Rethinking Production under Uncertainty

John H. Cochrane
Hoover Institution, Stanford University, and NBER

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 that is linked to productivity and that can be used to price a wide variety of assets, without regard to preferences. (*JEL* G12)

Received XXXX XX, XXXX; editorial decision XXXX XX, XXXX by Editor XXXXXXXXXXXX.

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, that is, productivity $\varepsilon(s)$, giving random output $y(s)$.

Figure 1 illustrates the production set implied by this technology for a two-period two-state world. A farmer has seeds W at time 0. The farmer may plant them as k, or sell 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 in spring to a bundle of contingent wheat in fall, but it has a kink across the states of nature. No matter how high the contingent claim price of wheat is in the low state relative to the high state, the farmer can do nothing to produce more in the low state at the expense of production in the high state.

I thank the editors, John Campbell, Wayne Ferson, and, especially, Frederico Belo for helpful and detailed comments. Send correspondence to John H. Cochrane, john.cochrane@stanford.edu.

doi:10.1093/rapstu/Sample Advance Access publication September 21, 2014

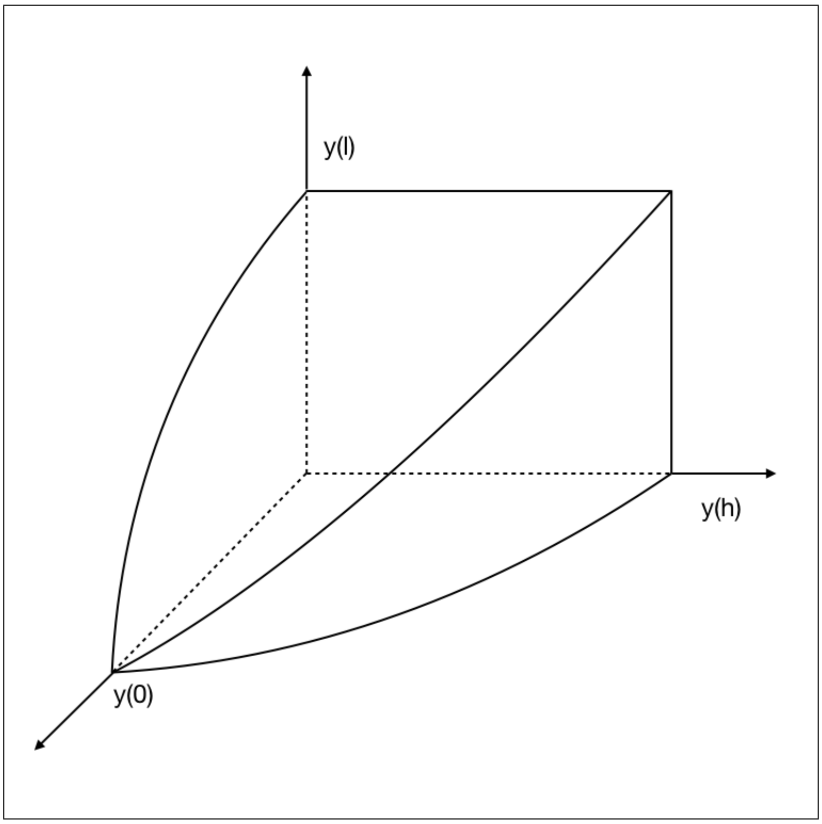

Figure 1
Standard production possibility set in a two-state two-date world
The technology is $y(s)=\varepsilon(s)f(k)$ for $s=h,l$, and $y(0)=W-k$.

This paper explores a 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 that 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, leaving sales at time 0, $y(0)=W-k$, 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

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

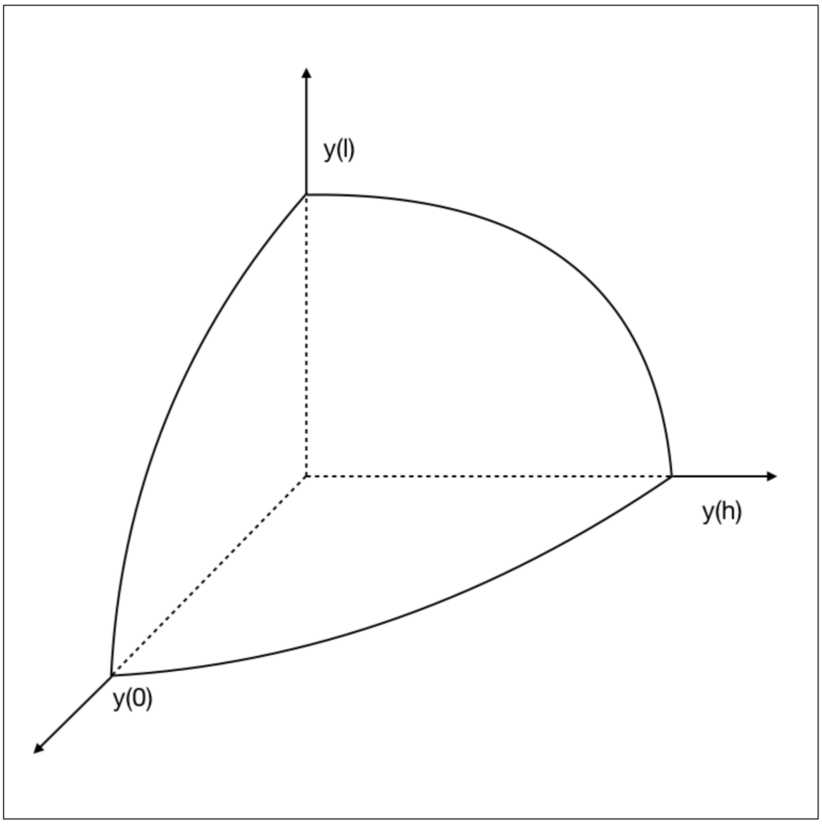

Figure 2
Smooth production possibility set
The firm can change the distribution of output across states of nature, that is, the pattern of $y(h)$ versus $y(l)$.

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 obtain higher productivity than natural in some states, by accepting lower than natural 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 p scaled by probabilities π, $m(s) = p(s)/\pi(s)$. If a firm maximizes contingent claim value

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

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 (2). In a dynamic extension of the model, we link the stochastic discount factor to productivity growth

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

where λ_t is a similar constant known at time t. The firm chooses to produce more in states of nature with high contingent claim prices or stochastic discount factors—higher marginal utility in general equilibrium—and in states in which natural productivity θ is larger.

