HOOVER INSTITUTION
ECONOMICS WORKING PAPERS

Rethinking Production Under Uncertainty

John H. Cochrane*

Economics Working Paper 19117

HOOVER INSTITUTION
434 GALVEZ MALL
STANFORD UNIVERSITY
STANFORD, CA 94305-6010

December 17, 2019

Conventional models of production under uncertainty specify that output is produced in fixed proportions across states of nature. I investigate a representation of technology that allows firms to transform output from one state to another. I allow the firm to choose the distribution of its random productivity from a convex set of such distributions, described by a limit on a moment of productivity scaled by a natural productivity shock. The model produces a simple discount factor linked to productivity, which can be used to price any asset, without regard to preferences.

The Hoover Institution Economics Working Paper Series allows authors to distribute research for discussion and comment among other researchers. Working papers reflect the views of the authors and not the views of the Hoover Institution.

---

* Hoover Institution, Stanford University and NBER. 434 Galvez Mall, Stanford CA 94305.
john.cochrane@stanford.edu. I thank Frederico Belo for many helpful and detailed comments. Data, code, and updates are at http://faculty.chicagobooth.edu/john.cochrane.

1. Introduction

Production possibilities in uncertain environments are usually modeled by augmenting standard production functions to include shocks. For example, we may write

$$y(s) = \varepsilon(s) f(k) \tag{1}$$

where $y(s)$ is output in state s, k is capital, and $\varepsilon(s)$ is random productivity. The firm chooses capital k, then nature chooses the state s, i.e. productivity $\varepsilon(s)$, giving random output $y(s)$.

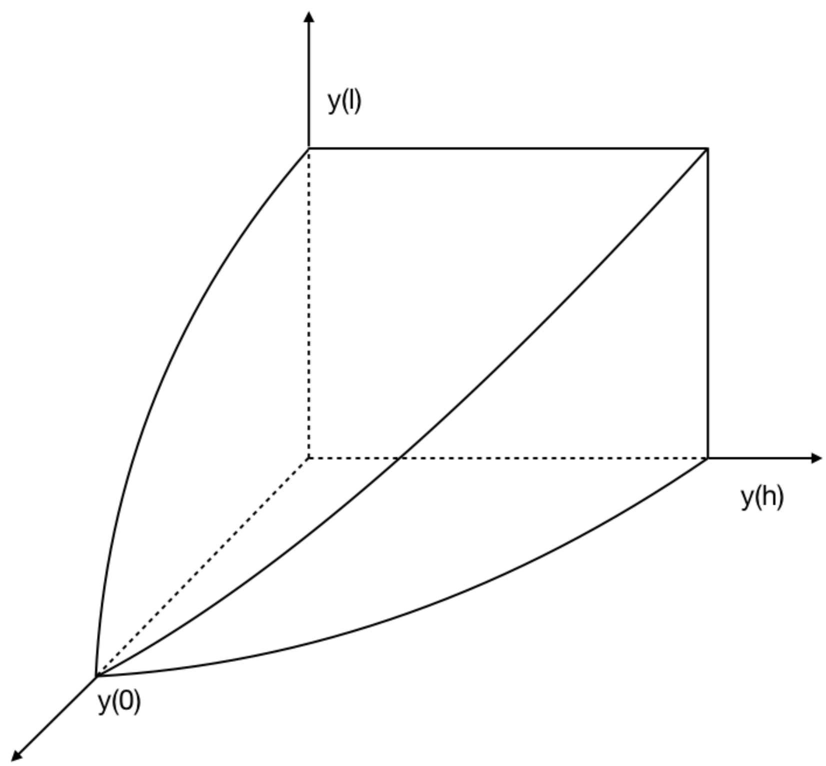

Figure 1: Standard production possibility set in a two-state world. The technology is $y(s) = \varepsilon(s)f(k)$ for $s = h, l$, and $y(0) = W - k$.

Figure 1 illustrates the production set implied by this sort of technology for a two-period world. A farmer has seeds W at time 0. He or she may plant them as k, or eat them as $y(0) = W - k$. At time 1, the field generates wheat $y(s) = \varepsilon(s)f(k)$ according to the state s, which can take on two values $s = h$ or $s = l$. The implied production set smoothly transforms wheat today to a bundle of contingent wheat tomorrow, but it has a kink across the states of nature. No matter how high the contingent claim price of wheat in the low state relative to the high state, there is nothing the producer can do to alter production in its favor.

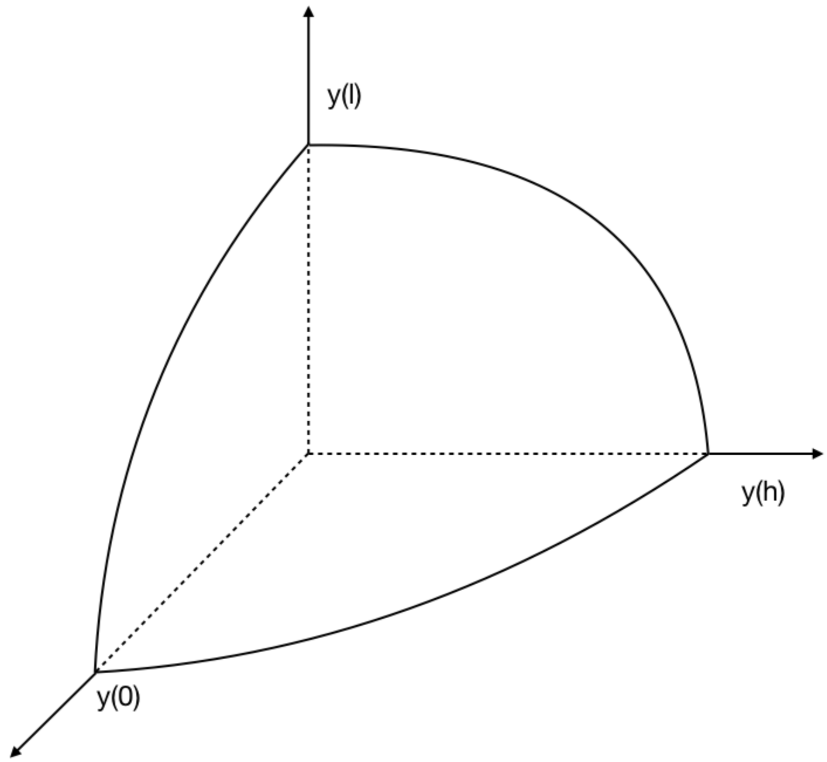

Figure 2: Smooth production possibility set. The firm can change the distribution of output across states of nature – the pattern of $y(h)$ vs. $y(l)$.

This paper explores a different representation for technology under uncertainty, in which the firm has a smooth choice over the state-contingent pattern of its output. Figure 2 illustrates the idea. Now, the farmer can also take actions which shift output from one state $s = h$ to another state $s = l$. If the relative contingent claim price of state l rises, for example, the farmer can produce more in state l and less in state h, potentially leaving consumption at time 0, $W - k$ unchanged.

I explore smooth production sets generated by adding a choice of the productivity distribution $\varepsilon(s)$ to the conventional description of technology (1), constraining the random variable ε to lie in a convex set with a smooth boundary. Most of this paper explores a parametric example, that random productivity ε is constrained by

$$E\left[\left(\frac{\varepsilon}{\theta}\right)^{1+\alpha}\right] = \sum_s \pi(s)\left[\frac{\varepsilon(s)}{\theta(s)}\right]^{1+\alpha} \leq 1 \tag{2}$$

where θ are a set of weights, and $\alpha \geq 0$ is a curvature parameter. We can think of the weights θ as natural or underlying random productivity. The firm may deviate from this natural productivity in some states, by accepting lower productivity in other states. I consider below whether we need the θ weights and how to measure and identify them.

Let m denote a stochastic discount factor, equivalent to contingent claim prices scaled by probabilities, $m(s) = p(s)/\pi(s)$. If a firm maximizes contingent claim value

$$\max E\left[m\varepsilon f(k)\right]$$

subject to (2), the first-order condition for choice of ε – choosing $\varepsilon(s)$ in every state of nature s – leads to

$$m = \lambda \frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}, \tag{3}$$

where λ is a constant that includes the Lagrange multiplier on the constraint. In the dynamic extension of the model,

$$m_{t+1} = \lambda_t \left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha} / \left(\frac{\theta_{t+1}}{\theta_t}\right)^{1+\alpha} \tag{4}$$

where λ_t is a similar constant known at time t. The firm chooses to produce more ε in states of nature with high contingent claim prices or discount factors, and states in which the natural productivity shock θ is larger.

Why is this representation of technology useful or interesting? My direct interest is the construction of production-based asset pricing models. The central question of asset pricing is to tie asset prices and returns to economic fluctuations. For example, such a tie is the only way to address whether observed risk premiums are "rational" compensation for "fundamental" risk. In production-based asset pricing we use firm first-order conditions and observations of real choice variables such as ε to infer what the prices m must have been to generate the choices ε. With that discount factor m we then understand economic fundamentals of asset prices, for example from $0 = E(mR^e)$ or equivalently $E(R^e) = -cov(m, R^e)/E(m)$ where R^e is an excess return. We relate risk premiums to the covariance of excess returns with the discount factor. Viewing assets as bundles of contingent claims, we understand asset prices from

$$\text{asset price} = E\left(m \times \text{ payoff}\right).$$

Expressions (3) and (4) give us a direct production-based model of the stochastic discount factor. By contrast most production-based asset pricing, surveyed below, can only identify from production data a set of returns which must be arbitraged against asset returns, and most general equilibrium asset pricing with production add a few production technologies which generate interesting asset returns and the underlying consumption stream that drives the discount factor. Formula (4) already has the form of several ad-hoc macro-asset pricing models that tie asset returns to a discount factor created from productivity growth, surveyed below. As we elaborate the framework, it can provide foundations for many additional models in this class, that tie asset risk premia to output growth, investment growth, growth of labor and labor shares, and so on.

Production-based asset pricing is deliberately parallel to the standard consumption-based approach to this central question. That theory starts with consumer first-order conditions, that contingent-claim price ratios are proportional to marginal rates of substitution,

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

With the usual power utility, but adding the possibility of a preference shock ϕ, $u(c) = (c/\phi)^{1-\gamma}$, we have

$$m_{t+1} = \beta \left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} / \left(\frac{\phi_{t+1}}{\phi_t}\right)^{1-\gamma}. \tag{5}$$

Once we observe equilibrium consumption, and assume away or identify preference shocks, we can link asset prices to the macroeconomy via consumption, ignoring the production technology that generates observed consumption. Most theories of asset pricing come down to different clever ways to substitute for consumption data in this expression. While a full understanding of the economy requires general equilibrium – knowing where that consumption came from – one can at least tie asset prices to half of the economy, and study preferences and the intermediation between consumers and prices in isolation while others work on production technology.

"Production-based asset pricing" takes the parallel approach: Tie asset prices to economic variables such as output, labor, and productivity via marginal rates of *transformation,* ignoring preferences. While a full understanding of the economy requires general equilibrium – understanding preferences, the consumer's probability assessments, and market structure – one can at least tie asset prices to half of the economy, and study production technology in isolation. One can at least find if firms react to prices optimally, and see if the cyclical tie of asset prices to firm data makes economic sense. Production-based asset pricing is also attractive because business cycles are essentially phenomena of production – declines in investment, durable goods output, and employment – and much less visible in nondurable and services consumption.

But Figure 1 makes clear why a pure production-based asset pricing model is not possible using standard representations of technology extending (1). Since there is a kink in the production set, many different contingent claims prices are consistent with any point the firm might choose. The producer's problem with a standard technology of the form (1) instructs the firm to optimally invest at time 0 to produce a fixed bundle of contingent claims. The firm should invest up to the point that the physical return on investment is correctly priced by the contingent claims/stochastic discount factor m, $1 = E\left[mR^I(k)\right]$, with $R^I(k) = \varepsilon f_k(k)$ in this simple example. We can therefore price payoffs that are perfectly spanned by investment returns. But we cannot infer anything about other returns, or a general discount factor.

When we specify a smooth production set as in (2), on the other hand, we can construct a pure production-based asset pricing model. The production-based discount factor (4) is obviously an attractive parallel to the consumption-based version (5). It appears we just substitute productivity for consumption and apply standard techniques.

The presence of random natural productivity θ raises some practical difficulties however – just as preference shocks ϕ would do if we allowed them. If we allow free shocks, we can explain anything, so allowing shocks means we need to think about their identification and measurement.

We need shocks *somewhere,* however. If neither preferences nor technology had shocks, asset prices would be constant. Preference shocks are becoming more common in asset pricing, including risk aversion and discount rate shocks. Behavioral finance argues that much fluctuation in asset prices comes from variation in probability assessments, that are equivalent to preference shocks. Yet the difficulty behavioral finance has had in generating rejectable content speaks to the difficulty of this identification and measurement question.

Basic correlations in the data argue that we need underlying technology shocks θ. If there were no such shocks, then firms would produce more (higher ε) in high discount-factor states. We usually associate high discount factors – high marginal utility – with low consumption and low stock prices. But output is low in recessions, not the other way around. The fact that stock prices, output, and consumption all comove positively tells us that the bulk of such fluctuations must come from technology shocks, not preference or irrational-probability shocks. Thus, production-based asset pricing (for the moment) seems to need underlying productivity shocks, and face the shock-identification question that is so often brushed under the rug in macroeconomics and finance. (And models with preference or belief shocks need to face the same question, along with getting the sign of the correlation between asset returns and the business cycle right.)

Though production-based asset pricing is my motivation and the focus of this paper, this representation of technology should also be useful in many other applications. Study of firm's choices of risk exposure, and how those choices respond to asset prices, including commodity futures and derivatives, is an attractive idea.

Why is a smooth representation of production possibilities such as (2) reasonable? First, producers do seem to have some ability to control the pattern of their output across states of nature, i.e. the distribution of the productivity shocks they face. A farmer may plant wheat on ground that does better in rainy or dry weather, choose seeds that prosper in different weather conditions, and so forth. Electric utilities may invest in equipment that produces electricity most efficiently given today's prices and regulatory treatment of coal, oil, gas, nuclear, solar, etc., or it may choose to invest in a variety of equipment, or more costly and flexible-fuel equipment that can adapt to different circumstances. This ability is not unlimited. Technology will naturally have kinks across states of nature completely unrelated to the production process. But technology will naturally not have kinks across many other states of nature that are related to the production process.

Second, one may simply view that the lack of rigid kinks is the most natural production set, and ask what empirical evidence there is for suck kinks. That is how we approach the choice of inputs and in studying nonstochastic production functions. We would start by modeling the farmer's choice to produce wheat vs. corn, or to produce more today and less tomorrow, as a smooth production set. So why is wheat in rainy weather vs. wheat in sunny weather naturally fixed? A reader of Debreu (1959), say, encountering the idea of contingent claims would surely start by writing down a smooth production set, mirroring smooth preferences across goods and states, and mirroring smooth technologies across inputs and over time. Static production theory in textbooks beautifully mirrors static preference theory. Why not production under uncertainty?

Historically, it seems that aggregate production functions with kinks across states of nature are not the result of such consideration and evidence. Instead, shocks were simply tacked on to deterministic intertemporal functions familiar from growth theory. Real business cycle models such as Kydland and Prescott (1982) and King, Plosser, and Rebelo (1988) use technologies of the form (1). None even considers the possibility of a smooth production set across states of nature. That choice is entirely understandable. A smooth production set adds complications. And these authors didn't need to generalize. Tacking productivity shocks on to standard intertemporal technologies was good enough for their uses. But that historical accident does not carve the decision in stone, or argue for kinks against smoothness.

Third, smooth production sets can occur when one aggregates standard production functions. Below, I explore a model in which a firm has access to several different technologies or processes, each of which has a different, but fixed distribution of shocks. By varying its input across the different processes, the firm can change the distribution of the shock in the aggregate production function that relates the firm's total output to its total input.

This approach is analogous the classic result that an aggregate of fixed-coefficient production functions can produce a smooth production function such as the Cobb-Douglas. The latter result is a standard justification for smooth input requirement sets, given that individual machines or production processes may require fixed inputs (Houthakker (1955)).

I apply the same logic to the multiple outputs (across states of nature) of a firm that operates in an uncertain environment. Since each firm, industry, or economy is an aggregate of an immense number of microscopic production activities, this aggregation view suggests a rich set of possibilities for transforming output across states.

One response to this observation might be to argue for disaggregated data and explicit aggregation theory. However, this is not practical advice. National, category, industry, firm and plant-level aggregates are a useful and informative source of data. If, say, individual machines are the ultimate fixed-coefficient producers, we don't have data on individual machines.

I delay a discussion of the literature until after the main body of the paper. It will be much easier to understand how this paper relates to other papers in the production-based enterprise after the reader has a better idea of what is in this paper.

2. Production functions and discount factors

Our goal is to write plausible and tractable aggregate production functions that allow transformation across states. There are many ways to write general concave functions that are differentiable across states of nature. However, it seems productive instead to incorporate standard production theory, and forms that have proved useful in the past, as far as possible.

For that reason, I specify a production function that describes the firm's ability to transform goods over time in a conventional way, but adds to it the ability to transform output across states. Additionally, I focus on and explore a particular CES functional form for this choice: output y is given by a standard production function combining capital k and labor n,

$$y = \varepsilon f(k, n) \tag{6}$$

$$y(s) = \varepsilon(s) f\left[k, n(s)\right]$$

where ε satisfies.

$$E\left[\left(\frac{\varepsilon}{\theta}\right)^{1+\alpha}\right] \leq 1 \tag{7}$$

$$\sum_s \pi(s) \left[\frac{\varepsilon(s)}{\theta(s)}\right]^{1+\alpha} = 1. \tag{8}$$

The second equation in each group expresses random variables as functions of finite states $s = 1, 2...S$. The finite state examples are easier to keep track of, but the analysis is valid for continuously distributed random variables.

