve looks at long-term treasuries and the TIPS spread to gauge inflation expectations, and at Federal Funds futures to gauge policy expectations (Piazzesi and Swanson (2008)) . These spreads forecast returns, not the intended objects.

- *Bonds/credit spreads.* Credit spreads are higher than default probabilities, and variation in credit spreads over time largely signals changes in expected returns, not changes in probabilities of default. (Berndt and Duffie (2010).)

- *Foreign exchange.*High foreign interest rates relative to domestic rates signal a higher return on foreign bonds, not a devaluation–the "carry trade." (Hansen and Hodrick (1980), Fama (1984).)

- *Sovereign debt.* High levels of sovereign or foreign debt do not signal high future government surpluses or trade surpluses. They signal low expected returns. (Gourinchas and Rey (2007).)

- *Houses.* High price/rent ratios signal low returns, not rising rents or prices that rise forever.

I made a graph and ran a regression, just because houses are in the news so much lately. Figure 3 presents prices and rents, and Table 4 presents the regression.

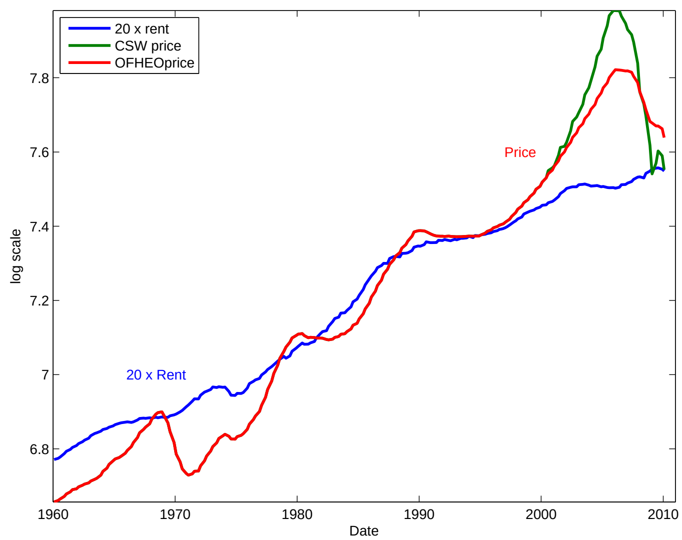

Figure 3: House pirces and rents. Data from http://www.lincolninst.edu/subcenters/land-values/rent-price-ratio.asp

Swings in the price-rent ratio resolve themselves by price changes rather than by rent changes. Regressions of housing returns (price plus rent) and rents on the rent/price ratio are almost exactly the same as for stocks in Table 3. (Housing returns are much more serially correlated than stock returns, so it's dangerous to take the analogy too far, however.)

	b	t	R^2
r_{t+1}	0.12	(2.52)	0.15
Δd_{t+1}	0.03	(2.22)	0.07
dp_{t+1}	0.90	(16.2)	0.90

Table 4 Regressions of log annual housing returns r_{t+1}, log rent growth Δd_{t+1} and log rent/price ratio dp_{t+1} on the price/rent ratio dp_t, $x_{t+1} = a + b \times dp_t + \varepsilon_{t+1}$ 1960:1-2010:1.

There is a strong common element and a strong business cycle association of these forecasts. (Fama and French (1989) among many others.) Low expected returns and high prices hold in "good times," when consumption, output and investment are strong, unemployment is low, interest rates are high, and the yield curve is flat. High expected returns and low prices hold in "bad times," when consumption, output and investment are low, unemployment is high, interest rates are low, and the yield curve is upward sloping.

1.3 Bubbles and volatility tests

These *facts* bring a good deal of structure to the argument over "bubbles." Many people refer to "bubbles" as matters of fact, e.g. "the housing bubble," or "the internet bubble." Alas, it's hard to find an operational *definition* of bubble, let alone a procedure for measuring one. "I wish I sold yesterday," "I don't understand why prices are so high" or "I can tell one when I see one" don't count. "Price higher than fundamentals" begs the question, "and how do you measure fundamentals?"

One possible definition of "bubble" is "prices that are high on expectations of subsequent higher prices; people buy just because they think they can sell to a greater fool." This is a "rational bubble," a violation of the transversality condition. In such a bubble expected returns are always the same, so higher valuations do not make it more likely to see a low return. As we saw above, the data speak strongly against this form of bubble. Higher valuations *do* correspond to lower returns, and "just enough" to fully "explain" price variation. Once we look at a 15 year horizon, high prices *do not* correspond to ever-higher prices.

To continue sensibly, then, a "bubble" must be defined as a high valuation, which corresponds to a discount rate that is "too low" in some sense. The *fact* is that high valuations correspond to low long-run returns, period. Now, we have something solid to talk about. Perhaps low risk premium in economic booms correspond to low macroeconomic risk premiums, and high risk premiums in recessions correspond to higher macroeconomic risk premiums. I survey those theories below. Perhaps the same variation in risk premiums corresponds to something "wrong" either in psychology, market frictions, monetary policy, etc. Good. At least we know what we're talking about. That's progress.

These regressions neatly sum up two decades of "volatility tests." Starting with Shiller (1981) we puzzled for two decades at why prices were so volatile. The essence of Shiller's plots is that high prices do not forecast higher dividend growth. Regressions (3) and (4) are exactly the most modern forms of these volatility tests, respecting the unit root in prices and dividends. And they make clear that volatility tests are *exactly* the same thing as long-run return forecasting regressions.

1.4 The multivariate challenge

A big challenge lies ahead here. We have a bunch of *univariate* regressions, on both left hand and right hand sides – one return on the left, and one variable on the right. We need to understand the *multivariate* counterparts. What happens when we put all the forecasting variables together on the right hand side? Which variables are really important? How are risk premiums connected across markets? Do the same variables forecast both stock and bond returns? In this extension we need to do more than just add variables; we need to think how multivariate information affects our understanding of price movements.

1.4.1 Multivariate challenge 1: understanding prices

Dozens of variables help the dividend yield to forecast returns. I use Lettau and Ludvigson's (2005) cay measure of a consumption/wealth ratio as a simple example, to explore how adding a right hand variable change return and dividend growth forecasts, and our consequent understanding of prices. Table 5 presents the regressions.

	Coefficients		t-statistics		Other statistics		
	dp_t	cay_t	dp_t	cay_t	R^2	$\sigma\left[E_t(y_{t+1})\right]\%$	$\frac{\sigma[E_t(y_{t+1})]}{E(y_{t+1})}$
r_{t+1}	0.12	0.071	(2.14)	(3.19)	0.26	8.99	0.91
Δd_{t+1}	0.024	0.025	(0.46)	(1.69)	0.05	2.80	0.12
dp_{t+1}	0.94	-0.047	(20.4)	(-3.05)	0.91		
cay_{t+1}	0.15	0.65	(0.63)	(5.95)	0.43		
$r^{lr} = \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}$	1.29	0.033					
$\Delta d^{lr} = \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}$	0.29	0.033					

Table 5. Forecasting regressions using dividend yield and consumption-wealth ratio, 1952-2009, annual data. Long-run coefficients are computed using a first-order VAR with dp_t and cay_t as state variables. Each regression includes a constant. The long-run regression coefficients are implied by the VAR. The appendix has additional details on the calculations.

Cay helps to forecast one-period returns. The t statistic is large, and it raises the variation of expected returns substantially. Cay only marginally helps to forecast dividend growth. (It works better in quarterly data.) Cay mean-reverts much more quickly than dp –0.65 rather than 0.94. In a graph of forecast vs. actual one year returns, Figure 4, we see that adding cay lets us forecast higher frequency "wiggles" while not much changing the "trend."

Consider by contrast the forecasts of long-run returns shown in Figure 5. The Figure contrasts the long-run return forecast from dividend yields, the long run return forecast from dividend yields with cay, and dividend yields themselves. Thus, we can see how dividend yields now break into long-run return vs. dividend growth forecasts, i.e. $dp_t = E_t r_t^{lr} - E_t \Delta d_t^{lr}$.

Though cay has a dramatic effect on *one-period* return r_{t+1} forecasts, it has almost no effect at all on *long-run* return $\sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}$forecasts. We can barely distinguish the univariate from the multivariate long-run forecasts, and essentially all price-dividend variation still comes from expected-return forecasts.

How can cay help to forecast one-year returns so much, but do so little to forecast long-run returns and long-run dividend growth? Looking at the Campbell-Shiller linearized present value formula,

$$dp_t = E\left[\sum_{j=1}^{\infty} \rho^{j-1} r_{t+j} | I_t\right] - E\left[\sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j} | I_t\right] \tag{5}$$

the answer is clear. Given dividend yields, a new variable which does not forecast long-run dividend growth can only raise the one-year return forecast if it lowers the forecast of returns further in the future, i.e. if it changes the term structure of risk premia. The appendix shows how this intuition applies to regression coefficients and the impulse response functions. In particular,

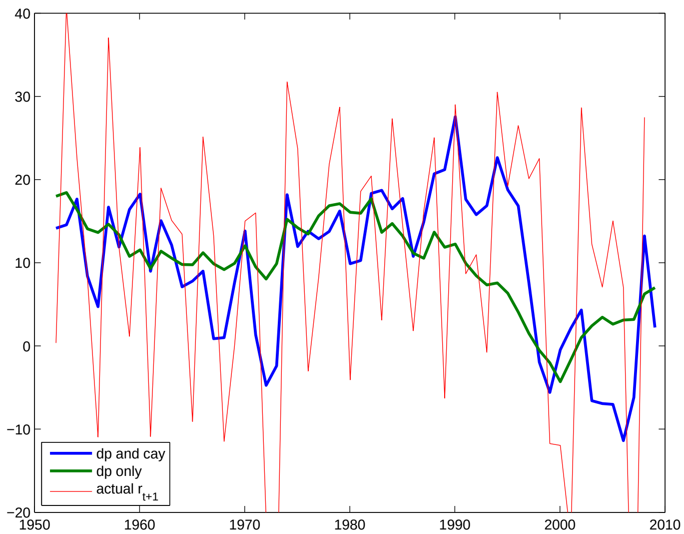

Figure 4: Fitted values of return-forecast regressions using dp only, and dp together with cay, along with the actual (ex post) return. Actual return r_{t+1} is graphed at time t along with its predictors.

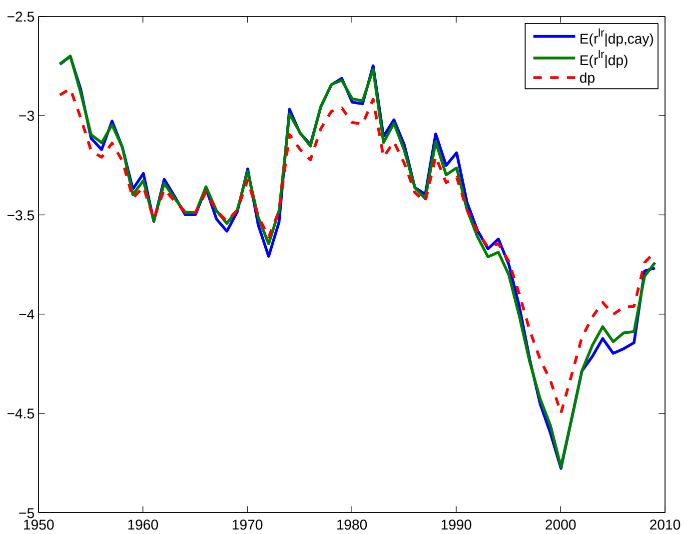

Figure 5: Dividend yield and expected long horizon returns, $E_t r_t^{lr} = E_t \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}$.

long-run responses to a dividend-yield shock must add up to one, while long run responses to other shocks must add up to zero.

$$\begin{aligned} 1 &= \sum_{j=1}^{\infty} \rho^{j-1} e_{dp \to r}^{(j)} - \sum_{j=1}^{\infty} \rho^{j-1} e_{dp \to \Delta d}^{(j)} \\ 0 &= \sum_{j=1}^{\infty} \rho^{j-1} e_{z \to r}^{(j)} - \sum_{j=1}^{\infty} \rho^{j-1} e_{z \to \Delta d}^{(j)}. \end{aligned}$$

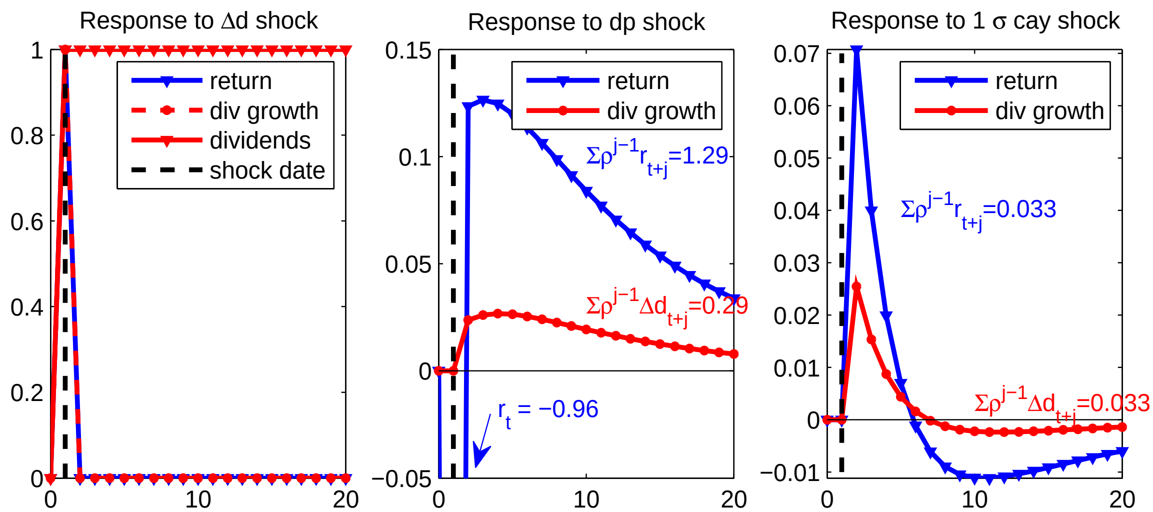

Figure 6: Response functions to dividend growth, dividend yield, and cay shocks. Calculations are based on the VAR of Table 5. Each shock changes the indicated variable without changing the others.

Figure 6 investigates the change in short-run and long-run forecastability and its implication for prices. I plot responses to a dividend growth shock, a dividend yield shock, and a cay shock, in each case setting movements of the other variables to zero, but including a contemporaneous return response to satisfy the return identity $r_{t+1} = \Delta d_{t+1} - \rho dp_{t+1} + dp_t$. (The left hand and middle panels are visually almost identical to the results from a bivariate system that ignores cay, so really cay only adds the rightmost panel.)

Don't be fooled by the "response" name, again we read causality backwards if at all. These plots answer the question, "what change in expectations corresponds to the given shock?"

The dividend growth shock corresponds to permanently higher expected dividends and prices, with no change in expected returns. It is a pure "expected cashflow" shock.

The dividend yield shock is essentially a pure discount rate shock, a rise in expected returns with little change in expected dividend growth.

The cay shock basically reflects a shift in expected returns from the distant future to the near future; with a small similar movement in the timing of a dividend growth forecast. We could label it *a shock to the term structure of risk premia,* a concept I think we will hear more from as we relate returns to prices. But the long-run expected return barely changes.

Other outcomes were possible. I had hoped for an increase in both long run return and dividend growth forecasts, which would have to be correlated. At the bottom of a recession, both risk premiums and real growth forecasts could be higher. It didn't turn out that way.

Here it seems that although *one* period return forecasts can be improved, the implications for stock *price* formation do no seem not much affected at all by extra variables, because the short-lived forecasting variable only changes the *term structure* of risk premia. Is this more generally true? I found the same result with the additional Goyal and Welch (2008) predictors, taken one at a time and with additional lags (see the appendix). But obviously, this is still more of a question than an answer. The point is, it's an important question.

1.4.2 More multivariate challenges: Right hand variables, factor structure, and covariance.

Dozens of additional variables forecast returns. Predictors that forecast *other* returns are an especially interesting category of additional variables. For example, the bond forecast factor in Cochrane and Piazzesi (2005) helps d/p a great deal to forecast stocks.

Figure 7 illustrates my sense of what we know and don't know. The red squares represent the predictability of each return from "its own" spread. The yellow squares represent variables which help to forecast returns. The question marks represent unknown possibilities.

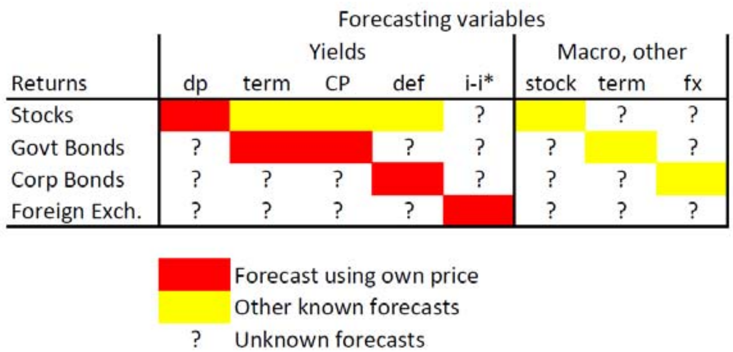

Figure 7: Schematic table of return forecast patterns. Colums: dp is the dividend yield, term is the term spread. CP is the Cochrane-Piazzesi (2005) tent shaped combination of forward rates. $i - i^*$ is the interest rate spread across countries. Macro other collects non-yield variables that forecast returns. Question marks represent unknown possibilities

First, we have to move past treating extra variables one or two at a time, and understand which of these variables are really important. Alas, huge multiple regressions is impossible. So the challenge is, how to answer the great multiple-regression question, without actually running huge multiple regressions?

Second, we want to understand the correlation of expected returns across assets. We know univariate forecasters are correlated with each other, most being high in recessions (Fama and French (1991) is the classic here.) Perhaps the multivariate structure will reveal greater correlation across assets. For example, suppose when we look at stocks and bonds, we find

$$\begin{aligned} r_{t+1}^{\text{stock}} &= a + 0.1 \times dp_t + 2.0 \times cp_t + \varepsilon_{t+1} \\ r_{t+1}^{\text{bond}} &= a + 0.05 \times dp_t + 1.0 \times cp_t + \varepsilon_{t+1} \end{aligned}$$

(These numbers are made up; this is just an example.) Then we would conclude that $E_t(r_{t+1})$ is perfectly correlated across stocks and bonds; there is a *one-factor model for expected returns.*.

My work on bond returns with Monika Piazzesi (Cochrane and Piazzesi (2005), (2008)) gives a blueprint that I hope will extend. Fama and Bliss (1987) had found that *each* bond's forward-spot spread forecast *that* bond's returns – red diagonals in Figure 7. We asked about the off-diagonals, by running multiple regressions. Surprise, we found that bond i's forward spread forecast bond j's returns, and regression coefficients were nearly exactly proportional across rows! There is a single factor that moves bond expected returns across all maturities.

Can we do something similar across asset classes? What is the *factor structure of expected returns?* Be careful, as this is a slippery concept, as we're used to thinking of factor structure in *ex-post* returns. But now that expected returns $E_t\left(r^i_{t+1}\right)$ vary over time, we can factor-analyze the covariance matrix of *expected* returns $cov(E_t r^i_{t+1}, E_t r^j_{t+1})$.,If the bond experience is any guide we will find a very different structure than by factor-analyzing the covariance of *ex-post* returns.

I expect we will not find a single factor, as in bonds. We might – in an economy driven only by changes in risk aversion, all premia would change at the same time. More likely, we will see that broad factors such as equity risk premium, term risk premium, and credit risk premium vary to some extent independently. But even three factors would be a great simplification.

In turn, factor structure in expected returns can help us to overcome the multiple-regression problem, since we have to estimate far fewer parameters. If we impose a (say) 1 factor structure on N assets and K forecasting variables, then we only have to estimate $N+K-1$ coefficients, not $N\times K$ coefficients.

The third empirical task is to relate time-varying expected returns to covariances of ex-post returns. In the end, most asset pricing theories relate expected returns to covariances with factors,

$$E_t\left(r^i_{t+1}\right) = cov_t(r^i_{t+1}\mathbf{f}'_{t+1})\boldsymbol{\lambda}_t$$

For example, Piazzesi and I (2008) found that covariances with a single "level" factor f_t, recovered as an average of all yields, with a time-varying market price of risk $\lambda_t = \lambda_0 + \lambda_1(\boldsymbol{\gamma}'\mathbf{x}_t)$ worked well to capture bond data:

$$E_t\left(r^i_{t+1}\right) = cov(r^i_{t+1}, f_{t+1})\left(\lambda_0 + \lambda_1(\boldsymbol{\gamma}'\mathbf{x}_t)\right);$$

Note the same state variables $\mathbf{x}_t$ enter here to describe the time-varying risk premium as do in describing expected returns. However, the ex-post return covariance "factor" (level) is completely different from the tent-shaped single factor describing *expected* returns.

This step is not an "*explanation*" of expected returns, but it is a necessary step of *data reduction*. There are many common factors in *ex-post* returns as well as expected returns. Relating patterns in *expected* returns to covariances with the common factors in ex-post returns reduces the theorists' job to explaining the behavior of a few factors rather than of many returns. This is the time-series counterpart to the standard cross-sectional investigation, described next, i.e. correlation of various expected return effects with additional factors such as value-growth.

Since the same state variables $\mathbf{x}_t$ appear again, simultaneously looking for this covariance structure while we look for factor structure in expected returns may further help to surmount the too-many-variables problem. As will lots of intuition and artistic skill.

At this point, it's clear that "time-series" has started to take on "cross-sectional" dimensions, so it's worth thinking about the cross section directly and then coming back to connect the two.

2 The Cross Section

Our understanding of facts in the cross section has undergone if anything more radical change. First, there was chaos. Then, came the CAPM. Every clever strategy to deliver high average returns ended up delivering high market betas as well.

It's worth remembering that the CAPM still works well for many phenomena .For example the average returns of size portfolios are reasonably captured by the CAPM, as figure 8 reminds us. The CAPM fell apart in looking at new phenomena, not that its old successes were overturned.

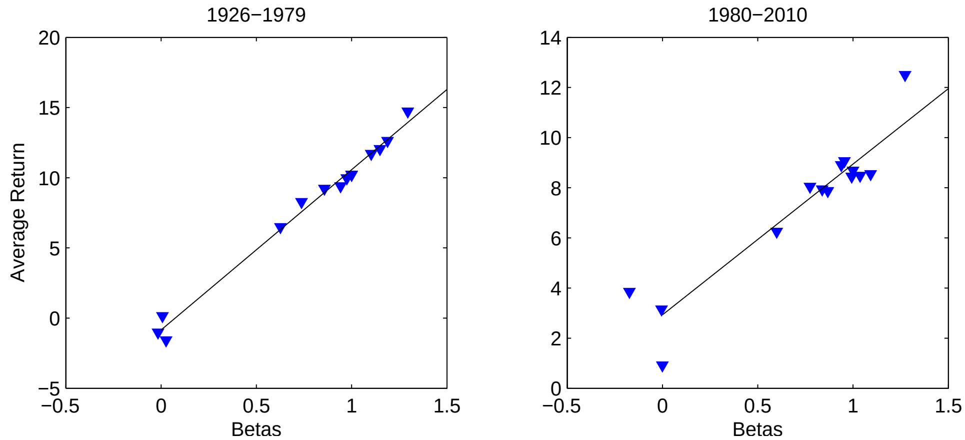

Figure 8: CAPM on Fama-French size portfolios, and , 10 and 30 year government bonds, montlhy data 1926-2009. The diagonal line is the fit of a cross-sectional regression.

Then there was chaos again; a wide number of strategies seemed to deliver high average returns *without* corresponding high betas. The "value effect" is the most prominent – stocks with low prices relative to book or other measures of value deliver high returns, *without* high betas.

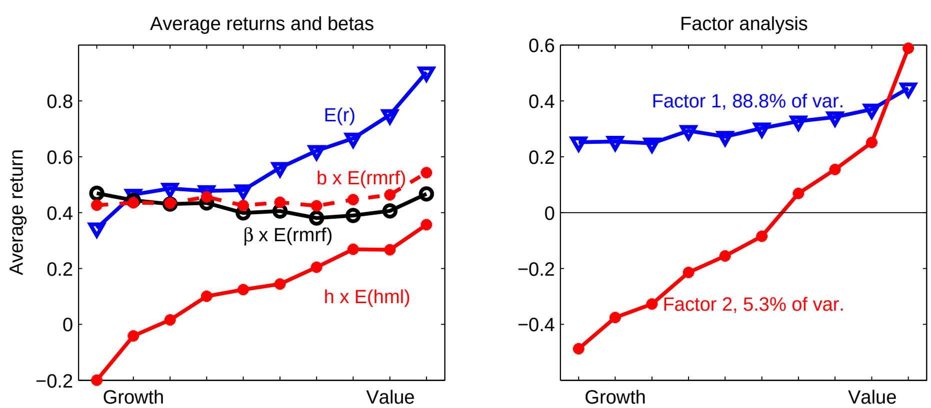

Figure 9: B/M sorted portfolios, monthly data 1963-2010. Left panel: Average returns, market beta $\times$ market premium, and two-factor betas times premiums. Right panel: eigenvectors of the largest two eigenvalues in the covariance matrix of excess returns.

To illustrate, Figure 9 plots the average returns and betas for the 10 Fama-French book/market sorted portfolios. Average excess returns rise as we go from growth to value. This by itself is not a puzzle. One would expect that "value" companies, with low prices relative to book value, are more sensitive to market movements. The puzzle is that they are not: market betas show no relation to the pattern of expected return. The puzzle is in the *joint* behavior of expected returns an betas. And it is a large puzzle – the unexplained variation across in expected returns across portfolios is again of the same magnitude as the overall level of the market expected return.

To emphasize that the puzzle is joint, Figure 10 contrasts the value puzzle in pre-1963 and post-1963 data. The expected return premium is there in both periods. But there is no value puzzle in the early period because average returns line up well with betas. (Ang and Chen (2007), Fama and French (2006)) Similarly, it's just as much a puzzle that where there *is* a spread in CAPM betas, we *do not* see larger returns. (Frazzini and Pedersen (2010), Fama and French (2006).)

As we think about theories to explain the value puzzle, clearly such theories need to address why *betas* are what they are and why they change, as much as they need to address why *expected returns* are what they are. Most theories are pretty weak on the latter point and only a few empirical papers have even asked the question. (Campbell and Mei (1993) is a notable exception.) (Figure 8 also breaks performance down to pre and post 1980, when size effects were first studied and small cap passive funds were started. The compression of *betas* is the most interesting difference.

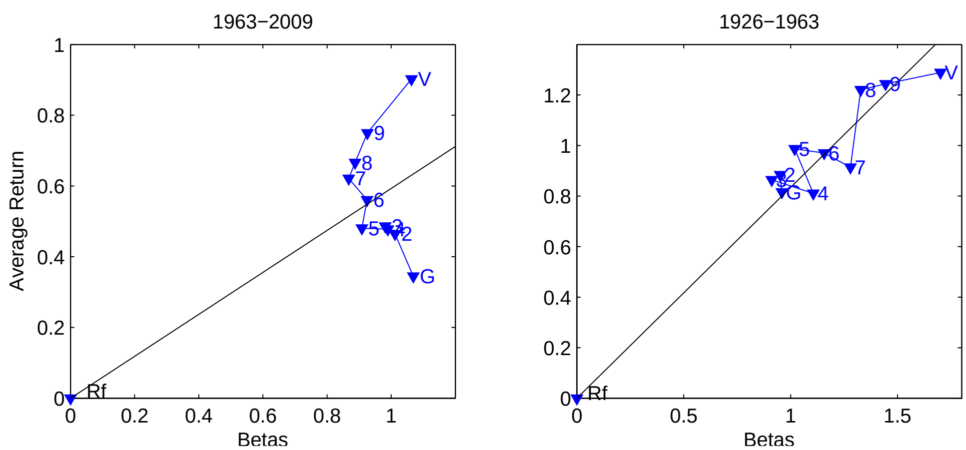

Figure 10: Value effect before and after 1963. Average returns on Fama - French 10 portfolios sorted by book/market equity vs. CAPM betas. Monthly data. Source: Ken French's website.

Fama and French (in a string of papers, but (1993) and (1996) are personal favorites) brought order once again with size and value factors. High expected return strategies correspond to higher betas on three (or four, including bonds) factors, formed by coarser sorts on size and book/market. To Illustrate, Figure 9 includes betas from a multiple regression of returns on the market and Fama and French's hml factor,

$$R_t^{ei} = \alpha_i + b_i rmrf_t + h_i hml_t + \varepsilon_t^i$$

As you can see, the average returns *do* line up well with larger values of the h regression coefficient.

Fama and French's papers are also a shining example of a huge methodological change in empirical finance. This, the most successful model in a generation, is rejected, dramatically so. Fama and French don't emphasize tests. This plot, and corresponding Fama and French (1996) tables of average returns and betas are the *data* for an implied cross-sectional regression. So what is the point of extensive tables showing the pattern of betas lines up with the pattern of expected returns, if the model is rejected? Well, we are all now much more interested in economic significance rather than purely statistical testing, and those tables are far more convincing of a model's basic usefulness than is the most artfully constructed probability value of a likelihood ratio or other overall test. McCloskey (1983) would cheer.

Alas, the world is once again spreading into chaos. A larger and larger number of expected return strategies has continued to pop up in the cross section of stocks bonds and other assets, as well as in the evaluation of funds and trading strategies. These include momentum, accruals and other accounting-related sorts, carry trade, put option writing, and various forms of "liquidity provision." These anomaly variables also produce expected return spreads, but do so without generating higher betas on the market, value or small factors.

2.1 Fama, French and Value

We'll have to repeat the Fama-French exercise, but it will be much harder this time. As in the time series, we face a huge multivariate challenge. To guide this effort, I think it's worth reflecting – with the benefit of much hindsight – on what Fama and French accomplished.

2.1.1 Factors; Where there is mean, there is covariance

Fama and French's first point, I think, is that an "anomalous" structure of mean returns corresponds to the structure of return covariances. All the "high expected return" value stocks (even correcting for market exposure) *will tend to fall together* if they fall. One can read this conclusion as an APT really. "Where there is mean return there must be common movement." If not, Sharpe ratios would explode.

One way to see this point is to factor-analyze the covariance matrix of Fama and French's 10 portfolio returns. The right hand panel of 9 presents the eigenvectors of the two largest eigenvalues of the return covariance matrix, together with the "fractions of variance explained" by each factor (I find $Q\Lambda Q' = cov(R^e, R^{e\prime})$ where R^e denotes the 25×1 vector of Fama-French returns. The Figure plots columns q_i of Q and the corresponding $\lambda_i / \sum_i \lambda_i$.) These eigenvectors q_i are simultaneously the "weights" by which one forms factors from returns, $f_t^i = q_i' R_t^e$, and they are the "loadings" or regression coefficients of returns on factors, $R_t^e = \alpha + q_1 f_t^1 + q_2 f_t^2 + .. + \varepsilon_t$.

The first component is the equally-weighted market (One could start by forcing the first factor to be the value-weighted market). The second factor is obviously a "value" factor – you form it by a grand portfolio with these weights, long value and short growth; and in turn these are the regression coefficients.

Finding multiple factors in return covariance is not a big surprise. The CAPM allows for multiple factors, such as industry portfolios. It just states that exposure to additional factors does not generate a risk premium. So, with the CAPM failing, Fama and French in retrospect did not have to manually or intuitively construct size and book/market factors. They could have just tried the next few factors in the covariance matrix of their returns, and noted that after including loadings on these factors there is little alpha left. Alternatively, since factors are orthogonal, they could have simply noted that the second and third factors have positive mean and zero beta. Since there is little ex-post variation left as well, the conclusion pretty much follows by APT logic. Fama and French's time-series R^2 are in the 90% range. (Unsurprisingly, the first three components of the more famous 25 size and book/market portfolios span those factors, and the fourth component is a "small growth factor." See the Appendix.)

This observation also give me confidence in the effect. A mean return which does not correspond to common movement gives rise to a huge Sharpe ratio. Thus, mean returns which survive must correspond to covariance; all the high mean return securities must rise and fall together ex-post. Covariance need not accompany mean – industry groups move together but seem to offer no alpha – but mean which is not accompanied by covariance – and hence implies astronomical Sharpe ratios – is suspect.

The point of Fama and French's factor model, then, is that practically all the variance and mean or pricing information in the 25 portfolios can be summed up in the mean and variance of the three factors. The opportunities in the 25 portfolios are just a repackaging of the three factors. I think this point has been missed by a generation of empirical researchers, who test other asset pricing models in the Fama and French 25 portfolios, as if they were independent. (Lewellen, Nagel and Shanken (2010), Daniel and Titman (2006).) The only way to do better than the FF 3 factor model on the 25 portfolios is to do better on the anomalies in those portfolios, in particular small growth, and one may be sceptical that this is the most interesting challenge to asset pricing.

2.1.2 Other sorts

Fama and French's second achievement is more important. They showed that many other anomalous expected returns also corresponded to high value betas. Sales growth is my favorite. "Contrarian" strategies that buy stocks of companies that have had 5 years of poor sales have high average returns. Though this strategy uses no price information, its resulting high returns correspond to high HML exposure. (Fama and French (1996.)) Just like the CAPM before it, Fama and French's "value" factor explains *other* expected return sorts.

Thus we reestablish a sense of the order brought by the CAPM: there are many strategies that deliver high expected returns, and there are many factors in the covariance of returns. Yet only a few covariance factors are necessary to understand the pattern of means. The HML and SMB factors can provide a generic risk-adjustment methodology for the zoo of potential new anomalies.

Alas, value betas do not explain momentum or a host of other anomalies that have appeared since then. Yes, each of these can be "explained" by adding additional "factors." For example, the average returns of 10 momentum portfolios can be "explained" by varying betas against a single "up minus down" factor, and one can generate a picture that looks very much like Figure 9. If the stocks that went up last year go down, they all go down together.

Similarly, bonds in countries that have high interest differentials relative to the US have higher ex post average returns than those with low interest differentials. But if these bets go sour, they all go sour together. (Lustig, Roussanov and Verdelhan (2010a).). A "high minus low" interest-differential factor "explains" the average return pattern. And so on.