Why is this representation of technology useful or interesting? My direct interest is the construction of production-based asset pricing models. These are models that link asset prices and returns to real economic variables through producer first-order conditions. Rather than focus on understanding firm behavior, the determinants of random productivity ε, given asset prices as summarized by a discount factor m, I am interested in turning the first-order conditions around to measure the stochastic discount factor from observed quantity choices. For example, once we infer the stochastic discount factor from productivity data via (4), we can relate risk premiums $E(R^e)$ to the covariance of excess returns R^e with the discount factor, $E(R^e) = -cov(R^e, m)/E(m)$, and we can understand the prices p of payoffs x from $p = E(mx)$. When we generalize production technologies, variables beyond productivity including output, investment, hours, and disaggregated production data enter the discount factor and contribute to this understanding.

This production-based approach ties the discount factor to marginal rates of transformation, ignoring and thus holding for any set of preferences. While a full understanding of the economy requires general equilibrium—understanding preferences, investors' probability assessments, and the consumer-facing market structure—one can at least tie asset prices to the production side of the economy, and study production technology and behavior in isolation. One can determine whether the cyclical relations between asset prices or returns and firm data make economic sense, while others work on preferences and market structure.

As the name implies, this approach to production-based asset pricing is deliberately parallel to the standard consumption-based asset pricing approach. The consumer first-order conditions are

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

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

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

In consumption-based asset pricing, we infer the stochastic discount factor from consumption data, or its proxies, via (5). We then understand risk premiums and asset prices by the covariance of payoffs with this discount factor.

This consumption-based approach infers the stochastic discount factor from marginal rates of substitution, ignoring and thus holding for any technology. While a full understanding of the economy requires general equilibrium—understanding production technology and its shocks, where cash flows come from—one can at least tie asset prices to the consumer-investor side of the economy, and study preferences, expectations, and consumer-facing market structure in isolation. One can determine whether the cyclical relations between asset prices or returns and consumer data make economic sense, while others work on production technology, and assembling production and consumption together in general equilibrium models.

While the approaches are parallel, production-based asset pricing is additionally attractive because business cycles are essentially a phenomenon of production—declines in investment, durable goods output, and employment—and much less visible in consumption. Indeed, formula (4) and its generalizations below have the form of many *ad hoc* macro-asset pricing models that tie asset returns to a discount factor created from productivity growth, investment growth, output, hours, and other production data, surveyed below. Thus, this production-based theory can provide foundations for many existing models in this class and the empirical success they already document.

Figure 1 illuminates why this direct approach to production-based asset pricing model is not possible using standard representations of technology extending (1). A kink in the production set across states of nature means that many different contingent claims prices are consistent with any production point the firm might choose. There is no marginal rate of transformation.

Much production-based asset pricing nonetheless uses standard technologies of the form (1). Firm first-order conditions in this case still contain useful information for asset pricing. The firm invests optimally, producing a fixed-coefficients bundle of contingent claims. The first-order condition for that investment says that the physical return on investment should be correctly priced by the stochastic discount factor m, $1 = E\left[mR^I(k)\right]$, with $R^I(k) = \varepsilon f_k(k)$ in a simple two-period example. We can therefore price asset payoffs that are perfectly spanned by investment returns, and we can check for arbitrage between asset and investment returns. This literature, surveyed below, has had considerable empirical success. But in this framework, we cannot infer anything about other returns, and we cannot back out a general discount factor, without making additional preference assumptions.

Since I add the choice of productivity ε to these representations of technology, all their predictions remain intact. This paper is a generalization of investment-return models, not an alternative to them.

The approach in this paper and its relatives, surveyed below, is thus distinctive in that by allowing and modeling a marginal rate of transformation across states, we can read the stochastic discount factor that prices a wide class of returns from production data directly, without preference assumptions, in exact analogy to the standard consumption-based model.

The word "production-based" is also sometimes used to describe any model that links its discount factor or pricing factors to production data, though the economic logic may involve consumer optimization or general equilibrium. I use it here to describe models that use of marginal rates of transformation alone.

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.

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

The presence of random natural productivity θ raises some practical Difficulties, 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.

Basic correlations in the data argue that we need underlying technology shocks θ as well, perhaps, as preference shocks. 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, low stock prices, and recessions. But output is low in recessions, not the other way around. Thus, the fact that stock prices, output, and consumption all comove positively suggests that the bulk of such fluctuations must come from underlying technology shocks, not preference or equivalent irrational probability shocks.

This conclusion is not ironclad. Productivity may rise in recessions, when output and stock markets fall, or when other risk factors fall. Not all macroeconomic variables and asset returns move in lockstep: we live in a multifactor world. But the logic is strong enough that we should keep natural productivity shocks in the model for now and think quantitatively about their need and how to identify them. Natural productivity shocks also act as a change-of-measure variable allowing us to treat probabilities flexibly.

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, that is, the distribution of the productivity shocks they face. A farmer may plant wheat in fields that do 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. Firms generically face questions of efficiency versus resilience. Choose one cheapest supplier or spread orders around multiple suppliers in different countries. Keep extra inventories around or order them just in time. "Real options" in management studies exactly this sort of production decision. Given that bankruptcy, adjustment costs, and reorganization costs are real, financial decisions, such as hedging input prices and equity versus debt financing, affect state-contingent outputs. In Spring 2020, decisions not to keep an inventory of face masks and ventilators around, and decisions to take on a lot of debt rather than equity finance are leading to much regretted state-dependence in output. The ability to produce during the pandemic state of nature is suddenly receiving great attention in industry and government, and hopefully better choices of ε will emerge before the next one hits.

This ability to transform output across states of nature 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, 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 to the classic result that an aggregate of production functions that demand fixed combinations of inputs can be smooth (Houthakker 1955). I apply the same logic to multiple outputs across states of nature.

Likewise, as one can span a full set of contingent claims by varying investment over time in two securities, as in Black-Scholes option pricing, so one could span contingent claims by time-varying physical investment in multiple fixed-coefficient technologies.

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. But aggregation theory is useful when we have detailed micro data that we wish to aggregate. The philosophy in this paper is to specify aggregated, and therefore smooth, firm, industry, sector, or economy production functions directly, corresponding to our data sources. This philosophy mirrors the specification of representative consumer preferences without spending a lot of time on aggregation in consumption-based asset pricing.

Third, one may simply view the lack of kinks as being the most natural production set and question the logic and evidence for such kinks. That is how we approach the choice of inputs and the study of nonstochastic multiple-output production functions. If we wish to model a farmer's choice to produce wheat versus corn, or to produce more today and less tomorrow, we start with a smooth production set. So if we wish to study wheat in rainy weather versus wheat in sunny weather, why would we start by assuming their proportions are immutably 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, outputs, 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 considers the possibility of a smooth production set across states of nature. That choice is entirely understandable. A smooth production set introduces complications. And these authors didn't need to generalize. Tacking productivity shocks onto standard intertemporal technologies was good enough for their uses. But that historical accident does not carve the decision in stone or argue for kinks and against smoothness. Here too, adding productivity choice generalizes rather than contradicts these models. One can always pick the underlying shock process θ so that the firm's equilibrium choice ε is the same as that specified in these models.