The firm can *choose* its productivity ε from the convex set of random variables described by (7). Nature hands the firm an underlying or natural productivity θ, and the firm may choose $\varepsilon = \theta$. But the firm can choose a higher value $\varepsilon(s)$ in some states s, if it accepts a lower value $\varepsilon(s')$ in some other state s'. The parameter α controls the firm's ability to transform across sates of nature. As $\alpha \to \infty$, productivity necessarily converges to the natural shock θ. As α decreases, it is easier and easier for the firm to transform output from one state to another. (Previous drafts of this paper used α in place of $1+\alpha$. I change notation here to more clearly mirror the risk aversion coefficient of power utility.)

An alternative way to think of (7) is that we generalize a certainty production function $y = \theta f(k, n)$ to a CES aggregate of output across states on the left hand side,

$$E\left[\left(\frac{y}{\theta}\right)^{1+\alpha}\right]^{\frac{1}{1+\alpha}} = \left[\sum_s \pi(s) \left(\frac{y(s)}{\theta(s)}\right)^{1+\alpha}\right]^{\frac{1}{1+\alpha}} \leq f(k, n).$$

Defining $\varepsilon = y/f(k, n)$ this is the same formulation as (7). This expression is perhaps more theoretically satisfying, as it describes a convex and smooth set of inputs and outputs. However, I find the idea of "picking productivity" maintains better the connection to well-studied production theory, so I use the former expression. The $\leq$ allows for free disposal, but with positive state prices the firm will always choose equality, so I drop the formalism.

Figure 3 plots the production set (8) in a two-state example, $s = \{h, l\}$ with $\theta(h) = 2$, $\theta(l) = 1$ and $\pi\ (h) = 0.5$. For $\alpha = 1$, you see how (8) induces a convex set of possible $\{\varepsilon(h), \varepsilon(l)\}$ possibilities, and with them a convex set of $y(s) = \varepsilon(s)f(k)$ possibilities we are looking for, as graphed in Figure 2. As we raise α, the curve is more convex, and as we lower α, the curve is flatter. Thus, higher α means that in response to a given contingent-claim price vector, the firm will deviate less from the initial θ, while for lower α it will deviate more. The parameter α plays a similar role to the risk aversion coefficient of utility theory. The natural shock θ biases the production set towards state h in this case.

Probabilities do not naturally enter production technologies. A farmer's ability to produce more in a rainy state and less in a dry state, by moving planting to a field that does better in rainy weather, does not have any natural connection to the probability that the rainy state occurs. Yet,

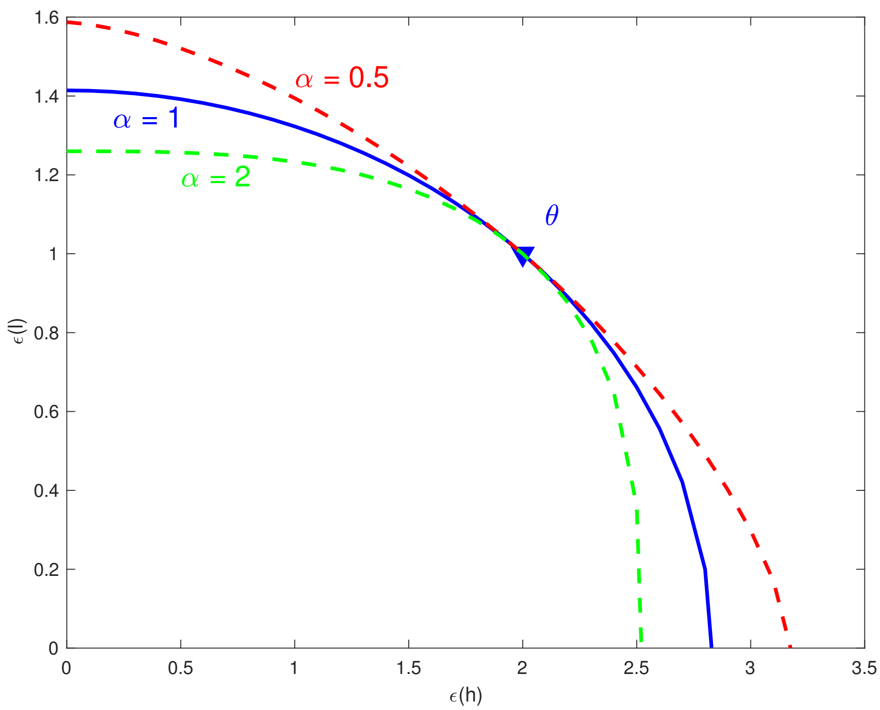

Figure 3: Shock choice sets. Each line gives the set of $\{\varepsilon(h), \varepsilon(l)\}$ that the firm can choose from, satisfying $E[(\varepsilon/\theta)^{1+\alpha}] \leq 1$. The base case is $\alpha = 1$, $\theta = (2, 1)$, and $\pi(h) = 0.5$. The dashed lines vary α to $\alpha = 0.5$ and $\alpha = 2$.

it is very convenient to sum across states of nature by some probability measure, and essentially mandatory to do so with continuously distributed random variables.

Thus, the probabilities in (7) and (8) are arbitrary. They are not necessarily (say) the firm manager's subjective probabilities, as the probabilities in the consumer first-order condition are the consumer's subjective (rational or not) probabilities.

This arbitrariness of probabilities is one reason to include the shock θ. One might wish for the simplicity of a model without natural productivity shocks, but then the probabilities themselves become the weights. Those probabilities might differ enormously from true or empirical probabilities used in analysis. Thus, the weights θ^{α} can serve as transformation between the probability weights, unrelated to actual probabilities, that define technological opportunities, and whatever probabilities we wish to use in analysis. The parameters θ and π are not separately identified, so any change in one can be made up by the other. In that sense the probabilities really do not enter the production set.

This seeming arbitrariness is a virtue. We do not have to worry about rational or irrational, conditional vs. unconditional, true vs. sample, real vs. risk-neutral probabilities, agents who see more than we do, and so forth.

To make these observations a bit more concrete, we could start by generalizing a technology $y = f(k)$ to a CES aggregate over states

$$\left(\sum_s \rho(s) y(s)^{1+\alpha} \right)^{\frac{1}{1+\alpha}} = f(k)$$

where $\rho(s)$ are a set of weights unrelated to probabilities. This describes a concave production set of outputs. But then divide by $f(k)$, and we have a constraint on productivity. And multiply and divide by a convenient set of probabilities, define

$$1/\theta(s)^{1+\alpha} = \rho(s)/\pi(s)$$

and we recover the original specification (2).

3. The simplest model

Consider a firm that maximizes the value of output ε. (I.e., fix $f(k, n) = 1$ to focus on the random variable choice.) The firm's problem is

$$\max_{\varepsilon} E(m\varepsilon) \text{ s.t. } E\left[(\varepsilon/\theta)^{1+\alpha}\right] \leq 1. \tag{9}$$

To be clear, with finite states s, the latter expression means

$$\max_{\{\varepsilon(s)\}} \sum_s \pi(s) m(s) \varepsilon(s) \text{ s.t. } \sum_s \pi(s) \varepsilon(s)^{1+\alpha} / \theta(s)^{1+\alpha} \leq 1.$$

The variable m is the stochastic discount factor, or contingent claim price divided by probability, $m(s) = p(s)/\left[p_0 \pi(s)\right]$, so the objective is the same thing as maximizing contingent claim value. The firm chooses the random variable $\varepsilon(s)$ in each state of nature s. Thus, a first-order condition operates state-by-state inside the expectation.

Introducing a Lagrange multiplier λ on the productivity-choice constraint, the first-order condition is

$$m = \lambda (1 + \alpha) \frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}, \tag{10}$$

i.e.

$$m(s) = \lambda (1 + \alpha) \frac{\varepsilon(s)^{\alpha}}{\theta(s)^{1+\alpha}}. \tag{11}$$

in each state of nature s. The first-order condition directs the firm to rearrange output towards states of nature with high discount factors or contingent claim prices, and towards states where it is easier to produce with high θ.

In standard theory of the firm, we want to solve for choices given prices, for ε given m. We do that by imposing the constraint (9) to eliminate the Lagrange multiplier λ, which yields

$$\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}} = \frac{m}{\left\{E\left[(m\theta)^{\frac{1+\alpha}{\alpha}}\right]\right\}^{\frac{\alpha}{1+\alpha}}}. \tag{12}$$

This condition expresses even more clearly the idea that the firm should produce more in states with high contingent claim prices m and high natural productivity θ.

However, our objective is a production-based asset pricing model: We want to infer what contingent claims prices m must have been in order to produce observed choices, ε. We want to solve first-order conditions and constraints for the discount factor m given choices ε and circumstances θ.

Equation (10) already gives us a discount factor that can price all zero-cost portfolios or excess returns. For this project, we need an m^* such that $0 = E(m^* R^e)$ for any excess return R^e. The level or scale of m^* is irrelevant. If $0 = E(m^* R^e)$ then $0 = E\left[(2m^*)R^e\right]$. Thus, the discount factor

$$m^* = \frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}} \tag{13}$$

immediately prices all zero-cost portfolios. The analogy to the consumption-based $m^* = c^{-\gamma}/\phi^{1-\gamma}$ with utility $u(c) = (c/\phi)^{1-\gamma}$ is attractive.

When using discount factors for zero-cost portfolios, it is often useful to normalize them so the mean discount factor and implied risk-free rate $E(m) = 1/R^f$ is reasonable. This normalization leads to

$$m^* = \frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}} / \left[R^f E\left(\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}\right)\right]. \tag{14}$$

This problem does not lead to a full characterization of the discount factor, because we have not given the firm any ability to transform output over *time*. Equation (12) gives the same choice ε for a discount factor $2m$ as it does for a discount factor m, so we cannot invert (12) to learn the level of the discount factor from ε.

4. A two-period model

To fully characterize the discount factor, we add the conventional $f(k)$ part of production theory, which allows the firm to transform output over time as well as across states. In this formulation the intertemporal and risk aspects of the problem separate so equations like (13) and

(14) continue to describe risk premiums. The intertemporal problem adds a single return which establishes the level of m and the level of returns.

Add capital and the possibility to invest at time 0. The firm's objective is to maximize contingent claim value,

$$\max_{\{k,\varepsilon\}} E\left[m\varepsilon f(k)\right] - k \text{ s.t. } E(\varepsilon^{1+\alpha}/\theta^{1+\alpha}) \leq 1.$$

The firm chooses capital k before the shock is realized. It chooses the value of productivity ε in each state of nature, e.g. $\varepsilon(s)$ for each s.

Introducing a Lagrange multiplier λ on the productivity-choice constraint, the first-order conditions are

$$\frac{\partial}{\partial k} : \ 1 = E\left[m\varepsilon f_k(k)\right] \tag{15}$$

$$\frac{\partial}{\partial \varepsilon} : \ mf(k) = \lambda(1+\alpha)\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}. \tag{16}$$

Equation (15) is the familiar condition that the discounted value of the production accruing to an additional unit of investment should equal its marginal cost. Equivalently, the firm should invest until the physical investment return is correctly priced. We can write it $1 = E(mR^I)$ with $R^I \equiv \varepsilon f_k(k)$ denoting the (random) investment return. This first-order condition is the same as it is in the standard case that the firm has no ε choice. By observing ε and k, we can learn one return R^I, and we can learn any returns that can be priced by arbitrage with R^I. But we cannot learn about other returns or payoffs.

Equation (16) is essentially the same as the productivity choice first-order condition of the simplest model without capital (10). A little more $\varepsilon(s)$ in state of nature s would raise the firm's objective by $\pi(s)m(s)f(k)$, at the cost of lowering output in some other states.

From (16), discount factor (13), $m^* = \varepsilon^{\alpha}/\theta^{1+\alpha}$, and its scaled version (14) that describe zero-price portfolios are unchanged with the addition of $f(k)$ to the production technology. Thus, this two-period model only adds the level of the discount factor to the previous description.

Now, let us incorporate (15) and fully solve for the discount factor. The level of the discount factor is determined in this model by the condition (15) that the discount factor prices the investment return.

$$1 = E\left[m\varepsilon f_k(k)\right] = E\left[\frac{\lambda\,(1+\alpha)}{f(k)}\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}\varepsilon f_k(k)\right] = \lambda\,(1+\alpha)\,\frac{f_k(k)}{f(k)}.$$

Equation (16) then becomes

$$m = \frac{1}{\theta f_k(k)} \left(\frac{\varepsilon}{\theta}\right)^{\alpha} = \frac{1}{\varepsilon f_k(k)} \left(\frac{\varepsilon}{\theta}\right)^{1+\alpha} = \frac{1}{R^I} \left(\frac{\varepsilon}{\theta}\right)^{1+\alpha}. \tag{17}$$

I can't decide which form on the right hand side is prettier. Take your pick.

Since any asset or claim to a payoff x is a bundle of contingent claims, we can write asset prices as price $= E(mx)$ or

$$\text{price } = E\left[\frac{1}{\varepsilon f_k(k)} \left(\frac{\varepsilon}{\theta}\right)^{1+\alpha} x\right].$$

The discount factor (17) is not the inverse of the investment return

$$m \neq 1/R^I = 1/\left[\varepsilon f_k(k)\right].$$

The discount factor (17) adjusts that investment return as the payoff x differs from the payoff R^I as the firm has chosen to distort its productivity from the underlying shock θ.

The investment return $R^I = \varepsilon f_k(k)$ is not riskfree. The model determines the riskfree rate indirectly, through the investment return together with the ε first-order condition that determines risk premiums. From (17), the riskfree rate is

$$\frac{1}{R^f} = E(m) = \frac{1}{f_k(k)} E\left[\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}\right] = E\left[\frac{1}{R^I}\left(\frac{\varepsilon}{\theta}\right)^{1+\alpha}\right].$$

This model separates the economics of intertemporal transformation and risk premiums. The first-order condition (15) governs the allocation of output over *time*, the tradeoff at the margin of an initial k for a risky bundle $\varepsilon f_k(k)$, and it determines the overall level of returns, the level of the discount factor. First-order condition (16) governs the allocation of output across *states of nature* and thus risk premiums. As we generalize the production technology $f(k)$, this simple calculation (13) for characterizing risk premiums turns out to remain essentially unchanged, while the characterization of the overall level of returns gets more complex.

That observation suggests that one may wish to separate levels of returns and risk premiums in empirical applications as well. Since application of (13) (slightly generalized below) is the same for different production technologies generalizing $f(k)$, one may wish to focus on risk premia and not tie the analysis to a specific intertemporal production technology, using (13) and excess returns. Likewise, the study of intertemporal technologies that determine the investment return $R^I = \varepsilon f'(k)$, which has comprised most "production-based asset pricing" to date, can continue relatively unaltered. The fact that the ε might have been chosen from a larger set does not change measurement of the investment return or arbitrage between investment returns and asset returns. One may also wish to avoid the intertemporal question in a risk-premium application by taking the risk-free rate from the data. One may simply treat the Lagrange multiplier λ as another free parameter of the model, chosen each date by the observation of the risk free rate or any other convenient return.

4.1 Production theory vs. asset pricing

In the theory of the firm, we usually solve such first-order conditions to give the producer's choices $\{k, \varepsilon\}$ in terms of prices, i.e. the discount factor m. To this end, we solve the pair of first-order conditions to give one equation describing k and another describes ε, each in terms of m and θ. The resulting expression for optimal k is[1]

$$1 = \left\{ E\left[(m\theta)^{\frac{1+\alpha}{\alpha}} \right] \right\}^{\frac{\alpha}{1+\alpha}} f_k(k) \tag{18}$$

while the optimal productivity ε is given by

$$\frac{\varepsilon^\alpha}{\theta^{1+\alpha}} = \frac{m}{\left\{ E\left[(m\theta)^{\frac{1+\alpha}{\alpha}} \right] \right\}^{\frac{\alpha}{1+\alpha}}}. \tag{19}$$

Equation (19) expresses the same intuition, produce more in high discount factor and high ε states, in purer form. It is the same expression as in the one-period model, (12).

Looking at (19), the choice $\varepsilon = \theta$ emerges if $m \propto 1/\theta$. In this case, we do have that $m = 1/[\varepsilon f_k(k)] = 1/R^I$, i.e. if the discount factor or contingent claim price vector equals the inverse of the firm's investment return. The $\varepsilon = \theta$ case does not emerge under risk neutrality or state prices proportional to probabilities, $m = \beta$.

Though my motivating application is production-based asset pricing, a theory of the firm with choice of productivity shocks would be interesting as well.

5. Labor

Adding labor changes the calculations in interesting ways. Adding other variable inputs, effort, prices (such as a relative price of investment and output goods), and other complications to the period production function has similar effects.

First, a disappointment: One might think that a firm which can adjust inputs after observing a shock can produce more or less output in response to that shock and thus achieve a

[1] From the first form of (17) write $m\theta f_k(k) = \epsilon^\alpha/\theta^\alpha$. Using the constraint $E[(\epsilon/\theta)^{1+\alpha}] = 1$, we have (18) Use (18) to substitute for $f_k(k)$ in the second form of (17) to obtain (19).

marginal rate of transformation. That intuition is false. The ability to produce more or less after a shock is observed does not allow the firm to *transform* output across states of nature. Producing more in one state does not make it harder to produce more in another (Belo (2010), footnote 4 makes this point.)