We don't know this yet for all expected return anomalies. For example, Fama and French (2010) establish that equity issuance gives a spread in mean returns, not related to market, size, b/m, and momentum "factors." But they did not show explicitly that 10 equity issuance portfolio average returns can be explained by an "equity issuance" long-short factor.

We should check this routinely. For every return anomaly, not explained by exposure to existing factors, we should check that it does correspond to a new factor formed on the same sorting variable. But it is most likely to be true of genuine return sorts, otherwise huge Sharpe ratios would emerge. (One reason it's a good check is that if it is not true, one is less inclined to believe that the expected return spread is real.)

So, like Fama and French's value analysis, at least we can usually reduce each anomaly to the presence of an extra priced factor, rather than need to keep track of 10 (or 20 or 100) anomaly portfolios. But the number of priced factors is exploding uncomfortably.

2.2 The multidimensional challenge

So, we are back to the something like situation before the arrival of Fama and French's small and value factors. The task before us is to repeat what they did – or find some equally clever way of restating the question.

1. Which of the dozens of variables that produce expected return spreads are really independently important, and which are subsumed by other variables?
2. Does each independent dimension of *mean* return correspond to an independent dimension of return *covariance*, or do mean returns line up with covariances on a reduced set of factors?

The first question is just the classic multiple regression question. Many sorts are basically different versions of the same thing; i/k, equity issuance, book/market, price/dividend and price/earnings have a strong common component at least, so their information is probably subsumed by a smaller set of variables.

The answer to the second question does not have to be yes. It's entirely possible that there are 27 priced dimensions of both risk and return. However, it would be lovely if, as in the CAPM world, multiple dimensions of mean return corresponded to a lower number of dimensions of risk, but those risks corresponded to large common movements in asset returns.

This vision is plausible as well. For example, in 2007 and 2008, hedge funds found to their dismay that portfolios they had constructed to exploit multiple "signals" all fell at the same time. This is exactly suggestive of a single source of "risk" corresponding to multiple signals of "return." It may be the case that "liquidity provision" in each market earns a return, but at the cost that liquidity dries up simultaneously in all markets.

Finally, of course, the same integration across markets I emphasized for time-series regression needs to happen in these "cross-sectional" characterizations as well. As I emphasize below, there really is no difference between time-series and cross section, so a grand unification of risk and return across countries, markets, characteristics, and factors is the goal. Asness, Moskowitz, and Pedersen (2010) are a courageous start in that direction.

Of course, there is always an equivalent single-factor representation, a single mean-variance efficient portfolio that captures the pricing information of an arbitrary number of factors. In this context the question is whether this single portfolio can be constructed in a simple manner, using fewer characteristics than are present in the independently important dimensions of mean return.

Though these questions are in a sense the same as those that Fama and French addressed, we won't be able to answer them in the same way. So, as well as reflecting on what they did, we need to reflect on how they did it, and how to extend the ideas.

2.3 Asset pricing as a function of characteristics

Fama and French bequeathed us a widely-followed methodology as well as a set of results: Take a variable, like book to market; form 10, 25, 50 or 100 portfolios based on that variable. Make a table of mean returns, and a table of betas and alphas. Look to see if the mean returns line up with the betas, and if the alphas are small. Then, if necessary form factors based on a coarser portfolio sort, like top 1/2 - bottom 1/2 or top 1/3 - bottom 1/3, and use these as additional "factors."

As successful as it was for size and value effects, and for documenting anomalies one at a time beyond these, this technique is clearly at its limits for making sense of my deeply multivariate questions. Two-way sorts are strained, and three- and four-way sorts are simply impossible. So let's go back to basics to think about what this technique is really accomplishing so we can generalize it to many more variables.

Why do we form portfolios anyway? There is an implicit and very useful auxiliary assumption here, not present in asset pricing theory:

Assumption *Economic properties such as means and covariances are stable functions of* **characteristics** *not of individual securities.*

The characteristics provide us information about means, variances, and covariances that are a) stable over time, allowing statistical inference, and b) more precisely measured than they are for individual securities. "Security name" is of course one characteristic. We use others, because we think other characteristics, such as size, book/market, momentum, etc. are more strongly and more stably attached to statistical properties of returns. The characteristic assumption also allows us to use "nearby" securities to estimate the properties of one value of the characteristic. In classic language, portfolios have lower variance and hence lower standard errors.

Given we are interested in characteristics, we do we look at portfolios? *Portfolio sorts are the same thing as nonparametric cross sectional regressions.* Figure 11 illustrates the idea. When we group securities by book/market ratio and then take (say) portfolio means we are in essence running a non-parametric cross-sectional regression of average returns on book/market ratios, using nonoverlapping histogram weights.

An econometrician seeing this picture would instantly tell us to use overlapping and smooth weights, as suggested in green. One security can inform more than one point of expected return. Such weighting does not lose the portfolio interpretation: portfolios can be weighted, and securities can appear in more than one portfolio.

I do not explore this idea deeply, because my goal is to think about multivariate techniques. Multivariate non-parametric regressions are a little easier than multiple portfolio sorts, but not a lot. To the multivariate end I think it's more useful to explore *parametric* estimates. Figure 11 naturally suggests a cross sectional regression

$$E(R^{ei}) = a + b \times \ln(beme_i) + \varepsilon_i$$

Just to be clear, this is not an "explanation" of average returns in the way that a cross-sectional regression of average returns on betas would be. It is a *characterization* of average returns, a compact way of summarizing the usual table of portfolio means.

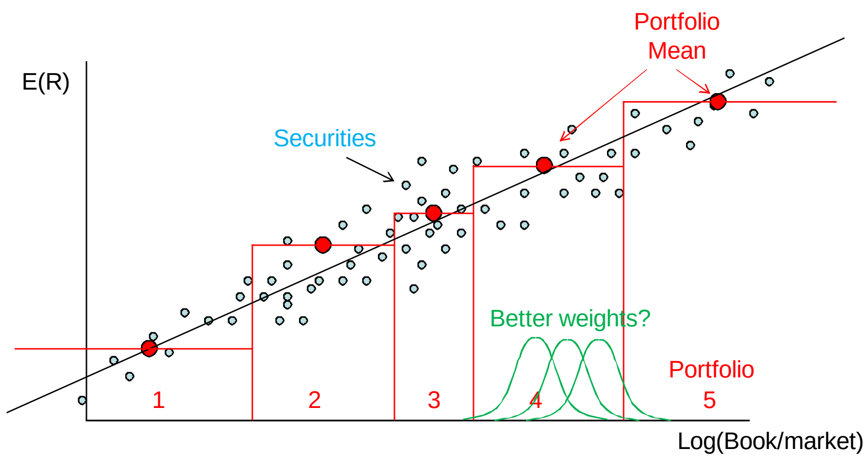

Figure 11: Sorted portfolios and cross-sectional regressions.

The difficulty of a parametric approach, of course is that we have to take a stand on functional form. It's a good idea to do some histograms or nonparametric univariate regressions to check units before jumping to multiple regressions. A few issues stand out

1. Units matter. I use log book/market after some simple histograms showed rather obviously this is the right transformation.

2. As Figure 11 illustrates, there are more securities in the middle, so the middle histograms are narrower. Thus, one has to decide whether expected returns are a linear function of log *book/market* itself, or whether they are a linear function of the *portfolio number*, of the security's *rank* in the distribution of book/market ratios.

In the illustration, expected returns are a linear function of book/market itself, so if we were to graph expected returns as a function of *portfolio number* we would see an S shaped relation, with more variation at the ends. Such a relation is typical of empirical work. You can see it clearly in Figure 9. It doesn't mean any skulduggery, or that "anomalies only occur in the extremes." An S shaped function of *portfolio number* happens naturally as a result of a linear function of *characteristic*, and a standard distribution of that characteristic in which there is more variation across quantiles in the tails of the distribution. Fat-tailed characteristic distributions will lead to even more extreme-weighted portfolio distributions.

Conversely, if indeed the expected return varies consistently and linearly across portfolio number, that means that the expected return is an s-shaped function of the underlying characteristic, matching the cumulative distribution function of that characteristic; or equivalently that rank really does matter.

3. If one runs OLS regressions, then there is a danger of fitting extremes, as illustrated in Figure 12. Extremes are also more volatile and less well measured. A GLS approach might make sense here. Value weightng might make sense as well. If two identical firms merge, they only count once for OLS. Note that too-fine portfolio sorts would also fall into the same trap. And one advantage of regressions is that they are not sensitive to within-portfolio value or equal weightng.

There is nothing new here – these are worries in fitting any regression. Theory doesn't help

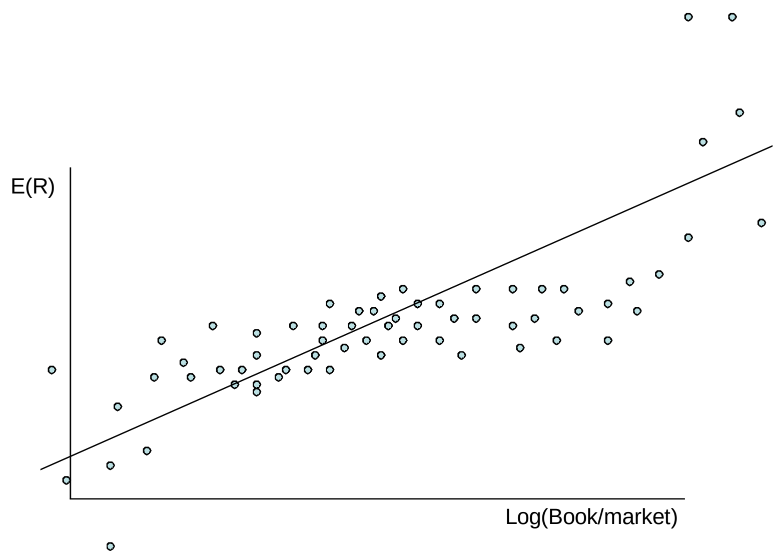

Figure 12: A warning on OLS equally-weighted cross-sectional regressions

(yet) since we are just exploring the functional form of the auxiliary assumption that statistics are stable functions of characteristic.

The goal of empirical asset pricing, then, seems to be to understand expected returns as a function of security characteristics. I'll use C to denote the large vector of such characteristics. They include size, book/market, momentum and other portfolio-sorting variables. They could also include betas and security names. Finally, of course, the contractual nature of the security matters, of course.

$$C_{it} = [size_i,\ b/m_i, d/p_i, mom_i, ..., \beta_i, name_i, stock/bond_i...]_t$$

Means of sorted portfolios cross-sectional and time-series regressions are just ways of estimating the *function* relating expected returns to this vector of characteristics,

$$ER^e(C) = E_t(R^e_{t+1}|C_{it}).$$

The challenge before us is to understand and simplify this function, to understand which characteristics really matter, now that we face a huge multidimensional vector of characteristics.

The second challenge is covariance – we want to line up the expected return *function* of characteristics with a similar covariance *function.* If one has a particular asset pricing model, be it CAPM, consumption-based, etc., that can be summarized by an observable discount factor m_{t+1}, i.e. a model of the form

$$E(R^e_{t+1}) = cov(R^e_{t+1}, m_{t+1}) \times \lambda,$$

we want to estimate the covariance *function*

$$cov(R^e_{t+1}, m_{t+1}) = cov_m(C).$$

The asset pricing model is successful if there is a single scalar constant λ which equates the mean *function* to the covariance *function.*

$$ER^e(C) = cov_m(C) \times \lambda?$$

Forming portfolios and then finding the covariance of the portfolio return with a factor is exactly a nonoverlapping histogram-weighted nonparametric estimate of this covariance function. It's easy enough to do this more efficiently, calculating overlapping kernel-weighted nonparametric estimates.

To address the multivariate challenge, however, once again it will probably be better to use a parametric approach. For example, If we have a parametric model for the expected return *function*, $ER^e(C)$,

$$ER^e(C) = a + b'C$$

which implies a panel regression

$$R^{ei}_{t+1} = a + b'C_{it} + \varepsilon_{it+1},$$

then it's natural to look for a similar linear covariance function $cov_m(C)$,

$$cov_m(C) = c + d'C. \tag{6}$$

This function could be estimated by a similar panel data regression. If we write

$$\varepsilon_{it} m_{t+1} = c + d'C_{it} + \delta_{it+1}$$

and model (6) is true, then $E_t(\delta_{it+1})$ so this model can be consistently estimated by regression methods. Since second moments usually swamp first moments for asset return data, running

$$R^e_{it} m_{t+1} = c + d'C_{it} + \delta_{it+1}$$

will generally provide an excellent approximation.

Then, the success of the asset pricing model comes down to whether these two *functions* are equal,

$$\begin{aligned} ER^e(C) &= cov_m(C)\lambda? \\ a + b'C &= \left(c + d'C\right)\lambda? \end{aligned}$$

Often we don't know the factor, and want to emulate Fama and French's construction of hml, smb, umd, etc. The natural way to construct factors is as a function of characteristics,

$$f_t = \sum_i w_i h(C_i) R^{ei}_t.$$

$h(C_i)$ forms the long-short portfolio. For Fama and French's construction of hml, for example, it is one in the top third of C_i, zero in the middle, and -1 in the bottom third. A smooth function including all securities, and a linear function

$$h(C_i) = e + fC_i$$

are natural alternatives to consider once the problem is stated this way. I include $w_i > 0$ to allow value weighting, or weighting to reflect the density of observed assets with characteristics near C_i.

More generally, we could follow my suggestion above to factor-analyze the covariance matrix of returns, and search for priced factors in turn. Here, we are thinking of the covariance *function* of characteristics, which can simply be measured as such

$$cov(R^i_{t+1}, R^j_{t+1}) = g(C_{it}, C_{jt}).$$

Throughout this discussion, I relate expected returns to covariances rather than betas. This expression offers the large simplification that there is no difference between whether a factor "is priced" and whether it "helps to price other assets," the "betas" are single regression betas and thus do not change when we add or subtract factors, and there is no difference between λ and weights in a discount factor representation $m = b'f$. Of course multiplying and dividing by a variance-covariance matrix yields the same thing.

2.4 Time series and cross section are the same thing.

We have grown used to thinking separately about "time-series" predictability, captured in forecasting regressions, and "cross-sectional expected return variation" captured in means of characteristic-sorted portfolio returns. Now that we realize portfolio sorts are the same thing as cross-sectional regressions, it's clear that time series regressions are the same thing as well. Just add t subscripts to the last section.

All time-series and cross sectional approaches add up to one big question – estimating

$$R^{ei}_{t+1} = a + b \times C_{it} + \varepsilon^i_{t+1} \tag{7}$$

in order to understand expected return variation as a function of characteristics,

$$E(R^{ei}_{t+1}|C_{it})$$

This is just a big panel-data regression. Then, relate this variation in expected returns to covariance, also a function of characteristics,

$$cov_t(R^{ei}_{t+1}, f_{t+1}) = g(C_{it})$$

At heart the portfolio technique as a way of turning a *time-series* phenomenon into a *cross-sectional* one. Suppose for example that all securities were ex-ante and unconditionally identical, yet had time-series predictability of the form (7). Then the "value" portfolio, consisting of stocks with temporarily higher beme, would have high average returns, and the "growth" portfolio would have the opposite.

This is the "managed portfolio theorem" of empirical asset pricing (Cochrane (2005c)). If we use an instrument in a conditional asset pricing test, that is the same as forming a managed portfolio based on that instrument and forming an unconditional asset pricing test.

$$E(m_{t+1} z_t R^e_{t+1}) = E(m_{t+1}(z_t R^e_{t+1}))$$

Equivalently, and more prosaically, a "static" or "cross-sectional" investment in hedge fund that market-times or carry-trades, is the same as "market timing." There is no conceptual difference between cross-sectional and time-series modes of investment.

This reinterpretation from "time-series" to "cross-sectional" facts was an important way to help us see the economic importance of previously small-seeming facts. Momentum is a great example. By forming momentum portfolios, an apparently "economically insignificant" $R^2 = 0.01$ can be turned into a very "economically significant" 1% per month 1-10 spread portfolio return. Suppose there is a small autocorrelation of individual returns, expressed in time-series as

$$R^{ei}_{t+1} = a + \rho R^{ei}_t + \varepsilon^i_{t+1}$$

Now, when we form a portfolio of the 1/10 stocks that have gone up the most last year, those stocks have typically gone up a lot. Suppose, for example, that the standard deviation σ_ε of individual log stock returns is 40% and normal. The mean of a standard normal, conditional on being in the 10th percentile, is 1.75 standard deviations[1]. Then, the spread between last year's 1-10 portfolio log returns is $\sigma \times 2.8 = 40 \times 2.8 = 112\%$. If the autocorrelation coefficient ρ is only 0.1, with an R^2 of 0.01, then the 1-10 portfolio will have a mean return of 11.2% or 1% per month.

Lustig, Roussanov and Verdelhan (2010) provide a great example of this equivalence between time-series and cross-sectional approaches. For 20 years, the "carry trade" had been investigated by regressions, following Hansen and Hodrick (1980) and Fama (1984): country-specific regressions

$$R^{ei}_{t+1} = a + b(r^i_t - r^{US}_t) + \varepsilon^i_{t+1}$$

where R^{ei}_{t+1} is the ex-post dollar return from buying i's currency, holding a bond, and converting back to dollars, less a US interest rate and r^i_t, r^{US}_t are the foreign and dollar interest rates. As usual b is about one, or sometimes larger.

Lustig, Roussanov and Verdelhan converted this regression to a "cross sectional" relationship by forming portfolios of countries, sorted on the basis of their interest differential to the US $r^i_t - r^{US}_t$, and finding higher average returns in the "high spread" portfolio.

Second, and more importantly, Lustig, Roussanov and Verdelhan formed a "carry trade factor," long the high-interest-spread countries and short the low-interest-spread countries. They found that the pattern of expected returns across carry-trade portfolios matched the exposure to this factor.

As above, this is easy enough to do or to reinterpret in a regression context: form the factor portfolio

$$f_{t+1} = \sum_j w(C_{jt}) R^{ej}_{t+1} \text{ or } f_{t+1} = \sum_j w(C_{jt}) \varepsilon^{ej}_{t+1}$$

where w is a set of weights. The weights can be linear in C; they can be the step function implied by typical high-low portfolio formation, or they can be chosen based on factor analysis of the covariance matrix of ε expressed as a function of characteristics. Then, find how the covariance of returns with the factor portfolio depends on the return characteristic

$$cov_t(R^{ei}_{t+1}, f_{t+1}) = g(C_{it}).$$

Portfolio betas are just a kernel estimate of this function. Finally, check if the average returns line up with the covariances.

However, two formally equivalent procedures lead to different intuition. In 20 years and perhaps a thousand papers investigating the carry trade, including 15 years since Fama and French (1996), nobody thought to do this obvious covariance analysis of the time-series regression residuals.

[1]

$$\frac{\int_{1.28}^{\infty} x e^{-\frac{x^2}{2}} dx}{\int_{1.28}^{\infty} e^{-\frac{x^2}{2}} dx} = 1.7537$$

$$\frac{1}{\sqrt{2\pi}} \int_{1.28}^{\infty} e^{-\frac{x^2}{2}} dx = 0.100$$

Now that we have seen the correspondence, however, there is no *statistical* advantage to forming portfolio means. For the momentum effect, the panel-data estimate of $\hat{\rho}$ will be statistically stronger, in fact, since it uses information from all assets and not just the 1-10 outliers. With the $\hat{\rho}$ estimate in hand, we can characterize the large variation in expected returns as easily with the regression model as with portfolio means. This story is just one more reminder that R^2 is not a good measure of the economic significance of expected return variation, either "over time" or "across assets."

Similarly, now that we understand by the portfolio characterization how to think about the cross-section of carry-trade returns and their factor structure, estimating that structure parametrically and using the regression tools might be more powerful.

Functional forms still need care, since again we are still characterizing the auxiliary hypothesis (association of expected return with a characteristic) rather than using any theory. If we unite time series and cross section in a panel, we are assuming that variation *over time* in book-market ratios has the same effect on average returns as variation *across assets at a given moment in time.* Maybe, maybe not. For example, Fama and French's portfolio construction implicitly assumes that these are not equal, or at least throws out a lot of information. By uniting the top 1/10 securities at one date with high average b/m, or more dispersed b/m with the top 1/10 securities at another date with low average b/m or less dispersed b/m, Fama and French implicitly assume that variation over time in a portfolio's b/m is not useful for understanding its expected return, and that expected returns attach to B/M rank rather than level, so that unusual dispersion in B/M across assets does not affect expected returns.

2.5 A look at a popular cross section

To explore the equivalence of portfolio and regression approaches, Table 5 presents panel data regressions using the Fama-French 25 portfolios. The first row gives a pure cross-sectional regression,

$$E\left(R_{t+1}^{ei}\right) = a + b \times E\left(size_{it}\right) + c \times E(bm_{it}) + \varepsilon_i;\ \ i = 1, 2, ..25$$

(Size and bm refer to logs. Since raw size grows over time, the size variable here is the log fraction of total market equity.) The fitted values of this regression fit the portfolio averages quite well, with a 77% R^2 (One does better still with a $size \times bm$ cross-term. This specification allows the growth portfolios to have a different slope on size than the value portfolios.)

	size_t	bm_t	Δsize_t	Δbm_t
1. Cross section	-0.030	0.27		
2. Pooled	-0.022	0.55		
3. Time dummies	-0.031	0.29		
4. Portfolio dummies	-0.087	1.48		
5. Differences	-0.030	0.46	-0.38	1.11

Table 5 Regressions of Fama-French 25 size and B/M portfolio returns on size and book/market characteristics. The regression specification is $R_{t+1}^{ei} = a + (a_t) + (a_i) + b \times size_{i,t} + c \times bm_{i,t} + d \times (size_{i,t} - size_{i,t-12}) + e \times (bm_{i,t} - bm_{i,t-12})$. Terms in parentheses only appear in some regressions. "Cross section" regresses time-averaged return on time-averaged portfolio characteristics. *size* is log(market equity/total market equity). *bm* is log(book/market). Monthly data 1947-2009. Data from Ken French's website.

The second row of Table 5 gives a pooled forecasting regression, which is the most natural way to integrate time series and cross section,

$$R^{ei}_{t+1} = a + b \times size_{it} + c \times bm_{it} + \varepsilon_{it+1};\ i = 1, 2...25; t = 1, 2, ..., T.$$

The size coefficient is a little smaller, and the b/m coefficient is much larger.

To diagnose the difference between the cross section and pooled, rows 3 and 4 present a regression with time dummies and a regression with portfolio dummies. The time-dummy regression is the same as taking out time averages for each portfolio, and looking only at variation across assets.

$$\begin{aligned} R^{ei}_{t+1} &= a_t + b \times size_{it} + c \times bm_{it} + \varepsilon_{it+1} \\ R^{ei}_{t+1} &= a + b \times \left(size_{it} - \overline{size}_t\right) + c \times \left(bm_{it} - \overline{bm}_t\right) + \varepsilon_{it+1} \end{aligned}$$

It's not quite the same as the pure cross-sectional regression, because it allows times with a higher cross-sectional spread in characteristics to show higher expected return spreads. Comparing first and third rows, this difference turns out not to matter. The regression with portfolio dummies is the same thing as taking out portfolio averages, and looking only at the variation over time of the characteristics,

$$\begin{aligned} R^{ei}_{t+1} &= a_i + b \times size_{it} + c \times bm_{it} + \varepsilon_{it+1} \\ R^{ei}_{t+1} &= a + b \times \left(size_{it} - \overline{size}_i\right) + c \times \left(bm_{it} - \overline{bm}_i\right) + \varepsilon_{it+1} \end{aligned}$$

Rows 3 and 4 show that variation *over time* in a *given* portfolio's bm is a much stronger signal of return variation than the same size variation *across portfolios* in average bm. (This finding may in part be due to data definitions. If book value means different things in different industries, then some industries will be perpetually stuck in "value" and others in "growth.")

This quick look is all a prelude to individual-return based regressions. For that application we can't use portfolio dummies or subtract portfolio means, since the average return of a security over the whole sample is meaningless. The last line of Table 5 gives a way to capture this difference between time-series and cross section without dummies – it allows an independent effect of recent changes in the characteristics. This specification accounts quite well for the otherwise unpalatable time and portfolio dummies. (It goes in the other direction from momentum).

Of course, this "time-series" result could be captured "cross-sectionally" as well – it suggests that if we form portfolios based on recent changes in size and book/market as well as levels, we will obtain portfolios with a higher spread in average returns.

2.6 Understanding the cross section of prices

Now that we understand that time series and cross section are really the same thing, we need to do a better job of tying "cross-sectional" variation in expected returns back to their implications about "cross-sectional" variation in valuations. How much of the difference between one asset's price-dividend, price-earnings, book-market, etc. ratio and another's is due to variation in expected returns, and how much to expected dividend growth or other cashflow expectations? Research in this area has been remarkably thin, compared to the vast amount of research on forecasting the market return and decomposing its source of variation, and the even vaster research characterizing cross-sectional variation in one-period average returns. (Vuolteenaho (2002), Chaves (2009) are a few of the exceptions.)

To get a little hint of this question, I'll examine the cross-section provided by Fama-French portfolios. Eventually, we want to state this as a function of *characteristics*, like everything else and use a full cross-section of individual securities. But using the portfolios lets me get a hint of where to go at little cost.

Figure 13presents the average return, dividend growth, and dividend yield of the Fama-French 10 book/market sorted portfolios. A long-term average return is approximately equal to the long-term average dividend yield plus the long-term average dividend growth rate. Value returns come roughly half from greater dividend growth and half from a larger dividend yield.

From a return perspective, this is unsurprising. But from a valuation perspective, this is a very surprising result. High prices – low dividend yields – should correspond to higher subsequent dividend growth, not lower.

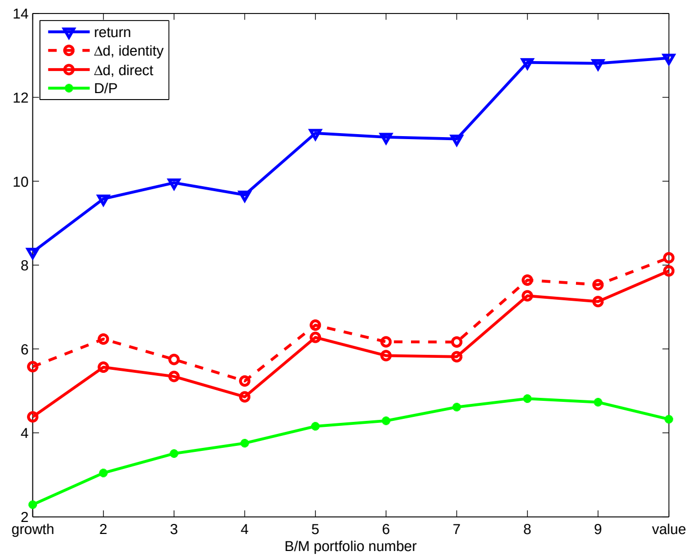

Figure 13: Average return r_{t+1}, dividend growth Δd_{t+1}, and dividend yield dp_t for the Fama - French 10 book/market portfolios, 1947-2008. The dashed Δd line gives mean dividend growth implied by the approximate identity $\Delta d_{t+1} = r_{t+1} - \kappa + \rho dp_{t+1} - dp_t$.

We can the observation in a regression, and do the same sorts of variance decompositions here as with the market return. Taking unconditional means of everything, the return identity $r_{t+1} = -\rho dp_{t+1} + dp_t + \Delta d_{t+1}$ (see the appendix) implies

$$E\left(r^i\right) = (1-\rho)\,E\left(dp^i\right) + E(\Delta d^i).$$

Flipping this around, we have

$$E\left(dp^i\right) = \frac{1}{1-\rho}\left[E\left(r^i\right) - E(\Delta d^i)\right]. \tag{8}$$

Now, we can see that a purely cross-sectional regression of average returns, dividend growth on dividend yields must obey

$$1 = \frac{b_r^{cs}}{1-\rho} - \frac{b_d^{cs}}{1-\rho},$$

where the b are the cross-sectional regression coefficients of the terms in (8). We can interpret these coefficients as the fraction of *cross-sectional* dividend yield variation

$$var_{cs}(dp) = \frac{1}{N}\sum_{i=1}^{N}\left[E(dp^i) - E(dp)\right]^2$$

driven by discount rates and driven by dividend growth.

(Vuolteenaho (2002) uses a different present value identity to understand variation in the book / market ratio directly, rather than use dividend yields as I have. This is a better procedure for individual stocks, which often do not pay dividends. I use dividend yields here for simplicity.)

The first column of Table 6 presents this cross-sectional regression. The results are quite similar to the time-series regressions for the market portfolio from Tables 2 and 3: *more than all of the cross-sectional variation in average dividend yields of these portfolios comes from cross-sectional variation in expected returns* (1.33). Expected dividend growth goes "the wrong way" – low prices correspond to high dividend growth.

	Cross section		Portfolio dummies		Time dummies		Pooled		Pooled	
	b	$\frac{b}{1-\rho}$	b	$\frac{b}{1-\rho\phi}$	b	$\frac{b}{1-\rho\phi}$	b	$\frac{b}{1-\rho\phi}$	dp_{t-1}	Δdp_{t-1}
r	0.053	1.33	0.107	0.903	0.044	0.325	0.095	0.967	0.090	0.074
Δd	0.036	0.89	-0.005	-0.041	-0.083	-0.611	0.004	0.042	-0.004	0.073
Δd^*	0.026	0.64	-0.011	-0.097	-0.092	-0.675	-0.003	-0.033	-0.012	0.076
dp			0.92		0.90		0.94		0.94	0.002

Table 6. Cross-sectional regression coefficients of average returns $E(r^i_{t+1})$, average dividend growth $E(\Delta d_{t+1})$, and dividend yield change on dividend yields $E(dp^i)$, 10 Fama-French B/M portfolios. Implied dividend growth Δd^* is calculated from the approximate identity $\Delta d^*_{t+1} = r_{t+1} - \kappa + \rho dp_{t+1} - dp_t$. I use $\rho = 0.96$

(The approximate return and present value identities do not have to be taken with an expansion point derived from the time-series average of dividend yield, though this is often done in applications. Therefore, it is perfectly OK to use a common expansion point for portfolios with permanently different dividend yields, and to use the linearized identity in this way to relate permanent differences in dividend yield to permanent differences in average returns and average dividend growth.

Sample means obey the identity

$$E\left(dp^i_t\right) = \frac{1}{1-\rho}\left[E\left(r^i_{t+1}\right) - E\left(\Delta d^i_{t+1}\right) + \rho\frac{1}{T}\left(dp^i_T - dp^i_1\right)\right].$$

The last term is fairly large with an 0.12 cross-sectional coefficient and 0.30 when divided by $1-\rho$. That term makes up the identity in the $b/(1-\rho)$ column here in this application.)

We can, of course, ask how much of the *time-variation* in these dividend yields around their *portfolio average* corresponds to return vs. dividend growth forecasts. A regression that includes portfolio dummies, shown next in Table 6, addresses this question. The 0.11 return-forecasting coefficient for *portfolios* is almost the same as the return forecasting coefficient for the market as a whole seen in Tables 2 and 3. The dividend growth forecast is now exactly zero. So *all variation in book/market sorted portfolio dividend yields over time, about portfolio means, corresponds to variation in expected returns.*

The regression with time dummies, next in Table 6, paints a different picture. The return coefficient is smaller at 0.044, and ϕ is smaller as well, so expected returns only account for 33% of the variation in dividend yields. Finally we see an important dividend growth forecast, with the right sign, -0.08-0.09, accounting for 61-68% of dividend yield volatility. The strong contrast of this result with the pure cross sectional regression means that *a time of unusually large cross-sectional dispersion in dividend yields corresponds to an unusually high dispersion in dividend growth forecasts.* Dividend yields of these portfolios move overall with a very slow trend. About this trend, there are times in which the dispersion widens and tightens. This widening and tightening does not add up to a large variation in individual dividend yields about their portfolio means, so we don't see it in the regression with portfolio dummies.

This is an important regression, in that it gives us a sense that there is a component of variation in valuations that does correspond to dividend growth forecasts. The unusual dispersion in dividend growth forecasts adds up to zero, so this kind of variation cannot be seen in the aggregate dividend yield and its forecasting relations. We get a sense here that there is individual-security variation in forecastable dividend yields, which drives some individual variation in prices, but which averages out across all securities, so that the aggregate dividend yield is driven primarily by expected returns.

As this table makes clear, there are lots of different kinds of variation in dividend yields, and they have different sources, with the time dummy and portfolio dummy results being the most different. One way to add them up is simply to pool the regression with no dummies. This regression looks at variation in dividend yields from all sources, and asks where such variation comes from. The results are very similar to the portfolio dummy results, once again showing all variation due to variation in expected returns.

The last column of Table 6 reminds us that other variables matter too of course, so the final project must be both multivariate and cross-sectional. Size portfolios, presented in the appendix show the same pattern of time vs portfolio and pooled results.

I don't want to overstate the usefulness of this kind of exercise. It's a nice data summary, but in the end the right thing to do is evaluate specific models: announce how discount rates vary, how cashflows vary, calculate a prediction for the book/market or dividend yield on the left hand side, and see how close the prediction is to actual values.

3 Agenda

In sum, we are faced with a wide variety of characteristics C_{it} which forecast both "time series" and "cross sectional" variation in returns R_{it+1}. Each characteristic seems to be associated with common movement – returns for which $E(R_{t+1}|C_{it})$ is large seem to move together; the covariance structure of returns seems to have as many "factors" as there are "characteristics" driving means, and many of those "factors" are priced.

This program is only really beginning. The list of return-forecasting characteristics is already long:

Market, size, value, momentum, other equity sorts (issues, for example); level, slope, curvature, credit, and return-forecasting factor in bonds; currency carry; put option writing and "liquidity provision" in hedge fund return attribution; returns from time-varying portfolios following each signal known to forecast returns.

Each of these seems independent of at least most of the others. Most of them seem to correspond to independent dimensions of risk, to "new factors" in covariance.

But not all, and not systematically. The first item on the agenda is to understand for every expected return anomaly whether it does in fact correspond to a different dimension of covariance. For example, is the equity issues anomaly (Fama and French (2010)) matched by an equity issues factor?

Second, the whole business needs to be integrated and understood in a unified, multivariate context. Which of the expected return characteristics are genuinely important in a multiple regression sense? Given that, which of them genuinely requires an independent dimension of covariance?