1. 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 constant elasticity of substitution (CES) functional form for this choice: output y is given by a standard production function combining capital k and labor n,

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

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

where ε satisfies

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

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

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 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 $y \leq \theta f(k,n)$ to a CES aggregate of output across states on the left-hand side,

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

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.

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 $\alpha = 1$, one sees 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, 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 toward 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 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 arbitrarily from true or empirical probabilities used in analysis. Thus, the weights $\theta^{1+\alpha}$ 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 versus unconditional, true versus

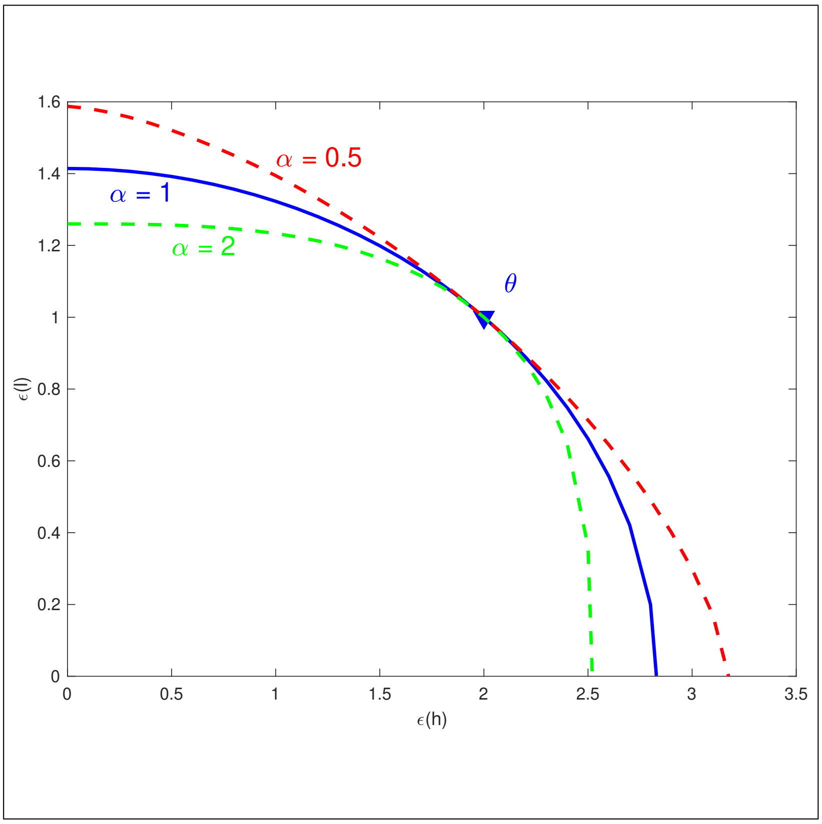

Figure 3
Shock choice sets
Each line represents the set of $\{\varepsilon(h),\varepsilon(l)\}$ that the firm can choose from, satisfying $E[(\varepsilon/\theta)^{1+\alpha}]\leq 1$. The base case is $\alpha=1$, $\theta=(2,1)$, and $\pi(h)=0.5$. The dashed lines vary α to $\alpha=0.5$ and $\alpha=2$.

sample, real versus risk-neutral probabilities, agents who see more than we do, and so forth.

To solidify these observations, 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 expression describes a concave production set of outputs. Divide by $f(k)$, and we have a constraint on productivity $\varepsilon=y/f(k)$. Given any convenient set of set of probabilities π, define

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

and we recover the original specification (2).

2. The Simplest Model

Now let us place this constraint in a model of the firm. Fix $f(k,n)=1$ to focus on the random variable choice, and consider a firm that maximizes the value of output, now just ε. The firm's problem is

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

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

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

The variable m is the stochastic discount factor, or contingent claim price divided by probability, $m(s)=p(s)/\pi(s)$, so the objective is the same 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

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

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

In standard theory of the firm, we solve for choices given prices, for ε given m. We do that by imposing the constraint in (9) to eliminate the Lagrange multiplier λ, which yields[1]

$$\frac{\varepsilon^{\alpha}}{\theta^{\alpha}}=\frac{m\theta}{\left\{E\left[(m\theta)^{\frac{1+\alpha}{\alpha}}\right]\right\}^{\frac{\alpha}{1+\alpha}}}. \tag{11}$$

[1] From (10),

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

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

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

Substitute out $\lambda(1+\alpha)$ in the top equation and rearrange to get (11).

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. (The point of this equation is to determine ε given the other variables, so one might reexpress it with just ε on the left-hand side to emphasize that point. I find the given expression prettier, as it describes how ε is distorted away from 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. Equation (10) already gives us a discount factor that can price all zero-cost portfolios or excess returns. For that goal, 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 $0=E[(2m^*)R^e]$. Thus, the discount factor

$$m^*=\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}} \tag{12}$$

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. (For symmetry, I include a preference shock ϕ here, discussed below.)

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

$$m^*=\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}\Big/\left[R^f E\left(\frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}\right)\right]. \tag{13}$$

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 (11) gives the same choice ε for a discount factor $2m$ as it does for a discount factor m, so we cannot invert (11) to learn the level of the discount factor from ε. Next, we will add time.

Expressions (10)-(13) relate random variables. They hold ex post state by state. The (s) notation in (10) emphasizes this fact. It is often convenient to give a name and number to states of nature. For example, s could denote inches of rainfall. Then (10) relates *functions.* The expression $m(s)$ is the function relating inches of rainfall to the discount factor, and (10) describes how that function comprises the functions $\varepsilon(s)$ and $\theta(s)$, or it describes the firm's optimal choice function $\varepsilon(s)$ in terms of the functions $m(s)$ and $\theta(s)$. Thinking this way is particularly convenient when one wants to construct a model, not just infer a discount factor from data. One typically specifies that s is a vector that follows a stationary Markov process, fully capturing all information.

3. A Two-Period Model

Next, 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 (12) and (13) continue to describe risk premiums. The intertemporal problem adds a single investment return which establishes the level of the discount factor m and the level of returns. Add capital and the possibility to invest at time 0. The firm maximizes contingent claim value,

$$\max_{\{k,\varepsilon\}} E[m\ \varepsilon f(k)] - k \text{ s.t. } E\left[\left(\frac{\varepsilon}{\theta}\right)^{1+\alpha}\right] \leq 1. \tag{14}$$

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

Again, introducing a Lagrange multiplier λ on the productivity-choice constraint of (14), the first-order conditions are

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

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

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 (15) $1 = E(mR^I)$ 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 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 (12), $m^* = \varepsilon^{\alpha}/\theta^{1+\alpha}$, and its scaled version (13) 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:

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

Equation (16) then becomes

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

The three forms on the right-hand side are equivalent. The reader may find that one or the other is more elegant.