To see this point, write the production function as

$$y(s) = \varepsilon(s) f\left[k, n(s)\right]$$

where $n(s)$ is labor input or effort in state s. Without productivity choice, the firm's problem is

$$\max_{\{k,n(s)\}} \sum_s \pi(s) m(s) \left\{\varepsilon(s) f\left[k, n(s)\right] - w(s) n(s)\right\} - k$$

where w can represent the wage, or the cost of providing effort. The first-order conditions are

$$m(s)\left[\varepsilon(s) f_n\left[k, n(s)\right] - w(s)\right] = 0 \tag{20}$$

$$\sum_s m(s)\varepsilon(s) f_k\left[k, n(s)\right] = 1. \tag{21}$$

Condition (20) does not help us to identify the discount factor $m(s)$, as $m(s)$ cancels from that equation. The firm adjusts $\varepsilon(s) f_n\left[k, n(s)\right] = w(s)$ separately in each state. This observation gives us no information linking states.

The contingent claim price is not the output price. The contingent claim price applies equally to output and wages. The wage is $w(s)$ *relative* to output in each state. Yes, the firm will hire more labor if the output price is higher – relative to the wage. But that consideration is already included in the real wage w. Written in terms of contingent claims prices $p(s) = m(s)/\pi(s)$, The first-order condition is not $p(s) f_n\left[k, n(s)\right] = w(s)$ –that's a different $p(s)$, an output price not a contingent-claim price. Intuitively, the action of hiring more labor in one state does not change the firm's options in another state, so this margin does not identify contingent claim prices.

Variable labor *does* however act like an additional productivity shock θ, so it gives us a measurable source of such shocks and will be important in quantitative exercises. To see these effects in the simplest model, return to the one-period model of Section 3. Now let the firm maximize

$$\max_{\{\varepsilon,n\}} E\left\{m\left[\varepsilon f(n) - wn\right]\right\} \text{ s.t. } E\left[(\varepsilon/\theta)^{1+\alpha}\right] \leq 1. \tag{22}$$

The labor decision and the wage are both stochastic, i.e. $w(s)$ and $n(s)$ are random variables, and the labor decision takes place after the firm observes the state of the world s. The first-order conditions are now the pair

$$\begin{aligned} \varepsilon f_n(n) &= w \\ m f(n) &= \lambda(1+\alpha)\varepsilon^{\alpha}/\theta^{1+\alpha}. \end{aligned}$$

With a standard power functional form

$$f(n) = n^{\sigma}$$

the first-order conditions become

$$\varepsilon \sigma n^{\sigma-1} = w \tag{23}$$

$$m n^{\sigma} = \lambda\,(1+\alpha)\,\varepsilon^{\alpha}/\theta^{1+\alpha}. \tag{24}$$

We can construct a discount factor for zero-cost portfolios from (24):

$$m^{*} = \frac{\varepsilon^{\alpha}}{\theta^{1+\alpha} n^{\sigma}}. \tag{25}$$

Comparing this result to (13), we add labor n^{σ}. Labor n appears in the discount factor formula just like another shock θ.

Alternatively, we may substitute from the first-order condition (23) to express the labor choice n as a function of wage w. From (23), the labor choice is

$$n^{\sigma} = \left(\frac{\varepsilon\sigma}{w}\right)^{\frac{\sigma}{1-\sigma}}. \tag{26}$$

Substituting for n^{σ} in (24), and solving for m,

$$m = \left[\frac{\lambda\,(1+\alpha)}{\sigma^{\frac{\sigma}{1-\sigma}}}\right] \frac{\varepsilon^{\alpha-\frac{\sigma}{1-\sigma}}}{\theta^{1+\alpha}} w^{\frac{\sigma}{1-\sigma}} \tag{27}$$

Thus, we have a discount factor for zero-cost portfolios

$$m^{*} = \frac{\varepsilon^{\alpha-\frac{\sigma}{1-\sigma}}}{\theta^{1+\alpha}} w^{\frac{\sigma}{1-\sigma}}. \tag{28}$$

This is a little more elegant, as now the discount factor is expressed as a function of the single choice variable ε and external circumstances w and θ. Times of high ε will induce the firm to hire more labor, so ε and n are really not two separate influences in (25). We see the effect of labor is to change the effective coefficient on ε, and to add the wage rate as a new shock. A measurement of the coefficient on ε with constant wages is not the pure coefficient of substitutability.

The discount factors (25) and (28) have important lessons going forward. The production-based discount factor is not necessarily as simple as just productivity raised to a power, and wages or labor inputs – or other inputs – appear as additional natural productivity shocks.

Solving (27) for ε, and using the constraint to find the Lagrange multiplier λ, we can express the productivity choice as[2]

$$\frac{\varepsilon^{1+\alpha}}{\theta^{1+\alpha}} = \frac{\left(m\theta^{\frac{1}{1-\sigma}} w^{-\frac{\sigma}{1-\sigma}}\right)^{\frac{1+\alpha}{\alpha-\frac{\sigma}{1-\sigma}}}}{E\left[\left(m\theta^{\frac{1}{1-\sigma}} w^{-\frac{\sigma}{1-\sigma}}\right)^{\frac{1+\alpha}{\alpha-\frac{\sigma}{1-\sigma}}}\right]}. \tag{30}$$

The firm chooses larger productivity in states with higher discount factors, higher natural productivity shocks, and lower wages. Wages act like the natural productivity shocks.

6. Intertemporal production

Next we generalize the idea to a standard intertemporal context. The firm's objective is

$$\max E \sum_{t=1}^{\infty} \Lambda_t (y_t - i_t).$$

where Λ_t is the stochastic discount factor, with $m_{t+1} = \Lambda_{t+1}/\Lambda_t$, $\Lambda_0 = 1$, y is output and i is investment. (It is more convenient in this dynamic setting to write the problem in terms of the level of the discount factor Λ, rather than the growth rate m. The level is more convenient when thinking about time zero contingent claims or continuous time. The growth m can be a bit more convenient in discrete time recursive formulations, but it's easier to write $m_{t+1} = \Lambda_{t+1}/\Lambda_t$, than

[2]From (27),

$$\begin{aligned} m\theta^{\frac{1}{1-\sigma}} w^{-\frac{\sigma}{1-\sigma}} &= \left[\frac{\lambda(1+\alpha)}{\sigma^{\frac{\sigma}{1-\sigma}}}\right] \frac{\varepsilon^{\alpha-\frac{\sigma}{1-\sigma}}}{\theta^{\alpha-\frac{\sigma}{1-\sigma}}} \\ \left(m\theta^{\frac{1}{1-\sigma}} w^{-\frac{\sigma}{1-\sigma}}\right)^{\frac{1+\alpha}{\alpha-\frac{\sigma}{1-\sigma}}} &= \left[\frac{\lambda(1+\alpha)}{\sigma^{\frac{\sigma}{1-\sigma}}}\right]^{\frac{1+\alpha}{\alpha-\frac{\sigma}{1-\sigma}}} \frac{\varepsilon^{1+\alpha}}{\theta^{1+\alpha}} \end{aligned} \tag{29}$$

Taking the expectation and using the productivity choice constraint,

$$E\left[\left(m\theta^{\frac{1}{1-\sigma}} w^{-\frac{\sigma}{1-\sigma}}\right)^{\frac{1+\alpha}{\alpha-\frac{\sigma}{1-\sigma}}}\right] = \left[\frac{\lambda(1+\alpha)}{\sigma^{\frac{\sigma}{1-\sigma}}}\right]^{\frac{1+\alpha}{\alpha-\frac{\sigma}{1-\sigma}}}$$

substituting back in (29) we have (30).

the corresponding cumulation.) Production possibilities are given by

$$
\begin{aligned}
y_t &= \varepsilon_t f(k_t) \\
k_{t+1} &= (1-\delta)k_t + i_t \\
1 &= E_\tau \left[\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}} \right)^{1+\alpha} \right].
\end{aligned} \tag{31}
$$

There is a separate constraint for each period $t+1$. In adapting the productivity choice set to this dynamic context, we have to specify when the expectation is taken. I allow the expectation in (31) to be conditional, $\tau = t$, unconditional, $\tau = 0$, or in between. The distinction matters for the dynamic properties of the chosen ε_t given a discount rate process, which I explore below, but it makes no difference to the discount rate formulas here.

Analogously to (15) and (16), the first-order conditions are

$$
\frac{\partial}{\partial k_{t+1}} : \; 1 = E_t \left\{ m_{t+1} \left[\varepsilon_{t+1} f_k(k_{t+1}) + (1-\delta) \right] \right\} \tag{32}
$$

and

$$
\frac{\partial}{\partial \varepsilon_{t+1}} : \; \Lambda_{t+1} f(k_{t+1}) = \lambda_{t+1} (1+\alpha) \frac{\varepsilon_{t+1}^{\alpha}}{\theta_{t+1}^{1+\alpha}}. \tag{33}
$$

The Lagrange multiplier λ_{t+1} applies to the time $t+1$ constraint in (31). It is known at time τ. If $\tau = 0$, λ_{t+1} varies over time, but is constant across states of nature at time$t$. If $\tau = t$, then λ_{t+1} is a time-t random variable.

Equation (32) again says that the investment return should be correctly priced,

$$
1 = E_t \left(m_{t+1} R^I_{t+1} \right); \; R^I_{t+1} \equiv \varepsilon_{t+1} f_k(k_{t+1}) + (1-\delta). \tag{34}
$$

Given the choice of productivity ε_{t+1}, invest in capital as usual. The investment return now includes the depreciated value of capital after a period. Equation (33) again says to produce more in high marginal utility, discount factor or contingent claim price states, and in high natural productivity θ states.

A discount factor that prices zero-cost portfolios at time t has

$$
0 = E_t(m^*_{t+1} R^e_{t+1})
$$

and can be scaled by any time t random variable; $b_t m^*_{t+1}$ also prices zero cost portfolios. A convenient zero-cost portfolio discount factor is thus

$$m_{t+1}^{*} = \left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha} / \left(\frac{\theta_{t+1}}{\theta_t}\right)^{1+\alpha}. \tag{35}$$

One could use the same discount factor as before,

$$m_{t+1}^{*} = b_t \varepsilon_{t+1}^{\alpha} / \theta_{t+1}^{1+\alpha}, \tag{36}$$

for any b_t. However, productivity ε_t, like consumption, typically is very persistent and grows over time, and θ_t should have similar properties. So, while (36) with $b_t = 1$, say, prices zero-cost portfolios, its conditional mean varies strongly over time, its implied risk-free rate is potentially counterfactual and time-varying, and it is potentially non-stationary violating the assumptions of all time-series empirical work. Choosing growth rates as at least an initial scaling, as in (35) is wise at least for typical time-series applications. Analogously, we typically use $m_{t+1} = \beta\left(c_{t+1}/c_t\right)^{-\gamma}$, though $m_{t+1}^{*} = c_{t+1}^{-\gamma}$ prices zero cost portfolios just as well.

One can scale further by any convenient time-t random variable. For example, one can produce a given shadow or measured risk-free rate R_t^f with

$$m_{t+1}^{*} = \frac{1}{R_t^f} \frac{\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha} / \left(\frac{\theta_{t+1}}{\theta_t}\right)^{1+\alpha}}{E_t\left[\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha} / \left(\frac{\theta_{t+1}}{\theta_t}\right)^{1+\alpha}\right]}, \tag{37}$$

with or without ε_t and θ_t in the denominators.

We can also scale the discount factor to price the investment return, and thereby display a full production-based discount factor that prices all returns,

$$m_{t+1} = \frac{\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha} / \left(\frac{\theta_{t+1}}{\theta_t}\right)^{1+\alpha}}{E_t\left[\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha} / \left(\frac{\theta_{t+1}}{\theta_t}\right)^{1+\alpha} R_{t+1}^I\right].}. \tag{38}$$

(The expectation in the denominator is a time-t random variable, so it fits in to the rubric of (36). It prices R_{t+1}^I and all zero cost portfolios R_{t+1}^e so it prices all returns R_{t+1}.)

More complex models of intertemporal production just make the formula for the investment return R^I more complicated. They only change the discount factor for zero-cost portfolios to the extent that variable inputs such as labor show up in the production function. For example, add adjustment costs and variable labor supply to the intertemporal production function. The firm's problem is now

$$\max E \sum_{t=1}^{\infty} \Lambda_t (y_t - i_t - w_t n_t)$$

subject to

$$\begin{aligned} y_t &= \varepsilon_t f(k_t, n_t) - \psi(i_t, k_t) \\ k_{t+1} &= (1-\delta)k_t + i_t \\ 1 &= E_\tau \left[(\varepsilon_{t+1}/\theta_{t+1})^{1+\alpha} \right]. \end{aligned}$$

The intertemporal first-order condition becomes

$$1 = E_t \left[m_{t+1} \frac{\varepsilon_{t+1} f_k(k_{t+1}, n_{t+1}) - \psi_k(i_{t+1}, k_{t+1}) + (1-\delta)\left[1 + \psi_i(i_{t+1}, k_{t+1})\right]}{1 + \psi_i(i_t, k_t)} \right] \tag{39}$$

which we can write as usual

$$1 = E_t \left(m_{t+1} R^I_{t+1} \right).$$

We now have a labor first-order condition,

$$\varepsilon_{t+1} f_n(k_{t+1}, n_{t+1}) = w_{t+1},$$

and productivity choice,

$$\Lambda_{t+1} f(k_{t+1}, n_{t+1}) = \lambda_{t+1} (1+\alpha) \frac{\varepsilon^{\alpha}_{t+1}}{\theta^{1+\alpha}_{t+1}}. \tag{40}$$

Relative to (33), this productivity choice condition is only different by the substitution of $f(k_{t+1}, n_{t+1})$ in place of $f(k_{t+1})$. This substitution means we have an extra source of natural productivity shock, which we can express in terms of employment n or wage, just as in the one-period model with labor. Using a standard Cobb-Douglas production function

$$f(k, n) = k^{1-\sigma}_{t+1} n^{\sigma}_{t+1}$$

and since we can scale by any variable known at time t, we can write zero-cost discount factors as

$$m^*_{t+1} = \frac{\varepsilon^{\alpha}_{t+1}}{\theta^{1+\alpha}_{t+1}} \frac{1}{n^{\sigma}_{t+1}};\; m^*_{t+1} = \frac{\varepsilon^{\alpha}_{t+1}}{\theta^{1+\alpha}_{t+1}} \left(\frac{k_{t+1}}{n_{t+1}} \right)^{\sigma};\; \text{ or } m^*_{t+1} = \frac{\varepsilon^{1+\alpha}_{t+1}}{\theta^{1+\alpha}_{t+1}} \frac{1}{y_{t+1}}.$$

using the first order condition for labor input,

$$\varepsilon_{t+1} \sigma k^{1-\sigma}_{t+1} n^{\sigma-1}_{t+1} = w_{t+1},$$

we can write

$$m^*_{t+1} = \frac{\varepsilon_{t+1}^{\alpha-\frac{\sigma}{1-\sigma}}}{\theta_{t+1}^{1+\alpha}} w_{t+1}^{\frac{\sigma}{1-\sigma}}.$$

These are the same formulas as the one-period problem (25) (28), with time subscripts. In this intertemporal context it is likely to be useful to use growth rates, e. g.

$$m^*_{t+1} = b_t \frac{(\varepsilon_{t+1}/\varepsilon_t)^{\alpha}}{(\theta_{t+1}/\theta_t)^{1+\alpha}} \frac{1}{(n_{t+1}/n_t)^{\sigma}}; \;\; m^*_{t+1} = b_t \frac{(\varepsilon_{t+1}/\varepsilon_t)^{1+\alpha}}{(\theta_{t+1}/\theta_t)^{1+\alpha}} \frac{1}{(y_{t+1}/y_t)}; \tag{41}$$

or

$$m^*_{t+1} = b_t \frac{(\varepsilon_{t+1}/\varepsilon_t)^{\alpha-\frac{\sigma}{1-\sigma}}}{(\theta_{t+1}/\theta_t)^{1+\alpha}} \left(\frac{w_{t+1}}{w_t}\right)^{\frac{\sigma}{1-\sigma}} \tag{42}$$

where b_t can also be set as convenient.

Belo (2006) investigates a CES production function

$$y_t = \varepsilon_t \left\{ (\omega k_t)^{\frac{\sigma-1}{\sigma}} + [(1-\omega)\, n_t]^{\frac{\sigma-1}{\sigma}} \right\}^{\frac{\sigma}{\sigma-1}}.$$

Using this definition and the first-order condition for labor, one can impute the productivity shock from the labor share and labor/output ratio, without needing capital data,

$$\varepsilon_t = \frac{1}{(1-\omega)} \left\{ \frac{w_t n_t}{y_t} \right\}^{\frac{\sigma}{\sigma-1}} \left\{ \frac{y_t}{n_t} \right\},$$

or in growth rates,

$$\frac{\varepsilon_{t+1}}{\varepsilon_t} = \left\{ \frac{w_{t+1} n_{t+1}/y_{t+1}}{w_t n_t / y_t} \right\}^{\frac{\sigma}{\sigma-1}} \left\{ \frac{y_{t+1}/n_{t+1}}{y_t/n_t} \right\}. \tag{43}$$

Together with a zero-cost discount factor

$$m^*_{t+1} = b_t \frac{(\varepsilon_{t+1}/\varepsilon_t)^{1+\alpha}}{(\theta_{t+1}/\theta_t)^{1+\alpha}} \frac{1}{(y_{t+1}/y_t)} \tag{44}$$

this is a particularly nice form, leading to a discount factor with output growth, labor share growth, and labor/output ratio growth as factors.

6.1 Dynamic productivity choice

In this section, I consider the choice of productivity ε given discount factors and natural productivity θ. We want some assurance that the productivity choice described by this model is sensible. The previous calculations just say how to construct a discount factor given the productivity choice. This section also analyzes the effect of the time τ at which expectations are taken, and the related question when firms make decisions to change the shock distribution.

