Lectures on asset pricing for macroeconomics

Jim Dolmas

February 21, 2013

Typeset in LaTeX, using Kile 2.1 by KDE.

Contents

Lecture 0

Introduction

These notes were written for a 'mini-course' on asset pricing that I gave for second-year Ph.D. students in macroeconomics at Southern Methodist University in the spring of 2012. The course was half a semester—seven three-hour lectures. I chose not to use a book for the course. A book like Duffie's *Dynamic Asset Pricing Theory* [Duf92] might have been useful for the very early material, on the basic theory surrounding stochastic discount factors, but after that the focus of the lectures was on published papers.

I also hadn't planned on writing over a hundred pages of notes: after the first couple lectures, the plan was to have the notes taper off to little more than lecture outlines. But, given that I compose in LaTeXonly a bit slower than I write by hand—and am better able to read what I wrote in the former case—it was difficult to stop once I got going.

0.1 Why study asset pricing?

Most grad students in macro need little motivation to study asset pricing. They can't help but notice the volume of work being done nowadays at the intersection of macro and finance.

Nevertheless, I think Sargent, as always, puts the motivation very well in the following few sentences. In a discussion of a paper by Ravi Bansal on 'long run risk', Sargent argued that this is work macroeconomists should pay attention to:

> Why? Because a representative agent's consumption Euler equation that links a one-period real interest rate to the consumption growth rate is the "IS curve" that is central to the policy transmission mechanism in today's New Keynesian models. A long list of empirical failures called puzzles come from applying the stochastic discount factor implied by that Euler equation. Until we succeed in getting a consumption-based asset pricing model that works well, the New Keynesian IS curve is built on sand. [Sar07]

One could also note that how we resolve asset pricing puzzles matters for how we think about the cost of business cycles, which is the source of my own interest in the subject.

0.2 Some stylized facts to aim for

If I were doing this again, I would begin with some facts up front—stylized facts that our models ought to be consistent with. As it was, the facts were introduced in the context of the various models. Some important facts would be:

1. The equity premium: The average real return on equity (says measured by a broad value-weighted index of stocks) has been historically large, on the order of 7 percent. The average real return on a relatively riskless asset, like a Treasury bill or commerical paper, has been low, on the order of 1 percent. The difference between the two—the equity premium—has averaged around 6 percentage points.

2. Return volatility: The standard deviation of the equity return is large, around 15 or 16 percent; the standard deviation of the riskless rate is much smaller, on the order of 3 percent.[1]

3. The market Sharpe ratio: Around 0.5 on average, but conditional Sharpe ratios are subject to considerable variation, with swings from 0 (near business cycle peaks) to 1 (at business cycle troughs) not uncommon [TW11, LN07].

4. Price-dividend ratios move around a lot, and have a lot of low frequency power. They appear to forecast returns (high P/D ratios imply low returns ahead), more so at longer horizons [Coc08]. Not everyone agrees with this, though [BRW08].

5. Short term interest rates are quite persistent.[2]

6. Nominal bond yields on average rise with maturity. Perhaps also for ex ante real yields.

7. The volatility of bond yields is fairly constant across maturities.

8. Consumption growth is not very volatile—a mean of about 2 percent and a standard deviation of about the same size (in annual data).

9. Consumption growth may be negatively autocorrelated/positively autocorrelated/*i.i.d.* The persistence isn't a robust fact across different samples. Mehra and Prescott [MP85] estimate a process with a first-order

[1] For facts 1 and 2, see [Koc96] among many, many others.

[2] An older, but I think still useful, reference for facts 5 through 7—and the problems they create for consumption-based models—is Wouter Den Haan's [Den95].

autocorrelation of -0.14 in annual data spanning nearly 100 years. In more recent samples one might find AC_1's on the order of $+0.30$. What *is* robust is that the AC_1 is small in absolute value. Consumption growth is not very persistent.

10. Aggregate dividend growth (or corporate earnings growth) is much more volatile than consumption growth. The correlation between the two is moderate—on the order of 50–60% [LP04].

0.3 Putting models on the computer

One of my aims for the class was to mix theory and computational work. Most of the exercizes—which are scattered throughout the text—involve writing some MATLAB code to solve some version of whatever model the lecture is about. With only seven weeks, though, and many models to cover, there's not much time to devote exclusively to computational techniques. So, the techniques involved are decidely not exotic. Most follow the basic template set by Mehra and Prescott's approach.

An appendix at the end of these notes covers some basics of using MATLAB. Guidance for solving particular problems appears at various points in the notes, wherever necessary.

0.4 Organization of the Lectures

The environment for lectures 1 and 2 is a two-period, today/tomorrow world. Decisions are taken today; uncertainty resolves tomorrow.

Lecture 1 covers portfolio choice in a mean-variance setting and the CAPM. In addition to deriving the CAPM in various versions—in terms of prices and payoffs, returns, and in beta form—the lecture also presents versions of the "Two Fund Theorem" (in the case of no riskless asset) and the "One Fund Theorem", a.k.a. the Tobin Separation Theorem (when there is a riskless asset).

Apart from a desire to "begin at the beginning", the point of lecture 1 is to illustrate some recurrent themes, within the context of the simple mean-variance model. Most important are the ideas of 'risk as covariance' and the pricing of risk.

From a technical standpoint, this lecture relies heavily on linear algebra, in contrast to the subsequent lectures.

Lecture 2 lays out the basic no-arbitrage approach to asset pricing in the context of a two period/N asset/S state environment. The keys results proven in the lecture are the Fundamental Theorem of Asset Pricing (relating no arbitrage to the existence of a state price vector) and the Representation Theorem (drawing an equivalence between the existence of a state price vector, a stochastic discount factor, and risk neutral probabilities). The proofs closely follow the approach of Dybvig and Ross [DR89] and Ross [Ros04].

The main takeaway from the lecture is the basic pricing relationship

$$p = \mathbb{E}(mx) \tag{1}$$

where $m \gg 0$ is the stochastic discount factor. This is the lens through which all the subsequent models are viewed.

With those results in hand, we turn to infinite-horizon consumption based models in **Lecture 3**, which begins with Lucas's 1978 paper [Luc78]. The focus here is less on Lucas's mathematical machinery and more on the structure of consumption based models. I also try to put Lucas's work into context, situating it against the backdrop of 1970s-style tests of market efficiency and the random walk hypothesis. In this and subsequent lectures, we take it for granted that (1) becomes

$$p_t = \mathbb{E}_t(m_{t+1}x_{t+1}) \tag{2}$$

in a many-period context.

Lucas's is the first consumption-based model we take to the computer—so the notes on Lucas include a description of the solution technique, including a first look at Markov chain approximations to $AR(1)$ processes.

From Lucas, the lecture turns to Mehra and Prescott [MP85] and the equity premium puzzle. The treatment is brief, given that the computational aspects have already been presented in the context of Lucas's model. This section also gives some alternative characterizations of the puzzle(s) in terms of second moment implications and bounds on attainable Sharpe ratios.

Lectures 4 and 5 then look at responses to the equity premium puzzle.

Lecture 1

The mean-variance model: Portfolio choice and the CAPM

This lecture treats the mean-variance model of portfolio choice and the equilibrium asset-pricing model based on it, the Capital Asset Pricing Model, or CAPM. Think of it as some 'pre-history' for the more modern approaches we will focus on for most of the course.

The mean-variance approach to portfolio choice—which emphasizes the reduction of risk by taking into account the covariances among asset payoffs—took a long time to emerge. Markowitz, in a history of portfolio choice [Mar99], notes that the theory had few real precursors. The few authors who did consider the problem of investment in risky assets—like Hicks [Hic35]—seemed to believe a version of the Law of Large Numbers meant all risk could be diversified away by investing in a large enough range of assets.

The insight of Markowitz [Mar52] was to recognize that asset payoffs were typically correlated and that no simple diversification could eliminate all risk. Rather, rational investors—if they care about the mean and variance of their final wealth—should take account of the covariance among asset payoffs in structuring their portfolios, so as to minimize the risk associated with any expected payoff. After Markowitz, it became clear that the risk of any asset was not a feature of the asset in isolation, but depended on how the asset's payoffs impacted the variance of the investor's portfolio. That, in turn, depends not just on the variance of the asset's payoff, but the covariances with other asset's payoffs.

The CAPM—the Capital Asset Pricing Model, developed independently by Sharpe [Sha64], Lintner [Lin65] and a couple others—followed very directly from Markowitz's model of portfolio choice. Like any model of equilibrium prices, it combines demands (Markowitz portfolios in this case) with supplies (exogenous supplies of risky assets) to determine prices. According to the

CAPM, an asset's price depends on the covariance of its payoffs with an aggregate payoff—that of the market portfolio. That role of covariance between individual asset payoffs and some aggregate is a feature that will carry over into all the more modern models we will examine throughout the course, though the aggregate will take different forms (like the marginal utility of consumption or wealth). In fact, once we've developed the more modern approach in the next lecture—which is based on stochastic discount factors, state prices or risk neutral probabilities—we'll reinterpret the CAPM as a stochastic discount factor model.

Before getting into the mean-variance model, though, we'll try to introduce some basic ideas through a series of simple examples, beginning with present discounted values and working our way to the role of covariance in pricing assets.

1.1 Basic ideas: The roles of arbitrage, risk aversion, and covariance

To begin, then, just think about present values. Imagine two periods (as we will through most of this introductory material); call them today and tomorrow. Agents can borrow and lend between the periods at a riskless interest rate r. Suppose there is an asset that pays x units of account tomorrow (per unit of the asset) with certainty. Agents can buy or sell units of the asset, and there is no restriction on short sales (agents can hold a negative position in the asset—someone who holds -1 units, for example, owes someone x tomorrow).

Under those conditions, the price p of the asset must obey

$$p = \frac{x}{1+r}$$

else an *arbitrage opportunity* is available. For example, if $p < x/(1+r)$, an agent could borrow p units of account, and buy a unit of the asset. Their net position today is zero, but tomorrow, they'll receive x and owe $(1+r)p < x$, gaining $x - (1+r)p$ essentially for free. Since they could do this at any scale, they have a 'money pump', a source of unlimited wealth. That's not an equilibrium. Thus $p \geq x/(1+r)$ must hold. A converse argument in which agents short the asset establishes $p \leq x/(1+r)$, so $p = x/(1+r)$ must hold.

Now, suppose there is uncertainty tomorrow—to be precise, suppose there are two possible states s tomorrow, $s \in \mathcal{S} = \{1,2\}$, which occur with probabilites $1/2$ and $1/2$. There is now also an asset whose payoff y is risky—in particular, $y : \mathcal{S} \to \mathbb{R}$ with $y(1) = 2x$ and $y(2) = 0$. This asset has the same expected payoff (x) as the first asset, but if agents care at all about risk, it shouldn't have the same equilibrium price. In fact, if agents dislike risk—and there are no other sources of uncertainty—its price $\hat{p}$ should obey

$$\hat{p} < \frac{E(y)}{1+r} = \frac{x}{1+r} = p.$$

Let's now add another risky asset, one whose payoffs (call them z) are in a sense opposite to those of the last asset, $z(1) = 0$ and $z(2) = 2x$. We'll also add another source of uncertainty that will give us some clue as to how y and z will be priced relative to one another: suppose that agents also receive endowments tomorrow, the endowments (call them e) are the same for all agents, and $e : \mathcal{S} \to \mathbb{R}$ with $e(1) \ggg e(2)$ (endowments if state 1 occurs are much larger than if state 2 occurs). In this case, even though both assets have expected payoffs equal to x, the z asset—which pays off in the state where endowments are low—will be regarded as more valuable by agents. Hence, we'd expect the z asset's price—call it $\tilde{p}$—to exceed the price of the y asset, $\hat{p}$.

We can say one more thing. Note that a portfolio consisting of one-half unit of the y asset and one-half unit of the z asset exactly replicates the payoffs of the riskless x asset: $(1/2)y(s) + (1/2)z(s) = x$ for $s = 1$ or $s = 2$. An arbitrage argument thus implies

$$(1/2)\hat{p} + (1/2)\tilde{p} = p$$

Note that an implication of the last equation (and the result that $\tilde{p} > \hat{p}$) is that $\tilde{p} > p > \hat{p}$—the risky z asset is more valuable than the riskless x asset. Even though it's payoff is uncertain, the z asset's price exceeds its expected present value because it covaries with the agents' endowments in a way that hedges their endowment risk.

These are very simple examples, but they illustrate a few concepts that will be important throughout these lectures—the roles played in the pricing of assets by arbitrage (precisely, the absence of arbitrage opportunities), risk aversion (more generally, curvature of the marginal utility of consumption or wealth), and covariances (between individual asset payoffs and more aggregate sources of uncertainty).

1.2 Mean-variance portfolio choice

1.2.1 No riskless asset

We begin with a framework in which there is no riskless asset. In addition to developing some basic concepts in mean-variance portfolio analysis, our main result will be the so-called *Two-fund Theorem*—all portfolios that are efficient (in a sense we will make precise below) can be constructed as linear combinations of precisely two portfolios (think of them as mutual funds, which explains the name of theorem).

There are N risky assets, with payoff vector $x = (x_1, x_2, \ldots x_N)$, which is a random variable taking values in $\mathbb{R}^N$. The assets have expected payoff vector given by

$$\mathbb{E}(x) = \mu$$

and a variance-covariance matrix Ω, which we assume to be positive definite.[1]

[1] An $N \times N$ matrix A is positive definite iff $u^\top Au > 0$ for all non-zero $u \in \mathbb{R}^N$. Note that any such A is non-singular.

Because Ω is a variance-covariance matrix, it is also symmetric. To be precise, $\Omega = [\omega_{i,j}]_{i,j=1,2,\ldots N}$ with

$$\omega_{i,j} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

In matrix notation, $\Omega = \mathbb{E}[(x-\mu)(x-\mu)^\top]$.

Remark 1.1. *Just a bit on notation. In contexts involving matrix algebra, think of all vectors as initially column vectors (so the mean vector μ, for example, is $N \times 1$). The superscript $\top$ denotes transposition. For two vectors x and y in $\mathbb{R}^N$, we'll use interchangeably the matrix product notation $x^\top y$ and the inner product notation $x \cdot y$—both denote $\sum_i x_i y_i$.*

Lastly, $x \geq 0$ means $x_i \geq 0$ for all i, equivalently $x \in \mathbb{R}^N_+$; $x > 0$ means $x \geq 0$ and $x \neq 0$, equivalently $x \in \mathbb{R}^N_+ \setminus \{0\}$; and $x \gg 0$ means $x_i > 0$ for all i, equivalently $x \in \mathbb{R}^N_{++}$.

We've already stated that there's no riskless asset, but even if we hadn't, the assumption that Ω is positive definite would rule that out. A riskless asset (say the ith) would have $\omega_{i,i} = 0$, in fact $\omega_{i,j} = 0$ for all j. If u is a vector with $u_i > 0$ and $u_j = 0$ for $j \neq i$, then $u \neq 0$ and $u^\top \Omega u = (u_i)^2 \omega_{i,i} = 0$. As we'll see below, once we define a portfolio, assuming that Ω is positive definite implies that there is no (non-zero) portfolio that has a zero variance.

There is also a vector of asset prices $p \in \mathbb{R}^N$, with $p \gg 0$, which is known to the investor at date 0. We'll be a bit vague for now about what units assets, payoffs and prices are in. Think of the payoffs as being in 'units of account' (eventually, they'll be in terms of consumption); assets as being in 'shares'; and prices in units of account per share.

A portfolio is a $z \in \mathbb{R}^N$ where z_i denotes the number of shares of the ith asset held by the investor. We do not rule out short sales—$z_i < 0$ is feasible for any i. The price of a portfolio z is $p \cdot z$ and its payoff is $z \cdot x$, a random variable. The portfolio's expected payoff is $\mathbb{E}(z \cdot x) = z \cdot \mathbb{E}(x) = z \cdot \mu \equiv \mu(z)$, using the linearity of the expectations operator. The variance of the portfolio's payoff is $\sigma^2(z) \equiv \mathbb{E}[(z \cdot x - z \cdot \mu)^2]$, which has a simple description in terms of z and Ω:

$$\begin{aligned}\sigma^2(z) &= \mathbb{E}[(z^\top(x-\mu))^2] \\ &= \mathbb{E}[z^\top(x-\mu)(x-\mu)^\top z] \\ &= z^\top \mathbb{E}[(x-\mu)(x-\mu)^\top] z \\ &= z^\top \Omega z\end{aligned}$$

The typical investor will have some initial wealth W_0 to allocate to his portfolio,

$$W_0 = p \cdot z, \tag{1.1}$$

and his final wealth will simply be the realized value of his portfolio

$$W_1 = z \cdot x.$$

There are no other sources of income. Thus, the mean and variance of his final wealth will simply be the mean and variance of his portfolio's payoff, $\mu(z)$ and $\sigma^2(z)$.

In the mean-variance portfolio choice approach, the investor is assumed to care only about the mean and variance of final wealth. Keeping variance the same, a higher mean is preferred, while keeping mean the same, a lower variance is preferred. The investor's mean-variance preferences may be justified in a number of ways, such as:

- This is a primitive assumption about the investor's utility function—it's simply defined as being over the mean and variance of final wealth, $v[\mu(z), \sigma^2(z)]$ say.

- The investor is an expected utility maximizer, and his von Neumann-Morgenstern utility function happens to be quadratic—e.g., $v(W_1) = aW_1 - (b/2)(W_1)^2$ with $a, b > 0$. In this case, the investor's expected utility takes the form $a\mu(z) - (b/2)\mu(z)^2 - (b/2)\sigma^2(z)$, which is decreasing in $\sigma^2(z)$ and increasing in $\mu(z)$ so long as $\mu(z) < a/b$.

- The investor has a utility function defined over the distribution of final wealth, but payoffs are normally distributed. Because the normal distribution is completely characterized by its first two moments, under some regularity conditions the investor's utility function will have a representation of the form $v[\mu(z), \sigma^2(z)]$.

We will assume investor preferences simply have the form $v[\mu(z), \sigma^2(z)]$, with v increasing in its first argument and decreasing in its second. For the main result of this sub-section, we don't need to be any more concrete than that. The typical investor's problem will be

$$\max_z v[\mu(z), \sigma^2(z)] \tag{1.2}$$

subject to the budget constraint (1.1).

We imagine that there are many investors with different levels of initial wealth W_0 and possibly different utility functions v; technically, we should superscript these as W_0^i and v^i, but as long as there is no confusion, we'll avoid introducing extra clutter to the notation.

A moment's reflection on the nature of the problem (1.2) will reveal that any z that solves it must be *mean-variance efficient*:

Definition 1.1. *A portfolio z is* mean-variance efficient *if any z' such that $\mu(z') > \mu(z)$ has $\sigma^2(z') > \sigma^2(z)$ and any z' such that $\sigma^2(z') < \sigma^2(z)$ has $\mu(z') < \mu(z)$. In words, a portfolio is mean-variance efficient if any portfolio with a higher mean has a higher variance, and any portfolio with a lower variance has a lower mean.*

Mean-variance efficient portfolios have the lowest variance among all portfolios having the same mean, and the highest mean among all portfolios having the same variance.

For a given initial wealth W_0, the set of mean-variance efficient portfolios can be found by solving the following family of problems, as we vary m:

$$\min_z \sigma^2(z) \tag{1.3}$$

subject to

$$\begin{aligned} \mu(z) &= m \\ W_0 &= p \cdot z \end{aligned}$$

Using the matrix results from above, we can re-write problem (1.3) in the following form:

$$\min_z z^\top \Omega z \tag{1.4}$$

subject to

$$z \cdot \mu = m \tag{1.5}$$

$$W_0 = p \cdot z \tag{1.6}$$

Note that this problem will only be interesting if the number of assets N exceeds 2. The reason is that, apart from the case where μ and p are collinear, the set of z that satisfy the two constraints (1.5) and (1.6) is a subset of dimension $N-2$, which in $\mathbb{R}^2$ is a singelton—i.e., a single portfolio z will satisfy both constraints, and we're left with really no problem to solve. Henceforth, we assume $N \geq 3$.

To solve the problem, form the Lagrangian

$$L = -\frac{1}{2} z^\top \Omega z + \lambda(z^\top \mu - m) + \pi(W_0 - z^\top p),$$

where the Lagrange mutlipliers are $\lambda, \pi \in \mathbb{R}$. The $1/2$ in front of the objective doesn't alter the solution and saves us from having some 2's floating around in the first-order conditions. The minus sign is there because I tend to write any optimization problem as a maximization problem—I find it economizes on the number of optimality conditions one needs to remember.

Differentiating L with respect to the vector z gives the first-order condition

$$-\Omega z + \lambda \mu - \pi p = 0$$

which is N equations stacked in vector form. Because Ω is positive definite, we can invert it to get

$$z = \lambda \Omega^{-1} \mu - \pi \Omega^{-1} p. \tag{1.7}$$

This is not yet a complete solution, because it still involves the multipliers λ and π. But, we can still see an important feature of the solution's structure. The efficient portfolio is a linear combination of two vectors, $\Omega^{-1}\mu$ and $-\Omega^{-1}p$, with weights λ and π.

Note, too, that the efficient portfolio only depends on m and W_0 through the scalars λ and π. For a given W_0—which is to say, for a given investor—as we vary m, λ and π will vary, but the variance-minimzing portfolio will always be a linear combination of the same two vectors, $\Omega^{-1}\mu$ and $-\Omega^{-1}p$. The same holds true as we vary W_0, which is in effect varying investors—all investors choose a linear combination of these same two vectors.

Since $\Omega^{-1}\mu$ and $\Omega^{-1}p$ are vectors in $\mathbb{R}^N$, we can think of them as portfolios (or funds). In this environment, all investors' asset demands could be met by two mutual funds, one offering the portfolio $\Omega^{-1}\mu$ and the other offering the portfolio $-\Omega^{-1}p$. Given what we've discussed above regarding equation (1.7), we may state:

Theorem 1.1 (The two-fund theorem)**.** *In the model of this section—no riskless asset, $N \geq 3$ assets, variance-covariance matrix Ω positive definite—there exist two portfolios such that any mean-variance efficient portfolio can be written as a linear combination of those two portfolios.*

We can actually say a bit more by solving for λ and π. This would be tedious algebra, but Exercize 1.1 asks you to show the following, by taking account of the constraints (1.5) and (1.6): at a solution to problem (1.4), the Lagrange multipliers λ and π are linear in W_0 and m. That is, there are coefficients $a_\lambda, a_\pi, b_\lambda$, and b_π—real numbers that themselves do *not* depend on W_0 and m—such that

$$\lambda = a_\lambda W_0 + b_\lambda m \tag{1.8}$$

$$\pi = a_\pi W_0 + b_\pi m \tag{1.9}$$

Exercize 1.1 (Linearity of λ and π)**.** *Combine the expression (1.7) for z with the constraints (1.5) and (1.6) to prove the claim from the previous paragraph. You can make this exercize less tedious by doing as much of it in matrix form as possible, using expressions like $\mu^\top\Omega^{-1}\mu$, $\mu^\top\Omega^{-1}p$, and $p^\top\Omega^{-1}p$.*

Given (1.8) and (1.9), we may re-write the expression (1.7) for a mean-variance efficient portfolio as follows:

$$\begin{aligned} z &= \lambda\Omega^{-1}\mu - \pi\Omega^{-1}p \\ &= (a_\lambda W_0 + b_\lambda m)\Omega^{-1}\mu - (a_\pi W_0 + b_\pi m)\Omega^{-1}p \\ &= W_0(a_\lambda\Omega^{-1}\mu - a_\pi\Omega^{-1}p) + m(b_\lambda\Omega^{-1}\mu - b_\pi\Omega^{-1}p) \end{aligned} \tag{1.10}$$

In (1.10), $a_\lambda\Omega^{-1}\mu - a_\pi\Omega^{-1}p$ and $b_\lambda\Omega^{-1}\mu - b_\pi\Omega^{-1}p$ are simply vectors in $\mathbb{R}^N$, which depend only on the parameters μ, Ω and p. We may think of these two vectors as portfolios, which for compactness, we'll define as $\hat{z}_1$ and $\hat{z}_2$, so that

$$z = W_0\hat{z}_1 + m\hat{z}_2 \tag{1.11}$$

Let's think now of many investors, assuming all of them have mean-variance preferences. Our work thus far has not identified the particular portfolio a utility-maximizing investor would choose, since the expected payoff m is itself a choice in the investor's utility-maximization problem—recall (1.2). We've only identified a family of portfolios to which any solution must belong—those that, for a given m have the lowest variance. The investor's preferences—the trade-off they make between additional mean payoff and additional variance—together with their initial wealth, will determine their choice of m. Without modeling that problem, let's simply let m^i denote investor i's choice of m, W_0^i his wealth, and z^i his resulting portfolio choice. Then,

$$z^i = W_0^i \hat{z}_1 + m^i \hat{z}_2. \tag{1.12}$$

Remark 1.2. *Note that even if all investors have identical preferences, their initial wealth level may affect their choices of m, so different investors would still make different choices. This would be the case if their preferences don't display constant relative risk aversion—for example if their preferences are quadratic (including both first- and second-order terms).*

Define the *market portfolio* as the sum of all investors' portfolios, and denote it by z^M. Summing (1.12) across all investors gives the market portfolio as

$$\begin{aligned} z^M &= (\sum_i W_0^i)\hat{z}_1 + (\sum_i m^i)\hat{z}_2 \\ &= W_0^M \hat{z}_1 + m^M \hat{z}_2 \end{aligned} \tag{1.13}$$

where W^M is the aggregate initial wealth of the market, and m^M is the aggregate expected payoff from all assets holdings in the market. Note that because the individual z^i satisfy the linear constraints $p \cdot z^i = W_0^i$ and $z^i \cdot \mu = m^i$, we have $p \cdot z^M = W_0^M$ and $z^M \cdot \mu = m^M$—that is, the market portfolio satisfies the constraints at the market initial wealth W_0^M and expected payoff m^M. Thus, (1.13) describes a mean-variance efficient portfolio, given W_0^M and m^M.

We've thus proved the other major result of this section:

Theorem 1.2 (Efficiency of the market portfolio)**.** *In the model of this section, if all investors have mean-variance preferences, the market portfolio is mean-variance efficient.*

Exercize 1.2 (The mean-variance frontier)**.** *This exercize asks you to use Matlab to find mean-variance efficient portfolios, then plot what's known as the* mean-variance frontier*, which is normally drawn in mean-standard deviation space. The data are for five risky assets:*

$$\mu = \begin{pmatrix} 1.38 \\ 1.29 \\ 1.47 \\ 1.65 \\ 1.25 \end{pmatrix}$$

and

$$\Omega = \begin{pmatrix} 26.63 & 13.63 & 10.86 & 6.15 & 8.55 \\ 13.63 & 34.46 & 15.26 & 7.59 & 11.01 \\ 10.86 & 15.26 & 38.32 & 7.69 & 13.49 \\ 6.15 & 7.59 & 7.69 & 24.7 & 6.50 \\ 8.55 & 11.01 & 13.49 & 6.50 & 19.01 \end{pmatrix}$$

The price vector is just $p = (1, 1, 1, 1, 1)^\top$, so you might think of the payoffs as gross returns*—payoffs per unit of account spent on the asset.*

Write some Matlab code to display two portfolios, or funds, from which all efficient portfolios can be constructed.

Assume an investor's initial wealth is $W_0 = 1$ and let m vary from the smallest expected payoff, $\min(\mu)$*, to the largest,* $\max(\mu)$*, in 50 evenly-spaced steps. You might use the Matlab command '*`linspace`*':* `m = linspace(min(mu),max(mu),50)`*. Write lines that solve for the multipliers λ and π at each of the the 50 points, calculate the efficient portfolio at each of the 50 points, and calculate the portfolio standard deviation $\sqrt{z^\top \Omega z}$ at each of the 50 points. You don't need to print out all those results, just show me the portfolios that correspond to the first and last m values, and all your code.*

Finally, plot m versus the portfolio standard deviations, with standard deviation on the horizontal axis. This is the mean-variance frontier (only its upper portion is relevant to investors). It's interesting to compare the mean-variance frontier with the means and standard deviations of the underlying assets. You can do this with Matlab's `hold` *command. After you've plotted the frontier, if* `mu` *and* `sd` *are the vectors of expected payoffs and standard deviations, enter* `hold;` *then* `scatter(sd,mu)`*. Note the standard deviations are just the square roots of the diagonal of Ω—*`sd = sqrt(diag(Omega))`*.*

1.2.2 Mean-variance portfolio choice when there is a riskless asset

We now add a riskless asset to the market. The riskless asset pays one sure unit of account next period for each unit of the asset purchased today. We'll let q denote the price of the riskless asset. An investor who buys $z_0 > 0$ units of the asset today pays qz_0 today and receives z_0 units of account tomorrow. An investor who shorts the riskless asset, choosing $z_0 < 0$ units, in effect borrows $-qz_0$ today and repays $-z_0$ next period. We can also describe the riskless asset by a (gross) riskless rate of return, denoted R_F, obeying

$$R_F = \frac{1}{q}.$$

Assumptions about the N risky assets remain the same as before: $\mu = \mathbb{E}(x)$ still denotes the expected payoff vector and $\Omega = \mathbb{E}[(x-\mu)(x-\mu)^\top]$ still denotes the variance-covariance matrix. We'll continue to let $z = (z_1, z_2, \dots z_N)$ denote a portfolio of the N risky assets, and p their prices.

A typical investor's budget constraint now takes the form

$$W_0 = qz_0 + p \cdot z \tag{1.14}$$

and his next-period wealth will be

$$W_1 = z_0 + z \cdot x. \tag{1.15}$$

We can use the budget constraint to eliminate z_0 from the expression for W_1, substituting $z_0 = (W_0 - p \cdot z)/q$ to get

$$\begin{aligned} W_1 &= \frac{1}{q} W_0 - \frac{1}{q} p \cdot z + z \cdot x \\ &= \frac{1}{q} W_0 + z \cdot (x - \frac{1}{q} p) \\ &= R_F W_0 + z \cdot (x - R_F p) \end{aligned} \tag{1.16}$$

where (1.16) uses $R_F = 1/q$. The investor's expected final wealth is $\mathbb{E}(W_1) = R_F W_0 + z \cdot (\mu - R_F p)$, and $W_1 - \mathbb{E}(W_1) = z \cdot (x - \mu)$, so the variance of final wealth will again take a quadratic form $\mathbb{E}[(W_1 - \mathbb{E}(W_1))^2] = z^\top \Omega z$. If an investor holds no risky assets ($z = 0$), his wealth simply grows at the gross riskless rate; if the investor spends all his initial wealth on his risky portfolio ($p \cdot z = W_0$), his expected final wealth is (as before) $z \cdot \mu$.

We could, as before, consider the problem of finding mean-variance efficient portfolios—for each m, the minimum-variance portfolio attaining $\mathbb{E}(W_1) = m$. Instead, though, we will go straight to the typical investor's utility maximization problem, assuming the investor has mean-variance preferences. The problem is

$$\max_z v[R_F W_0 + z \cdot (\mu - R_F p), z^\top \Omega z]. \tag{1.17}$$

Note that z in this problem is unconstrained—any difference between $p \cdot z$ and W_0 is made up for by long or short positions in the riskless asset.

Letting $v_1 > 0$ and $v_2 < 0$ denote the partial derivatives of v with respect to mean and variance, the first-order conditions are

$$v_1(\mu - R_F p) + 2v_2 \Omega z = 0 \tag{1.18}$$

which can be rearranged to give us:

$$z = -\frac{v_1}{2v_2} \Omega^{-1} (\mu - R_F p). \tag{1.19}$$

Note that the only term in (1.19) that is specific to the investor—that may reflect his wealth or tastes—is the *scalar* $-v_1/2v_2$, the marginal rate of substitution between mean and variance. The vector $\Omega^{-1}(\mu - R_F p)$ will be common to all investors, assuming they all have mean-variance preferences. We've thus derived—very simply—a 'one fund theorem' for our market including a riskless asset:

Theorem 1.3 (One fund theorem)**.** *Add a riskless asset to the model of the last subsection, and assume all investors have mean-variance preferences. Then, investors' optimal portfolios of risky assets are all scalar multiples of one another.*

Remark 1.3. *One can also see this result in the structure of the Lagrangian for the problem of finding a mean-variance efficient portfolio:*

$$L = -\frac{1}{2}z^\top \Omega z + \lambda(R_F W_0 + z \cdot (\mu - R_F p) - m)$$

The first-order conditions give $z = \lambda\Omega^{-1}(\mu - R_F p)$, and λ will be the only term that varies across investors.

Remark 1.4. *While* (1.19) *tells us something important about the structure of the solution to* (1.17)*, it does not necessarily constitute a complete solution for z, since it does not determine v_1/v_2, assuming this term is not a constant. Put differently, and maybe more precisely, the first-order conditions* (1.18) *are N equations in N unknowns, but not necessarily a system of* linear *equations.*

Remark 1.5. *Tobin [Tob58] proved a version of the One Fund Theorem in a 1958 paper on liquidity preference. The riskless asset in Tobin's model is money, and Tobin shows that investors' risky asset portfolios are all scalar multiples of a single portfolio. The result is thus sometimes known as the Tobin Separation Theorem.*

What about demand for the riskless asset? An investor's holdings of the riskless asset are obscured by the fact that we've substituted them out of the budget constraint. Once we solve the first-order conditions (1.18) for the risky asset demands z, we can find the investor's riskless asset demand by going back to the original budget constraint, $z_0 = (1/q)(W_0 - p \cdot z)$. In any case, the one fund theorem tells us that in a mean-variance environment with one riskless asset, every investor's portfolio can be represented as some amount of the riskless asset plus holdings of one risky portfolio (or fund) common to all investors.

1.3 The Capital Asset Pricing Model (CAPM)

The Capital Asset Pricing Model, or CAPM, was developed by Sharpe [Sha64] and Lintner [Lin65], building on the mean-variance portfolio analysis of Markowitz. It combines Markowitz's results on asset demands with some assumptions about asset supplies to derive equilibrium prices.

Again distinguish investors by their wealth W_0^i and preferences v^i. Let $\alpha_i = -2v_2^i/v_1^i$, which is positive under the assumption that $v_1^i > 0$ and $v_2^i < 0$. α_i is in units of mean per variance, and measures i's willingness to trade off higher (or lower) mean final wealth for higher (or lower) variance of final wealth. An investor with a high value of α_i, for example, would sacrifice a large reduction in mean for a small reduction in variance.

With that bit of notation, investor i's risky portfolio has the form

$$z^i = \frac{1}{\alpha_i}\Omega^{-1}(\mu - R_F p). \tag{1.20}$$

As we noted in Remark 1.4, α_i is not necessarily a constant, but we can still learn something about the nature of equilibrium prices even if we cannot explicitly solve it out of the investor's risky asset demands.

If we aggregate (1.20) across all investors, we will obtain the market portfolio z^M

$$\begin{aligned} z^M &= (\sum_i \frac{1}{\alpha_i})\Omega^{-1}(\mu - R_F p). \\ &= \frac{1}{\alpha_M}\Omega^{-1}(\mu - R_F p) \end{aligned} \tag{1.21}$$

where in (1.21) we've defined

$$\alpha_M = (\sum_i \frac{1}{\alpha_i})^{-1}.$$

The key insight to determining p is that z_j^M, the market's demand for shares of asset j must, in equilibrium, equal the supply of shares in asset j, which we take to be exogenous. Thus, we can treat z^M as exogenous, and rearrange (1.21) to get

$$p = \frac{1}{R_F}\mu - \frac{\alpha_M}{R_F}\Omega z^M. \tag{1.22}$$

Because it involves the potentially endogenous parameter α_M, a measure of market risk aversion, equation (1.22) does not necessarily constitute a complete solution for equilibrium prices (α_M may itself depend on p). Nevertheless, we can gain the key insights of the CAPM from (1.22).

The equation (1.22) is really N equations stacked in vector form. Let's consider the ith row—the expression for the ith asset price:

$$p_i = \frac{\mu_i}{R_F} - \frac{\alpha_M}{R_F}\sum_{j=1}^{N}\omega_{i,j} z_j^M. \tag{1.23}$$

The first term in (1.23), μ_i / R_F, is just the present value (using the gross riskless rate R_F) of the ith asset's expected payoff, μ_i. The term $\sum_j \omega_{i,j} z_j^M$ is just the covariance between the ith asset's payoff, x_i, and the payoff of the market portfolio, $z^M \cdot x$. For simplicity, call this $\text{cov}(i, M)$. Then,

$$p_i \begin{cases} < \mu_i / R_F & \text{if } \text{cov}(i, M) > 0 \\ = \mu_i / R_F & \text{if } \text{cov}(i, M) = 0 \\ > \mu_i / R_F & \text{if } \text{cov}(i, M) < 0 \end{cases}$$

An asset whose payoff covaries positively with that of the market portfolio is priced at less than the expected present value of its payoff, while an asset whose payoff covaries negatively with that of the market portfolio is priced at more than the expected present value of its payoff—analogous to what we saw in the very simple examples at the start of the this lecture. The middle case—$\text{cov}(i, M) = 0$—corresponds to an asset whose payoff risk is purely *idiosyncratic*, in the sense of being orthogonal to the market as whole. Under the CAPM, the price of such an asset, *no matter how large its variance*, is just equal to its expected present value—which would be the value placed on the asset by a completely risk neutral investor. To summarize:

Result 1.1 (The CAPM, price version)**.** *Under the assumptions of this section, an asset's equilibrium price is less than or greater than the expected present value of the asset's payoff, depending on whether the asset's payoffs covary postively or negatively with the payoffs of the market portfolio. Idiosyncratic risk is not priced.*

Remark 1.6. *The results for individual asset prices also hold for any arbitrary portfolio z of the assets. The price of the portfolio, $p \cdot z$, is less than, greater than, or equal to the present value of the portfolio's expected payoff, $(z \cdot \mu)/R_F$, depending on whether the covariance of the portfolio's payoff with the market portfolio payoff, $z^\top \Omega z^M$, is positive, negative, or zero.*

Because this must apply as well to the market portfolio, and because $(z^M)^\top \Omega z^M > 0$, the price of the market portfolio is necessarily less than the present value of its expected payoff.

1.3.1 The CAPM in return form

So far, we've worked with asset prices and payoffs, but the CAPM (as well as the other results) is more frequently cast in terms of asset *returns*. We've already introduced the riskless rate of return R_F and related it to the price q of a riskless asset. We now consider returns on the risky assets.

The gross return on the ith risky asset is defined as x_i/p_i, which we'll denote by R_i. We'll denote by R the $N \times 1$ vector of gross returns. An asset's return is basically it's payoff per unit of account invested in the asset. Just as x is random variable, so is R. The asset's expected return is $\mathbb{E}(R_i) = \mu_i/p_i \equiv R_i^e$. We can express this in matrix notation, for all N assets at once,by letting C be an $N \times N$ diagonal matrix with $(1/p_1, 1/p_2, \ldots 1/p_N)$ on the diagonal. Then,

$$R^e = \mathbb{E}[R] = \mathbb{E}[Cx] = C\mu$$

The variance-covariance matrix of returns—call it V—is then given by

$$\begin{aligned}
V &= \mathbb{E}[(R - R^e)(R - R^e)^\top] \\
&= \mathbb{E}[(Cx - C\mu)(Cx - C\mu)^\top] \\
&= C\mathbb{E}[(x - \mu)(x - \mu)^\top]C^T \\
&= C\Omega C^\top
\end{aligned}$$

Note too that as a diagonal matrix C is symmetric; that C^{-1} is also a diagonal matrix, with $(p_1, p_2, \ldots p_N)$ on the diagonal; and that $Cp = (1, 1, \ldots 1)^\top$, an $N \times 1$ vector of ones, which we will denote by $\mathbb{1}$. Having a linear map between payoffs and returns (for a given p) will allow us to easily toggle back and forth between results obtained in the 'payoffs/prices' framework and corresponding results in the 'returns' framework.

When working in terms of returns, a portfolio takes the form of a vector θ of *portfolio weights*, rather than a vector z of asset demands. The portfolio weights are fractions of the investor's initial wealth allocated to each asset. The θ_i's and z_i's are related by $\theta_i = p_i z_i / p \cdot z = p_i z_i / W_0$. Using the transformation matrix C,

$$\theta = \frac{1}{W_0} C^{-1} z$$

Letting $\theta_0 = q z_0 / W_0$, the fraction of initial wealth allocated to the riskless asset, the investor's budget constraint (1.14) becomes a constraint requiring the portfolio weights to sum to 1:

$$1 = \theta_0 + \sum_{i=1}^{N} \theta_i = \theta_0 + \theta \cdot \mathbb{1} \tag{1.24}$$

The return on the investor's portfolio of risky assets is $\theta \cdot R$, and the return on his whole portfolio of assets is $\theta_0 R_F + \theta \cdot R$. We can show that the investor's final wealth is given by

$$W_1 = (\theta_0 R_F + \theta \cdot R) W_0.$$

This follows from (1.15) and the equivalences $z_0 = (1/q) q z_0 = R_F \theta_0 W_0$ and

$$z^\top x = z^\top C^{-1} C x = (C^{-1} z)^\top (Cx) = W_0 \theta \cdot R.$$

Expected final wealth is just initial wealth times the expected portfolio return,

$$\begin{aligned} \mathbb{E}(W_1) &= (\theta_0 R_F + \theta \cdot R^e) W_0 \\ &= R_F W_0 + \theta (R^e - R_F \mathbb{1}) W_0 \end{aligned} \tag{1.25}$$

where (1.25) uses the fact that $\theta_0 = 1 - \theta \cdot \mathbb{1}$. The last line expresses the expected portfolio return in terms of the riskless return R_F and the portfolio-weighted *excess returns* on the risky assets, $R_i^e - R_F$.

The variance of final wealth is just $(W_0)^2$ times the variance of the portfolio of risky returns:

$$\begin{aligned} \mathbb{E}[(W_1 - \mathbb{E}(W_1))^2] &= (W_0)^2 \mathbb{E}[(\theta \cdot R - \theta \cdot R^e)^2] \\ &= (W_0)^2 \theta^\top \mathbb{E}[(R - R^e)(R - R^e)^\top] \theta \\ &= (W_0)^2 \theta^\top V \theta \end{aligned} \tag{1.26}$$

At this point, we could plug (1.25) and (1.26) into the investor's utility function, maximize with respect to θ, and retrace all the steps that led to our equilibrium price expression (1.22). Rather than take that long route to the return version of the CAPM, we'll begin with (1.20) and apply the transformation matrix C. Recall that (1.20) was an expression describing an investor's portfolio of risky assets:

$$z^i = \frac{1}{\alpha_i}\Omega^{-1}(\mu - R_F p)$$

Pre-multiply both sides by C^{-1} to get

$$\begin{aligned} C^{-1}z^i &= \frac{1}{\alpha_i}C^{-1}\Omega^{-1}(\mu - R_F p) \\ &= \frac{1}{\alpha_i}C^{-1}\Omega^{-1}C^{-1}C(\mu - R_F p) \\ &= \frac{1}{\alpha_i}C^{-1}\Omega^{-1}C^{-1}(C\mu - R_F C p) \end{aligned}$$

We can now use $C^{-1}z^i = W_0^i\theta^i$, $C\mu = R^e$, $Cp = \mathbb{1}$, and $V = C\Omega C^\top$—which, together with the symmetry of C, implies $C^{-1}\Omega^{-1}C^{-1} = V^{-1}$—to arrive at risky asset demands in terms of portfolio weights:

$$\theta^i W_0^i = \frac{1}{\alpha_i}V^{-1}(R^e - R_F\mathbb{1})$$

We now sum over all investors. Let $W_0^M = \sum_i W_0^i$ and

$$\theta^M = \sum_i (\frac{W_0^i}{W_0^M})\theta^i,$$

i.e., θ^M is an initial-wealth-weighted average of all investors' portfolios. Again let $1/\alpha_M = \sum_i(1/\alpha_i)$. We then obtain the following analogue to (1.21):

$$\theta^M W_0^M = \frac{1}{\alpha_M}V^{-1}(R^e - R_F\mathbb{1}).$$

Rearranging gives

$$R^e - R_F\mathbb{1} = \alpha_M W_0^M V\theta^M. \tag{1.27}$$

As before, we view the market portfolio θ^M (and aggregate market initial wealth W_0^M) as exogenous data, dictated by the supplies of the N risky assets. Note that the ith row of $V\theta^M$ is the covariance of the ith asset return with the return $\theta^M \cdot R$ on the market portfolio. In a slight abuse of our previous notation, let's call this magnitude $\mathrm{cov}(i, M)$. We then obtain the following analogue to Result 1.1:

Result 1.2 (The CAPM, return version)**.** *Under the assumptions of this section, an asset's equilibrium expected excess return is either positive ($R_i^e - R_F > 0$) or negative*

($R_i^e - R_F < 0$), depending on whether the asset's return covaries postively or negatively with the return on the market portfolio. Idiosyncratic risk is not priced—an asset whose return has a zero covariance with the return on the market portfolio has a zero expected excess return.