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

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

The discount factor (17) is not the inverse of the investment return, $m \neq 1/R^I = 1/[\varepsilon f_k(k)]$. The discount factor (17) adjusts that investment return as the firm has chosen to distort its productivity ε from the underlying shock θ. The investment return $R^I = \varepsilon f_k(k)$ is not risk-free. The model determines the risk-free rate indirectly, through the investment return together with the productivity ε first-order condition that determines risk premiums. From (17), the risk-free 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 $\varepsilon f(k)$, 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 (12) for characterizing risk premiums remains essentially unchanged, while the investment returns and therefore the characterization of the overall level of returns becomes more complex.

3.1 Production theory versus asset pricing

In the theory of the firm, we solve such first-order conditions to give the producer's choices $\{k, \varepsilon\}$ in terms of prices, that is, the discount factor m. To this end, we solve the pair of first-order conditions to give one equation describing k and another describing ε, each in terms of m and θ. The resultant expression for optimal capital k is[2]

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

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

while the optimal productivity ε is given by

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

Equation (19) expresses the same intuition as the first-order condition (16), produce more in high discount factor and high ε states, in purer form. It is the same expression as in the one-period model, (11).

Looking at (19), the choice $\varepsilon = \theta$ emerges if $m \propto 1/\theta$. In this case, we do have that $m = 1/[\varepsilon f_k(k)] = 1/R^I$, that is, 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, $m = \beta =$ constant.

Though my motivating application is production-based asset pricing, a theory of the firm with choice of productivity shocks would be interesting as well. However, the genius of consumption-based asset pricing is that we can infer discount factors from consumer first-order conditions without even solving the full consumer partial-equilibrium problem—without writing the budget constraint, income stream, and finding consumption in terms of prices and incomes, as, for example, rational expectations permanent income models do—and certainly without solving the whole general equilibrium. Here, production-based asset pricing follows the same path. We can infer the discount factor, or at least a discount factor for zero-cost portfolios, directly from firm first-order conditions without solving for the constraint as in (19), and without solving the full partial-equilibrium output, labor, and capital plan as in (18), let alone general equilibrium.

4. Labor

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

First, a disappointment: one might think that a firm that can adjust inputs after observing a shock can produce more or less output in response to that shock and thus achieve a marginal rate of transformation. That intuition is false. Producing more in one state does not make it more difficult to produce in another. The ability to produce more or less after a shock is observed does not allow the firm to *transform* output across states of nature. (Belo 2010, footnote 4 makes this point.)

To see this point, write the production function as

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

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

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

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

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

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

Condition (20) does not help us to identify the discount factor $m(s)$, as $m(s)$ cancels from that equation. The firm sets $\varepsilon(s)f_n[k,n(s)]=w(s)$ 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. Written in terms of contingent claims prices $p(s)=m(s)/\pi(s)$, the first-order condition is not $p(s)f_n[k,n(s)]= w(s)$; 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 2. Now let the firm maximize

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

The labor decision and the wage are both stochastic; that is, $w(s)$ 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

$$\varepsilon f_n(n)=w$$

$$mf(n)=\lambda(1+\alpha)\varepsilon^{\alpha}/\theta^{1+\alpha}.$$

With a standard power functional form of $f(n)=n^{\sigma}$, the first-order conditions become

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

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

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 (12), 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

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

Expression (28) using wages is a little more elegant than (25) using labor input, as now the discount factor is expressed as a function of the single choice variable ε and external circumstances w and θ. High ε will induce the firm to hire more labor, so ε and n are really not two separate influences in (25). In (28) labor changes the effective coefficient on productivity ε. A measurement of the coefficient on ε with constant wages is not the pure coefficient of transformation across states. However, the labor end of macroeconomics discourages the use of measured spot wages as equal to marginal products of labor, so the formulation using actual labor inputs may be more successful empirically.

The discount factors (25) and (28) have important lessons going forward. The production-based discount factor is not necessarily just productivity raised to a power. Here, wages or labor inputs appear as additional pricing factors in a discount factor formula. Additional material inputs or adjustment costs can appear similarly.

Solving (27) for ε, and using the constraint to find the Lagrange multiplier λ, we can express the productivity choice as[3]

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

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.

5. Intertemporal Production

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

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

where $\rho^{t-1}\Lambda_t$ is the stochastic discount factor, with $m_{t+1}=\rho\Lambda_{t+1}/\Lambda_t$, $\Lambda_0=1$, y is output and i is investment. I start with $y=f(k)$, and then generalize to add labor $y=f(k,n)$ and adjustment costs to 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 cumulated growth rate m.

Now, how do we extend the productivity choice constraint? We can approach this question in several ways.

5.1 A sum constraint

A natural way to extend the idea to a dynamic model is to write the constraint

$$E\left[(1-\rho)\sum_{t=0}^{\infty}\rho^t\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]\leq 1. \tag{31}$$

[3] From (27), we have

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

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

Taking the expectation and using the productivity choice constraint gives

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

Substituting this result back into (29), we have (30).

This formulation parallels the extension of power utility from a one-period setting $E\left(c^{1-\gamma}\right)$ to an intertemporal setting $E\sum_{t=1}^{\infty}\beta^t (c_t)^{1-\gamma}$. As in that case, however, this formulation allows the firm to substitute productivity over time, trading ε_t for ε_{t+1}, as well as across states of nature. I turn below to ideas that separate time versus risk.

Using a simple production technology $y=\varepsilon f(k)$, the firm's time-zero contingent claim problem is now

$$\max E\sum_{t=1}^{\infty}\rho^{t-1}\Lambda_t\left[\varepsilon_t f(k_t)-i_t\right]$$

$$\text{s.t. } k_{t+1}=(1-\delta)k_t+i_t,\ k_0, \tag{32}$$

together with (31). I scale the discount factor by the same time constant ρ as appears in the constraint (31). This is just a convenience to produce stationary solutions, as we often scale $\beta^t\Lambda_t$ so we can write $u'(c_t)=\Lambda_t$. Otherwise, we obtain growth in either the discount factor or productivity, which is fine but adds extra terms in powers of ρ. Growing θ can also make up any difference in growth rates.