We face some natural and related questions. First, macroeconomic models usually feature serially correlated productivity, say

$$\log \varepsilon_{t+1} = \rho \log \varepsilon_t + v_{t+1},$$

often including a random walk $\rho = 1$ or a unit root component, such as a random walk plus AR(1). Can our model produce this kind of productivity process? The answer is yes, but only under some specifications. Second, each time $t+1$ has a separate constraint (31). The constraints are not connected at all, and the distribution of ε_t can be entirely different from that of ε_{t+1}. This behavior is particularly troublesome in continuous time.

An immediate resolution of both concerns is similar: Even though the firm *can* choose wildly different distributions at adjacent moments of time, if circumstances move continuously and persistently, so will the choices. A constraint set that is smoother over time is still desirable, and I close with some thoughts on how to achieve it.

To express the productivity choice ε in terms of the discount factor in this intertemporal setting, we can impose the constraint and find [3]

$$\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{\alpha} = \frac{\Lambda_{t+1}\theta_{t+1}f(k_{t+1})}{\left\{E_\tau\left[\left(\Lambda_{t+1}\theta_{t+1}f(k_{t+1})\right)^{\frac{1+\alpha}{\alpha}}\right]\right\}^{\frac{\alpha}{1+\alpha}}}. \tag{46}$$

If we specify $\tau = t$, then $f(k_{t+1})$ drops out. But if the expectation is taken earlier, or when we generalize to $f(k_{t+1}, n_{t+1})$, it remains inside the expectation. At this level of generality I don't see a way to separately express the productivity ε and capital k choice given only the discount factor and natural shocks.

Let us simplify with $f(k_t) = 1$, as in general this will not be a major source of dynamics especially at high frequency. Start also with $\theta = 1$, so we can see how productivity responds to

[3] Algebra: From (33),

$$\frac{\Lambda_{t+1}\theta_{t+1}f(k_{t+1})}{\lambda_{t+1}(1+\alpha)} = \frac{\varepsilon^{\alpha}_{t+1}}{\theta^{\alpha}_{t+1}} \tag{45}$$

$$\left[\frac{\Lambda_{t+1}\theta_{t+1}f(k_{t+1})}{\lambda_{t+1}(1+\alpha)}\right]^{\frac{1+\alpha}{\alpha}} = \frac{\varepsilon^{1+\alpha}_{t+1}}{\theta^{1+\alpha}_{t+1}}$$

Taking time 0 expectations, imposing the constraint, and raising the result to the $\alpha/(1+\alpha)$ power,

$$\frac{1}{\lambda_{t+1}(1+\alpha)}\left[E_\tau\left\{\left[\Lambda_{t+1}\theta_{t+1}f(k_{t+1})\right]^{\frac{1+\alpha}{\alpha}}\right\}\right]^{\frac{\alpha}{1+\alpha}} = 1.$$

Substituting back in (45), we have (46).

discount factors alone. We have

$$\alpha \log \varepsilon_{t+1} = \log \Lambda_{t+1} - \left\{ \log \left[E_\tau \left(\Lambda_{t+1}^{\frac{1+\alpha}{\alpha}} \right) \right] \right\}^{\frac{\alpha}{1+\alpha}}.$$

If $\tau = t$, then indeed we will have an i.i.d. productivity *level* ε_t. If τ happens earlier though, productivity ε_t follows a moving average, accumulating shocks between $\tau + 1$ and $t + 1$. And fixing $\tau = 0$, $\log \varepsilon_{t+1}$ is as persistent as the level of the discount factor Λ_{t+1}. This is a practical argument for specifying $\tau = 0$ and generalizing the two period model to a dynamic model by using an unconditional expectation in (31) rather than conditional.

To see these statements explicitly, suppose that the discount factor follows a log random walk,

$$\log \Lambda_{t+1} - \log \Lambda_t = -\mu_\Lambda - \sigma_\Lambda v_{t+1},\ v_{t+1} \sim \mathcal{N}(0,1). \tag{47}$$

($\log \Lambda_{t+1} - \log \Lambda_t = \log m_{t+1}$ the usual one-period discount factor.) Now, working out the conditional expectation on its right hand side, (46) gives

$$\alpha \log \varepsilon_{t+1} = \log \Lambda_{t+1} - \log \Lambda_\tau + \left(\mu_\Lambda - \frac{1}{2} \frac{1+\alpha}{\alpha} \sigma_\Lambda^2 \right) (t + 1 - \tau) \tag{48}$$

or

$$\alpha \log \varepsilon_{t+1} = -\sigma_\Lambda \sum_{j=1}^{t-\tau+1} v_{\tau+j} - \frac{1}{2} \frac{1+\alpha}{\alpha} \sigma_\Lambda^2 (t + 1 - \tau).$$

If $\tau = t$, and the constraint is purely conditional $E_t(\varepsilon_{t+1}^{1+\alpha}) = 1$, then indeed the *level* of technology ε_{t+1} is i.i.d., and responds only to growth in the discount factor Λ_{t+1}, or the level of ε_{t+1} is proportional to the level of m_{t+1}. This behavior is unlike consumption-based models in which $\log m_{t+1} = \log \beta - \gamma \log \Delta c_{t+1}$, or $\log \Lambda_{t+1} = (t+1) \log \beta - \gamma \log c_{t+1}$. However, if τ comes earlier, then productivity follows a longer and longer moving average. And if $\tau = 0$, then productivity follows a random walk with drift, and productivity growth rates rather than levels are proportional to the one-period discount factor m_{t+1}.

We also obtain a persistent productivity process if we specify a persistent natural or underlying technology shock θ_t. Equation (46) then implies that ε_{t+1} will inherit the persistence of θ_{t+1}, modified in the direction of the discount factor. Now even if the constraint is conditional $\tau = t$, productivity will include the θ random walk component.

To see these points explicitly, suppose the natural productivity shock follows

$$\log \theta_{t+1} = \rho \log \theta_t + \mu_\theta + \sigma_\theta w_{t+1},\ w_{t+1} \sim \mathcal{N}(0,1),$$

and keep $\rho = 1$ in mind as an important case. Keep $f(k_t) = 1$, the random-walk discount factor process (47) and consider the $\tau = t$ conditional case in which chasing the discount factor alone produces an i.i.d. productivity choice ε_t. We then have

$$\frac{\alpha}{1+\alpha}\log\left\{E_t\left[(\Lambda_{t+1}\theta_{t+1})^{\frac{1+\alpha}{\alpha}}\right]\right\} = \log\Lambda_t + (\mu_\theta - \mu_\Lambda + \rho\log\theta_t) + \frac{1}{2}\left(\frac{1+\alpha}{\alpha}\right)\left(\sigma_\theta^2 + \sigma_\Lambda^2 - 2\sigma_{\theta\Lambda}\right),$$

so writing (46) in logs, the chosen productivity follows

$$\alpha\log\varepsilon_{t+1} = \alpha\log\theta_{t+1} + (\sigma_\theta w_{t+1} - \sigma_\Lambda v_{t+1}) - \frac{1}{2}\left(\frac{1+\alpha}{\alpha}\right)\left(\sigma_\theta^2 + \sigma_\Lambda^2 - 2\sigma_{\theta\Lambda}\right) \tag{49}$$

Log productivity ε_t is composed of a persistent component, $\alpha\log\theta_{t+1}$, which can be a random walk if $\rho = 1$, plus an i.i.d. component from the discount factor. To the extent that the firm wishes to distort ε_{t+1} away from θ_{t+1} for a period in order to chase a higher contingent claim price, it must pay by accepting a lower mean of ε_{t+1}, represented in the final term.

Why is the time τ when we take expectations that define the productivity-choice constraint important? Figure 4 illustrates a simple example. Let there be two periods, and two states in each period with probability 1/2 and an outcome $x = A, B, C, D$. If we impose a constraint conditioned at time 1, $E_1(x_2) = 1$ that constraint is

$$\frac{A+B}{2} = 1; \frac{B+C}{2} = 1. \tag{50}$$

If we impose a constraint conditioned at time 0, $E_0(x_2) = 1$ it is

$$\frac{A+B+C+D}{4} = 1. \tag{51}$$

Now, the conditional constraint (50) implies the unconditional constraint (51). But the conditional constraint is more stringent. The unconditional constraint (51) allows the firm to switch output from (A, B) to (C, D). For example, it allows

$$\frac{A+B}{2} = 1.5; \frac{B+C}{2} = 0.5.$$

The pure conditional constraint does not allow the latter substitution.

The unconditional constraint is equivalent to a set of conditional constraints,

$$E_1(x_2) = z_1;\ E_0(z_1) = 1. \tag{52}$$

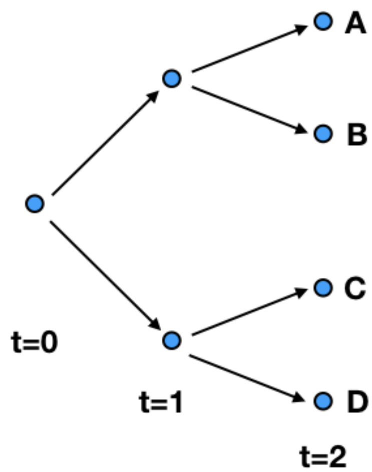

Figure 4: Information tree.

In our case, limiting the $1+\alpha$ moment, the unconditional constraint is similarly recursive. The constraint

$$\left[E_0(x_2^{1+\alpha})\right]^{\frac{1}{1+\alpha}} = 1$$

is equivalent to the sequence of constraints

$$z_1 = \left[E_1(x_2^{1+\alpha})\right]^{\frac{1}{1+\alpha}}$$
$$1 = \left[E_0\left(z_1^{1+\alpha}\right)\right]^{\frac{1}{1+\alpha}}.$$

So the $\tau < t$ assumption in the productivity choice constraint (31) amounts to applying the same productivity-choice idea to the constraint itself. We can write it

$$z_{t,t+1} = \left\{E_t\left[\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]\right\}^{\frac{1}{1+\alpha}} \tag{53}$$

$$z_{t-j,t+1} = \left\{E_{t-j}\left[\left(z_{t-j+1,t+1}\right)^{1+\alpha}\right]\right\}^{\frac{1}{1+\alpha}}; j = 1, 2, ...\tau \tag{54}$$

$$z_{\tau,t+1} = 1. \tag{55}$$

One naturally thinks of choosing a random variable at time 0 subject to a constraint on a time 0 expectation as involving less information, and choosing at time t as involving more information, but the opposite is true. Time zero contingent claim markets are subtle. When one chooses a time $t+1$ random variable at time t, there is only that one time t information set to consider. When one chooses a time $t+1$ random variable at time 0, one pre-plans what that time

t choice will be for every possible time t information set.

What is the more reasonable assumption? (Besides a desire to reverse-engineer the “right” answer.) This dynamic model brings up the issue of time. The firm's actions to transform across states of nature also involve time. The farmer plants seeds in different fields in the spring, but after that there is little he or she can do to transform output across weather states. The electric utility buys flexible or fuel-optimized equipment, but after that there is little it can do to transform output across states indexed by fuel costs. It makes sense to allow the firm more flexibility across states of nature if it has more time to rearrange things, and less flexibility as the time of a shock approaches.

Thinking about a shock at time 2, then, the firm will want to take action at time 0 that alter its time 2 distribution. The formulation (55) allows us to express that desire via state variables. The firm will take actions at time 0 that that expand or contract its constraint set z_1 that it will face at time 1 for time 2 shocks. The state variable $z_{s,t+1}$ is a shock much as productivity ε_{t+1} itself is, in the fashion of Mertonian state variables. The intertemporal firm will want to modify the distribution of the state variable z.

With this thought in mind, E_0 seems the right choice. The process of adjusting the time-t productivity may begin anytime before t, as captured by evolution of state variables $z_{s,t+1}$ as s proceeds forward. The firm may choose to start moving $z_{s,t+1}$ around earlier or later. One may wish to add θ shocks to the z choice in (54), or vary the value of α over horizon.

However, by writing $\tau = 0$, or more generally not advancing τ with t, and by imposing a different constraint at each time t, it is hard to write the problem in a recursive form with a limited number of state variables, which is crucial for solving more complex models numerically. The description here is most suited to a time-0 contingent claim perspective. I offer some thoughts below on how this might be accomplished.

The recursion (53)-(55) also allows us to answer the question, what if the firm re-optimizes at period $\tau < s < t$? With a constraint $\left\{E_s\left[(\varepsilon_{t+1}/\theta_{t+1})^{1+\alpha}\right]\right\}^{\frac{1}{1+\alpha}} = z_{s,t+1}$, the firm makes the same choices if it re-optimizes.

With $\tau < t$, and most easily $\tau = 0$, the dynamic problem generalizes to continuous time. The firm's problem is

$$\max E \int_{t=0}^{\infty} \Lambda_t \left[\varepsilon_t f(k_t) - i_t\right] dt \text{ s.t.}$$

$$\begin{aligned} dk_t &= -\delta k_t + i_t dt \\ 1 &= E\left[\left(\frac{\varepsilon_t}{\theta_t}\right)^{1+\alpha}\right]. \end{aligned} \tag{56}$$

There is a different constraint (56) for each time t.The productivity choice first-order condition is

$$\Lambda_t f(k_t) = \lambda_t (1+\alpha) \frac{\varepsilon_t^{\alpha}}{\theta_t^{1+\alpha}}.$$

Given ε and θ, this condition identifies the discount factor Λ_t up to differentiable functions of time $f(k_t)$and λ_t. Explicitly, we can evaluate the constraint as before, yielding

$$\frac{\varepsilon_t^{\alpha}}{\theta_t^{1+\alpha}} = \frac{\Lambda_t f(k_t)}{\left(E\left\{ \left[\Lambda_t \theta_t f(k_t)\right]^{\frac{1+\alpha}{\alpha}} \right\}\right)^{\frac{\alpha}{1+\alpha}}}.$$

The denominator varies as a differentiable function of time. Thus, in the basic asset pricing relation

$$E_t dR_t - r_t^f dt = -E_t \left(dR_t \frac{d\Lambda_t}{\Lambda_t} \right)$$

with

$$r_t^f dt = -E_t \left(\frac{d\Lambda_t}{\Lambda_t} \right),$$

we know the diffusion component, i.e.

$$\frac{d\Lambda_t}{\Lambda_t} = -r_t^f dt + d\left(\frac{\varepsilon_t^{\alpha}}{\theta_t^{\alpha}} \right) / \left(\frac{\varepsilon_t^{\alpha}}{\theta_t^{\alpha}} \right).$$

We can thus use any r_t^f to describe excess returns. Defining

$$dR_t^e = dR_t^i - dR_t^j$$

then

$$E_t \left(dR_t^e \right) = -E_t \left(dR_t^e \frac{d\Lambda_t}{\Lambda_t} \right) = -E_t \left[dR_t^e d\left(\frac{\varepsilon_t^{\alpha}}{\theta_t^{\alpha}} \right) / \left(\frac{\varepsilon_t^{\alpha}}{\theta_t^{\alpha}} \right) \right].$$

The intertemporal first order condition is standard.

7. Alternatives

This extension to dynamic problems is not as pretty as I would like it to be. Fundamentally, the constraint $E\left[(\varepsilon_t/\theta_t)^{1+\alpha}\right] = 1$ allows completely different random variables ε_t at each date t. One would suppose that the distribution of productivity at time t would not be that different from the distribution of productivity at time $t+\Delta$. In the farming and electric utility examples, the choice of fields and machines do not allow one exposure to shocks at one instant, and a different exposure 10 minutes later.

The above examples gives some reassurance that the firm will not *choose* wildly different productivity distributions at closely adjacent time periods, given that circumstances described by the distribution of discount factors Λ and natural productivity θ do not move that quickly. So the model will not generally produce crazy predictions. Still, the description of production sets is inelegant.

7.1 Alternative 1: A growth constraint

This situation is analogous to utility theory. The utility function or $E\sum\beta^t u(c_t)$ or $E\int e^{\delta t}u(c_t)dt$ can value consumption streams that have different distributions at each instant. Consumers facing continuous incentives do not choose such paths, but they could. The resolution of this sort of puzzle for consumption is to recognize that all consumption goods are durable at short enough horizon. Even a pizza is durable for 10 minutes.[4] This modification tends not to be used however, because the first-order conditions for durable goods are more complex.

A similar situation applies to production sets. We would like a productivity-choice set in which productivity at nearby dates must have similar distributions, and the distributions can more easily diverge from each other as the time between production events increases. Doing so, however, complicates the first-order conditions. Changing productivity ε_t at time t now influences the set from which future productivity $\varepsilon_{t+\tau}$ is chosen. Future discount factors as well as current ones enter the choice of ε_t, and inverting to find discount factors from productivity choices involves unwinding that intertemporal choice.

For example, we might write the choice set as a constraint on the *growth* of productivity:

$$E_\tau\left[\left(\frac{\varepsilon_{t+\Delta}}{\theta_{t+\Delta}}/\frac{\varepsilon_t}{\theta_t}\right)^{1+\alpha}\right]=1. \tag{57}$$

Somewhat analogously, with no depreciation, a durable purchase *changes* the flow of consumption services. This specification is equivalent to writing a natural shock $\theta_{t+\Delta}$ that incudes the previous actual productivity ε_t, as part of the natural starting point. If one buys machines with a given state-contingent output, then the natural starting point for next period is just to use those machines. (This discussion implicitly also writes the state of nature at time-t as a history plus a stationary and repeated shock."Rain" at time $t+1$ is the same addition to the cumulative state as "rain" at time t is to that cumulative state. This is a common and useful restriction on uncertainty facing the firm which I have not so far imposed.)