Early in the history of the CAPM, practitioners viewed the market risk parameter α_M as the only unobservable in (1.27); later, doubts were raised about whether the market portfolio was even observable. In any event, a version of the CAPM that eliminated α_M was seen as desirable. The result—the 'beta' form of the CAPM—eliminated α_M at the expense of making the CAPM a purely relative theory of equilibrium returns. This is the form in which the CAPM is most commonly expressed.

To arrive at it, note that (1.27) has implications for the excess returns on portfolios, including the market portfolio. The excess return on the market portfolio is $(\theta^M)^\top(R^e - R_F\mathbb{1})$. Pre-multiply both sides of (1.27) by $(\theta^M)^\top$ to get

$$(\theta^M)^\top(R^e - R_F\mathbb{1}) = \alpha_M W_0^M (\theta^M)^\top V \theta^M. \tag{1.28}$$

$(\theta^M)^\top V \theta^M$ is simply the variance of the return on the market portfolio, which we denote by $\mathrm{var}(M)$. Equation (1.28) then gives

$$\alpha_M W_0^M = \frac{1}{\mathrm{var}(M)}(\theta^M)^\top(R^e - R_F\mathbb{1}) \tag{1.29}$$

Now, plug (1.29) into (1.27) to get the following expression for the equilibrium expected excess return on the ith asset:

$$\begin{aligned} R_i^e - R_F &= \frac{\mathrm{cov}(i, M)}{\mathrm{var}(M)}(\theta^M)^\top(R^e - R_F\mathbb{1}) \\ &\equiv \beta_i(\theta^M)^\top(R^e - R_F\mathbb{1}) \end{aligned} \tag{1.30}$$

Result 1.3 (The CAPM, beta version)**.** *Under the assumptions of this section, the equilibrium expected excess return on any asset i depends only on the expected excess return on the market portfolio and the asset's beta,* $\beta_i = \mathrm{cov}(i, M)/\mathrm{var}(M)$*. In particular, the only source of systematic variation in the asset's expected excess return is variation in the expected excess return on the market portfolio, and the only source of variation in expected excess returns* across assets *is variation in their betas.*

Remark 1.7. *Analogous to our point in Remark 1.6, note that from (1.28), since* $\mathrm{var}(M) > 0$*, the market portfolio earns a positive expected excess return.*

Exercize 1.3 (Calculating betas)**.** *For this exercize, using the same data from Exercize 1.2, but interpret the μ from that exercize as R^e and the Ω as V. In Exercize 1.2 you calculated 50 efficient portfolios; assume the 25th portfolio which you obtained there is the market portfolio θ^M here. Write some Matlab code to calculate the betas for the 5 assets. Report the betas and turn in the code you wrote.*

Lecture 2

Some basic theory: Stochastic discount factors, state prices, and risk neutral probabilities

The common structure of almost all the models we'll look at throughout the course is summarized in the relation

$$p = \mathbb{E}(mx) \tag{2.1}$$

where p is an asset's price, x it's payoff (a random variable), and m is a random variable known alternatively as a *stochastic discount factor* or *pricing kernel*.[1] The existence of a stochastic discount factor is related to the absence of arbitrage opportunities—loosely, there is an m such that the above pricing equation holds for all assets if and only if there is no arbitrage. As we'll eventually see, economic models—describing agents' preferences, endowments, trading opportunities, *etc.*—will typically give rise to SDFs, and the asset-pricing implications of the different models can be usefully characterized by the different m's they imply.

SDFs are also closely linked to the concept of *risk-neutral probabilities*. Suppose there is a risk-free asset with price q and payoff 1 in all states, let $R_F = 1/q$ denote the gross risk-free rate of return. Risk-neutral probabilities—or a risk-neutral measure—are probabilities such that when we take expectations using those probabilities—call the expectation $\hat{\mathbb{E}}(\,\cdot\,)$—we get

$$p = \frac{1}{R_F}\hat{\mathbb{E}}(x) \tag{2.2}$$

That is, every asset is priced according to the present value of its expected payoff using the risk-neutral measure.

[1] In finance, m is often referred to as a *state-price density*.

We're going to take an indirect (but standard) route to equations (2.1) and (2.2), one that begins with the concept of state prices—loosely, a set of prices (distinct from asset prices) that value a marginal unit of wealth (or consumption or unit of account) in each possible state of the world. A key result—called the Fundamental Theorem of Asset Pricing—is going to link absence of arbitrage with the existence of state prices. A further result—the so-called 'Representation Theorem'—is going prove a type of equivalence between state prices, risk neutral probabilities, and stochastic discount factors.

The material in this lecture loosely follows the presentation in Duffie [Duf92]. Another good reference is John Cochrane's book [Coc01]. Dybvig and Ross's entry on "Arbitrage" in *The New Palgrave: Finance* [DR89] and the first chapter of Ross's book [Ros04] are also good sources for some of this material, especially the two theorems.

2.1 Basic structure of the two-period model

2.1.1 Assets and payoffs

As before, there are two periods, zero and one, or 'today' and 'tomorrow'. Decisions are taken today, and uncertainty resolves (and payoffs realized) tomorrow.

Tomorrow's uncertainty is *discrete*: there are S possible states of the world that may be realized, which we'll index by $s \in \{1, 2, \ldots S\} \equiv \mathcal{S}$. State s occurs with probability $\pi_s > 0$, and π denotes the vector of probabilities.

There are N risky assets, indexed by $i = 1, 2, \ldots N$. Assets pay off in units of account or consumption, depending on the context. Asset i's payoff in state s is denoted $x_{i,s}$, and X is the $N \times S$ matrix describing payoffs for all assets in all states. Thus, each row of X corresponds to an asset, and each column to a state. The expected payoff of asset i is $\sum_s x_{i,s}\pi_s \equiv \mathbb{E}(x_i)$, and in matrix notation

$$\mathbb{E}(X) \equiv X\pi. \tag{2.3}$$

A portfolio is just a vector $z \in \mathbb{R}^N$. There are no restrictions on short sales, unless otherwise noted. A portfolio z pays off $\sum_i z_i x_{i,s}$ in state s, so—in matrix notation—$z^\top X \in \mathbb{R}^S$ describes the portfolio's payoffs in every possible state. We may also write the portfolio payoff as $X^\top z$ when it is necessary to represent it as a column vector.

Having defined states, assets, payoffs, and portfolios, we're in a position to define two concepts that relate to them—*asset redundancy* and *asset market completeness*. A redundant asset is one whose payoffs can be replicated by a portfolio of other assets, while asset market completeness means any vector in $\mathbb{R}^S$ can be realized by some portfolio.

Definition 2.1 (Redundant assets)**.** *Asset i is* redundant *if there exists a portfolio z such that $z_i = 0$ and, $\forall s$, $x_{i,s} = \sum_j z_j x_{j,s}$. This is equivalent to saying there is a linear dependency in the rows of X.*

Definition 2.2 (Complete asset markets)**.** *Asset markets are complete if for any $y \in \mathbb{R}^S$ there is a portfolio $z \in \mathbb{R}^N$ with $X^\top z = y$. Since linear combinations of portfolios are again portfolios, we can also state market completeness as: there exist S portfolios, z^i, $i = 1, 2, \dots S$, such that their payoff vectors $\{X^\top z^i : i = 1, \dots S\}$ form a basis for $\mathbb{R}^S$.*

We'll often assume there are no redundant assets. We may or may not assume markets are complete. The following easy exercize is to help make sure you understand the concepts of redundancy and completeness.

Exercize 2.1. *Suppose there are two assets and three states. The payoff matrix is*

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \end{bmatrix}$$

Show that asset markets are incomplete, the simple way, by showing a $y \in \mathbb{R}^3$ that can't be attained by any portfolio $z \in \mathbb{R}^2$. Now, imagine adding a third asset, and consider two cases. First, add an asset that would be redundant, given the first two. Make sure to demonstrate that it is redundant. Second, add a third asset that completes the market. Demonstrate that it completes the market by showing portfolios z^1, z^2 and z^3 that attain the basis vectors $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.

2.1.2 Investors

There will be H investors, indexed by $h = 1, 2, \dots H$. Investors (potentially) differ in their preferences, initial wealth, and/or endowments. Investor h's preferences, in their most general form, will be described by a utility function $U^h : \mathbb{R} \times \mathbb{R}^S \to \mathbb{R}$. The arguments are either consumption values or units of account, as the context dictates.

In the most general case, we'll write $U^h(c_0, c_1)$ as shorthand for

$$U^h(c_0, c_{1,1}, \dots, c_{1,s}, \dots c_{1,S}).$$

A familiar special form that U^h might take is the time-separable, discounted expected utility form (which is separable across both dates and states):

$$U^h(c_0, c_1) = u_h(c_0) + \beta \sum_{s=1}^{S} \pi_s u_h(c_{1,s}). \tag{2.4}$$

In some cases, we may assume investors only care about final wealth or consumption: $U^h(c_1)$. In any case—whether preferences are over (c_0, c_1) or just c_1—we assume that investors' utilities are continuous (or at least upper semicontinuous). We'll also assume investors' utilities are strictly increasing in all their arguments—they prefer more to less.

Investors will have initial wealth W_0^h, and may receive a stochastic endowment $e^h = (e_1^h, e_2^h, \dots e_S^h)$ in period one. A feasible portfolio for investor h will obey the budget constraint

$$W_0^h \geq p \cdot z. \tag{2.5}$$

If they receive an endowment in period one, then the utility from their choice of portfolio will be

$$U^h(W_0^h - p \cdot z, e^h + X^\top z). \tag{2.6}$$

where (2.6) makes use of the fact that, when investor's prefer more to less, (2.5) holds as an equality.

Note that the budget set

$$B^h = \{(c_0, c_1) \in \mathbb{R}_+ \times \mathbb{R}_+^S : c_0 = W_0^h - p \cdot z, c_1 = e^h + X^\top z, \text{ some } z \in \mathbb{R}^N\} \tag{2.7}$$

is clearly closed and convex. If it is bounded—which we'll see has to do with the presence or absence of arbitrage—it is also compact, and therefore an optimal choice of z will exist.

When there's no chance of confusion—or a representative agent—we'll drop the superscript h to avoid unnecessary clutter.

2.1.3 Prices & arbitrage

We distinguish two sorts of prices—*asset prices* and *state prices*. The former are the usual prices investors face in the market: they specify the number of units of account (or consumption) at date zero an investor has to pay for a unit (or share) of each asset—which is a claim to the payoffs described above. The latter—usually just a theoretical construct—specify the value in units of account (or consumption) at date zero of an additional unit of account (or consumption) in each of the S possible states at date one. All our models will have asset prices, of course; state prices will exist under certain conditions, which we'll make clear shortly.

Investors take asset prices as given. The price of asset i is p_i and $p \in \mathbb{R}^N$ is the vector of asset prices. The price of a portfolio z is then just $p \cdot z$.

The state prices (when they exist) will be denoted by ψ_s for $s = 1, 2, \dots S$. The state price vector is $\psi \in \mathbb{R}^S$. If state prices exist, they can be used to price the payoffs of any of the assets. For example, asset i's payoffs in each state $(x_{i,1}, x_{i,2}, \dots, x_{i,S})$ form a vector in $\mathbb{R}^S$, and their value at state prices ψ is $\sum_s \psi_s x_{i,s}$. State prices, depending on the context, are sometimes referred to as *Arrow-Debreu* prices.

Since a unit of asset i is equivalent to a set of claims to i's payoffs in each of the S possible date-one states, when a state price vector exists, we say that it *prices assets correctly* if:

$$p_i = \sum_{s=1}^{S} \psi_s x_{i,s}. \tag{2.8}$$

holds for all $i = 1, 2, \dots N$.

Remark 2.1. *The existence of state prices is separate from the concept of asset market completeness. We can have a state price vector that tells us the value of a claim to one unit of account in state s (and only state s) for each s, even when a set of S assets with those payoffs—which would constitute a basis for $\mathbb{R}^S$—does not exist. Whether or not markets are complete* does *matter for the uniqueness of a state price vector.*

An *arbitrage*, or arbitrage portfolio, is—in words—a portfolio that either (a) costs nothing today, has a positive payoff in at least some states tomorrow, and no chance of a negative payoff in any state, or (b) costs less than nothing today (it has a negative price), and has a nonnegative payoff in every state tomorrow. Formally:

Definition 2.3 (Arbitrage)**.** *z is an* arbitrage *if either of the following two conditions hold:*

$$p \cdot z \leq 0 \text{ and } X^\top z > 0 \tag{2.9}$$

or

$$p \cdot z < 0 \text{ and } X^\top z \geq 0 \tag{2.10}$$

We say there is no arbitrage *if no such z exists.*

Remark 2.2. *The two conditions in Definition 2.3 can be combined into one if we stack $-p^\top$ (which is $1 \times N$) and $X^\top$ (which is $S \times N$) into a single $(S+1) \times N$ matrix. Then,* (2.9) *and* (2.10) *can be stated as: an arbitrage is a portfolio z such that*

$$\begin{bmatrix} -p^\top \\ X^\top \end{bmatrix} z > 0 \tag{2.11}$$

where the matrix product on the left is $S + 1 \times 1$, and you'll recall that for vectors $w > 0$ means all $w_i \geq 0$ and at least some $w_i > 0$. There is no arbitrage if no such z exists.

Since if z is a portfolio, so is αz for any real number α, an arbitrage (if one exists) can be run at any scale. Note that the presence or absence of arbitrage is a feature of asset prices and payoffs together.

Example 2.1. *Suppose asset i is redundant: there is a portfolio $z \in \mathbb{R}^N$ with $z_i = 0$ and $(x_{i,1}, x_{i,2}, \ldots x_{i,S}) = z^\top X$. If $p_i \neq p \cdot z$, there is an arbitrage opportunity. To see this, suppose $p_i > p \cdot z$, and consider the portfolio $z - e^i$, where e^i is the ith unit basis vector in $\mathbb{R}^N$. This portfolio corresponds to shorting one unit of asset i and buying the replicating portfolio z. Then,*

$$\begin{aligned} X^\top(z - e^i) &= X^\top z - X^\top e^i \\ &= X^\top z - \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,S} \end{bmatrix} \\ &= 0 \end{aligned}$$

so the payoffs in all states tomorrow are all zero. But, the cost of the portfolio is $p \cdot (z - e^i) = p \cdot z - p_i < 0$*, which therefore presents an arbitrage opportunity. For the case of* $p_i < p \cdot z$*, just consider the portolio* $-z + e^i$.

As the next section shows, arbitrage, asset prices and the existence state prices are intimately linked by a basic result in asset pricing. Once we have state prices, it will be easy to construct stochastic discount factors and risk neutral probabilities.

Exercize 2.2. *Suppose that there are two assets* $(N = 2)$ *and three states* $(S = 3)$*. The payoff matrix is*

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \end{bmatrix}$$

Characterize the set of prices vectors $p = (p_1, p_2)$ *that are consistent with there being no arbitrage opportunities. Show all the steps in your reasoning. Hint: There is one inequality that* p_1 *and* p_2 *need to obey for there to be no arbitrage.*

2.2 The Fundamental Theorem of Asset Pricing

This result is sometimes known as the Fundamental Theorem of Finance.[2] Loosely, it states an equivalence between the absence of arbitrage, the existence of a strictly positive state price vector, and the existence of an optimal portfolio choice for an investor who prefers more to less. Formally, in the context of our model:

Theorem 2.1 (The Fundamental Theorem of Asset-Pricing)**.** *The following are equivalent:*

1. *The asset price vector* p *and payoff matrix* X *satisfy no arbitrage—there is no portfolio* z *satisfying* (2.11).
2. *There exists a strictly positive state price vector* ψ *that correctly prices assets—that is, there is a* $\psi \gg 0$ *satisfying*

$$p = X\psi, \tag{2.12}$$

   the matrix version of (2.8).
3. *There exists a finite optimal portfolio choice for a (hypothetical) investor who prefers more to less.*

[2] See chapter one of Ross [Ros04], available from Princeton University Press here: `http://press.princeton.edu/chapters/s7834.pdf`. Other useful references are the beginning of Duffie's book [Duf92] or Dybvig and Ross's New Palgrave entry on "Arbitrage" [DR89].

PROOF. Condition (3) of the theorem clearly implies condition (1)—if some investor who prefers more to less has an optimal choice, there must be no arbitrage. If we can show that (1) implies (2), and (2) implies (3), we'll be done.

So, let's begin with (1) implies (2). Assume then that (1) holds: there is no $z \in \mathbb{R}^N$ satisfying (2.11). We'll use a separating hyperplane argument to contstruct a ψ that satisfies (2). Let's first define what we mean by a hyperplane in $\mathbb{R}^n$:

Definition 2.4 (Hyperplanes). *A* hyperplane *in $\mathbb{R}^n$ is a set of the form $H(p, \alpha) = \{x \in \mathbb{R}^n : p \cdot x = \alpha\}$, for a given $p \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Associated with any hyperplane are two closed half-spaces, $H(p, \alpha)^- = \{x \in \mathbb{R}^n : p \cdot x \leq \alpha\}$ and $H(p, \alpha)^+ = \{x \in \mathbb{R}^n : p \cdot x \geq \alpha\}$. The open half-spaces $H(p, \alpha)^{--}$ and $H(p, \alpha)^{++}$ are defined similarly, but with strict inequalities. Two sets $A, B \subset \mathbb{R}^n$ are said to be* separated by a hyperplane *$H(p, \alpha)$ if one of the sets is contained wholly in $H(p, \alpha)^-$ and the other set is contained wholly in $H(p, \alpha)^+$. The sets are said to be strongly separated if one of them is actually wholly contained in one of $H(p, a)$'s open half-spaces. A hyperplane $H(p, \alpha)$ is said to* support *a set A at a point $y \in A$ if $y \in H(p, \alpha)$ and A is wholly contained in either $H(p, \alpha)^-$ or $H(p, \alpha)^+$.*

Keeping that definition in mind, and referring back to equation (2.11), let

$$M = \begin{bmatrix} -p^\top \\ X^\top \end{bmatrix}$$

a $S + 1 \times N$ matrix. For any portfolio $z \in \mathbb{R}^N$, Mz is just a vector in $\mathbb{R}^{S+1}$, and the set

$$K = \{y \in \mathbb{R}^{S+1} : y = Mz \text{ for } z \in \mathbb{R}^N\}$$

is a subset of R^{S+1}. In fact it is a special kind of subset: K is a closed linear subspace, and is hence also a closed, convex cone.

Definition 2.5 (Linear subspaces and cones). *A set $A \subset \mathbb{R}^n$ is a* linear subspace *if $x, y \in A$ and $a, b \in \mathbb{R}$ imply $ax + by \in A$. A linear subspace A is a closed linear subspace if it's also a closed set (in the usual sense in $\mathbb{R}^n$). A set $B \subset \mathbb{R}^n$ is a cone if $x \in B$ and $a \in \mathbb{R}$, $a > 0$, imply $ax \in B$. A* cone *B is a closed, convex cone if it's also a closed, convex set. Clearly, any closed linear subspace of $\mathbb{R}^n$ is also a closed convex cone in $\mathbb{R}^n$.*

The no arbitrage condition is precisely the statement that the only point of intersection between K and the nonnegative orthant $\mathbb{R}^{S+1}_+$ is the origin—$K \cap \mathbb{R}^{S+1}_+ = \{0\}$. We are ready for the separation argument. We'll use (without proving) the following result, due to Samuel Karlin [Kar59]:[3]

Result 2.1. *Let V be a closed convex cone in $\mathbb{R}^n$ intersecting the nonnegative orthant only at the origin. Then, there is $\phi \in \mathbb{R}^n$, $\phi \gg 0$, such that $\phi \cdot y \leq 0$ for all $y \in V$.*

[3] See Theorem B.3.5 on page 404 of [Kar59]. A lot of the book is available for free preview on Google Books; see here: `http://books.google.com/books?id=qgMfct0YnFQC`

Essentially, Karlin's result is saying that there is a hyperplane $H(\phi, \alpha)$, with $\phi \gg 0$, $\alpha = 0$, that separates the cone V from the nonnegative orthant.

In our case, using Karlin's result, if there is no arbitrage, there is a strictly positive $\phi \in \mathbb{R}^{S+1}$ satisfying $\phi \cdot y \leq 0$ for all $y \in K = \{y \in \mathbb{R}^{S+1} : y = Mz \text{ for } z \in \mathbb{R}^N\}$. Since our K is a linear subspace, $y \in K$ implies $-y \in K$. This means ϕ must in fact satisfy $\phi \cdot y = 0$ for all $y \in K$, since $\phi \cdot y < 0$ for any $y \in K$ would mean $\phi \cdot (-y) > 0$, contradicting the fact that, since $-y \in K$, it should obey $\phi \cdot (-y) \leq 0$.

What does $\phi \cdot y = 0\,(\forall y \in K)$ mean? Using the definition of K, we must have, $\forall z \in \mathbb{R}^N$,

$$\begin{aligned} 0 &= \phi \cdot (Mz) \\ &= (\phi^\top M)z \end{aligned}$$

which is only possible—if you think about it, since it must hold for *every* choice of z—if $\phi^\top M = 0$.

Since $\phi \in \mathbb{R}^{S+1}$, we may abuse notation slightly and write it as $(\phi_0, \phi_1, \phi_2, \ldots \phi_S)$. Then, using the definition of M, and the rules for products of transposes,

$$0 = \begin{bmatrix} -p & X \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \vdots \\ \phi_S \end{bmatrix} = -p\phi_0 + X \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_S \end{bmatrix}$$

Now, let $\psi \equiv (1/\phi_0)(\phi_1, \phi_2, \ldots \phi_S)$—which is feasible since $\phi_0 > 0$—and rearrange to obtain

$$p = X\psi$$

which completes the proof that no arbitrage (condition 1 of the Theorem) implies the existence of a state price vector (condition 2).

Now, for (2) implies (3). This part is easy: we just construct a hypothetical investor whose marginal utilities of consumption in each of the S states tomorrow are given by the state price vector $\psi = (\psi_1, \ldots \psi_S)$. For example, define

$$\begin{aligned} U(W_0 - p \cdot z, e + X^\top z) &\equiv W_0 - p \cdot z + \psi^\top (e + X^\top z) \\ &= W_0 - p \cdot z + \psi \cdot e + \psi^\top X^\top z \\ &= W_0 - p \cdot z + \psi \cdot e + p \cdot z \\ &= W_0 + \psi \cdot e \end{aligned}$$

which clearly has a finite optimal portfolio choice—$z = 0$ would work—since the investor is indifferent amongst all choices of z. □

The theorem does not guarantee uniqueness of the state price vector. In fact, if asset markets are incomplete, it's easy to see how non-uniqueness may arise. When markets are incomplete, there are fewer than S linearly independent asset—*i.e.*, X has less than S linearly independent rows, so $p = X\psi$ represents fewer than S independent equations in the S variables $(\psi_1, \psi_2, \ldots \psi_S)$.

Suppose, for example, that the N assets have linearly independent payoff vectors (no redundant assets), but that $N < S$—fewer assets than states. Then, the rank of X is N, and the set of solutions to the linear system $p = X\eta$, that is, the set $\{\eta \in \mathbb{R}^S : X\eta = p\}$, has dimension $N - S > 0$.[4]

Exercize 2.3. *Using the same X from exercize 2.2, and assuming p is any p satisfying the condition you derived in exercize 2.2, characterize the set of possible state price vectors ψ satisfying $p = X\psi$.*

Showing that asset market incompleteness leads to non-uniqueness of the state price vector is not quite the same as showing that complete markets guarantee uniqueness. We'll do that now:

Result 2.2. *Suppose there is a strictly positive state price vector ψ which correctly prices assets ($p = X\psi$). If asset markets are complete, then ψ is the unique state price vector.*

PROOF. An easy way to show this is to note that when markets are complete, for each unit basis vector e^s in $\mathbb{R}^S$ there is a portfolio z^s such that $X^\top z^s = e^s$. Since ψ correctly prices assets, we must have

$$p \cdot z^s = \psi \cdot e^s = \psi_s \tag{2.13}$$

for all $s = 1, 2, \ldots S$. Since these equations must hold for any state price vector that correctly prices assets, state prices are uniquely determined.[5] □

2.3 The Representation Theorem: Going from state prices to SDFs and risk neutral probabilities

If we were working in infinite-dimensional spaces, the proof of this result would require some serious mathematical machinery. With $\mathcal{S}$ being finite-dimensional, though, it involves little more than some algebra. Like Theorem 2.1, it is an equivalence result. In this case existence of a state price vector is linked to the existence of a stochastic discount factor—see equation (2.1)—and the existence of risk-neutral probabilities (2.2).

[4] See any linear algebra text, for example Shilov [Shi77].

[5] Note that by no arbitrage—which, by Theorem 2.1, must hold if there is a $\psi \gg 0$ that correctly prices assets—if $\hat{z}^s$ also gives $X^\top \hat{z}^s = e^s$, then $p \cdot \hat{z}^s = p \cdot z^s$, so the left-hand side of (2.13) is a unique value for each s.

Before we can state it, though, it will be useful to have some notation that represents, for $v, w \in \mathbb{R}^S$, the vector $(v_1 w_1, v_2 w_2, \dots v_S w_S) \in \mathbb{R}^S$. Let vw—with no '$\cdot$', no $\top$—denote this vector. Recall that $\pi = (\pi_1, \pi_2 \dots \pi_S) \gg 0$ are probabilities over the S possible states. Then,

$$\mathbb{E}(vw) = \sum_{s=1}^{S} \pi_s v_s w_s.$$

Now, suppose we have a positive state price vector ψ that correctly prices assets. Let $m_s = \psi_s / \pi_s > 0$, which is feasible since all states have positive probability. Then, for any asset i,

$$\begin{aligned} p_i &= \sum_{s=1}^{S} x_{i,s} \psi_s && (2.14) \\ &= \sum_{s=1}^{S} \pi_s (\psi_s / \pi_s) x_{i,s} && (2.15) \\ &= \sum_{s=1}^{S} \pi_s m_s x_{i,s} && (2.16) \\ &= \mathbb{E}(m x_i) && (2.17) \end{aligned}$$

In the last line, x_i denotes the ith row of X. The equation shows that there is an $m \in \mathbb{R}^S$, $m \gg 0$, such that $p_i = \mathbb{E}(m x_i)$ for any asset i. By the linearity of expectations, this is also the case for any portfolio z of assets: $p \cdot z = \mathbb{E}[m(X^\top z)]$.

Conversely, suppose there is an $m \gg 0$ such that $p_i = \mathbb{E}(m x_i)$ holds for any asset i, and that all states have positive probability. Then ψ defined by $\psi_s = \pi_s m_s$ $(\forall s \in \mathcal{S})$ is a strictly positive state price vector that correctly prices assets. This is the first part of the representation theorem—there is a state price vector $\psi \gg 0$ that correctly prices assets if and only if there is a random variable $m \gg 0$ such that $p_i = \mathbb{E}(m x_i)$ for every asset i.

For the next part of the result—which relates to risk-neutral probabilities—we need another bit of notation, and a risk-free asset. Since we'll be considering expectations taken with respect to other probabilities (the risk-neutral probabilities that we'll be constructing), we need some notation for expectations that indicates the probabilities we're using. So, for any vector $y \in \mathbb{R}^S$ and any probabilities[6] $\phi \in \mathbb{R}^S$, let $\mathbb{E}_\phi$ denote the expectation of y with respect to ϕ:

$$\mathbb{E}_\phi(y) = \sum_{s=1}^{S} \phi_s y_s$$

With that notation, we can write expectations with respect to the true, or objective, probabilities π as $\mathbb{E}_\pi(\,\cdot\,)$. We may also write $\mathbb{E}(\,\cdot\,)$ for expectations with respect to π when there is no chance for confusion.

[6] $\phi \in \mathbb{R}^S$ are probabilities if $\phi \geq 0$ and $\sum_s \phi_s = 1$.

This part of the result also assumes that we have a risk-free asset. Let q denote the risk-free asset's price and $\mathbb{1} = (1, 1, \ldots 1)$, a row of ones, its payoff vector. I want to keep this vector distinct from the payoffs in X, but at the same time, statements of the form 'prices assets correctly' or 'for all assets' should be understood as applying not just to p and X, but to

$$\begin{bmatrix} q \\ p \end{bmatrix}$$

and

$$\begin{bmatrix} \mathbb{1} \\ X \end{bmatrix}$$

Now, suppose there is a stochastic discount factor $m \gg 0$ that prices assets according to $p_i = \mathbb{E}_\pi(m x_i)$ and $q = \mathbb{E}_\pi(m\mathbb{1})$. Note that $\mathbb{E}_\pi(m\mathbb{1}) = \mathbb{E}_\pi(m)$, so

$$\frac{1}{\mathbb{E}_\pi(m)} = \frac{1}{q} = R_F$$

the gross risk-free rate of return, as we defined it in Lecture 1. Note, too, that the vector

$$\phi \equiv \frac{1}{\mathbb{E}_\pi(m)} \pi m = \frac{1}{\mathbb{E}_\pi(m)} (\pi_1 m_1, \pi_2 m_2, \ldots \pi_S m_S)$$

obeys $\phi \geq 0$ and $\sum_s \phi_s = 1$. Thus, ϕ is a probability measure on $\mathcal{S}$. Now, divide both sides of $p_i = \mathbb{E}(m x_i)$ by $q = \mathbb{E}(m)$, and use the above considerations to obtain

$$\begin{aligned}
\frac{1}{q} p_i &= \frac{1}{\mathbb{E}(m)} \mathbb{E}(m x_i) \\
&= \mathbb{E}(\frac{1}{\mathbb{E}(m)} m x_i) \\
&= \sum_s \frac{\pi_s m_s}{\sum_r \pi_r m_r} x_{i,s} \\
&= \sum_s \phi_s x_{i,s} \\
&= \mathbb{E}_\phi(x_i)
\end{aligned}$$

Using, $q = 1/R_F$, we may write this more suggestively as saying that for any asset i,

$$p_i = \frac{1}{R_F} \mathbb{E}_\phi(x_i)$$

—that is, i's price is just the discounted present value of its expected payoff under the probabilities ϕ.

Conversely, suppose there are probabilities ϕ such that the last equation holds for all assets i. We can go in the other direction to derive a stochastic discount factor m, by defining $m_s = q\phi_s/\pi_s$. The steps in the transformation are

$$\begin{aligned} p_i &= q\sum_s \phi_s x_{i,s} \\ &= \sum_s \pi_s (\frac{q\phi_s}{\pi_s}) x_{i,s} \\ &= \sum_s \pi_s m_s x_{i,s} \\ &= \mathbb{E}_\pi(m x_i) \end{aligned}$$

In sum, we've shown:

Theorem 2.2 (Representation theorem)**.** *Suppose all the states $s = 1, 2, \dots S$ have positive probability. Then the following are equivalent (and, by Theorem 2.1, equivalent to no arbitrage):*

1. *There is a strictly positive state price vector ψ that prices assets correctly—i.e.*

$$p = X\psi.$$

2. *There is a strictly positive stochastic discount factor m that prices assets—i.e., for any asset i,*

$$p_i = \mathbb{E}(m x_i).$$

3. *There are strictly positive probabilities ϕ that price assets according to the risk-neutral pricing formula*

$$p_i = \frac{1}{R_F}\mathbb{E}_\phi(x_i).$$

Remark 2.3. *Since the SDF representation (2) holds for any asset—including a riskless asset with payoff vector $\mathbb{1}$—if q denotes the riskless asset's price, then (2) gives us $q = \mathbb{E}(m\mathbb{1}) = \mathbb{E}(m)$. Letting $R_F = 1/q$, we also have $R_F = 1/\mathbb{E}(m)$.*

Remark 2.4. *The SDF representation also has a covariance interpretation. For any asset i, let $\operatorname{cov}(x_i, m)$ denote the covariance between asset i's payoff x_i and the SDF m—i.e., $\operatorname{cov}(x_i, m) = \mathbb{E}(m x_i) - \mathbb{E}(m)\mathbb{E}(x_i)$. With this notation, and using the previous remark, the SDF representation can be written as*

$$\begin{aligned} p_i &= \mathbb{E}(m x_i) \\ &= \mathbb{E}(m)\mathbb{E}(x_i) + \operatorname{cov}(x_i, m) \\ &= \frac{\mathbb{E}(x_i)}{R_F} + \operatorname{cov}(x_i, m) \end{aligned}$$

That is, the price of asset i is the discounted present value of its expected payoff plus the covariance of its payoff with the SDF. Note the analogy to equation (1.23) from Lecture 1.

Exercize 2.4. *Suppose you observe an economy with two states and two assets. The price vector you observe is $p = (1, 1)$, and the asset payoffs are*

$$X = \begin{bmatrix} 1.008 & 1.008 \\ 0.905 & 1.235 \end{bmatrix}$$

so the first asset is riskless and the second risky. The probabilities of the two states are $\pi = (\frac{1}{2}, \frac{1}{2})$. Write a little MATLAB *code to find state prices ψ that correctly price the assets; the stochastic discount factor m; and the risk neutral probabilities ϕ. Turn in your results and your code.*

2.3.1 The representations in return form

The stochastic discount factor and risk-neutral pricing representations have useful versions when we work in terms of asset returns, as opposed to prices and payoffs. Let $R_i = (1/p_i)x_i \in \mathbb{R}^S$ denote the vector of gross returns for asset i; if state s occurs, a unit investment in i returns $R_{i,s}$. We'll continue to write $R_F = 1/q$ for the gross risk-free return.

In terms of returns, the stochastic discount factor representation is easily seen to be:

$$1 = \mathbb{E}(mR_i). \tag{2.18}$$

which follows simply from dividing both sides of (2) in Theorem 2.2 by p_i. This is true for R_F as well, since—as noted above—$R_F = 1/\mathbb{E}(m)$, implying $1 = R_F\mathbb{E}(m) = \mathbb{E}(mR_F)$.

As was the case with the prices and payoffs version, (2.18) can be written in terms of covariance:

$$\begin{aligned} 1 &= \mathbb{E}(mR_i) \\ &= \text{cov}(R_i, m) + \mathbb{E}(R_i)\mathbb{E}(m) \\ &= \text{cov}(R_i, m) + \mathbb{E}(R_i)\frac{1}{R_F} \end{aligned}$$

using $\mathbb{E}(m) = 1/R_F$ in the last line. Rearranging, we get:

$$\mathbb{E}(R_i) - R_F = -R_F\text{cov}(R_i, m) \tag{2.19}$$

That is, the expected excess return on asset i depends negatively on the covariance between the return on asset i and the stochastic discount factor.

Thinking about equation (2.19), suppose there is an asset whose return, call it R_m, is perfectly correlated with m. To keep things simple, suppose $R_m = \lambda m$

for some real scalar λ. Applying (2.19) to that asset's return gives

$$\begin{aligned}\mathbb{E}(R_m) - R_F &= -R_F \mathrm{cov}(R_m, m)\\ &= -\frac{R_F}{\lambda}\mathrm{cov}(R_m, R_m)\\ &= -\frac{R_F}{\lambda}\mathrm{var}(R_m) \qquad (2.20)\end{aligned}$$

Combining (2.19) and (2.20)—and using $\mathrm{cov}(R_i, m) = (1/\lambda)\mathrm{cov}(R_i, R_m)$—we obtain the *beta representation*

$$\mathbb{E}(R_i) - R_F = \beta_i(\mathbb{E}(R_m) - R_F) \qquad (2.21)$$

where $\beta_i = \mathrm{cov}(R_i, R_m)/\mathrm{var}(R_m)$. This is exactly anologous to the CAPM formula (1.30), with R_m in the role of the market portfolio.

The transformation of the representation in terms of risk-neutral probabilities is even simpler: divide both sides of condition (3) in Theorem 2.2 by p_i, and multiply both sides by R_F to get:

$$\mathbb{E}_\phi(R_i) = R_F \qquad (2.22)$$

—that is, the expected return on any asset, under the risk-neutral probability measure, is simply the riskless rate of return. Note that (2.22) implies that we can write i's return R_i (that is, the random variable, not the expectation) as

$$R_i = R_F + v_i$$

where v_i is a random variable that has a zero expectation under the probabilities ϕ.

The utility of either the stochastic discount factor representation or the risk-neutral probability representation is that, once we know m or ϕ in a given economy (assuming they are unique), we can price all assets by these very simple formulæ.

2.4 Investor utilities and the pricing representations

All our results thus far have been in an essentially preference-free context, apart from assuming investors prefer more to less. Now, we'll relate our pricing representations to investors' utility functions and optimal choices.

Recall that the typical investor's utility function is given by $U^h(c_0, c_1) : \mathbb{R} \times \mathbb{R}^S \to \mathbb{R}$, and the investor has the budget set

$$B^h = \{(c_0, c_1) \in \mathbb{R}_+ \times \mathbb{R}^S_+ : c_0 = W^h_0 - p \cdot z, c_1 = e^h + X^\top z, \text{ some } z \in \mathbb{R}^N\}$$

Assume that U^h is increasing and continuous. B^h is obviously convex and closed. If there is no arbitrage, it must also be bounded, hence compact.[7] Therefore, a solution to the problem $\max\{U^h(c_0, c_1) : (c_0, c_1) \in B^h\}$ must exist.

Our first result will relate the investor's vector of marginal utilities at his optimal choice to state prices. Assume U^h is also concave and differentiable. Denote the partial derivatives of U^h by

$$U_0^h(c_0, c_1) = \frac{\partial}{\partial c_0} U^h(c_0, c_1)$$

and

$$U_{1,s}^h(c_0, c_1) = \frac{\partial}{\partial c_{1,s}} U^h(c_0, c_1).$$

For compactness, let $U_1^h(c_0, c_1)$ denote the vector of S partial derivatives

$$(U_{1,1}^h(c_0, c_1), \ldots U_{1,S}^h(c_0, c_1)).$$

Our first result is:

Result 2.3. *Let $(c_0^*, c_1^*) \in B^h$, and $(c_0^*, c_1^*) \gg 0$. Then, (c_0^*, c_1^*) is an optimal choice if and only*

$$\frac{1}{U_0^h(c_0^*, c_1^*)} U_1^h(c_0^*, c_1^*)$$

is a state price vector.

This follows simply from the first-order conditions to the investor's utility maximization problem. With concave preferences and a convex feasible set, those conditions are necessary and sufficient for an interior optimum. Substitute the budget constraint into the utility function—as in (2.6)—and write the problem as

$$\max_z U^h(W_0^h - p \cdot z, e^h + X^\top z)$$

Differentiate with respect to the ith asset, z_i, to obtain

$$-U_0^h(c_0, c_1)p_i + \sum_{s=1}^{S} U_{1,s}^h(c_0, c_1)x_{i,s} = 0.$$

If a feasible (c_0^*, c_1^*) satisfies this condition for all $i = 1, 2, \ldots N$, it is an optimal choice; conversely a $(c_0^*, c_1^*) \gg 0$ that is an optimal choice will satisfy this condition for every i. But this means

$$p_i = \frac{1}{U_0^h(c_0^*, c_1^*)} \sum_{s=1}^{S} U_{1,s}^h(c_0^*, c_1^*)x_{i,s}$$

[7]This is easiest to see indirectly, by invoking Theorem 2.1, and using the state price vector ψ. Since $p = X\psi$, $p \cdot z = \psi^\top X^\top z$. Thus, $c_0 = W_0 - \psi^\top X^\top z$ and $c_1 = e + X^\top z$. It's clear that if $c_0 \to +\infty$, then we must have $\psi^\top X^\top z \to -\infty$. Since $\psi \gg 0$, this means some component of $X^\top z$ must go to $-\infty$. But then, some component of c_1 would eventually become negative. Conversely, if some $c_{1,s} \to +\infty$, some element of $X^\top z$ must diverge to $+\infty$. But, then $\psi^\top X^\top z$ would go to $+\infty$, eventually making c_0 negative.

or in vector form (treating U_1^h as an $S \times 1$ column vector)

$$p = \frac{1}{U_0^h(c_0^*, c_1^*)} X U_1^h(c_0^*, c_1^*).$$

Remark 2.5. *Under complete markets, any state price vector is unique. In fact, when markets are complete (and there is no arbitrage), we can write the investor's problem equivalently, with no reference to asset choices at all, as*

$$\max_{c_0, c_1} \{U^h(c_0, c_1) : W_0 + \psi \cdot e = c_0 + \psi \cdot c_1\}$$

where ψ is the unique state price vector.

Exercize 2.5. *Prove that last remark.*

Exercize 2.6 (Reinterpreting the CAPM)**.** *Doing this exercize with many investors would present a number of complications that would obscure the point, so suppose there is a single representative investor. His utility function over (c_0, c_1) is*

$$U(c_0, c_1) = c_0 + \sum_{s=1}^{S} \pi_s (a c_{1,s} - \frac{b}{2}(c_{1,s})^2).$$

In equilibrium he must hold an exogenous supply of assets, so you can treat his period-one consumption vector (in equilibrium) as some exogenous c_1^*—i.e., equilibrium prices are such that choosing c_1^* is optimal. Derive an expression for the state price vector, and use it to show that, for any asset i

$$p_i = \frac{\mathbb{E}(x_i)}{R_F} - b\,\mathrm{cov}(c_1^*, x_i),$$

where $\mathrm{cov}(c_1^*, x_i) = \mathbb{E}(c_1^* x_i) - \mathbb{E}(c_1^*)\mathbb{E}(x_i)$. *Hint: Also use the state price vector to price the riskless asset with price q and payoff $\mathbb{1}$.*

For the remainder of the discussion, let's specialize preferences to the time-separable, expected utility form,

$$U^h(c_0, c_1) = u_h(c_0) + \beta \sum_{s=1}^{S} \pi_s u_h(c_{1,s}).$$

Then,

$$\frac{1}{U_0^h(c_0^*, c_1^*)} U_1^h(c_0^*, c_1^*) = (\beta \frac{\pi_1 u_h'(c_{1,1}^*)}{u_h'(c_0^*)}, \beta \frac{\pi_2 u_h'(c_{1,2}^*)}{u_h'(c_0^*)}, \ldots \beta \frac{\pi_S u_h'(c_{1,S}^*)}{u_h'(c_0^*)}) \qquad (2.23)$$

is the form of the state price vector associated with the optimal choice (c_0^*, c_1^*).