The first-order conditions, varying investment and then productivity, are

$$\Lambda_t=E_t\left\{\rho\Lambda_{t+1}\left[\varepsilon_{t+1}f_k(k_{t+1})-(1-\delta)\right]\right\} \tag{33}$$

$$\Lambda_{t+1}f(k_{t+1})=\lambda(1+\alpha)(1-\rho)\frac{\varepsilon_{t+1}^{a}}{\theta_{t+1}^{1+\alpha}}, \tag{34}$$

where λ is the Lagrange multiplier on the productivity-choice constraint (31).

Condition (33) is the familiar intertemporal condition. It says to invest so that the one-period return from physical investment $R_{t+1}^{I}\equiv \varepsilon_{t+1}f_k(k_{t+1})-(1-\delta)$ is correctly priced by the discount factor. In this equation and below, keep in mind that capital k_{t+1} is known at time t.

Equation (34) ties the discount factor to productivity, just as the consumption-based discount factor is tied to consumption, $\Lambda_t=u'(c_t)$. It says to raise productivity in states with high contingent claim prices, in states in which it is easier to do so, and in states with higher output, since productivity multiplies output.

Dividing adjacent periods, Equation (34) leads to a discount factor for one-period returns comprising productivity growth and capital or productivity growth and output,

$$m_{t+1}=\frac{\rho\Lambda_{t+1}}{\Lambda_t}=\rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}\frac{f(k_t)}{f(k_{t+1})} \tag{35}$$

$$=\rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{1+\alpha}\left(\frac{y_{t+1}}{y_t}\right)^{-1}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \tag{36}$$

This expression is nicely analogous to the consumption-based discount factor proportional to consumption growth $\Lambda_{t+1}/\Lambda_t = (c_{t+1}/c_t)^{-\gamma}$. It leads to a multifactor macro-asset pricing model.

We can eliminate the multiplier λ to express productivity choice as before,[4]

$$\frac{\varepsilon_t^a}{\theta_t^\alpha} = \frac{\Lambda_t \theta_t f(k_t)}{\left\{E\sum_{t=0}^{\infty}(1-\rho)\rho^t\left[\Lambda_{t+1}\theta_{t+1}f(k_{t+1})\right]^{\frac{1+\alpha}{\alpha}}\right\}^{\frac{\alpha}{1+\alpha}}}. \tag{37}$$

This problem also allows a recursive statement, which is an easier basis for numerical solution of more complex models. Write the constraint (31) recursively as

$$W_t^{1+\alpha} \equiv E_t(1-\rho)\sum_{j=1}^{\infty}\rho^{j-1}\left(\frac{\varepsilon_{t+j}}{\theta_{t+j}}\right)^{1+\alpha} = E_t\left[(1-\rho)\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha} + \rho W_{t+1}^{1+\alpha}\right] \tag{38}$$

with $W_0 = 1$. At time t, the firm picks for each state at $t+1$ values for ε_{t+1} and W_{t+1}, subject to the constraint (38). The recursive problem is then

$$V(k_t, W_t, \varepsilon_t) = \max_{\{k_{t+1}, \varepsilon_{t+1}, W_{t+1}\}} \{\varepsilon_t f(k_t) - [k_{t+1} - (1-\delta)k_t]\} + E_t\left[\frac{\rho\Lambda_{t+1}}{\Lambda_t}V(k_{t+1}, W_{t+1}, \varepsilon_{t+1})\right]$$

subject to (38), k_0, and $W_0 = 1$. The first-order and envelope conditions of this recursive statement give the same results (33) and (34). See the Internet Appendix for algebra.

This recursive statement also allows us to think about the problem starting from time t, and its conditional information set. The problem is the same, with the constraint equal to $W_t^{1+\alpha}$ rather than one, as reopening a consumer problem at time t is the same, with conditional expectations and time-t wealth in the constraint.

An arbitrage argument offers insight. Since the constraint links date and time, we can synthesize any return, by producing a little more dy_{t+1} in the states of nature described by that return, at the cost of

[4] From (34),

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

Imposing the constraint,

$$E\sum_{t=0}^{\infty}(1-\rho)\rho^t\left[\Lambda_{t+1}\theta_{t+1}f(k_{t+1})\right]^{\frac{1+\alpha}{\alpha}} = \left[\lambda(1+\alpha)(1-\rho)\right]^{\frac{1+\alpha}{\alpha}}.$$

Substituting in (34) gives (37).

producing a little less dy_t, as described by the constraint. Differentiating the constraint (31) with respect to ε_t and all ε_{t+1} following the date and state ε_t,

$$\frac{\varepsilon_t^\alpha}{\theta_t^{1+\alpha}} d\varepsilon_t = E_t\left[\rho\left(\frac{\varepsilon_{t+1}^\alpha}{\theta_{t+1}^{1+\alpha}}\right) d\varepsilon_{t+1}\right].$$

The production process gives extra output according to $dy_t = f(k_t) d\varepsilon_t$, so the extra output dy_{t+1} and its cost dy_t obey

$$\frac{\varepsilon_t^\alpha}{\theta_t^{1+\alpha} f(k_t)} dy_t = E_t\left[\rho\left(\frac{\varepsilon_{t+1}^\alpha}{\theta_{t+1}^{1+\alpha}}\right)\frac{1}{f(k_{t+1})} dy_{t+1}\right].$$

Defining $R_{t+1} = dy_{t+1}/dy_t$, the firm can thus synthesize any return R_{t+1}

$$1 = E_t\left[\rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{1+\alpha}\left(\frac{y_{t+1}}{y_t}\right)^{-1}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)} R_{t+1}\right]. \tag{39}$$

We recognize the standard asset pricing equation with the discount factor (36).

In this way, we can view this production-based asset pricing model as an extension of the arbitrage argument of investment-based asset pricing. Investment-based asset pricing notices that the firm can synthesize a return R^I_{t+1} by varying investment at time t and $t+1$. Therefore any security with state-contingent payoffs R^I_{t+1} must have price one. This argument tells us how similarly to price any set of random payoffs, by synthesizing them via the productivity choice constraint.

5.2 Separating time and state by recursive production and CES

If we wish to separate the economics of time and state in production-based asset pricing, treated symmetrically in the sum constraint (31), we can naturally follow the approaches that separate time and state in utility theory. Unlike the case in utility theory, where the axioms of expected utility lead to state-separability, nothing (yet) but convenience indicates that production technology should be separable across sates. So non-state-separable and non-time-separable production functions are useful possibilities to consider.