[4]See Hindy and Huang (1992). This discussion makes light of serious mathematical questions, such as what does $c_t dt$, or $\varepsilon_t f(k_t)dt$ mean, and how do you integrate $\int u(c_t)dt$ or $E\int\varepsilon_t f(k_t)dt$ if the distribution c_t and ε_t can change discontinouously at every instant.

Specification (57) also leads to a natural continuous-time expression,

$$E_\tau\left[\left(\frac{d\left(\varepsilon_t/\theta_t\right)}{\theta_t/\varepsilon_t}\right)^{1+\alpha}\right]=0. \tag{58}$$

Specifications (57) or (58) naturally result in technology $\varepsilon_{t+\tau}$ that wanders further away from its initial value ε_t, and from the underlying shock θ, for a longer time horizon, even when $\tau = t$ which produced an i.i.d. productivity before.

So far so good, but the first-order conditions become more complicated, because changing ε_t changes the choice set for all subsequent $\varepsilon_{t+\tau}$. The resulting first-order conditions are harder to unwind to a discount factor. With a constraint on the growth of productivity (57), the firm's problem is

$$\max E\sum_{t=0}^{\infty}\Lambda_t\left[\varepsilon_t f(k_t)-i_t\right] \text{ s.t.}$$

$$k_{t+1}=(1-\delta)k_t+i_t$$

$$1=E_\tau\left[\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}/\frac{\varepsilon_t}{\theta_t}\right)^{1+\alpha}\right].$$

Now the first-order condition with respect to ε_{t+1} is

$$\Lambda_{t+1}f(k_{t+1})=\lambda_t(1+\alpha)\left(\frac{\varepsilon_{t+1}^{\alpha}}{\theta_{t+1}^{1+\alpha}}/\frac{\varepsilon_t^{1+\alpha}}{\theta_t^{1+\alpha}}\right)-\lambda_{t+1}(1+\alpha)E_{t+1}\left(\frac{\varepsilon_{t+2}^{1+\alpha}}{\theta_{t+2}^{1+\alpha}}\right)/\frac{\varepsilon_{t+1}^{2+\alpha}}{\theta_{t+1}^{1+\alpha}}. \tag{59}$$

This first-order condition is the same for the conditional $\tau = t$ and unconditional $\tau = 0$ specifications. The only difference is that in the conditional case λ_t is a time t random variable, where in the unconditional case it is a time-varying constant known at time 0.

We can use (59) recursively to write the productivity-choice first-order condition as

$$\sum_{j=1}^{\infty}E_{t+1}\left[\Lambda_{t+j}\varepsilon_{t+j}f(k_{t+j})\right]=\lambda_t(1+\alpha)\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}/\frac{\varepsilon_t}{\theta_t}\right)^{1+\alpha}.$$

In this form you can see more clearly that increasing ε_{t+1} at time $t+1$ makes the constraints easier for all future times, and thus has a present discounted benefit.

In these first-order conditions, you see effects similar to those of internal habit or durable goods models. For writing simulation or general equilibrium models, or even for estimation, they are no harder than those. But inferring the discount factor from productivity is not as pretty as in the time-separable cases.

Going forward, we want something tractable, which I think involves a state variable, something like the stock of durable goods in consumption theory or the value function in recursive utility theory, that tracks the distribution of productivity. Then the firm can invest resources to change that state variable. Such a model might also want the tradeoff α to be easier at longer horizons. I leave this as one of many hazy suggestions for future work.

7.2 Alternative 2: Treating time and state symmetrically

A second alternative goes back to first principles. To extend the production-based asset pricing idea to multiple dates, why not treat time and states of nature symmetrically? Write the firm's two-period problem as

$$\max c_0 + E(mc_1) \text{ s.t. } \left\{ \left(\frac{c_0}{\theta_0}\right)^{1+\alpha} + \rho E\left[\left(\frac{c_1}{\theta_1}\right)^{1+\alpha}\right] \right\}^{\frac{1}{1+\alpha}} \leq K, \tag{60}$$

where c denotes the firm's final output sold to consumers, i.e. $c = y - i$, and ρ is a constant parameter like α. This production set is concave and smooth across time and across states of nature. Explicitly, in the finite-state case,

$$\max c_0 + \sum_s \pi(s)m(s)c_1(s) \text{ s.t. } \left\{ \left(\frac{c_0}{\theta_0}\right)^{1+\alpha} + \rho \sum_s \pi(s) \left(\frac{c_1(s)}{\theta_1(s)}\right)^{1+\alpha} \right\}^{\frac{1}{1+\alpha}} \leq K.$$

The first-order conditions to this problem lead immediately to

$$m_1 = \rho \left(\frac{c_1}{c_0}\right)^{\alpha} / \left(\frac{\theta_1}{\theta_0}\right)^{1+\alpha}. \tag{61}$$

This discount factor prices all returns, levels as well as risk premia. It's simpler. The parallel to power utility is immediate. We can generalize this approach to multiperiod problems and continuous time transparently. The integral version of (60) becomes a Dixit-Stiglitz aggregator over time rather than across goods, e.g.

$$\left[E \int_{s=0}^{\infty} \rho^s \left(\frac{c_s}{\theta_s}\right)^{1+\alpha} ds\right]^{\frac{1}{1+\alpha}}.$$

It may be easier to transform across time than across states of nature, and we can capture such an effect as well by using different CES aggregators for state and time. For example,

$$\max c_0 + E(mc_1) \text{ s.t. } \left\{ \left(\frac{c_0}{\theta_0}\right)^{1+\alpha} + \rho \left\{ E\left[\left(\frac{c_1}{\theta_1}\right)^{1+\alpha^*} \right] \right\}^{\frac{1+\alpha}{1+\alpha^*}} \right\}^{\frac{1}{1+\alpha}} \leq K. \tag{62}$$

Then we obtain

$$m = \rho \left\{ E\left[\left(\frac{c_1}{\theta_1}\right)^{1+\alpha^*} \right] \right\}^{\frac{\alpha-\alpha^*}{1+\alpha^*}} \left(\frac{c_1^{\alpha^*}}{\theta_1^{1+\alpha^*}} \right) / \left(\frac{c_0^{\alpha}}{\theta_0^{1+\alpha}} \right)$$

Zero-cost portfolios are priced using α^* (drop all the time-0 variables). But $E(m) = 1/R^f$ is now distorted by the first term in brackets. Intertemporal transformation and risk-transformation are separated.

Why not follow this more elegant approach? One answer is, we then lose the connection to standard production theory. A standard intertemporal production function, say

$$y_t = \varepsilon_t f(k_t) \tag{63}$$

$$k_{t+1} = (1-\delta)k_t + (y_t - c_t) \tag{64}$$

does not produce a pretty CES intertemporal allocation (60). It is not the limit of (62) as $\alpha^* \to \infty$, for example. It does imply a smooth convex set for allocations over time. Derivatives $dc_{t+1}/dc_t = (1-\delta)\varepsilon_{t+1} f_k(k_{t+1})$ are well defined, and $dc_{t+1}^2/dc_t^2 = (1-\delta)\varepsilon_t f_{kk}(k_t) < 0$. But that set not expressible as a CES aggregator of final output $\{c_t\}$, or any other pretty functional form $g(c_0, c_1, ...) = 0$ that invites generalization to include states $c_t(s)$ in parallel with time – or at least I have not been able to express it in such a way and find that generalization.

So, we can follow elegance, and the beautiful symmetry of static utility and production theory exactly. But in so doing we have to throw out the contact with classic production theory. Alternatively, we can add productivity choice to standard production theory as I have so far. That leads to a less elegant but possibly more productive result. But perhaps better ways can be found to write smooth production sets integrating time and risk, and to connect them to the lessons of classic production theory without throwing the latter out and starting over.

However, there may also be good reason to abandon the symmetry between time and state. The underlying economic stories are quite different. We think of transformation over time with a story captured by the usual symbols – some output is put aside or invested to become capita that later produces more output. We think of transformation across states by stories such as planting in fields with different state-sensitivities, investing in machines with different sensitivities, and so on. Until the distribution across states can be expressed with a state variable similar to capital, perhaps keeping time and risk separate is wise.

8. General equilibrium, identification and calibration

In its simplest form, our discount factor is

$$m_{t+1} = \lambda_t \frac{\varepsilon_{t+1}^{\alpha}}{\theta_{t+1}^{1+\alpha}}.$$

What specifications of such a production-based model will be empirically successful? We know from Hansen and Jagannathan (1991) and its many extensions such as Cochrane and Hansen (1992) several properties that a successful discount factor must have. This section takes up this question, as well as the troublesome question of whether and what kinds of natural production shocks θ we need, and how to identify them.

The basic asset pricing formula for excess returns $0 = E(mR^e)$ implies that the expected return is proportional to the covariance of returns with the discount factor,

$$E(R^e) = -\frac{cov(m, R^e)}{E(m)}.$$

This relation implies

$$\frac{E(R^e)}{\sigma(R^e)} = -\frac{\sigma(m)}{E(m)}\rho(m, R^e). \tag{65}$$

To generate the market Sharpe ratio of about 0.5, the discount factor must be volatile, with $\sigma(m)$ on the order of 0.50 or more. That requirement has posed a challenge for consumption-based asset pricing, as consumption itself has a much lower than 50% volatility, and very high risk aversions are hard to swallow. It is hard to generate $\sigma(\Delta c^{-\gamma})$ on the order of 50%. Output and productivity are more volatile than consumption, however, and we have little a-priori feeling about the production curvature coefficient α. This paper is devoted to *lowering* α from its previously standard value, $\alpha = \infty$. So it is likely that achieving a high $\sigma(\varepsilon^{1+\alpha})$ will not be difficult, and the classic equity premium puzzle is not likely to cause much trouble for production-based asset pricing.

The discount factor should have a low and fairly stable conditional mean, to generate a low and relatively stable real riskfree rate $E_t(m) = 1/R_t^f$. Since the conditional mean and the risk premium are separated in these production-based formulas, with the level of the discount factor generated by conventional investment returns, we may anticipate few problems in generating a low and stable conditional mean discount factor.

Relation (65) also holds conditionally, with time t subscripts. Risk premiums vary over time. It is generally felt that time-varying conditional variance, $\sigma_t(m_{t+1})$ should vary over time, as conditional variance $\sigma_t(R^e_{t+1})$ operates on a different time scale and in response to different variables, and time-varying correlations $\rho_t(m_{t+1}, R^e_{t+1})$ are a headache. A time-varying variance of productivity ε may be plausible, time-varying opportunity sets α_t may be plausible, and one can imagine mechanisms parallel to habits in preferences that generate such variation in α_t endogenously.

The most obvious obstacle, however, is the sign of the covariance term. To generate a positive risk premium $E(R^e)$, the discount factor must covary negatively with the ex-post return. In consumption-based asset pricing, $m = \lambda c^{-\gamma}$, the positive correlation of consumption growth with asset returns is consistent with this negative correlation of the discount factor with asset returns and a positive risk premium. Its failure is one of magnitude, not of sign. In production-based asset pricing we have $m = \lambda\varepsilon^{\alpha}/\theta^{1+\alpha}$, however, with $\alpha > 0$. If there are no natural productivity shocks θ, productivity ε is positively correlated with the discount factor. A positive correlation of productivity growth ε with asset returns R^e produces counterfactual negative risk premium $E(R^e)$.

Intuitively, the discount factor, contingent claims price, or marginal utility is high in "bad times," when consumption is low, the stock market is low, and people would really value a marginal dollar. A firm without a θ, without a bias to one state or another, will rearrange its output to produce *m*ore in such high-price "bad times" states.

Now, there are many possibilities to avoid this conundrum. It presumes a one-factor model in which consumption, productivity, and asset returns all move together. Maybe productivity is indeed higher in bad times. Whether productivity is procyclical is a debated in macroeconomics. (Measuring productivity is a headache too.) Yes, real business cycle models generate recessions by productivity shocks, but the rest of macroeconomics in the new-Keynesian DSGE tradition is essentially devoted to disbelief of that proposition. In recessions, firms produce less, but they also shed workers and machines – and especially unproductive workers and machines. Maybe the extension to labor and other inputs will add wages, labor input or other cyclical variables to the discount factor formula with the right sign, as the discount factors with labor and wages (25) (28) suggest.

Asset returns, productivity and consumption are not perfectly correlated. Maybe large components of asset returns are not related to the business cycle, so asset returns can be negatively correlated with productivity and positively correlated with consumption. And not all asset returns are the same. Asset returns contain multiple orthogonal priced factors past the market, including value, size, momentum, term spread, default spread, and others. Maybe productivity is correlated negatively with these additional factors, generating their premiums at least, if not the market premium.

Moreover, the production-based discount factor formula applies to each firm, as the consumption-based discount factor applies to each individual. But, unlike the consumption case, we have detailed data on individual firms, industries, and sectors. The philosophy of production-based asset pricing already says to take these detailed data seriously, and construct many investment returns at a disaggregated level. Who knows where disaggregated information about production-based discount factors using firm-level productivity will lead.

Still, it is unpleasant that the basic model seems to produce the wrong sign. The simplest answer is to include natural productivity shocks θ. A model driven, at least predominantly, by natural productivity shocks θ and not preference shocks will produce the "right" sign – at the cost that now we must face the problem of how to identify natural productivity shocks θ. If there is a high productivity shock θ, other things constant, firms will produce more in that state. Consumers will consume more in that state, and drive down the discount factor or contingent claim price of that state. This lower price causes firms to back off – to lower productivity ε somewhat in the high-θ, low-price state, and to raise productivity somewhat in low-θ, high-price states. The firm essentially buys some insurance. But the insurance may not be (and, as we will see, typically is not) complete. The product $m = \lambda\varepsilon^{\alpha}/\theta^{1+\alpha}$ still moves negatively with θ, so the discount factor moves negatively with consumption and asset returns. Productivity ε and productivity shocks θ are not uncorrelated – in fact, in this example, we expect them to be perfectly negatively correlated.

By analogy, strawberry prices are higher in the winter, yet farmers produce fewer of them. Well, winter is a bad time for producing strawberries. Producers do what they can, building hothouses or growing strawberries in Chile. So they move production towards the high price state. But we still observe higher prices in times of lower output. We also can observe that the price of strawberries is equal to the marginal cost of producing them, and write a production-based strawberry pricing model. But in doing so, we must recognize that the strawberry market is dominated by natural productivity shocks, not preference or sentiment shocks.

8.1 A simple general equilibrium economy

To validate and flesh out this story, focusing on the novel and risk premium parts of these problems, I consider a general equilibrium of the simplest one period model, with a preference shock ϕ as well as a natural productivity shock θ. The bottom line of the model is a formula for equilibrium consumption, productivity and discount factor as a function of both shocks,

$$\log c = \log \varepsilon = \text{const.} + \frac{1+\alpha}{\alpha+\gamma}\log\theta + \frac{\gamma-1}{\alpha+\gamma}\log\phi. \tag{66}$$

$$\log m = \text{const.} - \gamma\frac{1+\alpha}{\alpha+\gamma}\log\theta + \alpha\frac{\gamma-1}{\alpha+\gamma}\log\phi. \tag{67}$$

To derive (66) and (67), add consumers with utility

$$Eu(c) = \sum_s \pi(s)u\left[c(s)\right]$$

where

$$u(c) = \frac{\left[c/\phi\right]^{1-\gamma} - 1}{1-\gamma}.$$

Marginal utility is

$$u'(c) = \frac{c^{-\gamma}}{\phi^{1-\gamma}} = \frac{c(s)^{-\gamma}}{\phi(s)^{1-\gamma}}.$$

The variable ϕ is a preference shock. For each $c(s)$, higher $\phi(s)$ lowers utility. For $\gamma > 1$, higher ϕ raises marginal utility. Thus a higher ϕ is a negative preference shock. Keeping the parallel with consumption-based asset pricing, driven by productivity shocks, it makes sense to investigate a production-based asset pricing model driven by preference shocks. Preference shocks are also increasingly popular in both finance and macroeconomics. Many new-Keynesian models now include preference shocks, at least as a stand-in for financial intermediation shocks. Changing risk aversion is sometimes modeled as a preference shock. Albuquerque et al. (2016) is a good example of a macro-finance paper arguing that preference shocks are needed to understand the data, and particularly the imperfect correlation between asset returns and economic variables. And behavioral finance is all about preference shocks. “Sentiment” or irrationally assessed probabilities are equivalent to ϕ shocks.

Consumers own the firms, and thus have a contingent claim that pays a random amount e. The consumer’s budget constraint is

$$E(me) = E(mc).$$

The consumer’s first-order conditions are

$$m = \lambda u'(c) = \lambda c^{-\gamma}/\phi^{1-\gamma} \tag{68}$$

so consumption is

$$c = m^{-\frac{1}{\gamma}} \phi^{\frac{\gamma-1}{\gamma}} \lambda^{\frac{1}{\gamma}}.$$

Evaluating λ via the budget constraint, the full solution to the consumer's problem is

$$c = E(me) \frac{m^{-\frac{1}{\gamma}} \phi^{\frac{\gamma-1}{\gamma}}}{E\left(m^{1-\frac{1}{\gamma}} \phi^{\frac{\gamma-1}{\gamma}}\right)}. \tag{69}$$

We cannot do better than (68) inverting m from c in this model. Doubling m has no effect on the budget constraint, and no effect on consumption in (69).