With (2.23) as the form of the state price vector, for any asset i we have

$$\begin{aligned} p_i &= \sum_{s=1}^{S} \beta \frac{\pi_s u_h'(c_{1,s}^*)}{u_h'(c_0^*)} x_{i,s} \\ &= \sum_{s=1}^{S} \pi_s m_s x_{i,s} \\ &= \mathbb{E}(m x_i) \end{aligned}$$

where

$$m = (\beta \frac{u_h'(c_{1,1}^*)}{u_h'(c_0^*)}, \beta \frac{u_h'(c_{1,2}^*)}{u_h'(c_0^*)}, \dots \beta \frac{u_h'(c_{1,S}^*)}{u_h'(c_0^*)}) \tag{2.24}$$

is the form taken by the stochastic discount factor under time-separable expected utility. In a slight abuse of notation, let $\beta u_h'(c_1^*)/u_h'(c_0^*)$ stand for the whole vector of marginal rates of substitution in (2.24). We then get the familiar form

$$p_i = \mathbb{E}\left(\beta \frac{u_h'(c_1^*)}{u_h'(c_0^*)} x_i\right).$$

And, in terms of excess returns and covariances, we have

$$\mathbb{E}(R_i) - R_F = -R_F \text{cov}(R_i, m) = -R_F \text{cov}(R_i, \beta \frac{u_h'(c_1^*)}{u_h'(c_0^*)}) \tag{2.25}$$

Equations such as the last two will play a prominent role in the next lecture, when we get to Lucas's model and, after that, the equity premium puzzle.

Next, we can derive the representation of pricing under risk neutral probabilities when we take into account investor preferences. As above, we can go from a stochastic discount factor to risk neutral probabilities using

$$\begin{aligned} \sum_s \pi_s m_s x_{i,s} &= \mathbb{E}(m) \sum_s \frac{\pi_s m_s}{\mathbb{E}(m)} x_{i,s} \\ &= \frac{1}{R_F} \sum_s \phi_s x_{i,s} \\ &= \frac{1}{R_F} \mathbb{E}_\phi(x_i) \end{aligned}$$

where $\phi = \pi m / \mathbb{E}(m)$, and we've used $\mathbb{E}(m) = 1/R_F$. The risk neutral probabilities under expected utility then have the form

$$\phi = \frac{1}{\mathbb{E}(\beta u_h'(c_1^*)/u_h'(c_0^*))} (\pi_1 \beta \frac{u_h'(c_{1,1}^*)}{u_h'(c_0^*)}, \pi_2 \beta \frac{u_h'(c_{1,2}^*)}{u_h'(c_0^*)}, \dots \pi_S \beta \frac{u_h'(c_{1,S}^*)}{u_h'(c_0^*)}) \tag{2.26}$$

Finally, note that the gross risk-free rate, under expected utility, obeys

$$R_F = \frac{1}{\mathbb{E}(\beta u_h'(c_1^*)/u_h'(c_0^*))} \tag{2.27}$$

Lecture 3

Some notes on Lucas (1978), Mehra-Prescott (1985), and the Equity Premium Puzzle

In this Lecture, we move beyond two periods to consider asset prices in economies with an infinite time horizon. Lucas's 1978 model [Luc78] is the seminal piece of work in this vein—almost all developments since can be thought of as adding particular bells and whistles to Lucas's simple framework.

Mehra and Prescott [MP85]—and the 'equity premium puzzle' they uncovered using Lucas's framework—provided much of the impetus for adding bells and whistles to the model.

The longer time horizon creates a few wrinkles in our approach, compared to the two-period framework in which we studied the basic theory.

For one, whereas in our two-period model asset payoffs and returns and the stochastic discount factor were random variables (and prices were just values to be determined at the initial date), in these models (and all that we'll look at after them), payoffs, returns, prices, and stochastic discount factors will all be stochastic processes.

What happens to representations like $p = \mathbb{E}[mx] = (1/R_F)\mathbb{E}_\phi[x]$? We're going to take it for granted—without going through all the mathematical niceties—that these translate in a straightforward way. The stochastic discount factor representation, for example, will become

$$p_t = \mathbb{E}_t[m_{t+1}x_{t+1}],$$

where $\mathbb{E}_t[\,\cdot\,]$ denotes expectation conditional on information available at date t. We'll discuss how risk-neutral pricing gets modified below.

Finally, we'll now need to take account of long-lived assets. Equity, for example, will be an asset that pays some dividend d (usually in units of consumption) at each date. The payoff at $t+1$ to holding a share of equity, though, is more than the dividend at $t+1$, since—after the dividend is paid—one still owns the equity share, which could be sold at the date $t+1$ price: the payoff becomes $x_{t+1} = d_{t+1} + p_{t+1}$. The return on equity is then

$$R_{t+1} = \frac{d_{t+1} + p_{t+1}}{p_t},$$

and the stochastic discount factor representation for p_t becomes

$$p_t = \mathbb{E}_t[m_{t+1}(d_{t+1} + p_{t+1})]. \tag{3.1}$$

Note that the last expression gives us a form of discounted present value. To see it, update (3.1) one period (to give an expression for p_{t+1}), then substitute this into the right-hand side of (3.1):

$$\begin{aligned} p_t &= \mathbb{E}_t[m_{t+1}(d_{t+1} + p_{t+1})] \\ &= \mathbb{E}_t[m_{t+1}d_{t+1}] + \mathbb{E}_t[m_{t+1}p_{t+1}] \\ &= \mathbb{E}_t[m_{t+1}d_{t+1}] + \mathbb{E}_t[m_{t+1}\mathbb{E}_{t+1}[m_{t+2}(d_{t+2} + p_{t+2})]] \\ &= \mathbb{E}_t[m_{t+1}d_{t+1}] + \mathbb{E}_t[m_{t+1}m_{t+2}d_{t+2}] + \mathbb{E}_t[m_{t+1}m_{t+2}p_{t+2}] \end{aligned}$$

In arriving at the last line, we used the Law of Iterated Expectations, one statement of which is $\mathbb{E}_t[\mathbb{E}_{t+1}[y_{t+2}]] = \mathbb{E}_t[y_{t+2}]$.[1] In any case, you can probably see the pattern that's developing. Let

$$\rho_{t,t+k} = m_{t+1}m_{t+2}m_{t+3}\cdots m_{t+k}\ (k = 1, 2, \ldots).$$

Then, assuming

$$\lim_{k\to\infty} \mathbb{E}_t[\rho_{t,t+k}p_{t+k} = 0]$$

we obtain

$$p_t = \sum_{k=1}^{\infty} \mathbb{E}_t[\rho_{t,t+k}d_{t+k}]. \tag{3.2}$$

3.1 Lucas (1978) "Asset prices in an exchange economy"

In classical general equilibrium theory (see, for example, [McK02]), an *exchange economy*, sometimes called an *endowment economy*, is one in which there is no production—agents trade from initial endowments, given prices, and equilibrium prices must be such that all markets clear.

Lucas's model is likewise one without production—there are productive assets that yield dividends (in the form of a consumption good) each period, and those dividends are an exogenous stochastic process. Aggregate output

[1] A more general statement is: If $\mathbb{E}[x|I]$ denotes expectation conditional on an information set I, then $\mathbb{E}[\mathbb{E}[x|I']\,|I] = \mathbb{E}[x|I]$ for $I \subset I'$. Over time, information necessarily accrues: $I_t \subset I_{t+1}$. Another way of stating it is: forecast errors are unforecastable—$\mathbb{E}[x - \mathbb{E}[x|I_{t+1}]\,|I_t] = 0$

(equal to aggregate consumption, because the good is nonstorable) is thus exogenous. Agents (a mass of identical agents, or one representative agent) trade claims to the assets, and equilibrium prices must be such that shares of all assets are held, agents are maximizing their utility, and the economy's resource constraint is satisfied.

3.1.1 Some historical context

We've talked about asset market completeness, but not about what, in finance, is known as *market efficiency*. This is distinct from economists' standard notions of efficiency (the Pareto criterion), and has to do with the extent to which asset prices reflect available information. The efficient markets hypothesis (EMH) as formulated by Fama [Fam65, Fam70] and others requires that, at a minimum, future asset prices cannot be forecast from past prices (the 'weak form' of the EMH). Stronger forms require that price changes not be forecastable using any publicly available information. Typical time series tests—of which there were a lot done in the 1970s—focussed on whether stock returns were predictable or whether stock prices followed random walks. Campbell, Lo and MacKinlay [CLM96] offer some intuition for these tests: Suppose

$$p_t = \mathbb{E}_t[F^*],$$

where F^* is the asset's fundamental value (which must come out of a model of price formation). If the expectation is conditioned on all publicly available information at t, and information accrues over time, then the expectation of p_{t+1} must obey

$$\begin{aligned} \mathbb{E}_t[p_{t+1}] &= \mathbb{E}_t[\mathbb{E}_{t+1}[F^*]] \\ &= \mathbb{E}_t[F^*] \\ &= p_t \end{aligned}$$

The second line uses the Law of Iterated Expectations. Take away the expectation, and one gets

$$p_{t+1} = p_t + u_{t+1} \text{ where } \mathbb{E}_t[u_{t+1}] = 0$$

—which is the definition of a random walk. A random walk is a particular example of a class of stochastic processes known as *martingales*, hence Lucas's reference to this sort of process in his introductory paragraphs.[2]

So, one point of Lucas's paper (in addition to the paper's methodological contributions) is to write down a perfectly good asset pricing model, in which information is used efficiently, and which nonetheless may give rise to equilibrium prices that don't follow random walks (or expected returns that move in predictable ways).

[2] A martingale obeys $\mathbb{E}[p_{t+1}|p_t, p_{t-1}, p_{t-2}\ldots] = p_t$.

Around the same time as Lucas's paper, Breeden [Bre79] formulated the closely related Consumption-Based CAPM.[3] In that model, an asset's excess return can be written in single beta form as

$$R_i^e - R^F = \beta_i^C [R_C^e - R^F],$$

where R_C^e is the expected return on an asset whose return equals the consumption growth rate (so R_C^e equals expected consumption growth), and β_i^C is asset i's beta with respect to asset C—i.e., $\beta_i^C = \text{cov}(R_i, R_C)/\text{var}(R_C)$.

Something very similar emerges from Lucas's model.

3.1.2 An economy with 'trees'

There are n assets—think of them as trees—indexed by $i = 1, 2, \ldots n$. Asset i produces y_{it} at date t (units of a nonstorable consumption good—*i.e.*, dividends, or 'fruit'). $y_t \equiv (y_{1t}, y_{2t}, \ldots y_{nt})$ follows a vector Markov process with a conditional distribution $F(y', y)$:

$$F(y', y) = \Pr\{y_{t+1} \leq y' | y_t = y\}.$$

Aggregate output is the sum of all the fruit produced, and this must equal aggregate consumption, since the good is nonstorable:

$$\sum_{i=1}^{n} y_{it} = c_t.$$

This is the economy's resource constraint.

Agents in the economy trade shares or claims to the trees: z_i shares of asset i at date t are a claim to $z_i y_{it}$ units of the asset's output. There is one perfectly divisible share of each asset in net supply, so another constraint the economy needs to obey in equilibrium is

$$z = (1, 1, \ldots 1) = \mathbb{1}.$$

3.1.3 The representative agent

You can think of the agents in Lucas's economy as a unit mass of identical agents or, more simply, one representative agent. The agent has additively separable expected-utility preferences over consumption. He seeks to maximize

$$\mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta^t U(c_t) \right]$$

subject to a sequence of budget constraints.

[3]The paper is available here: http://www.dougbreeden.net/uploads/Breeden_1979_JFE_Consumption_CAPM_Theory.pdf.

The agent's portfolio at the start of period t is $z_t = (z_{1t}, z_{2t}, \dots z_{nt})$, describing his holdings of shares of the n assets. This gives him resources at the start of t equal to

$$\sum_{i=1}^{n} z_{it}(y_{it} + p_{it}) = z_t \cdot (y_t + p_t)$$

where p_{it} is the *ex dividend* price of asset i (its price immediately after it pays its dividend).

The agent spends his resources on consumption (c_t) and asset holdings for next period $(p_t \cdot z_{t+1})$. Thus, he faces the sequence of budget constraints

$$z_t \cdot (y_t + p_t) \geq c_t + p_t \cdot z_{t+1}$$

Note that since there are no adjustment costs or transactions costs, the agent is indifferent between, say, holding an asset in both t and $t+1$ and selling it at the start of t just to buy it again to hold until $t+1$. So, the budget constraint is written as if the agent sells his whole portfolio at the start of the period, then uses that value (plus his dividend payments) to finance consumption and a new portfolio.

We must assume the agent starts with an initial portfolio z_0 at the start of $t = 0$.

3.1.4 Recursive formulation

Saying what it means for an agent to take prices as given in a stochastic environment requires some clarification. The formulation Lucas uses is a *recursive* one, and his competitive equilibrium concept is a *recursive competitive equilibrium*.

The agent takes as given the law of motion for the aggregate state (y) and a set of n *price functions* that map realizations of the state into realizations of prices:

$$p(y) = (p_1(y), p_2(y), \dots p_n(y)).$$

The agent's resources at each date depend on both the aggregate state y and the agent's individual state (in the case z). Given the nature of the agent's preferences, and the constraint he faces, his problem is recursive: it obeys Bellman's *principle of optimality*. If $v(z, y)$ is the maximized value of the agent's lifetime utility beginning from the state (z, y), then v obeys a Bellman equation:

$$v(z,y) = \max_{c,x} \left\{ U(c) + \beta \mathbb{E}\left[v(x, y')|y\right] : z \cdot (y + p(y)) \geq c + p(y) \cdot x \right\} \tag{3.3}$$

The optimal choices of c and x (next-period's portfolio) can be expressed as functions of the state (z, y): $c(z, y)$ and $x(z, y) = (x_1(z, y) \dots x_n(z, y))$. These are often called the agent's 'decision rules', 'policy functions' or 'optimal policies'.

My definition of equilibrium differs slightly from Lucas's, because I want to make the agent's choices a bit more explicit. A *recursive competitive equilibrium* consists of several objects: A value function $v(z, y)$, decision rules $c(z, y)$ and $x(z, y)$, and price functions $p(y)$ such that:

1. (Agent optimality) The decision rules solve the agent's maximization problem—they attain $v(z, y)$—given the price functions $p(y)$ and the law of motion for the aggregate state.

2. (Market-clearing) All shares are held in equilibrium—$x(\mathbb{1}, y) = \mathbb{1}$ at all y—and the economy's resource constraint is satisfied:

$$c(\mathbb{1}, y) = \sum_{i=1}^{n} y_i.$$

Remark 3.1. *Note that a version of Walras's Law holds here—from the budget constraint, starting from $z_0 = \mathbb{1}$, if all shares are held (so the portfolio x is $\mathbb{1}$), then $c = \mathbb{1} \cdot y$. And, conversely, if $c = \mathbb{1} \cdot y$ holds, and $n - 1$ of the assets are fully held, then so is the nth, again assuming the agent's $z_0 = \mathbb{1}$.*

3.1.5 Characterizing equilibrium

Characterizing equilibrium is simple in the Lucas environment, because equilibrium aggregate consumption must equal the exogenous output $\sum_i y_i$. Our procedure will be to derive the first-order conditions for agent optimality, plug in the exogenous consumption process, and see what those conditions imply for the equilibrium price functions $p(y)$.

Example 3.1. *As an example, imagine a static two-good exchange economy with many identical agents. All agents have the same utility function $u(c) = u(c_1, c_2)$ and the same endowment $e = (e_1, e_2)$. They trade at price vector $p = (p_1, p_2)$. Since everyone is identical, in equilibrium, everyone must decide that it's optimal to simply consume their own endowment. From this we can deduce the equilibrium price vector must satisfy*

$$(p_1, p_2) = \alpha \left(\frac{\partial U}{\partial c_1}(e), \frac{\partial U}{\partial c_2}(e) \right)$$

for some constant α.

To see the first-order conditions that Lucas derives simply, substitute c out of the utility function, using the budget constraint. This gives the agent's problem as:

$$v(z, y) = \max_x \left\{ U(z \cdot (y + p(y)) - p(y) \cdot x) + \beta \mathbb{E} \left[v(x, y') | y \right] \right\}$$

Differentiating the objective with respect to x_i yields

$$U'(c) p_i(y) = \beta \mathbb{E} \left[v_i(x, y') | y \right] \tag{3.4}$$

where v_i is the partial derivative of the value function with respect to the ith asset holding, and c is optimal consumption at (z, y). Applying a standard 'envelope' argument to $v(z, y)$ tells us that

$$v_i(z, y) = U'(c)(y_i + p_i(y)). \tag{3.5}$$

Advancing the expression in (3.5) by one period and plugging it into (3.4) gives

$$U'(c)p_i(y) = \beta\mathbb{E}\left[U'(c')(y_i' + p_i(y'))|y\right] \text{ for } i = 1, 2, \dots n \tag{3.6}$$

where we use the shorthand c' for optimal consumption at (x, y').

Equation (3.6)—a type of Euler equation—is a necessary condition for an optimal choice by the agent. Now, to derive an equilibrium asset pricing formula, we simply impose the resource constraint $c = \sum_i y_i$:

$$U'(\sum_i y_i)p_i(y) = \beta\mathbb{E}\left[U'(\sum_i y_i')(y_i' + p_i(y'))|y\right] \text{ for } i = 1, 2, \dots n \tag{3.7}$$

We can rewrite (3.7) to give

$$p_i(y) = \mathbb{E}\left[\beta\frac{U'(\sum_i y_i')}{U'(\sum_i y_i)}(y_i' + p_i(y'))|y\right] \text{ for } i = 1, 2, \dots n \tag{3.8}$$

or, stacking the n equations more compactly in vector form

$$p(y) = \mathbb{E}\left[\beta\frac{U'(\sum_i y_i')}{U'(\sum_i y_i)}(y' + p(y'))|y\right]. \tag{3.9}$$

As Lucas emphasizes, (3.8) or (3.9) can be viewed as a set of n *functional equations*—that is, mappings whose arguments are functions. In particular, imagine using an arbitrary function $f_0(\cdot) : \mathbb{R}^n_+ \to \mathbb{R}^n_+$ in place of $p(\cdot)$ on the right-hand side of (3.9). Performing the operations specified on the right-hand side of (3.9) then defines a function from $\mathbb{R}^n_+$ to $\mathbb{R}^n_+$, call it $f_1(\cdot)$. Its value at any $y \in \mathbb{R}^n_+$ is simply:

$$f_1(y) = \mathbb{E}\left[\beta\frac{U'(\sum_i y_i')}{U'(\sum_i y_i)}(y' + f_0(y'))|y\right]. \tag{3.10}$$

Equation (3.10) describes a mapping that takes a function as its argument and creates another function. We can write (3.10) more suggestively as

$$f_1 = T(f_0)$$

where T is our mapping defined on the space of functions from $\mathbb{R}^n_+$ to $\mathbb{R}^n_+$, taking values in the space of functions from $\mathbb{R}^n_+$ to $\mathbb{R}^n_+$. The equilibrium price function is a fixed point of this mapping: equation (3.9) is just saying

$$p = T(p).$$

Lucas—as always—is careful about specifying the conditions under which the mapping that I'm calling T has a fixed point.

3.1.6 The SDF implied by Lucas's model

Looking carefully at equation (3.10) or (3.9), we can see immediately what the stochastic discount factor must be—it is

$$m(y, y') = \beta \frac{U'(\sum_i y_i')}{U'(\sum_i y_i)}. \tag{3.11}$$

The stochastic discount factor applied to next-period payoffs depends on both the current and next-period state. Note that in general the return on an asset from the current period to next period also depends on both the current and next-period state:

$$R_i(y, y') = \frac{y_i' + p_i(y')}{p_i(y)}.$$

Returns then must satisfy the pricing relationship

$$\mathbb{E}\left[m(y, y')R_i(y, y')|y\right] \text{ for } i = 1, 2, \ldots n$$

An exception would arise if we introduced a one-period riskless bond into Lucas's environment (which we will do in Mehra and Prescott's model). It's price q and return R^F depend only on the current state: $q(y) = \mathbb{E}[m(y, y')|y]$, which only depends on y,[4] and $R^F(y) = 1/q(y)$.

Returning to the time series notation we began with—and using $c_t = \sum_i y_{it}$—we can write m in the more familiar form

$$m_{t+1} = \beta \frac{U'(c_{t+1})}{U'(c_t)}. \tag{3.12}$$

Asset prices obey

$$p_{it} = \mathbb{E}_t[m_{t+1}(y_{i,t+1} + p_{i,t+1})] = \mathbb{E}_t\left[\beta \frac{U'(c_{t+1})}{U'(c_t)}(y_{i,t+1} + p_{i,t+1})\right] \tag{3.13}$$

And if $R_{i,t+1}$ is the gross return on some asset between t and $t+1$—that is, $R_{i,t+1} = (y_{i,t+1} + p_{i,t+1})/p_{i,t}$—then

$$1 = \mathbb{E}_t[m_{t+1}R_{i,t+1}] = \mathbb{E}_t\left[\beta \frac{U'(c_{t+1})}{U'(c_t)} R_{i,t+1}\right] \tag{3.14}$$

Note that by the Law of Iterated Expectations, we can replace the conditional expecations in the last expression with unconditional expectations.[5]

Doing that, we can use the last expression—together with $\mathbb{E}[xy] = \text{cov}(x, y) + \mathbb{E}[x]\,\mathbb{E}[y]$—to derive a suggestive expression for expected asset returns:

$$\mathbb{E}[R_{i,t+1}] - \frac{1}{\mathbb{E}[m_{t+1}]} = -\frac{1}{\mathbb{E}[m_{t+1}]}\text{cov}\left(\beta \frac{U'(c_{t+1})}{U'(c_t)}, R_{i,t+1}\right)$$

[4] The y' gets 'integrated out' in taking the expectation.

[5] If $\mathbb{E}_t[m_{t+1}R_{i,t+1} - 1] = 0$ when we're conditioning on information that's accumulated through date t, then the unconditional expectation—when we know nothing at all—is surely zero as well.

3.1.7 Taking Lucas to the computer

There are a number of ways to implement Lucas's model computationally, but the simplest by far is to just replace Lucas's general Markov process $F(y', y)$ with a *Markov chain*. A Markov chain is a set of S states $(x_1, x_2, \dots x_S)$—which could simply be $(1, 2, \dots S)$—and a *transition matrix* $P = [P_{ij}]$, where

$$P_{ij} = \Pr\left\{x_{t+1} = x_j | x_t = x_i\right\}.$$

Since each row is a probability in $\mathbb{R}^S$, $P \geq 0$ and $\sum_j P_{ij} = 1$.

The invariant, or long-run, distribution of the Markov chain is a probability $\pi \in \mathbb{R}^S$ that satisfies $P^\top \pi = \pi$. We'll talk about calibrating them momentarily.

We'll sometimes write $P(i, j)$ for P_{ij} and $\pi(i)$ for π_i. That will help avoid confusion in places, and it also has the advantage of being closer to MATLAB conventions.

For now, let's suppose there is only one tree, so y is a scalar, and there's only one price function $p(y)$ to solve for. We can also solve for the price of a one-period riskless bond, $q(y)$, but that's easy enough to do once we see how to get p. Note that if aggregate output/aggregate consumption is assumed to follow an S-state Markov chain, then:

- y and the price function p will be vectors in $\mathbb{R}^S$. Instead of a function $p(y)$ to solve for, we only need to solve for S numbers, $p = (p(1), p(2), \dots p(S))$. Here, $p(i)$ is the asset price when today's state is i—i.e., when aggregate output is $y(i)$.[6]

- If we price a one-period riskless asset, it's price q and return R^F are also vectors in $\mathbb{R}^S$.

- The stochastic discount factor m and the return to the risky asset R are both $S \times S$ matrices. We'll write $m(i, j)$ for the discount factor applied between state i today and state j tomorrow (and similarly for R, writing $R(i, j)$). In particular,
$$m(i, j) = \beta \frac{U'(y(j))}{U'(y(i))}$$

If P is our transition probability matrix, then the SDF version of the pricing relationship (3.9) becomes

$$p(i) = \sum_{j=1}^{S} P(i, j) m(i, j) \left(y(j) + p(j)\right).$$

Let Ψ denote the $S \times S$ matrix with typical element $\Psi(i, j) = P(i, j)m(i, j)$.[7]

[6] In other words, we may as well identify the states with $(1, 2, \dots S)$, rather than writing $c(y(s))$.
[7] In MATLAB terms, `Psi = P.*m`.

Then, in vector form, we can write the pricing relationship compactly as

$$\begin{bmatrix} p(1) \\ p(2) \\ \vdots \\ p(S) \end{bmatrix} = \Psi \begin{bmatrix} y(1) + p(1) \\ y(2) + p(2) \\ \vdots \\ y(S) + p(S) \end{bmatrix}$$

$$p = \Psi\,(y + p)$$

If $I - \Psi$ is invertible, then we our solution is immediate

$$p = (I - \Psi)^{-1}\Psi y. \tag{3.15}$$

There are a number of other objects of interest one could then calculate, using the solution for p or the expression for the pricing kernel m. Some of them bear on the questions that motivate Lucas's paper—i.e., "Will prices follow random walks?" or "Will expected returns be predictable?" and so forth. The objects of interest one might construct include:

- The price of a riskless one period bond and its return:

$$q(i) = \sum_j P(i,j)m(i,j)$$
$$R^F(i) = \frac{1}{q(i)}$$

- Return on the risky asset, across every state transition:

$$R(i,j) = \frac{p(j) + y(j)}{p(i)}$$

- The expected return conditional on being in state i:

$$\mathbb{E}_i[R] = \sum_j P(i,j)R(i,j)$$

- The expected excess return conditional on being in i, $\mathbb{E}_i[R] - R^F(i)$.

- The (unconditional) expected return on the risky asset and the expected riskless rate:

$$\mathbb{E}[R] = \sum_i \pi(i)\mathbb{E}_i[R]$$
$$\mathbb{E}\left[R^F\right] = \sum_i \pi(i)R^F(i)$$

- The unconditional expected excess return on the risky asset (the 'equity premium'), $\mathbb{E}[R] - \mathbb{E}\left[R^F\right]$.

- Price/dividend ratios in each state, $P/D(i) = \dfrac{p(i)}{y(i)}$.

- Risk neutral probabilities:

$$\phi(i,j) = \frac{P(i,j)m(i,j)}{\sum_h P(i,h)m(i,h)}.$$

- Finally, one could check for mean-reversion in the asset price (which would be inconsistent with price following a random walk), by examining $\mathbb{E}_i[p] - p(i) = \sum_j P(i,j)p(j) - p(i)$. Loosely, the price would be mean reverting if $0 < (\mathbb{E}[p] - p(i))(\mathbb{E}_i[p] - p(i))$, where $\mathbb{E}[p]$ is the unconditional mean $\pi \cdot p$.[8]

An excercize below will ask you to calculate all these things, and describe some of the results. First, though, we need a little digression on approximating autoregressions with Markov chains. We'll focus on chains with $S = 2$, which can be done practically with pencil and paper.[9]

Making a two-state Markov chain that mimics a first-order autoregressive process is fairly simple. Suppose we want to mimic a process of the form

$$x_{t+1} - \mu = \rho\,(x_t - \mu) + \epsilon_{t+1} \tag{3.16}$$

We have estimates of the mean μ, the persistence parameter ρ, and the unconditional standard deviation of x_t, call it σ_x.[10] Let hats denote estimates, and assume that $|\hat{\rho}| < 1$.

A two-state Markov chain that mimics (3.16) will have a low-x state (x_l) and a high-x state (x_h), with $x_l < \hat{\mu} < x_h$. It will also have a 2×2 transition probability matrix P,

$$P = \begin{bmatrix} P_{ll} & P_{lh} \\ P_{hl} & P_{hh} \end{bmatrix}$$

say. Since the rows must sum to one, though, P really has only two parameters we would need to determine—for example, the probabilities of remaining in the low and high states, P_{ll} and P_{hh}. That's four parameters to determine, $(x_l, x_h, P_{ll}, P_{hh})$, with only three restrictions—the unconditional mean, unconditional standard deviation, and persistence—with which to determine them.

[8] This one may actually be immediate just from the fact that p inherits the autoregressive property of y.

[9] There are a couple popular methods for approximating autoregressions with Markov chains of any order. The most commonly used is due to Tauchen [Tau86]. Based on the results in a paper by Kopecky and Suen [KS10], I've lately switched, in my own work, to using Rouwenhorst's method [Rou95]. If you google these, you'll find papers with descriptions of them and probably some MATLAB code for implementing them as well. We may use Rouwenhorst's method in a subsequent lecture.

[10] σ_x is related to the standard deviation of ϵ_t by $\sigma_x^2 = \sigma_\epsilon^2/(1-\rho^2)$.

We have, in a sense, one free parameter, which one may usefully think of as the long-run probability of being in one or the other state—we could match the unconditional mean, standard deviation and persistence, and still have a parameter to play with that would influence the long-run distribution. Adding the long-run distribution as another 'target' of our calibration will pin down the free parameter.

Let π_l denote the long-run probability of being in the low state, and $\pi_h = 1 - \pi_l$ the probability of being in the high state. The symmetry inherent in the process (3.16) suggests setting $\pi_l = \pi_h = 1/2$. Once we make that assumption, as you'll see, we have enough conditions to pin down all the parameters. In particular, the assumption of equal long-run probabilities implies that the transition matrix P is symmetric—and in the 2×2 case, that means that if we can just pin down one entry, we can pin down all four entries.[11]

In any case, given equal long-run probabilities, one then shows that $(x_l, x_h) = (\hat{\mu} - \hat{\sigma}_x, \hat{\mu} + \hat{\sigma}_x)$ satisfies

$$\frac{1}{2}x_l + \frac{1}{2}x_h = \hat{\mu}$$
$$\frac{1}{2}(x_l - \hat{\mu})^2 + \frac{1}{2}(x_h - \hat{\mu})^2 = \hat{\sigma}_x^2$$

That leaves one parameter in P to be determined, say P_{ll}, by trying to match the persistence $\hat{\rho}$. What we want to try to match is basically either

$$\mathbb{E}_l[x - \hat{\mu}] = \hat{\rho}(x_l - \hat{\mu})$$

or

$$\mathbb{E}_h[x - \hat{\mu}] = \hat{\rho}(x_h - \hat{\mu}).$$

Let's examine the former, using $P_{lh} = 1 - P_{ll}$ and the definitions of x_l and x_h:

$$\begin{aligned} P_{ll}(x_l - \hat{\mu}) + (1 - P_{ll})(x_h - \hat{\mu}) &= \hat{\rho}(x_l - \hat{\mu}) \\ P_{ll}(-\hat{\sigma}_x) + (1 - P_{ll})(\hat{\sigma}_x) &= \hat{\rho}(-\hat{\sigma}_x) \\ -P_{ll} + 1 - P_{11} &= -\hat{\rho} \end{aligned}$$

implying

$$P_{ll} = \frac{1 + \hat{\rho}}{2}$$

[11] π satisfies $P^\top \pi = \pi$, while P must by definition obey $P\mathbb{1} = \mathbb{1}$. This implies, in the 2×2 case, a simple relation between π and the off-diagonal elements of P

$$\frac{\pi_l}{\pi_h} = \frac{P_{hl}}{P_{lh}}$$

This can be combined with $\pi \cdot \mathbb{1} = 1$ to give an expression for π in terms of P_{lh} and P_{ll}.

From that, we can fill in

$$
\begin{aligned}
P_{lh} = 1 - P_{ll} &= \frac{1-\hat{\rho}}{2} \\
P_{hl} = P_{lh} &= \frac{1-\hat{\rho}}{2} \\
P_{hh} = 1 - P_{hl} &= \frac{1+\hat{\rho}}{2}
\end{aligned}
$$

All that said, we're now ready for the promised computational exercize related to Lucas's model. We'll break it into two parts. The first is about setting up the Markov chain:

Exercize 3.1. *Using some some U.S. data on annual log consumption of non-durables and services (detrended using a Hodrick-Prescott filter), from 1950 to 2006, I estimated the following $AR(1)$*

$$\log(c_{t+1}) = \rho \log(c_t) + \epsilon_{t+1}$$

I didn't include a constant because the detrended series has mean zero. The estimates were $\hat{\rho} = 0.63775$ and $\hat{\sigma}_\epsilon = 0.0093346$, so an estimate of $\sigma_{\log(c)}$ would be $0.0093346/\sqrt{1 - 0.63775^2}$.

Either in MATLAB *or by hand (with* MATLAB *preferred), construct a 2-state Markov chain to approximate the process for $\log(c_t)$, assuming the long-run distribution is $(1/2, 1/2)$. Note this process has $\mu_{\log(c)} = 0$.*

Now, exponentiate your vector of $\log(c)$ values to get states in terms of c. That is, if `logc` is the 2×1 vector of values the chain can take on, set `c` = `exp(logc)`. Lastly, normalize `c` so it has a long-run mean of 1—if `pi` = `[0.5, 0.5]` is your vector of long-run probabilities, set `c` = `c/(pi * c)`.

Now that you have a Markov chain, you are ready for the main exercize itself:

Exercize 3.2. *Using the Markov chain you constructed in the last problem, write a* MATLAB *program to solve for the vector of asset prices p, and all the 'objects of interest' we listed after we derived equation (3.15). Assume that*

$$U(c) = \frac{c^{1-\alpha}}{1-\alpha}$$

and write a program that calculates the results for given values of α and β.

Report results for $\alpha = 5$ and $\beta = 0.95$. Discuss the extent to which the price/dividend ratio is useful for forecasting the asset return—i.e., what's the relationship between whether the P/D ratio is low or high and whether conditionally expected returns are low or high? Is this true for expected excess returns?

Now, jack up α to 20 and $\beta = 0.99$. Describe the differences between the risk neutral probabilities and the actual probabilities P in this case. Were they that different under the orginal parameters? How does the equity premium compare to the $(\alpha, \beta) = (5, 0.95)$ case? How do you interpret the returns (realized returns by state, not expected returns)?

3.2 Mehra and Prescott (1985) and the Equity Premium Puzzle

If you understood Lucas's model, Mehra-Prescott should be very simple, since they modify Lucas's model in just a couple ways. The meat of the paper is the calibration, and the findings they report in their Figure 4 on page 155.

This paper spawned a very large literature, one that's still very much active today. The paper makes the simple point that the representative agent model with constant relative risk aversion/constant elasticity of intertemporal substitution preferences—which is to say, the workhorse model of business cycle theory at the time, and to some extent even today—is inconsistent with the size of the historical risk premium stocks have commanded relative to (practically) risk-free bonds. There are some assumptions there, of course—for example, that aggregate consumption is a good proxy for the aggregate dividend from holding equity.

The essence of the result is that, historically, aggregate consumption is just 'too smooth'—when filtered through power utility, it yields an intertemporal marginal rate of substitution that doesn't have much volatility, consequently little covariance with the aggregate equity return. Little covariance in turn means a small excess return. Making the agent more risk averse—increasing the curvature of the marginal utility of consumption—can help get bigger excess returns, but only at the cost of lowering the expected $IMRS$—which then raises the model-implied riskless rate well above the historically low values we observe.

Like Lucas, it's an exchange economy, but this helps their cases—they give the model the best chance to succeed, because they don't even require that it generate, endogenously, a realistic consumption process. *They plug in a realistic consumption process*, and the model's price predictions fail.

3.2.1 Differences relative to Lucas's model

There are only a few noteworthy differences in Mehra and Prescott's model, as it compares to Lucas's framework—only one of them of great importance.

1. Mehra and Prescott employ a Markov chain model from the very outset. This is a possibility implicit in Lucas's framework, but since Mehra and Prescott's interest is quantitative, they make it explicit from the beginning.

2. They introduce a one-period riskless bond. This—as we know from the last section—is not a difficult addition. In terms of equilibrium, the assumption is that the bond is in zero net supply—prices must be such that in equilibrium, a representative agent holds zero units of the bond.

3. The one really substantive change is the modification they make to the consumption process. Rather than a Markov chain for the level of aggregate consumption, they assume that aggregate consumption evolves according to
$$y_{t+1} = x_{t+1} y_t$$
where the growth rate x_{t+1} follows a Markov chain. This implies that the log of consumption has a unit root.

The third modification has some implications for the structure of equilibrium prices and the stochastic discount factor. With the power utility function they use, the SDF depends only on the growth rate of consumption:

$$m_{t+1} = \beta \left(x_{t+1}\right)^{-\alpha}.$$

This means that, in writing down the pricing relationships (or putting them on the computer), m only depends on next period's state, not also on today's. It's distribution, and conditional expected value, still depend on today's state through the transition matrix P.[12]

As for the equilibrium equity price, it depends on both today's y and today's x, but it is homogeneous of degree one (in fact linear) in y. This follows from the present value relationship (3.2) we derived way back at the outset of the lecture:

$$p_t = \sum_{k=1}^{\infty} \mathbb{E}_t [\rho_{t,t+k} y_{t+k}].$$

Since each y_{t+k} is proportional to y_t and all the $\rho_{t,t+k}$ terms are products of m_{t+h}'s, p_t must be proportional to y_t. Thus, if $p(y, x)$ denotes the equilibrium price function, then $p(y, x) = yw(x)$ for some function w.

The equilibrium pricing relationship, in terms of the SDF, and not yet specializing to a Markov chain, is

$$p(y, x) = \mathbb{E}\left[m(x') \left(x'y + p(x'y, x')\right) | x\right]$$

[12] I'm sticking with the notation we've developed. Mehra and Prescott use ϕ_{ij} for elements of the transition matrix.

or—dividing out today's y—

$$w(x) = \mathbb{E}\left[m(x')\left(x' + x'w(x')\right)|x\right]. \tag{3.17}$$

The price of the riskless one-period bond is, as usual,

$$q(x) = \mathbb{E}\left[m(x') : x\right]. \tag{3.18}$$

And that's really it—(3.17) and (3.18) are the guts of the formal model. We specify a Markov chain for x—making x, w, m and q into vectors, and giving us a P to calculate the conditional expectations—and we're in a position where we know how to solve that model very simply.

3.2.2 Mehra and Prescott's calibration

Mehra and Prescott construct a two state Markov chain to mimic an estimated $AR(1)$ for annual per capita consumption growth, in particular a long-run mean of 1.018 (1.8% growth rate), standard deviation of 0.036, and autocorrelation coefficient -0.14. They assume the long-run distribution of the two states is $\pi = (1/2, 1/2)$. Using the techniques described above, this then gives

$$x = \begin{bmatrix} 1.018 - 0.036 \\ 1.018 + 0.036 \end{bmatrix} = \begin{bmatrix} 0.982 \\ 1.054 \end{bmatrix}$$

$$P_{11} = P_{22} = \frac{1 - 0.14}{2} = 0.43,$$

and

$$P_{12} = P_{21} = \frac{1 + 0.14}{2} = 0.57.$$

3.2.3 The target and the results

That's the process they feed in, at various combinations of the taste parameters α and β. What do they hope to get out of the model? The measure of success (or failure) is how close the model's average equity return, average risk-free return, and average equity premium—in our notation from above, $\mathbb{E}[R], \mathbb{E}\left[R^F\right]$, and $\mathbb{E}[R] - \mathbb{E}\left[R^F\right]$—come to the historical averages they document:

$$\mathbb{E}[R] = 1.07$$
$$\mathbb{E}\left[R^F\right] = 1.008$$
$$\mathbb{E}[R] - \mathbb{E}\left[R^F\right] = 0.062$$

—*i.e.*, 7%, 0.8%, and 6.2 percentage points.

Mehra and Prescott allow the risk aversion parameter α to vary between zero and ten, and the discount factor β between zero and one.[13] The results

[13]They give some arguments for not taking α above ten. In a subsequent lecture, we'll talk about disentangling risk aversion from intertemporal substitution, and what constitutes a 'plausible' degree of risk aversion.

are summarized in their Figure 4—getting the equity premium as high as 0.35 (which is presumably at $\alpha = 10$) entails pushing the riskless rate up to 4%.

What's going on? Consider the riskless rate first. Increasing α lowers the elasticity of intertemporal substitution (which is just $1/\alpha$). In an economy in which consumption, on average, is growing, a lower elasticity of intertemporal substitution will increase the compensation agents require for deferring consumption from today to tomorrow—in a deterministic economy, that would mean a higher interest rate.[14] There's more at work with regard to R^F than just that deterministic logic, but that mechanism appears to be dominant.[15]

With regard to the equity premium, given the low volatility of consumption, the stochastic discount factor has a small variance, unless we set α very high. The volatility of the discount factor puts a bound on how big a reward the economy can pay an investor for holding a risky asset. The following discussion draws on chapter 5 of Cochrane's book [Coc01]. To see the nature of the bound, consider

$$\mathbb{E}_t\left[m_{t+1}\left(R_{t+1} - R_t^F\right)\right] = 0.$$

By the Law of Iterated Expectations, this must hold unconditionally as well:

$$\mathbb{E}\left[m_{t+1}\left(R_{t+1} - R_t^F\right)\right] = 0.$$

Using $\mathbb{E}[xy] = \text{cov}(x,y) + \mathbb{E}[x]\,\mathbb{E}[y]$, re-write this as

$$\text{cov}(m_{t+1}, R_{t+1} - R_t^F) + \mathbb{E}[m_{t+1}]\,\mathbb{E}\left[R_{t+1} - R_t^F\right] = 0$$

or

$$\mathbb{E}\left[R_{t+1} - R_t^F\right] = -\frac{\text{cov}(m_{t+1}, R_{t+1} - R_t^F)}{\mathbb{E}[m_{t+1}]}.$$

Now, $\text{cov}(m_{t+1}, R_{t+1} - R_t^F) = \text{corr}(m_{t+1}, R_{t+1} - R_t^F)\sigma(m_{t+1})\sigma(R_{t+1} - R_t^F)$, where $\text{corr}(m_{t+1}, R_{t+1} - R_t^F) \in [-1, 1]$ is the correlation between m_{t+1} and the excess return $R_{t+1} - R_t^F$, and the $\sigma(\,\cdot\,)$'s are standard deviations. Thus,

$$\frac{\mathbb{E}\left[R_{t+1} - R_t^F\right]}{\sigma(R_{t+1} - R_t^F)} = -\text{corr}(m_{t+1}, R_{t+1} - R_t^F)\frac{\sigma(m_{t+1})}{\mathbb{E}[m_{t+1}]}.$$

The quantity on the left-hand side of the last expression is called a *Sharpe ratio*, and it measures an asset or portfolio's excess return per unit of volatility. It is, in some sense, a market measure of the price of risk. For the U.S. equity market as a whole, the mean in the numerator is around 0.062 and the standard deviation in the denominator is around 0.166 (from Mehra and Prescott's

[14] Recall that in a deterministic economy, $g \approx \text{EIS}(r - \eta)$,where g is the growth rate, r is the interest rate, and $\eta = 1/\beta - 1$ is the rate of time preference. Flipping this around gives, $r \approx \eta + (1/\text{EIS})g$. If $g > 0$, a lower EIS raises r.

[15] In a subsequent lecture, we'll see if preferences that separate risk aversion from intertemporal substitution can help with this.

calculations, reported in their Table 1). This implies a Sharpe ratio of around 0.37.