First, we can follow the Epstein and Zin 1989 recursive utility path. To get there, write the recursive statement of the sum constraint (38) as

$$R_{t+1} = \left[(1-\rho)\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha} + \rho W_{t+1}^{1+\alpha}\right]^{\frac{1}{1+\alpha}}$$

$$W_t = \left[E_t\left(R_{t+1}^{1+\alpha}\right)\right]^{\frac{1}{1+\alpha}}.$$

Here, I just introduce the notation R_{t+1} for the inside term of (38). Now we can generalize this constraint by changing the parameter of the first equation to $\sigma\neq\alpha$:

$$R_{t+1}=\left[(1-\rho)\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\sigma}+\rho W_{t+1}^{1+\sigma}\right]^{\frac{1}{1+\sigma}}$$

$$W_t=\left[E_t\left(R_{t+1}^{1+\alpha}\right)\right]^{\frac{1}{1+\alpha}}.$$

The parameter σ describes the firm's ability to transform productivity from one date to another. The parameter α describes its ability to transform productivity from one state to another.

The one-period discount factor becomes

$$m_{t+1}=\rho\frac{1+\sigma}{1+\alpha}\left(\frac{R_{t+1}}{\left[E_t\left(R_{t+1}^{1+\alpha}\right)\right]^{\frac{1}{1+\alpha}}}\right)^{\alpha-\sigma}\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{1+\sigma}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{1+\sigma}\left(\frac{y_{t+1}}{y_t}\right)^{-1}. \tag{40}$$

The Internet Appendix presents the algebra. As in recursive utility, a state variable, R_{t+1}, that combines current ε_{t+1} and future productivities now enters the discount factor, and it defines risk exposures. Identifying that state variable takes a lot of effort on the consumption side, and would likely require effort on the production side as well. But it also opens the door to an interesting menagerie of pricing factors. And, we can observe the firm's stock price where we cannot observe the consumer's utility, which may help.

Second, we can simply describe productivities as a CES aggregate, with distinct elasticities across time σ and states α,

$$\left\{\sum_{t=0}^{\infty}\rho^t\left[E\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]^{\frac{1+\sigma}{1+\alpha}}\right\}^{\frac{1}{1+\sigma}}\leq 1. \tag{41}$$

The one-period discount factor becomes

$$m_{t+1}=\left\{\frac{E\left[\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]}{E\left[\left(\frac{\varepsilon_t}{\theta_t}\right)^{1+\alpha}\right]}\right\}^{\frac{\sigma-\alpha}{1+\alpha}}\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{1+\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}\left(\frac{y_{t+1}}{y_t}\right)^{-1}. \tag{42}$$

The first term alters our description of the firm's ability to transform across time, and now controls the risk-free rate or other level of returns. Zero cost portfolios can be priced by the same discount factor as before, ignoring the first term. Productivity ε, still raised to $1+\alpha$, carries risk pricing. In the recursive formulation (40) the direct ε productivity term is raised to the $1+\sigma$ power and describes intertemporal substitution. The R_{t+1} term (which also contains ε_{t+1}) carries all the risk aversion, but one must calculate or measure that state variable somehow to measure risk premiums. So this CES aggregate is simpler and easier to operationalize, though it opens fewer doors to additional pricing factors.

5.3 A constraint each period

One can also allow productivity choice in a dynamic model by writing a separate constraint for each time period,

$$E\left[\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]\leq 1 \tag{43}$$

in place of the single constraint (31). This formulation does not allow the firm a marginal rate of transformation between time periods by productivity choice. All intertemporal transformation has to go through the investment return.

The firm maximizes the contingent claim value of output,

$$\max E\sum_{t=1}^{\infty}\rho^{t-1}\Lambda_t\{\varepsilon_t f(k_t)-[k_{t+1}-(1-\delta)k_t]\}$$

but now subject to the constraints (43) rather than (31). The first-order conditions are

$$\frac{\partial}{\partial i_t},\frac{\partial}{\partial i_{t+1}}:1=E_t\left\{\frac{\rho\Lambda_{t+1}}{\Lambda_t}[\varepsilon_{t+1}f_k(k_{t+1})+(1-\delta)]\right\} \tag{44}$$

$$\frac{\partial}{\partial \varepsilon_{t+1}}:\Lambda_{t+1}f(k_{t+1})=\lambda_{t+1}(1+\alpha)\frac{\varepsilon_{t+1}^{\alpha}}{\theta_{t+1}^{1+\alpha}}, \tag{45}$$

where $\rho^t\lambda_{t+1}$ is the Lagrange multiplier on the constraint (43). It varies over time, but not across states of nature.

Equation (44) again says that the investment return should be correctly priced. Equation (45) again says to produce more in high contingent claim price states, high natural productivity θ states, and high output states.

The difference between a separate constraint for each period (43) and the previous constraint on the sum (31) is that we have time-varying Lagrange multipliers λ_{t+1} in (45) rather than a constant λ. The multiplier λ_{t+1} measures the shadow value of transforming intertemporally, of trading some ε_t for the ability to increase all of the following ε_{t+1}.

We can quickly construct a discount factor that prices zero-cost portfolios. Such a discount factor m^* at time t satisfies

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

It 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

$$m^*_{t+1} = \rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \tag{46}$$

This discount factor is the same as the zero cost portfolio discount factor deriving from the sum constraint, (35), differing only by time t random variables.

One could use the same discount factor for zero cost portfolios as in the one-period model,

$$m^*_{t+1} = b_t \varepsilon^{\alpha}_{t+1} \theta^{-(1+\alpha)}_{t+1}, \tag{47}$$

for any b_t. However, productivity ε_t, like consumption, is typically very persistent and grows over time, and θ_t should have similar properties. So, while (47) with $b_t = 1$, say, prices zero-cost portfolios, its conditional mean $E_t(m^*_{t+1})$ and implied risk-free rate $R^f_t = 1/E_t(m^*_{t+1})$ vary strongly over time, and it is potentially nonstationary violating the assumptions of all time-series empirical work. Choosing growth rates as at least an initial scaling, as in (46), is wise for typical time-series applications. Analogously, we typically use $m_{t+1} = \beta(c_{t+1}/c_t)^{-\gamma}$, though $m^*_{t+1} = c^{-\gamma}_{t+1}$ also prices zero cost portfolios.

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^f_t with

$$m^*_{t+1} = \frac{1}{R^f_t} \frac{\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}}{E_t\left[\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}\right]}, \tag{48}$$

with or without ε_t and θ_t in the denominators. Scaling a discount factor to have a reasonable implied risk-free rate has proven wise in empirical work.

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,

$$m_{t+1} = \frac{\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}}{E_t\left[\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)} R^I_{t+1}\right].}. \tag{49}$$

The expectation in the denominator is a time-t random variable, so it fits in to the rubric of (47). This discount factor prices R^I_{t+1} and all zero cost portfolios R^e_{t+1}, so it prices all returns R_{t+1}.