Even this is the beginning of our empirical journey. In the multivariate extension, we have lost track of the central question, what is the source of *price* variation? Why are we focusing on one-period returns, and relating average returns to betas? When did our field stop being "asset pricing" and become "asset expected returning?" Why are betas – the covariance of tomorrow's price with tomorrow's market price – "exogenous" and on the right hand side?

Focusing on asset returns and betas makes some sense in a world with constant expected returns – and most sense of all if the world is completely i.i.d. In such a world, price variation comes exclusively from dividend growth news. Thus, the covariation of one asset's price with anothers measures the covariation of their cashflows – "return betas" are the same as "cashflow betas." Though in an economic general-equilibrium sense everything is endogenous, at least for finance purposes it makes a bit of sense to take such "cashflow betas" as "exogenous" and say one had "explained" expected returns if they line up with such covariances.

But now that we understand how much price variation comes from discount rate changes, the covariance of one asset's return with market return largely reflects covariance between their *discount rate* innovations. This is obviously a whole lot less "exogenous" to the question of discount rate (expected return) determination than cash-flow betas would be! It doesn't say much that we "explain" expected returns by the fact that innovations in expected returns correlate with innovation in market expected returns.

I think the eventual answer must be that we return to focus on asset *pricing.* Prices, price/dividend ratios, or book/market ratios should be our *left-hand* variable, the thing we're trying to *explain,* not a sorting characteristic for expected returns. Expected returns and dividend growth are the thing we should be trying to explain it *with.*

Part III

Theories

Having described a bit *how* discount factors vary across time and securities, let's think about *why* they do so.

4 Frictionless theories

4.1 Macroeconomics

I start with the standard model of macroeconomic risks. The typical ingredients are frictionless markets, and enough aggregation so that only aggregate risks matter. This class of theories therefore links discount rates to macroeconomic events. Risk premia are high for assets that tend to pay off poorly in "bad times" when aggregate marginal utility is high, and "bad times" are here related most of all to aggregate economic performance.

On a broad scale, of course, such a link is readily apparent. The collapses in asset prices and the macroeconomy in 2008 is certainly not a coincidence. We certainly have experienced a "bad time," a high marginal utility event, and many high average-return securities and strategies did much worse in this event than in good times. High expected-return assets and business cycles are correlated more generally – risky assets decline in recessions.

Moreover, all the signs of high expected returns occur in recessions – they are times of high dividend yields (low prices), high credit spreads, rising yield curve, high foreign-to-domestic interest spread. This rising *expected* return is largely responsible for the widespread decline in actual returns during the early stages of recessions.

The challenge is how to measure "bad times" and "times of high risk premium" in a theoretically coherent and empirically productive way.

4.1.1 Consumers and macroeconomics

The central assumption is that markets are frictionless, so we can exploit consumer/investor first order conditions to tie the discount factor to marginal utility growth,

$$\begin{aligned} E_t(m_{t+1}R^e_{t+1}) &= 0 \\ E_t(R^e_{t+1}) &= R^f cov(R^e_{t+1}, m_{t+1}) \\ m_{t+1} &= \delta \frac{u_c(t+1)}{u_c(t)}. \end{aligned}$$

I will also add the assumption that risks are adequately shared, so that only aggregate risks affect asset prices. Then, we can use aggregate consumption to infer the discount factor; in the canonical case of power utility that depends only on nondurable consumption (i.e. separable across goods),

$$m_{t+1} = \delta \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}.$$

(Perfect risk sharing is of course too strong an assumption; aggregation can work in far weaker settings. What I mean to emphasize is that the application of this branch of theory does not relate asset pricing to data that reflects dispersion of uninsured risks across people.)

Of course the simple version of this theory, combining power utility, nondurable and services consumption data, and high frequency evaluation, does not work well. Yet at a basic level, consumption must remain important. Consumption did fall like a stone along with the stock market in the recent great recession. For this reason, there has been a long search to find more sensible links at lower frequencies, by better measurement, and more careful thought about utility functions.

The basic Hansen-Jagannathan (1991) bound

$$\frac{E_t(R^e_{t+1})}{\sigma_t(R^e_{t+1})} = \frac{\sigma_t(m_{t+1})\rho_t(m_{t+1}R^e_{t+1})}{E_t(m_{t+1})} \approx \gamma\sigma_t(\Delta c_{t+1})\rho_t(\Delta c_{t+1}R^e_{t+1}) \tag{9}$$

is very useful for thinking about how this kind of theory matches data. (This interpretation of the bounds is from Cochrane and Hansen (1992).)

The equity premium remains a puzzle of course. Typical numbers need very high levels of risk aversion γ. $E(R^e) = 6\%$, $\sigma(R^e) = 18\%$, $\sigma(\Delta c) = 2\%, \rho < 1$ requires $\gamma > 17$. The heart of this puzzle is that consumption is so much smoother than stock returns. In an i.i.d. world, as in standard models like the log utility CAPM, consumption is proportional to wealth, so consumption volatility is as large as that of returns. The fact that so much short-run stock return volatility comes from discount rate variation, yet consumption, dividends and wealth must share the same trend suggests that long-run versions of this relation may be quite different, but that idea has not been systematically documented. (Note that aggregate consumption shares a trend with aggregate wealth and dividends, not to cumulated returns and dividends paid to an initial \$1 investment. The difference is important in making sense of long-run statistics.)

I emphasized the time-variation of expected returns rather than their level as a central empirical fact. Standard estimates of $\sigma_t(\Delta c)$ or $\sigma_t(R^e)$ don't vary much, and I wouldn't know what to make of a time-varying conditional correlation coefficient if there was one. These observations together with the pervasive business cycle behavior of many assets point to some form of time-varying risk premium, linked to the business cycle. Something about recessions makes people less willing to take systematic risks they otherwise happily would take, and this aversion is displayed in many asset markets. (The "systematic" qualifier is important. I don't know of any evidence that people become more risk averse generally – less willing to go bungee-jumping, say. However, the whole theory of finance is about distinguishing unwillingness to take systematic risks, that will pay off badly if the economy worsens, from unwillingness to take such idiosyncratic risks.)

John Campbell and I (1999) investigated what still seems to me a useful specification of that risk aversion coefficient: people get used to their consumption levels, so that rapid falls in consumption lead them to behave in a very risk averse manner. We specified a utility function

$$u(C) = \frac{(C_t - X_t)^{1-\gamma}}{1-\gamma}$$

where X_t represents the level of habit. The local curvature of this utility function is

$$-\frac{u''(C)}{Cu'(C)} = \gamma\frac{C_t}{C - X_t}$$

so it's natural that as consumption declines towards habit, people become more risk averse. In turn, habits adjust slowly to the level of consumption.

Figure 14 illustrates the idea. As you get closer to the disaster X, you become less willing to take bets that are the same *proportion* of consumption C.

Figure 15 presents the surplus consumption ratio $(C - X)/C$ inferred from the habit specification Campbell and I (1999) used, together with the price-dividend ratio. Our model predicts that the price-dividend ratio is a nearly log-linear function of the surplus consumption ratio. Obviously there is some residual – stylized one-state-variable models that predict 100% R^2 can

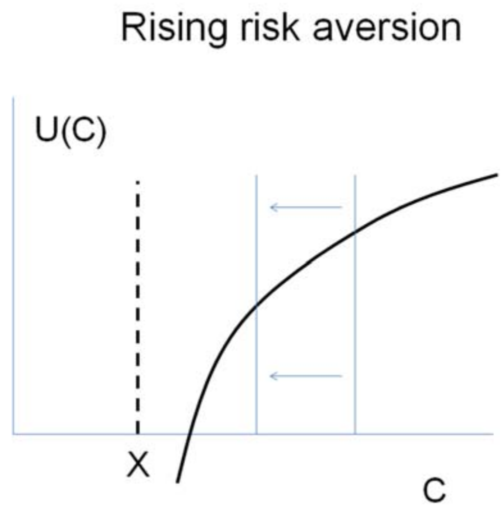

Figure 14: How utility with habits or subsistence level leads to varyng risk aversion.

always be rejected. But given the hue and cry that "the financial crisis invalidates all standard asset pricing," as well as the general impression in finance circles of the hopelessness of all macro based pricing, the general pattern is remarkably good. If anything, stocks fell *less* than suggested by the catastrophic fall in consumption.

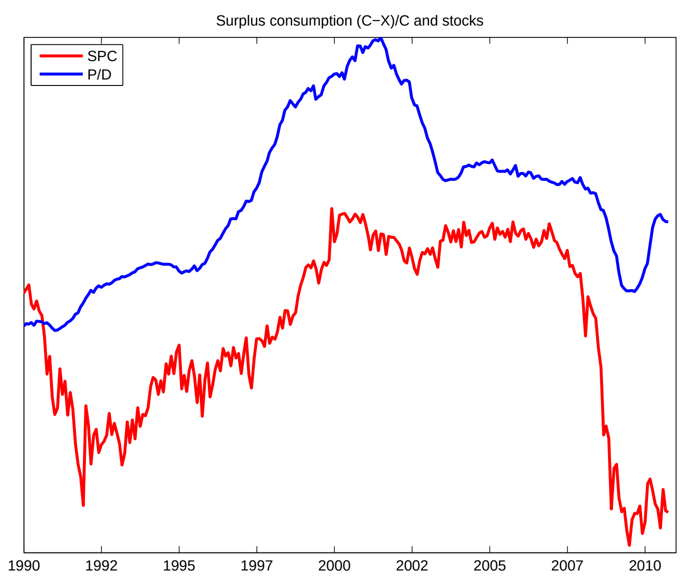

Figure 15: Surplus consumption ratio and price/dividend ratio. Surplus consumption is formed from real nondurable + services consumption using the Campbell and Cochrane (1999) specification and parameters. Price/dividend ratio is from the CRSP NYSE VW portfolio.

This is a great moment to be clear about causality. Asset pricing models are based on first-order conditions, which do not specify a direction of causation. One does not have to believe that "an exogenous consumption fall caused stock prices to fall," or "exogenous price variation caused consumers to cut back." The fact that we had a run on the shadow-banking system is at least an important amplification mechanism in a proper general-equilibrium understanding of this event! But assets can remain correctly *linked* to consumption growth through a financial crisis, and the frictionless consumption model can remain a useful piece of our understanding, even if a run or other "friction" is an important part of the causal chain describing why both stocks and consumption fell. The endowment-economy specification (Lucas (1978)) from macroeconomics, or the traditional linear-technology (fixed rate of return) specification from finance are each just useful abstractions for quickly arriving at consistent processes for consumption and asset returns that are valid in a richer general-equilibrium setting.

Campbell and I described habits as a pure aspect of preferences. However, a wide variety of mechanisms will give very similar behavior, including the accumulation of durable goods, houses, or knowledge of how to manage a different standard of living. It also can reflect leverage with any cost of bankruptcy. Both households and institutions become much more "risk averse" as they near bankruptcy induced by leverage. More broadly, models in which the cross-sectional distribution of risks induce greater average risk aversion (Constantinides and Duffie (1996)) have the same general feature. The central idea in this *class* of model is that there is a mechanism by which the average investor becomes more risk averse in macroeconomic bad times. Whether the precise nature of that mechanism is a habit in preferences, fear of bankruptcy, fear of losing a job or business, is an important specification issue to be worked out, but leaves the same general flavor.

4.1.2 Other consumer-based models

Four other modifications to the utility function are popular in current work 1) nonseparability across goods – durable and nondurable; traded and nontraded; 2) recursive utility and long run risks 3) rare disasters 4) imperfect risk sharing. (See Cochrane (2007a) for a fuller review of these models.)

Nonseparability is in principle important – the marginal utility of nondurable consumption may be affected by other goods and by leisure. Recessions are primarily features of durable goods purchases and employment. Nonseparable utility is one way of formalizing these evident empirical correlations with asset markets.

The long-run risks model, which basically comes form abandoning separability across states of nature, (Epstein and Zin (1989)) has become very popular following the work of Bansal and Yaron (2004), Hansen, Heaton and Li (2008) and others. Part of the popularity is that it gives a technically simple way of allowing higher risk aversion – resistance to substituting consumption across states of nature – without also imposing lower intertemporal substitution elasticity – resistance to substituting consumption over time. Habits accomplish the same thing, but are harder to work with. Its central empirical idea is that people care not just about current consumption growth, but separately about revisions in expectations of consumption growth in the far future. We obtain a "two factor model" with current consumption growth and innovations to news about future consumption growth as two, separate, priced factors. In this model, people fear assets that fall in times of bad news about future consumption, bad news that is *not* reflected in current consumption. For example, in this view the "state variable" that made 2008 such an awful year is *not* that current 2008 consumption fell like a stone, it's that 2008 brought horrible news about consumption in 2018. Obviously, to make this work you not only need that people *care* about long-run consumption with recursive utility, and you also need to believe and reliably measure that there *is* any news about long run consumption – that consumption really is not a random walk. In addition, to match *time varying* risk premia, these models suppose substantial variation in the conditional volatility of long-run consumption forecasts. Roughly speaking, in the context of (9), they assume substantial variation in $\sigma_t(\Delta c_{t+1})$ rather than γ_t.

I see three difficulties with this approach. First, the evidence for substantial news about long-run consumption growth is pretty weak. One can't prove it *isn't* there, but it's hard to independently measure that it *is* there. Second, ditto with substantial movements in conditional volatility. Third, and at a deeper level, do we really think that people would have been just as unhappy in 2008, and just as desperate to have investments that paid off well in that state of nature, if *current* economic conditions had not worsened, but they had the same news about poor growth in the far future? (See also Beeler and Campbell (2009) for a variety of empirical difficulties.)

Rare events models, introduced by Rietz (1988) and recently brought to great popularity by Barro (2006) allow us to imagine that the moments in (9) take substantially different values than sample estimates suggest, or vary over time in ways that sample based estimates do not show.

My main reluctance to embrace either class of models is that they tread dangerously close to unobservable, ex-post explanation. "News of consumption growth in the far future, which we can't see directly but is reflected in asset prices," or "time-varying probability of an infrequent huge disaster, which we can't see directly but is reflected in asset prices" can make sense of asset pricing data. But they are close to unobservable. It's hard to be critical of "sentiment" if this is the alternative "rational" explanation. These models also make application devilishly hard. In a rare disasters model, all that matters is covariance in the rare disaster. Monthly betas are basically meaningless measures of risk. If this model is true, it basically means applied asset pricing is hopeless.

The best hope, given that the underlying causes (long-run risk, rare disaster probability) will not be directly visible, is that models constructed on these lines will make sharp rejectable predictions *across* different assets; that the "long run risk" or "disaster" one invokes for stocks happens to provide an explanation for bonds as well. The number of predictions can be made fewer than the number of assumptions without directly measuring the fundamental shock. (Gabaix (2010)). Until then, these theories will be largely interpretive.

Given that nondurable consumption is time-aggregated and not well measured, and that much consumption comes from durable goods, I think that work emphasizing longer horizons (Daniel and Marshsall (1997)) and other goods may help to see the basic consumption-based links between macroeconomics and finance. Once one considers the variety of specification issues, rather than viewing the consumption-based model as "tried and failed," we have to recognize that it really is not yet deeply investigated.

4.1.3 Background risks, nontraded risks

An interesting, line of thought, ties asset pricing to the cross-sectional distribution of non-traded risks.

One idea emphasizes hedging non-market and non-tradeable risks including businesses and human capital. If we ask the "representative investor" in December 2008 why he or she is ignoring the higher expected returns in stock markets and the buying opportunity of a lifetime in fixed income, the answer might well be "that's nice, but I'm about to lose my job, and my business might go under. I understand that prices are low, but I can't take any more risks right now, especially in securities that will lose value if the recession gets worse."

Similarly, Fama and French's (1996) story for a value effect is that many investors have human capital tied to the fortunes of value firms. Asset markets are huge markets for insurance. Those whose outside risks are positively correlated with a security class sell, even if the securities have high average returns; those whose outside risks are negatively correlated buy. If there are more sellers than buyers, the price declines and average return rises. Then those who aren't exposed at all also write insurance.

The central assumption behind these ideas is that *markets are incomplete*, though still frictionless. Every investor can purchase or sell every asset, but the assets do not span the space of risks that investors care about. In this case, the complete-market theorem that "all risks are shared" $m^i_{t+1} = m^j_{t+1}$ becomes "all risks are shared as much as possible." If X is the set of available asset payoffs, then $p = E(m^i x) = E(m^j x)\ \forall x \in X$, but $proj(m^i_{t+1}|X) = proj(m^j_{t+1}|X) = x^*$. Marginal utility is not equated across people, but the marginal-utility-mimicking portfolio is equated. In turn, this does not mean that people share the same asset portfolio either, but adapt their asset portfolio to hedge the outside risks that drive differences in marginal utility. (Brandt, Cochrane and Santa-Clara (2006).)

It is still true here that "only aggregate risk is priced." If the same number of people are long as short a risk factor, they will use markets to transfer that risk, but it won't show up in pricing. And a risk that initially affects only a subset of the population directly becomes of aggregate concern once those not initially exposed use the markets to write insurance.

A second class of theory following Constantinides and Duffie (1996) describes the possibility that truly idiosyncratic risk affects asset pricing, and that widening in the dispersion of those risks is the factor or aggregate risk that drives asset pricing. If we define idiosyncratic risk as $\varepsilon^i = m^i - proj(m^i|X)$ so that $m^i = x^* + \varepsilon^i$, breaking marginal utility into the common component and "idiosyncratic risk" ε^i, then indeed adding ε^i has no effect on asset prices, since $p = E(m^i x)$ for any amount of ε^i. However, if marginal utility is not a linear function of consumption, then adding idiosyncratic consumption volatility can affect marginal utility m^i in a way that does affect asset prices.

The second style of model has not, in my estimation, yet worked out quantitatively (Cochrane (2007a) reviews this literature) and the first really has not bee investigated quantitatively. But the stories are compelling and both classes of models suggest interesting, sensible, and relatively unexplored links between discount rates and macroeconomic events.

They are especially unexplored relative to new utility functions. In the end, new utility functions can only do so much – they can suggest different transformations of the time series of aggregate consumption. From an empirical point of view, aggregate consumption is not very revealing about business cycles and financial crises. I think the relatively less explored area of research that tries to bring a fuller picture of the risks consumer/investors bear, and therefore a wider and more revealing fount of data, is more likely to bear asset pricing fruit than any transformation of the aggregate nondurable time series can hope to do. Of course this is as much a "measurement" as it is a "theory" issue.

4.1.4 Investment

Standard consumption-based asset pricing links asset price fluctuations to macroeconomics through consumer first-order conditions. One can and should also link asset prices to macroeconomic events through producer first order conditions as well. At a minimum, this step will have to be part of the larger goal, a general-equilibrium economic model that simultaneously generates quantity (business cycle) and asset pricing facts.

The Q theory of investment is the off-the-shelf analogue to the simple power-utility consumption-based model. The Q theory adds an adjustment cost to the standard production function. The simplest form uses quadratic adjustment costs,

$$\begin{aligned} \pi_t &= y_t - i_t = \theta_t f(k_t) - \frac{\alpha}{2}\left(\frac{i_t}{k_t}\right) i_t - i_t \\ dk_t &= -\delta k_t + i_t dt. \end{aligned} \tag{10}$$

With quadratic adjustment costs and constant returns in production $f_k(k)k = f(k)$, this model produces lovely classic results (see the appendix for derivations):

1) The investment/capital ratio is proportional to Q, known in finance as the market/book ratio,

$$1 + \alpha\left(\frac{i_t}{k_t}\right) = \frac{\partial V(k_t)}{\partial k_t} = \frac{V(k_t)}{k_t} \tag{11}$$

where $V(k_t)$ is the market value of the firm.

2) The asset return is equal to the marginal physical return on investment, i.e. invest one more dollar today, take out R dollars tomorrow, leave production at all other dates unchanged. The investment = asset return is

$$dR_t = \frac{\theta_t f'(k_t) - \delta - \frac{\alpha}{2}\left(\frac{i_t}{k_t}\right)^2}{1 + \alpha\frac{i_t}{k_t}} dt + \frac{\alpha\frac{i_t}{k_t}}{1 + \alpha\frac{i_t}{k_t}}\left(\frac{di_t}{i_t}\right). \tag{12}$$

The $1+\alpha\frac{i_t}{k_t}$ term in the denominator converts the marginal investment to marginal capital after paying adjustment costs. The numerator of the first term is the marginal product of capital, less depreciation, and then a term capturing the effect of more capital on adjustment costs. The second term is more interesting, as it is the stochastic term, which will correlated with risk premia rather than interest rate variation. *The investment return is proportional to investment growth.* Investing from a state with low investment, and hence low adjustment costs, allows one to save a great deal of investment in a future state with high investment and hence high adjustment costs.

Equations (11) and (12) accomplish our goals: they link macroeconomic variables to asset prices and returns. *Investment should be low when valuations (market to book) are low, and vice versa. Stock returns should be positively correlated with investment growth, both ex ante and ex post. Thus, expected stock returns should be high when expected investment growth is high, and ex-post stock returns should be high when there is good news about investment.*

These are 100% R^2 predictions; neither (11) nor (12) contains an error term. Thus, like the one state variable habit model, they are easy to formally reject. But like the habit model, the Q theory contains an important grain of truth, even (especially) in the recent traumatic period, as presented in Figure 16. When stock prices are high relative to dividends, or the market value is high relative to book value, which we now know entirely corresponds to a low risk premium and thus low cost of capital, investment rises strongly. When stock prices decline, as in the end of the tech era and the great recession, investment also tanks. Once again, despite the popular opinion that the financial crisis invalidates all of macro-finance, in fact classic finance works rather well in this extreme data point. Taking a time-difference of this sort of plot, the correlation between investment growth and stock returns is similarly evident (Cochrane (1991b)).

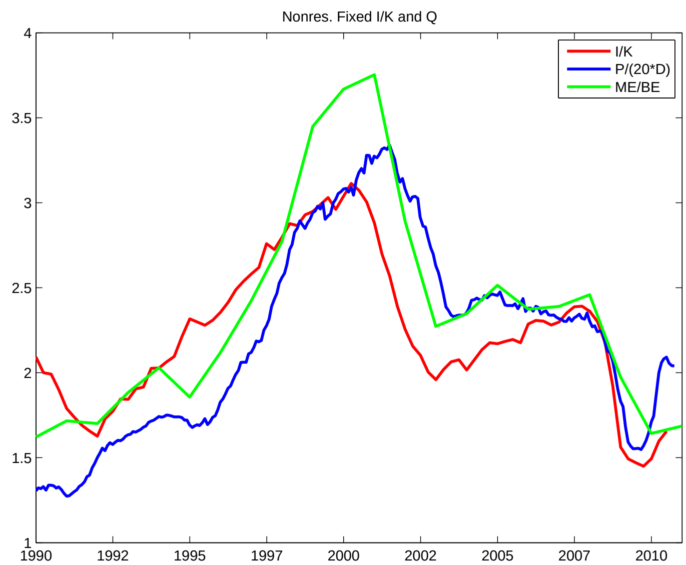

Figure 16: Investment/capital ratio, price/dividend ratio, and market/book ratio. Capital is cumulated from investment with an assumed 10% annual depreciation rate. Price/dividend from CRSP, market/book from Ken French's website.

Simple Q theory logic reminds us that the investment/capital ratio, book/market ratio, equity issuance, and price/dividend ratio are all basically the same thing, and united by the Q theory. (Li, Livdan and Zhang (2008), Liu, Whited, and Zhang, (2009).). Equity issuance is particularly useful to consider, as the fact that equity issuance forecasts returns has been interpreted as manager reaction to "overvaluation." Well, perhaps, but issuing equity and investing is exactly what the Q theory directs a "rational" manager to do.

The relationship between investment and asset prices described in (11) and (12) holds whether asset prices move on cashflow news or on discount rate news. If prices are high because, say, expected future productivity θ_{t+j} and hence profits π_{t+j} is high, it makes sense to invest. If prices are high because discount rates are low, it also makes sense to invest. Thus the movement of investment with stock prices is silent on this issue. To complete the picture that *discount rates* move, and investment responds to discount rate movements, we have to also add the evidence from return regressions that aggregate price movements (and investment/capital movements) largely forecast returns and not cashflows.

Like the consumption model, there is no sense of causality here. We are more used to thinking of investors as price-takers, as we used to think of the consumption first-order condition as running from price to quantity. However, general equilibrium does not respect this causality. As the endowment economy equilibrium example demonstrates, causality in general equilibrium can run from quantity to price, though each agent is a price taker. If adjustment costs are low, a rise in the price of capital is swiftly met by a huge increase in the supply of capital. In the limit, the rise in price can't happen in the first place.

As the power utility model only touches the surface, being really a simple textbook form, the quadratic adjustment costs model also only touches the surface, and we should expect more realistic functional forms to provide better empirical performance. Micro investment is irreversible, which can change the predicted relation between investment and asset prices (Kogan (2004), Gourio (2010)); investment can respond to blunt a rise in prices (q), but disinvestment cannot blunt a fall in prices. On the other hand, Thomas (2002) finds that irreversibility doesn't matter after aggregation for macroeconomics, and so it may not for finance. Investment certainly requires a lag after price rises – one has to commission architects, draw plans, and get zoning approval before starting to build. At a minimum, one would not therefore expect the model to work on a daily horizon, as one does not expect the consumption model with current data to work on a daily horizon. Adding more realistic dynamics should also therefore help. In this context, Lamont's (2000) finding that investment *plans* forecast stock returns better than actual investment, and forecast investment beautifully are an important confirming piece of evidence.

The general-equilibrium theory addressing the value effect has productively considered important modifications to production technology, valuing "growth options" and irreversibilities (Gomes, Kogan and Zhang (2003), Gala (2010), Gourio (2007)). Last, as I emphasize in Cochrane (2007a), the standard formulation of aggregate technology $\theta_t f(k_t)$ reflected in (10) is quite limited, in that firms can take action to transform goods over time, but they are very limited in their ability to transform goods over states of nature. Yet actual firms can choose production processes that pay off better in some states of nature than others; they can also choose the distribution of θ_t, and for example choose technologies that may do worse on average, but do much better in high contingent-claim states of nature. This rewriting of aggregate technology has really only been glimpsed (Belo (2010), Jermann (2010)).

Thinking about investment and the production side is important more generally. For example, a "bubble" can only exist if there are some frictions or adjustment costs to investment. If not, then Q must always be one; the minute a price bubble starts to erupt there is an infinite supply of new securities. As another example, Longstaff, Santa Clara and I (2003) investigated dynamics that appear in a two-tree world with no investment; after a good return in one tree, its expected return must also rise to keep agents from rebalancing away from that tree. Traditional finance models with linear technologies and instant investment/disinvestment do not have these dynamics, because we can all instantaneously rebalance, by turning trees into consumption. The extent of real two-tree dynamics, and at which horizon, depends on how quickly actual capital can flow to follow price signals. The right model combines two-tree intuition at short horizon, and becomes the linear-technology world at longer horizons.

4.2 Behavioral asset pricing and bubbles

My empirical survey stressed the large variation in expected returns over time and across assets. The "macroeconomic" view uses rational expectations (subjective probability = objective frequency) and tries to explain the fact by variation in risk premia. Behavioral asset pricing is, I think, centrally devoted to a different central assumption: *People's expectations are wrong.* Low prices are times or securities with irrationally pessimistic cashflow expectations and high prices vice versa. As a constructive definition, behavioral asset pricing takes lessons from psychology, lab experiments, and surveys to find systematic and predictable patterns to the wrong expectations.

I think this distortion in expectations is the crucial assumption behind behavioral asset pricing. (I acknowledge that any attempt here to state "the" assumptions of behavioral asset pricing is sure to run in to objections. However, the focus on mistaken expectations surely captures at least one important strain in behavioral asset pricing. See Barberis and Thaler (2003).) By and large, really, behavioral asset pricing maintains the frictionless market assumption and risk sharing, or at least these assumptions are not central. There are some frictions in many behavioral models, but these are largely secondary, put in place to keep "rational arbitrageurs" from coming in and undoing the behavioral biases. These assumptions are often not really needed. Expectational errors induce risk premiums but not arbitrage, so they could only be eliminated by risk neutral arbitrageurs. Risk averse "arbitrageurs" have limiting risk bearing capacity, so expectational errors by any mass of agents will still affect prices in a frictionless market.

The risk-neutral expectations theorem unites the "behavioral"and "rational" approaches. Any price that can be written as an expected discounted payoff,

$$p = E(mx) = \sum_s \pi(s)m(s)x(s),$$

can also be written as a pure expected payoff, using a distorted set of probabilities E^*:

$$p = \frac{1}{R^f}E^*(x) = \frac{1}{R^f}\sum_s \pi^*(s)x(s).$$

We can construct the "distorted" probabilities from true probabilities and discount factor:

$$\pi^*(s) \equiv R^f \pi(s)m(s).$$

(And $1/R^f \equiv \sum_s \pi(s)m(s)$.)

Conversely, *"Irrational" expectations are the same thing as a distorted discount factor* Given "irrational" expectations $\pi^*(s)$, we can construct a "discount factor" $m(s)$ that produces the same prices.

Stated at this level, "rational" and "behavioral" theories are observationally equivalent – and both are vacuous. Whether you call it $\pi^*(s)$ or $\pi(s)m(s)$ makes no difference, at least until you put some restrictions on how to construct π^* or m. If you say "prices went down, sentiment must have turned" that is the same thing as saying "prices went down, the risk premium must have risen."

Observational equivalence goes a long way to explaining how a 40-year old argument can continue with no resolution – and why further argument on the same lines is pointless. If we want to get somewhere, we need first to agree to rule out ex-post story-telling and residual-naming contests, and bring additional data or rejectable theory to the table. None of these considerations are scientific proof, of course, they are just my motivation for searching in one garden rather than another.

Most obviously, each side needs to

- *Tie discount factors, or (equivalent) distorted expectations to other data.*

One needs to say, "given (say) consumption growth, here is what I think the discount factor is, and therefore here is what my model says stock prices should be today." My habit model with John Campbell (1999) does this – given a time series of consumption, it generates a time series of what we think the stock price should be at each date. Of course it fails – it does not match the actual stock price data point for data point. We see the glass half full, and see many patterns matched in other stock prices. But one can also see the same glass half empty. There is a residual, asking to be named. Well, that's the way with rejectable theories.

A big challenge to behavioral theories is to forge this link. Give us *an external* measure of "sentiment," or "optimism and pessimism," not drawn from the same asset prices; construct a model of what you think prices *should* be, and contrast actual prices with those values. Escape the charge that this is ex-post storytelling.

- *Do something else to reduce the ratio of assumptions to facts*

This is a pretty high standard, but lower goals will help. Really, the point is to reduce the ratio of assumptions to facts explained. I suggested that long-run risk or rare disasters theories, also grounded on fundamentals that are close to unobservable, could escape the charge of residual naming by *connecting* the behavior of multiple markets or prices. The habit model also connects a variety of phenomena with one structure, and can be judged on that ground. Behavioral models could do the same thing. Show that the theory makes a tight connection that if fact x (say, value effect) holds, then fact y must also hold (say, momentum), in a way that would not be easy to overturn with equally plausible assumptions if the facts went the other way. This, I think is what Fama (1998) was asking for, in criticizing the use of "overreaction" in one context and "underreaction" in another.

While waiting for models and connections to data, here are some observations that have motivated me to spend my time looking for answers on the "rational" side of the fence.

- *Correlation with macroeconomics.*

The strong correlation of time-variation in discount rates with macroeconomic events is surely suggestive, and it is a central part of the "rational" story for time-varying risk premiums. When prices are high and expected returns are low (1996), the "rationalist" explains, "an economic boom is going on. People aren't worried about jobs, businesses, or financing. They're happy to "chase yield," as they feel they can absorb systematic risks. They're holding high-priced assets knowing full well expected returns are low." The behaviorist says "no, this is a case of contagious overoptimism; returns will be low when they wake up," but there is nothing intrinsically connected to macroeconomics in this view. Similarly, in, say, December 2008, the "rationalist" might say "yes, people are leaving the buying opportunity of a generation on the table (high expected returns), but that makes sense – they're all worried their jobs or businesses might fail, or we might be in a great depression. Nobody feels they can take risks now." The behaviorist might answer that this is a moment of collective excessive pessimism, and returns will be high as people calm down. Again, there is nothing intrinsic about a recession in that view. Why is it that people become irrationally pessimistic in recessions, just when "rational" theories say they should, and vice-versa in booms? This requires some psychological mechanism that makes people overoptimistic based on economic conditions a thousand miles away, or perhaps via news media. I don't know of psychological foundations or lab experiments that tie irrational optimism and pessimism to macroeconomic events, so this view requires some new psychology.

Alternatively, it requires some new theory of macroeconomics in which the waves of optimism and pessimism *cause* recessions.

- *Pervasive and correlated across assets.*

There is a strong common component to the time-varying expected return across all asset classes. This also is suggestive of a rise in economywide risk aversion. The basic model is $E_t(R^{ei}) = \gamma_t cov(R^{ei}_{t+1}\Delta c_{t+1})$, and the fact that expected returns are correlated across assets suggests γ_t is at work. It would be a bit embarrassing if we had to believe in big variation in conditional covariances or factor-specific market prices of risk.

Many psychological biases lead one to narrow framing, salience of recent experience, etc., which more naturally predict that optimism and pessimism would be focused on specific assets or asset classes.

- *Association of mean returns with common movement.*

Each "anomalous" dimension of mean returns is associated with a new dimension of covariance. The value stocks all move together, momentum stocks all move together, and so on. Behavioral theories can be good at explaining patterns of *mean* return. For example, theories of slow reaction to information can generate the positive autocorrelation in individual returns that is behind the momentum effect. But such theories haven't really tried to address the *covariance* of returns. If psychological biases generate positive autocorrelation in returns, why do all the positive momentum stocks *fall together* next year if they fall? What is the behavioral explanation for a value or momentum factor in the *covariance* of returns? As I have emphasized, every anomaly is equally an anomaly of expected returns and betas or covariance structure. As Figures 8 and 10 remind us, the appearance and disappearance of anomalies can be as much about changes in covariance structure as changes in expected returns, which behavioral or learning stories might address.

As the pattern of mean returns over time looks just like what a rational model, driven by business cycles, might predict, so the pattern of mean returns across assets looks just like what a rational model, driven by many sources of *common* risk across assets might predict.

- *Survey evidence.*

Behaviorists often point to very unsettling survey evidence of people's actual expectations. In boom times, people report wildly enthusiastic return and cashflow forecasts and vice versa.)