It's straightforward to show—since correlation is between -1 and 1—that the last expression implies

$$\frac{|\mathbb{E}\left[R_{t+1}-R_t^F\right]|}{\sigma(R_{t+1}-R_t^F)} \leq \frac{\sigma(m_{t+1})}{\mathbb{E}[m_{t+1}]}. \tag{3.19}$$

If m prices all assets, then the absolute value of the Sharpe ratio for any asset is bounded by the magnitude on the right, which is a measure of the volatility of the stochastic discount factor relative to its mean. For a given mean, the less volatile a model's SDF, the smaller the maximum Sharpe ratio that the model can generate.

When $m_{t+1} = \beta(x_{t+1})^{-\alpha}$, we can say more, especially if we assume x_{t+1} is lognormally distributed—*i.e.*, $\log(x_{t+1}) \sim N(\mu, \sigma^2)$, so $\log\left[(x_{t+1})^{-\alpha}\right] = -\alpha \log(x_{t+1}) \sim N(-\alpha\mu, \alpha^2\sigma^2)$. A property of lognormally distributed random variables is that, for $\log(y) \sim N(\mu_y, \sigma_y^2)$,

$$\sigma\left(y\right) = \mathbb{E}[y]\sqrt{\exp(\sigma_y^2)-1}$$

Thus,

$$\frac{\sigma(m_{t+1})}{\mathbb{E}[m_{t+1}]} = \sqrt{\exp(\alpha^2\sigma^2)-1}.$$

Using the approximation $\sqrt{\exp(\alpha^2\sigma^2)-1} \approx \alpha\sigma$, we can express the bound in (3.19) as approximately:

$$\frac{|\mathbb{E}\left[R_{t+1}-R_t^F\right]|}{\sigma(R_{t+1}-R_t^F)} \leq \alpha\sigma.$$

Under power utility—and assuming lognormality is a good approximation—generating a significant premium for bearing aggregate asset return risk requires either a lot of risk (a big σ) or a lot of aversion to risk (a big α).

Exercize 3.3. *Write a* MATLAB *program to replicate Mehra and Prescott's exercize. Report results for nine parameter combinations—eight given by pairs of* $\beta \in \{0.95, 0.99\}$ *and* $\alpha \in \{1, 5, 10, 20\}$, *plus one with* $(\beta, \alpha) = (1.125, 18)$. *For each combination, report* $100 * (\mathbb{E}[R]-1)$, $100(\mathbb{E}\left[R^F\right]-1)$, $100(\mathbb{E}[R]-\mathbb{E}\left[R^F\right])$ *(all long-run, unconditional expectations), and* $\sigma(m)/\mathbb{E}[m]$, *the ratio of the unconditional standard deviation of the SDF to the unconditional mean of the SDF.*

3.3 Additional dimensions of the equity premium puzzle: Second moment aspects

In solving exercize 3.3, you hopefully discovered that setting $\alpha = 18$ and $\beta = 1.125$ gets you pretty close to the first moments of asset returns that Mehra and Prescott set as the target of their exercize—an average risk-free rate just under 1% and an average equity return of around 7%. With $\alpha = 18$, and the Mehra-Prescott Markov chain for consumption, the maximum allowable Sharpe ratio in the model—given by the bound

$$\frac{\mathbb{E}\left[R - R^F\right]}{\sigma\left[R - R^F\right]} \leq \frac{\sigma(m)}{\mathbb{E}[m]} \tag{3.20}$$

that we derived in section 3.2.3—is actually higher than the average Sharpe ratio we observe in the data (0.56 versus about 0.4).

Are we therefore done with the equity premium puzzle—should we declare success and move on to a different topic? Even if we set aside the plausibility of α so large and a $\beta > 1$—which says that, in a deterministic environment, starting from a constant consumption path, and agent would be willing to save at a significantly negative real interest rate—we probably shouldn't declare victory just yet.[16]

As it turns out, the second moment implications of our model with $(\alpha, \beta) = (18, 1.125)$ are off—our model has an unconditional standard deviation of the risk-free rate which is too high and an unconditional standard deviation of the equity return which is too low.

Mehra and Prescott estimate from their historical data that $\sigma(R^F) = 0.056$ and $\sigma(R) = 0.165$—*i.e.*, 5.6 percentage points for the risk-free rate, and 16.5 percentage points for the equity return. For the model of exercize 3.3, with $(\alpha, \beta) = (18, 1.125)$, we get

$$\sigma(R^F) = 0.079$$
$$\sigma(R) = 0.139$$

These numbers may seem close to the historical data, but we should first bear in mind that the volatility of the risk-free rate in Mehra and Prescott's 100-year sample is quite high compared to estimates over more recent samples—in fact, Campbell and Cochrane [CC99] (discussed more below) look at the more recent data, and set a constant riskless rate as one of the targets of their model.

Even taking the Mehra-Prescott volatility data at face value, though, the distribution of our model's volatility across the low- and high-growth states is such that the model produces only negligible variation in the conditional excess return to equity and its volatility, hence little variation in the conditional

[16] You may wonder whether $\beta > 1$ poses a problem for existence. If consumption has a long-run growth rate of x, the relevant sufficient condition for existence of optimal paths is $\beta x^{1-\alpha} < 1$, as proved by Brock and Gale in 1969 [BG69] and later re-stated by Kocherlakota in 1990 [Koc90].

Sharpe ratio. This is in contrast to the data, which show large swings in conditional Sharpe ratios from business cycle peaks (where the conditional Sharpe ratio is low) to business cycle troughs (where it is high). A difference on the order of 1.0 between highs and lows is not uncommon.[17]

Our model with $(\alpha, \beta) = (18, 1.125)$ produces:

	$\mathbb{E}\left[R - R^F \vert x\right]$	$\sigma\left[R - R^F \vert x\right]$	$\mathbb{E}\left[R - R^F \vert x\right] / \sigma\left[R - R^F \vert x\right]$
Low x state	0.069	0.114	0.605
High x state	0.051	0.099	0.516

Are these discrepancies between model and data all part of the same puzzle? One could think of three, or perhaps even four separate puzzles we might give names to—an equity premium puzzle (it's hard to get a big equity premium), a risk-free rate puzzle (it's hard to get a low risk-free rate), a volatility puzzle (it's hard to get a big standard deviation of equity returns and a low standard deviation for the risk-free rate), and a Sharpe ratio puzzle (it's hard to get a strongly countercyclical Sharpe ratio). Giving them different names may help to clarify the phenomena we're trying to explain, but in some sense they are all part of one puzzle, since we would not want to say we've solved one, if in doing so we're failing along the other dimensions.

3.4 The Melino and Yang insight: Getting SDFs consistent with first and second moments of returns

Time permitting we will talk about Melino and Yang's [MY03] proposed resolution of the equity premium, and related, puzzles, which relies on state-dependent preferences. Of interest to us here, though, is a trick they use to characterize the process for returns that would be consistent with the consumption process in the Mehra-Prescott model and consistent with Mehra and Prescott's estimates for the unconditional means and standard deviations of returns.

To see the value of that, imagine that you had risk-free rates for the low- and high-growth states

$$\hat{R}^F = \begin{bmatrix} \hat{R}^F(1) \\ \hat{R}^F(2) \end{bmatrix}$$

and a matrix of equity returns for all state transitions

$$\hat{R} = \begin{bmatrix} \hat{R}(1,1) & \hat{R}(1,2) \\ \hat{R}(2,1) & \hat{R}(2,2) \end{bmatrix}$$

[17]See, for example, Ludvigson and Ng [LN07], Lettau and Ludvigson [LL10] or Tang and Whitelaw [TW11].

that, under the Mehra-Prescott transition matrix P for consumption growth, and its long-run distribution $\pi = (1/2, 1/2)$, had unconditional first and second moments that matched Mehra and Prescott's historical estimates. Assuming these didn't imply a collinearity, you could then solve

$$\sum_{j=1}^{2} P(i,j)m(i,j)\hat{R}(i,j) = 1 \qquad (i = 1, 2)$$

$$\sum_{j=1}^{2} P(i,j)m(i,j) = \hat{R}^F(i) \qquad (i = 1, 2)$$

for the stochastic discount factor m that would exactly match those first two moments of returns in the historical data.

Knowing what such a stochastic discount factor m looks like—or what the implied risk-neutral probabilities look like—would be of great help in identifying models that can (or cannot) match the data. Given the behavior of m, the question would be—What sort of model of preferences would map the consumption growth process into that stochastic discount factor?[18]

How do they construct their $\hat{R}^F$ and $\hat{R}$? They assume that consumption growth is a sufficient statistic for the risk-free rate and the equity 'price-dividend' ratio—which is the w in the formula (3.17). This means (as we've seen in our solution) that R^F takes on two values and w takes on two values. They then calculate values of $\hat{R}^F = \left(\hat{R}^F(1), \hat{R}^F(2)\right)$ and $\hat{w} = (\hat{w}(1), \hat{w}(2))$ that produce unconditional means and standard deviations for returns that match Mehra and Prescott's historical estimates. The equity return from state i to state j is related to w by $\hat{R}(i,j) = x(j)\,(1 + \hat{w}(j))\,/\hat{w}(i)$.

The risk-free rate is the easier of the two. $\hat{R}^F$ solves:

$$\frac{1}{2}\hat{R}^F(1) + \frac{1}{2}\hat{R}^F(2) = 1.008$$

$$\sqrt{\frac{1}{2}(\hat{R}^F(1) - 1.008)^2 + \frac{1}{2}(\hat{R}^F(2) - 1.008)^2} = 0.056$$

This is a quadratic equation, which has two solutions:

$$\hat{R}^F_1 = \begin{bmatrix} 1.064 \\ 0.952 \end{bmatrix} \text{ and } \hat{R}^F_2 = \begin{bmatrix} 0.952 \\ 1.064 \end{bmatrix} \qquad (3.21)$$

Assuming that state 1 is the low-growth state, $\hat{R}^F_1$ implies a countercyclical risk-free rate, $\hat{R}^F_2$ a procyclical one. At this point, we can't say which solution is the relevant one, though we will be able to do so after finding the candidate $\hat{R}$'s.

Finding the candidate $\hat{R}$'s is considerably harder—it's a much more complex quadratic equation to solve. Using

$$\hat{R}(i,j) = \frac{x(j)(1 + \hat{w}(j))}{\hat{w}(i)}$$

[18]In fact, as we'll see, Melino and Yang find that preferences must display counter-cyclical risk aversion—this motivates their focus on state-dependent preferences.

and the Mehra-Prescott values for the Markov chain (x, P),

$$x = \begin{bmatrix} 0.982 \\ 1.054 \end{bmatrix} \tag{3.22}$$

and

$$P = \begin{bmatrix} 0.43 & 0.57 \\ 0.57 & 0.43 \end{bmatrix} \tag{3.23}$$

Melino and Yang find the values of $\hat{w}$ that satisfy

$$\frac{1}{2}\sum_{j=1}^{2} P(1,j)\hat{R}(1,j) + \frac{1}{2}\sum_{j=1}^{2} P(2,j)\hat{R}(2,j) = 1.07$$

$$\sqrt{\frac{1}{2}\sum_{j=1}^{2} P(1,j)(\hat{R}(1,j) - 1.07)^2 + \frac{1}{2}\sum_{j=1}^{2} P(2,j)(\hat{R}(2,j) - 1.07)^2} = 0.165$$

As with $\hat{R}^F$, they find two solutions for $\hat{w}$:

$$\hat{w}_1 = \begin{bmatrix} 23.467 \\ 27.839 \end{bmatrix} \text{ and } \hat{w}_2 = \begin{bmatrix} 27.839 \\ 23.467 \end{bmatrix}$$

which implies two possibilities for $\hat{R}$:

$$\hat{R}_1 = \begin{bmatrix} 1.02385 & 1.29528 \\ 0.86306 & 1.09186 \end{bmatrix} \text{ and } \hat{R}_2 = \begin{bmatrix} 1.01727 & 0.92633 \\ 1.20680 & 1.09891 \end{bmatrix} \tag{3.24}$$

This gives four possible combinations of $\hat{R}^F$ and $\hat{R}$, but—as Melino and Yang note—three of the four can be ruled out on the grounds of violating no arbitrage. The combination that remains after eliminating the three that allow arbitrage is:

$$\hat{R}^F = \begin{bmatrix} 1.064 \\ 0.952 \end{bmatrix} \tag{3.25}$$

$$\hat{R} = \begin{bmatrix} 1.02385 & 1.29528 \\ 0.86306 & 1.09186 \end{bmatrix} \tag{3.26}$$

Exercize 3.4. *Show that the other three combinations from (3.21) and (3.24) imply arbitrage opportunities. You can do it in* MATLAB *if you like. Since these are returns, treat the price vector as $p = (1, 1)$. The wrinkle here, compared to Lecture 2, is the two possible states today. For each combination of $\hat{R}^F_i$ and $\hat{R}_j$, you'll want to ask—Is there an arbitrage opportunity if today's state is state 1? Is there an arbitrage opportunity if today's state is state 2? 'Yes' to either of those questions is enough to constitute an arbitrage opportunity—an arbitrage need not be available in both states. Each of those questions is answered using the definition of arbitrage 2.3 from Lecture 2.*

$\hat{R}$ and $\hat{R}^F$ are all you need to back out implied risk neutral probabilities. Routledge and Zin [RZ10] perform this calculation, and get

$$\hat{\psi} = \begin{bmatrix} 0.85 & 0.15 \\ 0.61 & 0.39 \end{bmatrix} \tag{3.27}$$

Compare this with the transition matrix P (3.23). If today's state is the high-growth state, the risk-neutral probabilities (the second row of $\hat{\psi}$) are not that different from the objective probabilities (the second row of P), indicating little risk aversion. By contrast, if today's state is the low-growth state, the risk-neutral probabilities put a much bigger weight on remaining in the low-growth state (and a much smaller weight on moving to the high-growth state), as compared to P. That indicates significant risk aversion. In other words—the structure of the asset returns $\hat{R}$ and $\hat{R}^F$ implies *countercyclical risk aversion*.

This is consistent with observations on the countercyclicality of the conditional Sharpe ratio (and in fact is a sort of confirmation of it). The sort of story one could tell goes like this—in recessions, risk aversion is high, and consequently the price of risk (measured as the compensation in excess return required to bear risk) is high.

It also tells you why the unadorned Mehra-Prescott model we solved in exercize 3.3 fails to generate a strongly countercyclical Sharpe ratio. The SDF for that model depends only on the consumption growth rate realized next period, and is independent of this period's state. In other words—it's a vector, not a matrix. The next exercize asks you to calculate the SDF consistent with Melino and Yang's $\hat{R}$ and $\hat{R}^F$, from (3.25). It's exactly identified, and you'll see it has two non-collinear rows.

Exercize 3.5. *Using the Melino-Yang returns (3.25) and the Mehra-Prescott Markov chain transition matrix (3.23), calculate conditional Sharpe ratios in the two states, and solve for the stochastic discount factor $\hat{m}$ consistent with the returns and the transition matrix P.*

As we go through the various responses to the equity premium puzzle, we'll want to keep the Melino-Yang characterization in mind. The first set of responses we'll look at are models with habit formation.

Lecture 4

Responses to the Equity Premium Puzzle, I: Modifying the Representative Agent

Over the next two lectures, we'll look at various responses to the equity premium puzzle. Given time constraints, this won't be an exhaustive catalog of the models spawned by Mehra and Prescott's observation, though we'll try to hit the ones that are currently the most significant (in terms of both their success and the amount of work currently being done on them).

I've grouped the responses into two broad categories. In this lecture, we'll look at responses that modify the representative agent's preferences. In the next, we'll look at models that tinker in some way with the consumption process.

We'll begin with models that incorporate habit formation in the agent's preferences.

4.1 Models with habit formation

Models with habits were among the earliest responses to Mehra and Prescott's puzzle. Constantinides [Con90] is an early example. See the review of the literature (as of 1999) in Campbell and Cochrane [CC99] for more examples.

The idea is quite simple. Replace the agent's preference over consumption streams—

$$\mathbb{E}_0\left[\sum_t \beta^t u(c_t)\right]$$

—with

$$\mathbb{E}_0\left[\sum_t \beta^t u(c_t - h_t)\right]$$

where h_t, the *habit stock*, is a function of past consumption, and is predetermined at date t. For a given curvature of $u(\,\cdot\,)$, the presence of the habit stock will translate fluctuations in c_t of a given size into larger fluctuations in the marginal utility of consumption. We'll see momentarily that large fluctuations in the intertemporal marginal rate of substitution are possible as a result.

As noted, the habit stock is assumed to be a function of past consumption. For example, h might take the form

$$h_t = D(L)c_{t-1} \tag{4.1}$$

where $D(L)$ is a polynomial in the lag operator. For the most part, we'll work with a simple version of (4.1), in which only last period's consumption matters:

$$h_t = \delta c_{t-1} \tag{4.2}$$

where $\delta \in (0, 1)$ measures the response of the habit stock to past consumption.

4.1.1 Internal versus external habits

Note that if the agent takes account of the effect of today's consumption on tomorrow's habit stock (an *internal habit*), marginal rates of substitution can get quite complicated, even in the simple case of (4.2).

To see what happens, forget about uncertainty (and asset pricing) for a moment, and just think about calculating marginal utilities from

$$U(C) = \sum_{t=0}^{\infty} \beta^t u(c_t - \delta c_{t-1}). \tag{4.3}$$

where $C = \{c_t\}_{t=0}^{\infty}$. Let $U_t(C)$ denote the marginal utility of date t consumption, c_t. Then,

$$U_t(C) = \beta^t u'(c_t - \delta c_{t-1}) - \delta\beta^{t+1} u'(c_{t+1} - \delta c_t), \tag{4.4}$$

$$U_{t+1}(C) = \beta^{t+1} u'(c_{t+1} - \delta c_t) - \delta\beta^{t+2} u'(c_{t+2} - \delta c_{t+1}) \tag{4.5}$$

and the marginal rate of substitution between c_t and c_{t+1}, call it $MRS_{t,t+1}$ is

$$MRS_{t,t+1} = \beta \frac{u'(c_{t+1} - \delta c_t) - \delta\beta u'(c_{t+2} - \delta c_{t+1})}{u'(c_t - \delta c_{t-1}) - \delta\beta u'(c_{t+1} - \delta c_t)} \tag{4.6}$$

which is a bit unpleasant to look at, and would pose some difficulties in working with computationally.

Beginning at least with Abel [Abe90], and especially since the success of Campbell and Cochrane's paper [CC99], it's become increasingly routine to avoid the complexities inherent in (4.6) by assuming the habit is *external*—*i.e.*, the agent takes as given the evolution of h_t, which is viewed as a function of past aggregate or average consumption. As such, external habits reflect a type of 'keeping up with the Joneses' phenomenon.[1]

[1] http://en.wikipedia.org/wiki/Keeping_up_with_the_Joneses

In our simple context, with utility $u(c_t - \delta c_{t-1})$, the agent assumes the habit stock δc_{t-1} is out of his control—we would want to write this more precisely as $u(c_t - \delta \bar{c}_{t-1})$, with a bar to denote an aggregate or average quantity. The marginal rate of substitution relevant for the agent's choices then becomes

$$MRS_{t,t+1} = \beta \frac{u'(c_{t+1} - \delta \bar{c}_t)}{u'(c_t - \delta \bar{c}_{t-1})}$$

In a representative agent model (or a model with a unit mass of identical agents), we would take first order conditions for the agent's intertemporal choice problem, then impose the equilibrium requirement $c_t = \bar{c}_t$ to get

$$MRS_{t,t+1} = \beta \frac{u'(c_{t+1} - \delta c_t)}{u'(c_t - \delta c_{t-1})} \tag{4.7}$$

This differs only slightly—but nonetheless significantly—from the marginal rates of substitution we've worked with this far.

4.1.2 Putting habits in the Mehra-Prescott model

Henceforth, we're going to work with external habits. Putting them in the Mehra-Prescott model is straightforward. We won't re-derive everything, but let's at least state the agent's problem. After that, we'll jump to how the stochastic discount factor gets modified, and from there we'll be able to price assets.

First, we need to expand the aggregate state to include last period's aggregate consumption, call it y_{-1}. The aggregate state, call it s for short, is now $s \equiv (y, y_{-1}, x)$. From one period to the next, the aggregate state evolves according to

$$\begin{aligned} s' &= (y', y'_{-1}, x') \\ &= (x'y, y, x') \end{aligned}$$

where we assume that x' evolves according to a Markov chain $\{x(1), \ldots, x(S); P\}$, as in the original Mehra-Prescott model.

The individual state is (z, b), where z is the agent's beginning-of-period equity holding and b his riskless asset holding. The Bellman equation becomes

$$v(z, b, s) = \max \left\{ u(c - \delta y_{-1}) + \beta \mathbb{E} \left[v(z', b', s') | s \right] \right\}$$

subject to

$$z(y + p(s)) + b \geq c + p(s)z' + q(s)b'$$

Since the habit is external, there's not much change to the agent's first-order conditions; combined with an envelope condition, they yield, for the choice of z':

$$u'(c - \delta y_{-1})p(s) = \beta \mathbb{E} \left[u'(c' - \delta y)(y' + p(s')) | s \right]. \tag{4.8}$$

For the choice of b' we get:

$$u'(c - \delta y_{-1})q(s) = \beta \mathbb{E} \left[u'(c' - \delta y) | s \right]. \tag{4.9}$$

We now impose the equilibrium condition $c = y$ (together with $z' = 1$, $b' = 0$) and re-arrange to get the pricing relationships

$$p(s) = \beta \mathbb{E}\left[\beta \frac{u'(x'y - \delta y)}{u'(y - \delta y_{-1})}(x'y + p(s'))|s\right] \tag{4.10}$$

and

$$q(s) = \beta \mathbb{E}\left[\beta \frac{u'(x'y - \delta y)}{u'(y - \delta y_{-1})}|s\right]. \tag{4.11}$$

The form of the stochastic discount factor is easily seen. Note that if $u'(r) = r^{-\alpha}$, then the SDF has form

$$\begin{aligned}
\beta \frac{u'(x'y - \delta y)}{u'(y - \delta y_{-1})} &= \beta \left(\frac{x'y - \delta y}{y - \delta y_{-1}}\right)^{-\alpha} \\
&= \beta \left(\frac{x'y}{y} \frac{1 - \delta/x'}{1 - \delta/x}\right)^{-\alpha} \\
&= \beta (x')^{-\alpha} \left(\frac{1 - \delta/x'}{1 - \delta/x}\right)^{-\alpha}
\end{aligned} \tag{4.12}$$

The stochastic discount factor thus depends on both this period's and next period's consumption growth—from the Melino-Yang characterization, we know that dependence is a critical feature, if we're to be consistent with countercyclical risk aversion.

Since the discount factor just depends on x and x', and because the present value relationship is linear, it is again the case that the equity price can be written as $p(s) = w(x)y$—there's no dependence on y_{-1} and the dependence on y is linear. The riskless asset price q depends only on x—abusing notation a bit, $q(s) = q(x)$.

The key pricing relations become

$$w(x) = \mathbb{E}\left[m(x, x')x'(1 + w(x'))|x\right] \tag{4.13}$$

and

$$q(x) = \mathbb{E}\left[m(x, x')|x\right] \tag{4.14}$$

where, from (4.12)

$$m(x, x') = \beta (x')^{-\alpha} \left(\frac{1 - \delta/x'}{1 - \delta/x}\right)^{-\alpha}. \tag{4.15}$$

Based on equations (4.13) to (4.15), it's easy to operationalize the model in MATLAB:

- Using the Mehra-Prescott Markov chain (3.22) and (3.23), x, w, and q are 2×1 vectors, m is a 2×2 matrix.
- Form the stochastic discount factor using (4.15) to fill in `m(i,j)` at all the pairs `x(i)` and `x(j)`.

- After that, everything else follows as in exercize 3.3.

You of course need to specify values for the preference parameters, which are now a triple (α, β, δ). In an exercize, you'll experiment with a few combinations to see the impact of habits, and the new parameter governing their strength. Note that since we need consumption to exceed the habit stock in all states (else the argument of the utility function becomes negative), this means the size of δ is limited by the requirement that $1 - \delta/x(i) > 0$ for all i, or $\delta < \min\{x(i)\}$.

Note, too, that in Mehra and Prescott's two-state framework, the term in the expression for the stochastic discount that is added when habits are introduced—that is, the

$$\left(\frac{1-\delta/x'}{1-\delta/x}\right)^{-\alpha}$$

—will be 1 for $x' = x$, in other words for transitions from state one to state one or from state two to state two. Its values for the 'off-diagonal' state transitions—from one to two and from two to one—are inversely related:

$$\left(\frac{1-\delta/x(2)}{1-\delta/x(1)}\right)^{-\alpha} = \frac{1}{\left(\frac{1-\delta/x(1)}{1-\delta/x(2)}\right)^{-\alpha}} \tag{4.16}$$

If we assume—as we did above—that we're fixing state one as the low-growth state, then the term on the left is between zero and one for $0 \le \delta < x(1)$.[2] Its inverse for the other transition ranges over $[1, \infty)$.

This suggests a parsimonious way to parametrize the model with the simple external habit—at least for purposes of exploring the impact of habit. Let θ denote the quantity on the left side of (4.16), and let $m = (m(1), m(2))$ represent the stochastic discount factor from the *original*, no-habit model—*i.e.*, the model of exercize 3.3. The stochastic discount factor for the habit model is then given by

$$\begin{bmatrix} m(1) & \theta \times m(2) \\ \theta^{-1} \times m(1) & m(2) \end{bmatrix} \tag{4.17}$$

It turns out that because habit only modifies the stochastic discount factor in the very limited way given by (4.17), it can't produce an m that matches the Melino-Yang stochastic discount factor you derived in exercize 3.5. Its effects in the low- and high- growth states are simply too symmetric. It *can* deliver the first moments of equity premium data, with minimal utility curvature and a β that's only *slightly* bigger than one. It fails on the second moments, though—which motivates the more complex habit introduced by Campbell and Cochrane. The next exercize asks you to explore this.

[2] The term inside the parentheses varies has a minimum of one (at $\delta = 0$) and diverges to $+\infty$ as δ approaches $x(1)$. Raising it to the $-\alpha < 0$ flips that range into $(0, 1]$.

Exercize 4.1. *For this exercize, set $\alpha = 1$ and $\beta = 1.01$, and use the Mehra-Prescott Markov chain (3.22) and (3.23). You're going to incorporate habit into the stochastic discount factor as in (4.17). Set* `theta` $=$ `linspace(.01,1,100)`*— note you have to start at a $\theta > 0$. Write a* MATLAB *program that calculates, for each* `theta(i)`*, the average returns, the average equity premium, the unconditional volatility of the risk-free rate, and the difference between the conditional Sharpe ratio in the low-growth state and the high-growth state.*

First, (a) find the `theta(i)` *that gets you closest to an equity premium of 0.062, or 6.2%. A simple way to do this is to use* MATLAB'S `min` *and* `abs` *functions—if* `EP` *is the vector of average equity premia, use* `min(abs(EP − 0.062))`*. What do the other results look like at that* `theta(i)`*? What's the implied habit parameter δ at that value of* `theta(i)`*?*

Repeat this for (b) the `theta(i)` *that gets the unconditional volatility of the risk-free rate closest to 0.056, and (c) the* `theta(i)` *that gets the change in the conditional Sharpe ratio closest to 1.*

Make some plots of `theta` *versus the average equity premium, versus the volatility of the risk-free rate, and versus the change in the Sharpe ratio.*

4.1.3 Habits and countercyclical risk aversion

Under external habits, the agent's Arrow-Pratt measure of relative risk aversion is non-constant, and in fact increases as consumption gets closer to the habit stock. Thus, the agent's risk aversion will be higher in those states where consumption growth is low.

Recall that the coefficient of relative risk aversion is essentially the (absolute value of the) elasticity of the marginal utility of consumption with respect to consumption—*i.e.*, the proportionate change in the marginal utility of consumption for a given proportionate change in consumption. It's a measure of the curvature of marginal utility.

Formally, for the usual case without habits, we have

$$\begin{aligned}
\text{Coefficient of RRA} &= \left| \frac{\Delta(u'(c))/u'(c)}{\Delta c/c} \right| \\
&= \left| \frac{u''(c)\Delta c/u'(c)}{\Delta c/c} \right| \\
&= \left| \frac{u''(c)c}{u'(c)} \right|.
\end{aligned}$$

When $u'(r) = r^{-\alpha}$, this gives us α as the coefficient of relative risk aversion.

When an external habit is present (so we treat h as fixed when we vary c, and changing c today has no perceived consequences for tomorrow), we obtain

$$\begin{aligned}\text{Coefficient of RRA } &= \left|\frac{u''(c-h)c}{u'(c-h)}\right| \\ &= \left|\frac{u''(c-h)(c-h)}{u'(c-h)}\right| \frac{c}{c-h}\end{aligned}$$

or $\alpha \dfrac{c}{c-h}$ in the case of $u'(r) = r^{-\alpha}$.

If we think about habits of the simple form $h_t = \delta c_{t-1}$, and $c_t = x_t c_{t-1}$, we have

$$\alpha \frac{c_t}{c_t - h_t} = \alpha \frac{c_t}{c_t - \delta c_{t-1}} = \alpha \frac{1}{1 - \delta / x_t}$$

—low consumption growth brings δ / x_t closer to one, thus raising the agent's relative risk aversion.

The Arrow-Pratt measure of relative risk aversion can be motivated with a thought experiment along the following lines. Imagine the agent faces uncertainty over today's consumption. His level of consumption will be θc, where θ is the realization of a positive random variable with mean 1 and variance σ_θ^2. We're interested in calculating the fraction of certain consumption $c = \mathbb{E}[\theta c]$ he would be willing to give up, call it s, such that he's indifferent between having $(1-s)c$ for sure and having uncertain consumption θc:

$$\mathbb{E}[u(\theta c)] = u\left((1-s)c\right)$$

Taking a second-order Taylor series expansion of $u(\theta c)$ around $\theta = 1$, then passing the expectation operator through, gives

$$\mathbb{E}[u(\theta c)] \approx u(c) + \frac{1}{2} u''(c) c^2 \sigma_\theta^2.$$

A first-order Taylor series expansion of $u((1-s)c)$ around $s = 0$ gives

$$u\left((1-s)c\right) \approx u(c) - u'(c)cs.$$

Combining these gives us an approximate expression for the relative risk premium s, assuming $u' > 0$ and $u'' < 0$:

$$s \approx \frac{1}{2} \left|\frac{u''(c)c}{u'(c)}\right| \sigma_\theta^2 \tag{4.18}$$

4.1.4 Campbell and Cochrane's model

Campbell and Cochrane [CC99] take as a starting point the stochastic discount factor, which we'll write with time subscripts as

$$m_{t+1} = \beta \left[\frac{c_{t+1} - h_{t+1}}{c_t - h_t}\right]^{-\alpha}.$$

Rather than specify how the habit stock h_t depends on past consumption, they define what they call the 'surplus consumption ratio',

$$S_t = \frac{c_t - h_t}{c_t}$$

and make assumptions about its dynamics. The pricing kernel, with this definition, becomes

$$m_{t+1} = \beta \left(\frac{c_{t+1}}{c_t}\right)^{-\alpha} \left(\frac{S_{t+1}}{S_t}\right)^{-\alpha}.$$

Like the stochastic discount factor we derived above, it has a 'standard' part (depending on a discount rate, consumption growth and a curvature parameter) and something 'non-standard', the part which depends on growth of the surplus ratio.

Just as habit led to time-varying risk aversion above, it does as well in Campbell and Cochrane's model, with the coefficient or relative risk aversion inversely related to the surplus ratio:

$$\left|\frac{u''(c_t - h_t)c_t}{u'(c_t - h_t)}\right| = \frac{\alpha}{S_t}.$$

The process the specify for the surplus ratio has consumption growth as its driving impulse, but allows for much richer dynamics than we could achieve in the context of the simple habit model we laid out above. In particular, they assume that

$$\log(S_{t+1}) = (1-\phi)\bar{s} + \phi \log(S_t) + \lambda(\log S_t)[\log(c_{t+1}/c_t) - g] \tag{4.19}$$

where g is the mean of log consumption growth, ϕ controls the persistence of the surplus ratio process, and the crucial function $\lambda(S_t)$ controls the sensitivity of changes in the surplus ratio to shocks to consumption growth.

Log consumption growth is assumed to be a very simple *i.i.d.* process with normal innovations—$\log(c_{t+1}/c_t) = g + v_{t+1}$, with $v_{t+1} \sim N(0, \sigma^2)$. Much of the analysis is carried out in term of logarithmic approximations and exploits the properties of lognormal random variables. Conditional on S_t, for example, S_{t+1} inherits lognormality for log consumption growth. They can write the stochastic discount factor as

$$\begin{aligned} m_{t+1} &= \beta G^{-\alpha} e^{-\alpha(\log(S_{t+1}/S_t) + v_{t+1})} \\ &= \beta G^{-\alpha} e^{-\alpha((\phi-1)(\log(S_t)-\bar{s}) + [1+\lambda(\log(S_t))]v_{t+1})} \end{aligned}$$

where $G = e^g$. The SDF is also conditionally lognormal.

This means, in particular, that conditional on S_t, the ratio of the standard deviation of m_{t+1} to its mean is approximately equal to the conditional standard deviation of $\log(m_{t+1})$, following the same results we used subsequent to equation (3.19) near the end of Lecture 3. Thus, the Sharpe ratio bound ((3.19) or (3.20)) from this discount factor is approximately $\alpha\sigma[1 + \lambda(\log(S_t))]$. Since low values of the surplus ratio correspond to bad times—periods with low consumption—having λ be a decreasing function of the surplus ratio makes possible a countercyclical Sharpe ratio.

The function λ also plays a role in transmitting consumption volatility into volatility of the risk-free rate. Here again, a decreasing λ works in the right direction. In fact, Campbell and Cochrane 'reverse engineer' the λ function so as to guarantee a constant risk-free rate. Their choice of λ at the same time guarantees that the habit stock is unresponsive to consumption innovations at the model's steady state,and moves positively with consumption near the steady state.

How does the Campbell-Cochrane model translate into the two-state Mehra-Prescott framework we've been using throughout the last two Lectures? The first thing to note is that if, within our model, consumption growth is a sufficient statistic for the surplus ratio—so S_t takes on two values, one in the low-growth state and another in the high-growth state, then we can add nothing beyond what we achieved with the simple habit of the previous subsection. As was the case with the simpler habit, the resulting SDF using the surplus ratio would necessarily relate to the basic Mehra-Prescott SDF by (4.17). What we called θ in the discussion would simply be $S(2)/S(1)$.

There's really no easy way to get the rich dynamics inherent in Campbell and Cochrane's specification (4.19) into our two-state version of the Mehra-Prescott model without adding another state variable (probably S_t itself). Adding a state variable isn't necessarily a bad idea—we'll need to do it eventually, to incorporate disasters or long-run risk—but it's hard to see how to create a simple Markov chain that would mimic (4.19).

To do anything more in the two-state framework, we would need to assume that growth of the surplus ratio differs across the four possible state transitions, which means S_{t+1}/S_t must be a more complicated function of x_t and x_{t+1}. Without specifying that function, we might nonetheless ask what it would have to look like to match the exactly identified stochastic discount factor we found using the Melino-Yang returns (exercize 3.5).

That is, simply write, say, γ_S for S_{t+1}/S_t, and solve for the

$$\gamma_S = \begin{bmatrix} \gamma_S(1,1) & \gamma_S(1,2) \\ \gamma_S(2,1) & \gamma_S(2,2) \end{bmatrix}$$

such that $\beta x(j)^{-\alpha}\gamma_S(i,j)^{-\alpha}$ equals the $\hat{m}(i,j)$ we derived in exercize 3.5.

Exercize 4.2. *Do that, assuming $\alpha = \beta = 1$. You can (sort of) interpret*

$$\lambda_1 \equiv \frac{\log(\gamma_S(1,2)/\gamma_S(1,1))}{\log(x(2)/x(1))}$$
$$\lambda_2 \equiv \frac{\log(\gamma_S(2,2)/\gamma_S(2,1))}{\log(x(2)/x(1))}$$

as analogous to values of the λ function in Campbell and Cochrane's (4.19). *What do you find?*

4.1.5 Some additional issues regarding habits

There are a couple additional issues regarding habits that deserve some mention.

First, regarding the assumption of external habits, to paraphrase some remarks I heard Lars Hansen give in a talk at a meeting of the Econometric Society, either external habits are just a gimmick, or the externality is a real one, and we should be thinking about the optimal policy response (as we would with any externality). If it's the latter case, then Lars Ljungqvist and Harald Uhlig's "Optimal Endowment Destruction under Campbell-Cochrane Habit Formation" [LU09] makes for a cautionary analysis.

Ljungqvist and Uhlig approach the model laid out by Campbell and Cochrane from the perspective of social planners. They solve the social planner's problem and find, for the model calibrated as in Campbell-Cochrane, a society of agents with Campbell and Cochrane's preferences and endowment process would experience a welfare gain equal to about a permanent 16% increase in consumption if the planner could enforce one month of fasting each year.

A second issue pertains to what happens if we introduce production. It's turning out that we're probably going to have little time to consider asset-pricing in production economies, but the issue is still worth mentioning. The problem is that the bells and whistles that 'help', when consumption is exogenous, may create problems when consumption is endogenously determined by agents' production and investment decisions. In production economies we face the added constraint that whatever we come up with should not only be consistent with the asset return facts, but also with business cycle facts. That's a tall order to fill.

Specifically regarding habits, the problem entailed by allowing production is this: agents with habits strong enough to generate high equity premia in endowment economies will want to keep their consumption so smooth that—when allowed the chance to do so in production economies—they choose consumption paths that are extremely smooth. Wildly volatile investment—the residual—then absorbs the effects of technology, and other, shocks. Beginning with Jermann [Jer98], this has meant: If we augment the basic stochastic growth model to include habits (to match asset returns data), we must also add capital adjustment costs (to foil agents' attempts to achieve extremely smooth consumption). To paraphrase Jermann, in order for a model to get both the asset return and business cycle facts right, agents must really, really want to smooth consumption—but be prevented from doing it.

Related to the desire to keep consumption smooth in the presence of habits, habits can also lead to perverse labor supply responses, as agents respond to negative labor productivity shocks—which push consumption toward the habit level—with higher work effort. See Graham [Gra08] or some of the impulse responses plotted in [Dol11].

Nevertheless, habits are now quite common in many DSGE models far removed from asset-pricing; see for example the medium-scale model of Smets and Wouters [SW03].

4.2 Epstein-Zin preferences

Epstein-Zin preferences, also sometimes known as Epstein-Zin-Weil preferences or Kreps-Porteus preferences, preserve some of the most attractive features of standard time-additively separable, constant discounting, expected utility preferences—in particular, their recursivity, which makes dynamic programming possible—while breaking (at least to some extent) the tight link between risk aversion and intertemporal substitution implied by the standard preferences.[3]

4.2.1 Basic properties

We've already seen that with preferences of the form

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t u(c_t)\right] = \mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t \frac{c_t^{1-\alpha}}{1-\alpha}\right] \tag{4.20}$$

the parameter α is the coefficient of relative risk aversion, and $1/\alpha$ is the elasticity of intertemporal substitution.[4] With Epstein-Zin preferences, two separate parameters govern the degree of risk aversion (for timeless gambles) and willingness to substitute consumption over time (in deterministic settings).

To motivate the Epstein-Zin form, think about writing (4.20) in a recursive way, letting U_t denote lifetime utility from date t onward, which is a stochastic process, assuming consumption is a stochastic process. We get

$$U_t = u(c_t) + \beta\mathbb{E}_t[U_{t+1}] \tag{4.21}$$

—lifetime utility from today on is an aggregate of within-period utility from c_t, and discounted expected lifetime utility starting tomorrow. It's the additivity in (4.21) that gives us separability over time; that we're taking expectations of future utility implies separability over states.[5]

[3] The main references for the theoretical development are [KP78], [EZ89], and [Wei90].

[4] The elasticity of intertemporal substitution is the elasticity of the ratio c_{t+1}/c_t with respect to the relative price of consumption in periods t and $t+1$. It is only cleanly defined for deterministic consumption paths.

[5] There are a number of ways to define separability. For a differentiable $U(x_1, x_2, \ldots, x_n, \ldots)$, a standard definition is that x_i and x_j are separable from x_k if the marginal rate of substitution $U_i(x)/U_j(x)$ does not depend on x_k. See [BPR78] for many more definitions.

We can imagine relaxing the separability over time by replacing the linear 'aggregator' in (4.21) with something more general:

$$U_t = W(c_t, \mathbb{E}_t[U_{t+1}]). \tag{4.22}$$

This is in fact the form created by Kreps and Porteus [KP78], who gave axioms on a primitive preference ordering over temporal lotteries such that the ordering was representable by a recursive function of the form (4.22).

Epstein and Zin [EZ89] and, independently, Weil [Wei90] wrote down parametric versions of Kreps-Porteus preferences. These preferences give a specific CES form to the aggregator. They also make a convenient monotone transformation of the utility process U_t from (4.22), in such a way that the parameter governing (timeless) risk aversion is very clearly separated from the parameter governing (deterministic) intertemporal substitution. The Epstein-Zin, or Epstein-Zin-Weil, form is:

$$U_t = \left[(1-\beta)c_t^{\rho} + \beta\mu_t\left(U_{t+1}\right)^{\rho}\right]^{\frac{1}{\rho}}, \tag{4.23}$$

where $\mu_t(\,\cdot\,)$ is, in the language of Epstein and Zin, a 'certainly equivalent' operator, conditional on information at date t, having the form:

$$\mu_t\left(U_{t+1}\right) = \left(\mathbb{E}_t\left[U_{t+1}^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}}. \tag{4.24}$$

We assume that $\rho \leq 1$ in (4.23) and $\alpha > 0$ in (4.24).[6] Some key features of the preferences described by (4.23) and (4.24) are:

- The certainty equivalent operator and the CES aggregator are both homogeneous of degree one; thus U_t is homogeneous of degree one in consumption.
- If there's no uncertainty, $\mu_t(U_{t+1}) = U_{t+1}$, and the resulting preferences are ordinally equivalent to time-additively separable preferences with a constant elasticity of intertemporal substitution given by $\dfrac{1}{1-\rho}$.
- The utility of a *constant* consumption path c is just c (think about it); call it $U(c) = c$. Thus, for gambles over constant consumption paths (timeless gambles), preferences are expected utility, with a constant coefficient of relative risk aversion equal to α:

$$\mu_0\left[U(\tilde{c})\right] = \left(\mathbb{E}_0\left[\tilde{c}^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}}.$$

[6] In the case of $\rho = 0$, we get $U_t = c_t^{1-\beta}\mu_t\left(U_{t+1}\right)^{\beta}$, which is ordinally equivalent to log additivity. For $\alpha = 1$, the certainty equivalent can be thought of as $\exp(\mathbb{E}_t[\log(U_{t+1})])$.

- If $\alpha = 1 - \rho$—*i.e.,* if the risk aversion coefficient is the inverse of the intertemporal substitution elasticity, the preferences collapse to a form ordinally equivalent to (4.20).[7]

- In contrast to (4.20), the timing of the resolution of uncertainty matters for agents with Epstein-Zin preferences. The standard preferences (4.20) conform to the 'reduction of compound lotteries' axiom of expected utility. This is not so with (4.23) and (4.24). Consider two lotteries over consumption streams. Lottery A gives c today, c tomorrow, and then either $(c, c, c, \ldots)$ or $(c', c', c', \ldots)$ with probabilities δ and $1-\delta$. Lottery B gives c today, and then either $(c, c, c, \ldots)$ or $(c, c', c', \ldots)$ with probabilities δ and $1-\delta$.[8] Lottery A has a later resolution of uncertainty than lottery B. An agent with preferences given by (4.20) would view them as equivalent. Under Epstein-Zin preferences, early resolution is preferred if $1-\alpha-\rho < 0$, while late resolution is preferred if $1-\alpha-\rho > 0$.