Producers have a stock of capital with $f(k) = 1$. They maximize

$$E\left[m\varepsilon f(k)\right] \text{ s.t. } E(\varepsilon^{1+\alpha}/\theta^{1+\alpha}) \leq 1.$$

Producers' first-order conditions are

$$m = \frac{\lambda \varepsilon^{\alpha}}{\theta^{1+\alpha}} \tag{70}$$

Using the constraint to eliminate the Lagrange multiplier λ, the solution to the producer's problem is

$$\frac{\varepsilon^{\alpha}}{\theta^{\alpha}} = \frac{m\theta}{\left\{E\left[(m\theta)^{\frac{1+\alpha}{\alpha}}\right]\right\}^{\frac{\alpha}{1+\alpha}}}.$$

We cannot do better than (70) to invert the discount factor from ε. Doubling m does not affect the constraint, and has no effect on the productivity ε choice.

In equilibrium, consumers own the firm so their endowment equals the firm profit, $e = \varepsilon$, and consumption equals the firm's output, $c = \varepsilon$. (This equality is an important limitation of this static analysis. In a dynamic model, equilibrium requires $c_t = y_t - i_t$. I leave general equilibrium of a dynamic model as one of many loose ends for future work.)

We can find the equilibrium from the planning problem

$$\max Eu(c) = (c/\phi)^{1-\gamma} \text{ s.t. } E\left[(c/\theta)^{1+\alpha}\right] \leq 1.$$

The first-order condition is

$$c^{-\gamma}/\phi^{1-\gamma} = \lambda c^{\alpha}/\theta^{1+\alpha}$$

Imposing the productivity choice constraint, the full solution of the planning problem is

$$(\alpha+\gamma)\log c = \frac{\alpha+\gamma}{1+\alpha}\log\left\{E\left[(\theta/\phi)^{(1-\gamma)\frac{1+\alpha}{\alpha+\gamma}}\right]\right\} + (1+\alpha)\log\theta + (\gamma-1)\log\phi.$$

The constant is not interesting for us, however, so I write the equilibrium as (66)

$$\log c = \log \varepsilon = \text{const.} + \frac{1+\alpha}{\alpha+\gamma} \log \theta + \frac{\gamma-1}{\alpha+\gamma} \log \phi. \tag{71}$$

Using either discount factor formula (68) or (70), the equilibrium discount factor is as given by (67)

$$\log m = \text{const.} - \gamma \frac{1+\alpha}{\alpha+\gamma} \log \theta + \alpha \frac{\gamma-1}{\alpha+\gamma} \log \phi. \tag{72}$$

A claim to consumption, or the output of the firm, has price

$$p = E(mc) = E(m\varepsilon)$$

and thus excess return

$$R^e = \frac{c}{E(mc)} - \frac{1}{E(m)}.$$

In this model the return is perfectly positively correlated with consumption. Scaling by the risk free rate to obtain a quantity independent of the level of the discount factor m,

$$\frac{E\left(R^e\right)}{R^f} = \frac{E(m)E(c)}{E(mc)} - 1$$

Assuming normal distributions, we have

$$\frac{E\left(R^e\right)}{R^f} = \gamma\sigma^2 \left[\frac{1+\alpha}{\alpha+\gamma} \log \theta\right] - \alpha\sigma^2 \left[\frac{\gamma-1}{\alpha+\gamma} \log \phi\right] + (\gamma-\alpha) cov \left[\frac{1+\alpha}{\alpha+\gamma} \log \theta, \frac{\gamma-1}{\alpha+\gamma} \log \phi\right]. \tag{73}$$

Together, (71) (72) and (73) characterize the general equilibrium of this economy.

8.2 Analysis

Under log utility $\gamma = 1$, the economy uses the productivity θ shocks unchanged, ignoring the preference shock ϕ: $\log c = \log \varepsilon = \log \theta$, and $\log m = -\log \theta$. Log utility also generates a positive premium $E(R^e)/R^f = \sigma^2 (\log \theta)$. This case shows immediately how ε and θ can be positively correlated – perfectly, here – and how a production-based model can generate a negative correlation between productivity and the discount factor and thus a positive risk premium.

Next, free up risk aversion, but turn off preference shocks ϕ, and consider an economy driven only by natural productivity shocks θ. Depending on risk aversion γ, equation (71) shows that chosen productivity ε can be either more or less volatile than natural productivity θ. If risk aversion $\gamma > 1$, the (71) term in front of $\log \theta$ is less than one. Producers reduce productivity ε volatility compared to the natural θ shocks, reducing productivity ε in good, high θ, states, in order to raise productivity ε in bad, low θ states. If $\gamma < 1$, however, producers actually choose *more* volatile ε productivity shocks than the natural θ shocks.

In all of these cases however, ε and θ move together, the discount factor m is negatively related to observed productivity and consumption, and the equity premium is positive. The signs are normal, but we have to deal with unmeasured underlying productivity shocks θ, perfectly correlated with measured productivity ε in $m = \lambda\varepsilon^{\alpha}/\theta^{1+\alpha}$.

In the limit $\alpha \to \infty$, the standard case without productivity choice, we also have $\varepsilon = c = \theta$, and the standard consumption-based model with $\log m = -\gamma \log \theta$ and a positive equity premium. The model with productivity choice maintains the same signs and intuition.

Suppose we parallel consumption-based asset pricing by modeling an economy with pure preference shocks ϕ and no underlying technology shocks θ. For the realistic $\gamma > 1$ case, equilibrium consumption in (71) rises with the preference shock. However, the discount factor in (72) also rises with the preference shock so the discount factor is high when consumption is high. Firms have done what they can – with $\alpha < \infty$, the rise in consumption is greater and the rise in marginal utility less than it would be otherwise. But firm responses do not change the sign. The equity premium is negative.

This case verifies our conjecture: a production-based asset pricing model driven by preference shocks and no underlying productivity shocks delivers the wrong sign – output is higher in high marginal utility states, so a claim to consumption or the productivity shock provides insurance and generates a negative risk premium.

This simple result suggests that productivity (and whatever complexities of the production process that stands for) rather than preferences (and whatever complexities the latter stand for, including intermediation and time-varying irrational probability assessments) must be the dominant shock driving the joint behavior of asset returns and macroeconomic fluctuations. This observation is not limited to models with an active production-based margin. If $\alpha = \infty$ but $\theta = 1$, consumption is constant. The discount factor is

$$\log m = \text{const.} + (\gamma - 1)\log \phi.$$

and the expected return on the consumption claim is

$$\frac{E\left(R^{e}\right)}{R^{f}} = -\sigma^{2}\left[(\gamma - 1)\log \phi\right].$$

Pure preference (or "sentiment") shocks can give variation in prices, but they cannot generate the positive association we see between returns and the macroeconomic cycle.

In reality, we should be prepared to see a mixture of preference and productivity shocks. Identification then follows exactly the standard supply and demand parable of economics 101 – and reminds us of just how difficult identification really is, if we do not sweep it under the rug by assuming one side lives in a shock-free world.

What relation between consumption or productivity and discount factors we will see in equilibrium? If there are preference shocks but no productivity shocks, then (71) and (72) obey

$$\log m = \text{const.} + \alpha \log c = \text{const.} + \alpha \log \varepsilon.$$

The data trace out the marginal rate of *transformation* curve and identify production curvature α, for any value of risk aversion γ. The upward sloping line of Figure 5 illustrates this case.

If there are productivity shocks θ but no preference shocks, then (71) and (72) imply

$$\log m = \text{const.} - \gamma \log c = \text{const.} - \gamma \log \varepsilon. \tag{74}$$

The data trace out the marginal rate of substitution curve and identify risk aversion γ, whether or not firms have a technology choice, i.e. for any α. The downward-sloping line of Figure 5 illustrates this case. How did we lose the production-based discount factor and α? The production-based discount factor formula

$$\log m = \text{const.} + \alpha \log \varepsilon - (1 + \alpha) \log \theta \tag{75}$$

is still there. However, with productivity shocks and no preference shocks, ε and θ are perfectly correlated in equilibrium, by (71).

I stress this case, because it seems like an important parable for what we may see in the data. Several papers, discussed in the literature review below, use ad-hoc discount factors based on productivity, and find such negative coefficients. If underlying productivity shocks dominate, then although the discount factor has a positive and structural coefficient on productivity, in (75), an approximate discount factor that uses productivity but does not (somehow) control for the shock θ, will see a negative coefficient. That coefficient is the risk aversion coefficient in this pure case, but in general, with both productivity and preference shocks, it is a mongrel combination of parameters.

When there are both preference and productivity shocks, the data fill out between these two options. The forward regression of $\log m$ on $\log c$ produces a line that is too flat, while the reverse regression of $\log c$ on $\log m$ produces a line that is too steep. Since the errors are not orthogonal to right hand variables, neither regression offers a consistent estimate of either parameter. The left-hand panel of Figure 5 shows an example in which productivity shocks dominate, while the right hand panel shows an example in which preference shocks dominate. With preference and productivity shocks of the same size, the data fill a cloud.

Conventional GMM estimation of consumption-based models hinges crucially on the assumption that there are no preference shocks, and thus all variation comes from productivity shocks. The difficulties of that model, and the imprecision and paradoxically high values of γ it reports suggest that preference shocks may indeed be part of the story.

(The plot simplifies the story by graphing the relation between discount factor m and productivity or consumption. We usually do not have data on m, but the same idea holds for standard estimates. Examining the GMM objective $E(c^{-\gamma}R^e) = 0$, the GMM estimate of γ is inconsistent when the true discount factor includes preference shocks, and likewise that a GMM objective $E(\varepsilon^{\alpha}R^e) = 0$ leads to an inconsistent estimate of α when the discount factor includes productivity shocks.)

To estimate γ or α consistently, we need identifying assumptions. Since all variables shown so far depend on both shocks, and those shocks are unobservable, such assumptions are not immediately obvious. It's easy to make identifying assumptions. It's harder to make correct identifying assumptions. The literature review includes some identification efforts, and the final section includes some thoughts on how to address the issue.

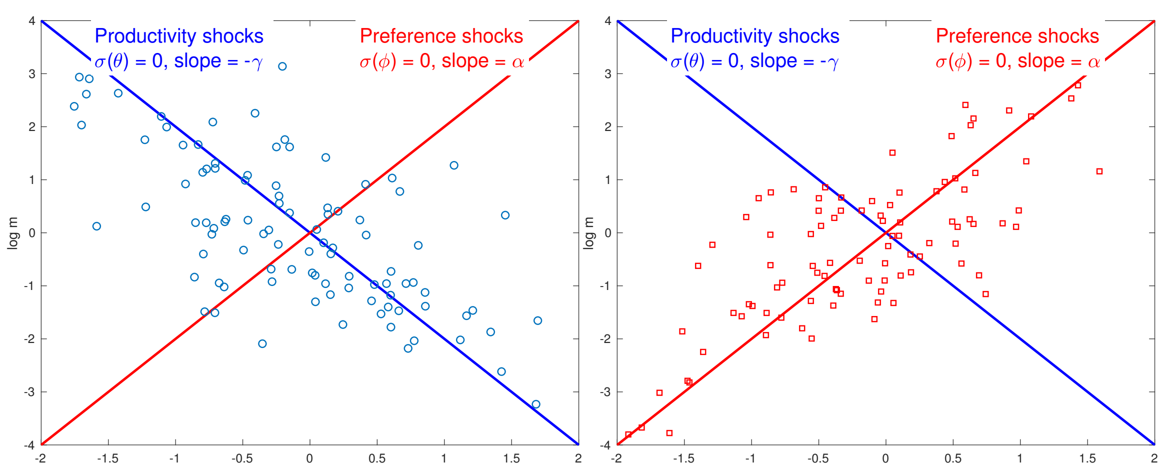

Figure 5: Discount factor $\log m$ vs. consumption $\log c$ in the general equilibrium model. The downward sloping line plots the case of no preference shocks. The upward sloping line plots the case of no productivity shocks. Left panel: The circles show artificial data from a case with larger productivity than preference shocks, $\sigma(\theta) = (\gamma - 1) \times 1$, $\sigma(\phi) = (1 + \alpha) \times 0.4$. Right panel: The squares shod artificial data from a case with larger preference than productivity shocks, $\sigma(\theta) = (\gamma - 1) \times 0.4$, $\sigma(\phi) = (1 + \alpha) \times 1$.

The discount factors with labor inputs (41), (42), (43) and (44) provide additional sources of variation that might help to produce a procyclical discount factor. Labor growth n_{t+1}/n_t and output growth y_{t+1}/y_t lower the discount factor in (41) and (44). This seems to be a helpful step towards the quest to produce a model with a countercyclical discount factor. Equation (43) adds labor share and output per worker, with interesting cyclical properties. However, wages enter positively in (42), which suggests another force pulling us towards a procyclical discount factor. How can the two expressions suggest different directions? Because these are all endogenous variables and all correlated with each other, as θ and ε are. In response to a natural productivity shock, firms hire more labor, drive up wages, and lower actual productivity ε. Both (25) and (28) can hold at the same time because of the different coefficient on ε.

I do not pursue general equilibrium with wages, or the necessarily more ambitious technologies one will need to seriously address data. Still, these simple expressions point to the possibility that with labor and other inputs, including additional shocks or wedges, the natural productivity shocks θ may not have to do all the work of producing a procyclical discount factor.

Equation (42), in which the discount factor is proportional to $(\varepsilon_{t+1}/\varepsilon_t)^{\alpha-\frac{\sigma}{1-\sigma}}$, seems to offer another approach: perhaps for small α and large $\sigma/(1-\sigma)$, the coefficient on ε can become negative, so that higher productivity comes with lower discount factors, even without natural shocks θ or contemporaneous wage movement w. However, $\alpha > \sigma/\left(1-\sigma\right)$ is the condition for a convex problem. (If $\alpha \leq \sigma/(1-\sigma)$, then the firm chooses all of its production in one state, and one should state and impose the condition $\varepsilon \geq 0$.)

8.3 An endowment-economy analogy for production-based asset pricing

This sort of general equilibrium excursion helps us to understand the problems we will face when confronting data, and what kind of measurement or identifying assumptions for shocks θ and ϕ might be useful. However, the guiding philosophy of a production-based asset pricing model is to avoid computing a full general equilibrium. Figure 6 illustrates the idea.

One can approach data with a full general equilibrium economy, incorporating a production function, productivity choice ε and a utility function. Then one finds contingent claim prices or the discount factor from the tangency point of marginal rate of transformation or substitution, represented by the straight line.

Consumption-based asset pricing simplifies the computation. If one models the consumption process, or the productivity ε process *as if* it were an endowment, or a fixed random variable, then one can still read asset prices off marginal rates of substitution alone. Specifically, start with a general equilibrium with natural productivity θ, a curvature parameter α and a

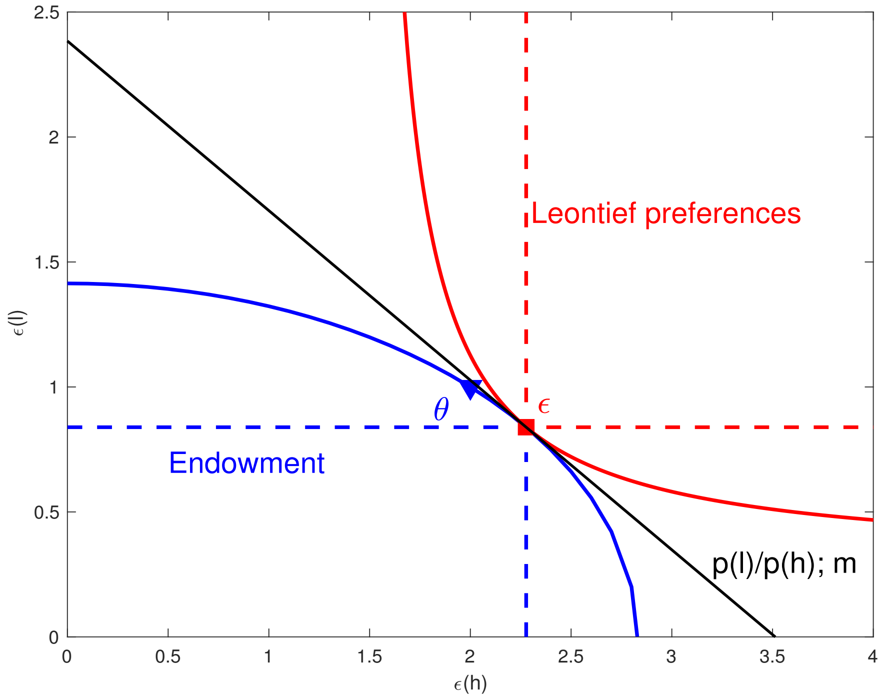

Figure 6: General equilibrium. The outward-bowed curve is a shock-choice set $E\left[(\varepsilon/\theta)^{1+\alpha}\right] \leq 1$ with $\alpha = 1$, $\theta = [2, 1]$. The inward-bowed curve is an indifference curve for a power utility consumer with $u = (c/\theta)^{1-\gamma}$, $\phi = [5, 1]$, $\gamma = 2$. The dashed lines give equivalent endowment economies, i.e. fixed shocks ε or Leontief preferences that deliver the same equilibrium quantities and prices. $p(l)/p(h)$ is a contingent claim price ratio and m is a stochastic discount factor.

chosen productivity ε. Construct a new economy consisting of a fixed-proportions production function calibrated to the observed ε, $\varepsilon^* = \theta^* = \varepsilon$, and $\alpha^* = \infty$, but keeping preferences and the preference shock $\phi^* = \phi$ unchanged. This new economy has the same asset pricing implications as the old one, read off marginal rates of substitution alone. In Figure 6 one can model the production side as the Northeast pointing box outlined by the dashed lines, keep the indifference curve, and asset prices are unchanged.