Managers, including venture capitalists, seem to follow the (incorrect) advice we have taught them for years, at best discounting cashflows with a very constant 6 or 7% market expected return, ignoring time-varying expected returns and multiple factors. Despite our bad advice, they seem to get the right answer, as dramatized by Figure 16, investing more when the cost of capital is lower and vice versa. They do that, it seems, by inserting optimistic cash flow forecasts when times are good, and pessimistic cash flow forecasts when times are bad.

However, in interpreting surveys or lab experiments, we always have to be very careful about language. It doesn't take long in teaching MBAs to realize that the colloquial meaning of "expectation" and "risk" is entirely different from our idea of "conditional mean."

If people report the *risk-neutral* expectation in response to a survey, then we're entirely back where we started, and the expectations are in fact completely rational. And, though they surely will not know the language, thinking about and reporting this expectation makes enormous sense. The risk-neutral expectation is the expectation weighted by marginal utility, $E^*(x) = k\ E(u_c(t+1)x_{t+1})$. For many decisions, this is the right sufficient statistic. If you're deciding between two options, the one with the higher *risk-neutral* or *marginal-utility-weighted* expectation is the better option, not the one with higher expected value. Rational decision makers should always think in terms of risk-neutral probabilities, not objective probabilities!

Related, the colloquial concept of "risk" has little to do with "variance." We all know the difficulty of explaining to an introductory class that "upside risk" is the same as "downside risk." "Expectation" often means "what I think will happen if everything goes as I hope" and "risk" is all "downside." This actually makes sense as well for many economic or engineering projects which have severely skewed distributions. The conditional mean and variance is a poor descriptor of the risk of a rocket launch.

This is how I decode the fact that venture capitalists typically say they discount cashflows with absurdly high discount rates, on the order of 30% or so, and require 30% "hurdle rates" to accept projects. Of course their cashflow "expectations" are not conditional means, they are "if everything goes right" forecasts. The 30% "discount rate" is a good way to account for the left tail of "risk." In sum, perhaps the responses to surveys simply reflect usual practice of jiggering the numbers to get the answer you want out of any model – a splendid example of "as if" rationality. Perhaps the wildly positive expectations in good times are simply ways of saying "it's a good time for me to invest" – which it is if one's discount rate is low – and vice versa.

At a minimum, the statement of a survey about expectations should clarify, "is that the true-probability-measure expectation or the risk-neutral-probability measure expectation you just quoted me." The absurdity of asking that question shows the equal absurdity of assuming the answer was given one way or another.

- *But let's not overstate the evidence on the other side either*

It is also a mistake to conclude because either consumer or producer first-order conditions work well, as suggested by my plots (16) and (15) that "markets are rational." Producers might be reacting completely rationally to fads or bubbles induced by consumer irrationality, or vice versa. Only a full general-equilibrium understanding of asset prices (along with quantities) really answers that question.

4.2.1 Boundaries

In this discussion, I focus on *asset pricing*, not on *individual portfolio formation.* The whole field of documenting biases in individual decisions is fascinating, but beyond the scope of this essay.

The boundary between "behavioral" and "rational" is not well defined, as the concept of "rational" is not well defined. There is much work on strange preferences in finance and macroeconomics. In the end, a "utility function" is just a way of modeling empirical facts about how people make decisions confronting risk and delay in rewards, as seen in macroeconomic and financial data, as well as to incorporate insights from psychology including surveys and lab experiments. These include the development of recursive or non-state-separable utility (Epstein and Zin (1989)), dynamic "utility functions" to capture the insights of prospect theory (Barberis, Santos, and Huang (2001)), habit persistence preferences, robust control theory (for example, Hansen and Sargent (2005)), hyperbolic discounting (Laibson (1997)) and others. I can't think of an operational definition of "rational" which neatly categorizes any of this work. The only empirical way to measure which of these gets called "behavioral" and which "rational" comes down to an unsavory analysis of which citations authors choose to make and which conferences they go to, rather than some discernible intellectual content.

So in my view, the whole "behavioral" vs. "rational" debate is just silly, as any debate must be that has gotten nowhere in 30 years. We're arguing about nothing. All asset pricing does is try to model human behavior, figure out consequent supply and demands for financial assets, and try therefore to understand patterns in price movements. Give us your model, its predictions, and the data. We judge models by how well they fit the data, along with the relative paucity of assumptions relative to predictions, and a sense of how hard it would have been for the modelers to make the opposite prediction if the data had come out the other way. I am sure even the most ardent behavioralist would not want to define "rational" as "anything that can be written down, formalized, and able to make quantitative rejectable predictions" leaving his own field perpetually only for ex-post residual naming. Thus, even framing what asset pricing is about as a debate between "rational" and "irrational" or "behavioral" explanations is pointless. This framing does not fruitfully describe what theoretical or empirical financial economists *do.* Framing what we do as "understanding the sources of discount rate variation" is, I think, much more fruitful.

4.3 Finance models

Most of applied finance uses a different sort of model, in which discount rates are determined in terms of a few large factors, such as market, size and value portfolios, or principal components of the term structure.

The theoretical development of these multifactor models have long roots as "macroeconomic," models purporting to offer deep explanations for discount rates. All of the textbook derivations of the CAPM, ICAPM, and other multifactor models are specializations of the consumption-based model, and maintain its central assumptions. By adding specializations (i.i.d., log utility, no outside income, etc.) they characterize the composition of the marginal utility mimicking portfolio of asset returns, allowing us to bypass the use of consumption data.

The central assumptions behind the CAPM are 1) Returns are i.i.d. and 2) Investors don't have jobs or outside income. The evident fact that both are false means we should not be at all surprised that the CAPM fails and new factors emerge. The surprise is that it took so long to find what theory had anticipated all along. Given the huge variation of expected returns over time, we *should* see ICAPM factors. If DP forecasts returns, then innovations to DP should be a second priced factor. Of course, DP innovations are highly correlated with market returns, so disentangling the orthogonal component is not so easy. (Campbell and Vuolteenaho (2004).) In turn this observation suggests a natural route to making sense of new factors – perhaps they are correlated with innovations to variables that forecast returns Even the simple CAPM implies multiple factors, as the *conditional* CAPM with time-varying betas and market risk premiums implies a multiple - beta *unconditional* representation. (However, it's not clear why one would expect the CAPM to hold since its derivation with long-lived investors needs i.i.d. returns.)

The presence of human and nontraded capital is a very intuitively plausible source of additional factors, as explained above. As I sniff through many of the multiple factors in the cross section, a whiff of "recession risk," aversion to holding assets that do poorly when people might lose jobs, companies might fail, and so on, above and beyond market portfolio, permeates many of them.

However, empirical work seems hardly constrained by these derivations. Since Stambaugh (1982), nobody has seriously tried to use a wealth portfolio return in the CAPM, and even bond returns are seldom included. Derivations of the CAPM say consumption growth should move one for one with the market return, a prediction never checked. The ICAPM is often mentioned as inspiration for multiple factors, though few actually check whether additional factors do in fact represent innovations to state variables that forecast market returns as the theory requires, (Liew and Vassalou (2000), Brennan, Wang and Xia (2004), Brennan and Xia (2006) are notable exceptions), and less still whether multiple factor risk premia make sense as partial derivatives of a value function.

It's easy to forget, but in none of these theories is factor *structure* either necessary or sufficient for factor *pricing*. If the consumption-based model were true, it is possible and indeed likely that innovations in consumption would still have little explanatory power for ex-post returns. Conversely strong comovements such as industries need not be priced. The coincidence of factor structure with factor pricing is convenient, of course, and certainly worth pursuing as a very useful data reduction fact if it's true.

Applied work draws most closely really from the APT (Ross, (1976), (1978)), which does make this link. It just states that patterns of mean return must be matched by patterns of covariance, or large Sharpe ratios will emerge. If we have a factor representation

$$R^e_{t+1} = \alpha + \beta f_{t+1} + \varepsilon_{t+1}$$

where R^e, α, and ε are $N \times 1$ vectors and β is an $N \times K$ matrix, then the maximum Sharpe ratio we can obtain by portfolios of returns and factors is given by

$$SR^2 = E(f)' cov(f, f')^{-1} E(f) + \alpha' cov(\varepsilon, \varepsilon')\alpha$$

small ε must imply small α or large Sharpe ratio. Even here, though, the connection is tenuous as time-series R^2 or formal Sharpe ratio bounds are seldom computed.

For all these reasons, Fama (1991) called the typical application of theory a fishing, just as he and French were landing a whale.

Yet if we are going to ignore this theory, we have to deduce some other rules of the game. What alternative "theory" or logic gives any rules to what is or is not allowed as a "factor" and when one can or cannot say that factor pricing achieves an "explanation" in the introduction or conclusion? This is an honest question, and not a critical comment. The right approach to empirical work is to figure out what authors were really trying to accomplish, and then maybe think about how to do it better, not to write starchy essays inveighing impractical purity. Methodology should be an empirical science too.

4.3.1 Different purposes, different models

So what is the "right" form for finance models? If we try to understand what people do, it's pretty clear that authors have different purposes, different questions in mind, and that the right model depends deeply on the author's purpose. I think there was an optimistic vision, in the optimistic 1960s and 1970s when finance models were being developed, that one model would emerge and become the standard for every purpose. To understand current empirical work, continue it and improve it, we have to recognize that dream fell apart.

4.3.2 Data reduction for deeper explanation

One useful view of what empirical work is actually achieving is data reduction in preparation for "deep explanation." Fama and French (1993, 1996) tell us that all the pricing information in 25 size and book/market portfolios can be summarized in two factors It's silent on why those factors receive premia. Now deeper theories just have to explain the premiums of the size and book/market factors. Similarly, one can view the CAPM as a one-factor reduction, leaving for economics the (hard) question why the market premium is so large.

A little more generally,empirical work with finance models documents expected returns as a function of characteristics $ER^e(C)$, and then relates those patterns to the factor structure of returns, also a function of characteristics, $cov(R^e, f) \times \lambda_f = cov_f(C) \times \lambda_f$, where factors are also formed on the same characteristics. If (as with size and book/market), all we do is boil down the puzzle posed by the whole expected return function of N characteristics $ER^e(C)$ to N factor risk premia $\lambda_f = E(f)$, that is already progress. If in addition, the dimensionality of priced f with $\lambda_f = E(f) > 0$ is less than that of C, we achieve even more data reduction.

With data reduced this way, "macro" pricing needs only to figure out the *factor* discount rates, or why the *factors* are priced, $E(f) = cov(f, m)$, or $p_f = E(mf)$. It's much easier to focus a theory on why hml and smb are priced than it is to develop, calibrate, and test a theory on the entire Fama French 25 portfolios, to say nothing of the underlying individual stocks and their large vector of accounting characteristics.

So we have a very nice picture of asset pricing research: empirical work boils down the alarming set of anomalous expected return claims to a corresponding set of priced factors, large-scale systematic risks that generate rewards. Hopefully, that number of factors is smaller than the number of expected-return variables, either by multiple regression logic that some expected-return characteristics are just proxies for others or that some expected-return characteristics correspond to covariances with old factors, not needing new factors. "Macro" "behavioral" or other "deep" theories can then focus on why the factors are priced. In the end we meet in the middle.

This is a sensible vision for our end goal. Even if the consumption-based model with power utility were literally and perfectly true, lags and difficulty in measurement would make risk-adjustment by consumption betas an error-prone exercise at best. When examining an anomaly, evaluating a fund manager, or any of the other workday tasks of discount factor models, one would be far better off estimating and using the mimicking portfolio for consumption growth than the government's data filtered through a utility function. Furthermore, the mimicking portfolio for consumption growth might be a poor "factor," it would likely leave a low R^2 in time-series regressions and therefore large uncertainties in measured alphas. Thus, even if the consumption-based model were perfectly true, we would still want to understand the factor structure of returns, and it would still be useful to relate individual returns to multiple priced risk factors. We would merely have a deeper understanding of what the multiple risk factor's premiums were and where they came from.

Conversely, in this vision it doesn't make much sense to subject macro models to horse races, i.e. to test them by whether they can produce smaller alphas on something like Fama French 25 portfolios than do factor models such as Fama and French's. Even if we had the perfect model, it would be beat by its mimicking portfolios in such an exercise due to measurement and time aggregation alone. (Campbell and Cochrane (2000) give a quantitative example, in which the CAPM beats the true consumption-based model in simulated data.) The point of the macro model is to explain the factors.

I think Fama and French (1996) were in part aiming towards deep explanation of this sort, that size and value factors would eventually be shown to be "mimicking portfolios for state variables of concern to investors," i.e. $E(mf) \neq 0$, or $proj(m|X)$ includes f. This is probably a good part of the reason that Fama and French stopped at size and book/market factors in 1996, though they saw momentum as an anomaly and that a "momentum factor" would "explain" it. However, though they had some paragraphs of elegant prose on what kinds of macroeconomic models might lie behind the value factor, neither they nor anyone since has been able to come up with similar prose for momentum factors. This, along with some hope that momentum might go away with deeper study, led them to focus only on size and book/market factors. So we infer an important restriction on the number and nature of factors one should consider with this end in mind: you need at least a plausible story that the factor is a mimicking portfolio for some aggregate risk.

4.3.3 More factors for less lofty purposes

Performance evaluation. Many practical applications have less lofty purposes, however. For example, Carhart (1997) did use a momentum factor for evaluating the performance of fund managers. Even with no macroeconomic justification, this is the right thing to do *for Carhart's question.* Carhart wanted to know whether fund performance reflected stock-picking skill, or simply reflected momentum in the underlying stocks. When he says that he can "explain" fund returns with a momentum factor, he means that one can achieve the same returns as actively-managed portfolios do without the judgment or fees of the active manager, simply by programming a computer. Even if the momentum factor is "irrational," spurious, sample-specific, or otherwise wrong – if the phenomenon *in stocks* is unexplainable – it remains true that he reduces the puzzling behavior of managers to a much less puzzling, and already well-known behavior of securities. *For this purpose* any factor that is well established to work in the assets which managers trade can be included. (However, the explosion of factors is making the whole active/passive, alpha/beta, inefficiency/systematic distinction meaningless, a point I take up in section 7.4 below.)

Anomaly digestion. Similarly, the day to day job of empirical asset pricing is often digestion of new anomalies. After figuring out if an anomalous expected return survives the usual significance and selection biases, one wants to know, is this a genuinely new dimension of systematic risk and premium, or is it just another manifestation of a known anomaly? Fama and French's (1996) analysis of sales growth is a great example. Sales losers have high average returns. However this fact is completely "explained" by higher betas on hml. "Explained" here does not mean in a deep macro sense, because we don't know why the value premium is there. But it does mean that "this is just another manifestation of a known anomaly." For this purpose we again do not need to be too picky about the nature of the factors. What matters is only that the factors are well established.

In both cases, one also has to be more careful about writing introductions and conclusions of course; if one has "explained" a new anomaly or a manager return by momentum betas, that means only that it is reduced to an existing puzzle.

This sort of anomaly reduction was part of Fama and French's purpose, and their greatest success – to provide a benchmark for routine everyday risk-adjustment of new anomalies to a smaller set of known factors. Now that we know how well momentum has held up, momentum is and should be added to this sort of exercise, though it is still singularly lacking in "deep" explanations.

Price taking applications; relative pricing. Much of applied finance comes down to "price taking applications." Cost of capital, portfolio formation, pricing of new or illiquid securities (pricing a swap or an option), and project valuation, are all cases in which we just want to know what prices or risk premia are, and we don't really care that much about where they came from.

To decide how many tomatoes to buy, you only need to know the price of tomatoes, not whether this price is high because there was a frost in California or because there is a bubble in the tomato market. Similarly, portfolio theory or investment theory only need to know the amount of expected return and factor exposures to make the right decisions, not whether the discount rates announced by the market are arrived at "rationally" or not, and if so how.

More generally, these questions are "relative pricing " questions rather than "absolute pricing" questions. "Absolute pricing" seeks to understand the determinants of a discount rate via correlation to basic macroeconomic risks. But most of applied finance cares nothing about such large questions. In these applications we only care about "relative pricing," is this asset priced consistently with all the others, especially relative to assets that one might use to hedge or finance a given position. Even most applications of the CAPM work this way. "Relative pricing" takes known prices (discount rates) and tries to extend them a bit, without asking where they come from. The rules for "factors" used in such exercises are a good deal less stringent. And the conclusions so obtained are more modest.

4.3.4 The mean-variance warning

This discussion does not mean that anything goes, and one can ignore theory all together in factor-fishing. Finance models must always but up against the Roll (1977) theorem; *there is a single-factor representation that will explain any data set*: An ex-post mean-variance efficient portfolio will always price assets perfectly in any sample. In discount-factor language, one can always construct a discount factor x^* as a linear combination of asset payoffs that perfectly prices those assets $p = E(x^*x)$. Exactly. In sample. One can always construct a single factor in such a way to generate *zero* alphas, exactly.

Only the restrictions one puts in the factor search process stops one from rediscovering this tautology. As I have described the exercise, there is very little in the rules of the game that keeps us from accidentally rediscovering the theorem and producing nice tables of small alphas, or pretty "actual" vs. "predicted" plots.

One of those restrictions, of course, is that one should be reaching towards something that is vaguely plausible as the mimicking portfolios of a macro (or behavioral) model. But that restriction is hard to enforce, and is not needed for many practical applications.

Reading the methodology that has developed in empirical finance, we can see that lots of what empirical researchers do is designed to allay this fear in an informal way. Why present tables of average returns and betas, and stress the consistency of those tables as a function of characteristics, rather than just present an overall test of model fit based on joint zero alphas, which is all that formal statistics would suggest? Well, if we were just finding a spurious ex-post mean-variance efficient portfolio, those tables would look much different, with arbitrary coefficients suggestive of overfitting. This is why we look for confirmation of results in longer samples, international data, or other markets. It's why referees demand 10 tables where one would do.

4.3.5 Challenges for finance models

1. Taming the zoo of factors The first challenge for traditional finance model is to bring order to the exploding number of dimensions of expected returns and corresponding priced factors, as explained in section 3. Ideally, we hope for two things: First, that the many priced factors can be understood and their explanatory power in applications replicated by a smaller number, as (for a while) first the market, and then size and value factors were able to do. Second, even if that goal cannot be reached, we would like to understand the many priced factors as reflections of a simpler underlying factor. For example, it would be lovely if a single macro or consumption-based model could explain the size, value, momentum, and other factors.

In some sense, of course, this is easy – there is always a mean-variance efficient portfolio, so even size, value, momentum, etc. can already be boiled down to one factor. But the nature of that factor has to become clearer for such a boiling down to be useful, for us to understand there are not fundamentally 27 separate sources of risk. I think that clarity will only come if the single or lower dimensional factor has an independent source rather than just being an empirically determined sum of market, size, value, momentum, etc. factors.

2. Betas, covariances. Second, betas are a huge gaping hole in the situation I have described. Even if we understand expected returns as compensation for a value premium, and if we understand the macroeconomic determinants of the value premium, we don't have a clue why one portfolio has a higher covariance with hml than another portfolio. Efforts to understand even value betas in terms of underlying fundamentals are few (Fama and French (1995) for example) and nonexistent for other puzzles such as momentum.

In the past we might have been happy presuming that betas are cashflow betas, covariances are driven by cashflow covariances, and direct measurement of covariances via asset returns suffices. But since betas are largely discount-rate betas, the beta question is just as endogenous as the expected return question. We haven't explained anything until we understand the betas.

A nasty fact is that betas and covariances are hard to measure, and unstable. In theory we should learn second moments much more quickly than first moments; in the diffusion limit any nonzero data sample perfectly reveals second moments. Practice certainly does not work out that way! In fact, practice has worked out the other way: The fact that small stocks and value stocks move together did not come from a generation of factor analysis of covariance matrices. It was discovered *after* value and small were found to be characteristics driving mean returns. Apparently *means* are, so far, a better way of understanding covariance structure, or at least the priced factors.

All this suggests to me, first, that characteristic-based measures of covariance should be as useful as characteristic-based measure of mean have been. If we look for $cov(R^i, R^j) = g(C^i, C^j)$, we may find it, and it may be more stable than measures based on firm name as the only characteristic.

Dynamics in expected returns means there are dynamics in covariances as well. Daily, monthly, annual and 10 year betas will be different, conditional betas will vary over time. It's widely remarked that "correlations are larger in big market moves," though the standard market model $R^{ei} = \alpha_i + \beta_i f_t + \varepsilon_i$ makes exactly this prediction – a large f movement increase the correlation between R^i and R^j. But this observation really means that one has to understand the factor structure of the covariance matrix.

However as we think we see an underlying stability in expected returns as a function of characteristics, $R^{ei}_{t+1} = a + bC_{it} + \varepsilon_{it+1}$ implies $ER^e(C)$, we may also find underlying stability in thinking of covariances and betas explicitly as a function of characteristics.

3. Prices and betas

More deeply, I think that understanding covariances in the discount rate world will require the same shift from "expected returning" back to "asset pricing" that I suggested in section 3. In fixed income, we don't find covariances empirically, and study one-period average returns relative to empirically-determined covariances. We derive "betas" – duration, exposure to level slope curvature and default factors – by starting from present value models which find *prices* and then differentiating that price model. In a non-i.i.d. world, and especially in a world with large discount rate variation, this is really the only hope for saying we understand covariances.

4.3.6 Prices

More deeply, as I argue throughout, I think that the facts of large discount rate variation means we will have to recast asset pricing theory as asset pricing – with price, book/market, dividend yield as the "left hand variable" being explained rather than sorting variables or predictor variables for expected returns, with no special status.

Easy to say, hard to do. It's easy to say "focus on prices, cashflows and discount rates." But it takes both artistic and technical advances to know how to do it in a useful simple and productive way.

In a formal sense, of course, it doesn't matter whether you look at returns or prices. $1 = E_t(m_{t+1}R_{t+1})$ and $P_t = E_t \sum_{j=1}^{\infty} m_{t,t+j} D_{t+j}$ each imply the other. However, we have long seen in applied work that how you look at things makes a great deal of difference. For example, treating momentum as a portfolio-sort rather than a time-series regression let us see its economic importance in a way we did not see before, even though the two representations are equivalent. Mathematical equivalents don't mean intuitive equivalents.

Also, small errors and appropriate approximations are not invariant to the integration or first differencing that transforms a return view to a price view. Small, easily ignored but very persistent mistakes or approximations in understanding expected returns can add up to huge mistakes in understanding prices. For example, at a 2% dividend yield, $D/P = (r - g)$ means a "large" 12% price error is the same thing as a "small" 10bp/month return error. Conversely, small price errors can have a huge impact on returns. A tiny i.i.d. price error induces the appearance of mean reversion where there is none, and common procedures amount to taking many differences of prices, which amplify the error/signal ratio. A forward spread $f_t^{(n)} - y_t^{(1)} = p_t^{(n-1)} - p_t^{(n)} + p_t^{(1)}$ is already a triple-difference of price data.

We should not make light of the technical and artistic challenge. Sure, it's easy to write down a present value equation, and say "OK, now give us your dynamic model of dividend growth and discount rate and we'll test it." That approach leads to a huge uninterpretable black box. The affine term structure literature is justly famous for finding *useful specializations* of a present value model. That is our goal for pricing of risky payoffs as well.

The "art" part of all applied economics is figuring out how to use two-period or two-good intuition in a multiperiod or multi-good world, so that we can apply complex models. I think that's how we got to the focus on one-period expected returns in the first place. It's worth remembering that fact to think about how we can make a similar but more productive mapping for a discount-rate driven world.

Finance is easy in a one-period world. Then $R_{t+1} = D_{t+1}/P_t$ so all return uncertainty and betas come from cash flows. Returns are also stationary, so we can easily focus empirical work on returns if we have repeated observations of one-period worlds. It's easy enough then to translate back to prices once we understand returns in such a world, $P_t = E_t(D_{t+1})/E_t(R_{t+1})$.

This easy thought continues if the world is i. i. d. Then

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} = \frac{(1 + P_{t+1}/D_{t+1})\, D_{t+1}/D_t}{P_t/D_t}.$$

In a truly i.i.d. world, P_t/D_t is constant so again returns only depend on dividend growth. If expected returns are constant, then P_t/D_t only varies on news of future dividend growth, and we almost have the same simplification. But this happy state of affairs no longer works, now we know that at least a substantial part of P_{t+1}/D_{t+1} reflects discount-rate news.

I think we got where we are because the constant expected return assumption looked so good for so long, so understanding prices and present values by understanding monthly returns made sense. It allowed us to use one-period logic in a difficult multiperiod world. But if we had known all along as the theory of finance was being developed that discount rate variation was so large, we might never have gone down this path, and studied prices and present values from the outset.

Similarly, Markowitz's (1952) brilliant insight was to put mean and variance of portfolio return on the axes in place of apples and oranges from demand theory. This step reduced preferences over a distribution to two understandable dimensions. Arrow and Debreu contingent claims allow a different way of mapping two-period intuition to preferences over random variables. While in some sense formally equivalent, mean-variance and discount-factor/contingent claim views of the world lead to surprisingly different insights in applied work.

The translation of time-series puzzles to cross-sectional puzzles by forming sorted portfolios has a similar beauty. Understanding time-varying expected returns with time-varying covariances with some discount factor or time-varying risk premia is really head-spinning. Translating the whole thing to an i.i.d. cross-section is much easier, and quickly leads to greater insights. Lustig, Roussanov and Verdelhan's (2010) finding of factor structure across country portfolios sorted by interest rate differentials is a beautiful example. These regressions had been around 20 years, but "conditional covariance" was hard enough that nobody thought to uncover this obvious factor.

So, the theoretical task ahead is, how to focus on prices and payoffs, without getting bogged down in big black boxes that hide intuition?

The Campbell-Shiller approximation is a useful way to start. Write it as

$$p_t - d_t = \kappa + \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j} - \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j} \tag{13}$$

and compare with the one-period version

$$p_t - d_t = \Delta d_{t+1} - r_{t+1}$$

Obviously, the Campbell-Shiller decomposition allows us to use two-period intuition directly in multiperiod, dynamic problems – but it tells us to focus on *long run* dividend growth and returns, not *one-period* returns. This may provide a useful simplification. For example, if we turn the Campbell-Shiller approximation around,

$$\sum_{j=1}^{\infty} \rho^{j-1} r_{t+j} = \kappa + \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j} - (p_t - d_t)$$

it tells us that *long-run* return uncertainty all comes from cashflow uncertainty. Hence, *long-run* betas are in fact all "cashflow betas." But in a dynamic world, such "long-run" betas may have little to do with one-period betas, either $cov(r_t, r_t^m)$, or $cov(\Delta d_t, \Delta d_t^m)$.

The world of fixed income has already made this step. Very little fixed income research is done with expected returns and betas, taking betas as statistically determined and "exogenous" at least in interpretation. Instead, fixed income research uses explicit (usually "affine") models that derive prices from expected discounted payoffs, and derives factor exposures.

The technical challenge is to do the same thing, in a productive way, for pricing of risky streams. Several generalizations of the affine structure have been investigated, and this is a promising line of research. (Ang and Liu (2004), Lettau and Wachter (2007), Gabaix (2010)) As in fixed-income pricing, this line of inquiry does not model discount rates in terms of expected returns, in the mode of (13). The affine term structure model analogy suggests instead exponential-affine models of the discount factor, m in $P_t = E_t \sum_{j=1}^{\infty} m_{t+j} D_{t+j}$. The cost is that we then lose direct contact with the vast archive of empirical work focused on expected returns.

It's worth remembering that most affine models do not capture risk premia as well; that "integrating risk premia in affine models" is only a recent inquiry (my favorite, of course, is Cochrane and Piazzesi (2008)). The purpose of affine term structure models is largely arbitrage pricing; to pretend that there is an exact factor structure with no error so that one can fill in the term structure or extend pricing from term structure to term structure options by arbitrage alone. They are often derived under the risk neutral measure. This is fine for their purposes, and even for deriving betas. For thinking about discount rates, however, we will have to incorporate risk premia and solve models under the real measure.

That all sounds hard, but a focus on prices and long-run payoff streams may end up being simpler as well. State variables for dynamics can appear in the description of one period returns, but disappear in a prices and payoffs perspective. For example, a 10 year indexed zero coupon bond is a complex one-month security, justified to a long run investor by the fact that its one-period returns are highly correlated with the state variable (yield) for its long run investment opportunities. At a 10 year, price and payoff perspective it is simply the risk free rate. More generally, as I explain in section 7.3 below, state variables for time-varying investment opportunities can disappear from long run *pricing* results. Similarly, we have already seen in the cay example how a forecasting variable can be very important for short run returns, but hardly contribute anything to pricing at all.

Often, when something seems really hard, the secret to success is to avoid asking the question in the first place. If we can avoid having to model the details of time-varying expected returns, betas, and premiums, we may end up at useful results with much less effort. Thus, I have at least a hope that a prices and payoffs perspective will also cut down somewhat on the zoo of factors.

5 Theories with frictions

The three categories of theory I mentioned so far really maintain the frictionless assumption. A burgeoning new class of theories emphasizes frictions to obtain deviations from the "macro" style of modeling.

At heart, these models basically maintain the "rational" assumption, instead deriving departures from classic asset pricing by the imposition of frictions. Admittedly, there are often "irrational" agents in such models. However these agents are usually just convenient shortcuts rather than central to the vision. A model may want some large volume of trade, or not to specify completely the motives of "noise traders," while focusing clearly on the institutional mechanism that must adapt to demands from various traders who are really outside the model. For such a purpose, it's easy to simply allude to a slightly irrational class of trader and not spell out their motives from first principles. However those assumptions are not motivated by deep reading of psychology or lab experiments.

I think it's useful to distinguish three categories of frictions: 1) Segmented markets 2) Institutional finance or intermediated markets and 3) Liquidity premiums. Consistent with my overall theme, each class of theory in the end comes down simply to a specification of a discount rate.

5.1 Segmented markets.

In a segmented market, some investors participate in some markets, and other investors participate in other markets. Figure 17 illustrates. This assumption means that the basic first order condition

$$p_t^j = E_t(m_{t+1}^i x_{t+1}^j)$$

holds only for investors i who are matched to security j.

Risks are then shared only between investors in a specific group, not across groups. This feature limits risk-bearing activity and therefore leads to the emergence of premia that are not related to aggregate risks. Thus one basic consequence of "segmented markets" is that *risks are not shared* across investors as they are in the standard model.

A shock to an asset class will then be a shock to the wealth and marginal utility of all the investors in that class, generating a premium, even if the shock is insignificant relative to aggregate wealth. As a result, segmented markets generate risk premia for local pricing factors. For example, if stocks are segmented from bonds, then we can see stocks that are priced with stock market factors, and bonds that are priced with bond market factors. Similarly, if country markets are segmented, then each country's stocks may be priced with a country index, but overall premiums will not correspond to the overall world market index. Segmented markets can be invoked to explain the wide variety of premiums associated with their own factors. If groups of investors specialize in holding small stocks, value stocks, writing put options, etc.,then each of those securities will see premiums in proportion to "their" factors, but those factor premiums will not make sense relative to aggregate premiums. For example, Gabaix, Krishnamurthy and

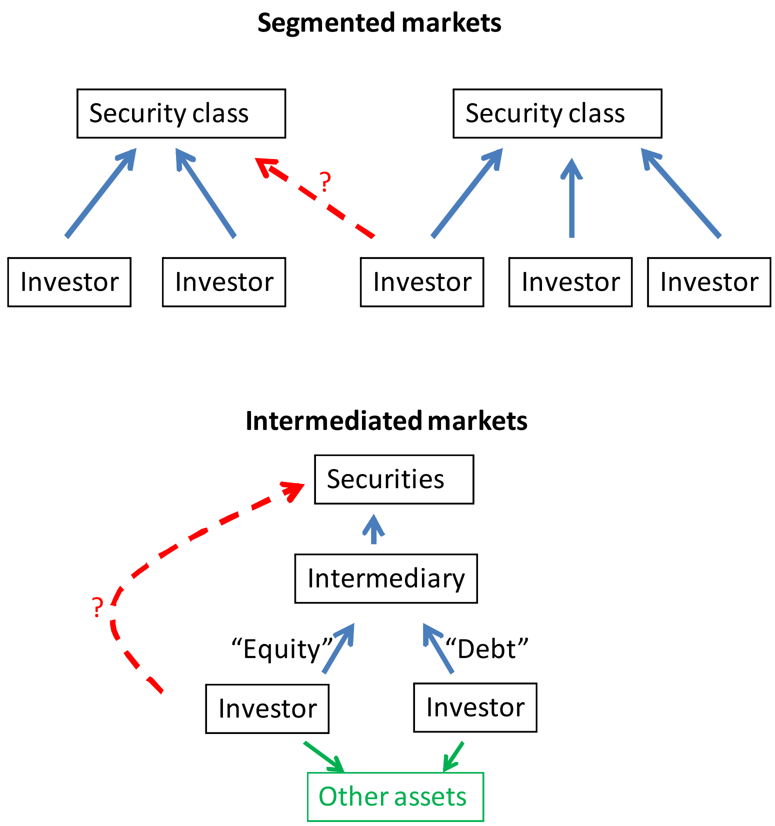

Figure 17: Segmented markets vs. intermediated markets.

Vigneron (2007) find that mortgage-backed securities are priced by a pricing kernel that reflects MBS risk, but poorly priced with kernels based on wider aggregates, suggesting this market is integrated within its class, but segmented from other securities.

Segmentation can help to make sense of the emergence of so many factors in the first place. If security price movements were all due to cash flows, then there is no particular reason why prices within a segmented group should move together (unless of course the group is set up to have securities with similar cashflows, which is often realistic.) However, given that price movements seem to reflect discount rate shocks, they naturally propagate across the securities held by a small group of investors. If a small group loses money in one investment, and this raises their risk aversion, the prices of other securities held by this small group will also decline, and we will see a "common factor" emerge.

Segmented markets, and consequent limited risk-bearing ability, also gives rise to "downward sloping demand curves." In such markets, discount rates will rise (prices will fall) if their investors are suddenly asked to bear more risk. This story has been told, for example, to understand the effects of Federal Reserve asset purchases in commercial paper and mortgage backed security markets, and segmentation by maturity underlies its belief that buying long-term treasury bonds will lower their yields.

Why not?

The natural theoretical puzzle is why markets should be segmented, as illustrated by the red dashed line in Figure 17. What keeps investors from accessing different markets? Truly segmented markets require some transactions or information costs, as there are profits and utility to be made by bringing investors to markets.