4.2.2 Asset pricing with Epstein-Zin preferences

We can pretty easily put Epstein-Zin preferences into the Mehra-Prescott model. The key question, of course, is—What does the stochastic discount factor look like?

Because Epstein-Zin preferences are recursive, we can do dynamic programming as before. As is out discussion of the habit model, let s stand for the aggregate state, which in this case is $s = (x, y)$, just as in the basic Mehra-Prescott model. We write $\mu_s(\,\cdot\,)$ for the certainty equivalent conditional on the state s. The agent's dynamic program is then given by:

$$v(z, b, s) = \max_{c, z', b'} \left[(1-\beta)c^\rho + \beta \mu_s \left(v(z', b', s') \right)^\rho \right]^{\frac{1}{\rho}} \tag{4.25}$$

subject to

$$z(p(s) + y) + b \geq c + p(s)z' + q(s)b'. \tag{4.26}$$

For compactness, let $W(c, \mu) = [(1-\beta)c^\rho + \beta\mu^\rho]^{1/\rho}$, and note that the partial derivatives of W obey $W_1(c, \mu) = (1-\beta)W(c, \mu)^{1-\rho}c^{\rho-1}$ and $W_2(c, \mu) = \beta W(c, \mu)^{1-\rho}\mu^{\rho-1}$. Note too that, under some mild assumptions, we can pass derivatives through μ as follows. Consider the case of z' (the case of b' is analogous):

$$\begin{aligned}
\frac{\partial}{\partial z'}\mu_s(v(z', b', s')) &= \frac{\partial}{\partial z'} \left(\mathbb{E}\left[v(z', b', s')^{1-\alpha} | s \right] \right)^{\frac{1}{1-\alpha}} \\
&= \frac{1}{1-\alpha} \left(\mathbb{E}\left[v(z', b', s')^{1-\alpha} | s \right] \right)^{\frac{1}{1-\alpha} - 1} \frac{\partial}{\partial z'} \mathbb{E}\left[v(z', b', s')^{1-\alpha} | s \right] \\
&= \mu_s(v(z', b', s'))^\alpha \mathbb{E}\left[v(z', b', s')^{-\alpha} \frac{\partial}{\partial z'} v(z', b', s') | s \right]
\end{aligned}$$

[7] Utility function V is *ordinally equivalent* to utility function U if $U = f(V)$ for some increasing function f. In this case define $V = (1/\rho)U^\rho$.

[8] See Figure 1 in Weil [Wei90].

With all that in mind, the first-order condition for the choice of z' is

$$W_1(c, \mu_s(v(z', b', s')))p(s) = W_2(c, \mu_s(v(z', b', s')))\frac{\partial}{\partial z'}\mu_s(v(z', b', s'))$$

which we can write as

$$(1-\beta)c^{\rho-1}p(s) = \beta\mu_s(v(z', b', s'))^{\rho+\alpha-1}\mathbb{E}\left[v(z', b', s')^{-\alpha}\frac{\partial}{\partial z'}v(z', b', s')|s\right] \tag{4.27}$$

We can use an envelope argument to get an expression for the derivative of v with respect to z. At state (z, b, s), this is given by

$$\begin{aligned}\frac{\partial}{\partial z}v(z, b, s) &= W_1(c, \mu_s(v(z', b', s')))(p(s) + y)\\ &= (1-\beta)W(c, \mu_s(v(z', b', s')))^{1-\rho}c^{\rho-1}(p(s) + y)\\ &= (1-\beta)v(z, b, s)^{1-\rho}c^{\rho-1}(p(s) + y)\end{aligned}$$

where the last line uses that fact that $v(z, b, s) = W(c, \mu_s(v(z', b', s')))$ at the optimum. We now advance this expression one period and plug it into the right hand side of (4.27) to get

$$c^{\rho-1}p(s) = \beta\mu_s(v(z', b', s'))^{\rho+\alpha-1}\mathbb{E}\left[v(z', b', s')^{1-\alpha-\rho}(c')^{\rho-1}(p(s') + y')|s\right]$$

The first-order condition for holdings of the riskless asset is analogous, simply plugging in $q(s)$ for $p(s)$ and 1 for the payoff, rather than $p(s') + y'$—

$$c^{\rho-1}q(s) = \beta\mu_s(v(z', b', s'))^{\rho+\alpha-1}\mathbb{E}\left[v(z', b', s')^{1-\alpha-\rho}(c')^{\rho-1}|s\right]$$

As we've done before, we now impose equilibrium ($c = y$, $z = z' = 1$, $b = b' = 0$) and rearrange to obtain the model's pricing formulas. It'll be convenient to let $V(s) = v(1, 0, s)$—*i.e.*, $V(s)$ is the agent's value function in equilibrium. With the imposition of equilibrium and some re-arranging, we get, for the equity price,

$$p(s) = \mathbb{E}\left[\beta\left(\frac{V(s')}{\mu_s(V(s'))}\right)^{1-\alpha-\rho}\left(\frac{y'}{y}\right)^{\rho-1}(p(s') + y')|s\right] \tag{4.28}$$

We can immediately see the form of the stochastic discount factor, which for now we'll write as depending on both s and s'. In a moment we'll think more carefully about which variables it depends on. We have:

$$m(s, s') = \beta\left(\frac{V(s')}{\mu_s(V(s'))}\right)^{1-\alpha-\rho}\left(\frac{y'}{y}\right)^{\rho-1} \tag{4.29}$$

This stochastic discount factor—like the stochastic discount factor in the habit model—incorporates a 'standard' part and a 'non-standard' part. The standard part is $\beta(y'/y)^{\rho-1}$—the utility discount factor times a decreasing power function of aggregate consumption growth.[9] This is analogous to the $\beta(x')^{-\alpha}$ piece in the stochastic discount factors we've previously encountered.

The non-standard part is the part involving the agent's value function. Suppose that $1-\alpha-\rho<0$. Then, other things the same, payoffs in states where realized lifetime utility falls short of its conditional certainty equivalent value will be weighted more heavily in pricing assets then payoffs in states where realized lifetime utility exceeds its conditional certainty equivalent value. This has the potential to increase the volatility of the stochastic discount factor, by reinforcing the volatility coming from the standard $\beta(x')^{\rho-1}$ channel.

Note that in the special case of $\alpha = 1-\rho$, the stochastic discount factor reduces to the standard $\beta(x')^{-\alpha}$.

So, what does m really depend on?—(x,y,x',y')? Or maybe (x,x')? Or just x'? Make the Mehra-Prescott assumption that $y' = x'y$, with x' following a Markov chain. Under that assumption, m depends on just x and x', with the dependence on x coming through the conditioning in the certainty equivalent present in (4.29). There is no dependence on the level of consumption y, because of the degree-one homogeneity of preferences. That homogeneity allows us to write the equilibrium value function as

$$V(s) = \phi(x)y$$

for some function ϕ. The 'non-standard' term in the stochastic discount factor then becomes

$$\begin{aligned}\left(\frac{V(s')}{\mu_s(V(s'))}\right)^{1-\alpha-\rho} &= \left(\frac{\phi(x')y'}{\mu_x(\phi(x')y')}\right)^{1-\alpha-\rho}\\ &= \left(\frac{\phi(x')x'y}{\mu_x(\phi(x')x'y)}\right)^{1-\alpha-\rho}\\ &= \left(\frac{\phi(x')x'}{\mu_x(\phi(x')x')}\right)^{1-\alpha-\rho}\end{aligned}$$

where the last line uses the homogeneity of μ.

Plugging this into (4.29) gives

$$m(x,x') = \beta(x')^{\rho-1}\left(\frac{\phi(x')x'}{\mu_x(\phi(x')x')}\right)^{1-\alpha-\rho} \tag{4.30}$$

Under the assumption that x' follows a Markov chain $\{x(1),\ldots x(S),P\}$, x' and ϕ are in fact just vectors, and m is a matrix:

$$m(i,j) = \beta(x(j))^{\rho-1}\left(\frac{\phi(j)x(j)}{\mu_i(\phi'x'))}\right)^{1-\alpha-\rho} \tag{4.31}$$

[9] Recall $\rho \leq 1$.

where

$$\mu_i(\phi' x') = \left(\sum_{j=1}^{S} P(i,j)\,(\phi(j)x(j))^{1-\alpha} \right)^{1/(1-\alpha)}. \tag{4.32}$$

As in the basic Mehra-Prescott model, the equity price here is again homogeneous in aggregate consumption—we have $p(x,y) = w(x)y$—and the riskless asset price depends only on x, $q = q(x)$. Under the Markov chain assumption, these are vectors, and the pricing equations become

$$w(i) = \sum_{j=1}^{S} P(i,j)m(i,j)(w(j)x(j) + x(j)) \tag{4.33}$$

$$q(i) = \sum_{j=1}^{S} P(i,j)m(i,j) \tag{4.34}$$

Once we know the matrix $m = [m(i,j)]$, we know how to solve (4.33) and (4.34) for w and q, and from them, all the objects we might be interested in.

Finding m poses a new challenge, though, because of it's dependence on the value function, through ϕ. We first need to find ϕ, and that will require something we've not had to do thus far—solving for something iteratively, rather than just inverting some matrices.

4.2.3 Solving for ϕ

Recall that $\phi(x)y$ is the value function in equilibrium—*i.e.*, the agent's maximized lifetime utility from state (x,y) on, given that $c = y$ at all dates and states. It thus follows the Bellman-like equation

$$\phi(x)y = \left[(1-\beta)y^\rho + \beta\mu_x\left(\phi(x')y'\right)^\rho \right]^{1/\rho}$$

or—dividing both sides by y, using $y' = x'y$ and the homogeneity of μ—

$$\phi(x) = \left[1 - \beta + \beta\mu_x(\phi(x')x')^\rho \right]^{1/\rho}$$

Under the Markov chain assumption, this becomes

$$\phi(i) = \left[1 - \beta + \beta\mu_i(\phi' x')^\rho \right]^{1/\rho} \tag{4.35}$$

where $\mu_i(\phi' x')$ is the same object as in (4.32).

To solve for ϕ, we treat (4.35)—together with (4.32)—as a mapping that takes a vector ϕ_0 into a new vector ϕ_1. Given a ϕ_0, the steps in that mapping, in MATLAB, are:

1. Use (4.32) to create a vector of certainty equivalents—call it, for example, `mu`. If you've set up ϕ_0 and x so they are both column vectors, you can do this in one 'vectorized' step:

```
mu = (P * ((phi0. * x).∧(1 − alpha))).∧(1/(1 − alpha));
```

2. Given mu, update ϕ using (4.35). Again, this step can be done in one operation, without resorting to a 'for' loop:

$$\texttt{phi1} = (\texttt{1} - \texttt{beta} + \texttt{beta} * (\texttt{mu.} \wedge \texttt{rho})). \wedge (\texttt{1/rho});$$

You would perform those two steps repeatedly, inside a 'while' loop, until the successive iterates are changing only negligibly from one iteration to the next. Before starting the loop, you need to set initial conditions and create some variables to control the loop's behavior. You start with some ϕ_0—say a column of ones. I create one variable called chk which will record the distance between iterates, and a variable tol that is my threshold for stopping the loop (when chk gets less that tol). You can use MATLAB's norm function to measure the distance between iterates. I set tol to be something very small, 10^{-5}, say, or 10^{-7}.[10]

I also create a variable called maxits—the maximum number of times to iterate, regardless of whether the iterates converge. This is just in case something is not right, so the loop doesn't go on forever. There's a variable t which starts at 0 and gets incremented by 1 with each pass through the loop. The loop breaks when either chk $<$ tol or t $>$ maxits.

The general form would be:

```
phi0 = ones(2,1);
chk = 1;
tol = 1e-7;
maxits = 1000;
t = 0;
  while chk>tol && t<maxits;
    [Steps to make φ1 here]
    chk = norm(phi0-phi1)
    t = t+1;
    phi0 = phi1;
  end;
```

Note the last line, which sets $\phi_0 = \phi_1$—we're repeatedly mapping a ϕ into a new ϕ, then plugging that result into the mapping to get another iterate.

Exercize 4.3. *Mehra-Prescott meet Epstein-Zin—or, reproducing some results in Weil [Wei89]. Write some* MATLAB *code to solve the Mehra-Prescott model with Epstein-Zin preferences. Report the average equity premium and the average risk-free rate for the following parameter combinations:*

[10] This choice depends a bit on what you're trying to find and what you plan on using it for.

	β	ρ	α
1	0.95	1/2	5
2	0.95	1/2	45
3	0.95	−9	5
4	0.95	−9	45
5	0.98	1/2	5
6	0.98	1/2	45
7	0.98	−9	5
8	0.98	−9	45

These are some of the parameter combinations for which Weil reports results in the tables on page 413 of his 1989 J.M.E. paper. His γ is our α, and his ρ is our $1-\rho$; in the table above, $\rho = 1/2$ corresponds to an EIS of 2, and $\rho = -9$ corresponds to an EIS of 0.1.

When I did this, I found some minor differences (at the first or second decimal places) between my results and Weil's, so don't be surprised if you don't obtain exact *matches. If you're doing it right, though, your numbers should be very, very close to Weil's.*

4.2.4 An alternative for eliminating the value function from the pricing rules altogether

When you're working in the 'laboratory' of computational experiments, the dependence of the equilibrium pricing rules on the value function is not a big deal. One can simply solve for the value function computationally. The presence of the value function in expressions like (4.28) is more problematic if one wants to do econometrics. The value function is unobserved.

Epstein and Zin [EZ89] first noted that it was possible to eliminate the value function from the stochastic discount factor, replacing it with a term related to the return on the representative agent's wealth. The steps are somewhat involved, so let's state the result up front, then go through the derivation. It's convenient to use time subscripts for much of this. We're going to show how to re-write the stochastic discount factor

$$m_{t+1} = \beta \left(\frac{c_{t+1}}{c_t} \right)^{\rho-1} \left(\frac{v_{t+1}}{\mu_t(v_{t+1})} \right)^{1-\alpha-\rho} \tag{4.36}$$

as

$$m_{t+1} = \left(\beta \left(\frac{c_{t+1}}{c_t} \right)^{\rho-1} \right)^{\frac{1-\alpha}{\rho}} (R_{t+1})^{\frac{1-\alpha}{\rho}-1} \tag{4.37}$$

where R_{t+1} is the equilibrium equity return.

Ignoring holdings of the riskless bond (which is assumed to be in zero net supply, anyway), define the agent's wealth at the start of a period by

$$W_t = z_t(p_t + y_t).$$

We can then relate wealth to wealth one period earlier by

$$\begin{aligned} W_{t+1} &= z_{t+1}(p_{t+1} + y_{t+1}) \\ &= \frac{z_{t+1}(p_{t+1} + y_{t+1})}{p_t z_{t+1}} p_t z_{t+1} \\ &= R^z_{t+1}[z_t(p_t + y_t) - c_t] \\ &= R^z_{t+1}(W_t - c_t) \end{aligned} \tag{4.38}$$

Here R^z_{t+1} is the return on the agent's portfolio; in equilibrium this is, of course, the market equity return, which we denote as before by R_{t+1}.[11]

Now write the agent's dynamic program with W as the individual state (s still denotes the aggregate state, and we assume R is a function of s):

$$v(W, s) = \max_c \left[(1-\beta)c^\rho + \beta\mu_s\left(v(R'(W-c), s')\right)^\rho\right]^{1/\rho}.$$

In this expression, we'll assume the agent holds the market portfolio, and we write R' as shorthand for $R(s, s')$, the market return between states s and s'.[12] The degree-one homogeneity of utility implies that v is degree-one homogeneous in W—*i.e.,* we can write

$$v(W, s) = \xi(s)W$$

for some function ξ. Note that since $(W - c)$ is non-stochastic (conditional on s),

$$\begin{aligned} \mu_s\big(v(R'(W-c), s')\big) &= \mu_s\big(\xi(s')R'(W-c)\big) \\ &= \mu_s\big(\xi(s')R'\big)(W-c) \end{aligned}$$

Thus, the agent's maximization problem has the simple form:

$$\max_c \left[(1-\beta)c^\rho + \beta\mu_s\big(\xi(s')R'\big)^\rho (W-c)^\rho\right]^{1/\rho}. \tag{4.39}$$

The problem looks almost static, but the term $\beta\mu_s(\xi(s')R')^\rho$ is capturing the trade-off between consumption today and wealth tomorrow. The first-order condition for the problem is

$$(1-\beta)c^{\rho-1} = \beta\mu_s\big(\xi(s')R'\big)^\rho (W-c)^{\rho-1}$$

[11] It should be clear that including b, or applying this to the Lucas case of multiple trees, makes no difference for the form.

[12] We'll briefly discuss portfolio choice below, in section 4.2.5.

which has a solution of the form

$$c = \kappa(s)W \tag{4.40}$$

where

$$\kappa(s) \equiv \left(1 + \left[\frac{\beta\mu_s(\xi(s')R')^\rho}{1-\beta}\right]^{1/(1-\rho)}\right)^{-1}.$$

Of use later will be the following relation between $\kappa(s)$ and $\mu_s(\xi(s')R')^\rho$ implied by the first order conditions:

$$\frac{1-\kappa(s)}{\kappa(s)} = \left(\frac{\beta\mu_s(\xi(s')R')^\rho}{1-\beta}\right)^{\frac{1}{1-\rho}}. \tag{4.41}$$

Also, substituting $c = \kappa(s)W$ back into the Bellman equation gives a direct relationship between ξ and κ. Making this substitution gives us:

$$\begin{aligned}
\xi(s)^\rho &= (1-\beta)\kappa(s)^\rho + \beta\mu_s\left(\xi(s')R'\right)^\rho (1-\kappa(s))^\rho \\
&= (1-\beta)\kappa(s)^\rho \left[1 + \frac{\beta\mu_s(\xi(s')R')^\rho}{1-\beta}\left(\frac{1-\kappa(s)}{\kappa(s)}\right)^\rho\right] \\
&= (1-\beta)\kappa(s)^\rho \left[1 + \left(\frac{1-\kappa(s)}{\kappa(s)}\right)^{1-\rho}\left(\frac{1-\kappa(s)}{\kappa(s)}\right)^\rho\right] \\
&= (1-\beta)\kappa(s)^\rho \left[1 + \frac{1-\kappa(s)}{\kappa(s)}\right] \\
&= (1-\beta)\kappa(s)^{\rho-1}
\end{aligned} \tag{4.42}$$

where the third line uses (4.41).

The equations (4.41) and (4.42), plus the budget constraint (4.38) (with the equilibrium market return R_{t+1}), and the decision rule (4.40) are all the pieces we need. We proceed by asking what all these relationships imply for the growth rate of consumption between states s_t and s_{t+1}. There are several steps, at various times utilizing (4.38), and (4.40)–(4.42):

$$\begin{aligned}
\frac{c_{t+1}}{c_t} &= \frac{\kappa(s_{t+1})W_{t+1}}{\kappa(s_t)W_t} \\
&= \frac{\kappa(s_{t+1})R_{t+1}(1-\kappa(s_t)W_t}{\kappa(s_t)W_t} \\
&= R_{t+1}\kappa(s_{t+1})\frac{1-\kappa(s_t)}{\kappa(s_t)} \\
&= R_{t+1}\left(\frac{\xi(s_{t+1})^\rho}{1-\beta}\right)^{\frac{1}{\rho-1}}\left(\frac{\beta\mu_{s_t}(\xi(s_{t+1})R_{t+1})^\rho}{1-\beta}\right)^{\frac{1}{1-\rho}} \\
&= \beta^{\frac{1}{1-\rho}} R_{t+1}^{1-\frac{\rho}{\rho-1}}\left(\xi(s_{t+1})R_{t+1}\right)^{\frac{\rho}{\rho-1}}\mu_{s_t}(\xi(s_{t+1})R_{t+1})^{\frac{\rho}{1-\rho}} \\
&= \beta^{\frac{1}{1-\rho}} R_{t+1}^{\frac{1}{1-\rho}}\left[\frac{\xi(s_{t+1})R_{t+1}}{\mu_{s_t}(\xi(s_{t+1})R_{t+1})}\right]^{\frac{\rho}{\rho-1}}
\end{aligned}$$

Now, rearrange to solve for the terms involving the value function—*i.e.*, the terms in ξ:

$$\frac{\xi(s_{t+1})R_{t+1}}{\mu_{s_t}(\xi(s_{t+1})R_{t+1})} = \left(\frac{c_{t+1}}{c_t}\right)^{1-\frac{1}{\rho}} (\beta R_{t+1})^{\frac{1}{\rho}}. \tag{4.43}$$

But, note, the term on the left is precisely $v_{t+1}/\mu_t(v_{t+1})$:

$$\frac{\xi(s_{t+1})R_{t+1}}{\mu_{s_t}(\xi(s_{t+1})R_{t+1})} = \frac{\xi(s_{t+1})R_{t+1}(W_t - c_t)}{\mu_{s_t}(\xi(s_{t+1})R_{t+1}(W_t - c_t))} = \frac{v_{t+1}}{\mu_t(v_{t+1})}.$$

Now, just substitute the right-hand side of (4.43) for $v_{t+1}/\mu_t(v_{t+1})$ in (4.36) and simplify the resulting expression to obtain (4.37).

4.2.5 A note on portfolio choice in the EZ framework

We assumed at the start of this section that the agent held the market portfolio, but the treatment of portfolio choice in the Epstein-Zin framework is itself interesting.

Suppose there are many assets, with $R' = \left(R'_1, R'_2, \ldots R'_n\right)$ the vector of returns. One asset, perhaps the first, may be the risk-free asset. Let θ denote the vector of portfolio weights; as in our treatment of mean-variance portfolio choice, the budget constraint becomes a constraint that the weights sum to one—$\theta \cdot \mathbb{1} = 1$. The portfolio return is $\sum_i \theta_i R'_i = \theta \cdot R'$.

The agent's dynamic programming problem becomes:

$$v(W,s) = \max_{c,\theta} \left\{ \left[(1-\beta)c^{\rho} + \beta \mu_s \left(v(\theta \cdot R'(W-c), s') \right)^{\rho} \right]^{1/\rho} : \theta \cdot \mathbb{1} = 1 \right\}$$

As before, we still have (from the homogeneity of preferences) that $v(W,s) = \xi(s)W$. The choice of c still solves a maximization problem of the form (4.39), given a choice of θ.

Since the aggregator $W(c,\mu)$ is increasing in μ, the choice of θ just maximizes $\mu_s\left(\xi(s')\theta \cdot R'\right)$, that is, θ solves:

$$\max_{\theta} \left\{ \mu_s \left(\xi(s')\theta \cdot R(s,s') \right) : \theta \cdot \mathbb{1} = 1 \right\}.$$

This problem appears to be essentially a static one, but the normalized value function $\xi(s')$ encodes information about the marginal value to the agent of consumption today versus wealth in different states tomorrow.

Taking account of the form of the certainty equivalent operator μ_s, the problem can be written as

$$\max_{\theta} \left\{ \mathbb{E}_s \left[\left(\xi(s') \sum_i \theta_i R_i(s,s') \right)^{1-\alpha} \right]^{\frac{1}{1-\alpha}} : \sum_i \theta_i = 1 \right\}. \tag{4.44}$$

The first-order condition for the kth portfolio weight is:

$$\mathbb{E}_s\left[\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{-\alpha}\xi(s')R_k(s,s')\right] = \lambda(s) \tag{4.45}$$

where $\lambda(s)$ is the Lagrange multiplier on the constraint $\sum_i \theta_i = 1$. Since this condition holds for every asset k, we have, for k and h,

$$\mathbb{E}_s\left[\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{-\alpha}\xi(s')\left(R_k(s,s') - R_h(s,s')\right)\right] = 0.$$

Alternatively, multiply (4.45) by θ_k and sum over all k to get

$$\begin{aligned}\lambda(s) &= \mathbb{E}_s\left[\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{1-\alpha}\right]\\ &= \mu_s\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{1-\alpha}\end{aligned}$$

Then, for any asset k,

$$\mathbb{E}_s\left[\frac{\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{-\alpha}\xi(s')}{\mu_s\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{1-\alpha}}R_k(s,s')\right] = 1 \tag{4.46}$$

But then, (4.46) indicates that

$$\frac{\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{-\alpha}\xi(s')}{\mu_s\left(\xi(s')\sum_i \theta_i R_i(s,s')\right)^{1-\alpha}} \tag{4.47}$$

is a stochastic discount factor. In fact, it is the same stochastic discount factor as in (4.36) or (4.37), assuming the agent holds the market portfolio—*i.e.*, $\sum_i \theta_{i,t} R_{i,t+1}$ equals the equilibrium equity return $R_{i,t+1}$.

Exercize 4.4. *Use* (4.43) *and* $\sum_i \theta_{i,t} R_{i,t+1} = R_{t+1}$ *to show that* (4.47) *is the same stochastic discount factor described by either* (4.36) *or* (4.37).

Forgetting about asset pricing momentarily, within the framework of the last two sections, if an agent faces *i.i.d.* returns, then it's straightforward to show that his consumption-wealth ratio κ and marginal utility of wealth ξ must be constants. In that case, the portfolio problem (4.44) does not depend on ξ and consists simply in choosing a portfolio with a maximal certainty equivalent return, $\mu(\sum_i \theta_i R_{i,t+1})$.[13]

[13]Because returns are *i.i.d.*, there is no dependence of μ on t.

Given that return, the agent chooses consumption today in the same way he would if he faced a constant certain return of $\bar{R} = \mu(\sum_i \theta_i R_{i,t+1})$. Under certainty, Epstein-Zin preferences are ordinally equivalent to

$$\sum_{t=0}^{\infty} \beta^t c_t^{\rho}/\rho.$$

An agent with those preferences facing a constant rate of return on savings equal to $\bar{R}$, would choose a path of consumption satisfying the Euler equations

$$c_t^{\rho-1} = \beta \bar{R} c_{t+1}^{\rho-1}$$

which are solved by a decision rule of the form $c_t = \kappa W_t$, for a constant κ.[14] The κ that solves these first order conditions is given by

$$1 - \kappa = (\beta \bar{R}^{\rho})^{\frac{1}{1-\rho}}.$$

This is the same κ one obtains from combining (4.41) and (4.42) under the assumption of *i.i.d.* returns.

If there is one risky asset—so $\bar{R} = \mu(R_{t+1})$—then a mean-preserving spread of the distribution of R_{t+1} lowers $\mu(R_{t+1})$. This can either increase or decrease the agent's savings rate $(1-\kappa)$, depending on whether $\rho/(1-\rho)$ is positive or not. Recalling that $\rho = 1 - 1/\text{EIS}$, an increase in rate-of-return uncertainty, in an *i.i.d.* world, increases the agent's savings rate so long as his EIS > 1.

This result is worth keeping in mind when we get to the long-run risk model of Bansal and Yaron [BY04]—an EIS greater than one is critical to their results. This property of Epstein-Zin preferences—and others—were shown by Weil in his *Q.J.E.* paper [Wei90].

4.3 State-dependent preferences: More on Melino and Yang (2003)

In section 3.4, we saw Melino and Yang's [MY03] argument for some form of state-dependent preferences as a way of matching the first and second moments of asset returns in the simple two-state Mehra-Prescott model. In this section, we'll look at the form of state-dependence they studied.

Their approach is basically to take the Epstein-Zin preferences described by (4.23) and (4.24), and allow the three parameters ρ, β and α to vary with the aggregate state—*i.e.*, letting $\rho = \rho(s)$, $\beta = \beta(s)$ and $\alpha = \alpha(s)$. In the context of the Mehra-Prescott model, each parameter (potentially) takes on two values, depending one whether consumption growth is low, $x(1)$, or high, $x(2)$.

They experiment with allowing one, two or all three of the parameters to vary with the growth state. Allowing two of the parameters to depend on

[14] You can verify this by plugging in κW_t for c_t and $\kappa \bar{R}(1-\kappa)W_t$ for c_{t+1}.

x (while the third is a constant) gives exactly the degrees of freedom necessary make the model's stochastic discount factor consistent with the return process—(3.25)—that Melino and Yang derived as consistent with the first two moments of asset returns data, given the two-state Mehra-Prescott consumption process.[15] But, as Melino and Yang show, just having the necessary degrees of freedom doesn't mean the resulting state-dependent parameters will satisfy natural requirements like $\rho \leq 1$ or $\alpha > 0$.

If only one parameter is allowed to vary with the state, they have to choose what to target; in these cases, they seek values of the state-dependent parameters that are consistent with the pricing relation for the equity return, $\mathbb{E}_t(m_{t+1}R_{t+1}) = 1$, then see what those parameters imply for the values taken on by the risk-free rate. When all three parameters are allowed to vary with the state, they have an extra degree of freedom, and there are consequently many ways to match the data.

Their notation is almost identical to ours, though they use α for our $1-\alpha$, and denote the consumption growth rate (our x) by g. The price-dividend ratio we have called w, they denote by P—*i.e.*, whereas we have written the equity return from state i to state j as $R(i,j) = x(j)(1+w(j))/w(i)$, they would write $g(j)(1+P(j))/P(i)$. One major difference, though—and this becomes apparent if you try to derive the form of their stochastic discount factor within our model of Epstein-Zin preferences—is that whereas we write the CES aggregator $W(c,\mu)$ as

$$W(c,\mu) = \left[(1-\beta)c^\rho + \beta\mu^\rho\right]^{\frac{1}{\rho}},$$

they write

$$W(c,\mu) = \left[c^\rho + \beta\mu^\rho\right]^{\frac{1}{\rho}}.$$

With a constant utility discount factor, these preferences just differ by a factor of proportionality—they imply the same marginal rates of substitution, hence are equivalent representations. This is not the case when β can vary with the state.

The key expression is their equation (6.9), which describes the stochastic discount factor when all three parameters are allowed to vary with the state. Translated it into our notation, and with our form for the aggregator, it reads:

$$m_{t+1} = \beta(s_t)x_{t+1}^{-\alpha(s_t)}\left(\frac{w_t}{\beta(s_t)}\right)^{1-\frac{1-\alpha(s_t)}{\rho(s_t)}}(1+w_{t+1})^{\frac{1-\alpha(s_t)}{\rho(s_{t+1})}-1} \times \left(\frac{(1-\beta(s_{t+1}))^{1/\rho(s_{t+1})}}{(1-\beta(s_t))^{1/\rho(s_t)}}\right)^{1-\alpha(s_t)} \tag{4.48}$$

The last piece—the one involving $\beta(s_t)$ and $\beta(s_{t+1})$—is one that comes about because of the form we use for the aggregator. Setting that last term equal to one gives the pricing kernel studied by Melino and Yang.

[15] Alternatively, they have the degrees of freedom to equate the model's stochastic discount factor to the $\hat{m}$ you derived in exercize 3.5.

How do they get this? If none of the parameters vary with the state, (4.48) is just a version of our (4.37)—simply plug in x_{t+1} for c_{t+1}/c_t, $x_{t+1}(1+w_{t+1})/w_t$ for R_{t+1}, and rearrange.

When the parameters vary, there is another route, that relies on an equilibrium relationship between the consumption-wealth ratio—the κ of the last section—and the price-dividend ratio w. At the start of the section 4.2.4, we defined the agent's wealth as $z_t(p_t + y_t)$. In equilibrium, with $z = 1$ and $c = y$, the consumption-wealth ratio κ obeys

$$\begin{aligned}\kappa_t &= \frac{c_t}{W_t} \\ &= \frac{y_t}{p_t + y_t} \\ &= \frac{1}{p_t/y_t + 1} \\ &= \frac{1}{w_t + 1}\end{aligned}$$

If you begin with the version (4.47) of the stochastic discount factor, taking account of the dependence of the preference parameters on the state—that is, begin with

$$m_{t+1} = \frac{\xi_{t+1}^{1-\alpha(s_t)} R_{t+1}^{-\alpha(s_t)}}{\mu_t(\xi_{t+1}R_{t+1})^{1-\alpha(s_t)}}$$

—and (a) use (4.41) to replace the $\mu_t(\xi_{t+1}R_{t+1})$ with an expression in w_t, $\rho(s_t)$, and $\beta(s_t)$; (b) use (4.42) to replace the ξ_{t+1} with an expression in w_{t+1}, $\rho(s_{t+1})$, and $\beta(s_{t+1})$; and (c) use $x_{t+1}(1+w_{t+1})/w_t$ to replace R_{t+1}, and re-arrange some terms, you should obtain (4.48).

In terms of the the Markov chain representation for the evolution of the state, we can re-write (4.48) as

$$\begin{aligned} m(i,j) = \beta(i)x(j)^{-\alpha(i)} \left(\frac{w(i)}{\beta(i)}\right)^{1-\frac{1-\alpha(i)}{\rho(i)}} & (1+w(j))^{\frac{1-\alpha(i)}{\rho(j)}-1} \\ & \times \left(\frac{(1-\beta(j))^{1/\rho(j)}}{(1-\beta(i))^{1/\rho(i)}}\right)^{1-\alpha(i)} . \end{aligned} \tag{4.49}$$

To perform the sort of computational experiments that Melino and Yang perform, one would plug the Mehra-Prescott growth rates (x) and the Melino-Yang price-dividend ratios (w) into (4.49). The Melino-Yang price-dividend ratios are (see the P_l and P_h they settle on toward the bottom of page 810):

$$\begin{bmatrix} w(1) \\ w(2) \end{bmatrix} = \begin{bmatrix} 23.467 \\ 27.839 \end{bmatrix}$$

Also, plug in parameter values for any of the taste parameters that are *not* going to be state-varying. Then, seek values of the state-dependent parameters to try to satisfy one or more of the four pricing relations

$$\sum_{j=1}^{2} P(i,j)m(i,j)\hat{R}(i,j) = 1 \tag{4.50}$$

$$\sum_{j=1}^{2} P(i,j)m(i,j) = 1/\hat{R}^F(i) \tag{4.51}$$

where P is the Mehra-Prescott transition matrix and $\hat{R}$ and $\hat{R}^F$ are the Melino-Yang returns (3.25).

Melino and Yang consider several combinations of state-dependence in one or more parameters, while keeping the other(s) constant. The cases they report results for are:

1. β and ρ fixed, α state-dependent.
2. β and α fixed, ρ state-dependent.
3. ρ fixed, α and β state-dependent.
4. β fixed, α and ρ state-dependent.
5. All three parameters state-dependent.

In (1) and (2) they seek values to satisfy (4.50), then check what those values imply for the risk-free rate—*i.e.*, they check how badly they miss on (4.51). In (3) and (4), they can hit both sets of targets, but not necessarily with plausible parameter values.

Probably their most striking finding is that countercyclical risk aversion *alone*—that is, with constant ρ and β—doesn't help: there's no real improvement over what you could get with all parameters constant. A procyclical elasticity of intertemporal substitution allows you to match first moments of returns, keeping the other parameters fixed. *But*, countercyclical risk aversion together with a slightly procyclical willingness to substitute intertemporally allows you to match both first and second moments of the returns data.

Since the risk-free rate is countercyclical—in the data as well as in their $\hat{R}^F$—it's not surprising that we get a procyclical EIS: in bad times agents must be demanding more compensation to substitute consumption over time. Not surprisingly, given the risk neutral probabilities implied by their returns $\hat{R}$ and $\hat{R}^F$, the risk aversion parameter α alternates between extreme risk aversion and approximate risk neutrality. For a constant utility discount factor of $\beta = 0.98$, for example, they match first and second moments exactly with risk aversion alternating between $\alpha \approx 23$ and $\alpha \approx 0$, while ρ moves slightly between about -1.98 and -2.10—*i.e.*, the agent's EIS alternates between roughly 0.34 and 0.32.[16]

[16] And, in a way, that's the truly striking part. Just countercyclical risk aversion?—no real gain. Countercyclical risk aversion plus an almost imperceptible change in the EIS?—Bingo.

4.4 Preferences displaying first-order risk aversion

Habits and Epstein-Zin preferences both relax the state separability characteristic of the standard time-additive, expected-utility preferences used by Mehra and Prescott (and in most RBC/DSGE models). For timeless gambles, they still conform to the axioms of expected utility (in particular, state separability). The Melino-Yang preferences are a bit harder to categorize, but conditional on the state of the economy, they are simply EZ preferences, and thus conform to expected utility as well (again, conditional on the state). The preferences in this section and the next, each, in different ways, dispense with the independence axiom, and thus depart from expected utility, even for timeless gambles.[17]

The first form we'll look at goes by different names—'first order risk aversion', 'rank dependent expected utility', 'expected utility with rank dependent probabilities', 'anticipated utility', or 'Yaari' preferences, to name a few. We'll refer to them as first-order risk averse, or FORA, preferences, though bear in mind that's a bit vague (the disappointment aversion preferences of the next section also display first-order risk aversion).

It's probably easiest to describe *what* they are, and then talk about *why* they are what they are. FORA risk preferences were first applied to the equity premium puzzle by Epstein and Zin [EZ90]. In terms of functional structure, they are like standard EZ preferences—a CES aggregator plus a certainty equivalent operator. What's different is the form of the certainty equivalent operator.

4.4.1 FORA: The 'what'

Suppose ϕ follows an n-state Markov chain with transition probabilities P, and importantly, suppose that states are ordered such that

$$\phi(1) \leq \phi(2) \leq \phi(3) \cdots \leq \phi(n).$$

Then, we define the certainty equivalent, for $\alpha \geq 0$, by:

$$\hat{\mu}_i(\phi) = \left[\sum_{j=1}^{n} G_P(i,j)\phi(j)^{1-\alpha} \right]^{1/(1-\alpha)}, \tag{4.52}$$

where the $G_P(i,j)$ are defined, for $\gamma \in (0,1]$, by:

$$G_P(i,1) = P(i,1)^\gamma, \qquad (j=1) \tag{4.53}$$

$$G_P(i,j) = \left(\sum_{h=1}^{j} P(i,h) \right)^\gamma - G_P(i,j-1), \qquad (j = 2,3,\ldots n). \tag{4.54}$$

Thus, for example, in the two-state case, we have

$$G_P(i,1) = P(i,1)^\gamma$$
$$G_P(i,2) = 1 - P(i,1)^\gamma$$

[17] The best reference work on issues concerning choice under uncertainty is, to my mind, Krep's 'underground classic', *Notes on the Theory of Choice* [Kre88].

Note that the $G_P(i,j)$ have all the properties of Markov chain probabilities: $G_P(i,j) \geq 0$ for all i and j, and $\sum_j G_P(i,j) = 1$ for all i. Thinking of them as probabilities, note that their 'CDF' obeys:

$$\begin{aligned}
\Pr\{\phi_{t+1} \leq \phi(j) : \phi_t = \phi(i)\} &= \sum_{k=1}^{j} G_P(i,k) \\
&= \left(\sum_{k=1}^{j} P(i,k)\right)^{\gamma} \\
&\geq \sum_{k=1}^{j} P(i,k)
\end{aligned}$$

The last line follows from the fact that when $\gamma \in (0,1]$, $r^\gamma \geq r$ for all $r \in [0,1]$.

In a sense, G_P gives more weight to the lower valued outcomes than does the true probability P. This is most apparent, again, in the two-state case. Suppose that $P(i,1) = P(i,2) = \frac{1}{2}$, and that $\gamma = 0.9$. Then

$$(G_P(i,1), G_P(i,2)) = \left((1/2)^{0.9}, 1-(1/2)^{0.9}\right) \approx (0.54, 0.46).$$

In terms of asset-pricing, everything works here as in our treatment of the basic Epstein-Zin model—we just replace the original $\mu(\cdot)$ with the new $\hat{\mu}(\cdot)$, and all the expectations that *the agent* takes with expectations using the 'probabilities' G_P.[18]

Let $\hat{\mathbb{E}}$ denote expectation taken with respect to G_P. The key pricing equation becomes (for any asset return R)

$$\hat{\mathbb{E}}_t\left[\left(\frac{c_{t+1}}{c_t}\right)^{-\frac{1}{\psi}}\left(\frac{v_{t+1}}{\hat{\mu}_t(v_{t+1})}\right)^{\frac{1}{\psi}-\alpha} R_{t+1}\right] = 1 \tag{4.55}$$

The rank-ordering of the outcomes is based on the values of v_{t+1}—the Markov states need to be ordered so that state 1 has the lowest value of v_{t+1} and state n has the highest.

Because the aggregator and certainty equivalent are still homogeneous of degree one, we get all the useful homogeneity properties we had before—v_t still equals $\Phi(x_t)c_t$ in equilibrium when consumption follows a Mehra-Prescott-type process, and the price of a consumption claim still has the form $p(x_t, c_t) = w(x_t)c_t$.

If you've written good code to solve the Mehra-Prescott model with Epstein-Zin preferences, it's easy to modify that code to solve the model with FORA preferences. An exercize below will ask you to do just that.

[18] And note well that when we get to the steps where *we're* calculating expectations—*e.g.*, the expected return on equity—we use the true probability P, *not* G_P.

4.4.2 FORA: The 'why'

At first glance, FORA preferences may seem a bit *ad hoc*—we've basically just made the agent more pessimistic about the evolution of the state of the economy than is really the case. But, these preferences actually have a very good theoretical pedigree. First, though, let's understand what's meant by 'first-order' risk aversion.

The 'first-order' in the name refers to the risk premia these preferences imply—that is, the compensation an agent requires to be indifferent between a risky consumption outcome and a certain one. With standard expected utility preferences—as we saw back in equation (4.18)—the risk premium associated with a small gamble is proportional to the gamble's variance (which is second-order small); those are 'second-order risk averse' preferences. For gambles with a small standard deviation, σ^2 is much smaller than σ—consumption growth, for example, has a standard deviation around 0.03 and a variance around 0.009. In fact, as σ gets small, $\sigma^2 \approx 0$, so expected utility agents are approximately risk neutral for sufficiently small gambles.[19]

Yet, evidence suggests that people are non-negligibly averse to small gambles.[20] Getting expected utility maximizing agents to care about small risks requires extreme amounts of curvature in their utility functions. In contrast, preferences that display 'first order' risk aversion generate risk premia that (for small gambles) are proportional to gambles' standard deviations—so risk premia decline linearly with the size of the gamble, rather than quadratically.

A first-order risk averse agent might think about extending the warranty on his refrigerator. An expected utility agent never would—or, if he would do that, then he would necessarily behave in a wildly risk averse way with respect to sizeable gambles.[21]

Epstein and Zin's specification of first-order risk aversion is based on the non-expected utility formulations of Yaari [Yaa87] and Quiggin [Qui82]. Risk preferences of this sort can be derived under various sets of axioms (see Wakker [Wak94], and the references therein). A key feature of these preferences—like many other alternatives to expected utility—is that they are non-linear in probabilities, hence will violate the independence axiom underlying expected utility. Among the aims of the authors who originally formulated risk preferences of this form was to elaborate models of choice under risk capable of rationalizing the apparent fact that individuals often make choices that are inconsistent with the independence axiom—for example, the Allais paradox, or the com-

[19] The distinction between first-order and second-order risk aversion was first made by Segal and Spivak [SS90].

[20] Not everyone agrees on the evidence, but people do routinely insure against small risks—*e.g.*, paying \$50 or so to extend the warranty on a \$500 refrigerator. More seriously, see, for example, the application in Bernasconi [Ber98], which uses first-order risk aversion to rationalize what appear (from an EU perspective) to be puzzlingly high rates of tax compliance in most developed economies.