One can create an analogous production-based asset pricing model. Again, measure or model the consumption or productivity process $\varepsilon^* = \varepsilon$. Leave productivity shocks $\theta^* = \theta$ and keep the smooth production set $\alpha^* = \alpha$. Marry this production process to fixed-coefficient *preferences.* In place of the smooth utility function and preference shocks – both of which may be quite complex these days – let

$$u\left[c(h), c(l)\right] = \min\left[\frac{c(h)}{\varepsilon(h)}, \frac{c(l)}{\varepsilon(l)}\right]. \tag{76}$$

Then, measure contingent claim prices or the discount factor from the marginal rate of *transfor-*

mation alone. This new economy has the same asset prices and quantity implications as the full general equilibrium – but spares the researcher having to model and measure the entire consumption and intermediation side of the economy.

Fixed coefficient preferences (76) act like endowments. They generate a simplified general equilibrium economy with the same asset pricing and quantity implications as the full equilibrium – if one models the equilibrium consumption and productivity processes correctly. This approach may be useful for simulation economies, as the endowment economy formulation has been successful in consumption-based asset pricing.

9. A simple aggregation model

The main philosophy in this paper is to model the aggregated (smooth) production possibility set directly, rather than to derive the structures of such sets from primitive traditional specifications. The primitives are typically unobservable, and, again, there was no particular reason for specifying fixed patterns across states in the first place. However, it is useful as motivation, and to help think about what a smooth production set might look like, to sketch a model in which a smooth aggregated production set is derived from underlying traditional technologies.

Consider a two-state world in which the firm has *two* technologies. For example, a farmer can plant in two fields. One field does well in wet weather, the other in dry weather. The farmer can then shape the risk-exposure of his or total output to weather by varying the amount planted in each of the two fields. I've told the story before, let's write it in equations and make an accurate picture.

Let the technologies of field i be

$$y_i(s) = \varepsilon_i(s) k_i^{\eta};\ s = h \text{ or } l,\ i = 1 \text{ or } 2.$$

Total output is then

$$y(s) = y_1(s) + y_2(s);\ s = \{h, l\}$$

and total inputs are constrained by initial capital less initial sales,

$$k = k_1 + k_2.$$

We want to know what this structure implies for the aggregates k and $y(s)$. (Or, if we wish to characterize the production set by outputs alone, $y(0) = W - k$ and $y(s)$.) Figure 7 plots the answer. To produce the figure, I vary k_1 from 0 to $k = 1$, I let $k_2 = k - k_1$. Then, I calculate

$y(s) = \varepsilon_1(s)k_1^\eta + \varepsilon_2(s)k_2^\eta$ with $\varepsilon_1(h) = 2$, $\varepsilon_1(l) = 1$ and $\varepsilon_2(h) = 1$, $\varepsilon_2(l) = 2$. The far lower right point on the curve, for example, puts all initial capital into technology 1 that does well in the h state. The far upper left point puts all initial capital into technology 2 that does well in the l state. The aggregate production possibility set is smooth. Free disposal allows the aggregate production set to fill out the area indicated by dashed lines.

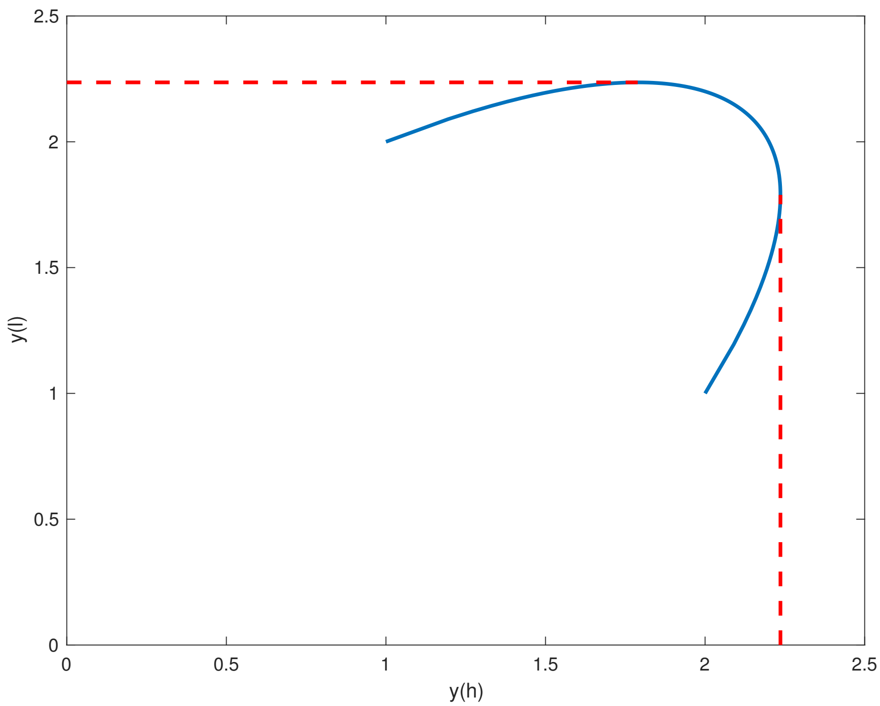

Figure 7: Aggregate production set $\{y(h), y(l)\}$ induced by two technologies, $y(s) = y_1(s) + y_2(s)$; $y_i(s) \leq \theta_i(s)k_i^{0.5}$; $i = 1, 2$; $s = h, l$; $k_1 + k_2 = 1$.

For this construction to work – for the marginal rate of transformation $\partial y(h)/\partial y(l)$ to exist, so we can equate it to contingent claims price ratios $\partial y(h)/\partial y(l) = p_h/p_l$ for any such ratio – we need a spanning or invertibility condition, in this case that the matrix

$$\begin{bmatrix} \varepsilon_1(h) & \varepsilon_2(h) \\ \varepsilon_1(l) & \varepsilon_2(l) \end{bmatrix}$$

is non-singular. If there are more than two technologies, we need the rank of a larger shock matrix to be at least two, the number of states. We also need sufficient concavity of the underlying production function $f(k)$. If not, the curve of Figure 7 is a straight line, and production ends up at one of the corners for all but one contingent-claim price.

For continuous-state economies, we subdivide technology into finer units of analysis.

Each square foot of land may have slightly different sensitivity to weather. Thus, consider technologies indexed by z, and states of nature indexed by ω. Aggregate output is

$$y(\omega) = \int dz \varepsilon(\omega, z) f\left[k(z)\right]$$

The corresponding invertibility or spanning condition is that we can invert this relation to

$$f\left[k(z)\right] = \int d\Pi(\omega)\lambda(\omega, z) y(z).$$

Loosely, this condition expresses the idea that the number of states equals the number of technologies, or that there exists a distribution of capital $k(z)$ that achieves any given state-contingent output $y(z)$. Here I integrate across states with a probability measure $d\Pi(\omega)$.

Alternatively, we can derive smooth production sets by allowing the firm to vary its investment in a few technologies continuously over time, extending the classic Black-Scholes option pricing approach to multiple risky and concave investment strategies.

I do not belabor necessary and sufficient conditions for spanning, as the major point of this paper is to write down smooth technologies directly, just as we write down aggregate technologies $y = f(k, l, .)$ that are smooth across inputs rather than derive them from deeper fundamentals. You can see from this discussion where such an aggregation theory would go.

We use an aggregation theory when we have good detailed evidence on foundations – individual preferences, machine-specific production functions – and we wish to use that knowledge to figure out representative-agent preferences or representative-firm production functions. That is not (yet) the case here, so the aggregation theory just makes the point that the aggregate production function may be smooth when machine-specific production functions are not.

With this basic idea, you can see many potential microfoundations for active tradeoffs across states. The firm could invest in capital or R&D to shift its output across states – buying solar cells, or multifuel engines, for example, change the distribution of profits across states indexed by energy price shocks. And thinking about such aggregation stories may be a useful way to improve on the description of shock choices in an intertemporal context, as outlined above.

The aggregation story emphasizes two points, however: First, technologies generated in this way will vary only across states of nature that are related somehow to the production process. The firm cannot transform output across states of nature that depend on a pure preference shock, or other exogenous random variable such as who wins the Super Bowl. Second, since probabilities do not enter the technology, probabilities do not enter the marginal rates of transformation. There is no counterpart to the risk neutral benchmark in which marginal rates of substitution are proportional to probabilities.

10. Literature

The idea of linking asset prices to quantities via producer first-order conditions, and thereby studying the production side of the economy without having to specify preferences, goes back a long way. My first effort was Cochrane (1988).

The word "production-based" has (in my view) become somewhat confused with "general equilibrium models that include production." A vast literature writes models with (often interesting and elaborate) preferences, along with detailed (interesting and elaborate) production technologies, and sometimes market frictions as well, calibrated to match asset pricing facts. Though some such papers use the word "production-based" to describe their efforts (Croce (2014), for example), I'll limit discussion to the effort to link asset prices to production data without, or with at most light, assumptions about consumers and market structure.

This effort swiftly ran in to the problem outlined in the introduction: standard production technologies do not give a marginal rate of transformation across states.

Standard technologies do, however, give rise to investment returns, and production-based asset pricing has to date largely linked macroeconomics to asset pricing via investment returns. One can characterize the prediction of an essentially Q-theory model, as outlined in Section 6, that production technology gives rise to a physical return R^I, measurable from investment, capital, output, and labor decisions (39). With adjustment costs, the investment return is dominated by investment, and thus is approximately proportional to investment growth. As a result, models based on investment returns are often called "investment-based asset pricing," and a cross-sectional extension (discussed below) an "investment CAPM." The investment return should be priced by any discount factor, $1 = E(mR^I)$. Any asset returns or prices that can be determined by arbitrage with the investment return should be so priced. When marginal Q equals average Q, the firm's stock (or stock and bond) return should equal the investment return, ex-post as well as ex-ante.

This effort was remarkably successful, at least compared to the widespread view that Q theory doesn't work at all. Cochrane (1991) shows that an investment return based on aggregate investment data is well correlated with stock returns at business-cycle frequencies, and that variation in expected stock returns as forecast by the dividend yield, term spread, investment to capital ratio and other variables matches well variation in expected investment returns. Lamont (2000) shows that measures of investment plans offer even better correlations. When stock prices rise it takes time to put investment into motion, but investment plans move quickly. One could also specify a time-to-build technology, but investment plans show the correlation quickly and transparently. Unlike many theories, the investment-return approach works better for big movements than small ones: the 1990s stock boom corresponded to an investment boom; the 2008 stock price plummet coincided with an investment collapse. (See Cochrane (2017) Figure 4.)

This branch of production-based asset pricing is exactly the same as a very simple version of Q theory. Yet it seems to work much better. This experience reflects an important lesson: how theories are implemented empirically matters a lot. Traditional Q theory focuses on detailed treatment of corporate taxes and measures of book values; it focuses on interest rates as the central driver of cost of capital; it relates the level of investment to the level of Q; it includes more complex production technologies (with marginal not equal to average Q, for example); it often uses cash-flow forecasts and other detailed measurements beyond investment and stock prices. It also focuses on failure: The theory predicts a 100% R^2 – investment should be proportional to Q, exactly, with no error. Any error is a formal rejection of the theory. That research focuses on the correlation of Q theory errors with cashflow. Much of the research has a goal of using Q theory only as a control to show what it can not explain, in order to advance a cashflow-constraint agenda.

By contrast, investment-return work focuses on equity premiums as the central driver of cost of capital, and we now know that equity premiums vary over time far more than riskfree rates, and in the opposite direction. Equity premiums are high in recessions with low stock prices, and low investment; interest rates are low in recessions. Investment-return work relates business-cycle frequency measures of investment growth to stock returns, ignoring the obvious high frequency failure (5 minute stock returns do not correlate with 5 minute investment growth) and ignoring low frequencies and the cross-section of levels where measurement issues allow prices to diverge persistently from book values. And, admitting that anything less than 100% R^2 is a formal rejection, it looks for the part of the glass that is half full. And finds it.

This lesson will be important in using the cross-sectional branch of production-based asset pricing described in this paper. There are hundreds of implementation decisions. Formal rejections of specific implementations will be easy. Figuring out where the theory is most useful will be harder.

Relating variation in the market return over time to investment growth is interesting, but the variation in average returns across assets and (especially) across portfolios sorted on various characteristics is the heart of the asset pricing empirical challenge. Extending production-based asset pricing to describe the *cross-section* of returns is the crucial next step.

The investment-return based literature took that step, constructing multiple investment returns to extend asset pricing predictions to a larger cross-section. We have a wealth of data on industry, portfolio, and firm-level production that can construct similarly detailed investment returns. Though we still can only price by arbitrage from this set of returns, the more cross-sectional information the better.

The literature that Zhang (2017) calls the "investment CAPM" made remarkable progress by this approach. (Again, the normally forbidden adjective is merited, I think, relative to expectations.) Each *firm's* investment return should equal that firm's asset return. Firms with higher investment growth have higher investment returns and higher stock returns, both actual and expected. The same prediction holds of portfolios of firms. Zhang (2017) shows that cross-sectional variation in expected investment returns line up well with many of the "anomalous" cross-sectional patterns in expected stock returns. (The iceberg of which this survey is a tip includes Lyandres, Sun, and Zhang (2008) Li, Livdan, and Zhang (2009), Liu, Whited, and Zhang (2009), Wu, Zhang, and Zhang (2010), Li and Zhang (2010), Liu and Zhang (2014) and Goncalves, Xue, and Zhang (2019). There are lots of anomalies and measurement issues to work out!)

Whether one can say this approach "explains" the anomalies and if so "rationally" is a contentious question. It documents that firms adjust decisions properly in response to expected returns, so investment decisions and expected returns are connected as economics says they should be. But both investment and expected returns are endogenous variables. Both could be driven by fads and irrationalities on the parts of consumers. Still, if expected returns line up with consumption or factor betas, one can make the same objection to the word "explain," as returns, consumption and its betas are also endogenous variables. So, one can say that the investment CAPM "explains" as well as a standard consumption CAPM would, if a consumption CAPM were successful.

I also think the word "investment CAPM" is a bit misleading. "CAPM" suggests that all expected returns line up with covariances of returns with investment growth or investment returns, and promises a theory that in principle can explain any asset return as the CAPM does. That is not the case. The "investment CAPM" theory remains arbitrage between each return and each investment return, by arbitrage, in isolation. But he or she who does the detailed work gets to baptize the results, so just understand how the fundamental structure of an "investment CAPM" remains different from that of a CAPM or consumption CAPM.

We still desire a general purpose model then, one that could in principle price a larger set of returns. Cochrane (1996) investigates one way to extend a cross-section of investment returns to price lots of assets. It uses a discount factor formed from two investment returns,

$$m = a + b_r R_{t+1}^{I,r} + b_{nr} R_{t+1}^{I,nr} \tag{77}$$

where r denotes residential investment and nr denotes nonresidential investment, in order to price a cross-section of stocks. (It is also where I first thought about conditional vs. unconditional factor models, scaling factors in GMM, and the somewhat dangerous plots of average returns vs. predicted average returns.) Obviously, one can extend this approach with a larger set of investment returns on the right hand side. Li, Vassalou, and Xing (2006) take an important step, considering investment by households, corporate, noncorporate and financial businesses, and they price the Fama French 25 size and book to market portfolios as well as the Fama and French factor models do.

Why are we allowed to extend observation of two returns to price other returns, which are not connected by pure arbitrage? Arbitrage pricing theory, a limit on Sharpe ratios of strategies that profit from the difference between asset returns and investment returns, leads to such an approximate discount factor of the form (77) for asset returns highly correlated with combinations of the two investment returns. (See Cochrane (2005), Chapter 9.4.) Or, the paper speculates, if the investment returns span the investment opportunity set then consumption and marginal utility must be driven by the two investment returns. On p. 577,

> Why should investment returns be factors for asset returns? Factor pricing models are derived by arbitrage assumptions or by preference assumptions. We can assume that the firms on the … NYSE are claims to different combinations of N production technologies, plus idiosyncratic components that have small prices. Alternatively, we can invoke preference assumptions under which the returns on the N active production precesses, which are the only nondiversifiable payoffs in the economy and add up to aggregate wealth, drive marginal utility growth and hence price assets…

Zhang (2005), and, citing Zhang, Jones and Tüzel (2013), İmrohoroğlu and Tüzel (2014), Belo and Lin (2012) and Belo and Yu (2013) follow a similar approach. They estimate or simulate “production-based” models with discount factors

$$\log m_{t+1}^{*} = \text{constant} - \gamma_t \varepsilon_{t+1} \tag{78}$$

where ε_{t+1} is the shock to aggregate productivity and γ_t is a coefficient. (Belo, Lin, and Bazdresch (2014) add a cost-shock second factor.) However, as presented, it is a bit of a stretch to call these models "production based," at least by the definition given here of pricing assets from producer first-order conditions, leaving out preferences. These models really follow in the mode of the second suggestion in Cochrane (1996), loosely suggesting that consumption should be a function of the aggregate productivity shock. They really uses consumption-based asset pricing to extend the discount factor from a single investment return to multiple returns. For example, Zhang (p. 71) writes

> "Suppose there is a fictitious consumer side of the economy featuring one representative agent with power utility and a relative risk averse coefficient, A. The log pricing kernel is then $\log M_{t+1} = \log \beta + A(c_t - c_{t+1})$, where c_t denotes log aggregate consumption. Since I do not solve the consumer's problem that would be necessary in a general equilibrium, I can link c_t to the aggregate state variable in a reduced-form way by letting $c_t = a + bx_t$ [ε_t in my notation] with $b > 0$."

This paper can give a truly production-based view of where the model (78) comes from. As we saw in Section 8, when underlying productivity shocks θ dominate, then actual productivity ε is chosen in a way that is positively correlated with underlying productivity θ. The discount factor $\log m =$ const. $+ \alpha \log \varepsilon - (1+\alpha) \log \theta$ then may load negatively on the observed ε in a reduced form model that does not include θ. (I.e. θ is estimated from ε.)