To some extent, the answer to this question is time. Investors do specialize in particular classes of securities, and there are fixed costs of learning how to analyze and trade classes, especially the more obscure classes. For example Mitchell, Pedersen, and Pulvino (2007) describe "slow-moving capital:" A small group of hedge funds specialized in convertible debt, which takes some setup costs. Large losses in convertible debt left those investors unable to bear risks so the prices fell and discount rates rose. It took several months for multi-strategy funds who had not suffered the convertible debt losses to gear up and move in to the market. Which they eventually did. As an extreme example, everyone in this room holds some version of a market index, and would happily invest more if we saw a clearly temporary halving of price. Yet in the "flash crash," exactly this event occurred, and it took a few minutes for the "slow moving" capital to respond.

The other answer to this question is financial innovation, which removes the barriers to segmentation. For example, a reasonable story to tell of the small firm anomaly is that it represented a somewhat segmented market before the late 1970s. To hold a diversified portfolio of small stocks required time, effort, trading skill and accounting skill. The invention of the passively-managed small stock fund meant this segment could be cheaply accessed by all investors, and it is perhaps no coincidence that the small stock premium disappeared about the same time. (However, the small factor in covariance has not disappeared). Similarly, catastrophe reinsurance was once a business with few players and high rates of return (Froot and O'Connell (2008)). Catastrophe options and catastrophe linked bonds help to spread the risk and bring down the premiums.

More generally, as Fama and French (2010) emphasize, many anomalies are limited to very small cap stocks, what I like to call "anomalies in dusty corners of the market." The natural implication (mine, they don't say this) is that very small stocks are hard to trade, and small size means it's not worth creating institutions to lower trading costs and more widely share their risks.

Of course I have told too many stories here. Segmented markets theory has to be evaluated quantitatively, and the costs and barriers documented, or at least the presence or absence of investors documented. Like all theories, this one cannot become a basis for ex-post storytelling.

5.2 "Institutional finance" "intermediated markets"

I think "intermediated markets" or "institutional finance" applies to a different, vertical, separation of investor from payoff. Investors use delegated managers to handle their assets. Then, frictions or principal/agent problems in the delegated management relationship spill over into market prices for the assets. (Among many others in this exploding literature, Garleanu and Pedersen (2010), Brunnermeier (2009), Brunnermeier and Pedersen (2009), He and Krishnamurthy (2010).)

For example, the investors may split their investments to the managers into "equity" and "debt" claims, and managers are therefore leveraged. When losses appear, the managers stave off bankruptcy by trying to sell risky assets. In essence, their debt levels act like the habit level to raise risk aversion. But since all the managers are doing the same thing, nobody can sell in the end and prices are driven down, or, equivalently, discount rates rise. Colorful terms like "fire sale," "liquidity spirals" and so forth are used to describe this process.

Intermediation and segmentation are often combined, naturally, as some classes of asset require the specialized knowledge of intermediaries.

This class of models faces, if anything, an even more pressing "why?" challenge.

First, why would investors split their investments with managers into equity and debt categories, and then suffer when managers try to sell assets after losses? We need some asymmetric information or moral hazard reason for this inefficient contract between fundamental investors and the managers. Often, authors rely on simple institutional observation or allude to behavioral stories, but again I think these are shortcuts rather than essential parts of the theory.

Second, why don't investors offset these problems by jumping in directly when discount rates rise inefficiently? If the investors going through intermediaries don't wish to invest directly, where are the sovereign wealth funds, the defined benefit pension funds, the university or and charitable endowments, the Warren Buffets and other wealthy individuals?

There remain huge pools of unintermediated, unleveraged, directly managed investments who can and do move around across most security classes looking for deals. In the end it is always these who we depend on to move in and take advantage of excessive excess returns, or to lower their holdings of "overvalued" securities. And when they don't – as when most university endowments joined the panic and tried to sell in the crash of 2008 – it's hard to blame intermediation on the result. (And government policy has an unintended consequence here. Your fire sale is my buying opportunity. The net effect of the pervasive policy to bail out troubled institutions and support prices has been to lower the rewards for large Warren-Buffet style investors to keep a big pool of cash waiting on the sidelines to come in and offer the same service.)

5.2.1 Information, arbitrage, and common fallacies

In the context of both segmented and intermediated markets, a widespread fallacy holds that pricing is determined by the "marginal investor," the last one to buy. This is false. Any investor who has the potential to buy and sell is "marginal" at any moment. "Marginal investor" means one on his first order condition, not the last one to buy. In classic asset pricing everyone is "marginal" all the time – and also never trades. Any paper using the article "the" marginal investor (singular) is likely to be making this error. Prices are set by the *average* investor (weighted by risk aversion), not the last one to buy, because if they're not happy with prices, everybody can buy or sell at any time.

Segmented markets also raise the possibility that prices are informationally inefficient. The traders who are not active in a market are not able to express their information in that market.

For example, short sales constraints have long been interpreted as possibly leading to "overvaluation," as then only the optimists' views are expressed. Miller (1977) describes this situation, and Lamont and Thaler (2003) cite this theory as an important interpretation of apparent "overpricing" related to short sales constraints. Lamont (2004) also describes one of the largest alphas around, from selling stocks whose managers are undertaking actions to stop short sellers. Of course, the point of a short constraint is that you can't trade on it, but why don't the current owners try to sell?

From a theoretical point of view *limited risk bearing* in a segmented market is much easier to support than *informational inefficiency.* After all, all it should take is one informed trader to sneak in and release his information to the market to solve the latter problem. Actually, all it takes is for one informed but excluded trader to have one cocktail party conversation with one included uninformed trader. The baseline Milgrom and Stokey (1982) model of information inclusion requires no trading and no risk-bearing by informed traders. Informed traders should simply bid prices up with no volume, because uninformed traders should refuse to trade with them. This doesn't take any smarts, uninformed traders should simply be passively indexing and refusing to unbalance their portfolios for anyone. (If this isn't clear now, I return to the issue in section 5.3.)

The process of arriving at equilibrium is fundamentally different for informational efficiency than for discount rates. The former can (and should, but doesn't) happen with no trade, and can happen with no one bearing additional risk. The latter requires large, often institutional, changes to allow *all* investors to properly share risks.

If some "systematic" factor (momentum, carry trade, put option writing) has an unwarranted risk premium, the only cure is for that risk to be more widely shared. The *average* investor must change his demands. This is much harder, so markets can maintain "segmented" risk premia for a long time, even while trading within each market quickly removes any informational inefficiencies.

This insight can help to explain why many discount-rate anomalies are not quickly "arbitraged away," even if they do not correspond to macroeconomic risks. It may also help to explain why tests for informational inefficiencies were such a quick and resounding success in the 1970s, but understanding discount rate variation across different markets and strategies is so elusive; why we find endless new premia, each associated with a common source of variance, but each bearing strained relations to overall, average-investor, macroeconomic risk.

In any case, in categorizing theories based on segmentation, it's important to distinguish *limited risk bearing with full information* from a potential *informational friction*, and for segmented-markets theorists at least thinking about how to tell the two stories apart.

5.3 Liquidity premia; trading value

A long tradition in asset pricing recognizes that some assets have higher or lower discount rates in compensation for greater or lesser liquidity. Defining liquidity – there may be several different kinds – modeling it and understanding it deeply are still open questions however. (See the longer review in Cochrane (2005b).)

Individual liquidity

The on-the-run / off-the-run spread in US government bonds, made famous by Long Term Capital management (Krishnamurthy (2002)), the similar "benchmark" effect in Japanese government bonds (Boudoukh and Whitelaw (1991)), and the spread between agency and identical Treasury debt (Longstaff (2004)) are familiar examples. The idea that venture capital offers a higher risk premium, i.e. a lower discount rate, to compensate for its illiquidity is common, though hard to document empirically.

Systemic liquidity

More recently, both theories and empirical work have come about that emphasize liquidity as a *systemic* factor, not just an individual-security characteristic. (Acharya and Pedersen (2005), Pastor and Stambaugh (2003)). Times when all assets become illiquid are high marginal utility events, so assets that pay off well in times of aggregate illiquidity should offer lower expected returns.

Defining liquidity;

Many hedge funds and active managers, including university endowments, describe their strategy as "providing liquidity," presumably in expectation of higher returns. Many of these investors also found themselves selling in a panic along with everyone else in the Fall of 2008, so perhaps our understanding of "providing liquidity" needs some deepening. It's easy to say one is in the business of taking the other side of noise traders, but one has to avoid taking the other side of informed traders in the process. Simply posting bid and ask prices is just a bet on low volatility (and "liquidity provision" needs to be so benchmarked).

Liquidity, trading and market microstructure have often been ignored in mainstream finance. There was, I think, a justifiable feeling that the big economic issues are the first 6 percent of premium, and if we need some liquidity excuse for the last 5 to 10 basis points, that's not terribly important. In this view, prices are set *as if* trading were unimportant. Trading and microstructure, though interesting, don't add that much to the big picture.

"Convenience yield"

I suspect this view is wrong, and variation in discount rates due to "liquidity," whether individual or systemic, will be seen to be much more important than we now think. Securities *usefulness in information trading* will be seen to constitute an important, rather than trivial amount of security price variation.

To many observers, of course, the financial crisis of 2008 suggests a huge liquidity event, bringing more attention to this class of discount rate variation.

In thinking about this, I have been inspired by one of the most obvious "liquidity" and "transactions" related premium of all: money. Money and bonds are claims to identical payoffs – a dollar bill and a T bill are each claims to a dollar in 3 months time. Yet they carry different prices or discount rates, an apparent violation of arbitrage. Belying the usual view that liquidity is only about the last 10 basis points, the liquidity spread for money can be large. In hyperinflations, nominal interest rates, which measure the discount rate distortion due to the liquidity value of money, can be hundreds of percent per year.

We know the reason of course: money is useful in making transactions. The precise nature of this liquidity benefit still eludes clean modeling in macroeconomics. Many models simply put money in the utility function, a "behavioral' shortcut similar to many in liquidity theories of asset pricing. Others study very stylized trading setups, such as cash-in-advance models.. Money demand makes several predictions:

1. The high-priced security must be "special," providing some service that other ways of obtaining the same or similar cashflows do not.

2. Supply must be limited: there must be a constraint limiting people from supplying more of the high-priced security or substitutes for its "special" purposes.

3. The liquidity premium or price distortion is larger when the quantity of the special security is lower

Money is "special" in its use in transactions. You can't buy coffee with a treasury bill. The money supply is limited by the government, as is the supply of substitutes. Higher interest rates correspond to lower amounts of money (either supplied or demanded).

In Cochrane (2003) I asked whether analogous forces are at work in liquidity spreads, and in particular whether equity shares can display similar effects.

Now, securities are not desirable for facilitating transactions. But on-the-run or benchmark government bonds are "special" in *trading*. They're easy to buy or sell, and easy to repo. Most of all, information trading (or betting!) on the direction of interest rates takes place in the on-the-run market.

I asked whether the 3Com/Palm spread analyzed by Lamont and Thaler (2003) could have a similar interpretation. (3Com had done a 5% carve-out of Palm, and was going to distribute the rest of Palm to its shareholders in 6 months. Buying 3Com and waiting 6 months was as much as a third cheaper than buying Palm shares directly.) Table 7, copied from that paper, highlights the similarities between 3Com/Palm "mispricing" and Dollars/T bills "mispricing".

	3Com/Palm	Dollar/Tbill
Law of one price violated	x	x
Restrictions on long-term short	x	x (no banknotes)
High turnover, short horizon in "expensive" end	x	x
Turnover higher as price spread higher	x	x (velocity)
High price security is "special" for trading	x	x
Price spread higher as quantity lower	x	x
Price spread lower as substitutes arise	x	x
Much shorting. substitutes despite cost	x	x (checking accounts)
Size can be large	x	x (hyperinflations)

Table 7 Similarities between Palm/3Com and Dollar/Tbill

Trading volume in Palm was astronomical. As much as 20% of the outstanding shares changed hands each day. To a day-trader, who thinks he has some information about the 15% daily Palm volatility, the 1% per week loss due to "overpricing" of Palm shares relative to 3Com shares, is a small cost of doing business on the order of the transactions costs. Short-selling creates additional "shares," expanding supply, but Lamont and Thaler documented extreme constraints on short selling. Despite these constraints, short-selling eventually doubled the outstanding supply, and the overpricing eased as this happened.

Again, one may object that money is about transactions and Palm shares are not. But Palm shares were "special" in another way. The market for betting on information requires an adequate supply of shares. In a frictionless theory, all information trading could be handled by extremely fast trading of one share of Palm, with the remainder stuck in 3Com's corporate treasury, just as in frictionless theory all transactions in the US economy can be accomplished by electronic claims to a single dollar bill framed at the Treasury department. In our world, that is not true. Even a short seller has to obtain physical shares for some time. Substitutes such as 3 com or options did not suffice. 3Com and synthetic options prices were de-linked at high frequency.

One of the most puzzling facts of the Palm/3Com event is that 3Com shares *fell* as Palm prices exploded during the IPO day. Since 3Com retained 95% of Palm shares this really is puzzling. However, it makes sense as a liquidity premium. If all the traders betting on the future of the Palm device move from trading 3Com stock to Palm stock on the day of the carve-out, then Palm stock gets the liquidity premium originally embedded in 3Com stock. This pattern is exactly what happens to Japanese government bonds when the benchmark changes (Boudoukh and Whitelaw (1991))

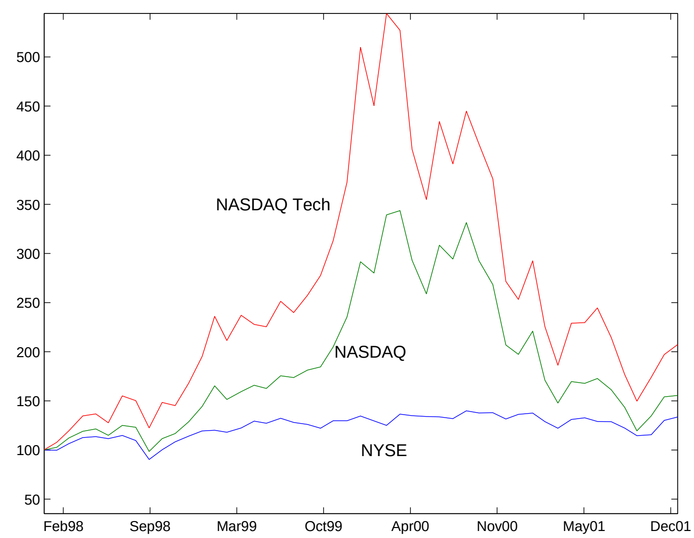

Figure 18: Price indices in the tech boom and bust.

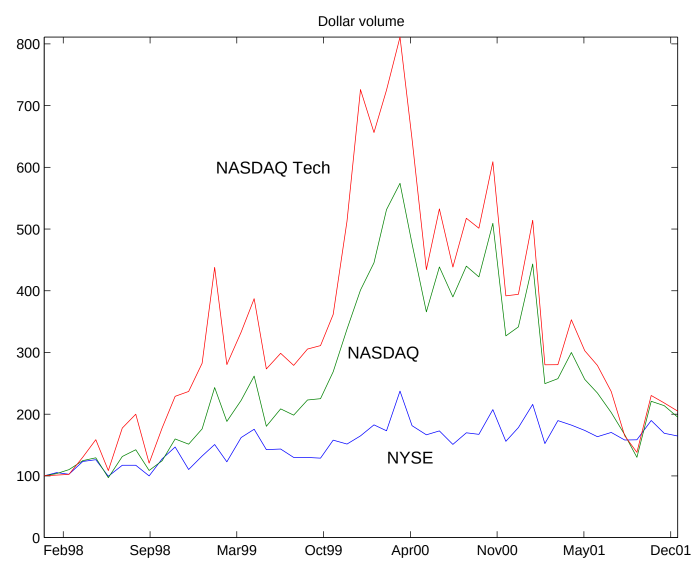

Figure 19: Dollar volume in the tech boom and bust.

How widespread are large liquidity premia in equity markets? Figures 18 and 19 are suggestive. The *price* raise and fall in the tech boom was concentrated in Nasdaq and Nasdaq Tech. The *volume* rise and fall was concentrated in the same place. Are these correlations possibly a mere coincidence? Or is the price rise at least in part due to a low discount rate, reflecting a huge liquidity premium? If you want to bet on the future of internet companies, either way, you need a certain supply of shares. And to high-frequency information traders, the small *return* premium corresponding to a large long-lasting *price* premium is a small cost of doing business – exactly as, to a purchaser of coffee, the small interest cost of holding a \$20 bill for a week is a small cost of doing business.

Back to bubbles. Table 8 tries to put together the central facts about "bubbles" and theories that account for them. The theories are, frictionless asset pricing driven by earnings forecasts; frictionless asset pricing driven by macro discount rate variation, "rational bubble" violations of the transversality condition, "irrational" expectations, "fads," "sentiment," prices set by optimists due to short sale constraints, and finally the liquidity/convenience yield idea I have been pursuing here.

Every asset price "bubble" coincides with a trading frenzy. Houses turned over quickly in 2006. (Leamer (2010), p. 41 Figure 2 shows a beautiful correlation between price and volume for Los Angeles houses.) Tulip bulb future turned over quickly in 1620 (Garber (2000), entertainingly reviewed in Cochrane (2001)). This is a coincidence in every other theory, both rational and irrational. I think it's central – and if we accept that, then we accept that liquidity premia can explain a large part of price fluctuation. (This does not mean the premia have to be that large. Small long-lasting changes in expected return, especially among growth stocks with high valuations already, can imply large changes in prices.)

	Theory					
	Frictionless		Rational		Optimists/	Conv.
Fact	Earnings	Discount	Bubble	Fads	Pessimists	Yield
Prices rise, decline	x	x	x	x	x	x
Prices do not forecast earnings		x	x	x	x	x
Prices forecast long-run returns		x		x		x
Long term short difficult			x	x	x	x
Large dispersion of opinion					x	x
Price high with low shares					x	x
Price high with high volume						x
Price high with volatility high						x
Biggest in growth stocks						x

Table 8. Summary of bubble theories and facts for which they account.

Liquidity implications

The presence of a liquidity or convenience yield premium makes sense of "downward sloping demands" or a "demand for shares." "Demand" is really a poor term in this context, as the pheonomenon is just as easily described as "supply." It's really the equilibrium price as a function of external buying or selling. However, liquidity demand disappear once satiated. Then, the price is the frictionless discounted present value of dividends. Figure 20 illustrates. Thus, liquidity cannot make sense of a globally "downward sloping demand."

This is a useful warning. For example, our Federal Reserve found that buying commercial paper and mortgage-backed securities had an effect on prices during the financial crisis. Alas, it is a mistake to conclude from that event that it can continue to lower rates once the big liquidity crisis is over.

The puzzle of trading

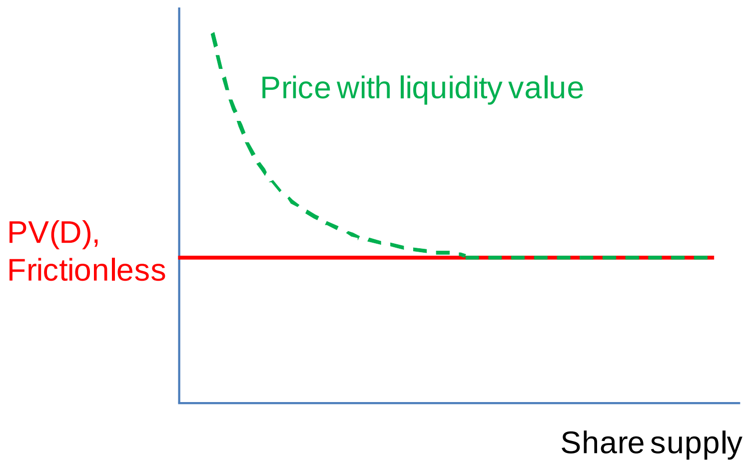

Figure 20: Price as a function of share supply when there is a convenience yield due to trading.

At a deeper level, why do people need to trade so much away? The average investor holds the market portfolio, passively, so prices should adjust so that portfolio is optimal. Life-cycle accumulation and decumulation, and any sense of optimal hedging of outside risks requires orders of magnitude less trading than we see. Optimal portfolio theory in the presence of transactions costs (Constantinides (1979)) just recommends that one trade even less frequently, but then demands little premium for illiquidity.

Verbally, we know the answer: Most trading is trading on information. The process of information discovery and its revelation in prices seems to take a lot of trading. The markets we study *exist* to support information-based trading. If assets existed only to raise capital, share risks, and support life-cycle investment, they could easily be sold at a retail level, as insurance or bank accounts are.

Yet we really do not yet have good models of information-based trading. The passive traders should index. An informed trader, trying to buy an underpriced security, should run in to a wall of indexers who simply refuse to take the other side of any such trade, and the price should rise with no volume. Anyone contemplating trading should realize it's a zero sum game, and half those who participate must lose. This is the famous "no-trade" theorem of asset pricing. (Milgrom and Stokey (1982).) Yet we know prices change in association with enormous volume (Brandt and Kavajecz (2004), Evans and Lyons (2002)). Brandt and Kavajecz show cleverly that this pattern follows the "information revelation" view: Informed traders make money while their information disperses, thus prices rise to where they would go anyway along with a large flow of "buy" orders. The alternative is the "price pressure" view that exogenous order imbalances run down sloping "demand" (or supply) curves. Yet why anyone trades with the informed traders at all is a bit of a mystery. (Or, why there are any "noise traders" which allow such "information traders" to hide.)

Liquidity models usually hard-wire the huge volume of trading. The typical tricks are to assume agents live only a week or two, so must turn over their portfolios in overlapping generations style (Acharya and Pedersen (2005)), or to assume some sort of "overconfidence" in one's signal (Scheinkman and Xiong (2003).) . Both assumptions are useful shortcuts allowing the authors to focus on the important parts of their models which lie elsewhere, but not particularly deep theories of trading.

I'm not sure how we will understand information-based trading and its price effects. We don't have good economic models yet. It seems a stretch to me to claim that there is no such model – that the mere existence of the NYSE proves once and for all that people are "irrational," whatever that means. But once again, the facts are pushing us to think harder about it. And the result will nicely tie up the loose end of "market efficiency": we should understand how information becomes reflected in prices.

Government debt, Macroeconomics, and other definitions of "liquidity"

The financial crisis in 2008 and subsequent recession saw movements in the dollar and spreads between government and private bonds that were also suggestive of a "flight to quality" or "flight to liquidity." Such a flight helps to make sense of many macroeconomic events, the huge fall in aggregated demand, and policy interventions of the period (Cochrane 2010, "understanding policy") In the 1930s people ran from bonds to money, meaning cash. In 2008-2010 people ran from private and foreign debt to US government debt.

This event is quite likely understandable as a change in discount rate (there's nothing else for US government debt!) Though US government debt is singularly useful in information-trading about the future course of interest rates, that story does not seem like a compelling source of it's "specialness" in this episode. Something about absolute security from nominal default, transparency on balance sheets, or repoability, seems a more likely source of the premium.

All of which goes to say that there are lots of kinds of "liquidity" yet to be distinguished.

6 Frictions in the financial crisis

Naturally, the 2008 financial crisis and subsequent great recession have dramatically increased interest in asset pricing with frictions, along with a lot of irresponsible "the financial crisis proves markets are inefficient" journalism.

This is an interesting event, and a context in which to consider how branches of the theory I have outlined will prove useful.

A conventional baseline and deeper puzzle.

I suspect my plots of surplus consumption and investment/capital ratios Figures 15 and 16 through the crisis already come as a surprise to many, who may have concluded that frictionless asset pricing has nothing to say about the crisis.

Figure 21 presents government and corporate rates through the great recession. We see a familiar pattern – government rates declined, while private rates rose. The credit spread widened. Historically, and certainly ex-post in this case, the credit spread corresponded to a higher expected return rather than a massive increase in default probability. Figure 22 plots the BAA spread along with stock prices. (A valuation ratio like the dividend yield is conceptually better here, but since it uses trailing dividends reinvested at the market return the exact timing of the dividend yield is a bit suspect.) We see that the stock price decline and credit spread widened contemporaneously.

These price movements are so uniform across markets and consistent in time, that they look very much like a large uniform rise in risk premium, a rise in γ_t in $E_t(R^e_{t+1}) = \gamma_t cov_t(m_t R^e_{t+1})$, and especially a rise in the premium for default risk. Even in the commercial paper market, high

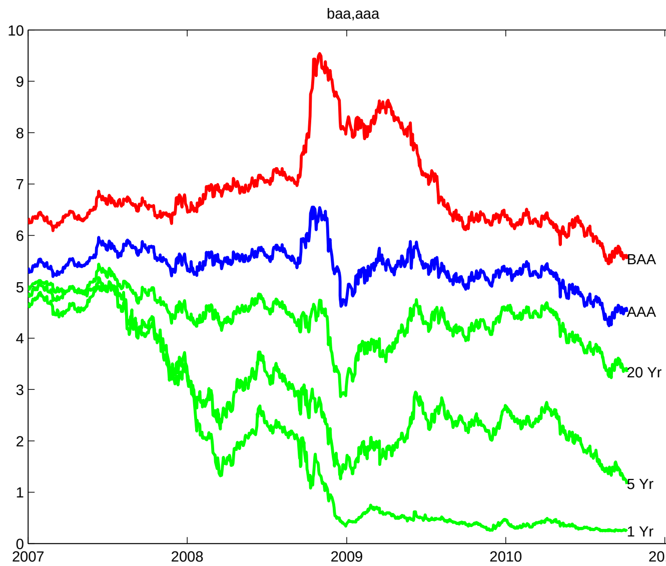

Figure 21: Interest rates through the financial crisis and recession. 20Yr, 5Yr, and 1Yr are constant maturity government bond rates. Source: Fred data base.

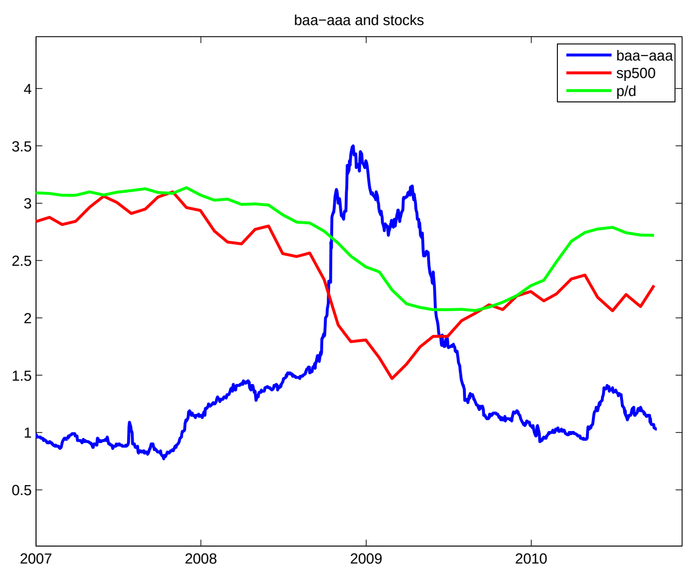

Figure 22: Credit spread, dividend yield, and S&P500 stock price.

grade short term nonfinancial commercial paper rates were unaffected by the crisis. It's very hard to argue that intermediaries set the prices in these markets. We all are "marginal" for stock and bond indices. Every single one of us can increase or decrease our exposure to the market or corporate and government bonds any time we want to, through widely available index funds. Most of us have those funds in our 401(k) plans. Attentive, unlevered, "long-term" investors including endowments, sovereign wealth funds, wealthy individuals are all aware of these price movements and to quickly buy and sell. They were largely "on their margins." The puzzle is that so many were trying to sell in winter 2008-2009. There is no pattern here that large indices fell more that were "more intermediated" or more subject to frictions. A behavioral story that affects stocks, bonds, and structured products so uniformly seems hard to construct.

On the other hand, we know that the fundamental financial shock was something like a run, concentrated in the shadow banking system, and verbal reports of a "flight to quality," "flight to liquidity," etc. are worth paying attention to. The question, I think, is how such a flight produces a pattern in prices that looks just like a rise in aggregate risk aversion.

What about the arbitrages?

Now, the financial crisis also gave rise to numerous fascinating asset pricing puzzles, and "arbitrages" in particular. In these cases, familiar arbitrage relationships were violated. In each case, unwinding the arbitrage opportunity required one to borrow dollars. But the essence of the crisis is that borrowing dollars had become suddenly much harder. For example, repo haircuts increased dramatically (Gorton and Metrick (2009)). Typical "arbitrageurs" active in these markets had lost money, faced redemptions, or were otherwise unable to perform their usual functions. The marginal utility of a dollar, delivered to a hedge fund, was particularly high. Here are two prominent examples:

1. CDS-Bond spread. Buying a treasury bond plus writing a CDS is the same thing as buying a corporate bond. They should have the same price. In the financial crisis, they didn't. Corporate bonds were cheaper. To arbitrage the difference, you need to borrow money against the corporate bond. Figure 23 presents a graph from Fontana (2010). You can see that the blue bond spread is not the same as the CDS, and typically lower. However, Euro sovereign CDS bond spread stayed high after the "financial crisis" reflected in short-term debt market dislocations had long passed. Greece CDS and bonds in 2010 still look like Figure 23. (Fontana 2010)

2. Covered interest parity. There are two ways to save riskfree dollars for three months: invest in US securities, or buy Euros, invest in Euro bonds, and buy a forward contract to return to dollars. These should give the same rate of return, and usually do. During the financial crisis, the Euro end paid more. To arbitrage this, you have to borrow dollars. Figure 24 presents a graph from Baba and Parker (2008). Deviations from covered interest parity (black dashes, right scale) grew from less than 0.1 percent (10 basis points) to 0.3 percent (30 bp) during the financial crisis. As illustrated by the separate levels, (solid black and red lines) the dollar rate was cheaper. Borrowing at US libor to invest abroad would make an "arbitrage" profit.

Now the emergence of an arbitrage opportunity is always a dramatic event. But these pictures do put the events in a good perspective.

- Note in each case that the *difference* between the two ways of getting the same cashflow is dwarfed by the *overall* change in prices. Bond yields increased dramatically, as did the CDS prices. Interest rates on both dollar and Euro dropped in half.

- In each case, the "arbitrage," though attractive to traditional "arbitrageurs" who might be able to take a highly leveraged position, is not large enough to attract "long only" "deep pocket" money. If you're buying bonds for a pension fund, you buy bonds anyway, which

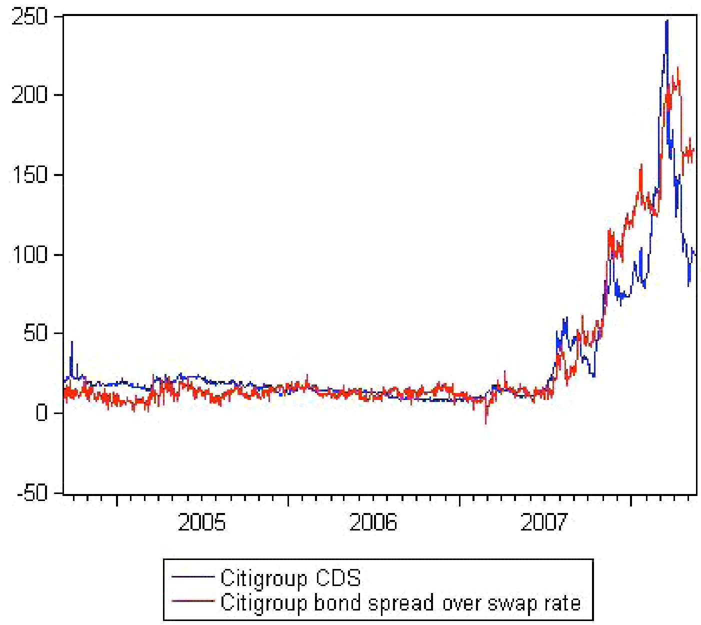

Figure 23: Citigroup CDS and bond spreads. Source: Fontana (2010)

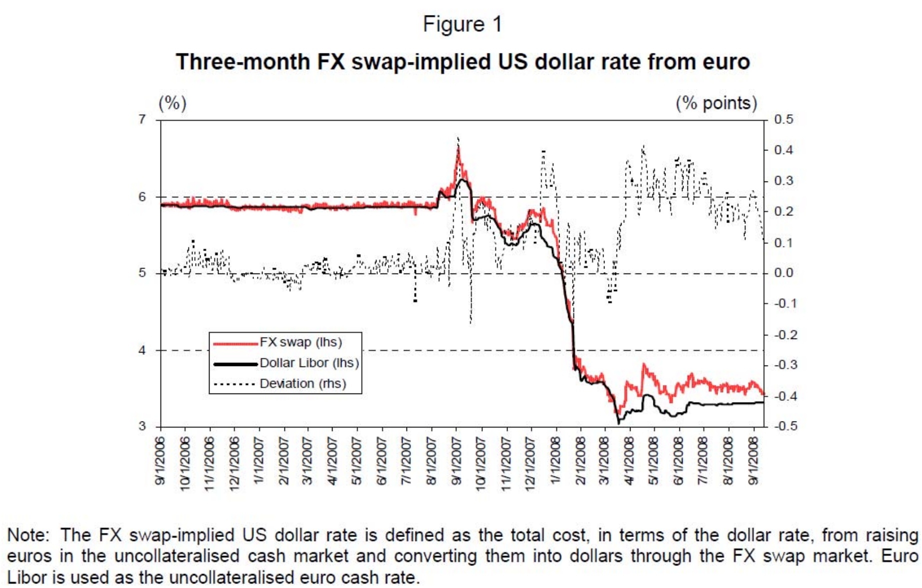

Figure 24: Covered interest parity violations in the fiancial crisis. Source: Baba and Parker (2008).

is the "cheap" end. If you have a cash management account that invests at something like US libor, 30 basis points may not be enough in the depths of a financial crisis to attract your attention to FX swaps or foreign investing and forward contracts.

So the question for us, really is which is the dog, and which is the tail? Is the story of *overall* price change largely told by the "macro" view, based on investors connected to markets through first order conditions, with interesting spreads opening up between traditional and slightly innovative ways of doing the same thing? Or does the appearance of these arbitrages somehow feed back and inform us about the level of price movements? The size and uniformity of the price movements suggests the former to me, but obviously this is the important question.