[21] This point—which we discuss in more detail below—is formalized in Rabin's [Rab00] 'calibration theorem'. See also [SS08].

mon ratio effect documented by Kahneman and Tversky [KT79].[22]

The *Allais Paradox* is the following observation about choices individuals typically make. Consider the following choice problems (parameters from Kahneman and Tversky [KT79]):

(P1) Choose between:

- **A** \$2,500 with probability 0.33, \$2,400 with probability 0.66, \$0 with probability 0.01.
- **B** \$2,400 with probability 1.

(P2) Choose between:

- **C** \$2,500 with probability 0.33, \$0 with probability 0.67.
- **D** \$2,400 with probability 0.34, \$0 with probability 0.66.

Most people choose B in problem P1 and C in problem P2, in violation of expected utility.
The *Common Ratio Effect* is the following observation about the choices people typically make. Again, consider the following two choice problems (parameters again from Kahneman and Tversky):

(P1) Choose between:

- **A** \$4,000 with probability 0.80, \$0 with probability 0.20.
- **B** \$3,000 with probability 1.

(P2) Choose between:

- **C** \$4,000 with probability 0.20, \$0 with probability 0.80.
- **D** \$3,000 with probability 0.25, \$0 with probability 0.75.

Again, most people choose B in problem P1 and C in problem P2, in violation of expected utility.

When $\alpha \neq 0$ in (4.52) and $\gamma \neq 1$ in (4.53) and (**??**), $\hat{\mu}$ incorporates aspects of preferences featuring both first-order risk aversion and more standard constant relative risk aversion (CRRA). If $\gamma = 1$, we are in the case of basic Epstein-Zin preferences, while if $\alpha = 0$ we have risk preferences that match the formulation of Yaari. Of course, if $\gamma = 1$ and $\alpha = 1/\psi$, we obtain the case of time-additively separable expected utility.

The fact that risk preferences of this form are non-linear in probabilities gives them another attractive feature: the ability to at least partially divorce agents' attitudes towards risk from their attitudes towards wealth.[23] Under expected utility, aversion to risk is equivalent to diminishing marginal utility of wealth, and the intimate connection between the two concepts has been

[22] Starmer [Sta00] is an excellent recent survey of this literature.

[23] As Yaari [Yaa87] puts it: "At the level of fundamental principles, risk aversion and diminishing marginal utility of wealth, which are synonymous under expected utility, are horses of different colors." In Yaari's theory the divorce of the two concepts is complete.

shown to be problematic for the EU model. For example, Chetty [Che06] has shown that estimates of labor supply elasticity (and the degree of complementarity between consumption and leisure) can put sharp bounds on admissible coefficients of relative risk aversion, since both values are linked to the curvature of agents' von Neumann-Morgenstern utilities over consumption. Chetty finds that the mean coefficient of relative risk aversion implied by 33 studies of labor supply elasticity is roughly unity, which would mean that the EU model is incapable of rationalizing both observed labor supply behavior and the degrees of risk aversion observed in many risky choice settings, many of which imply double-digit coefficients of relative risk aversion.

As we hinted at above, another attractive feature of FORA preferences is the fact that they can be parametrized to give a reasonable amount of risk aversion for both large and small gambles. This is in contrast to the standard expected utility specification. In the CRRA class, for example, if the coefficient of risk aversion is calibrated so that an agent with those preferences gives plausible answers to questions about large gambles, the agent will be roughly risk neutral for small gambles. If, on the other hand, the coefficient of risk aversion is set sufficiently large that the agent gives plausible answers to questions about small gambles, he will appear extremely risk averse when confronted with large gambles.[24]

One way to visualize the approximate risk neutrality of the standard expected utility specification with constant relative risk aversion is to note that it's "smooth at certainty"—the agent's indifference curves between consumption in different states of nature are smooth and tangent (at the certainty point) to the indifference curves of a risk neutral agent. This is true for EU with any differentiable von Neumann-Morgenstern utility function.

FORA preferences introduce a kink into agents' indifference curves at the certainty point; the kink is what allows for a plausible calibration of risk aversion for small gambles.[25] The parameter γ—which makes outcome rankings matter—is the source of the kink. The parameter α, analogous to the risk aversion coefficient in CRRA preferences, governs curvature away from the certainty point and allows for a plausible calibration of risk aversion for large gambles.

[24]This point was made formally by Rabin [Rab00], though EU preferences aren't the only form susceptible to this critique. Indeed, as Safra and Segal show in a recent paper [SS08], almost all common alternatives to expected utility are susceptible to this criticism. The one exception noted by Safra and Segal is Yaari's dual theory of choice under risk, which is the special case of (4.52) when $\alpha = 0$. A useful perspective is offered by Palacios-Huerta, Serrano and Volij [PSV04]: "[I]t is more useful not to argue whether expected utility is literally true (we know that it is not, since many violations of its underpinning axioms have been exhibited). Rather, one should insist on the identification of a useful range of empirical applications where expected utility is a useful model to approximate, explain, and predict behavior."

[25]See figure 1 in [EZ90]. The "disappointment aversion" preferences used by Routledge and Zin [RZ10] and Campanale *et al.* [CCC10] share this feature.

4.4.3 FORA: Some results

Blah blah.

4.5 Models with disappointment aversion

This is a promising approach that, unfortunately, we won't have time to go over in class. The key papers are: Gul [Gul91], for the axiomatic background; Routledge and Zin [RZ10], for the asset-pricing application; and Campanale, Castro and Clementi [**?**], for situating preferences of this form in a production economy.

Lecture 5

Responses to the Equity Premium Puzzle, II: Modifying the Consumption Process

The models of the last lecture all modified, in some way, the preferences of Mehra and Prescott's representative agent.[1] This lecture will focus on two models that alter the consumption process faced by the representative agent—Bansal and Yaron's [BY04] 'long run risk' approach, and the 'rare disasters' model, originally due to Rietz [Rie88], but lately revived by Barro [Bar06], Gourio [Gou08], and Gabaix [Gab08].

Your first thought might be—"The consumption process is whatever it is in the data; you can't just plug in another one." That would be true if the data spoke definitively on the subject, but, given a limited number of observations, the data may not be sufficient to discriminate between alternatives that, while close to one another in some statistical sense, have dramatically different implications for the behavior of economic models. Is the distribution of log consumption growth rates better described by a normal distribution or by a distribution with fatter tails? Whether a consumption disaster is likely to occur once every couple hundred years or a couple times every hundred years is difficult to decide with just 100 years' worth of data, but makes an enormous difference for pricing claims to aggregate consumption.

Looking ahead to Bansal and Yaron's long-run risk, it's difficult to distinguish between a consumption process with log differences that are *i.i.d*. about a

[1] Unfortunately, we didn't have time to look at models that dispense with the representative agent altogether—for example, [Guv09]. Our treatment was also far from exhaustive, omitting the interesting work on disappointment aversion by Routledge and Zin [RZ10] or Campanale *et al.* [CCC10].

constant mean, and a process with *i.i.d.* fluctuations about a conditional mean subject to very small but very persistent fluctuations. Figure 5 plots some artificial data I created in MATLAB. In one of the series in the top panel I used a constant mean growth rate, in the other a fluctuating conditional mean. The fluctuating conditional mean—the difference between the two series in the top panel—is shown in the lower panel. The parameters are Bansal and Yaron's, so the standard deviation of the innovations to the conditional mean is very small compared to the innovations *around* the conditional mean. The *i.i.d.* in-

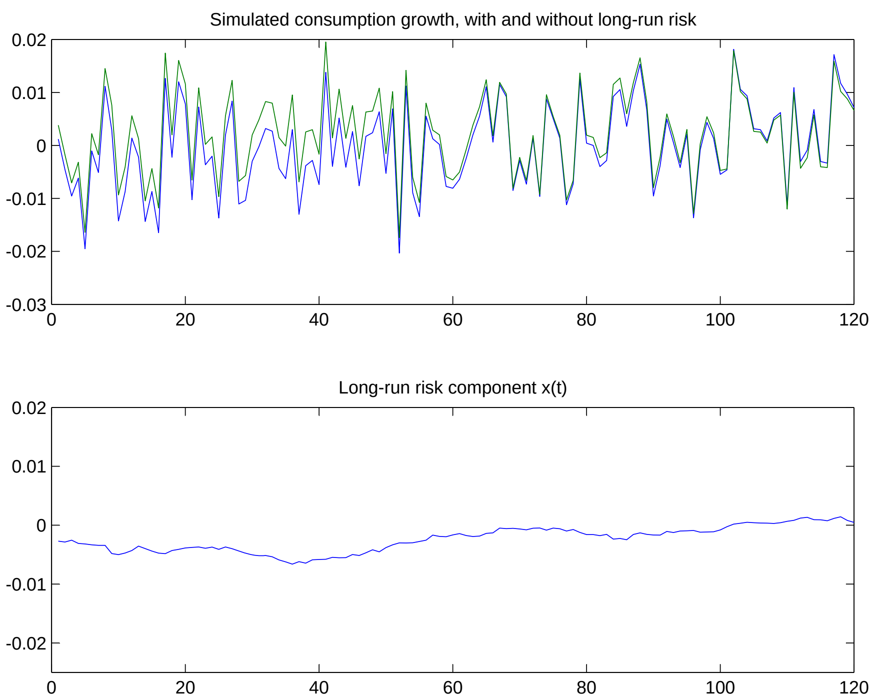

Figure 5.1: Simulated data using parameters from Bansal and Yaron's consumption process. Top panel shows log growth rates with and without long-run risk component. Bottom panel shows long-run risk component.

novations swamp the innovations to the conditional mean, making it difficult to see much difference between the series. The Bansal-Yaron parameters are calibrated for monthly data, so the 120 observations in the figure correspond to ten years' worth of data. Given enough data, the difference between the two processes is readily apparent in the *level* of consumption.

Figure 5 plots the log-levels of consumption—*i.e.*, cumulated log growth rates—for series with and without a long-run risk component. To highlight the differences, in this case I've plotted simulated series of $10,000$ observations each—about 800 years' worth. The differences are clear even over shorter sub-

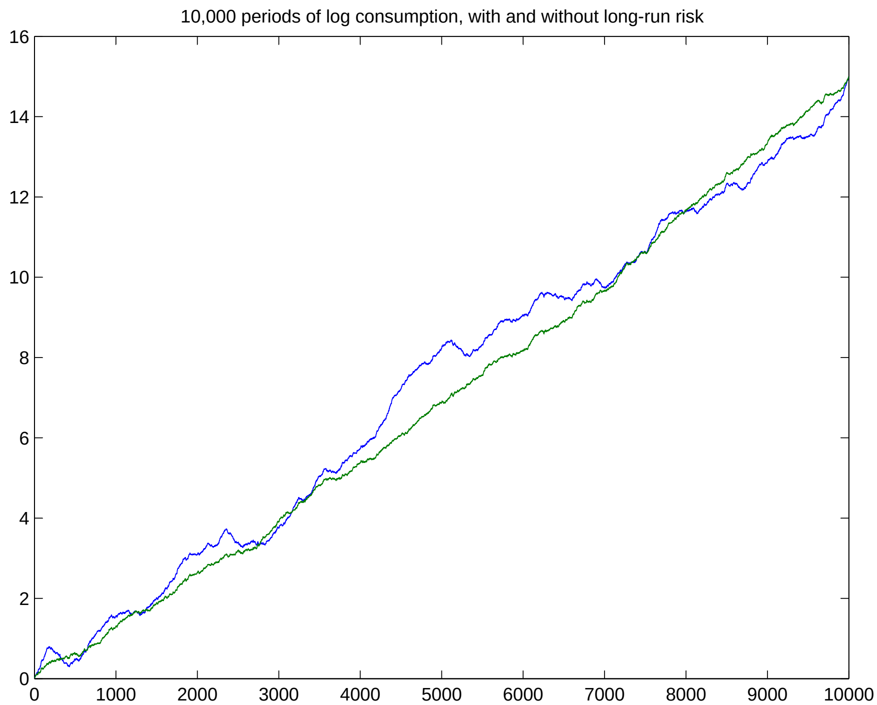

Figure 5.2: Simulated data using parameters from Bansal and Yaron's consumption process. The figure shows cumulated log growth rates for series with and without a long-run risk component.

samples, but I've chosen to show $10,000$ periods just because anything shorter might look like I'm cherry-picking the data. And, it's certainly *not* movements at that low a frequency that are driving the difference between the asset pricing implications with and without long-run risk. In an exercize to follow, you'll be asked to solve a version of the Bansal-Yaron model. That solution will involve iterating on the representative agent's (equilibrium) value function. By checking how many iterations it takes for the iterates of the value function to converge to within a reasonably small tolerance of one another—say 10^{-7}—you'll see that the agent effectively looks into the future far fewer than $10,000$ periods.[2]

[2]To be sure, technically the agent looks over the whole infinite horizon; practically speaking, though, anything beyond 150 or so periods is discounted so heavily that it has only a negligible impact on the agent's utility.

5.1 Long-run risk in consumption growth: Bansal and Yaron (2004)

5.1.1 Some context: Explaining movements in price-dividend ratios

Let's begin with a little bit in the way of context, though. Long-run risk emerged as a potential resolution of the equity premium puzzle only after movements in the long-run growth rate of dividends were suggested as an explanation for the apparent excessive volatility of price-dividend ratios.

That 'excess volatility' literature begins in the early 1980s, with the papers of Shiller [Shi81] and Leroy and Porter [LP81]. It was initially framed in terms of the volatility of stock prices relative to the present discounted value of dividends. While that initial approach suffered from some econometric flaws that were quickly pointed out, it's still worth a quick look, since its descendants still frame current discussions.[3]

Shiller's paper was probably the more influential of the two. Shiller argued that stock prices should be expectations of what he called their 'ex post rational value', or $p_t = \mathbb{E}_t(p_t^*)$. His 'ex post rational value' was the discounted sum of actual future dividend payments, discounted at a constant rate γ:

$$p_t^* = \sum_{j=1}^{\infty} \gamma^j d_{t+j} \tag{5.1}$$

Note that the d_{t+j}'s represent actual, realized dividends.

Then, if $p_t = \mathbb{E}_t(p_t^*)$, it follows that we can remove the expectation and write

$$p_t^* = p_t + u_t \tag{5.2}$$

where u_t is orthogonal to p_t. If we take the variance of both sides of (5.2), we obtain

$$\mathrm{var}(p_t^*) = \mathrm{var}(p_t) + \mathrm{var}(u_t)$$

since p_t and u_t are orthogonal. Since variance is always nonnegative, we obtain the variance bound

$$\mathrm{var}(p_t^*) \geq \mathrm{var}(p_t) \tag{5.3}$$

which states that the variance of the stock price cannot exceed the variance of the 'ex post rational price'.

Shiller then constructed an empirical counterpart to p_t^* using data on actual dividends, an assumption about γ, and approximating the infinite series in (5.1) with a truncated one. He then compared detrended p_t and p_t^* and concluded that p_t was much more volatile that p_t^*, in apparent violation of (5.3). See figure 5.1.1, which reproduces Figure 1 from Shiller's paper.

[3] Leroy's Palgrave entry on "Excess Volatility Tests" [Ler08] gives a nice history of this literature. A version of this short, informative paper is available from Leroy's website, here `http://www.econ.ucsb.edu/~sleroy/downloads/excess.pdf`.

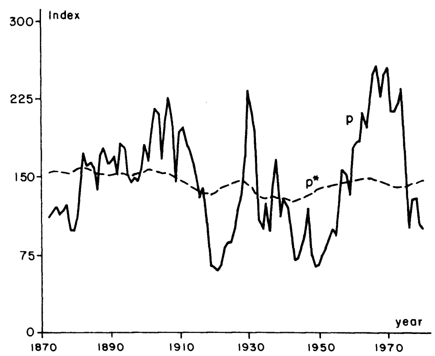

FIGURE 1
Note: Real Standard and Poor's Composite Stock Price Index (solid line p) and *ex post* rational price (dotted line p^*), 1871–1979, both detrended by dividing a long-run exponential growth factor. The variable p^* is the present value of actual subsequent real detrended dividends, subject to an assumption about the present value in 1979 of dividends thereafter. Data are from Data Set 1, Appendix.

Figure 5.3: Figure taken from Shiller [Shi81].

While this approach had some econometric problems associated with it—having to do with the possible nonstationarity of both prices and dividends—as well as some economic problems,[4] the point still stuck. Eventually, the idea of excess price volatility was formulated in more robust way as a question of what explains the volatility of price-dividend ratios.

Campbell and Shiller [CS88] provided a useful framework—referred to now as the Campbell-Shiller approximation—for thinking about the sources of volatility of price-dividend ratios. They start with an identity based on the definition of an *ex post* return on a stock (or index of stocks):

$$R_{t+1} = (d_{t+1}/d_t)(1 + p_{t+1}/d_{t+1})/(p_t/d_t).$$

They then take a Taylor series approximation of this identity to get

$$r_{t+1} = \kappa_0 + g_{t+1} + \kappa_1 z_{t+1} - z_t \tag{5.4}$$

where $g_{t+1} = \log(d_{t+1}/d_t)$ and $r_{t+1} = \log(R_{t+1})$ are the continuous dividend growth rate and continuous return, and z_t is the log price-dividend ratio $\log(p_t/d_t)$. The coefficients κ_0 and $\kappa_1 > 0$ come out of the Taylor series approximation, and involve means of r, z and g.

As it stands, (5.4) has no economic content—it's an approximation of an identity that must hold *ex post*. One can give it economic content by turning

[4] What model predicts a constant-discounted, expected present value as the price of an asset? Oddly, Shiller's paper came out *after* Lucas's 1978 paper, though was apparently not at all informed by it.

it around, applying rational expectations, and treating it as determining today's price-dividend ratio given expectations of tomorrow's return, dividend growth, and price-dividend ratio. Recursively substituting for the future price-dividend ratio gives an expression relating today's price-dividend ratio to the expected value of all future returns and dividend growth rates:

$$\begin{aligned} z_t &= \mathbb{E}_t[\kappa_0 + g_{t+1} - r_{t+1} + \kappa_1 z_{t+1}] \\ &= \frac{\kappa_0}{1-\kappa_1} + \mathbb{E}_t\left[\sum_{j=1}^{\infty} \kappa_1^{j-1}\left(g_{t+j} - r_{t+j}\right)\right] \end{aligned} \tag{5.5}$$

The expression (5.5) reveals two sources of variation in an asset or portfolio's price-dividend ratio, changes in expected dividend growth rates or changes in expected returns. For example, a little manipulation of (5.5) shows that $z_{t+1} - z_t$ is given by

$$\begin{aligned} z_{t+1} - z_t = &-\frac{1}{\kappa_1}\mathbb{E}_t(g_{t+1} - r_{t+1}) \\ &+ \frac{1-\kappa_1}{\kappa_1}(z_t - \frac{\kappa_0}{1-\kappa_1}) + (\mathbb{E}_{t+1} - \mathbb{E}_t)\left[\sum_{j=2}^{\infty} \kappa_1^{j-1}\left(g_{t+j} - r_{t+j}\right)\right] \end{aligned}$$

I think it's fair to say that for the past two decades, the assumption has been that most of the action here had to come from changing expectations of returns—if g_{t+1} is *i.i.d.*, or close to it, then $(\mathbb{E}_{t+1} - \mathbb{E}_t)\left[\sum_{j=2}^{\infty} \kappa_1^{j-1}\left(g_{t+j}\right)\right] \approx 0$. This is the perspective, for example, of Campbell and Cochrane [CC99], who assume the growth rates are in fact *i.i.d.*, and seek a mechanism for generating swings in the expected returns.

However, Barsky and DeLong [BDL93], and later Bansal and Lundblad [BL02], pointed out that permanent or very highly persistent changes in the expected growth rate, even if very small, could have large effects on asset prices. From (5.5), if κ_1 is close to one, a small permanent increase in the conditional mean of dividend growth can lead to a large increase in the log price-dividend ratio, holding fixed the returns r_{t+j}. Suppose $\mathbb{E}_{t+1}(g_{t+j}) - \mathbb{E}_t(g_{t+j}) = \Delta$ for all j. Then,

$$(\mathbb{E}_{t+1} - \mathbb{E}_t)\left[\sum_{j=2}^{\infty} \kappa_1^{j-1}\left(g_{t+j}\right)\right] = \frac{\kappa_1 \Delta}{1-\kappa_1},$$

which can be much larger than Δ if κ_1 is close to one.[5]

A similar conclusion holds if the increase is not permanent, but highly persistent—say $\mathbb{E}_{t+1}(g_{t+j}) - \mathbb{E}_t(g_{t+j}) = \rho^{j-1}\Delta$ for $j = 2, 3, \ldots$, which ρ near one. This is, in fact, the assumption made by Bansal and Yaron—in particular, they assume that log consumption growth has both an *i.i.d.* component and a

[5] Recall that κ_1 was a Taylor series coefficient in the Campbell-Shiller approximation. It's given by $1/(1+\exp(-\bar{z}))$, where $\bar{z}$ is the mean of z. Since prices are typically at least a few multiples of current dividend, it's reasonable to assume that κ_1 will in fact be very close to one.

persistent component. They model dividends as sharing the persistent component in consumption growth, but subject to larger fluctuations.

5.1.2 The model

Barsky-DeLong and Bansal-Lundblad are not equilibrium models. The challenge that Bansal and Yaron take on is to write down a consumption-based equilibrium model that incorporates the mechanism described above, whereby small persistent movements in expected dividend growth rates cause large movements in price-dividend ratios. Such a model needs to have two features: (1) the change in expected growth rate has to dominate any resulting change in the expected return; (2) the model's stochastic discount factor has to price the expected growth rate risk. This leads them to a model with Epstein-Zin preferences with an elasticity of intertemporal substitution greater than one.

Why an EIS greater than one? It's not reasonable to expect that a change in the conditional mean growth rate of dividends, in (5.5), will not also entail a change in expected returns. The assumption of an EIS greater than one is a way of guaranteeing the net effect is still positive. The intuition for this can be had by considering the long-run relationship between the consumption growth rate and the real interest rate under certainty:

$$r = -\log(\beta) + \frac{1}{\psi} g \tag{5.6}$$

where ψ is the elasticity of intertemporal substitution, g is the log consumption growth rate, and r is the log, or continuous, real interest rate. Equation (5.6) is one we've seen several times before. It implies

$$\Delta r = \frac{1}{\psi} \Delta g,$$

so that

$$\Delta g - \Delta r = \left(1 - \frac{1}{\psi}\right) \Delta g$$

which is positive if $\psi > 1$.

The need for Epstein-Zin preferences will become shortly.

Bansal and Yaron assume the following process for log consumption growth, $\log(c_{t+1}/c_t) \equiv g_{t+1}$:

$$g_{t+1} = \nu + x_t + \sigma \eta_{t+1} \tag{5.7}$$
$$x_{t+1} = \rho x_t + \phi_e \sigma e_{t+1} \tag{5.8}$$

where η_{t+1} and e_{t+1} are both *i.i.d.* standard normal—*i.e.*, $N(0,1)$—variables. I've used ν for their μ, since we've already been using μ to denote the Epstein-Zin certainty equivalent operator. Thus, the log consumption growth rate is conditionally normal with conditional mean $\mathbb{E}_t(g_{t+1}) = \nu + x_t$ and constant conditional variance $\text{var}_t(g_{t+1}) = \sigma^2$. The process for x_t has an unconditional mean of zero; its conditional variance is $(\phi_e \sigma)^2$, and its unconditional variance—much larger if ρ is near one—is $(\phi_e \sigma)^2/(1-\rho^2)$.

They also model dividends separately from consumption, and price both claims to the consumption process and claims to the dividend process. Log dividend growth, $\log(d_{t+1}/d_t) \equiv g_{d,t+1}$, is assumed to obey

$$g_{d,t+1} = \nu_d + \phi x_t + \phi_d \sigma u_{t+1} \tag{5.9}$$

where u_{t+1} is $N(0,1)$ and independent of η_{t+1} and e_{t+1}. They will assume ϕ_d is large, so that dividend growth has a much higher conditional variance than consumption growth. ϕ will also be large, so innovations to x_t—the e_{t+1}'s—will have a larger impact on the conditional mean of dividend growth than on the conditional mean of consumption growth.[6]

Some features of the Bansal-Yaron process:

- ϕ_e is going to be calibrated so that the variance of x_{t+1} is small compared to σ^2, making the white noise component of consumption growth dominant over short horizons.

- ρ is going to be calibrated close to one, so that the fluctuations in the conditional mean $\mathbb{E}_t(g_{t+1}) = \nu + x_t$ will be highly persistent. This also means that the unconditional variance of x_{t+1} will be large.

- Our Mehra-Prescott process matches up with the *i.i.d.* part of g_{t+1}. To be sure, we modeled consumption growth as having a slight negative autocorrelation, but Mehra and Prescott's process is not that far from *i.i.d.*.

- The state variables—on which prices and the agent's value will depend—will be c_t (or d_t) and x_t. As before, homogeneity will allow us to divide out the dependence on levels. The *i.i.d.* shocks are not state variables—while the stochastic discount factor will be seen to depend on x_t, x_{t+1}, and η_{t+1}, the *i.i.d.* disturbance will integrate out when we take expectations.

- Bansal and Yaron also incorporate time-varying volatility (σ varies over time). Time permitting, we'll add that in after we examine the more basic model with constant volatility.

Let $p_c(x_t, c_t)$ denote the price of a claim to the consumption process, $p_d(x_t, d_t)$ the price of a claim to the dividend process, and $q(x_t)$ the price of a riskless claim to one unit of consumption next period. Using the price of the consumption claim as an example, the assets are priced in the by-now-familiar way

$$p_c(x_t, c_t) = \mathbb{E}_t \left[m_{t+1} \left(p_c(x_{t+1}, c_{t+1}) + c_{t+1} \right) \right]. \tag{5.10}$$

[6] Aggregate dividends are in fact much more volatile than aggregate consumption.

With Epstein-Zin preferences, we'll continue to obtain our homogeneity result, that we may write $p_c(x_t, c_t) = w_c(x_t)c_t$ for some function $w_c(x_t)$, the price-dividend ratio for the consumption claim. Then,

$$\begin{aligned} w_c(x_t) &= \mathbb{E}_t\left[m_{t+1}\frac{c_{t+1}}{c_t}\left(w_c(x_{t+1})+1\right)\right] \\ &= \mathbb{E}_t\left[m_{t+1}e^{\nu+x_t+\sigma\eta_{t+1}}\left(w_c(x_{t+1})+1\right)\right] \qquad (5.11) \\ &= e^{\nu+x_t}\mathbb{E}_t\left[m_{t+1}e^{\sigma\eta_{t+1}}\left(w_c(x_{t+1})+1\right)\right] \qquad (5.12) \end{aligned}$$

where (5.11) uses the fact that $c_{t+1}/c_t = e^{g_{t+1}}$, and (5.12) relies on the fact that x_t can be taken outside the date-t conditional expectation.

Bansal and Yaron solve their model using the return form of the Epstein-Zin stochastic discount factor, as we derived in section 4.2.4, equation (4.37). They describe their results in terms of log-linear approximations of the Campbell-Shiller type, though the computational method they use to obtain the numbers in their tables is a polynomial projection method.[7]

I prefer to examine the model using the Epstein-Zin stochastic discount factor in its form

$$m_{t+1} = \beta\left(\frac{c_{t+1}}{c_t}\right)^{-\frac{1}{\psi}}\left(\frac{v_{t+1}}{\mu_t\left(v_{t+1}\right)}\right)^{\frac{1}{\psi}-\alpha}$$

where I've written $1/\psi$ for our previous $\rho - 1$, both to make the dependence on the value of the EIS more explicit, and to avoid confusion with the persistence parameter ρ from (5.8).

The (equilibrium) lifetime utility process v_t still obeys the recursion

$$v_t = \left[(1-\beta)c_t^{1-\frac{1}{\psi}} + \beta\mu_t\left(v_{t+1}\right)^{1-\frac{1}{\psi}}\right]^{\frac{1}{1-\frac{1}{\psi}}},$$

where c_t is equilibrium consumption. The only variables at t that are informative about $t+1$ and beyond are the level of consumption, c_t, and the persistent part of consumption growth, x_t. Moreover, nothing that Bansal and Yaron have added to the basic Epstein-Zin model changes the essential homogeneity properties of the model. Thus, we can write

$$v_t = \Phi(x_t)c_t$$

for some function Φ.

Taking account of this homogeneity, we can divide c_t out of the utility pro-

[7]We don't have time to go into polynomial methods, but this model is well-suited for that approach. Judd [Jud98] discusses computationally efficient methods for solving dynamic models using polynomial projections.

cess to get

$$\Phi(x_t) = \left[1-\beta+\beta\mu_t\left(\Phi(x_{t+1})\frac{c_{t+1}}{c_t}\right)^{1-\frac{1}{\psi}}\right]^{\frac{1}{1-\frac{1}{\psi}}}$$
$$= \left[1-\beta+\beta\mu_t\left(\Phi(x_{t+1})e^{\nu+x_t+\sigma\eta_{t+1}}\right)^{1-\frac{1}{\psi}}\right]^{\frac{1}{1-\frac{1}{\psi}}}$$
$$= \left[1-\beta+\beta e^{\left(1-\frac{1}{\psi}\right)(\nu+x_t)}\mu_t\left(\Phi(x_{t+1})e^{\sigma\eta_{t+1}}\right)^{1-\frac{1}{\psi}}\right]^{\frac{1}{1-\frac{1}{\psi}}} \quad (5.13)$$

Likewise, the stochastic discount factor can be written as

$$m_{t+1} = \beta\left(e^{\nu+x_t+\sigma\eta_{t+1}}\right)^{-\frac{1}{\psi}}\left(\frac{\Phi(x_{t+1})e^{\sigma\eta_{t+1}}}{\mu_t\left(\Phi(x_{t+1})e^{\sigma\eta_{t+1}}\right)}\right)^{\frac{1}{\psi}-\alpha} \quad (5.14)$$

Note that both (5.13) and (5.14) involve $\mu_t\left(\Phi(x_{t+1})e^{\sigma\eta_{t+1}}\right)$. Because η_{t+1} is independent from x_{t+1}, we can split the certainty equivalent of $\Phi(x_{t+1})e^{\sigma\eta_{t+1}}$ as

$$\mu_t\left(\Phi(x_{t+1})e^{\sigma\eta_{t+1}}\right) = \mu\left(e^{\sigma\eta_{t+1}}\right)\mu_t\left(\Phi(x_{t+1})\right).$$

This follows from the fact—which you can verify from the definitions of μ and statistical independence—that if y and z are independent, then $\mu(yz) = \mu(y)\mu(z)$. The certainty equivalent of $e^{\sigma\eta_{t+1}}$, using the rules for expectations of lognormal random variables,[8] is

$$\mu(e^{\sigma\eta_{t+1}}) = e^{\frac{1}{2}(1-\alpha)\sigma^2}.$$

Applying these results to (5.13) and (5.14), after some algebra, we can derive the two expressions that will be the basis for our computational solution of the model:

$$m_{t+1} = m_0 e^{-\frac{1}{\psi}x_t}e^{-\alpha\sigma\eta_{t+1}}\left(\frac{\Phi(x_{t+1})}{\mu_t\left(\Phi(x_{t+1})\right)}\right)^{\frac{1}{\psi}-\alpha} \quad (5.15)$$

$$\Phi(x_t) = \left[1-\beta+\beta e^{(1-\frac{1}{\psi})\left(\nu+x_t+(1/2)(1-\alpha)\sigma^2\right)}\mu_t\left(\Phi(x_{t+1})\right)^{1-\frac{1}{\psi}}\right]^{\frac{1}{1-\frac{1}{\psi}}} \quad (5.16)$$

where the m_0 in the expression for the stochastic discount factor collects together several constants, $m_0 = \beta\exp\left\{-(1/\psi)\nu + (1/2)(\alpha-(1/\psi))(1-\alpha)\sigma^2\right\}$.

From (5.16), we can see that $\Phi(x_t)$ is increasing in x_t.[9] Also, the effect on $\Phi(x_t)$ of a given increase in x_t is bigger the more persistent the x process is—*i.e.*, the closer ρ is to one.[10]

[8] If $z \sim N\left[\mathbb{E}(z), \mathrm{var}(z)\right]$, then $\mathbb{E}[e^z] = \exp\left(\mathbb{E}(z) + \frac{1}{2}\mathrm{var}(z)\right)$.

[9] This should be intuitive. Because of the positive autocorrelation in x, an increase in x_t raises future growth rates (conditionally), and a higher growth rate must produce higher lifetime utility

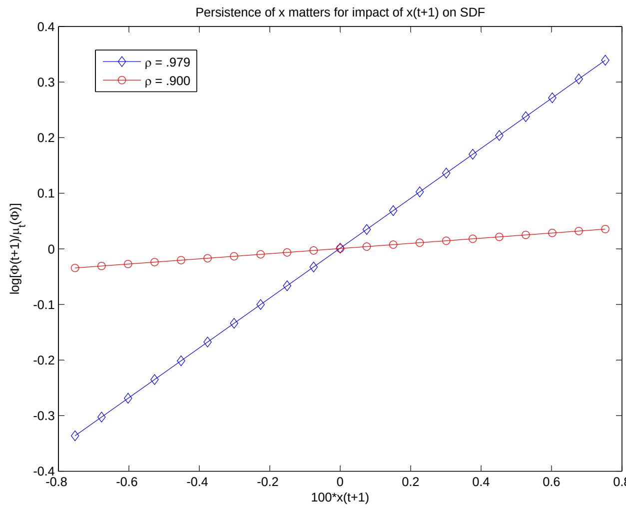

Figure 5.4: The figure plots $\log\left(\Phi(x_{t+1})/\mu_t(\Phi(x_{t+1}))\right)$ for $\rho = .979$ and $\rho = .9$, illustrating the role of persistence in the pricing of x_{t+1} risk. Φ is solved for using the Markov chain method described in the next section. The parameters are as in Bansal and Yaron

All of this tells how the stochastic discount factor given by (5.15) is going to price risk associated with innovations to the conditional mean of of the consumption growth rate. The *i.i.d.* risk is priced by the $e^{-\alpha\sigma\eta_{t+1}}$ term, which is standard—a high growth realization makes the agent less 'hungry' (in Cochrane's terminology), so assets whose payoffs covary positively with η_{t+1} are less valuable than assets whose payoffs covary negatively with η_{t+1}. The key parameter here is the risk aversion parameter α.

The x_{t+1} risk—risk associated with the innovations e_{t+1} in $x_{t+1} = \rho x_t + \phi_e \sigma e_{t+1}$—is priced by the scaled value function $\Phi(x_{t+1})$. A positive realization of e_{t+1} raises the conditional mean rate of consumption growth x_{t+1}, generating an increase in $\Phi(x_{t+1})$; the increase is larger the higher is the degree of persistence in the x process. If $\frac{1}{\psi} - \alpha < 0$, which it will be if $\psi > 1$ and $\alpha > 0$, a positive innovation to x_{t+1} makes the agent less 'hungry'. Assets whose payoffs covary positively with x_{t+1} will require a risk premium to convince the agent to hold them in equilibrium. This characterizes both the consumption

per unit of consumption today (which is what $\Phi(x_t)$ measures). Proving it is a bit more subtle and involves treating (5.16) as a mapping that (one can show) preserves monotonicity. And, the limit of a sequence of increasing functions is, at the least, a nondecreasing function. Note that all this is true independent of whether $\psi > 1$ or not.

[10]At the other extreme ($\rho = 0$), the x process is *i.i.d.*, and $\mu_t\left(\Phi(x_{t+1})\right) = \mu\left(\Phi(x_{t+1})\right)$ is a constant for all t. In this case, the effect of changes in x_t on $\Phi(x_t)$ is fully captured by the $\exp\left(\left(1 - \frac{1}{\psi}\right) x_t\right)$ term in (5.16).

and dividend claims, under the assumption that $\psi > 1$.

We can write the return on the consumption claim as

$$R^c_{t+1} = \frac{p_c(x_{t+1}, c_{t+1}) + c_{t+1}}{p_c(x_{t+1}, c_{t+1})} = e^{\nu + x_t + \sigma\eta_{t+1}} \frac{w_c(x_{t+1}) + 1}{w_c(x_t)} \tag{5.17}$$

Under the assumption that $\psi > 1$, for the reasons described above at the start of this section, the consumption price-dividend ratio $w_c(x_{t+1})$ will be increasing in x_{t+1}, so the consumption claim is exposed to the x_{t+1} risk. Similarly, the return on the dividend claim is

$$R^d_{t+1} = e^{\nu_d + \phi x_t + \phi_d \sigma u_{t+1}} \frac{w_d(x_{t+1}) + 1}{w_d(x_t)}. \tag{5.18}$$

The dividend claim is exposed to the x_{t+1} as well, through $w_d(x_{t+1})$. Note that u_{t+1}, the *i.i.d.* part of the dividend claim's return, is independent of everything in the SDF m_{t+1}. This means that the *i.i.d.* risk in dividend growth is *not* priced—if the only risk associated with the dividend claim were from u_{t+1}, then the dividend claim's return would equal the risk-free rate.

5.1.3 Computation

As mentioned above, introducing polynomial methods is beyond the scope of this course, so our computational approach here will be similar to ones we've used before—namely, discretizing the state space using Markov chains. In particular, we're going to use a Markov chain to approximate the process for x_t, given in (5.8). Let $(x(1), x(2), \dots x(S); P)$ denote the Markov chain.

As before, under the Markov chain assumption, the stochastic discount factor m will be a matrix, while the scaled value function Φ, and the prices w_c, w_d and q, will be vectors. The certainty equivalent of Φ conditional on $x = x(i)$ today is

$$\mu_i(\Phi) = \left[\sum_{j=1}^{S} P(i,j)\Phi(j)^{1-\alpha} \right]^{\frac{1}{1-\alpha}}.$$

Φ itself we can solve for iteratively, using the Bellman-like equation

$$\Phi(i) = \left[1 - \beta + \beta e^{(1-\frac{1}{\psi})(\nu + x(i) + (1/2)(1-\alpha)\sigma^2)} \left[\sum_{j=1}^{S} P(i,j)\Phi(j)^{1-\alpha} \right]^{\frac{1-\frac{1}{\psi}}{1-\alpha}} \right]^{\frac{1}{1-\frac{1}{\psi}}} \tag{5.19}$$

The next step is to evaluate the pricing relations

$$\begin{aligned} w_c(x_t) &= \mathbb{E}_t\left[m_{t+1} e^{\nu + x_t + \sigma\eta_{t+1}} \left(w_c(x_{t+1}) + 1 \right) \right] \\ w_d(x_t) &= \mathbb{E}_t\left[m_{t+1} e^{\nu + \phi x_t + \phi_d \sigma u_{t+1}} \left(w_d(x_{t+1}) + 1 \right) \right] \\ q(x_t) &= \mathbb{E}_t[m_{t+1}] \end{aligned}$$

One thing to note is that, because u_{t+1} and η_{t+1} are independent of x_{t+1}, the terms involving these innovations (including the η_{t+1} term in m_{t+1}) can be collected together, and their expectations evaluated separate from the terms in x_{t+1}. Under the Markov chain assumption, these relations then become—using (5.15)—

$$w_c(i) = A_c e^{(1-\frac{1}{\psi})x(i)} \sum_{j=1}^{S} P(i,j) \left(\frac{\Phi(j)}{\mu_i(\Phi)} \right)^{\frac{1}{\psi}-\alpha} (w_c(j)+1) \tag{5.20}$$

$$w_d(i) = A_d e^{(\phi-\frac{1}{\psi})x(i)} \sum_{j=1}^{S} P(i,j) \left(\frac{\Phi(j)}{\mu_i(\Phi)} \right)^{\frac{1}{\psi}-\alpha} (w_d(j)+1) \tag{5.21}$$

$$q(i) = A_q e^{-\frac{1}{\psi}x(i)} \sum_{j=1}^{S} P(i,j) \left(\frac{\Phi(j)}{\mu_i(\Phi)} \right)^{\frac{1}{\psi}-\alpha} \tag{5.22}$$

where A_c, A_d, and A_q collect together various constant terms:

$$\begin{aligned} A_c &= m_0 e^{\nu} \mathbb{E}\left[e^{(1-\alpha)\sigma\eta_{t+1}}\right] \\ A_d &= m_0 e^{\nu_d} \mathbb{E}\left[e^{-\alpha\sigma\eta_{t+1}}\right] \mathbb{E}\left[e^{\phi_d\sigma u_{t+1}}\right] \\ A_q &= m_0 \mathbb{E}\left[e^{-\alpha\sigma\eta_{t+1}}\right] \end{aligned}$$

With all this machinery in hand, it's straightforward to put the model on the computer and see what it spits out. At this point, you should be familiar enough with MATLAB and models of this sort to put together a program that solves the model as described above. Most of the parameters, I'll take from Bansal and Yaron.[11] We still need to specify the Markov chain, though.

Rather than use a two-state Markov chain, as we've been doing, I wanted the flexibility to add more states, so I used Rouwenhorst's method.[12] I experimented with the number of states ranging from 3 to 101—adding any more than around 21, though, didn't change the results much at all.

5.1.4 Calibration and results

You should consult Bansal and Yaron for the justifications they give for their parameter choices. They calibrate the model to a monthly frequency—that's new for us–and the parameters they pick for the consumption, long-run risk, and dividend processes in their basic version of the model are given in Table 5.1.

[11] An exercize will ask you to solve the model with some parameters estimated by Constantinides and Ghosh [CG11], available here: http://faculty.chicagobooth.edu/george.constantinides/documents/AssetPricingTestsKEEPAugust11_11.pdf

[12] I've mentioned Rouwenhorst's [Rou95] method before in passing. In a way, it just extends the technique we learned for two states to many states. The invariant distribution that results is exactly binomial, so with enough states, it approximates a normal distribution. Kopecky and Suen [KS10] describe the method, and show that it does a good job of approximating very persistent $AR(1)$ processes.

Table 5.1: Parameters for stochastic consumption, long-run risk, and dividend processes in Bansal-Yaron's basic model.

ν	σ	ρ	ϕ_e	ν_d	ϕ	ϕ_d
0.0015	0.0078	0.979	0.044	0.0015	3.0	4.5

I approximated the $AR(1)$ for x_t with a Markov chain—using Rouwenhorst's method—with 31 states.

The preferences parameters β, ψ and α, we'll play around with a bit. Bansal and Yaron report some results for α either 7.5 or 10, and ψ either 1.5 or 0.5, with $\beta = 0.998$. I have a hard time getting the riskless rate as low as they report with $\beta = 0.998$, and do much better with $\beta = 0.999$—almost exactly matching their numbers, in fact. This could be for a couple reasons—one, we're using different approximation methods, and two, I'm not quite sure I'm getting the time aggregation right.[13] In any case, the results here are for $\beta = 0.999$. That's relatively high—it annualizes to 0.988.

Like Bansal and Yaron, I'll report results for the risk-free rate and the return on the dividend claim (what they refer to as the market return) rather than the aggregate consumption claim. I solved for Φ iteratively from (5.19), then solved the pricing equations (5.20)–(5.22), and then formed returns according $R^F = 1/q$, and R^d given by (5.18). I focused on log returns, $\log(R^F_t) = -\log(q_t)$, and $\log(R^d_{t+1}) = \nu_d + \phi x_t + \phi_d \sigma u_{t+1} + \log(w_{d,t+1} + 1) - \log(w_{d,t})$. I also looked at the properties of $\log(w_{d,t})$, the log price-dividend ratio for the dividend claim.