However, the point of this model is to separate the economics of time and risk. Therefore, it may be clearer to examine its implications for risk premiums via (46) or (48) and separately to examine its investment returns, rather than to cloud the latter economics by constructing the grand discount factor of (49).

The expectation in (43) is unconditional, as in the sum constraint (31). The Internet Appendix considers whether the expectation in (43) should be conditional, $E_t(\cdot)$, unconditional, $E(\cdot)$ or somewhere in between $E_\tau(\cdot)$, $0<\tau<t$. The distinction matters for the dynamic properties of the chosen ε_t given a discount rate process, but it makes no difference to the discount rate formulas here. In the Internet Appendix, I conclude that an unconditional or $\tau << t$ formulation is more reasonable. Absent serially correlated natural productivity θ, a conditional constraint leads to a productivity level proportional to discount factor growth, and thus to a productivity level that is poorly serially correlated. An unconditional constraint or $\tau << t$ more naturally produces serially correlated productivity and a one period discount factor m_{t+1} related to productivity growth. The unconditional constraint also generalizes more easily to continuous time.

We can write the constraint $E_\tau\left[(\varepsilon_{t+1}/\theta_{t+1})^{1+\alpha}\right]\leq 1$ recursively as

$$z_{t,t+1}=\left\{E_t\left[\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]\right\}^{\frac{1}{1+\alpha}} \tag{50}$$

$$z_{t-j,t+1}=\left\{E_{t-j}\left[(z_{t-j+1,t+1})^{1+\alpha}\right]\right\}^{\frac{1}{1+\alpha}};j=1,2,...\tau \tag{51}$$

$$z_{\tau,t+1}=1. \tag{52}$$

A specification $\tau<t$ thus amounts to applying the productivity-choice idea to the constraint itself. The firm can take actions at $t-1$ to loosen the time-t constraint on ε_{t+1} in one time-t state, though tightening that constraint in another time-t state. The firm may begin the process of adjusting the time-$t+1$ productivity anytime after τ, as captured by evolution of state variables $z_{s,t+1}$ as s proceeds forward.

The firm's actions to transform across states of nature involve time. The farmer plants seeds in different fields in the spring, but after that he or she can do little to transform fall output across weather states. The electric utility buys flexible or fuel-optimized equipment, but after that it can do little 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. The $\tau<t$ specification allows that idea. Ideally, we would like to capture the changing difficulty of making choices as

the data approaches by adding θ shocks to the z choice in (51) or by varying the value of α over horizon.

5.4 More detailed production processes

More complex and realistic models of intertemporal production make the formula for the investment return R^I more complicated. They also 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

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

subject to

$$y_t = \varepsilon_t f(k_t, n_t) - \psi(i_t, k_t)$$

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

and either

$$E\left[(1-\rho)\sum_{t=0}^{\infty} \rho^t \left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right] \leq 1 \tag{53}$$

or

$$1 = E_\tau \left[\left(\frac{\varepsilon_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]. \tag{54}$$

The intertemporal first-order condition becomes

$$1 = E_t \left(m_{t+1} R^I_{t+1}\right) \tag{55}$$

with

$$R^I_{t+1} \equiv \frac{\varepsilon_{t+1} f_k(k_{t+1}, n_{t+1}) - \psi_k(i_{t+1}, k_{t+1}) + (1-\delta)[1 + \psi_i(i_{t+1}, k_{t+1})]}{1 + \psi_i(i_t, k_t)}.$$

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, either

$$\Lambda_{t+1} f(k_{t+1}, n_{t+1}) = \lambda (1+\alpha)(1+\rho) \frac{\varepsilon^\alpha_{t+1}}{\theta^{1+\alpha}_{t+1}} \tag{56}$$

with the single sum constraint (53) or

$$\Lambda_{t+1} f(k_{t+1}, n_{t+1}) = \lambda_{t+1} (1+\alpha) \frac{\varepsilon^\alpha_{t+1}}{\theta^{1+\alpha}_{t+1}} \tag{57}$$

with the period by period constraint (54).

Relative to the cases with no labor and adjustment costs, (34) and (45), the productivity choice conditions (56) and (57) differ by the generalization $f(k_{t+1},n_{t+1})$ in place of $f(k_{t+1})$. This substitution adds employment or wage to the discount factor formula, just as in the one-period model with labor. Adjustment costs would also enter the discount factor formula if we wrote productivity to multiply them, that is, $y_t=\varepsilon_t[f(k_t,n_t)-\psi(i_t,k_t)]$. From (56), the sum-constraint model's discount factor is similar to the forms (35),

$$m_{t+1}=\rho\frac{\Lambda_{t+1}}{\Lambda_t}=\rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{f(k_{t+1},n_{t+1})}{f(k_t,n_t)}\right)^{-1}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}$$
$$=\rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{1+\alpha}\left(\frac{y_{t+1}}{y_t}\right)^{-1}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \tag{58}$$

Using a Cobb-Douglas production function

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

we can write the first expression of the sum-constraint model's discount factor as

$$m_{t+1}=\rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{k_{t+1}}{k_t}\right)^{-(1-\sigma)}\left(\frac{n_{t+1}}{n_t}\right)^{-\sigma}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \tag{59}$$

Using the first-order condition for labor input,

$$\varepsilon_{t+1}\sigma k_{t+1}^{1-\sigma}n_{t+1}^{\sigma-1}=w_{t+1},$$

we can write the same discount factor in terms of wages,

$$m_{t+1}=\rho\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha-\frac{\sigma}{1-\sigma}}\left(\frac{k_{t+1}}{k_t}\right)^{-1}\left(\frac{w_{t+1}}{w_t}\right)^{\frac{\sigma}{1-\sigma}}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \tag{60}$$

From the separate-constraint first-order condition (57), we can write similar zero-cost portfolio discount factors. The one-period formulas (25) and (28) remain valid, just add time $t+1$ subscripts. Growth rate formulas are likely to be more useful, for example,

$$m^*_{t+1}=b_t\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha}\left(\frac{n_{t+1}}{n_t}\right)^{-\sigma}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}, \tag{61}$$

$$m^*_{t+1}=b_t\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{1+\alpha}\left(\frac{y_{t+1}}{y_t}\right)^{-1}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}, \tag{62}$$

or

$$m^*_{t+1}=b_t\left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{\alpha-\frac{\sigma}{1-\sigma}}\left(\frac{w_{t+1}}{w_t}\right)^{\frac{\sigma}{1-\sigma}}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \tag{63}$$

where b_t can be set as convenient.

(Equation (63) seems to offer the possibility of a discount factor negatively correlated with productivity. However, $\alpha > \sigma/(1-\sigma)$ is the condition for a convex problem. Otherwise, the firm chooses all of its production in one state. I have left implicit the restriction $\varepsilon \geq 0$.)