Part IV
Implications

Finance is not just about theory and data, it's about practical use. The *facts* of varying risk premiums change the practical uses and procedures in finance dramatically, and these changes have only started to be digested.

7 Portfolio theory

The facts of strong time-variation in risk premiums and strong expected return variation unrelated to market betas obviously attracts the attention of portfolio theorists! The former in particular has attracted a huge literature on how traditional investors who maximize lifetime expected utility should exploit the market-timing and intertemporal-hedging opportunities implicit in time-varying expected returns. It turns out that while Merton's (1971a) *characterization* of optimal portfolios in terms of value-function derivatives is easy, actually *calculating* those derivatives to quantitatively implement that advice is not so easy, especially if one wants to get past large numerical exercises and obtain intuitive answers (A small list: Brennan, Schwartz and Lagnado (1997), Campbell and Viceira (1989), (2002).) "Cross-sectional" facts such as value offer greater Sharpe ratios but have so far been the focus of less formal attention.

The hole in this approach is that *the average investor must hold the market portfolio.* If any mass of investors follows new portfolio theory, then the return facts on which the advice is built will vanish. We can't all market-time, and we can't all buy value. The average investor can't even rebalance, despite common advice that we all should do so. If discount-rate variation is "irrational," well we can't all be smarter than average either. (Behavioral finance, and behavioral economics in general, is prone to the conceit that only the other guy is behavioral; we smart investors, smart academics, or benevolent bureaucrats are, of course above it all and not swayed by emotions and biases.)

A useful and durable portfolio theory must be consistent with the fact that the average investor holds the market portfolio.

One might say then there is no portfolio theory. Just hold the market portfolio. The job of academics is then to figure out why – to figure out what interaction of preferences, risks, and price dynamics makes it optimal for investors do to nothing.

I think this is wrong, and that portfolio theory is more important than ever. At a minimum, figuring out *why* the average investor should do nothing is very useful advice to persuading him to do so. "Sure, these anomalies look great but there must be something in your preferences or risks that means you should ignore it" is just not very convincing to anyone. Much more importantly, we can build a portfolio theory consistent with the average investor theorem by building it on *differences* between people. People of different risk aversion and different risk exposure *should* hold different portfolios. And the finding that there are so many dimensions of risk makes this project much more interesting than it used to be.

A puzzle

In this line of thought, I also want to address an empirical puzzle. In the 1970 world view, there is really nothing much for industry to do, and correspondingly little reason why we should be able to charge high tuition or pay fancy salaries. The "tailored portfolio" claims of the 1960s are exploded by the two-fund theorem. There is no need for a professional manager to tailor individual risky investments to your risk tolerance, because splits between the market portfolio and riskfree investments offer all investors better rewards. The entire industry is reduced to chasing alpha, but generations of empirical work find essentially no alpha in active management. And alpha is a zero sum game by definition, so half the people chasing it must lose.

As economists, it's untenable to view a multibillion dollar industry as useless. And to those of us of a "rational" bent, it is a bit inconsistent to explain the continued success of non-indexed investing and active management as simple manifestations of human folly, forces we deny when explaining variation in asset prices. At least the behaviorists are consistent on this score!

The implications for portfolio theory and practice of widespread *discount rate* variation in asset prices offer a way to surmount this puzzle. As I will make clear, there is an economic function for individual-specific tailored portfolio formation, which earns a respectable fee in equilibrium. And there is an economic function for teachers of the required skills. All without breathing the word "alpha" – understood as exploiting informational asymmetries.

7.1 Multiple factors, multiple risks, multiple agents

Both theory and empirical work point to much greater importance of *differences between people* which justify differences between their portfolios, and make portfolio theory much more interesting and important, even while respecting the average-investor-must-hold-the-market theorem.

Even traditional mean-variance theory recognized such differences. It recognized one source of risk, the market portfolio. It suggested that people should hold more or less of this risk in inverse proportion to their risk aversion. People more risk averse than the market hold less market portfolio risk and vice versa. The condition that the average investor holds the market is useful in making this portfolio decision. Rather than understand the mean and variance of the market and try to implement $w = E(R^e)/\left[\gamma\sigma^2(R^e)\right]$ an investor could think about whether he or she is more or less risk averse than average.

Well, now we have many dimensions of risk, encompassing size, value, momentum etc., as well as state variables for time-varying expected returns. The optimal portfolio decision must split over many of these factors. This decision can also be represented graphically, in a natural multidimensional generalization of the standard mean-variance frontier diagram. The optimal portfolio (continuous time, multiple priced assets, state variables for non-traded assets and time-varying moments) is equivalent to a multifactor efficient frontier: minimize variance of return holding the mean constant *and* holding constant the covariance of return with additional factors. (See Fama (1996), Cochrane (2007b)). This is a quadratic objective with linear constraints, so the risky asset frontier is a hyperbolic cup, as shown in Figure 25. The mean-variance frontier including a risky asset is a cone, also shown.

Investors care about mean, variance and exposure to the additional factors. Their indifference curves are now surfaces as shown in the right hand side of Figure 25.

In order to span the multifactor-efficient frontier we need *three* portfolios. For example, the risk free rate, market index and HML could be three portfolios to span the frontier that minimizes variance while controlling covariance with value stocks. The average investor wishes to control exposure to the extra factor, and therefore accepts a portfolio that has less mean and more variance in return for less exposure to factor risk, as shown in the right hand panel of 25. This leads to the most celebrated but in my mind most boring implication of multifactor pricing – there are portfolios which beat the market in mean-variance space. Yes, but only measure zero agents only care about mean and variance, otherwise the factor would not be priced.

Much more interesting, investors now *differ* in multiple dimensions. Some are less risk averse and want portfolios further "out," which they achieve as usual by leveraging up a tangency portfolio. But now some investors are more or less worried about the risks posed by the extra state variable. For example, Fama and French's (1996) story for value is that the average investor is worried that value stocks tend to fall at the same time his or her human capital will fall in value as well. But some investors ("steelworkers") will be more worried about that; some investors ("fat cats") will be simply sitting on investment portfolios and not care, and others ("tech nerds") will have human capital correlated with growth stocks. The steel workers should be *short* value, buying more insurance despite the fee; while the fat cats and especially the tech nerds should be long and very long value, selling insurance to the steelworkers and earning the value premium for providing such insurance.

The rather boring two fund theorem has become a much more interesting three fund theorem. Now that we have 27 factors, in fact, the two fund theorem has become a 27 fund theorem! The most plausible and concrete source of individual differences is difference in outside, nontraded income; either from human capital (job) or from privately held businesses or narrow investments including real estate, as well as demographics (age, tax treatment, bequests, etc.) Each individual needs to evaluate which of the "factor portfolios" they should be holding or selling.

The *priced* factors, including "cross-sectional" factors such as value or put-option writing and "state variable" factors for intertemporal portfolio allocation, are only the second step of this process, really. In this portfolio context, hedging out individual risks that line up with *unpriced* factors is just as important and maybe more so. For example, the steelworker should be short a steel-industry portfolio, even though the steel industry portfolio may have zero alpha with respect to every model under the sun. This kind of insurance for outside risks is *free* (costs no alpha) and really comes before any "interesting" investment decision; it's like buying fire insurance on a house. It's probably the single most important investment decision the average investor can make, before even thinking about chasing alphas and premiums.

In this context, it's weird (or maybe a matter of habit) that we have spent 40 years searching for *priced* factors only, factors that would be interesting for the portfolio of the measure-zero mean-variance efficient investor, and completely ignoring the *unpriced* factors which are potentially far more beneficial to the majority of people. Understanding the *variation* across people in exposure to systematic but unpriced factors, and finding low-cost portfolios that most easily span that variation, is probably the most important thing we could do, and yet we really have never even asked the question.

All of this sounds rather hard. It is hard, and that's good! We finally have something to do, and a reason for the industry to exist, for whom we charge fat tuitions to our MBA students. Figuring all this out is a difficult problem, requiring analytical expertise; it is a retail business, and needs to charge a substantial fee. Based on the two-fund theorem, we make fun of the "tailored portfolio" advertised by high-fee financial advisers. Well it turns out there is an important place for tailored-portfolio theory, empirical work, and fee-based advice, and I haven't even breathed the word "alpha."

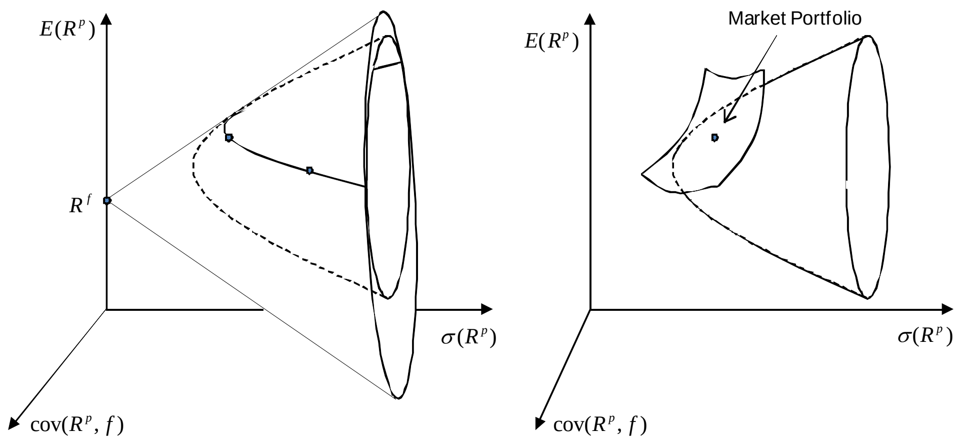

Figure 25: Multifactor efficient frontiers

At a minimum, this consideration is an important reality check for any portfolio advice. If you're asked to buy some factor risk, say value, you can ask the question, why am I different? The average investor must hold the market. For me to be long value, someone else must be short value. This advice can't be right for everyone.

(Of course many investors think they should deviate from market weights because they're smarter than average. Exactly half of them are deluded.)

Is understanding the economics of asset pricing important to portfolio theory? In one sense, no. Just as you don't care at the grocery store whether tomatoes are on sale because there is a speculative bubble in the tomato futures market, or because there is a bumper crop, you don't really care why prices are the way they are in forming a portfolio. However, asset prices are a bit more nebulous than tomato prices. I argued that "am I more or less risk averse than the average person" is a useful way to think about the stock bond allocation – at least a complement to "what is my risk aversion coefficient, and estimate of mean and variance of returns." Analogously, if we could understand, say, the value effect as a compensation for nontraded risk to human capital in some industries, that insight would help a great deal to answering "am I more or less exposed to value risk in my outside income than is the average person." If you know what insurance is really for, it helps to decide if you should be a buyer or a seller .

7.2 Mertonian State Variables.

The fact that prices move on discount rate news, and expected returns vary over time, means that state variables for intertemporal opportunities are essential parts of portfolio theory. I think it's common to ignore this fact, and I want to bring it to life.

As one measure of how much it's ignored, reflect on the fact that almost all hedge funds, active managers, and active institutions such as endowments and sovereign wealth funds still use some version of mean-variance portfolio theory to decide how to spread money over their many hot ideas. Reflect on how silly this is: you have a strategy that is *predicated* on the idea that expected returns and variances vary over time, in ways can model, so you should be altering your portfolio monthly to take advantage of these shifts. And yet you use a portfolio optimizer that assumes returns are i.i.d. and ignores state variable risk.

Bond investing gives a stark horror story. Suppose you are a highly risk averse investor, with a 10 year horizon, and are investing to cover a defined payment, say a child's tuition at the University of Chicago. The right investment is a 10-year zero-coupon TIP. (Campbell and Viceira (2001) Wachter (2003), Sangvinatsos and Wachter (2005).)

Now, suppose interest rates rise, interest rate volatility surges, and the value of your investment plunges. What should you do? How do you answer the advisers who call and say "you need to bail out of this horrible investment. You can't afford to take any more losses. Your risk-adjusted return is terrible, come to us, we have some alpha. You need to rebalance your portfolio. Volatility means the Sharpe ratio of your investment is terrible, you need to scale back. You need to increase your rate of return target in order to cover your spending plans" and so on. Obviously the answer is simple: you should tear up the envelope.

You ignore the rotten one-year mean-variance behavior of long-term bonds because long term bonds are very good hedges for their intertemporal opportunity set. Figure 26 dramatizes the fact. In a year that long-term bond prices decline, long-term bond yields and expected returns rise a great deal. Evaluating a bond investment in a one-period mean-variance, alpha-beta framework is silly. (Far too many real bond investors still evaluate portfolios by one-year returns benchmarked to a bond index, but that's another story.)

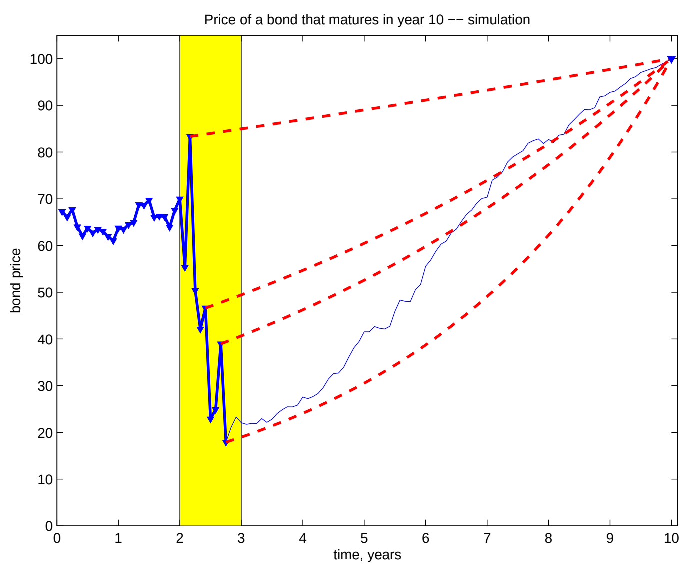

Figure 26: Bond price through time. A cautionary example.

This all seems pretty obvious, but consider the position of many long-term stock investors in December 2008. The market had plummeted, and stock volatility rose dramatically, from 16% usually to as much as 70 -80%, reflected both in realized volatility and in option-implied volatility. Figures 27 and 28 illustrate the situation they faced. What should they do?

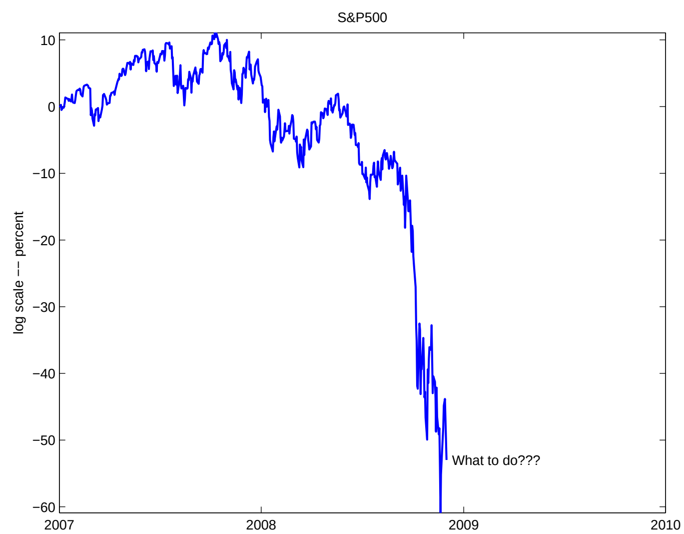

Figure 27: S&P500 index in the financial crisis.

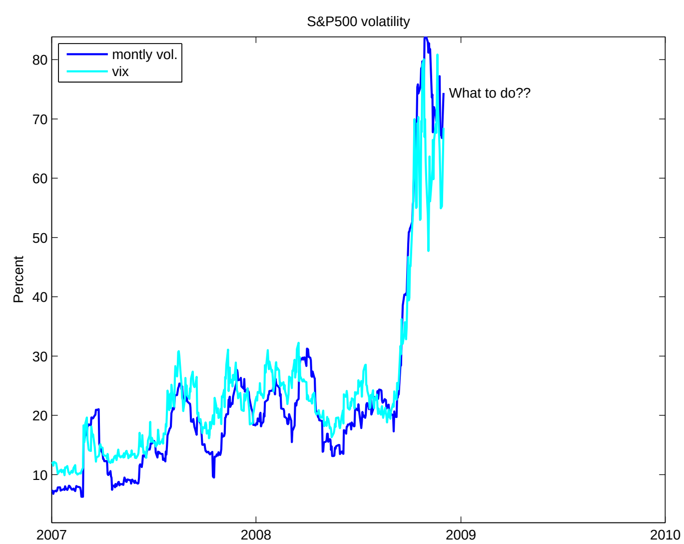

Figure 28: VIX and implied volatility in the financial crisis.

Well, the standard formula gives some standard advice

$$\text{stock share} = \frac{1}{\text{risk aversion}} \frac{\text{expected return - riskfree rate}}{\text{return variance}}$$

For example, you might have been happy with a conventional 60/40 allocation with risk aversion of 2 and usual numbers for equity premium and volatility.

$$\text{was: } 0.6 = \frac{1}{2.0} \times \frac{0.06 - 0.02}{0.18^2}$$

Now that volatility has gone up to 70%, however, the standard formula says *sell everything now!*

It recommends a 4% equity position!

$$\text{is: } \frac{1}{2.0} \times \frac{0.06 - 0.02}{\mathbf{0.70}^2 \Leftarrow \text{change}} = 0.04!$$

On your way to the phone, however, stop and think a minute: The market did not fall $1 - .04/.6 = 0.93\%$. Everyone else is not down to 4% equity. The average investor is holding the market portfolio. One period mean-variance analysis is not so unknown that it makes you dramatically smarter than average. Is there not some reason that perhaps the average investor has not panicked as you're about to do?

A glimmer of the answer should appear. If volatility can change from 0.18^2 to 0.70^2, why are you using a formula that assumes i.i.d. returns? More importantly, we learned above that dividend yields forecast stock returns. *Stocks are a bit like bonds; price drops that increase yield also increase expected returns.* This means that *like bonds, stocks are a good hedge against their own state-variable risk.*

Don't panic. Don't sell everything. You were using an outdated formula. And of course, in this case, "don't sell" was well-rewarded advice.

I realize I'm not breaking new ground here. Mertonian state variables have been recognized in portfolio theory for 40 years, and the fact that dividend-yield predictability makes stocks more attractive to long-horizon investors is the subject of a huge literature. But it isn't sinking in. I think the hedge funds, endowments, and portfolio managers who persist in mean-variance applications haven't digested that *Mertonian state variable effects are big, important, and need to be incorporated into portfolio theory.*

7.3 Prices and payoffs

This discussion naturally brings me back to a theme of this essay – that our multifactor, non-i.i.d. world should bring us back to thinking about *prices* and *streams of payoffs* rather than one-period returns.

The bond example is particularly stark. Yes, it's a mistake for a 10-year investor to think about a 10-year bond in terms of one-period mean and variance. But it's just as nutty to think about a 10 year bond as an investment whose one-year mean and variance is matched by an equally large one-year exposure to a state variable (yield) that forecasts its expected return. If I wrote down the one-year problem, it would probably take any of us an hour to realize that the ten year bond is the answer. You have to solve a value function and take its derivative. Good luck convincing a client or a student of the proposition. If you look at the ten-year problem – 10 year prices, and 10 year payoffs– it's just obvious that the ten-year TIP strip is the relevant riskfree asset, and that highly risk averse investors should just tear up the envelopes in the meantime.

The corresponding proposition for a lifetime investor is equally obvious once we look at *payoffs: An indexed perpetuity is the risk-free asset for an infinitely lived investors* (i.e. one with objective $E\sum_{t=0}^{\infty}\delta^t u(c_t)$). It's not at all obvious if you just look at one-period Mertonian problems. Exactly what asset has just the right duration? What is the exact model for the time-variation of returns?

Strikingly obvious to us, but not to industry and applications who still think of overnight rates as "risk free." Though easy to create by stripping TIPS, indexed perpetuities are really not even available. (Indexed annuities are available, but suffer a lot of problems.) Such assets should be the first option next to "total market index" on the Vanguard website. And people need to understand that price fluctuation of their "riskfree asset" are irrelevant.

Now that we're looking at *payoff streams* that you receive in return for an initial price, it's pretty obvious that the claim to consumption, or claim to aggregate dividends is the relevant "risky" asset. And there must be a question to which the answer is "every investor splits his portfolio between an indexed perpetuity and a claim to the aggregate dividend stream."

There is. It turns out (Cochrane (2008)) that the quadratic utility derivation of the mean-variance frontier extends in just this way. *Every investor wants to hold a portfolio that is a combination of an indexed perpetuity and a claim to the aggregate dividend stream.* Furthermore, *every investor holds a "long-run" mean-variance efficient portfolio*, defining "long-run" expectations that sum over time and states of nature $\widetilde{E}(x) = \frac{1}{1-\beta}\sum_t \beta^t \sum_{s^t} \pi(s^t) x_t$. Less risk averse investors hold more of the claim to aggregate dividends, and vice versa.

Moreover, *the two-fund characterization of optimal portfolios holds for arbitrary variation in investment opportunities. Mertonian state variables disappear* once you look at final payoffs, just as they did once our 10 year TIP investor stopped worrying about monthly returns and looked only at his final payoff. Once you look at prices and consumption streams, time-varying investment opportunities become just a way of obtaining a final consumption stream. That's natural at a deep level – time and cross section really are one; dynamic strategies can be repackaged as buy and hold investments. That's what a hedge fund does!

Outside-income state variables do not disappear. If an investor has a nontradeable job or business in this setup, he first finds a portfolio whose *payoffs* have maximum long-run correlation with the *payoffs* of his outside income, and shorts that portfolio. Then he invests in a mean-variance efficient portfolio as described. Notice "payoffs." The hard part of hedging outside income has always been to find valuations. Correlating labor income growth with asset returns is a mistake of units. More deeply it assumes that the labor income discount factor is constant, an assumption we know is false for asset returns. Using dividend growth to proxy for returns and vice versa would be a huge mistake for traded assets. This approach suggests that fundamentally we should not care about valuations. At least the characterization of the optimal hedge portfolio involves correlations of payoffs as it should.

Once again, looking a *prices* and *streams of payoffs* allows us to repackage two-period intuition to understand multiperiod problems. And doing so gets us to focus back on underlying rewards and risks measured at the cashflow level, rather than worrying about betas that are correlations between discount rate changes.

Furthermore, a price-payoff approach helps to overcome some of the difficutlies that parameter uncertainty poses to conventional Mertonian dynamic portfolio theory. (As Barberis (2000) and Pastor (2000) show, parameter uncertainty is real risk to the investor and affects portfolio choices.) Investors are not sure just how much expected return varies with a state variable, hence how exactly to form high-frequency hedging portfolios. The desired stream of payoffs is much simpler to calculate.

However, none of this is easy. And that's a feature, not a bug. Classic mean-variance portfolio theory was partly successful for stopping just as the hard part arrived. Sure, hold a mean-variance efficient portfolio, but how? In a sense, active management and hedge funds have been selling solutions to that problem for 60 years since the mean-variance *characterization* of the solution was announced.

A simple *characterization* of optimal final payoffs doesn't tell you the *financial engineering* of how best to construct them. Finding an optimal hedge portfolio for the cashflows of an outside risk is easy to state, and hard to do. It needs academic research in finding *non-priced* factors useful for hedging, and not just alphas for one-period investors. It needs academic research in characterizing the non-market but hedgeable risks that lots of investors face. (Heaton and Lucas (2000) is a beginning.) The financial engineering of *forming* long-run mean-variance efficient portfolios is no simpler than the engineering of forming one-period mean-variance efficient portfolios, or forming portfolios of options that tailer to nonsymmetric preferences. And as in all such engineering, which of many ways to accomplish the same thing – static investment in a market-timing hedge fund, vs. explicit market timing; buy put options or dynamically synthesize them for a loss-averse investor – depend on transactions cost and market structure details.

This is all good news again for the financial services industry and those of us who make our money preparing students for jobs in that industry. Here is a very important function, requiring individual specific tailoring, and analytical skill that we know how to provide, and commanding legitimate fees. And I haven't breathed a word of alpha.

7.4 Alphas, betas, performance evaluation

Discount rate variation and a world with multiple dimensions of risk fundamentally change our understanding of the active-management industry, and may help to explain the puzzle of its persistence.

The 1970 view was very simple. There is one source of "priced" or systematic risk, the market index. You can take more or less of it. Investors understand it and have chosen their desired level with passive investments, so they know how to offset any implicit beta risk in active management. Active management is devoted to chasing "alpha," which means exclusively to uncover inefficiently priced assets. Alas alpha-chasing is by definition a zero sum game, and therefore rather unsurprisingly unsuccessful at least in aggregate.

Our new view, including new studies of mutual fund and especially hedge fund performance, change all of that, in ways I don't think we have fully appreciated as yet.

Suppose the value effect is real, and corresponds to hedging outside labor income as Fama and French suggest. Still, to an investor who doesn't know about the value effect and is holding the market index, it is pure alpha.

Now consider our view of the facts. Not only is there "cross sectional" value, small, momentum, accruals, issuance etc. factors in stocks, there are the "market timing" factors (or their cross-sectional equivalents) such as the international carry trade, term premium, and credit spreads in bonds. Hedge fund performance analysis has awakened us to other possibilities: many of their strategies amount to writing index puts or other option strategies, which seem to offer high premiums that potentially offset the occasional disasters. (Of course they really look good if you ignore the disasters, but that's another story.) Whatever "liquidity provision" means, certainly has the same character. The performance of the HFR "equity-market-neutral" hedge fund category plotted in figure 29 is a stark reminder. (Whatever "equity-market-neutral" means it sure is not "zero beta." But it does not have a static beta either, being seeming more exposed both for large increases and large decreases.)

Each of these strategies amount to taking "systemic" risk, in that bets in the categories line up with additional "factors." We don't yet understand those factor premia as covariances with

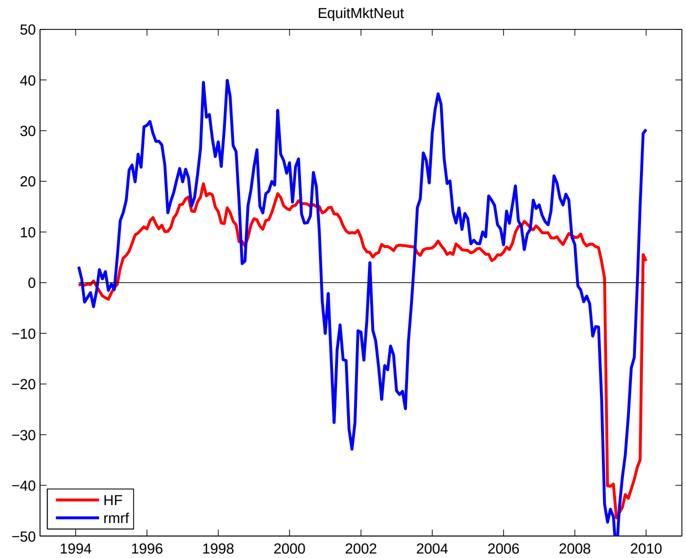

Figure 29: One year returns of the HFR equity market neutral edge fund index and the crsp value-weighted excess return

some underlying risk (marginal utility); and even if we did, their association with strong factors in the covariance of asset returns would leave us with a highly multidimensional view of risk and return anyway.

The vast majority of quantitatively-focused hedge funds and active managers – at least the ones who talk to me – basically are dynamically investing in these sorts of factors. Typical investors don't even know these sources of risk exist, and have not even begun to think about exposure to them. If, for example, we regard a rise in credit-default swap premium as a signal that a category of credit risk is "on sale" and needs greater risk sharing from the rest of us, how is this risk sharing going to happen? How are you and I going to increase our demands? The only realistic hope is that our investment is mediated by a high-tech institutional investor that can quickly spot, process, and take on systematic risks on our behalf. I have just described a hedge fund. As a Chicago Economist I am delighted to find some use for a rapidly growing institution that evidently meets some market test, rather than just continue to deplore active management.

More generally, multidimensional risks call in to question the whole alpha-beta distinction in a deep way.

Distinguishing "alpha" from "beta" by deep understanding of macroeconomic source of risk vs. deep understanding of an informational inefficiency never made sense in portfolio evaluation. Investors don't care where low prices come from, they just want to buy cheap. The point of performance evaluation is, sensibly, to distinguish strategies that require the skill, intuition and information gathering abilities of the manager, for which a fee must be paid, from easy passive indexing that the investor could do on his or her own.

But now we have 27, poorly understood sources of systemic risk. Suppose we analyze the performance of a hedge fund, and find its "alpha" can all be explained by exposures to value, momentum, and a carry-trade strategy that dynamically rebalances into high interest rate currencies. "That's not alpha" the investor might claim, "I can program my computer to do that." "Go ahead and try it," the manager might respond "writing that program took me a long time. It is my "alpha."

More generally, defining "beta" as "that which we can program a computer to do," i.e. execute trades as a function of public signals, and "alpha" as "that we cannot program a computer to do" – which still is just executing trades as a function of public signals – makes no sense at all. We don't want alpha and beta to come down to a philosophical debate about computability and the limits of artificial intelligence programming to mimic human decisions!

In a world with 27, poorly understood sources of "systemic" risk, each with time-varying rewards, the whole alpha (information) vs. beta (systemic) distinction just doesn't make sense any more. Looking at all this, it seems to me there is no such thing as "alpha," there is only "beta you understand and beta you don't understand." Active managers are good at selling "beta you don't understand."

Of course, I wish they were a bit more transparent about this fact, and I think there is much to be done to better match investors with the right dimensions of risk they should be buying or selling – everything other than market weights is still a zero sum game. And I am stretched to see 2+20 as an optimal contract in this view. But at least it makes more sense than just deploring active management for another 40 years.

7.5 Capital budgeting in Corporate finance

The first lesson of capital budgeting is to use the CAPM to discount cash flows for investment decisions, typically assuming a constant (6%) market expected return. The first slide in a corporate finance class looks something like this

$$\text{value of investment} = \frac{\text{expected payout}}{R^f + \beta\left[E(R^m) - R^f\right]}.$$

Which, we now must recognize, is completely wrong. The market premium isn't always 6%, but varies over time by as much as its mean. Expected returns don't line up with CAPM betas, but rather with multifactor betas. And since expected returns change over time, the discount rate is different for cashflows at different horizons.

It's interesting that investment decisions got so close to right anyway, with high investment following high stock prices. (Remember figure 16) Evidently, a generation of our MBAs figured out how to jigger the numbers and get the right answer despite a dramatically wrong model.

One might conclude that the right procedure is to use full multifactor models with time-varying market premiums and time-varying market betas in an explicit present value model. Not yet, I think. While we know these forces exist, the exact specification is unsure, and all the important numbers are rather poorly estimated. Even using the CAPM, $\sigma/\sqrt{T}$ means that the standard error of the market return was at least 2-3 percentage points. The standard error of the conditional mean market return and its persistence is substantially larger.

The demise of the i.i.d. CAPM world view also leads us away from using one model for every purpose. Capital budgeting is a "relative pricing" exercise – we want to use available information in asset markets to help us decide what the discount rate for a project should be. We really don't care on the deep source of those discount factors. For this purpose, simply looking at average returns of "similar" securities is enough.

I argued above that there is a two-step heart to current empirical procedures: Understand expected returns as a function of characteristics; understand covariance with factors as a function of characteristics, and then see if you can equate the two: $ER(C_i) = cov(R, f|C_i)\lambda$. I argued that this "finance" exercise was a subset of a deeper exercise, to understand the factor risk premiums by covariance with deeper measures of marginal utility (say) , $\lambda = cov(f, m)$. This application is an even smaller subset. Really all we want to know is $ER(C_i)$. Even with a successful factor model it's not obvious you get a better *measure* of $ER(C_i)$ from the left hand than you do from an estimate of the right hand side. So, even when we do understand "finance" factor models and their connection to "economic" factor models, we still may prefer to use estimates of expected returns as a function of characteristics in applied capital budgeting.

This idea still leaves out dynamics. Do discount a cashflow, we need expected returns at different horizons. Fundamentally, discounting cashflows is a "price and present value" sort of calculation, not an "expected return" calculation. That suggests that the fundamental answer will come once we develop price and present value asset pricing models. If we have P/D or B/M as a function of characteristics on the left hand side, and our models understand those as present values of cashflows, then we have directly the quantity we're looking for.

This is not so abstract a suggestion. Really, the Q theory gives a completely different procedure for capital budgeting. It says, make investment decisions by looking at Q, or B/M. When you don't know Q for a given project (just as when you don't know $E(R)$ in the conventional calculation), but you understand $Q(C)$, then you are staring at the valuation directly. When we understand $Q(C)$ as a covariance of long-run cashflows with factors, all the better, and perhaps the right hand side of that equality will also be fruitful in measurement. The emerging"affine" models for equity valuation also offer a promising possibility for explicit valuation.

7.6 Accounting, regulation, and corporate finance

All sorts of procedures in accounting, regulation, and corporate finance implicitly assume returns are independent over time, and hence prices are good forecasters of cashflows.

Suppose a firm has a single cashflow in 10 years, and is funded by a zero-coupon bond and equity, which owns the residual. In most accounting and corporate finance, we would use the stock and bond prices to calculate the probability and distance to default. Implicitly, however, we assume that prices move on cashflow news and returns are i.i.d. when we do so. When a price declines because of a change in discount rates, that fact has *no* implication for the probability or distance to default.

Many uses of mark-to-market accounting thus implicitly assume i.i.d. returns. For example, the heart of bank capital regulation is that we want to make sure banks have enough assets in the future to cover their future liabilities – if the stream of interest payments on their mortgages will be enough to pay the stream of interest payments due their depositors. Mark-to-market regulation forces them to come up with more capital or sell assets if the *price* of their assets (mortgages) falls, even if the *current payments* of those mortgages are fully able to meet current interest payments. That all makes sense if a lower price corresponds to lower asset cashflows – if those mortgages, though presently current, are more likely to default in the future. but what if the lower price simply reflects a higher credit premium? Now the *probability* of default is unaffected.

Banks often complain that low asset prices do not reflect "insolvency" they just represent "illiquidity" or "temporarily depressed markets." Well, we are learning there is a lot to "illiquidity" in discount rates, so there is some sense to this complaint. Rules that allow "hold to maturity" accounting treatment for long term illiquid assets may make some rough and ready sense after all.