Note that, when you put these in terms of the Markov chain, $\log(R^F_t) = \log\left(R^F(i)\right) = -\log\left(q(i)\right)$ and $\log\left(w_{d,t}\right) = \log\left(w_d(i)\right)$. The log return on the dividend claim has both a Markov chain piece, $\nu_d + \phi x(i) + \log\left(w_d(j) + 1\right) - \log\left(w_d(i)\right)$ and an *i.i.d.* piece, $\phi_d \sigma u_{t+1}$. For the expectation of $\log(R^d_{t+1})$, the *i.i.d.* piece is zero. If π^* is the invariant probability distribution of the Markov transition matrix P, then

$$\mathbb{E}\left[\log\left(R^F\right)\right] = \sum_i \pi^*(i) \log\left(R^F(i)\right),$$

$$\mathbb{E}[\log(w_d)] = \sum_i \pi^*(i) \log\left(w_d(i)\right),$$

[13] If you just want to *annualize* a monthly rate, you either multiply it by 12 (if it's a log rate) or raise the gross monthly rate to the 12th power, and subtract one. If you want the *average for an annual period*, then, if we're working with log rates, and they're *i.i.d.*, the annual mean and variance are both 12 times the monthly mean and variance. If they're not *i.i.d.*, then it's more complicated, but I didn't really have the time to work it out before class. The issue may be the approximation, though. They do have another paper [BKY07], which employs another solution technique, where they too assume β is higher—$\beta = 0.9989$.

and

$$\mathbb{E}\left[\log(R^d)\right] = \sum_i \pi^*(i) \sum_j P(i,j) \left\{ \nu_d + \phi x(i) + \log\left(\frac{w_d(j)+1}{w_d(i)}\right) \right\}$$

For standard deviations, computing $s.d.(\log(R^F))$ and $s.d.(\log(w_d))$ is simple: we just use the invariant probabilities π^* and the means we calculated in the previous step. For $\log(R_d)$, we take the square root of the sum of (1) the variance of the Markov chain part (using P and π^*) and (2) the variance of the $i.i.d.$ part, $(\phi_d \sigma)^2$.

To put the return-related quantities into annual average percent terms we multiply the means by 1200 and the standard deviations by $\sqrt{12} \times 100$.[14] We leave $\log(w_d)$ in monthly units. The results are in Table 5.2; lower-case letters denote logs, and $p_d - d$ denotes $\log(w_d)$.

Table 5.2: Some results for the long-run risk model. Parameters for consumption, long-run risk, and dividends are as in Table 5.1. For all cases here, $\beta = .999$.

α	ψ	$\mathbb{E}[r^d - r^F]$	$\mathbb{E}[r^F]$	$s.d.(r^d)$	$s.d.(r^F)$	$s.d.(p_d - d)$
7.5	0.5	0.56	4.82	13.16	1.17	0.07
7.5	1.5	2.93	1.63	16.90	0.39	0.17
10.0	0.5	1.16	4.93	13.08	1.17	0.07
10.0	1.5	4.25	1.36	16.49	0.38	0.16
10.0	0.1	−5.6	15.55	33.95	5.83	0.44
10.0	2.0	4.68	0.88	17.05	0.29	0.17

Exercize 5.1. *Constantinides and Ghosh [CG11] estimated an annual version of the Bansal-Yaron model using GMM. One set of results they obtain (they try several specifications), estimating only the parameters of the stochastic processes, is given by:*

ν	σ	ρ	ϕ_e	ν_d	ϕ	ϕ_d
0.02	0.006	0.437	5.20	0.01	2.06	15.8

Write a MATLAB *program to solve the model—following the steps outlined above—and report results for the same quantities as in Table 5.2, for* $\alpha = 10$, $\psi = 1.5$ *and* $\beta = 0.99$.

They also do an estimation of the preference parameters and stochastic process parameters together. In this case, they obtain:

[14] See footnote 13.

β	α	ψ	ν	σ	ρ	ϕ_e	ν_d	ϕ	ϕ_d
0.968	9.34	1.41	0.021	0.012	0.482	0.90	0.018	5.14	3.06

Solve the model and report results for these parameters. Note since this is an annual model, you only need to multiply things by 100*, not* 1200 *or* $\sqrt{12} \times 100$.

Compare the results in both cases to row 4 of Table 5.2. What might explain the differences in the results you obtain? (Answer—a lot of things; just try to get some intuitive feel for how the model works.)

You should use Rouwenhorst's method for calibrating the Markov chain. I'll post the code on the website.

Optional. *Investigate the sensitivity of the results to the number of states in the Markov chain—if you tried* $n = 3, 5, 7 \ldots$*, is there an* n *above which the results stop changing by very much? (Note: using odd numbers of states guarantees that the mean of zero is one of the states of the Markov chain.)*

5.1.5 Time-varying volatility

We don't have time to go into this in any detail, but basically, the model as described above produces almost no variation in the conditional expected excess returns (across $x(i)$ states)—the conditionally expected dividend return and the risk-free return have, in effect, roughly the same 'slope' with respect to variation in x. Since the dividend return is conditionally *i.i.d.*, it's conditional standard deviation is constant at $\phi_d \sigma$. This means the model produces virtually no variation in the conditional Sharpe ratio.

As a remedy to this, Bansal and Yaron introduce time-varying volatility. In particular, the standard deviation σ in (5.7), (5.8) and (5.9) is replaced with a time-varying σ_t, the square of which is assumed to follow an $AR(1)$ process:

$$\sigma_{t+1}^2 = \sigma^2 + \rho_1(\sigma_t^2 - \sigma^2) + \sigma_\epsilon \epsilon_{t+1} \tag{5.23}$$

where it is assumed that ϵ_{t+1} is *i.i.d.*, $N(0, 1)$, and independent of the other *i.i.d.* innovations.[15]

In terms of the pricing framework we've developed, σ_t is now a second state variable. In particular, $\Phi(x_t)$ becomes $\Phi(x_t, \sigma_t)$, which is *decreasing* in σ_t, the more so the closer is the persistence parameter ρ_1 to one.[16] So the pricing kernel will now price volatility risk in the sense that a high σ_{t+1} realization means greater 'hunger' (if $1/\psi - \alpha < 0$).

Bansal and Yaron show that when $\psi > 1$, the model's price-dividend ratios are *decreasing* in σ_t, the more so the greater is the persistence of the volatility process. This means that the consumption and dividend claims are exposed to

[15]The standard deviation σ_ϵ must be chosen to be very small, since we don't want the process to violate $\sigma_t^2 > 0$. Technically speaking, the ϵ_{t+1}'s shouldn't really be assumed to be normal.

[16]Basically, more volatile consumption is less valuable given a concave certainty equivalent.

the volatility risk (they're worth less when the agent is more hungry), so the time-varying volatility raises their conditionally expected returns when σ_t is high.

Of course, variation in σ_t also affects the risk-free rate, but the effects are much smaller—thus conditionally expected excess returns vary positively with σ_t.

σ_t will also affect the conditional standard deviation of returns (positively, of course), so for this mechanism to produce a time-varying Sharpe ratio, the movements in conditionally expected excess returns need to be larger than the movements in the conditional standard deviation. That turns out to be the case for Bansal and Yaron's calibration of the model.

5.2 Rare consumption disasters

Rare consumption disasters were one of the first proposed potential resolutions of the equity premium puzzle—Thomas Rietz's "The equity risk premium: A solution" [Rie88] appeared in the *J.M.E.* just three years after Mehra and Prescott's paper. Rietz's idea was that equity claims are subject to occasional large crashes that are not well-captured by the Mehra-Prescott consumption process. In fact, writing the process, as Mehra and Prescott do, as a two-state Markov chain, the best and worst outcomes are just one standard deviation above and below the mean.

One could use a chain with more states, but techniques like Rouwenhorst's or Tauchen's assume normality of the innovations to the underlying $AR(1)$ process. Adding states at $\pm 2\sigma$, $\pm 3\sigma$, $\pm 4\sigma$, *etc.* will have little effect because the probabilities of being in those states will be negligibly small. Disasters matter only if the tails of the distribution of growth rates are fatter than normal.

If you read the background in Barro [Bar06] you'll see that Rietz's resolution was eventually dismissed because no one really thought the tails of the distribution were as fat as Rietz needed them to be to get his results. Rietz considered many combinations of disaster size (as the percent decline in aggregate consumption) and frequency. He found a number of combinations such that the model's first moment implications roughly match the data. The least extreme examples were for a roughly one in one hundred chance of a 25% decline in consumption; with risk aversion around ten and β very close to one, the model roughly reproduced the average risk-free rate and equity premium.

Another criticism of Rietz's model was that it assumed the risk-free asset would remain risk-free even in the event of disaster. This didn't seem right, as crises often lead to defaults on even normally riskless government debt (sometimes implicitly, through unexpected inflation).[17]

Two decades after Rietz, Barro revived interest in disaster models, not so much by offering anything new on the modeling front (though he did include partial default on the 'riskless' asset), but rather by documenting empirically

[17] Mehra and Prescott responded with all these criticisms in a subsequent issue of the *J.M.E* [MP88] with the cleverly-titled "The equity risk premium: A solution?".

that disasters may in fact be as large and as common as Rietz supposed. Given that we have only short time series for any one country, Barro uses data from many countries to characterize the probability of large declines in consumption (actually per capita GDP). You should read through his evidence. He sums up his findings as implying a disaster probability of between 1.5–2.0% with associated declines in per capita GDP ranging from 15% to 64%. Wars account for many of the disaster episodes he catalogs.

Barro's analysis was subsequently refined by Gourio [Gou08], Gabaix [Gab08], and others. The model we'll look at in this section is the one presented in Gourio's paper. First, though, we'll look at an approach closer to Rietz's.

5.2.1 Modeling rare disasters

Disasters are actually quite easy to incorporate into our basic model. For example, consider augmenting the Mehra-Prescott process in the following way. In either 'normal' state $i = 1, 2$, the distribution of next period's growth rate is given by: with probability $1 - f$, x_{t+1} is distributed according to $P(i, \{1, 2\})$—*i.e.*, $x_{t+1} = x(j)$ with probability $P(i, j)$; and, with probability f, $x_{t+1} = x^d$, the disaster outcome, where $x^d \lll \{x(1), x(2)\}$.[18] And, if the current state is x^d, next period's state is either $x(1)$ or $x(2)$, with probabilities given by the invariant distribution of P—in our case, $\{\frac{1}{2}, \frac{1}{2}\}$.

This is similar to Rietz's approach. It implies a three-state Markov chain for x_{t+1}, given by

$$x \in \{x(1), x(2), x^d\} \equiv X$$

and

$$\Pi = \begin{bmatrix} (1-f)P(1,1) & (1-f)P(1,2) & f \\ (1-f)P(2,1) & (1-f)P(2,2) & f \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

One can show that with a symmetric P, the invariant distribution of Π is given by

$$\pi^* = \left(\frac{1}{2(1+f)}, \frac{1}{2(1+f)}, \frac{f}{(1+f)} \right).$$

Given the new Markov chain (X, Π), one just proceeds to solve the model exactly as before, as we did in the basic Mehra-Prescott model.

Remark 5.1. *Actually, what Rietz did was to recalibrate P and $\{x(1), x(2)\}$ for each choice of f and x^d, so as to guarantee that the three-state Markov chain (X, Π) has mean, standard deviation and first-order autocorrelation consistent with Mehra-Prescott's estimates of $E(x) = 1.018$, s.d.$(x) = 0.036$, and $AC_1(x) = -0.14$. This has always struck me as the wrong approach. Apart from the Great Depression—which is big, but not 'a-25%-decline' big— the Mehra-Prescott numbers are estimated for an economy operating in 'normal' times. So I think it's better to use those estimates for*

[18]Note that since we're moving on from Bansal and Yaron, we'll go back to letting x_{t+1} denote the gross rate of growth of consumption from t to $t + 1$.

the P and $\{x(1), x(2)\}$ as I've put them in the three-state chain above—they describe the behavior of growth conditional on 'no disaster'.

Maybe it's intuitive that disaster risk (which affects the payoff from a claim to aggregate consumption) raises the equity premium. How, though, does it keep the risk-free rate at a reasonable level, as Rietz was able to achieve?

If, as I want to suppose, P and $\{x(1), x(2)\}$ correspond to the Mehra-Prescott process, then you can see immediately that disasters will lower the risk-free rate—it will be lower than in the comparable Mehra-Prescott economy with the same α and β.

If there is no default on the riskless asset in a disaster, then its price in state $i = 1, 2$ will be

$$\begin{aligned} q(i) &= \mathbb{E}_i\,(m) \\ &= (1-f)\sum_{j=1}^{2} P(i,j)\beta x(j)^{-\alpha} + f\beta(x^d)^{-\alpha} \\ &= (1-f)q^{\mathrm{MP}}(i) + f\beta(x^d)^{-\alpha} \end{aligned} \tag{5.24}$$

where $q^{\mathrm{MP}}(i)$ is the riskless asset price from the corresponding Mehra-Prescott economy. If $\beta(x^d)^{-\alpha}$ is large, even a small f can greatly increase $q(i)$ compared to $q^{\mathrm{MP}}(i)$, and thus lower $R^F(i) = 1/q(i)$ in the 'normal states'.

In state 3, the disaster state,

$$q(3) = \frac{1}{2}\beta x(1)^{-\alpha} + \frac{1}{2}\beta x(2)^{-\alpha}$$

which one can show is between $q^{\mathrm{MP}}(1)$ and $q^{\mathrm{MP}}(2)$.

To incorporate default on the riskless asset, we would no longer assume it pays one unit in every state, but rather $(1, 1, 1-d)$ in states 1, 2, and 3. The parameter $d \in [0, 1]$ can be thought of as either (a) in a disaster, the issuer of the riskless claim only pays d, or (b) conditional on a disaster, the issuer defaults entirely with probability d (and with probability $1-d$ pays one unit).

It turns out that, in the Rietz-like model described here, adding default on the normally riskless asset harms the predictions of the model—unless d is very small, it will greatly increase the risk-free rate. In terms of equation (5.24), as $d \to 1$, it's essentially eliminating the $f\beta(x^d)^{-\alpha}$ term that was holding the risk-free rate down. The next exercize asks you to explore some of these phenomena.

Exercize 5.2. *This should be simple, given that you know how to solve the basic Mehra-Prescott model, as in Exercize 3.3. Use the Mehra-Prescott process for $\{x(1), x(2); P\}$. Write a* MATLAB *program and solve the model described above for the following three sets of parameters:*

α	β	f	x^d	d
9	.97	0	0	0
9	.97	0.01	0.75	0
9	.97	0.01	0.75	0.30

Report the average (unconditional mean) equity premium, risk-free rate and equity return for each combination (in percent terms). Line one of the table is just the basic Mehra-Prescott model. You should find that results look pretty good for the middle row, but bad for the first and last.

Now, keeping f, x^d, and d as in the last row of the table, explore the effects of varying α and β. Are there any combinations that will get your results back in the range of the nice ones you obtained for the middle configuration? Document what you find, maybe with a table or graph. I don't know the answer to this part, so I'm curious to see what you find.

Gourio's approach simplifies this in one dimension and complicates it in another. Making things simpler, Gourio assumes that consumption growth in non-disaster periods is *i.i.d.*. In the event of a disaster, consumption growth has both an *i.i.d.* component and a (negative) disaster component. This eliminates the need to model the escape from the disaster state to the set of 'normal' states (the third row in the Rietz-style Π above). Complicating things a bit, Gourio assumes that the probability of a disaster varies over time. Given that the Rietz (or Barro) models can match the unconditional means of the equity and riskless asset returns, Gourio is exploring whether a model with time-varying disaster risk can match other features of the data—in particular the volatility of the equity price-dividend ratio, and the price-dividend ratio's predictive power for returns and excess reutnrs.

For preferences, Gourio considers both the standard case (time-additively separable) and Epstein-Zin. He touches only briefly on the issue of default on the riskless asset.

Formally, the consumption growth process is given in logs by[19]

$$\log(c_{t+1}/c_t) = \begin{cases} \nu + \sigma\eta_{t+1} & \text{with probability } 1 - f_t, \\ \nu + \sigma\eta_{t+1} + \log(1-b) & \text{with probability } f_t \end{cases}$$

where $\eta_{t+1} \sim N(0,1)$ and $b \in (0,1)$ is the size of the disaster. The disaster probability f_t is assumed to follow a first-order Markov process with transition probabilities $P(f_t, f_{t+1})$—the probability of going from f_t to f_{t+1}—and such that $f_t \in [\underline{f}, \overline{f}]$ for all t. In one case, Gourio assumes the Markov process is actually *i.i.d.*—$P(f_t, f_{t+1}) = P(f_{t+1})$. Consistent with the way I've written the transition probability, we'll be treating the Markov process as a Markov chain throughout our discussion of the model.

[19] We modify the notation a bit compared to Gourio.

5.2.2 Standard preferences

The time-varying disaster probability doesn't add that much complication when preferences take the standard form,

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t \frac{c_t^{1-\alpha}}{1-\alpha}\right].$$

The stochastic discount factor is

$$m_{t+1} = \beta\left(\frac{c_{t+1}}{c_t}\right)^{-\alpha},$$

or

$$m_{t+1} = \begin{cases} \beta e^{-\alpha(\nu+\sigma\eta_{t+1})} & \text{if there is no disaster at } t+1 \\ (1-b)^{-\alpha}\beta e^{-\alpha(\nu+\sigma\eta_{t+1})} & \text{if there is a disaster at } t+1 \end{cases}$$

Conditional on the probability of disaster f_t, the price of the riskless asset—call it $q(f_t)$—is easily computed as

$$\begin{aligned} q(f_t) &= \mathbb{E}[m_{t+1} : f_t] \\ &= (1-f_t)\beta e^{-\alpha\nu+(1/2)(\alpha\sigma)^2} + f_t(1-b)^{-\alpha}\beta e^{-\alpha\nu+(1/2)(\alpha\sigma)^2} \\ &= \left(1-f_t+f_t(1-b)^{-\alpha}\right)\beta e^{-\alpha\nu+(1/2)(\alpha\sigma)^2} \end{aligned}$$

so the log risk-free rate is given by

$$\log R_t^F = -\log(\beta) + \alpha\nu - \frac{1}{2}(\alpha\sigma)^2 - \log\left(1-f_t+f_t(1-b)^{-\alpha}\right). \tag{5.25}$$

Note that because $0 < b < 1$, $(1-b)^{-\alpha}$ is bigger than one, and possibly very large—with, say, $b = 0.25$ and $\alpha = 10$, $(1-b)^{-\alpha} \approx 18$. For f_t in the neighborhood of 0.01, $\log\left(1-f_t+f_t(1-b)^{-\alpha}\right) \approx f_t\left((1-b)^{-\alpha}-1\right)$, so small changes in f_t will have large consequences for the log risk-free rate.[20]

Likewise, the price-dividend ratio for a claim to aggregate consumption—call it $w_c(f_t)$—obeys

$$\begin{aligned} w_c(f_t) &= \mathbb{E}\left[m_{t+1}\frac{c_{t+1}}{c_t}\left(w_c(f_{t+1})+1\right) : f_t\right] \\ &= \left(1-f_t+f_t(1-b)^{1-\alpha}\right)\beta e^{(1-\alpha)\nu+(1/2)((1-\alpha)\sigma)^2}\mathbb{E}(w_c(f_{t+1})+1 : f_t) \end{aligned} \tag{5.26}$$

Note that in obtaining this expression (and the expression for $q(f_t)$), we're taking expectations with respect to several things that, conditional on f_t, are

[20] Remember, you'd multiply $\log(R_t^F)$ by 100 to put it in percent terms. For the b and α mentioned in the text, a 0.01 increase in f_t, say from 0.01 to 0.02, subtracts roughly 17 percentage points off the log risk-free rate. Assuming that you can get the level right at some value of f, you would want f_t to have a very small variance around that value.

independent—expectations with respect to η_{t+1}, the occurrence (or not) of a disaster, and the new probability of a disaster given the current probability. See footnote 6 in Gourio.

An important feature of (5.26) is that we can re-arrange it to give

$$\mathbb{E}\left(\frac{w_c(f_{t+1})+1}{w_c(f_t)} : f_t\right) = \frac{1}{\left(1-f_t+f_t(1-b)^{1-\alpha}\right)\beta e^{(1-\alpha)\nu+(1/2)((1-\alpha)\sigma)^2}} \tag{5.27}$$

This is useful because, given the assumptions about the consumption process, the expected return on the consumption claim (conditional on f_t) obeys

$$\mathbb{E}(R^c_{t+1} : f_t) = \mathbb{E}\left(\frac{c_{t+1}}{c_t}\right)\mathbb{E}\left(\frac{w_c(f_{t+1})+1}{w_c(f_t)} : f_t\right)$$

where

$$\begin{aligned}\mathbb{E}\left(\frac{c_{t+1}}{c_t}\right) &= (1-f_t)e^{\nu+(1/2)\sigma^2} + f_t(1-b)e^{\nu+(1/2)\sigma^2} \\ &= \left(1-f_t+f_t(1-b)\right)e^{\nu+(1/2)\sigma^2}\end{aligned}$$

Thus, in this simple version of the model, we can evaluate the conditionally expected equity return without having to solve for the price-dividend ratios. It is

$$\begin{aligned}\mathbb{E}(R^c_{t+1} : f_t) &= \frac{\left(1-f_t+f_t(1-b)\right)e^{\nu+(1/2)\sigma^2}}{\left(1-f_t+f_t(1-b)^{1-\alpha}\right)\beta e^{(1-\alpha)\nu+(1/2)((1-\alpha)\sigma)^2}} \\ &= \beta^{-1}e^{\alpha\nu+(1/2)\sigma^2-(1/2)((1-\alpha)\sigma)^2}\left(\frac{1-f_t+f_t(1-b)}{1-f_t+f_t(1-b)^{1-\alpha}}\right)\end{aligned}$$

In log terms,

$$\begin{aligned}\log\mathbb{E}(R^c_{t+1} : f_t) = -\log(\beta) + \alpha\nu + \frac{1}{2}\sigma^2 - \frac{1}{2}((1-\alpha)\sigma)^2 \\ + \log\left(\frac{1-f_t+f_t(1-b)}{1-f_t+f_t(1-b)^{1-\alpha}}\right)\end{aligned} \tag{5.28}$$

If you differentiate the term inside the last $\log(\cdot)$ you'll find that the expected return is increasing in f_t iff $\alpha > 1$.

Combining (5.25) and (5.28), we can immediately obtain an expression for the conditional equity premium, in log terms:

$$\log\left(\frac{\mathbb{E}(R^c_{t+1})}{R^F_t} : f_t\right) = \alpha\sigma^2 + \log\left(\frac{\left(1-f_t+f_t(1-b)^{-\alpha}\right)\left(1-f_t+f_t(1-b)\right)}{1-f_t+f_t(1-b)^{1-\alpha}}\right) \tag{5.29}$$

Not surprisingly, it's a typical *i.i.d.* piece ($\alpha\sigma^2$) à la Mehra-Prescott, plus a disaster-related piece. Using (5.29), we can explore the effects of variation in f_t on the conditional equity premium (as well as the roles of b and α).

Gourio's Proposition 2 proves establishes some results on how magnitudes in the model vary with f_t: the risk-free rate is decreasing in f_t; the equity return are increasing in f_t iff $\alpha > 1$; the price-dividend ratio is increasing in f_t iff $\alpha > 1$; the equity premium is increasing in f_t if f_t is small.

Thinking about time series behavior, there is a problem with the directions of change proven in the proposition—since we'll likely be assuming $\alpha > 1$, the proposition describes a model where the conditional equity return and conditional excess return covary positively with the price-dividend ratio, which is the opposite of what's found in the data.

This leads Gourio to examine the behavior of the model using Epstein-Zin preferences, to which we now turn.

5.2.3 Epstein-Zin preferences, *i.i.d.* disaster risk

Gourio initially specializes the process for the disaster probability f_t to be *i.i.d.*—$P(f_t, f_{t+1}) = P(f_{t+1})$. It's still a state variable that will affect current prices, but the *i.i.d.* assumption means that the current realization is not informative about future realizations. Thinking about Epstein-Zin preferences, this mean that the certainty equivalent of *scaled* lifetime utility from tomorrow onward—that is, v_{t+1} once we divide out the level of current consumption, or what we've been calling Φ_{t+1}—will be a constant.

Specifically, the (equilibrium) lifetime utility of the representative agent, normalized by c_t, will follow:

$$\Phi_t = \left[1 - \beta + \beta\mu\left(\Phi_{t+1}\frac{c_{t+1}}{c_t}\right)^{1-\frac{1}{\psi}}\right]^{\dfrac{1}{1-\frac{1}{\psi}}}, \tag{5.30}$$

where

$$\begin{aligned}\mu\left(\Phi_{t+1}\frac{c_{t+1}}{c_t}\right) &= \mu\left(\Phi_{t+1}e^{\log(c_{t+1}/c_t)}\right)\\ &= e^{\nu+\frac{1}{2}(1-\alpha)\sigma^2}\left(1 - f_t + f_t(1-b)^{1-\alpha}\right)^{\dfrac{1}{1-\alpha}}\mu\left(\Phi_{t+1}\right) \end{aligned} \tag{5.31}$$

and $\mu(\Phi_{t+1})$ is the *constant*

$$\mu(\Phi_{t+1}) = \left[\sum_{f'} P(f')\Phi(f')^{1-\alpha}\right]^{\dfrac{1}{1-\alpha}}. \tag{5.32}$$

Note that these steps utilize the various independences built into the process for $\log(c_{t+1}/c_t)$. Equation (5.31), for example, uses the fact that η_{t+1} is *i.i.d.* and independent of the realization of the disaster state. And the realization of the disaster state, in turn, is independent of the realization of next-period's disaster probability.[21] In (5.32), I've assumed that the distribution for f_{t+1} is

[21] Hence, we've really done a $\mu(xyz) = \mu(x)\mu(y)\mu(z)$ split of the certainty equivalent.

discrete—$P(f')$ is the probability of f', and $\Phi(f')$ is the value of Φ_{t+1} in that state.

Combining these expressions, we can write the (scaled) utility process as

$$\Phi(f_t) = \left[1-\beta+\beta e^{(1-\frac{1}{\psi})(\nu+\frac{1}{2}(1-\alpha)\sigma^2)}\left(1-f_t+f_t(1-b)^{1-\alpha}\right)^{\frac{1-1/\psi}{1-\alpha}}\mu\left(\Phi\right)^{1-\frac{1}{\psi}}\right]^{\frac{1}{1-1/\psi}}$$

More compactly, letting $\theta = \nu + (1/2)(1-\alpha)\sigma^2$ and $A = \mu(\Phi)$, we can write the process as

$$\Phi(f_t) = \left[1-\beta+\beta\left(e^{\theta}A\right)^{1-\frac{1}{\psi}}\left(1-f_t+f_t(1-b)^{1-\alpha}\right)^{\frac{1-1/\psi}{1-\alpha}}\right]^{\frac{1}{1-1/\psi}}. \tag{5.33}$$

Note that Φ is decreasing in f_t, independent of whether $\alpha \gtrless 1$ or $\psi \gtrless 1$. This follows because the term

$$\left(1-f_t+f_t(1-b)^{1-\alpha}\right)^{\frac{1}{1-\alpha}}$$

is decreasing in f_t for both $\alpha > 1$ and $\alpha < 1$.[22]

Computationally, (5.33) defines a function $\Phi(f)$ for an arbitrary choice of A; the equilibrium Φ is the one that satisfies (5.33) and $A = \mu(\Phi)$. So, one would solve this iteratively—from an initial Φ_0, calculate $A = \mu(\Phi_0)$, use (5.33) to calculate Φ_1, and repeat with $\Phi_0 = \Phi_1$.

The recursion (5.33) is one of two building blocks for pricing assets in the model. The other, of course, is the stochastic discount factor. The general form is the by-now-familiar

$$m_{t+1} = \beta\left(\frac{c_{t+1}}{c_t}\right)^{-\frac{1}{\psi}}\left(\frac{v_{t+1}}{\mu_t(v_{t+1})}\right)^{\frac{1}{\psi}-\alpha}.$$

In pricing assets, m_{t+1} involves three independent random variables with respect to which expectations must be taken—whether or not a disaster occurs, with probability f_t; the *i.i.d.* innovation to consumption growth, η_{t+1}; and the draw of next-period's disaster probability, f_{t+1}, from the distribution $P(f_{t+1})$.

The price-dividend ratio for a claim to aggregate consumption will be (as before) a function of the disaster probability, and the return on the consumption claim will be

$$R^c_{t+1} = \frac{c_{t+1}}{c_t}\frac{1+w_c(f_{t+1})}{w_c(f_t)}.$$

The price-dividend ratio obeys the pricing relation

$$w_c(f_t) = \mathbb{E}_t\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{1-\frac{1}{\psi}}\left(\frac{v_{t+1}}{\mu_t(v_{t+1})}\right)^{\frac{1}{\psi}-\alpha}(1+w_c(f_{t+1}))\right] \tag{5.34}$$

[22] If $\alpha < 1$, $(1-b)^{1-\alpha} < 1$, and so $1-f+f(1-b)^{1-\alpha}$ is decreasing in f; since $1/(1-\alpha) > 0$, the whole expression is decreasing in f. On the other hand, if $\alpha > 1$, then $(1-b)^{1-\alpha} > 1$, and the stuff inside the parentheses is increasing in f, but the whole thing is raised to the $1/(1-\alpha) < 0$ power.

It's left as an exercize for you to show that (5.34) specializes to

$$w_c(f_t) = \beta e^{(1-\frac{1}{\psi})\theta} A^{\alpha-1/\psi} \left(1 - f_t + f_t(1-b)^{1-\alpha}\right)^{\frac{1-(1/\psi)}{1-\alpha}} \times \sum_{f'} P(f')\Phi(f')^{\frac{1}{\psi}-\alpha} \left(1 + w_c(f')\right) \quad (5.35)$$

The price of the riskless asset q obeys

$$q(f_t) = \beta e^{\theta_1} A^{\alpha-1/\psi} \left[\frac{1 - f_t + f_t(1-b)^{-\alpha}}{\left(1 - f_t + f_t(1-b)^{1-\alpha}\right)^{\frac{(1/\psi)-\alpha}{1-\alpha}}} \right] \sum_{f'} P(f')\Phi(f')^{\frac{1}{\psi}-\alpha} \quad (5.36)$$

where $\theta_1 = (\alpha - (1/\psi))\,\theta - \alpha\nu + (1/2)(\alpha\sigma)^2$.

Exercize 5.3. *Derive the expressions (5.35) and (5.36). Then, show that when the elasticity of intertemporal substitution, ψ, is equal to one, future utility (Φ) doesn't matter for the equity premium—we obtain the same result for $\log\left[\mathbb{E}(R^c_{t+1})/R^F_t\right]$ as we did in the case of standard preferences—i.e., equation (5.29). A key step is to show that a constant w_c solves (5.35) when $\psi = 1$.*

As complicated as these last expressions appear, Gourio is still able to prove some results analytically. The key is that, under the *i.i.d.* assumption on f, terms like $\sum_{f'} P(f')\Phi(f')^{\frac{1}{\psi}-\alpha}$ and $\sum_{f'} P(f')\Phi(f')^{\frac{1}{\psi}-\alpha}\left(1 + w_c(f')\right)$ are constants with respect to variation in f_t. Thus, for example

$$\log R^F_t = -\log\left(q(f_t)\right) = \text{ constants } + \log\left(\frac{\left(1 - f_t + f_t(1-b)^{1-\alpha}\right)^{\frac{(1/\psi)-\alpha}{1-\alpha}}}{1 - f_t + f_t(1-b)^{-\alpha}}\right)$$

Gourio shows that if $\alpha \geq 1$, then $\log R^F_t$ is decreasing in f_t for small values of f_t.

It's only slightly more complicated to characterize the dependence on f_t or the price-dividend ratio and return on the consumption claim. First, note that from (5.35)

$$w_c(f_t) = (\text{positive constants}) \times \left(1 - f_t + f_t(1-b)^{1-\alpha}\right)^{\frac{1-(1/\psi)}{1-\alpha}} \quad (5.37)$$

This is decreasing in f_t if $1 - (1/\psi) > 0$—*i.e.*, if the EIS $\psi > 1$. Having $w_c(f_t)$ decreasing in f_t is a desirable feature, at least on intuitive grounds—if there's an increase in the probability of disaster striking tomorrow, one would expect today's equity price to fall.

Conditional on f_t, the expected return on the consumption claim is

$$\mathbb{E}\left(R^c_{t+1} : f_t\right) = e^{\nu+\frac{1}{2}\sigma^2}\left(1 - f_t + f_t(1-b)\right)\frac{\sum_{f'} P(f')(1 + w_c(f'))}{w_c(f_t)}. \tag{5.38}$$

Using (5.37), and collecting together constants (relative to f_t), we can write the log expected return as

$$\log \mathbb{E}\left(R^c_{t+1} : f_t\right) = \text{ constants } + \log\left(\frac{1 - f_t + f_t(1-b)}{\left(1 - f_t + f_t(1-b)^{1-\alpha}\right)^{\frac{1-(1/\psi)}{1-\alpha}}}\right)$$

Whether $\log \mathbb{E}\left(R^c_{t+1} : f_t\right)$ is increasing or decreasing in f_t depends on whether $\psi \leq 1$. If $\psi \leq 1$, the log expected return is definitely decreasing in f_t. If $\psi > 0$, it can be increasing or decreasing depending on the precise parameter values, and the value of f_t.

Intuitively, there are two effects on $\log \mathbb{E}\left(R^c_{t+1} : f_t\right)$ of an increase in f_t. On the one hand, an increase in f_t lowers the asset's expected payoff next period (the $1 - f_t + f_t(1-b)$ term). But an increase in f_t also lowers the asset's price today—*i.e.*, $w_c(f_t)$ falls, which is captured in the $\left(1 - f_t + f_t(1-b)^{1-\alpha}\right)^{\frac{1-(1/\psi)}{1-\alpha}}$ term. Whether the expected return increases or decreases will depend on whether the expected falls by less or more than the fall in price. If $\psi < 1$, the answer is always 'by less'. If $\psi > 1$, the answer depends on parameter values and on f_t.

It turns out that, for α sufficiently large, ψ need only be a bit bigger than one to obtain the result that $\log \mathbb{E}\left(R^c_{t+1} : f_t\right)$ goes up when f_t goes up. Note that the increase in $\log \mathbb{E}\left(R^c_{t+1} : f_t\right)$ in that case is entirely due to a wider risk spread—we've already seen that higher f_t *lowers* R^F_t for $\alpha \geq 1$.

Figure 5.2.3 plots my calculation of the set of (α, ψ) pairs consistent with an increasing log expected equity return. I calculated this assuming $b = 0.43$ (as Gourio does) and for $f_t = 0.012$. It turns out that for f_t of that order of magnitude—around a 1% chance of disaster, give or take a half a percent or so—I see little noticeable difference in the curve shown in the figure.

From the figure we also see that if risk aversion α is too low—especially around $\alpha = 2$—the necessary ψ values get unreasonably high.[23]

In any case, these hopeful results in the *i.i.d.* case prompt Gourio to explore, computationally, the behavior of the model with Epstein-Zin preferences and persistence in the process describing the evolution of f_t.

5.2.4 Epstein-Zin preferences, persistent probability of disaster

If you've followed Gourio's model thus far, incorporating persistence in f_t is really not a big deal. What changes? First, and most importantly, the constant

[23] As you may have gathered from reading Bansal and Yaron, even an EIS as high as $\psi = 1.5$—their assumption—is pushing the envelope, as far as EIS values that most people would consider plausible.

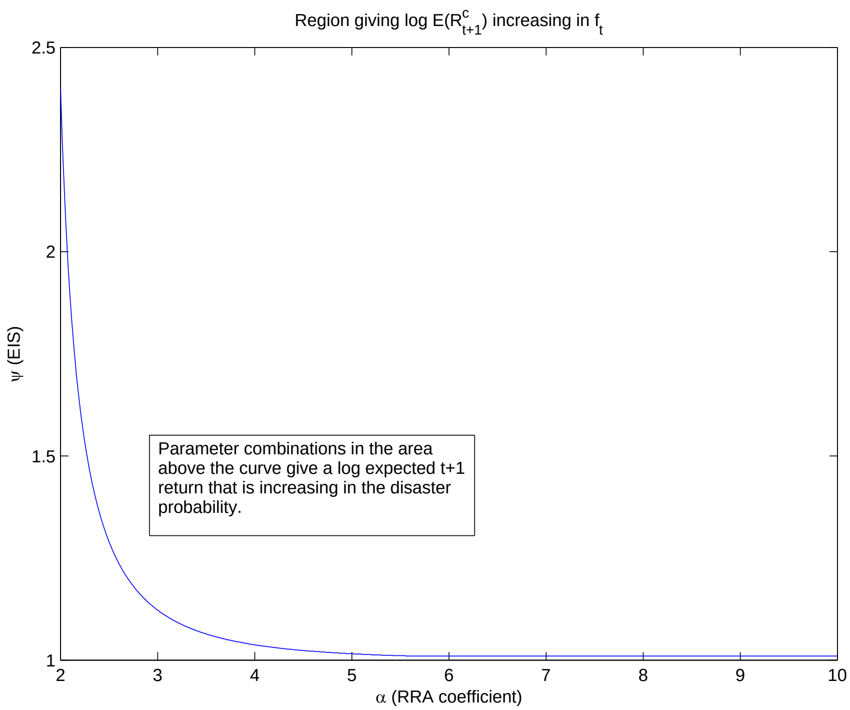

Figure 5.5: Combinations of α and ψ that produce a log expected equity return that is increasing in the probability of disaster. The pairs above the curve have this property. The calculations were made assuming $b = 0.43$ and $f_t = 0.012$. From the model of Gourio [Gou08], with EZ preferences and $f_t \sim i.i.d.$

that we'd labelled A—that is, $\mu\left(\Phi\right)$—is no longer a constant. Rather, it's now

$$\mu\left(\Phi : f_t\right) = \left[\sum_{f'} P(f_t, f')\Phi(f')^{1-\alpha}\right]^{\frac{1}{1-\alpha}}.$$

Also, expressions like $\sum_{f'} P(f')\Phi(f')^{\frac{1}{\psi}-\alpha}$, $\sum_{f'} P(f')\Phi(f')^{\frac{1}{\psi}-\alpha}\left(1 + w_c(f')\right)$, and $\sum_{f'} P(f')\left(1 + w_c(f')\right)$ become

$$\sum_{f'} P(f_t, f')\Phi(f')^{\frac{1}{\psi}-\alpha},$$

$$\sum_{f'} P(f_t, f')\Phi(f')^{\frac{1}{\psi}-\alpha}\left(1 + w_c(f')\right),$$

and

$$\sum_{f'} P(f_t, f') \left(1 + w_c(f')\right),$$

resulting in obvious modifications of the key equations (5.35), (5.36), and (5.38).

While these modifications are not especially difficult to make, they do make it impossible to draw any conclusions about the model on a purely analytical basis, as Gourio did with the first two versions of the model. That makes it necessary to solve the model computationally. Gourio assumes a simple, symmetric two-state Markov chain for f_t—$f_t \in \{f_l, f_h\}$, and

$$P = \begin{pmatrix} 1-\pi & \pi \\ \pi & 1-\pi \end{pmatrix}$$

Gourio calibrates f_l, and f_h so as to give an unconditional mean probability of disaster equal to 1.7%.[24] This gives $f_l = 0.017 - \epsilon$ and $f_h = 0.017 + \epsilon$, where ϵ is the unconditional standard deviation of f_t. Gourio experiments with different values for ϵ, as well as for π, which governs the persistence of the process.[25] Initially, Gourio chooses $\epsilon = 0.01$ and $\pi = 0.1$—the latter choice implies a first-order autocorrelation of $AC_1 = 1 - 2\pi = 0.8$. As Gourio notes, there's very little empirical guidance for the choice of either ϵ or π.

The choice of b is guided by Barro's work—Gourio sets $b = 0.43$, which is about the mid-point of the range of disasters Barro catalogs. Gourio also assumes that partial default on the riskless bond (only the fraction $1 - b$ is repaid), which occurs with probability 0.4 in the event of a disaster. In terms of the variable 'd' we introduced at the beginning of this section on disasters, Gourio is effectively setting $1 - d = 0.6(1 - 0.43)$.

The other parameters are the taste parameters, β, α and ψ, and the parameters of the *i.i.d.* part of consumption growth, ν and σ. Gourio sets $\nu = 0.025$, $\sigma = 0.02$, and $\beta = 0.97$. For the most part, he sets $\alpha = 4$, apart from one case with $\alpha = 3.17$.[26] He experiments with values for the EIS ranging from $\psi = 0.25$ to $\psi = 1.5$.

Some variations that Gourio also considers include:

- Leverage—as in Bansal and Yaron, Gourio also prices a dividend stream that's much more volatile than aggregate consumption. In particular, $\log(d_{t+1}/d_t) = \lambda \log(c_{t+1}/c_t)$, with $\lambda = 3$. Leverage proves to be important for matching the volatility of the price-dividend ratio and the price-dividend ratio's ability to forecast future returns.

- Having the probability of disaster tomorrow depend on the occurrence (or not) of a disaster today. The f_t process above is independent of whether disasters actually occur. Gourio tries some experiments where a disaster today makes a disaster tomorrow either more or less likely. This

[24] Since P is symmetric, its invariant distribution is $(1/2, 1/2)$.

[25] Recall that a symmetric two-state Markov chain that mimics an autoregressive process will have $\pi = (1 - AC_1)/2$, where AC_1 is the process's first-order autocorrelation.

[26] Our α is his θ.

amounts to having different P matrices depending on the occurrence or non-occurrence of a disaster. This channel turns out to have little effect on the results.

- Picking parameters—in particular, α, ϵ, and π—that best match the following moments in the data: the volatility of the price-dividend ratio, the mean equity premium, and the coefficient of a regression of excess returns on the (inverse of) the price-dividend ratio. Those results match many features of the data, though the model implies too low a volatility of the riskless rate and too high a volatility of dividend growth.

Lecture 6

Bond Pricing and the Term Structure of Interest Rates

So far, we've worked with models where we've priced either infinitely-lived equity or one-period bonds.

References

[Abe90] Andrew B. Abel. Asset prices under habit formation and keeping up with the Joneses. *American Economic Review*, 80(2):38–42, 1990.

[Bar06] Robert J. Barro. Rare disasters and asset markets in the twentieth century. *Quarterly Journal of Economics*, 121(3):823–866, 2006.

[BDL93] Robert B. Barsky and J. Bradford De Long. Why does the stock market fluctuate? *The Quarterly Journal of Economics*, 108(2):291–311, 1993.

[Ber98] Michele Bernasconi. Tax evasion and orders of risk aversion. *Journal of Public Economics*, 67:123–134, 1998.

[BG69] William A. Brock and David Gale. Optimal growth under factor augmenting progress. *Journal of Economic Theory*, 1(3):229–243, 1969.

[BKY07] Ravi Bansal, Dana Kiku, and Amir Yaron. Risks for the long run: Estimation and inference. Unpublished manuscript, 2007.

[BL02] Ravi Bansal and Christian Lundblad. Market efficiency, asset returns, and the size of the risk premium in global equity markets. *Journal of Econometrics*, 109(2):195–237, 2002.

[BPR78] Charles Blackorby, Daniel Primont, and R. Robert Russell. *Duality, Separability, and Functional Structure: Theory and Economic Applications*. North-Holland, 1978.

[Bre79] Douglas T. Breeden. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics*, 7(3):265–296, 1979.

[BRW08] Jacob Boudoukh, Matthew Richardson, and Robert Whitelaw. The myth of long-horizon predictability. *Review of Financial Studies*, 24(4):1577–1605, 2008.