Likewise, the combination (43)-(44) gives a discount factor with output growth, labor share growth, and growth in the labor/output ratio as pricing factors. This result provides an alternative theoretical foundation for a wide variety of asset pricing models that include such variables as risk factors. Among many others, Campbell (1996) Jagannathan and Wang (1996) find that a labor income growth factor helps to price the cross section of returns. Lettau, Ludvigson, and Ma (2019) find that the change in capital share, which is one mins the labor share in the discount factor formula (44), prices a cross section of returns.

Jermann (2013) used the idea that with two investment returns, one can span two states of nature, by pure arbitrage with no reference to preferences. In essence, he implemented the model of Section 9. He created a two-state simulation model, which captures salient features of the term structure. The trouble is, this approach is limited to simulation economies as reality seems to have more states of nature than investment returns.

This paper explores a fundamentally different approach to understanding marginal rates of transformation across states of nature, than elaborating on investment returns and trying to extend pricing from investment returns to other payoffs by preferences, arbitrage, or approximate arbitrage. It is a revision of the first part of Cochrane (1993). That paper sat a long time, as I hoped to complete an empirical counterpart and cleanly solve the θ identification question. But bringing such a model to data, or constructing simulation models that may be compared to data, is an extensive project in its own right, with numerous measurement, specification and identification issues to face. So this will, for now, have to stand as it is.

Belo (2010) is the first paper to use this production technology with productivity choice empirically. Belo proposed a clever approach to the identification problem, which could (and should) be generalized to much larger groups of investment returns, of the sort used by Zhang and coauthors. Discount factor formulas such as $m^* = \lambda\varepsilon^{\alpha}/\theta^{(1+\alpha)}$ hold separately for each technology, just as $m_{t+1} = \beta u'(c_{i,t+1})/u'(c_{i,t})$ holds separately for each individual i. Taking logs of the discount factor (13),

$$\log\left(m^*_{t+1}\right) = \alpha\log(\varepsilon_{i,t+1}) - (1+\alpha)\log(\theta_{i,t+1})$$

separately for each technology i. (Belo multiplies and divides by $\theta_t^{1+\alpha}$ and ε_t^{α} to express the model in growth rates. I simplify here to make the point clearer. Belo also uses α where I use $1+\alpha$.) Belo then assumes that multiple technologies have a factor structure,

$$(1+\alpha)\log(\theta_{i,t}) = \sum_{j=1}^{J}\lambda_{ij}F_{j,t}.$$

With a single factor F, and two technologies 1 and 2, then,

$$\log\left(m^*_{t+1}\right) = \alpha\log(\varepsilon_{1,t+1}) - (1+\alpha)\,\lambda_1 F_{t+1} \tag{79}$$

$$\log\left(m^*_{t+1}\right) = \alpha\log(\varepsilon_{2,t+1}) - (1+\alpha)\,\lambda_2 F_{t+1} \tag{80}$$

Now, we can eliminate the latent factor F, to express the discount factor.

$$\log\left(m^*_{t+1}\right) = \frac{\alpha}{\lambda_1-\lambda_2}\left[\lambda_1\log(\varepsilon_{2,t+1}) - \lambda_2\log(\varepsilon_{1,t+1})\right]. \tag{81}$$

We observe $\log(\varepsilon_i) = \log(y_i) - \log f(k_i)$. Anything time t is soaked up into the constant and identified by pricing the risk free rate or the investment returns. Normalizing $\lambda_1 = 1$, we can estimate λ_2.

The model is identified, though we do not directly observe the natural productivity shock θ. Intuitively, since the firms have different loadings on a common θ, they will choose productivity shocks ε that are perfectly correlated, but one moves more than the other. Then the difference between the shocks reveals the discount factor. Or, solving (79) and (80) for the shocks ε, the shocks move by the same amount in response to m, but one moves more than the other in response to F. Thus, watching the differences between the shocks, we can disentangle the two sources of ε movement, m and F.

The assumption may appear more plausible with more technologies. Across J technologies, there are J sources of unobserved movement θ_j and one additional source of movement m. Reducing the dimensionality of the θ by only one via a factor structure assumption, we can identify m. To generalize, we need a $J-1$ factor structure of technology shocks, not a single-factor structure. (Belo's online Appendix C pursues a $J=3$ factor model.)

Since Belo assumes $y_t = \varepsilon_t f(k_t)$ with k_t predetermined, he uses y_t in place of ε_t in (81). The bottom line is a two-factor macro-pricing model, using output growth,

$$\log(m_t^*) = a - b_1 \Delta y_t^1 - b_2 \Delta y_t^2.$$

This bottom-line result is the same form as the Cochrane (1996) investment-based model, with output growth in the place of investment growth. But Belo derives that otherwise ad-hoc model from the pure production-based pricing idea with the clever factor structure assumption to identify natural productivity shocks. He also adds a relative price of output and investment goods, which adds a second set of factors, and prices a more up to date set of asset returns.

11. Concluding comments and speculation

This paper is clearly an exploratory step. There is lots to do to create production-based asset pricing models that can unite asset pricing and macroeconomic facts.

I explored one particular functional form. Other functional forms, and a more general theoretical treatment, beckon. We have already seen that once labor is included, the discount factor includes either labor or wages, and not just productivity and its underlying shock. More detailed production functions may well change that form.

One needs curvature across states, but one can put in that curvature in many ways. Generalizing the one-period model to many periods, I assumed that the productivity choice sets hold independently at each point in time. I assumed $\{E_\tau\left[(\varepsilon_{t+1}/\theta_{t+1})^\alpha\right]\} \le 1$ and $\{E_\tau\left[(\varepsilon_{t+2}/\theta_{t+2})^\alpha\right]\} \le 1$ separately. One could envision production choice sets that allow the firm to transform output across states and time simultaneously. Here, such margins exist, but go entirely first across states, then over time via capital accumulation, then back across states again. I offered a joint CES form that treats state and time symmetrically, but that seemed to abandon too much the capital accumulation story.

I write a concave set of random variables $E\left[(\varepsilon/\theta)^{1+\alpha}\right] \le 1$ by summing across states with artificial probabilities. One might want to consider formulations that are not separable across states, as recursive utility does for preferences.

Bringing this production-based approach to data requires many choices, and most of all addressing the question how to identify the underlying productivity shocks θ, or to find a specification that does not need them. Initially, this task looks daunting. If θ is completely unobserved, and likely to be correlated with ε, then how can we implement $m = \lambda\varepsilon^{\alpha}/\theta^{1+\alpha}$? One can find a θ at any date to generate any discount factor one wishes. The example from Section 8 in which θ and ε are perfectly correlated, so the reduced form is $m = \lambda\varepsilon^{-\gamma}$ is not encouraging.

On second glance, this identification problem is no different or worse than the similar identification issues that haunt all of macroeconomics and finance. That example is exactly the same, with only a change in Greek letters, as the example in Cochrane (2011) in which the interest-rate rule of New-Keynesian models has a right hand variable (inflation) perfectly correlated with its (monetary policy) shock, so yields exactly the wrong coefficient. VARs are plagued by the question whether interest rates cause inflation or expected inflation causes interest rates. Yet new-Keynesian models and VARs are a thriving industry, solved and compared to data. How? By thinking hard and making identification assumptions, finding something orthogonal or exogenous somewhere. All economic models include shocks somewhere, and usually must do so if they want to avoid 100% R^2 predictions. Yet a shock in any equation usually means that equation cannot be directly estimated – we need a shock somewhere else to do that, and an exclusion restriction. Yet shocks have to be somewhere, and, if we are honest, most likely everywhere. Medium scale empirical macro models contain shocks in every equation. The increasing popularity of preference shocks (risk aversion, discount factor, financial frictions) or their observational equivalents (taste, sentiment, probability) raises exactly the same identification problem for conventional asset pricing.

The other approach to identification is to construct simulation economies. One may not be able to measure natural productivity θ, but one can specify a θ process, simulate data, and see what it takes for the simulated moments to match actual moments. That process includes lots of unstated identification assumptions, or in fact isn't identified at all – there may be other assumptions that produce the same moments. But it is how we construct models. Getting a model that can match the data is hard enough, and valuable, even if one cannot prove that some other model or parameterization might fit the data as well.

We have really just begun to properly explore the cross-sectional richness of production data. Zhang (2017) makes the most progress, computing the investment returns of sorted portfolios by computing the investment returns of their component firms, and comparing the cross-section of investment returns to the cross-section of asset returns. (Belo (2010) online Appendix C also encapsulates a wide cross section of sector and industry output data.)

In the project of extending asset pricing from the investment returns we can observe to asset returns, surely we want to use as many investment returns as possible. In applying APT logic – price a given asset return by finding close by, or near-replicating investment returns – there is no reason to apply the APT philosophy, that hoped for a small number of factors. Conversely, in Zhang (2017), each firm's investment return is a primitive. But surely there is factor structure or other common movement in the investment returns, that may restore something like the APT philosophy, and not require 3000 separate fundamental quantities underlying asset pricing.

The productivity choice approach here is fundamentally different from investment returns in this respect. Each firms' investment return $R^I_{i,t+1}$ is a separate object, giving us a separate measurement and prediction for one part of the payoff space. A discount factor using investment returns loads on all of them, $m = a + b_1 R^I_{1,t+1} + b_2 R^I_{2,t+1} + ... + b_i R^I_{i,t+1}$. However, each firms' productivity choice $m = \lambda_i \varepsilon_i^{\alpha_i} / \theta_i^{1+\alpha_i} = \lambda_j \varepsilon_j^{\alpha_j} / \theta_j^{1+\alpha_j}$ should equal the common m. This proposition mirrors the proposition that each individual consumer should set marginal utility growth to equal the common discount factor, $m = \lambda c_i^{-\gamma} / \phi_i^{1-\gamma}$. Thus, while APT logic and investment returns lead us to a discount factor m loading on many objects – essentially each firm's investment return in the Zhang (2017) approach – productivity-choice logic leads us to many measurements of a single discount factor. Disaggregated data should be useful for constructing that discount factor. Individual firm data may have measurement error, of course, and as Belo (2010) shows us, disaggregated data can help us to surmount the shock identification issue. Moreover, as Constantinides and Duffie (1996) show us for consumers, the common discount factor can look very different from aggregate productivity raised to a power. As in that case, cross-sectional dispersion in productivity can show up in the aggregate. Moreover, one should ideally integrate the investment-return and productivity-choice approaches, using both the cross-sectional information of many investment returns, and the many sources of cross-sectional information on the common discount factor. The aggregation model of Section 9 already points to interesting productivity choice in the aggregate production function that may not exist in firm-level production. Now, extend that to multiple technologies that also have productivity choices.

Clearly, the investigation has just begun.

References

Albuquerque, Rui, Martin Eichenbaum, Victor Xi Luo, and Sergio Rebelo. 2016. "Valuation Risk and Asset Pricing." *The Journal of Finance* 71 (6):2861–2904. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12437.

Belo, Frederico. 2006. "A Pure Production-Based Asset Pricing Model." Ph.D. Dissertation, University of Chicago Booth School of Business.

———. 2010. "Production-based measures of risk for asset pricing." *Journal of Monetary Economics* 57 (2):146 – 163. URL http://www.sciencedirect.com/science/article/pii/S0304393209001822.

Belo, Frederico and Xiaoji Lin. 2012. "The Inventory Growth Spread." *The Review of Financial Studies* 25 (1):278–313. URL https://doi.org/10.1093/rfs/hhr069.

Belo, Frederico, Xiaoji Lin, and Santiago Bazdresch. 2014. "Labor Hiring, Investment, and Stock Return Predictability in the Cross Section." *Journal of Political Economy* 122 (1):129–177. URL http://www.jstor.org/stable/10.1086/674549.

Belo, Frederico and Jianfeng Yu. 2013. "Government investment and the stock market." *Journal of Monetary Economics* 60 (3):325 – 339. URL http://www.sciencedirect.com/science/article/pii/S0304393213000056.

Campbell, John Y. 1996. "Understanding risk and return." *Journal of Political Economy* 104:298–34.

Cochrane, John H. 1988. "Production-Based Asset Pricing." Working Paper 2776, NBER. URL https://www.nber.org/papers/w2776.pdf.

———. 1991. "Production-Based Asset Pricing and the Link Between Stock Returns and Economic Fluctuations." *The Journal of Finance* 46 (1):209–237. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1991.tb03750.x.

———. 1993. "Rethinking production under uncertainty." URL https://faculty.chicagobooth.edu/john.cochrane/research/papers/Cochrane%20Rethinking%20Production%20Under%20Uncertainty.pdf. Manuscript.

———. 1996. "A Cross-Sectional Test of an Investment-Based Asset Pricing Model." *Journal of Political Economy* 104:572–621.

———. 2005. *Asset Pricing, Revised Edition.* Princeton, N.J.: Princeton University Press.

———. 2011. "Determinacy and Identification with Taylor Rules." *Journal of Political Economy* 119:565–615.

———. 2017. "Macro-Finance." *Review of Finance* 21:945–985. URL https://doi.org/10.1093/rof/rfx010.

Cochrane, John H. and Lars Peter Hansen. 1992. "Asset Pricing Explorations for Macroeconomics." *NBER Macroeconomics Annual* 7:115–165. URL http://www.jstor.org/stable/3585001.

Constantinides, George M. and Darrell Duffie. 1996. "Asset Pricing with Heterogeneous Consumers." *Journal of Political Economy* 104 (2):219–240. URL http://www.jstor.org/stable/2138925.

Croce, Mariano Massimiliano. 2014. "Long-run productivity risk: A new hope for production-based asset pricing?" *Journal of Monetary Economics* 66:13 – 31. URL http://www.sciencedirect.com/science/article/pii/S0304393214000506.

Debreu, Gérard. 1959. *The theory of value: an axiomatic analysis of economic equilibrium*. New York, N.Y.: WIley.

Goncalves, Andrei S., Chen Xue, and Lu Zhang. 2019. "Aggregation, capital heterogeneity, and the investment CAPM." *Review of Financial Studies* Forthcoming. URL http://theinvestmentcapm.com/uploads/1/2/2/6/122679606/vitae_luzhang2019december.pdf.

Hansen, Lars Peter and Ravi Jagannathan. 1991. "Implications of Security Market Data for Models of Dynamic Economies." *Journal of Political Economy* 99 (2):225–262. URL http://www.jstor.org/stable/2937680.

Hindy, Ayman and Chi-fu Huang. 1992. "Intertemporal Preferences for Uncertain Consumption: A Continuous Time Approach." *Econometrica* 60 (4):781–801. URL http://www.jstor.org/stable/2951566.

Houthakker, H. S. 1955. " The Pareto Distribution and the Cobb-Douglas Production Function in Activity Analysis 1." *The Review of Economic Studies* 23 (1):27–31. URL https://doi.org/10.2307/2296148.

İmrohoroğlu, Ayşe and Şelale Tüzel. 2014. "firm-level productivity, risk, and return." *management science* :2073–2090.

Jagannathan, Ravi and Zhenyu Wang. 1996. "The Conditional CAPM and the Cross-Section of Expected Returns." *The Journal of Finance* 51 (1):3–53. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb05201.x.

Jermann, Urban J. 2013. "A production-based model for the term structure." *Journal of Financial Economics* 109 (2):293 – 306. URL http://www.sciencedirect.com/science/article/pii/S0304405X13000767.

Jones, Christopher S. and Şelale Tüzel. 2013. "Inventory investment and the cost of capital." *journal of financial economics* 107:557–579.

King, Robert G., Charles I. Plosser, and Sergio T. Rebelo. 1988. "Production, growth and business cycles: I. The basic neoclassical model." *Journal of Monetary Economics* 21 (2):195 – 232. URL http://www.sciencedirect.com/science/article/pii/030439328890030X.

Kydland, Finn E. and Edward C. Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50 (6):1345–1370. URL http://www.jstor.org/stable/1913386.

Lamont, Owen A. 2000. "Investment Plans and Stock Returns." *The Journal of Finance* 55 (6):2719–2745. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00304.

Lettau, Martin, Sydney Ludvigson, and Sai Ma. 2019. "Capital Share Risk in U.S. Asset Pricing." *The Journal of Finance* 74 (4):1753–1792. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12772.

Li, Dongmei and Lu Zhang. 2010. "Does q-theory with investment frictions explain anomalies in the cross-section of returns?" *Journal of Financial Economics* 98:297–314.

Li, Erica X. N., Dmitry Livdan, and Lu Zhang. 2009. "Anomalies." *Review of Financial Studies* 22:4301–4334.

Li, Qing, Maria Vassalou, and Yuhang Xing. 2006. "Sector Investment Growth Rates and the Cross Section of Equity Returns." *The Journal of Business* 79 (3):1637–1665. URL http://www.jstor.org/stable/10.1086/500687.

Liu, Laura Xiaolei, Toni M. Whited, and Lu Zhang. 2009. "Investment-based expected stock returns." *Journal of Political Economy* 117:1105–1139.

Liu, Laura Xiaolei and Lu Zhang. 2014. "A neoclassical interpretation of momentum." *Journal of Monetary Economics* 67:109–128.

Lyandres, Evgeny, Le Sun, and Lu Zhang. 2008. "The new issues puzzle: Testing the investment-based explanation." *Review of Financial Studies* 21:2825–2855.

Wu, Jin (Ginger), Lu Zhang, and X. Frank Zhang. 2010. "The q-theory approach to understanding the accrual anomaly." *Journal of Accounting Research* 48:177–223.

Zhang, Lu. 2005. "The value premium." *Journal of Finance* 60:67–103.

———. 2017. "The investment CAPM." *European Financial Management* 23:545– 603.