One may object, rightly, that the *risk-neutral* probability of default has risen. Perhaps the regulator wants to control the risk-neutral probability of failure, not the actual probability. Doing so assumes (at least) that the regulator has the same discount rate as is reflected in the market. Perhaps so, and perhaps not, but at least we're having a dramatically different conversation.

I am not arguing here that mark-to-market accounting is bad, or that fudging the numbers is a good idea. The point is that what you *do* with a mark to market number can be quite different in a world driven by discount rate variation than one driven by cashflow variation.

The standard capital structure (debt vs. equity financing) trades the tax benefits of debt against the costs of default. A firm whose equity value declines is considered closer to default, and thus should recapitalize. But this prediction implicitly assumes that the decline in equity value does in fact correspond to likelier default. If equity declined because the discount rate rose, then the firm is no closer to default at all. Again, we can argue whether the risk-neutral or actual probability should govern the capital structure decision, but again that's a very different argument. This insight may help to explain why firms in fact respond very slowly if at all to changes in the market value of equity (Welch (2004)).

The view that the stock price is driven by earnings expectations lies behind stock-based executive compensation as well. The fact that a large fraction of stock market value reflects variation in discount rates over time, and new-factor beta exposure in the cross section might change that idea. It's already a bit of a puzzle that executives should hold the "systematic" risks due to market or commodity-price exposures. Now that a good part of such risk is discount rate risk or correlation of discount rate risk with that of pricing factors, the logic of such incentives changes. Perhaps the point has less to do with effort and operating performance, but with incentives for risk management.

7.7 Macroeconomics

If large variation in risk premia implies basic changes for macroeconomics, which has barely started to think about risk premia at all.

Macroeconomics is focused on a single intertemporal price, "the" interest rate, which crucially intermediates saving and investment. Finance already knows that the relevant interest rate for a typical borrower is many percentage points above the Federal Funds rate. Moreover, the *variation* across time and across borrowers in the relevant interest rates is almost entirely due to variation in the *risk premium* for corporate debt, a fact screaming from Figure 21.

For one timely example, compare this essay with Bob Hall's (2011) parallel speech at the American Economic Association. To Hall, the central problem in the current recession is that "the" interest rate is stuck at zero. With deflation (?) real rates are too high, so people are saving too much and firms are not investing enough. Hall's analysis centers on the idea that asset *rates of return* are "sticky." The picture of recessions that comes from finance centers on the *willingness to bear risk* and variations in *risk premiums,* having little to nothing to do with the level of short-term interest rates.

The difference between standard pictures of consumer and producer behavior is already pretty stark between macroeconomics and finance. Despite my beautiful Figure 16 linking investment to stock prices, the Q theory is considered a failure in macroeconomics, because investment is poorly linked to "the" interest rate. Well of course not, the proper cost of capital is driven entirely by the risk premium not "the" interest rate. The difference in thinking about consumer (we call them investors) behavior is starker still. For example, the consumers in the habit model (Cambpell and Cochrane (1999)) consider very strong precautionary saving motives and very strong intertemporal substitution motives, along with large time varying risk aversion. Their behavior is very far from the permanent-income intuition (or "liquidity constrained" alternative) in macroeconomic thinking. As one simple story, macroeconomists often think about how consumers will respond to a change in "wealth," coming from a change in stock prices or house prices. We would point out that such a change might engender a very different response if it corresponds to a *discount rate rise* – a temporary change in price with no change in capital stock or cashflow – than if it corresponds to a change in cashflows, or destruction of physical capital stock.

Formal macroeconomics has started to introduce some of the same ingredients that macro-finance researchers are using to understand discount rate variation, including "new" preferences, adjustment cost or other frictions in capital formation, and financial frictions in credit markets. For example, empirically-oriented new-Keynesian DSGE models (for example Christiano, Eichenbaum and Evans (2005)) now often include habits in utility and some adjustment-cost friction in capital formation. Though Kydland and Prescott (1982) had a "time to build" friction in production, most of the subsequent "real business cycle" literature following King, Plosser and Rebelo (1988) left out adjustment costs. They didn't need them to match the basic quantity correlations on which they focused, but such modes lead to the prediction $q = 1$ always and so don't do well on asset prices. The first round of "new Keynesian" literature abstracted from capital altogether.

This process is incomplete, however. For example, the habits in these models only include one period lags $(c_t - bc_{t-1})^{1-\gamma}$, and so will not generate long-run predictability. And the estimation or calibration of these models typically pays little attention to matching discount rate variation facts. The models do not generate anything like the level or variation of the equity premium, to say nothing of value effects.

Of course, macroeconomists might point out that general-equilibrium finance models don't replicate the macroeconomic facts they're interested in such as responses to monetary policy shocks either. And it's not as if anyone is unaware of the importance of the question. The job is just hard. At a technical level, macro models are complicated, and first-order approximations are hard enough to work out. To capture risk premia, you need second-order approximations, and to capture time-varying risk premia you need third order approximations. And that's just to put "macro" risks in business cycle models. Putting "financial frictions" in such models is harder still. At a deeper level, everyone recognizes that "grand synthesis" models do not consist of just mixing all the popular ingredients together and stirring the pot.

Still, the fact remains that asset markets *are* where marginal rates of substitution are equated (or not) to marginal rates of transformation, and that the current generation of macroeconomic models, though making progress, still does not capture the fundamental facts revealed by asset markets.

Monetary economics and asset pricing also have a lot to share. The basic financial system considered in monetary economics has two assets, "money," which pays no interest and is used for transactions, and "bonds" or "savings accounts," which are a store of value but must be exchanged for money at a substantial transactions cost before being used for transactions. Obviously, though this may represent something of what consumers faced in the 1930s, it is nothing like our modern financial system, which includes a dizzying array of highly liquid, interest-paying assets. Based on asset pricing insights, I and others (see Cochrane (2005a), (2010) and the citations therein) have been working an alternative foundation for inflation, the "fiscal theory of the price level."

The central insight here is straight from finance: Nominal government debt is exactly the same thing as "equity" in the government; it is the residual claim to primary fiscal surpluses. From this insight, the price level can be determined even in a completely frictionless economy from the standard asset pricing equation,

$$\frac{\text{Debt}}{\text{Price level}} = E_t \sum_{j=0}^{\infty} m_{t,t+j} \text{ (primary surplus}_{t+j}\text{)}. \tag{14}$$

This approach determines the price level with no special or transactions status for money, no need for limited money supply, or even no money at all, so long as the government issues nominal debt. More importantly, this analysis points to variation in the *discount rate* $m_{t,t+j}$ for government debt as the primary determinant of business cycle variation in "aggregate demand." A "flight to quality" such as in the recent recession lowers the discount rate for government debt. This change raises the real value of government debt, which implies deflation on the left side of (14). This analysis offers a very simple way to link the "rising risk premium" which finance people see as the core of a recession with the "decline in aggregate demand" which macroeconomists see. An audience used to the assaults of imperialistic macroeconomists bearing strange utility functions might appreciate a small bit of imperialistic finance, bringing our central present value equation, and the insight that *discount rate* variation accounts for the bulk of its movement over time.

8 Conclusions

Since the late 1970s, the bulk of research in asset pricing is best described as characterizing and understanding discount rate variation. This change is prompted by facts: discount rate variation is much larger than we thought, accounting for the bulk of the movement of valuation ratios over time and across assets. The puzzles and anomalies that we face all trace back to discount rate variation rather than problems of information incorporation. That doesn't mean the discount rates are "right", 'rational" or "socially efficient." It just means that they are the central question we're all investigating.

As I hope my survey made clear, we are really only beginning the task. The simple empirical task of characterizing discount rate (expected return) variation in the time series and cross section is begging for a consolidation. The first consolidation, in which Fama and French reduced a range of anomalies to size and book/market factors needs to be repeated. It will be much harder, as the number of factors has increased dramatically. Achieving the consolidation will, I think, require a methodological consolidation as well, in which we explicitly describe expected returns and covariances as functions of characteristics, going beyond multidimensional portfolio sorts, and formally and fully uniting the time-series regression and cross-sectional portfolio formation approaches. As the last time around, the challenge is really clearly stating the *question*.

Throughout, I emphasized my view that discount rate variation will force us to recast our thinking in terms of prices and payoff streams rather than one-period expected returns. Valuation ratios should be the left hand variable, the thing we explain, not ad-hoc characteristics for sorting portfolios. One-period betas, driven at least in large part by covariation among discount rates, are just as "endogenous" as the expected returns they conventionally "explain." Furthermore, taking a long-horizon perspective can dramatically simplify both asset pricing ad portfolio theory in a highly dynamic world. Mertonian state variables for one-period returns disappear in both cases, at least as a first approximation.

I offer a categorization of theories, in self-conscious analogy to Fama's (1970) categorization of informational efficiency. Surely that categorization is incomplete, and subdivisions are surely warranted, especially in the fast-evolving theories positing financial frictions and liquidity effects. Still, I hope that the basic issues, and some of the basic unsolved questions, apply at this larger scale of aggregation.

Like the empirical work, this is not so much a compilation of settled theories. Like facts, theories of discount rate variation are in their infancy. The categorization I hope will help us to think clearly about how to construct more successful theories in the future. In particular, though we have a plethora of new utility functions to play with, investigation of non-traded risks, and how asset markets function to share risks, seems promising and unexplored. Research on the production side is lagging, and the theories of segmented markets, financial frictions, behavioral finance, and liquidity effects are in their infancy, often as much stories as theories.

Application, which requires well developed intuition more than formal theorizing, lags even further behind in digesting the facts that discount rates vary so much.

In particular, portfolio theory seems stuck on alpha opportunities for the one remaining mean-variance investor, which cannot apply to more than measure zero investors without undermining its own premises. The view of the world that there are multiple dimensions of risk, with shifting premiums, opens up a much richer portfolio theory. Asset markets are big insurance markets, in which people more exposed to a dimension of risk sell, and those more exposed buy. Viewed this way, our focus on "priced" factors, of interest only to the last remaining mean-variance investor – is unfortunate. To the typical investor, detailed understanding of the *unpriced* (mean excess return = 0) factors which are nonetheless maximally correlated with shocks to his outside risks, would be much more useful. We have not really even started looking for these. I also emphasized that long-run investors in a market with time-varying investment opportunities might do much better to think about portfolio theory in terms of prices and streams of payouts, rather than the financial engineering of dynamic portfolio strategies that achieve those payouts.

Other areas of financial application, including valuation, cost of capital, measuring distance to default, accounting, and regulation, implicitly assume i.i.d. returns much more than one might think – and recognizing the large variation of discount rates will fundamentally change those applied procedures.

9 References

Acharya, Viral and Lasse Heje Pedersen, 2005, "Asset Pricing with Liquidity Risk," Journal of Financial Economics, 77, 375-410.

Amihud, Yakov, Haim Mendelson, and Lasse Heje Pedersen, 2005, "Liquidity and Asset Prices," Foundations and Trends in Finance, 1, 269-364

Ang, Andrew, and Jun Liu, 2004, "How to Discount Cashflows with Time-Varying Expected Returns," Journal of Finance 59, 2745-2783.

Ang, Andrew, and Joe Chen, 2007, "CAPM Over the Long Run: 1926-2001," Journal of Empirical Finance 14, 1-40.

Asness, Cliff, Toby Moskowitz and Lasse Pedersen, 2010, "Value and Momentum Everywhere" Manuscript, AQR Capital Management, University of Chicago, and NYU.

Baba, Naohiko and Frank Packer, 2008, "Interpreting deviations from covered interest parity during the financial market turmoil of 2007–08," BIS Working Paper 267.

Bansal, Ravi, and Amir Yaron, 2004, "Risks For the Long Run: A Potential Resolution of Asset Pricing Puzzles," Journal of Finance 59, 1481-1509.

Barberis, Nicholas, 2000, "Investing for the Long Run when Returns are Predictable," Journal of Finance, 55, 225-264.

Barberis, Nicholas, Tano Santos, and Ming Huang, 2001,"Prospect Theory and Asset Prices," Quarterly Journal of Economics 116, 1-53.

Barberis, Nicholas, and Richard Thaler, 2003, "A Survey of Behavioral Finance" in George Constantinides, René. Stulz and Milton Harris eds., Handbook of the Economics of Finance, North Holland, Amsterdam, 2003.

Barro, Robert J., 2006, "Rare Disasters and Asset Markets in the Twentieth Century," Quarterly Journal of Economics 121, 823-866.

Beeler, Jason, and John Y. Campbell, 2009, "The Long-Run Risks Model and Aggregate Asset Prices: An Empirical Assessment", NBER Working Paper 14788, http://kuznets.fas.harvard.edu/~campbell/papers/BeelerCampbell030209.pdf

Belo, Frederico, 2010, "Production-Based Measures of Risk for Asset Pricing,"Journal of Monetary Economics, 57(2), 146-163.

Boudoukh, Jacob, and Robert Whitelaw, 1991, "The Benchmark Effect in the Japanese Government Bond Market ," *Journal of Fixed Income,* 1, 52-59.

Brandt, Michael, John H. Cochrane and Pedro Santa-Clara, 2006, "International Risk Sharing is Better Than You Think, or Exchange Rates are Too Smooth," Journal of Monetary Economics 53, 671-698.

Brandt Michael and Kenneth A. Kavajecz, 2004, "Price Discovery in the U.S. Treasury Market: The impact of Orderflow and Liquidity on the Yield Curve," Journal of Finance 59, 2623-2654.

Brennan, Michael J., Eduardo S. Schwartz and Roland Lagnado, 1997, Strategic asset allocation, *Journal of Economic Dynamics and Control* 21, 1377-1403.

Brennan, Michael J., Ashley W. Wang, and Yihong Xia, 2004, "Estimation and test of a simple model of intertemporal asset pricing," Journal of Finance 59, 1743–1775.

Brennan, Michael J., and Yihong Xia, 2006, "Risk and valuation under an intertemporal capital asset pricing model," Journal of Business 79, 1–35.

Brunnermeier, Markus, and Lasse H. Pedersen , 2009, "Market Liquidity and Funding Liquidity," The Review of Financial Studies, 22, 2201-2238.

Brunnermeier, Markus, 2009, "Deciphering the Liquidity and Credit Crunch 2007-2009" Journal of Economic Perspectives 23, 77-100.

Campbell, John Y., and John Ammer, 1993, "What Moves the Stock and Bond Markets? A Variance Decomposition for Long-Term Asset Returns," Journal of Finance 48, 3–37.

Campbell, John Y., and John H. Cochrane, 1999 "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior," Journal of Political Economy 107, 205-251.

Campbell, John Y., and John H. Cochrane, 2000, "Explaining the Poor Performance of Consumption-Based Asset Pricing Models," Journal of Finance 55, 2863-78.

Campbell, John Y., and Jianping Mei, 1993, "Where do betas come from? Asset pricing dynamics and the sources of systematic risk," Review of Financial Studies 6, 567–592.

Campbell, John Y. and Robert J. Shiller, 1987, "Cointegration and Tests of Present Value Models" Journal of Political Economy 95, 1062–1088.

Campbell, John Y., and Robert J. Shiller,1988, "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," Review of Financial Studies 1, 195–228.

Campbell, John Y., and Robert J. Shiller, 1991, "Yield Spreads and Interest Rate Movements: A Bird's Eye View," Review of Economic Studies 58, 495–514.

Campbell, John Y., and Luis Viceira, 1999, "Consumption and Portfolio Decisions When Expected Returns are Time Varying", Quarterly Journal of Economics 114, 433-495.

Campbell, John Y., and Luis M. Viceira, 2001, "Who Should Buy Long-Term Bonds?," American Economic Review 91, 99-127.

Campbell, John Y., and Luis M. Viceira, 2002, Strategic Asset Allocation: Portfolio Choice for Long-Term Investors, Clarendon Lectures in Economics, Oxford University Press.

Campbell, John Y., and Tuomo Vuolteenaho, 2004, "Bad Beta, Good Beta,"American Economic Review 94, 1249-1275.

Carhart, Mark M., 1997, "On Persistence in Mutual Fund Performance," Journal of Finance 52, 57-82.

Chaves, Denis, 2009, "What Explains the Variance of Prices and Returns?: Time-series vs. Cross-section," Ph. D. Dissertation, University of Chicago, http://ssrn.com/abstract=1508433

Christiano, Lawrence, Martin Eichenbaum and Charles Evans, 2005, "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," Journal of Political Economy, 113 1–45.

Cochrane, John H., 1991a, "Volatility Tests and Efficient Markets: A Review Essay," Journal of Monetary Economics 27, 463-485.

Cochrane, John H., 1991b, "Production-Based Asset Pricing and the Link Between Stock Returns and Economic Fluctuations," Journal of Finance 46, 207-234.

Cochrane, John H., 1992, "Explaining the Variance of Price-Dividend Ratios," Review of Financial Studies 5, 243-280.

Cochrane, John H., 1993, "Rethinking Production Under Uncertainty," Manuscript, University of Chicago.
http://faculty.chicagobooth.edu/john.cochrane/research/Papers/
Cochrane%20Rethinking%20Production%20Under%20Uncertainty.pdf

Cochrane, John H., 1994, "Permanent and Transitory Components of GNP and Stock Prices," Quarterly Journal of Economics CIX, 241-266

Cochrane, John H., 1996, "A Cross-Sectional Test of an Investment-Based Asset Pricing Model," Journal of Political Economy 104, 572-621.

Cochrane, John H., 2001, "Review of 'Famous First Bubbles: The Fundamentals of Early Manias'," Journal of Political Economy 109, 150-1154.

Cochrane, John H., 2003, "Stock as Money: Convenience Yield and the Tech-Stock Bubble," in William C. Hunter, George G. Kaufman and Michael Pomerleano, Eds., Asset Price Bubbles Cambridge: MIT Press.
http://faculty.chicagobooth.edu/john.cochrane/research/Papers/
cochrane_stock_as_money.pdf

Cochrane, John H., 2005a, "Money as Stock," Journal of Monetary Economics 52, 501-528.

Cochrane, John H., 2005b, "Liquidity, Trading and Asset Prices," NBER Asset Pricing Program Report, NBER Reporter,
http://faculty.chicagobooth.edu/john.cochrane/research/Papers/
Cochrane_NBER_program_report_winter05.pdf

Cochrane, John H., 2005c, Asset Pricing, Revised Edition. Princeton: Princeton University Press.

Cochrane, John H., 2006, "The Dog That Did Not Bark: A Defense of Return Predictability," Review of Financial Studies 21, 1533-1575.

Cochrane, John H., 2007a, "Financial Markets and the Real Economy," in Rajnish Mehra, Ed. Handbook of the Equity Premium Elsevier, 237-325.
http://faculty.chicagobooth.edu/john.cochrane/research/Papers/
financial_and_real_proofs_aug_07.pdf

Cochrane, John H., 2007b, "Portfolio Theory," Manuscript, University of Chicago,
http://faculty.chicagobooth.edu/john.cochrane/research/Papers/portfolio_text.pdf.

Cochrane, John H., 2008, “A Mean-Variance Benchmark for Intertemporal Portfolio Theory,” manuscript, University of Chicago, http://faculty.chicagobooth.edu/john.cochrane/research/Papers/mean_variance_benchmark.pdf

Cochrane, John H., 2010, “Understanding policy in the great recession: Some unpleasant fiscal arithmetic,” Forthcoming, European Economic Review http://faculty.chicagobooth.edu/john.cochrane/research/Papers/EER_2262.pdf;

Cochrane, John H., and Lars Peter Hansen, 1992, “Asset Pricing Explorations for Macroeconomics,” In Olivier Blanchard and Stanley Fisher, Eds.,1992 NBER Macroeconomics Annual, 115-165.

Cochrane, John H., Francis Longstaff and Pedro Santa-Clara, 2003, “Two Trees,” Review of Financial Studies 21, 347-385.

Cochrane, John H., and Monika Piazzesi, 2005, “Bond Risk Premia,” American Economic Review 95, 138-160.

Cochrane, John H., and Monika Piazzesi, 2008, “Decomposing the Yield Curve,” Manuscript, University of Chicago, http://faculty.chicagobooth.edu/john.cochrane/research/Papers/interest_rate_revised.pdf

Constantinides, George M. 1979, “Multiperiod Consumption and Investment Behavior with Convex Transactions Costs,” Management Science 25, 1127-1137

Constantinides, George M. and Darrell Duffie, 1996, “Asset Pricing with Heterogeneous Consumers,” Journal of Political Economy 104, 219-240.

Daniel, Kent, and David Marshall, 1997, “The Equity Premium Puzzle and the Risk-Free Rate Puzzle at Long Horizons,” Macroeconomic Dynamics1, 452-484.

Daniel, Kent, and Sheridan Titman, 2006, “Testing Factor-Model Explanations of Market Anomalies,” Manuscript, Columbia University and University of Texas, http://www.columbia.edu/~kd2371/papers/unpublished/dt5.2.pdf

Duffie, Darrell, and Antje Berndt, 2011, “Explaining Corporate Debt Returns ” Manuscript, Stanford University.

Epstein, Larry G. and Stanley Zin, 1989,“ Substitution, risk aversion and the temporal behavior of consumption and asset returns: A theoretical framework,” Econometrica 57, 937–969.

Evans, Martin D.D., and Richard K. Lyons, 2002, “Order Flow and Exchange Rate Dynamics” Journal of Political Economy, 110, 170-180.

Fama, Eugene F., 1970, “Efficient Capital Markets: A Review of Theory and Empirical Work," Journal of Finance 25, 383-417.

Fama, Eugene F., 1984, “Forward and Spot Exchange Rates,” Journal of Monetary Economics 14, 319-38.

Fama, Eugene F., 1986, “Term Premiums and Default Premiums in Money Markets.” Journal of Financial Economics 17, 175-96.

Fama, Eugene F., 1991, “Efficient Capital Markets: II,” Journal of Finance 46, 1575-617.

Fama, Eugene F. 1996, "Multifactor Portfolio Efficiency and Multifactor Asset Pricing," Journal of Financial and Quantitative Analysis, 31, 441-465.

Fama, Eugene F.,1998, "Market Efficiency, Long-Term Returns, and Behavioral Finance," Journal of Financial Economics 49, 283-306.

Fama, Eugene F. and Robert R. Bliss, 1987, "The Information in Long-Maturity Forward Rates," American Economic Review 77, 680-92.

Fama, Eugene F. and Kenneth R. French, 1988, "Dividend Yields and Expected Stock Returns," Journal of Financial Economics 22, 3-25.

Fama, Eugene F. and Kenneth R. French, 1989, "Business Conditions and Expected Returns on Stocks and Bonds," Journal of Financial Economics, 25, 23-49.

Fama, Eugene F. and Kenneth R. French , 1993, "Common Risk Factors in the Returns on Stock and Bonds." Journal of Financial Economics 33, 3-56.

Fama, Eugene F. and Kenneth R. French, 1995, Size and Book-to-Market Factors in Earnings and Returns," Journal of Finance 50, 131-55.

Fama, Eugene F. and Kenneth R. French, 1996, "Multifactor Explanations of Asset Pricing Anomalies," Journal of Finance 51, 55-84.

Fama, Eugene F., and Kenneth R. French, 1997, "Industry Costs of Equity," Journal of Financial Economics 43, 153-93.

Fama, Eugene F., and Kenneth R. French, 2000, "Forecasting Profitability and Earnings," Journal of Business 73, 161-75.

Fama, Eugene F., and Kenneth R. French, 2006, "The Value Premium and the CAPM," The Journal of Finance, 61, 2163-85.

Fontana, Alessandro, 2010, "The persistent negative CDS-bond basis during the 2007/08 financial crisis" Manuscript, Unversitá Ca Foscari, Venice,
http://venus.unive.it/fontana/wp1.pdf

Fontana, Alessandro, and M. Scheicher, (2010) "An Analysis of Euro Area Sovereign CDS and their Relation with Government Bonds," European Central Bank, Working Paper No. 1271 .

Frazzini, Andrea and Lasse Hege Pedersen, 2010, "Betting against beta," Manuscript, New York University

Froot, Kenneth. A., and P. O'Connell, 2008, "On the Pricing of Intermediated Risks: Theory and Application to Catastrophe Reinsurance," Journal of Banking and Finance 32, 69-85.

Gabaix, Xavier, 2010, "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance," Manuscript, New York University,
http://pages.stern.nyu.edu/%7Exgabaix/papers/tenPuzzles.pdf

Gabaix, Xavier, 2010, "Linearity-Generating Processes: A Modelling Tool Yielding Closed Forms for Asset Prices," Manuscript, New York University,
http://pages.stern.nyu.edu/%7Exgabaix/papers/linearity.pdf

Gabaix, Xavier, Arvind Krishnamurthy, and Olivier Vigneron, 2007, "Limits of Arbitrage: Theory and Evidence from the Mortgage Backed Securities Market," *Journal of Finance,* 62(2), 557-596.

Garber, Peter M. 2000, Famous First Bubbles: The Fundamentals of Early Manias. Cambridge, Mass.: MIT Press

Gala, Vito, 2010, "Irreversible Investment and the Cross-Section of Stock Returns in General Equilibrium," Manuscript, London Business School

Garleanu, Nicolae, and Lasse Heje Pedersen, 2009, "Margin-Based Asset Pricing and Deviations from the Law of One Price", Manuscript, New York University, Forthcoming Review of Financial Studies.

Gomes, Joao, Leonid Kogan, and Lu Zhang, 2003, "Equilibrium Cross-Section of Returns," Journal of Political Economy 111, 693-732.

Gourio, Francois, 2007, "Labor Leverage, Firms Heterogeneous Sensitivities to the Business Cycle, and the Cross-Section of Returns" Manuscript, Boston University, http://people.bu.edu/fgourio/xspaper.pdf.

Gourio, Francois, 2010, " Putty-Clay Technology and Stock Market Volatility," Manuscript, Boston University, http://people.bu.edu/fgourio/submissionPCSM2010.pdf

Gorton, Gary B., and Andrew Metrick, 2009, "Securitized Banking and the Run on Repo" Yale ICF Working Paper 09-14.

Gourinchas, Pierre-Olivier and Hélène Rey, 2007, "International Financial Adjustment," Journal of Political Economy 115, 665-703.

Goyal, Amit, and Ivo Welch, 2008, "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," Review of Financial Studies 21, 1455-1508.

Hall, Robert E., 2011, "The Long Slump," Manuscript, Stanford University, http://www.stanford.edu/~rehall/Long%20Slump.pdf

Hansen, Lars Peter., John C. Heaton and Nan Li, 2008, "Consumption Strikes Back?: Measuring Long Run Risk," Journal of Political Economy, 116, 260-302.

Hansen, Lars Peter, and Robert J. Hodrick 1980,. "Forward Exchange-Rates As Optimal Predictors of Future Spot Rates - An Econometric-Analysis," Journal of Political Economy 88, 829-853.

Hansen, Lars Peter and Ravi Jagannathan, 1991, "Implications of Security Market Data for Models of Dynamic Economies," Journal of Political Economy 99, 225-262.

Hansen, Lars Peter, and Thomas J. Sargent, 2005, "Recursive Robust Estimation and Control under Commitment," Journal of Economic Theory, 124, 258-301.

Heaton, John, and Deborah J. Lucas, 2000, Portfolio choice and asset prices: The importance of entrepreneurial risk, *Journal of Finance* 55, 1163–1198.

He, Zhiguo and Arvind Krishnamurthy, 2010, "Intermediary Asset Pricing" Manuscript, University of Chicago and Northwestern University.

Jermann, Urban, 2010, "The Equity Premium Implied by Production," Journal of Financial Economics 98, 279-296.

King, Robert G., Charles I. Plosser and Sergio T. Rebelo, 1988, "Production, growth and business cycles : I. The basic neoclassical model," Journal of Monetary Economics 21,195-232.

Koijen, Ralph and Jules van Binsbergen, 2009, "Predictive regressions: A present-value approach," Manuscript, University of Chicago and Stanford University, Forthcoming Journal of Finance.

Kogan, Leonid, 2004, "Asset Prices and Real Investment," Journal of Financial Economics 73, 411-432.

Krishnamurthy, Arvind, 2002, "The Bond/Old-Bond Spread," Journal of Financial Economics 66, 463-506.

Krishnamurthy, Arvind, 2008, "Fundamental Value and Limits to Arbitrage," Manuscript, Kellogg Graduate School of Management.

Krishnamurthy, Arvind, 2010, "How Debt Markets Have Malfunctioned in the Crisis" Journal of Economic Perspectives 24, 328.

Kydland, Finn, and Edward C. Prescott, 1982, "Time to Build and Aggregate Fluctuations," Econometrica 50, 1345-1370.

Lacerda, Filipe, and Pedro Santa-Clara, 2010, "Forecasting Dividend Growth to Better Predict Returns, " manuscript Universidade Nova de Lisboa, http://docentes.fe.unl.pt/~psc/ForecastingDividendGrowthToBetterPredictReturns.pdf

Laibson, David, 1997, "Golden Eggs and Hyperbolic Discounting," Quarterly Journal of Economics 62, 443-77.

Lamont, Owen, 2000, "Investment Plans and Stock Returns," Journal of Finance 55, 2719-2745.

Lamont, Owen, 2004, "Go Down Fighting: Short Sellers vs. Firms," Manuscript, Yale University.

Lamont Owen, and Richard Thaler, 2003, "Can the Market Add and Subtract?: Mispricing in Tech-Stock Carve-Outs," Journal of Political Economy 111, 227-268.

Leamer, Edward, 2010, "Tantalus on the Road to Asymptopia" *Journal of Economic Perspectives* 24 (2) 31-46. http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.24.2.31

Lettau, Martin, and Sydney Ludvigson, 2001a, "Consumption, Aggregate Wealth and Expected Stock Returns," Journal of Finance, LVI, 815-849.

Lettau, Martin and Sydney Ludvigson, 2001b, "Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying," Journal of Political Economy, 109, pp. 1238-1287.

Lettau, Martin, and Jessica Wachter, 2007, "Why is Long-horizon Equity Less Risky? A Duration-based Explanation of the Value Premium," Journal of Finance LXII, 55-92

Lettau, Martin, and Sydney Ludvigson, 2005, "Expected Returns and Expected Dividend Growth," Journal of Financial Economics 76, 583-626.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, "A skeptical appraisal of asset pricing tests,"Journal of Financial Economics 96, 175-194.

Li, Erica X. Dmitry Livdan, and Lu Zhang, 2008, "Anomalies," Review of Financial Studies 22, 4301-4334

Liew, Jimmy, and Maria Vassalou, 2000, "Can book-to-market, size and momentum be risk factors that predict economic growth? Journal of Financial Economics 57, 221–245.

Liu, Laura Xiaolei, Toni M. Whited, and Lu Zhang, 2009, "Investment-Based Expected Stock Returns," Journal of Political Economy 117, 1105-1139.

Longstaff, Francis A., 2004, "The Flight to Liquidity Premium in U.S. Treasury Bond Prices," Journal of Business 77, 511-526.

Lucas, Robert E. Jr. 1978, "Asset Prices in an Exchange Economy," Econometrica 46, 1429-45.

Lustig, Hanno, Nikolai L. Roussanov, and Adrien Verdelhan, 2010a, "Common Risk Factors in Currency Markets," Manuscript UCLA, MIT, and Wharton School

Lustig, Hanno, Nikolai L. Roussanov, and Adrien Verdelhan, 2010b, "Countercyclical Currency Risk Premia," Manuscript UCLA, MIT, and Wharton School.

Markowitz, Harry, 1952, "Portfolio selection," *Journal of Finance*, 7, 77-99.

Merton, Robert C., 1971a, Optimum consumption and portfolio rules in a continuous time model, *Journal of Economic Theory* 3, 373-413.

Merton, Robert C., 1971b, An intertemporal capital asset pricing model, *Econometrica* 41, 867-87.

McCloskey, Deirdre, 1983, "The Rhetoric of Economics," Journal of Economic Literature 31, 482-517.

Miller, Edward M., 1977, "Risk, Uncertainty, and Divergence of Opinion," Journal of Finance 32, 1151–68.

Milgrom, Paul, and Nancy Stokey, 1982, "Information, trade and common knowledge," Journal of Economic Theory 26, 17-27.

Mitchell, Mark, Lasse Heje Pedersen, and Todd Pulvino, 2007,"Slow-Moving Capital," American Economic Review 97, 215-220.

Pastor, Lubos, 2000, "Portfolio selection and asset pricing models," Journal of Finance 55, 179-223.

Pastor, Lubos, and Robert F. Stambaugh, 2003, "Liquidity Risk and Expected Stock Returns," Journal of Political Economy 111, 642-685.

Piazzesi, Monika and Eric Swanson, 2008, “Futures Prices as Risk-Adjusted Forecasts of Monetary Policy,” Journal of Monetary Economics 2008, 55, 677-691.

Rietz, Thomas A. 1988. “The Equity Risk Premium: A Solution,” Journal of Monetary Economics, 22, 117–31.

Roll, Richard, 1977, “A Critique of the Asset Pricing Theory's Tests,” Journal of Financial Economics, 4, 129-176.

Ross, Stephen A., 1976, “The Arbitrage Theory of Capital Asset Pricing,” Journal of Economic Theory 13, 341-360.

Ross, Stephen A., 1978, “A Simple Approach to the Valuation of Risky Streams," Journal of Business 51, 453-475.

Sangvinatsos, Antonios and Jessica Wachter, 2005,“Does the failure of the expectations hypothesis matter for long-term investors?” Journal of Finance 60, 179–230.

Scheinkman, José A. and Wei Xiong, 2003, “Overconfidence and Speculative Bubbles,” Journal of Political Economy 111, 1183-1219.

Shiller, Robert. 1981, “Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?,” The American Economic Review, 71, 421-436.

Stambaugh, Robert, 1982, “On the Exclusion of Assets from Tests of the Two-Parameter Model: A Sensitivity Analysis,” Journal of Financial Economics 10, 237–268.

Thomas, Julia, 2002, “Is Lumpy Investment Relevant for the Business Cycle?” Journal of Political Economy, 110, 508–534.

Vuolteenaho, Tuomo, 2002, “What Drives Firm-Level Stock Returns?,” The Journal of Finance 57, 233–264.

Wachter,Jessica, 2003,“Risk aversion and allocation to long-term bonds,” Journal of Economic Theory 112,325–333.

Welch, Ivo, 2004, “Capital Structure and Stock Returns,” Journal of Political Economy 112, 106-131.