[BY04] Ravi Bansal and Amir Yaron. Risks for the long run: A potential resolution of asset-pricing puzzles. *Journal of Finance*, 59(4):1481–1509, 2004.

[CC99] John Y. Campbell and John H. Cochrane. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, 107(2):205–251, 1999.

[CCC10] Claudio Campanale, Rui Castro, and Gian Luca Clementi. Asset pricing in a production economy with Chew-Dekel preferences. *Review of Economic Dynamics*, 13(2):379–402, 2010.

[CG11] George M. Constantinides and Anisha Ghosh. Asset pricing tests with long run risk in consumption growth. Unpublished manuscript, University of Chicago Booth School of Business, August 2011.

[Che06] Raj Chetty. A new method of estimating risk aversion. *American Economic Review*, 96:1821–1834, 2006.

[CLM96] John Y. Campbell, Andrew W. Lo, and A. Craig MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1996.

[Coc01] John H. Cochrane. *Asset Pricing*. Princeton University Press, 2001.

[Coc08] John H. Cochrane. The dog that did not bark: A defense of return predictability. *Review of Financial Studies*, 21(4):1533–1575, 2008.

[Con90] George M. Constantinides. Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy*, 98(3):519–543, 1990.

[CS88] John Y. Campbell and Robert J. Shiller. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1(3):195–228, 1988.

[Den95] The term structure of interest rates in real and monetary economies. *Journal of Economic Dynamics and Control*, 19(5-7):909–940, 1995.

[Dol11] Jim Dolmas. Risk preferences, intertemporal substitution, and business cycle dynamics. Unpublished manuscript, May 2011.

[DR89] Philip H. Dybvig and Stephen A. Ross. Arbitrage. In John Eatwell, Murray Milgate, and Peter Newman, editors, *The New Palgrave: Finance*, pages 57–71. W. W. Norton Publishing Company, 1989.

[Duf92] Darrell Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, 1992.

[EZ89] Larry Epstein and Stanley Zin. Substitution, risk aversion and the temporal behavior of asset returns: A theoretical analysis. *Econometrica*, 57:937–969, 1989.

[EZ90] Larry Epstein and Stanley Zin. 'First-order' risk aversion and the equity premium puzzle. *Journal of Monetary Economics*, 26:387–407, 1990.

[Fam65] Eugene F. Fama. The behavior of stock market prices. *Journal of Business*, 38(1):34–105, 1965.

[Fam70] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970.

[Gab08] Xavier Gabaix. Variable rare disaster: A tractable theory of ten puzzles in macro-finance. *American Economic Review*, 98(2):64–67, 2008.

[Gou08] François Gourio. Time-series predictability in the disaster model. *Finance Research Letters*, 5(4):191–203, 2008.

[Gra08] Liam Graham. Consumption habits and labor supply. *Journal of Macroeconomics*, 30(1):382–395, 2008.

[Gul91] Faruk Gul. A theory of disappointment aversion. *Econometrica*, 59(3):667–686, 1991.

[Guv09] Fatih Guvenen. A parsimonious macroeconomic model for asset pricing. *Econometrica*, 77(6):1711–1750, 2009.

[Hic35] John R. Hicks. A suggestion for simplyfying the theory of money. *Economica*, pages 1–19, February 1935.

[Jer98] Urban Jermann. Asset pricing in production economies. *Journal of Monetary Economics*, 41(2):257–275, 1998.

[Jud98] Kenneth L. Judd. *Numerical Methods in Economics*. MIT Press, 1998.

[Kar59] Samuel Karlin. *Mathematical Methods and Theory in Games, Programming, and Economics*. Addison-Wesley Publishing Company, Inc., 1959.

[Koc90] Narayana R. Kocherlakota. On the 'discount' factor in growth economies. *Journal of Monetary Economics*, 25(1):43–47, 1990.

[Koc96] Narayana R. Kocherlakota. The equity premium: It's still a puzzle. *Journal of Economic Literature*, 34(1):42–71, March 1996.

[KP78] David Kreps and Evan Porteus. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, 46(1), 1978.

[Kre88] David M. Kreps. *Notes on the Theory of Choice*. Westview Press, 1988.

[KS10] Karen A. Kopecky and Richard M. H. Suen. Finite state Markov-chain approximations to highly persistent processes. *Review of Economic Dynamics*, 13:701–714, 2010.

[KT79] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[Ler08] Stephen F. Leroy. Excess volatility tests. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave MacMillan, 2nd edition, 2008.

[Lin65] John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37, 1965.

[LL10] Martin Lettau and Sydney C. Ludvigson. Measuring and modeling variation in the risk-return tradeoff. In Yacine Aït-Sahalia and Lars Peter Hansen, editors, *Handbook of Financial Econometrics*, volume 1, chapter 11, pages 617–690. Elsevier Science B.V., 2010.

[LN07] Sydney C. Ludvigson and Serena Ng. The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222, 2007.

[LP81] Stephen F. Leroy and Richard D. Porter. The present value relation: Tests based on implied variance bounds. *Econometrica*, 49(3):555–574, May 1981.

[LP04] Francis A. Longstaff and Monika Piazzesi. Corporate earnings and the equity premium. *Journal of Financial Economics*, 74(3):400–421, 2004.

[LU09] Lars Ljungqvist and Harald Uhlig. Optimal endowmwnet destruction under Campbell-Cochrane habit formation. NBER Working Paper No. 14772, 2009.

[Luc78] Robert E. Lucas, Jr. Asset prices in an exchange economy. *Econometrica*, 46(6):1429–1445, November 1978.

[Mar52] Harry M. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[Mar99] Harry M. Markowitz. The early history of portfolio theory: 1600-1960. *Financial Analysts Journal*, 55(4):5–16, 1999.

[McK02] Lionel W. McKenzie. *Classical General Equilibrium Theory*. MIT Press, 2002.

[MP85] Rajnish Mehra and Edward C. Prescott. The equity premium: A puzzle. *Journal of Monetary Economics*, 15:145–161, 1985.

[MP88] Rajnish Mehra and Edward C. Prescott. The equity risk premium: A solution? *Journal of Monetary Economics*, 22:133–136, 1988.

[MY03] Angelo Melino and Alan X. Yang. State-dependent preferences can explain the equity premium puzzle. *Review of Economic Dynamics*, 6(4):806–830, 2003.

[PSV04] Ignacio Palacios-Huerta, Roberto Serrano, and Oscar Volij. Rejecting small gambles under expected utility. *Economics Letters*, 91:250–259, 2004.

[Qui82] John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3:323–343, 1982.

[Rab00] Matthew Rabin. Risk aversion and expected-utility theory: A calibration result. *Econometrica*, 68:1281–1292, 2000.

[Rie88] Thomas A. Rietz. The equity risk premium: A solution. *Journal of Monetary Economics*, 22(1):117–131, 1988.

[Ros04] Stephen A. Ross. *Neoclassical Finance*. Princeton University Press, 2004.

[Rou95] K. Geert Rouwenhorst. Asset pricing implications of equilibrium business cycle models. In Thomas F. Cooley and Edward C. Prescott, editors, *Frontiers of Business Cycle Research*, pages 294–330. Princeton University Press, Princeton, NJ, 1995.

[RZ10] Bryan R. Routledge and Stanley E. Zin. Generalized disappointment aversion and asset prices. *Journal of Finance*, 65(4):1303–1332, 2010.

[Sar07] Thomas J. Sargent. Commentary. *Federal Reserve Bank of St. Louis Review*, 89(4):301–303, 2007.

[Sha64] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.

[Shi77] Georgi E. Shilov. *Linear Algebra*. Dover Publications, 1977.

[Shi81] Robert J. Shiller. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review*, 71(3):421–436, 1981.

[SS90] Uzi Segal and Avia Spivak. First order versus second order risk aversion. *Journal of Economic Theory*, 51(1):111–125, 1990.

[SS08] Zvi Safra and Uzi Segal. Calibration results for non-expected utility theories. *Econometrica*, 76:1143–1166, 2008.

[Sta00] Chris Starmer. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 38:333–382, 2000.

[SW03] Frank Smets and Raf Wouters. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association*, 1(5):1123–1175, 2003.

[Tau86] George Tauchen. Finite state Markov chain approximations to univariate and vector autoregressions. *Economics Letters*, 20:177–181, 1986.

[Tob58] James Tobin. Liquidity preference as behavior toward risk. *Review of Economic Studies*, 25(2):65–86, 1958.

[TW11] Yi Tang and Robert F. Whitelaw. Time-varying Sharpe ratios and market timing. *Quarterly Journal of Finance*, 1(3):465–493, 2011.

[Wak94] Peter Wakker. Separating marginal utility and probabilistic risk aversion. *Theory and Decision*, 36:1–44, 1994.

[Wei89] Philippe Weil. The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics*, 24(3):401–422, 1989.

[Wei90] Philippe Weil. Non-expected utility in macroeconomics. *Quarterly Journal of Economics*, 105(1):29–42, 1990.

[Yaa87] Menachem Yaari. The dual theory of choice under risk. *Econometrica*, 55:95–115, 1987.

Appendix A

An introduction to using MATLAB

A.1 Introduction

MATLAB is a matrix-based numerical calculation and visualization tool. It is much like an extremely powerful calculator, with the ability to run scripts—*i.e.*, programs—and generate high-quality plots and graphs. It is also an extremely easy package to learn.

When you start up MATLAB by double-clicking on the MATLAB icon, the 'MATLAB desktop' opens. The layout you see will depend on who's used it last, and whether they've tinkered with the desktop layout. At the very least, you'll see the Command Window and (maybe) views of the workspace, file directory, or a list of recently used commands. MATLAB uses standard Windows conventions—*e.g.*, typing ALT+F brings down the FILE menu, ALT+E brings down the EDIT menu, etc.
The 'prompt' in the Command Window—the spot where you type in commands—looks sort of like this: `>>`.

One way to familiarize yourself quickly is to enter `demo` or `help` at the prompt. Entering `demo` gives you access to a video tour of MATLAB's features (enter `demo 'matlab''getting started'` to see a menu of videos about basic features). Entering `help` gives you a long list of topics you can get help on.

A.2 Creating matrices

There are several ways to create matrices in Matlab. You can create a matrix by typing in the elements of the matrix, inside square brackets. Use a space or comma to separate the elements of a row and use a semicolon to separate rows. Thus, either `A = [1 2 3; 4 5 6]` or `A = [1, 2, 3; 4, 5, 6]` will create the

2×3 matrix `A` which is

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

You'll note that if you enter `A = [1 2 3;4 5 6];`—*i.e.*, this time ending the line with a semi-colon—MATLAB apparently does nothing. This is not the case. MATLAB still records that `A` now denotes the above matrix; the semi-colon at the end of the line simply tells MATLAB to suppress displaying the result. You'll want to do this in most cases, especially with large matrices or vectors, or when you are running an iterative program—displaying the execution of each line greatly slows the program down.

The colon (':') is useful for creating certain types of vectors: in MATLAB, `b = 1:5;` produces a row vector $b = (1, 2, 3, 4, 5)$. Using the colon, you could write our matrix `A` above by typing `A = [1:3; 4:6];`

Matlab also has several built-in functions for creating specialized matrices, among them: `zeros` (matrix of zeros), `ones` (matrix of ones), `eye` (identity matrix), and `rand` (matrix of pseudorandom variables drawn from a uniform distribution on $[0, 1]$).[1] The syntax for all these functions is the same: `zeros(N,M)` creates an $N \times M$ matrix of zeros, and `zeros(N)` makes a square $N \times N$ matrix of zeros. For these, or any function, typing `help` *function_name* will display help related to the function *function_name*.

Having created a matrix, say `A`, you can call an element of it—say the $(2, 1)$ element—by typing `A(2,1)`. If `A` is the matrix described above, typing `A(2,1)` at the command line returns `ans = 4` ('`ans`' is Matlab's shorthand for 'answer'). Entering `A(2,1)` is treated as the question 'What's the 2–1 element of `A`?', the answer of which is four. If `A` is still the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

entering `A(3,1)` will produce an error message, '`Index exceeds matrix dimensions`', since `A` only has two rows and we have asked about the first element in its (non-existent) third row.

You can change an element of a matrix by giving it a new value: if you type `A(2,1)=0`, Matlab returns

$$\mathtt{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 6 \end{bmatrix}$$

You can call or assign values to all the elements of a row or column or some submatrix of a matrix using the colon. In MATLAB, `A(i:j,h:k)` is the submatrix of A consisting of A's rows i through j and columns h through k, `A(i,:)` is the ith row, and `A(:,k)` is the kth column.

Type `A(1,:)` and Matlab will return `ans = [1 2 3]`, all the elements of the first row. Likewise, `A(:,2)` returns

$$\mathtt{ans} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

[1] Purely deterministic computers can't create real random variables, but they can create things *like* random numbers. See http://en.wikipedia.org/wiki/Pseudorandom_number_generator.

all the elements of the second column. `A(1:2,2:3)` returns the 2×2 submatrix

$$\mathtt{ans} = \begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix}$$

In the same way, entering `A(2,:)=ones(1,3)` will change the second row of `A` to a row of ones, and returns

$$\mathtt{A} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix}$$

A convenient feature of MATLAB's indexing of matrices is its use of the word '`end`': `A(i,end)` is the last element of the ith row of `A`, `A(i,j:end)` is the jth through last elements of the ith row, and `A(end,end)` or just `A(end)` is the last element of the last row and column.

A.3 Basic matrix operations

Multiply matrices with '`*`', add matrices with '`+`', subtract matrices with '`-`'. The matrices have to be conformable for these operations in the usual linear algebra sense: adding (`A+B`) and subtracting (`A-B`) requires the matrices `A` and `B` to be of the same dimension. Matrix multiplication (`A*B`) is only feasible if the number of columns of `A` equals the number of rows of `B`.

Rules for conformability are relaxed for scalar multiplication or scalar addition. You can multiply a matrix by a scalar using '`*`' or add a scalar to every element of a matrix using '`+`', so `5*A` multiplies every element of `A` by `5`, and `5+A` adds `5` to every element of `A`.

Transpose matrices using the prime (`'`) symbol: if `A` is the original matrix we were manipulating above, entering `B = A'` returns

$$\mathtt{B} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Other transformations which are occasionally useful are those that reorient a matrix or vector in various ways. For example, if you enter `x=1:5` you will get the vector

$$\mathtt{x} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

If you then enter `y=fliplr(x)` you'll get the vector

$$\mathtt{y} = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \end{bmatrix}$$

—`fliplr` is short for 'flip left-to-right'. More information about these is available by entering `help elmat`, where the '`elmat`' stands for 'elementary matrices and matrix manipulation'.

To find the inverse of a matrix A in MATLAB, you can type `inv(A)`. Of course, A must be a square matrix. For example, enter `A = [1,2;3,4];` then `B = inv(A)`. You should get

$$\texttt{B} = \begin{bmatrix} -2.0000 & 1.0000 \\ 1.5000 & -0.5000 \end{bmatrix}$$

If you then enter `C = B*A` or `C = A*B`, you should see an identity matrix. Well, actually, you'll see

$$\texttt{C} = \begin{bmatrix} 1.0000 & 0 \\ 0.0000 & 1.0000 \end{bmatrix}$$

which leads us to the issue of precision. Given the precision of the calculations MATLAB uses here, `C(1,2)` is zero, but `C(2,1)` is not. You can see this by using MATLAB's method for checking equality, the double equal sign. In general, entering `x==y` for scalars x and y returns a one if $x = y$ and a zero otherwise. If x and y are vectors or matrices of the same size, then `x==y` returns a vector or matrix consisting of ones for the indices where the elements of x and y are equal and zeros elsewhere.[2] So, entering `C(1,2)==0` returns a one, while `C(2,1)==0` returns a zero. You can see what `C(2,1)` actually is by just entering `C(2,1)`. It is a very, very small number, but nonetheless not equal to zero.

MATLAB also has operations for matrix division, which use the slash '/' and backslash '\'. For scalars a and b, `a/b` and `b\a` are ordinary division of a by b. When A and B are matrices, `A/B` is matrix 'right division' of B into A, which is the same as AB^{-1}, while `B\A` is matrix left division, or $B^{-1}A$. Enter `help slash` for more information on these operations. The methods MATLAB uses for these calculations are more precise than the ones it uses for `inv`, and one can use the slashes for finding inverses—if A is an $n \times n$ matrix, then `B = eye(n)/A` calculates A^{-1}. Using the previous `A`, set `B = eye(2)/A`. If you then enter `C = A*B`, you should see the more exact[3]

$$\texttt{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A.3.1 An example—ordinary least squares

Suppose you wanted to calculate the ordinary least squares estimate of β in the regression

$$y = X\beta + u$$

where y is an $N \times 1$ vector of N observations on the 'dependent variable', X is an $N \times k$ matrix of N observations on each of k 'independent' variables, β is a

[2]Type `help relop` for more information about '`==`' and MATLAB's other 'relational operators'—*e.g.*, operators for checking whether $x \geq y$ or whether x has any non-zero elements and so forth.

[3]Calculating inverses using slashes is also faster than using `inv`. This could be an important consideration if you have an iterative program that needs to calculate a large number of inverses. Just as a test, I made a random 100×100 matrix, and calculated its inverse 1000 times, first as `inv(A)`, then as `eye(100)\A`. The amount of time required for the slash method was just 3% of the time required for the `inv` method.

$k \times 1$ vector of 'coefficients', and u is an $N \times 1$ vector of disturbances, whose expectation conditional on X is zero. The idea is that the y_i's are random variables, whose expected values conditional on X, have the following linear form

$$\mathbb{E}[y_i : X_i] = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k$$

If this is the true underlying process generating the data, then realizations—our observations—should obey the regression equation above. The point of least-squares regression analysis is to use the observations X and y to construct an estimate of the (unobserved) vector of coefficients β. In a purely mathematical sense, what least squares does is pick β to minimize the Euclidean distance between the vector $y \in \mathbb{R}^N$ and the subspace of $\mathbb{R}^N$ generated by the k columns of X.[4] The least-squares estimate of β—call it $\hat{\beta}_{\text{OLS}}$—is thus the solution to the following quadratic minimization problem:

$$\min_b (y - Xb)^\top (y - Xb).$$

If you multiply the quadratic form out, you'll see that the problem can be written as

$$\min_b y^\top y - 2b^\top X^\top y + b^\top X^\top Xb$$

The first-order condition is

$$-2X^\top y + 2X^\top Xb = 0,$$

or

$$X^\top Xb = X^\top y,$$

which gives

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top y$$

if $X^\top X$ is nonsingular. In any case, in MATLAB one can perform OLS regression easily. If `y` and `X` are matrices containing your observations of the dependent and independent variables, then entering `beta = inv(X'X)*X'y`, or `beta=(X'X)\X'y` will yield the vector of OLS estimates.

A good exercize to try is to create some data `X` and `y` and apply the regression formula to it. Try `X = [ones(50,1),2*rand(50,1)]` and set the true beta as `beta_true = [5;5]`. Make `y` by adding some normally distributed disturbances to the conditional mean `X*beta_true`. MATLAB's `randn` function makes $N(0, 1)$ pseudorandom variables, so let's create `y = X*beta_true + 2*randn(50,1)`. The mean of `X` should be about $[1, 1]$, so the conditional mean of our `y` will be about 10 and the disturbances will then have a standard deviation about 20 percent of the mean. Now use either of the expressions from the last paragraph to calculate the OLS beta, and compare it to `beta_true`. It should be different, but close.

[4] If x_1 and x_2 are two vectors in $\mathbb{R}^n$, the subspace generated by x_1 and x_2 is the set of all vectors of the form $\alpha x_1 + \beta x_2$ for $\alpha, \beta \in \mathbb{R}$. So, $\{z = Xb : b \in \mathbb{R}^k\}$ is the subspace in $\mathbb{R}^N$ one gets from taking all possible linear combinations—with coefficients $b = (b_1, b_2, \ldots b_k)$—of the columns of X.

A.4 Array operations and elementary functions

MATLAB also has so-called 'array' operations, which are done by putting a period '.' in front of a regular operation. If `A` and `B` are two matrices of the same dimension, then `A.*B` multiplies every element of `A` by the corresponding element of `B`. That is, if $A = [a_{ij}]$ and $B = [b_{ij}]$ then `A.*B` is the matrix whose ijth element is $a_{ij}b_{ij}$. `A./B` performs a similar sort of division—the typical element of the resulting matrix is a_{ij}/b_{ij}. If a is a scalar and B is a matrix, then `a./B` results in a matrix whose typical element is a/b_{ij}.

Other useful operations or functions are raising matrices to powers, taking their square roots, logarithms or exponentials. If A is a square matrix and n an integer, then `A^n` multiplies A by itself n times—*i.e.*, this is A^n. The array version, `A.^n`, raises every element of A to the nth power, for which operation A need not be square.

With square roots, logarithms and exponentials, the convention is reversed in that `B = sqrt(A)` produces a matrix B consisting of the square roots of the elements of A; `B = log(A)` produces a matrix B consisting of the natural logarithms of the elements of A; and `B = exp(A)` creates a matrix B whose ijth element is $e^{a_{ij}}$. That is, these unadorned functions operate in an 'array' sense.

The 'matrix' versions of these functions, `sqrtm`, `logm` and `expm`, perform matrix square roots, logs and exponentials—*e.g.*, if A is a matrix, `B = sqrtm(A)` tries to calculate a matrix B satisfying $BB = A$, and `B = expm(A)` tries to approximate the matrix exponential

$$e^A = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots$$

These functions don't get used too often in most of the applications we normally do.

Help on these topics can be had by entering `help ops` for basic operations and `help elfun` for elementary functions like logs and exponentials. As usual, just typing `help` gives you a list of specific categories about which you can look for help.

A.5 Multi-dimensional arrays

MATLAB allows you to work with matrices, or 'arrays', with more than two dimensions. Three-dimensional arrays are sometimes useful in solving dynamic programming problems or asset pricing problems. Arrays with more than three dimensions may be useful, but are less intuitive, or at least less capable of being visualized.

Enter `B = rand(2,2,3)` at the command line (no semi-colon at the end) to see what a three-dimensional array looks like. It's just like three 2×2 matrices. MATLAB will display them in a vertical list, `B(:,:,1)` then `B(:,:,2)` then `B(:,:,3)`. When programming with three-dimensional arrays, I find it useful to visualize the third dimension as 'depth' (the first two dimensions, rows and columns, corresponding to 'height' and 'width'), so the three dimensional array is like three 2×2 matrices printed on three cards, the three cards arranged so that `B(:,:,1)` is in front, `B(:,:,2)` is behind it, and `B(:,:,3)` is behind `B(:,:,2)`.

A.6 Structure arrays

Structure arrays are a convenient form for collecting together, and moving around, groups of objects of different dimensions. I use them a lot in setting the parameters of a model to be solved or collecting the results.

It's easiest to describe structure arrays just by creating one as an example. Suppose we have a model to solve that will depend on an elasticity of intertemporal substitution (a scalar, η say), a discount factor (a scalar β), and a Markov chain describing the process for consumption (which is a vector of consumption states c and a transition probability matrix P). Our program to solve the model will need those parameters, and our program may call functions that we've written that also need those parameters. It would be convenient to have an array `parameters` that contains all those variables, so we can pass them simply to the program and the functions used by the program. Since they are all different dimensions, we can't use a matrix or multi-dimensional array. So, we'll make `parameters` a structure array.

Suppose we want to set $\eta = 1/2$, $\beta = .94$, $c = (.96, 1.04)$ and

$$P = \begin{pmatrix} 0.975 & 0.025 \\ 0.025 & 0.975 \end{pmatrix}$$

To create a structure `parameters` with the objects in it, we just enter:

```
parameters.eta = 1/2;
parameters.beta = .94;
parameters.c = [.96; 1.04];
parameters.P = [0.975 0.025; 0.025 0.975];
```

That's all there is to it. If you now type `parameters` and hit ENTER, you'll see a list of what's in the structure. You can call an object in the structure just by typing `parameters.`*object*. If the object is a vector or matrix you can call an element of the object just by typing `parameters.`*object*`(i,j)`. For example, entering `parameters.P(2,1)` returns 0.025.

A.7 Eigenvalues and eigenvectors

MATLAB allows you to easily calculate things like eigenvalues and eigenvectors and to diagonalize or otherwise decompose matrices. If A is a square matrix—say

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

—then `eig(A)` will calculate the eigenvalues of A and display them. The `eig` function is one which can have multiple outputs. If you enter `[P,D] = eig(A);` MATLAB calculates both the eigenvectors and eigenvalues of A. The eigenvectors get stored in the matrix `P`, as `P`'s columns, and the eigenvalues get stored in `D`, where `D` is a diagonal matrix with the eigenvalues on the diagonal. The matrices are arranged so that eigenvalues and eigenvectors are matched in the sense that the eigenvector associated with the eigenvalue in the ith column of `D` is in the ith column of `P`.

Try entering `[P,D]=eig(A)` for the 2×2 matrix `A` defined above; then, compare `A*P(:,1)` with `D(1,1)*P(:,1)`. They should be the same, or at least their difference should fairly close to zero (on the order of 10^{-16}), allowing for some imprecision in the calculations.

A.8 Max and min

MATLAB has a '`max`' function and a '`min`' function for finding the biggest or smallest elements in matrices or vectors. If `A` and `B` are matrices of the same dimension (say $n \times k$), then entering `C = max(A,B)` returns the $n \times k$ matrix `C` whose i, jth element is the maximum of `A(i,j)` and `B(i,j)`. The operation of `min` is analogous. Both `max` and `min` will also work in this manner if `A` is a matrix and `B` is a scalar (or vice versa).

More typical for what we do, though, is the following use of `min` and `max`. If `x` is a vector, then `max(x)` by itself returns the maximal value in `x`. This function has multiple outputs, too. `[M,m] = max(x)` returns the maximal value in `M` and the number of the element of `x` where the maximum occurs in `m`. The `min` function can do the same for finding minima. Thus if you enter `x = -5:5;` and `y = x.^2;`—which creates the vectors

$$x = \begin{pmatrix} -5 & -4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

and

$$y = \begin{pmatrix} 25 & 16 & 9 & 4 & 1 & 0 & 1 & 4 & 9 & 16 & 25 \end{pmatrix}$$

—then `[M,m] = min(y)`, you should get that `M = 0` and `m = 6`. If you enter `[M,m] = max(y)`, you'll find that when MATLAB is 'indifferent'—here both `y(1)` and `y(11)` are maxima—it opts for the first occurrence in the vector, in this case returning `M = 25` and `m = 1`.

When `max` and `min` are applied to matrices, the operation described above (by default) looks for the max and min in each column of the matrix—so it's operating along, or searching up and down, the rows of each column. Thus, if A is an $n \times k$ matrix, `[M,m] = max(A)` finds the biggest element in each of the k columns of A and stores them in `M` which is a $1 \times k$ row vector. For each column, the row number where the maximum in that column occurs is stored in `m`. Thus, if

$$A = \begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}$$

then `[M,m] = max(A)` returns `M = [2 4]` and `m = [2 1]`.

You can make MATLAB change the dimension `max` and `min` operate on by entering the desired dimension as an additional argument. Since, as we saw above, a second argument already has a specified purpose in `max` and `min`, the dimension is entered as a third argument, and we put an empty matrix—squares braces with nothing inside—as the second argument. Enter `[M,m] = max(A,[],1)` and MATLAB returns the same answer as before—taking the max along the rows (dimension 1), in each column, is the default. `[M,m] = max(A,[],2)` takes the max along the columns, in each row. Note that `M` and `m` are column vectors in that case. You can also use `max` and `min` along the third (or higher) dimensions of multi-dimensional arrays. Trying entering `A = randn(2,2,2)`, then `[M,m]=max(A,[],3)`. Do the results make sense to you?

I should add that, though I keep using `M` and `m` for the output, you can use whatever variable name you like, as long as it begins with a letter, not a number. MATLAB will accept almost anything as the name of a variable, so in applications you should feel free to give things mnemonic names—*e.g.*, you can use, say, `K2L` for the capital-labor ratio in some model or `wage` for the wage rate. There are rules about avoiding certain special characters, and, I think, a limit on the number of characters in the variable name.

A.9 Special scalars

MATLAB has a few special scalars which come up often. One example is `Inf`, for plus infinity. `Inf` can result from division by zero—try entering `10/0`—or from 'overflow'—when a number is so big that to machine precision it is infinite.[5] Loosely, `Inf` satisfies the properties of $+\infty$—if you add something to `Inf`, the answer's still `Inf`, and dividing something by `Inf` (other than `Inf` itself) yields an answer of zero.

If you enter `Inf/Inf` or `0/0`, you will get to see another special scalar, MATLAB's `NaN`, or 'not a number'. It is usually not good to see this in your output, unless you put it there intentionally.[6]

MATLAB also uses `i`, `j` and `pi` to denote particular numbers—`i` and `j` stand for $\sqrt{-1}$ and `pi` stands for the irrational number π. Within a given session, you can make `i`, `j` and `pi` whatever you'd like. This is useful as `i` and `j` make nice indices for so-called 'for loops'—a loop that executes some operation 'for

[5] Any computer has to describe a number in a certain finite number of bytes, so numbers larger than some critical size are effectively infinite while numbers smaller than some size are regarded as zero. On the computer I'm using right now, for example, anything much over `1.7e+308` is just `Inf`, and anything less than around `4.9e-324` is just `0`. Note that machine precision varies with the order of magnitude of the numbers you're working with. On most systems, the next biggest number than 0 which MATLAB can recognize is 4.9407e-324. The difference between 1 and the next biggest number than 1 (called `eps`) is 2.2204e-016. Try entering `1+eps==1`; MATLAB will return a `0`, indicating the statement is false. Now try `1+eps/2==1`; MATLAB returns a `1`, indicating the statement is true. But, enter `eps/2>0`, and you'll see that that's true, too.

[6] Like with `zeros` or `ones`, `NaN(N,M)` creates an $N \times M$ matrix of `NaN`'s. These are sometimes useful in programming.

$i = 1, 2, \ldots n$' or 'for $j = 1, 2, \ldots n$'. The next time you start up MATLAB—or if you 'clear' the renamed variables—`i`, `j` and `pi` will be restored to their default values. For example—and this also illustrates the use of `clear`—type `pi = 3` at the prompt. You've now set `pi` equal to 3. If you subsequently want to use the real π, or rather the computer's approximation to it, type `clear pi`. This removes your variable `pi` from the workspace. If you then type `pi` at the prompt, you will see that `pi` again denotes π.

MATLAB's `clear` command, just illustrated, is used to remove variables from your 'workspace'—which is, roughly, the memory where MATLAB keeps track of what matrices you've created. You can clear particular variables, as we did above—`clear X Y Z` removes variables `X`, `Y` and `Z` from the workspace—or remove all variables from the workspace, by typing `clear` by itself. You can find out what variables are currently in the workspace, and some information about them such as size and memory allocated to them, by typing `whos`.

A.10 Loops and such

MATLAB has all the standard sorts of 'loops' you might use in programming—'for' loops, 'while' loops, and 'if–then' structures. Typing `help for`, `help while` or `help if` gives MATLAB's help on each of these. Personally, I think learning through examples is a good way to quickly see how these loops work.

A.10.1 A 'while' loop example

Typically, a 'while' loops performs some specified set of commands repeatedly as long as—'while'—something is less than or greater than something else. Here's a really simple example. Enter `t = 0;` at the MATLAB prompt. Then, enter `while t<10000;`. You'll note that as soon as you hit ENTER, the prompt disappears. Do not be alarmed. This is normal; the prompt will come back after you've ended your loop with an `end;`. Now, type `t = t + 1;`—*don't forget the semi-colon!*—and hit ENTER. Finally, type `end;` and hit ENTER. The prompt should quickly reappear. You have just written and executed a 'while' loop. What your loop did was increase `t` by increments of 1, starting from `t` equal to zero, until `t` was at least as big as $10,000$. If you now type `t`, you should see `t = 10000`. A bit silly, but you get the idea.

Here's a richer example of a 'while' loop—maybe also a silly one, though. We know that the transcendental number e is defined as

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

Suppose we wanted to approximate this by terminating the series at some n—*i.e.*, approximating e with

$$1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{n!}$$

for some n. How big an n do we need to get within some ϵ of e?—or, rather, how big an n do we need to get within ϵ of MATLAB's approximation to e, since computers can't handle nonterminating, nonrepeating decimals either. In MATLAB, e is `exp(1)`. What our while loop will do is start with $e = 1$ and add terms of the form $1/n!$ for $n = 1, 2, \ldots$, continuing until the difference between our e and MATLAB's `exp(1)` is less than some ϵ.

Let me describe roughly how the loop will work, and then show you how to execute it in MATLAB. There are two things we'll need to keep track of, the value of e and the value of n. We'll also want to calculate $n!$, which we can do by introducing a third variable—call it x—which will begin at $x = 1$ and be updated to n times its current value at each pass through the loop. Thus, on the first pass it will be 1; on the second $(n = 2)$, it will be $2 \cdot 1$; on the third $(n = 3)$, it will be $3 \cdot 2 \cdot 1$; and so on. Starting with $e = n = x = 1$, each pass through the loop will perform the following commands: $e = e + (1/x)$; $n = n + 1$; and $x = nx$. The 'while' statement will be such that this will go on as long as e is more than ϵ away from `exp(1)`. If you think about the commands I listed, you'll see that the first pass will set

$$e = 1 + (1/1) = 1 + \frac{1}{1!},$$

$$n = 1 + 1 = 2,$$

and

$$x = 2 \cdot 1 = 2!.$$

The next pass will set

$$e = 1 + 1 + (1/2!) = 1 + \frac{1}{1!} + \frac{1}{2!},$$

$$n = 2 + 1 = 3,$$

and

$$x = 3 \cdot 2! = 3!$$

and so forth.

So, this loop will do exactly what it is supposed to do. Now, what about the while statement. Suppose our ϵ is 10^{-15}, which can be entered as `1e-15` in MATLAB. Our while statement will be: `while abs(e-exp(1))>1e-15;`. The '`abs`' is for absolute value—our loop will go on as long as the absolute value of the difference between `e` and `exp(1)` exceeds 10^{-15}. Don't worry—this doesn't take long either.

All that said, we'll now write out the steps as we would perform them in MATLAB. First, we need our 'initial conditions':

```
e = 1;
n = 1;
x = 1;
```

Then, we have the loop itself:

```
while abs(e-exp(1))>1e-15;
e = e + (1/x);
n = n + 1;
x = n*x;
end;
```

Of course, our interest in writing this loop was to find out the value of n which brings our approximation within ϵ of MATLAB's `exp(1)`. To see what the value is, just type `n`. The answer should be `n = 18`.

Do not despair if a loop goes haywire, seeming never to end. To break a loop that's gone haywire or is going on longer than you'd like, just hit CTRL+C, the 'Control' key together with the letter 'C'.

A.10.2 A 'for' loop example

Again, rather than try to describe abstractly how a 'for' loop works, we will try to learn through an example.

Suppose you wanted to generate an artificial time series $\{x_t\}_{t=1}^T$ which obeys the first-order autoregression, or $AR(1)$,

$$x_t = \rho x_{t-1} + \xi_t$$

starting from some initial value x_0, where the ξ_t's are i.i.d. normal random variables with mean zero and variance σ^2. Let's take $T = 100$, $\rho = .9$, $x_0 = 0$ and suppose that the standard deviation of the ξ_t's is $\sigma = .035$. The first step is to construct the ξ_t's, of which there will be 100. Recall that if z is a standard normal random variable—that is, z is normal with mean zero and variance one—then $x = \mu + \sigma z$ is distributed normally with mean μ and variance σ^2. So, '`xi = .035*randn(1,100);`' creates a 1×100 row vector ξ consisting of realizations of random variables drawn from a normal distribution with mean zero and standard deviation .035. If you want to see what ξ looks like, enter '`plot(xi);`'.

Now we want to create the x_t's. Let's begin by creating a vector x consisting of x_0 followed by T zeros. Our loop will then turn the zeros into realizations following the $AR(1)$ process above. Since $x_0 = 0$ itself, you can accomplish this with '`x = zeros(1,101);`' which creates a 101-element row vector `x` with all zeros.[7] Note that given the way MATLAB numbers elements of a vector, x_0 corresponds to `x(1)` in the vector `x`, and the T realizations of x_1 through x_T will correspond to elements `x(2)` through `x(101)`. Then, the following 'for' loop creates x_1 through x_T according to the $AR(1)$ process above:

```
for t = 1:100;
x(t+1) = .9*x(t) + xi(t);
end;
```

[7] If x_0 were something other than zero, 1 say, you would want to create a vector `x` consisting of x_0 followed by 100 zeros. You could do this with `x = [1,zeros(1,100)];`.

You'll notice that, just as with the 'while' loop, as soon as you enter the 'for' statement here, the prompt disappears until the loop is completed with an 'end' statement. You can see what `x` looks like by entering `plot(x)` at the prompt. It's interesting to compare what the disturbances `xi` look like as compared to `x`. One way to put both on the same plot is by using MATLAB's `hold` command. If you enter `plot(xi)`, then `hold`, you'll get a message like '`Current plot held`', which means that any additional plot statements will superimpose their plots on the existing plot. Alternatively, create a vector `t=(1:100);` Then enter the command `plot(t,xi,t,x(2:end))`. The '`x(2:end)`' is the second through last elements of `x`—that is, the 100 values after x_0 that we created with our loop.

On your own, you could play around a little bit and see what a random walk looks like—set $\rho = 1$—and also how difficult it must be to distinguish between that sort of process and the $AR(1)$ with ρ, say, equal to .98. You can also see what a what a random walk with drift looks like—the process is

$$x_t = \delta + x_{t-1} + \xi_t$$

where δ is the 'drift' term. In terms of the loop, you would just change the '`x(t+1) =...`' line to '`x(t+1) = delta + x(t) + xi(t);`' where `delta` is whatever you want to set the drift to—you could put an actual number there, or preface the whole loop with a definition `delta =` *[number]*, then write the line exactly as I have written it here.

A.11 Writing programs, scripts and function files

A.11.1 Scripts

Programs, or scripts, are collections of MATLAB commands stored in a file. When the program is run, MATLAB simply executes whatever commands are contained in the file. As you'll see, writing programs in MATLAB is very easy.

To create a new script, which MATLAB refers to as an 'M-file', open the FILE menu, select NEW, then M-FILE. As soon as you do this, the MATLAB editor opens, with a blank document, in which you can type your program.

MATLAB ignores spaces, and carriage returns or semi-colons are used to separate commands. MATLAB also ignores anything on a line that begins with the percent '%' sign, which is useful for adding comment lines to your programs. To write a program, then, simply type the commands you want MATLAB to execute—with semi-colons at the end of each if you want to suppress displaying the execution of each command—and separate the commands with carriage returns.

As a simple example, let's write a program that executes the same commands as in our $AR(1)$ example. Type the following:

```
T = 100;
rho = .98;
```

```
x0 = 0;
sigma = .035;
xi = sigma*randn(1,T);
x = [x0,zeros(1,T)];
for t = 1:T;
x(t+1) = rho*x(t) + xi(t);
end;
figure;
plot(x);
```

This program will generate the $AR(1)$ process we looked at above. The only slight differences are that we've created variables for T, ρ, x_0 and σ. This will make it easy to go back and change any of their values, since we'll only have to change it one place. Another difference is the line '`figure;`' which will open a new figure window before the plot commands are executed. We'll need this if we want to run the program a couple of times and compare the figures. You will have noticed by now that if no figure windows are open, a '`plot`' command opens a new figure window. What you may not have noticed is that if a figure window is already open, a '`plot`' command will put its plot in the open window, replacing whatever was in that window before. If multiple figure windows are open, the plot command will puts its plot in whichever window is the 'current' one—*i.e.*, whichever one was most recently viewed. See `help figure` for more information. In any case, if we do not want our program to disrupt any of our existing figures, we will want it to open a new window before it executes the plot command.

Now you want to leave the editor and go back to the command window, but first you'll want to first save your program. From the editor's FILE menu, select SAVE. You'll need to give the file a name, and the extension must be '.m'—*i.e.*, if I wanted to name this program 'ar1maker', I would enter 'ar1maker.m' as the filename. I assume that the default directory where the editor will save the file is somewhere on MATLAB's 'path'—that is, in a place where MATLAB can find it. If you get a message that a file with this name already exists—which is very likely if you all run out and try and name it 'ar1maker.m'—please be courteous, and do not overwrite the existing file (unless it's your own). Come up with a new name—maybe `ar1maker2.m` or `ar1maker3.m`. Whatever name you come up with—let's say it's 'ar1maker.m'—after you save it and return to MATLAB's command window, you simply type `ar1maker` at the prompt (no need for the '.m'). Your program will then be run. You'll know that it worked if a figure window opens up and you see x_t plotted.

You can now check out one of the main advantages to writing scripts—the ability to go back and change some parameter quickly and easily, and re-execute the commands, without a lot of needless typing. To tinker with your program, just open the file using the FILE menu. You might change the value of ρ or T, save the changes to the file, and run the program again.

A.11.2 Function files

MATLAB lets you create your own functions, which is very useful in solving or simulating models in economics. A function file is a special type of M-file, one that declares in the first line that it is a function, and what the syntax of the function is.

An extremely simple introduction to creating function files is to turn the M-file you just created into a function file. So, go back to the M-file you just created in the last sub-section using the MATLAB editor. If the filename you gave it was 'ar1maker', then type the following as the first line of the file:

```
function x = ar1maker(rho,sigma,x0,T)
```

We've declared our file is a function file and specified the function's syntax. Our function will produce (and plot) the sample path `x`, given inputs for the parameters `rho` and `sigma`, the initial value `x0` and the number of periods `T`.

Next, since our original M-file specified values for `rho`, `sigma`, `x0` and `T`, we need to either delete or comment out those lines. The editor allows you to easily comment out a block of text by selecting it with the cursor, then choosing EDIT¿COMMENT from the editor's menu bar. If you take this route (rather than deleting), your finished product will look like this:

```
function x = ar1maker(rho,sigma,x0,T)
% T = 100;
% rho = .98;
% x0 = 0;
% sigma = .035;
xi = sigma*randn(1,T);
x = [x0,zeros(1,T)];
for t = 1:T;
x(t+1) = rho*x(t) + xi(t);
end;
figure;
plot(x);
```

Now, save and return to the command window. Test your function by typing something like `x = ar1maker(.9,.05,1,20);` The function should create the simulated series (`x`) and make the plot, just as it did before.

If you're really eager to test how well you're learning MATLAB, create a simulated x_t series, and—referring back to the OLS regression example—figure out how to calculate $\hat{\rho}_{\text{OLS}}$ in the regression $x_t = \rho x_{t-1} + u_t$.

A.12 Last tips

This document is just meant as a quick introduction to get you started using MATLAB. There are a lot of topics we haven't covered—notably, using 'strings' (pieces of text) in programs or functions, set operations (union, intersection), and the whole area of logical operations. You can learn more about these topics from MATLAB's help facility.

As you write increasingly complex programs, an important lesson to keep in mind is that MATLAB is designed to do array operations really, really fast. Things like 'for' loops—well, not so much. Anywhere you can replace a loop with an array operation will greatly speed up your code.

Also, look around for what's available on the internet to learn from. Most researchers who use MATLAB will post their programs on their websites along with their papers. Read how other people have solved problems or written programs to perform different tasks. If you borrow someone's code, make sure to give them credit in your work, and—as important—make sure you understand what their code is doing, if you're going to use it. You don't need to always be reinventing the wheel (unless you're assigned to reinvent a wheel as homework), but you don't want the code you use to just be a 'black box', either.