Part V

Appendix

10 Present values, volatility, bubbles, and long-horizons

I start with a derivation of the Campbell-Shiller (1988) linearization. The return is by definition,

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} = \frac{\left(1 + \frac{P_{t+1}}{D_{t+1}}\right)}{\frac{P_t}{D_t}} \frac{D_{t+1}}{D_t}$$

Therefore the log return is

$$\begin{aligned} r_{t+1} &= \log\left(1 + \frac{P_{t+1}}{D_{t+1}}\right) - \log\left(\frac{P_t}{D_t}\right) + \log\frac{D_{t+1}}{D_t} \\ r_{t+1} &= \log\left(1 + e^{-dp_{t+1}}\right) + dp_t + \Delta d_{t+1} \end{aligned}$$

I Taylor expand the first term about a constant PD. This constant need not be the mean.

$$\begin{aligned} r_{t+1} &= \log(1 + PD) - \frac{PD}{1+PD}\left(dp_{t+1} - dp\right) + dp_t + \Delta d_{t+1} \\ r_{t+1} &= \log(1 + PD) + \frac{PD}{1+PD}dp - \frac{PD}{1+PD}dp_{t+1} + dp_t + \Delta d_{t+1} \end{aligned}$$

Denoting $\rho = \frac{PD}{1+PD}$ then $PD = \frac{\rho}{1-\rho}$; $1 + PD = \frac{1}{1-\rho}$ and we can write the approximation

$$r_{t+1} = \kappa - \rho dp_{t+1} + dp_t + \Delta d_{t+1}$$

where

$$\kappa = \log(\frac{1}{1-\rho}) + \rho\log(\frac{1-\rho}{\rho}) = -(1-\rho)\log(1-\rho) - \rho\log(\rho).$$

In time-series applications where we will only consider second moments we interpret symbols as deviations from means, ignore κ, and write

$$r_{t+1} = -\rho dp_{t+1} + dp_t + \Delta d_{t+1}.$$

To find regression coefficients implied by a first-order VAR, I run

$$\begin{aligned} r_{t+1} &= b_r dp_t + \varepsilon^r_{t+1} && (15) \\ \Delta d_{t+1} &= b_d dp_t + \varepsilon^d_{t+1} \\ dp_{t+1} &= \phi dp_t + \varepsilon^{dp}_{t+1} && (16) \end{aligned}$$

Then I report

$$b_r^{(k)} = b_r \frac{1 - (\rho\phi)^k}{1 - \rho\phi}.$$

Adding returns, dividends or more lags to the VAR does not change the basic picture. Table 3 gives the estimated VAR.

I use dividend growth implied by the identity $r_{t+1} \approx -\rho dp_{t+1} + dp_t + \Delta d_{t+1}$. Actual dividend growth gives very similar results.

A second kind of "variance decomposition" takes variance of both sides of the Campbell-Shiller identity rather than regress both sides on dp_t as I have.

$$dp_t \approx \sum_{j=1}^{\infty} \rho^{j-1} \left(r_{t+j} - \Delta d_{t+j} \right)$$

$$var\left(dp_t\right) \approx var\left[E_t \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}\right] + var\left[E_t \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}\right] - 2cov\left[E_t \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}, E_t \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}\right]$$

Campbell and Ammer (1993) use this decomposition, where E_t is generated from a VAR. The trouble with this decomposition is that the covariance term is typically large. When the dividend yield is the only forecaster, the two terms are by construction perfectly correlated, for example. Also, the dividend growth term is substantial even when prices forecast dividend growth with the "wrong" sign. However, properly interpreted both decompositions provide the same information.

10.1 Multivariate challenge 1: understanding prices

I identified the shocks in Figure 6 by setting changes to the other variables in turn equal to zero. The return identity $r_t = -\rho dp_t + dp_{t-1} + \Delta d_t$ means that therefore some of the shocks must come with contemporaneous shocks to returns.

The dividend growth shock is a shock to dividend growth with no change in dividend yield or cay. If you change dividend growth and not yield, that means prices increase with dividends and there is a contemporaneous return $r_{t+1} = \Delta d_{t+1}$

The dividend yield shock has no change in dividend growth or cay. If dividends don't change and dp does, the price must change a lot. By the return identity $r_t = -\rho dp_t + dp_{t-1} + \Delta d_t$, a dividend yield shock implies a huge shock to *current* returns r_t, which decline by $\rho = 0.96$, as shown. Intuitively, a pure rise in discount rates lowers current return so it can raise subsequent returns. Expected *future* returns all rise.

The cay shock is a change in cay_t with no change in dividend yield dp_t or dividend growth Δd_t, and hence no change in return r_t.

I choose this definition of shocks. because it leads to nicely interpretable responses, e.g. "cashflow" and "discount rate." The resulting shocks are nearly uncorrelated, which is also convenient.

This VAR is very simple, since I left dividend growth out of the right hand side. My purpose is to distill the essential message of more complex VARs, and in such VARs, coefficients on dividend growth are small.

In the regressions, I rescaled cay by its standard deviation, so it has $\sigma(cay) = 1$. In its original units, the variation of cay was much smaller than that of $\sigma(dp) = 0.42$ leading to much larger regression coefficients. The scale of cay doesn't matter, since its regression coefficients add to zero. By this scaling my unit shocks are one-standard-deiviation shocks.

I also use dividends computed from the return identity (2): $\Delta d_{t+1} = r_{t+1} + \rho dp_{t+1} - dp_t$. This definition ensures that approximate identities hold exactly, but dividend growth forecasts rather than return forecasts reflect approximation errors. The results are almost the same with actual dividend growth.

Annual dividend growth also includes some return information, because one reinvests dividends to the end of the year at the market return. Sums of dividends are even less foreastable and a good deal less volatile; and have the conceptual advantage of not mixing in any return information, focusing only on decisions truly made by firms. However, if one uses sums of dividends over the year, then the identity $R_{t+1} = (P_{t+1} + D_{t+1})/P_t$ does not hold. It's nice to use data definitions for which identities hold!

The standard definition of dividends, which cumulates returns, has a second practical advantage. Consider the sharp fall in stocks in the Fall of 2008. Now, using a simple sum of past dividends, we would see a large decline in price/dividend ratio. But much of that decline surely reflected news that dividends in 2009 were going to fall dramatically. In this way, the sum-of-dividend definition gives a measure that should forecast dividend growth as well as returns. By reinvesting dividends to the end of the year, the "dividend" series is much lower than the sum; this price-dividend ratio already includes the information that dividends will decline next year and therefore produces a better return forecast.

The identity (5), reproduced here,

$$dp_t = E\left[\sum_{j=1}^{\infty} \rho^{j-1} r_{t+j} | I_t\right] - E\left[\sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j} | I_t\right],$$

suggests that an extra variable can only help to forecast long horizon returns if it forecasts long horizon dividend growth; it can help to forecast one year returns by changing the term structure of return forecasts as well. Here I show how that intuition applies algebraically to multiple regression coefficients and the impulse response function.

A bit more formally, we can generalize identity (3) for to multiple regression coefficients. Regressing both sides of (5) on dp_t and z_t, we obtain the generalized restriction on long-run multiple regression coefficients,

$$1 = b_r^{lr} - b_d^{lr} \tag{17}$$
$$0 = c_r^{lr} - c_d^{lr}, \tag{18}$$

where the notation refers to regressions

$$\begin{aligned} r_t^{lr} &= b_r^{lr} dp_t + c_r^{lr} z_t + \varepsilon^r \\ \Delta d_t^{lr} &= b_d^{lr} dp_t + c_d^{lr} z_t + \varepsilon^d \end{aligned}$$

Equation (17) is the same as before, now applied to the multiple regression coefficient. Equation (18) expresses the idea that a new variable can only help to forecast long-run returns if it also helps to forecast long-run dividend growth.

In terms of individual long-horizon regressions

$$r_{t+j} = b_r^{(j)} dp_t + c_r^{(j)} z_t + \varepsilon_{t+j}^r$$

etc., (5) similarly implies

$$
\begin{aligned}
1 &= \sum_{j=1}^{\infty} \rho^{j-1} b_r^{(j)} - \sum_{j=1}^{\infty} \rho^{j-1} b_d^{(j)} \\
0 &= \sum_{j=1}^{\infty} \rho^{j-1} c_r^{(j)} - \sum_{j=1}^{\infty} \rho^{j-1} c_d^{(j)}.
\end{aligned}
$$

A variable can help to forecast *one-year* returns, $c_r^{(1)} \neq 0$ only if it correspondingly changes the forecast of longer-horizon returns, or dividend growth.

Since impluse-response functions are the same as regression coefficients of future variables such as r_{t+j} on shocks at time t, the impulse-response functions

$$
(E_t - E_{t-1})\, r_{t+j} = e^{(j)}_{dp \to r} \varepsilon_t^{dp} + e^{(j)}_{z \to r} \varepsilon_t^{z}
$$

where $E_t \equiv E(\cdot|dp_t, z_t)$ must obey the same relation,

$$
\begin{aligned}
1 &= \sum_{j=1}^{\infty} \rho^{j-1} e^{(j)}_{dp \to r} - \sum_{j=1}^{\infty} \rho^{j-1} e^{(j)}_{dp \to \Delta d} \\
0 &= \sum_{j=1}^{\infty} \rho^{j-1} e^{(j)}_{z \to r} - \sum_{j=1}^{\infty} \rho^{j-1} e^{(j)}_{z \to \Delta d}.
\end{aligned}
$$

This fact lets me easily interpret the change in forecastability by adding cay, in the context of the present value identity, by plotting the impulse responses.

To see if the same pattern holds more generally, I tried a number of the forecasting variables in Goyal and Welch (2008). The results are in Table A1. Each of these variables helps substantially to forecast one-period returns. Yet they mean-revert quickly and don't forecast dividends much, so the contribution to the variance of dividend yields is still almost all from the variance of long-run expected returns.

	dp	cay	eqis	svar	ik	dfy
$b_{r,z}$		2.21	-0.71	1.48	-5.30	5.25
$t(b_{r,z})$		(1.73)	(-2.53)	(3.40)	(-0.85)	(1.86)
$b_{r,dp}$	0.13	0.10	0.19	0.15	0.11	0.13
$t(b_{r,dp})$	(2.61)	(1.82)	(3.75)	(3.05)	(2.16)	(2.53)
R^2	0.10	0.16	0.19	0.15	0.11	0.13
$\sigma(E_t \Sigma_{j=1}^{\infty} \rho^{j-1} r_{t+j})$	0.52	0.46	0.49	0.42	0.53	0.49
$\sigma(E_t \Sigma_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j})$	0.17	0.13	0.16	0.11	0.17	0.14

Table A1. Multiple return-forecasting regressions and implied variance of long-horizon returns. Data are from Goyal and Welch, 1947-2009. I calculate the variance of long-horizon expected returns and dividend growth from a bivariate VAR, and using actual (not identity) dividend growth forecasts. Equis, Percentage Equity Issuance is the ratio of equity issuing activity as a fraction of total issuing activity. Svar is stock variance, computed as sum of squared daily returns on the S&P 500.

Ik is the investment to capital ratio, the ratio of aggregate (private nonresidential fixed) investment to aggregate capital for the whole economy. Dfy, the default yield spread, is the difference between BAA and AAA-rated corporate bond yields.

An obvious first source of additional variables is less restrictive VARs than the simple first-order VAR with zero coefficients on lagged return and dividend growth that I have presented. Even in the information set of lagged $\{dp_t, r_t, \Delta d_t\}$, there is more information.

For example, the second lag of dividend yields is at least economically important. Table A2 presents the regressions

	dp_t	Δdp_t	$t(dp_t)$	$t\,(\Delta dp_t)$	R^2	$\sigma(E_t(y))\%$	$\frac{\sigma(E_t(y))}{E(y)}$
r_{t+1}	0.13	0.26	(2.45)	(1.83)	0.15	6.76	0.65
Δd_{t+1}	0.03	0.35	(0.62)	(3.27)	0.14	4.98	0.90
dp_{t+1}	0.93	0.10	(24.7)	(0.85)	0.91		
Δdp_{t+1}	-0.07	0.10	(-1.85)	(0.85)	0.06		

Table A2. Forecasts using dividend yield and change in dividend yield. CRSP value weighted return 1947-2009.

The change in dividend yield helps the return forecast, increasing R^2 from 0.09 to 0.15, and correspondingly increasing the more interesting measures of expected return variation.

The change in dividend yield really helps to forecast dividend growth, with a 3.27 t statistic, 5% standard deviation of forecast and forecast that varies by 90% of the mean. Of course, it is natural by the identity that a variable which forecasts dividend growth should also forecast returns. (Though an offsetting long-run dividend growth forecast is of course a logical possibility).

The 0.10 autoccorrelation in Δdp_t however suggests that this will be a very short-lived signal, one with little impact on forecasts of long-run dividend growth or returns, and thus to our view of the sources of price-dividend ratio volatility.

Similarly, while *individual* r_{t-j} and Δd_{t-j} coefficients don't look big, they can nonetheless help as a *group*, or by sensibly restricting the pattern of lagged coefficients. In this vein, Lacerda and Santa-Clara (2010) and Koijen and van Binsbergen (2009) find that moving averages of past dividend growth help to forecast both returns and dividend growth (as they must, given the present value identity), almost doubling the return-forecast R^2.

In sum, even within the $\{dp_t, r_t, \Delta d_t\}$ information set, we should regard the simple AR(1) structure as mainly pedagogical, and do real empirical work with more complex VAR representations.

10.2 Multivariate challenge 2: Right hand variables and factor structure

To express the factor structure of expected returns in equations, collect all the N_x forecasting variables for all N_r returns, as in Figure 7, into a big vector $\mathbf{x}_t$. Then, the multiple regressions we hope to end up with are of the form

$$\begin{aligned} r^i_{t+1} &= a_i + \mathbf{b}_i'\mathbf{x}_t + \varepsilon^i_{t+1}. \quad i = 1, 2, ...N_r \\ E(r^i_{t+1}|\mathbf{x}_t) &= a_i + \mathbf{b}_i'\mathbf{x}_t. \end{aligned}$$

Collect expected returns across asset classes into a big $N_r \times 1$ vector

$$\mathbf{Er}_t = \left[\begin{array}{cccc} E(r^1_{t+1}|\mathbf{x}_t) & E(r^2_{t+1}|\mathbf{x}_t) & ... & E(r^N_{t+1}|\mathbf{x}_t) \end{array} \right]'$$

Then, the question is whether we then write

$$\mathbf{Er}_t = \boldsymbol{\beta}_1 y_{1t} + \boldsymbol{\beta}_2 y_{2t} + ...$$

in a way that the first few factors capture almost all the common movement? Can we write the covariance matrix of expected returns,

$$cov(\mathbf{Er}_t, \mathbf{Er}_t') = \boldsymbol{\beta}_1 \boldsymbol{\beta}_1' \sigma^2(y_1) + \boldsymbol{\beta}_2 \boldsymbol{\beta}_2' \sigma^2(y_2) + ..$$

in a way that the first few terms are the most important?

The y_{jt} are in turn linear combinations of the $\mathbf{x}_t$, $y_{jt} = \boldsymbol{\gamma}_j' \mathbf{x}_t$. Thus, if there is a factor structure, we can estimate the expected returns with far fewer parameters. For example, if there is a one-factor model,

$$\mathbf{Er}_t = \left(\boldsymbol{\beta}_1 \boldsymbol{\gamma}_j'\right) \mathbf{x}_t$$

So rather than $N_r \times N_x$ unconstrained coefficients, we only have to estimate $N_r + N_x$ coefficients.

10.3 Cay

To calculate long run regressions, with $z = cay$, the table is

$$\left[\begin{array}{c} r_{t+1} \\ \Delta d_{t+1} \\ dp_{t+1} \\ z_{t+1} \end{array} \right] = \left[\begin{array}{cccc} 0 & 0 & b_r & c_r \\ 0 & 0 & b_d & c_d \\ 0 & 0 & \phi_{dp,dp} & \phi_{dp,z} \\ 0 & 0 & \phi_{z,dp} & \phi_{z,z} \end{array} \right] \left[\begin{array}{c} r_t \\ \Delta d_t \\ dp_t \\ z_t \end{array} \right] + \left[\begin{array}{c} \varepsilon^r_{t+1} \\ \varepsilon^d_{t+1} \\ \varepsilon^{dp}_{t+1} \\ \varepsilon^z_{t+1} \end{array} \right]$$

or

$$\begin{aligned} \left[\begin{array}{c} dp_{t+1} \\ z_{t+1} \end{array} \right] &= \left[\begin{array}{cc} \phi_{dp,dp} & \phi_{dp,z} \\ \phi_{z,dp} & \phi_{z,z} \end{array} \right] \left[\begin{array}{c} dp_t \\ z_t \end{array} \right] + \left[\begin{array}{c} \varepsilon^{dp}_{t+1} \\ \varepsilon^z_{t+1} \end{array} \right] \\ X_{t+1} &= \Phi X_t + \varepsilon^x_{t+1} \\ \left[\begin{array}{c} r_{t+1} \\ \Delta d_{t+1} \end{array} \right] &= \left[\begin{array}{cc} b_r & c_r \\ b_d & c_d \end{array} \right] \left[\begin{array}{c} dp_t \\ z_t \end{array} \right] + \left[\begin{array}{c} \varepsilon^r_{t+1} \\ \varepsilon^d_{t+1} \end{array} \right] \end{aligned}$$

then

$$\begin{aligned} E_t \left[\begin{array}{c} r^{lr}_{t+1} \\ \Delta d^{lr}_{t+1} \end{array} \right] &= \left[\begin{array}{cc} b_r & c_r \\ b_d & c_d \end{array} \right] \sum_{j=1}^{\infty} \rho^{j-1} \Phi^{j-1} \left[\begin{array}{c} dp_t \\ z_t \end{array} \right] \\ E_t \left[\begin{array}{c} r^{lr}_{t+1} \\ \Delta d^{lr}_{t+1} \end{array} \right] &= \left[\begin{array}{cc} b_r & c_r \\ b_d & c_d \end{array} \right] (I - \rho\Phi)^{-1} \left[\begin{array}{c} dp_t \\ z_t \end{array} \right] \end{aligned}$$

from which we can recover

$$\begin{aligned} E_t \left[\begin{array}{c} r^{lr}_{t+1} \\ \Delta d^{lr}_{t+1} \end{array} \right] &= \left[\begin{array}{cc} b^{lr}_r & c^{lr}_r \\ b^{lr}_d & c^{lr}_d \end{array} \right] \left[\begin{array}{c} dp_t \\ z_t \end{array} \right] \\ \left[\begin{array}{cc} b^{lr}_r & c^{lr}_r \\ b^{lr}_d & c^{lr}_d \end{array} \right. &= \left. \left[\begin{array}{cc} b_r & c_r \\ b_d & c_d \end{array} \right] (I - \rho\Phi)^{-1} \right] \end{aligned}$$

In these regressions, I rescaled cay by its standard deviation, so it has $\sigma(cay) = 1$. In its original units, the variation of cay was much smaller than that of $\sigma(dp) = 0.42$ leading to much larger regression coefficients. The scale of cay doesn't matter, since its regresion coefficients add to zero. By this scaling my unit shocks are one-standard-deiviation shocks.

I also use dividends computed from the return identity (2): $\Delta d_{t+1} = r_{t+1} + \rho dp_{t+1} - dp_t$. This definition ensures that approximate identites hold exactly, but dividend growth forecasts rather than return forecasts reflect approximation errors. The results are almost the same with actual dividend growth.

Actual annual dividend growth also includes some return information, because one reinvests dividends to the end of the year at the market return. Sums of dividends are even less foreastable and a good deal less volatile; and have the conceptual advantage of not mixing in any return information. However, if one uses sums of dividends over the year, then the identity $R_{t+1} = (P_{t+1} + D_{t+1})/P_t$ does not hold.

11 Fama French and the cross section

Fama and French (1996) use two factors, size and book/market of course. I emphasized book/market in the text as I think it is the most important. One can do a similar graphical presentation of the more famous two factor model and 25 portfolios.

Figure 30 plots the full sample average returns of the Fama French 25 portfolios, and their market betas. You can see there is little relation between the two, and again a puzzle.

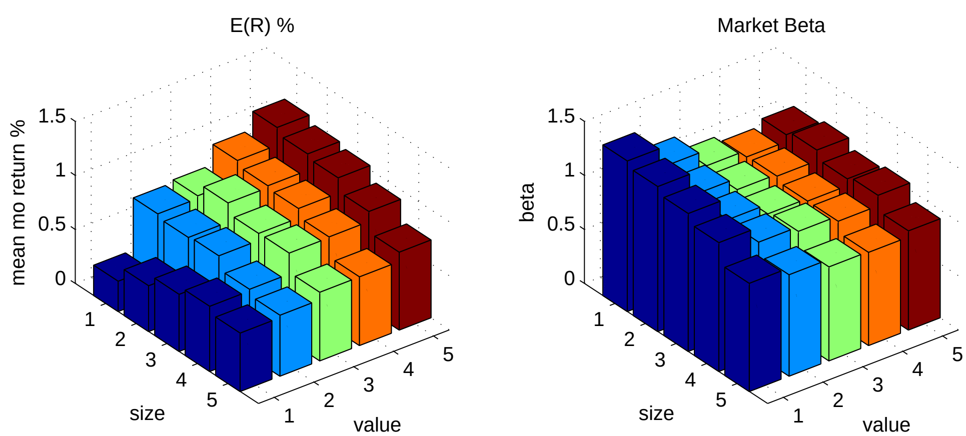

Figure 30: Average returns and single-regression market betas for Fama French 25 size and b/m sorted portfolios, monthly1926-2009.

Figure 31 presents the average returns with market, hml and smb exposures. Here you can see visually that portfolios with high average returns have high exposures to the two new factors.

Figure 32 presents the eigenvalue decomposition for the Fama French 25.

The first component is the market – though weighted towards small caps even a bit more than the equal-weighted market. The second and third components span size and book to

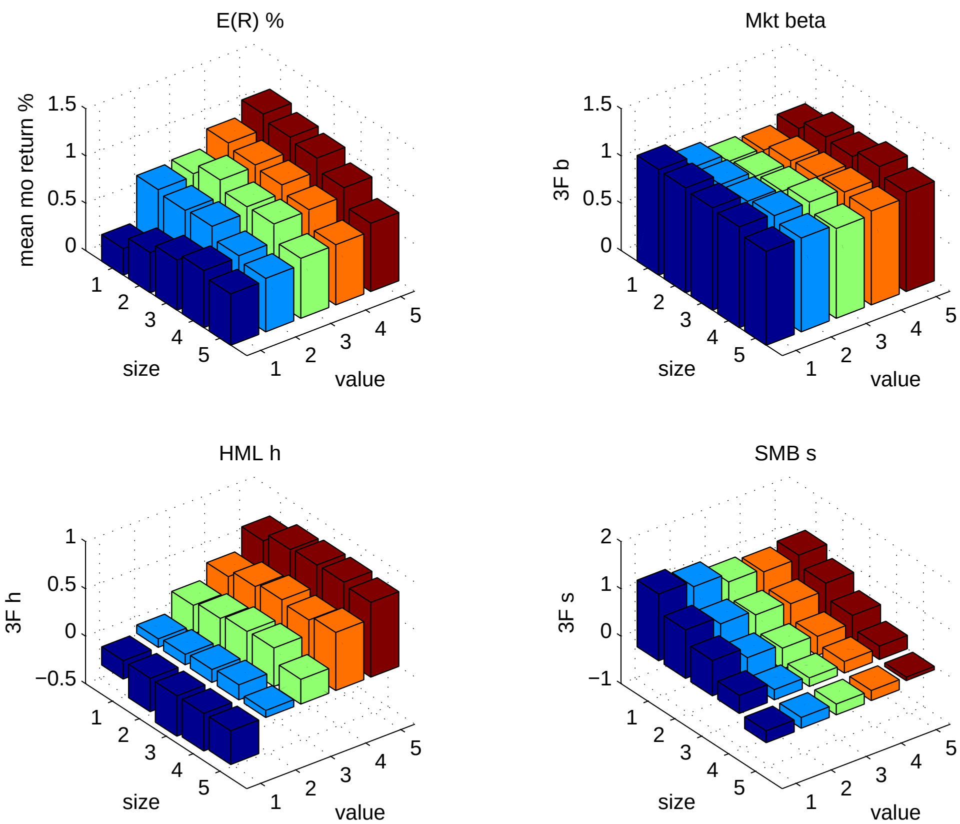

Figure 31: Average returns and four-factor betas, Fama-French 25 size and b/m sorted portfolios, monthly data 1926-2009.

market. (Eigenvalue decompositions rank by variance, so do not resolve well factors of about the same variance). Perhaps unsurprisingly, the fourth component is long small growth and short everything else, corresponding to the biggest anomaly in the Fama French model.

12 Asset Pricing as a function of characteristics

We want to know whether *differences* in mean returns across portfolios are significant. To address this question, many authors look at the 1-10 (or 1-20, or 1-100) portfolio return spread, and whether the corresponding alpha is zero. The number of portfolios seems to be determined by "sort finely enough so that the 1–N portfolio mean return spread is 100 bp/month with a t stat over 2."

There is some sense to this procedure, but we can do much better. Consider the ideal world for such an investigation: Expected returns rise with a characteristic C_i (for example, $C_i = \log(bm_i)$),

$$E(R^i) = a + b \times C_i,$$

and this variation corresponds exactly to a factor (for example, hml)

$$R^i_t = \beta^i \times f_t + \varepsilon_t,$$

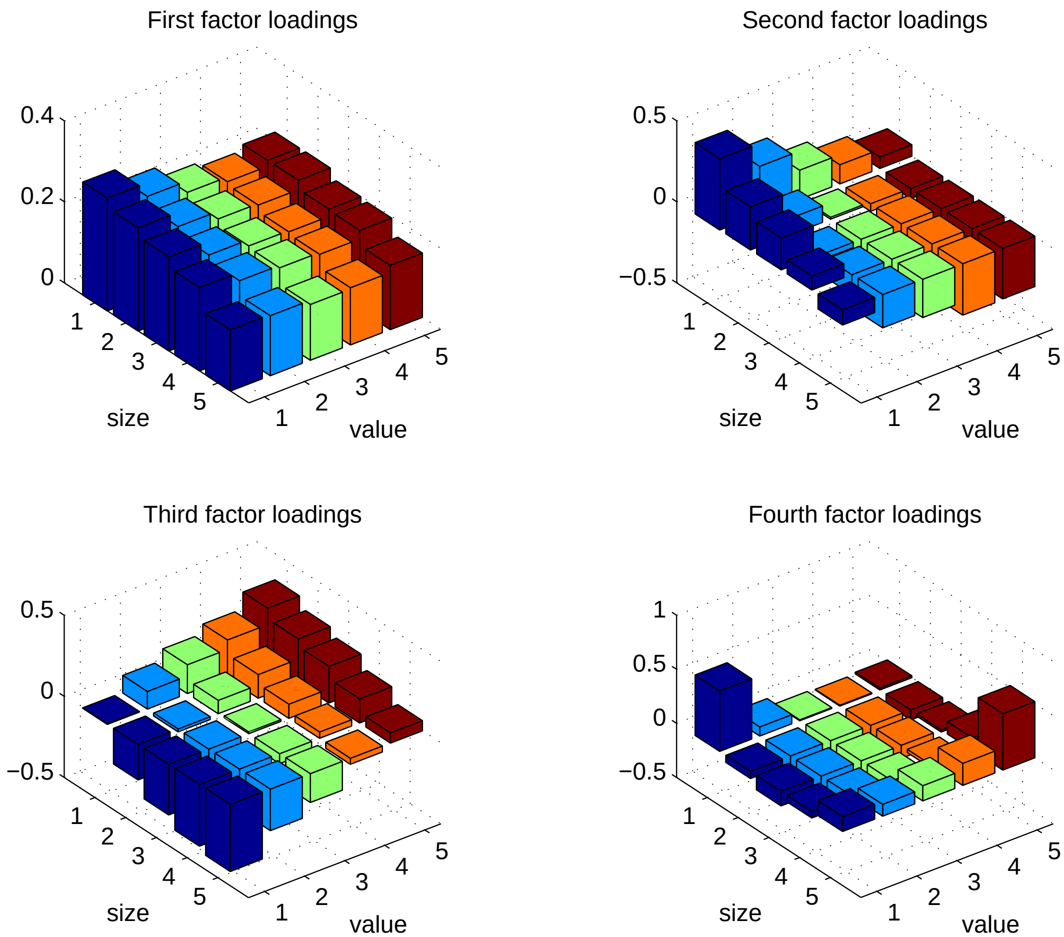

Figure 32: Eigenvalue factor analysis of the Fama-French 25 size and book/market portfolio returns

with betas that also rise with the characteristic

$$\beta^i = \frac{a}{E(f)} + \frac{b}{E(f)} \times C_i,$$

and uncorrelated residuals

$$cov(\varepsilon^i, \varepsilon^j) = 0.$$

Now, consider the usual 1-10 or 1-20 portfolio difference

$$\begin{aligned} E(R^i - R^j) &= b(C_i - C_j) \\ \sigma^2(R^i - R^j) &= \left(\beta^i - \beta^j\right)^2 \sigma_f^2 + 2\frac{\sigma_\varepsilon^2}{N} = \frac{b^2}{E(f)^2}\left(C^i - C^j\right)^2 \sigma_f^2 + 2\frac{\sigma_\varepsilon^2}{N} \end{aligned}$$

where N is the number of securities in each portfolio. Therefore, the Sharpe ratio, which is proportional to the t statistic $\sqrt{T}E/\sigma$for the mean spread-portfolio return, is

$$\frac{E(R^i - R^j)}{\sigma(R^i - R^j)} = \frac{E(f)}{\sigma_f} \frac{b(C_i - C_j)}{\sqrt{b^2\left(C^i - C^j\right)^2 + 2\frac{\sigma_\varepsilon^2}{N}\frac{E(f)^2}{\sigma_f^2}}}$$

This Sharpe ratio rises as we look at further separated portfolios. As $C_i - C_j$ increases, it approaches the pure Sharpe ratio of the factor $E(f)/\sigma(f)$. It does not increase forever – there is no sin, really, in looking at 1-100 portfolios if need be. Of course splitting the portfolio more finely reduces N, so finer splits end up reducing the Sharpe ratio.

Having seen this analysis, of course, we can see a better way. The extreme portfolio differences are an inefficient measure of the factor Sharpe ratio. It's much more efficient simply to examine the statistical significance of the cross-sectional regression coefficient $\hat{b}$, which uses information in all the securities, not just the end portfolios. Since $\hat{b} = cov(E(R^i), C^i)/var(C^i) = E(R^i \times \left[C^i - E(C^i)\right])/var(C^i)$ this regression coefficient is the same thing as testing the mean $E(f)$ of a factor which is also formed as a linear function of the characteristic $f_t = R_t^i \times \left[C^i - E(C^i)\right])$.

12.1 Time series and cross section

Here's how the regression of returns on dividend yields works out for Fama French size portfolios. Table A3 corresponds to Table 6 in the text, which examines the Fama-French book/market portfolios. Small-cap portfolios have lower dividend yields and higher returns than large-cap portfolios, so the pure cross-sectional regression shows the opposite sign of the return forecast, though the coefficient is an order of magnitude smaller than usual, 0.01 rather than 0.1. Here, higher pieces (low dividend yields) do seem associated with higher subsequent dividend growth. Again, with portfolio dummies, practically all dividend yield variation *over time for a given portfolio* comes from expected return variation, while variation *across portfolios in a given time period* is split between return and dividend growth forecasts. The pooled regression, weighting all kinds of variation equally still finds that on that basis all variation in dividend yield corresponds to return forecasts. Again, neither approach is satisfactory – we want to pursue this investigation as a function of multiple characteristics in individual securities, not via sorted portfolios.

	Cross section		Portfolio dummies		Time dummies		Pooled	
	b	$\frac{b}{1-\rho}$	b	$\frac{b}{1-\rho\phi}$	b	$\frac{b}{1-\rho\phi}$	b	$\frac{b}{1-\rho\phi}$
r	-0.014	-0.355	0.077	1.022	0.023	0.267	0.067	0.947
Δd	-0.030	-0.757	0.016	0.207	-0.045	-0.525	0.011	0.150
Δd^*	-0.048	-1.197	0.002	0.022	-0.063	-0.733	-0.004	-0.053
dp			0.963		0.952		0.968	

Table A3. Cross-sectional regression coefficients of average returns $E(r^i_{t+1})$, average dividend growth $E(\Delta d^i_{t+1})$, and dividend yield change on dividend yields $E(dp^i)$, 10 Fama-French ME (size) portfolios. Implied dividend growth is calculated from the approximate identity $\Delta d_{t+1} = r_{t+1} - \kappa + \rho dp_{t+1} - dp_t$.

13 Investment

Here, I derive (11) and (12).

The firm's problem is to maximize value

$$V(k_t) = \max_{\{i_{t+j}\}} E_t \int_{j=0}^{\infty} m_{t+j}\pi_{t+j} dj$$

with a technology given by

$$\begin{aligned}
\pi_t &= \theta_t f(k_t) - \left(1 + \frac{\alpha}{2}\left(\frac{i_t}{k_t}\right)\right) i_t \\
\frac{dk_t}{dt} &= -\delta k_t + i_t
\end{aligned}$$

and hence

$$k_{t+j} = e^{-\delta j} k_t + \int_{s=0}^{j} e^{-\delta(j-s)} i_{t+s} ds.$$

Differentiating with respect to i_t, the first order condition is

$$1 + \alpha\left(\frac{i_t}{k_t}\right) = \frac{\partial V(k_t)}{\partial k_t} = \frac{V(k_t)}{k_t}$$

The second equality follows because the technology is constant returns. The rate of return is

$$dR_t = \frac{dV_t}{V_t} + \frac{\pi_t}{V_t} dt$$

Substituting from the first-order condition,

$$\begin{aligned}
dR_t &= \frac{-\delta k_t dt + i_t dt + \alpha di_t + \theta_t f(k_t) dt - \left(1 + \frac{\alpha}{2}\left(\frac{i_t}{k_t}\right)\right) i_t dt}{k_t + \alpha i_t} \\
dR_t &= \frac{\theta_t f'(k_t) - \delta - \frac{\alpha}{2}\left(\frac{i_t}{k_t}\right)^2}{1 + \alpha\frac{i_t}{k_t}} dt + \frac{\alpha\frac{i_t}{k_t}}{1 + \alpha\frac{i_t}{k_t}}\left(\frac{di_t}{i_t}\right).
\end{aligned}$$