Measuring productivity is difficult. Belo 2016 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 productivity 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,

$$\frac{\varepsilon_{t+1}}{\varepsilon_t} = \left(\frac{w_{t+1} n_{t+1}/y_{t+1}}{w_t n_t / y_t} \right)^{\frac{\sigma}{\sigma-1}} \left(\frac{y_{t+1}/n_{t+1}}{y_t/n_t} \right). \tag{64}$$

Substituting these expressions into (58) or (62), we obtain a discount factor with output growth, labor share growth, and labor/output ratio growth as factors, and no explicit productivity, for example,

$$m_{t+1} = \rho \left(\frac{w_{t+1} n_{t+1}/y_{t+1}}{w_t n_t / y_t} \right)^{\frac{\sigma(1+\alpha)}{\sigma-1}} \left(\frac{y_{t+1}/n_{t+1}}{y_t/n_t} \right)^{1+\alpha} \left(\frac{y_{t+1}}{y_t} \right)^{-1} \left(\frac{\theta_{t+1}}{\theta_t} \right)^{-(1+\alpha)}. \tag{65}$$

An important bottom line: as we make the production function more detailed and realistic, a wide variety of production variables, not just productivity ε, enter the discount factor. Now, as in the case of labor input, these variables are also chosen given the discount factor and productivity shock, so in principle they do not offer additional information and a multifactor model is not needed. But wages here induce independent movement in labor input. The situation is much like that of multiple nonseparable goods $u(c_a, c_b, \ldots)$ in consumption theory, where their relative quantities or relative prices enter the discount factor.

5.5 Constraints on net output

We might also go back to first principles. To extend the production-based asset pricing idea to multiple dates, why not proceed exactly in analogy to consumption-based asset pricing? Rather than apply analogs to consumption-based asset pricing to the choice of *productivity,* as I have done so far, why not apply those analogs to the firm's final output net of investment directly? Microeconomics textbooks treat production and consumption with beautiful symmetry. Why not us?

We can write the firm's two-period problem as

$$\max c_0+E(mc_1) \text{ s.t. } \left\{\left(\frac{c_0}{\theta_0}\right)^{1+\alpha}+\rho E\left[\left(\frac{c_1}{\theta_1}\right)^{1+\alpha}\right]\right\}^{\frac{1}{1+\alpha}}\leq K, \quad (66)$$

where c denotes the firm's final output sold to consumers, that is, $c=y-i$, and K is a constant. This production set is concave and smooth across time and across states of nature.

Explicitly, in the finite-state case, the firm's problem is

$$\max c_0+\sum_s \pi(s)m(s)c_1(s)$$

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

The first-order conditions to this problem lead to

$$m_1=\rho\left(\frac{c_1}{c_0}\right)^{\alpha}\left(\frac{\theta_1}{\theta_0}\right)^{-(1+\alpha)}. \quad (67)$$

The parallel to power utility is immediate.

We can generalize this approach to multiperiod problems and continuous time transparently,

$$\max E\sum_{t=1}^{\infty}\rho^{t-1}\Lambda_t c_t \text{ s.t. } E\left[(1-\rho)\sum_{t=1}^{\infty}\rho^{t-1}\left(\frac{c_t}{\theta_t}\right)^{1+\alpha}\right]\leq 1, \quad (68)$$

and

$$\max E\int_{t=0}^{\infty}\rho^t\Lambda_t c_t dt \text{ s.t. } E\left[\frac{1}{\rho}\int_{t=0}^{\infty}\rho^t\left(\frac{c_t}{\theta_t}\right)^{1+\alpha}dt\right]\leq 1.$$

We are used to Dixit-Stiglitz aggregators across goods. This formulation applies the same idea over time. The resultant discount factor is simply

$$\Lambda_t=\lambda\frac{c_t^{\alpha}}{\theta_t^{1+\alpha}};\, m_{t+1}=\rho\frac{\Lambda_{t+1}}{\Lambda_t}=\left(\frac{c_{t+1}}{c_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \quad (69)$$

We have an output-based macro-factor model, not one based on productivity. That productivity loomed so large in the previous analysis was entirely a modeling choice.

Writing analogs to nonseparable utility that distinguish transformation over time from transformation across states of nature is straightforward as well. We can quickly write an Epstein and Zin 1989 style recursive non-state-separable constraint on final net output,

$$R_{t+1} = \left[(1-\rho)\left(\frac{c_{t+1}}{\theta_{t+1}}\right)^{1+\sigma} + \rho W_{t+1}^{1+\sigma} \right]^{\frac{1}{1+\sigma}} \tag{70}$$

$$W_t = \left[E_t\left(R_{t+1}^{1+\alpha}\right)\right]^{\frac{1}{1+\alpha}}. \tag{71}$$

Again, the discount factor will include the state variable R_{t+1} as in (40).

We can write a CES constraint that separates time and risk by simply aggregating over time and states with different coefficients,

$$(1-\rho)\sum_{t=1}^{\infty}\rho^{t-1}\left\{E\left[\left(\frac{c_t}{\theta_t}\right)^{1+\alpha}\right]\right\}^{\frac{1+\sigma}{1+\alpha}} \leq 1. \tag{72}$$

Then we obtain

$$\Lambda_t = \lambda(1-\rho)(1+\sigma)\left\{E\left[\left(\frac{c_t}{\theta_t}\right)^{1+\alpha}\right]\right\}^{\frac{\sigma-\alpha}{1+\alpha}} \frac{c_t^{\alpha}}{\theta_t^{1+\alpha}}$$

and

$$m_{t+1} = \rho\left\{\frac{E\left[\left(\frac{c_{t+1}}{\theta_{t+1}}\right)^{1+\alpha}\right]}{E\left[\left(\frac{c_t}{\theta_t}\right)^{1+\alpha}\right]}\right\}^{\frac{\sigma-\alpha}{1+\alpha}} \left(\frac{c_{t+1}}{c_t}\right)^{\alpha}\left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)}. \tag{73}$$

Zero-cost portfolios can still be priced using (69). But $E(m) = 1/R^f$ is now distorted by the first term in brackets of (73). Intertemporal transformation and risk-transformation are separated.

Incidentally, similar CES preferences seem like a useful alternative to recursive preferences for consumption-based asset pricing. Write the consumer's objective

$$\max\sum_{t=0}^{\infty}\beta^t\left[E\left(c_t^{1-\gamma}\right)\right]^{\frac{1-\sigma}{1-\gamma}}.$$

The consumer's first-order conditions lead to

$$m_{t+1} = \beta\left[\frac{E\left(c_{t+1}^{1-\gamma}\right)}{E\left(c_t^{1-\gamma}\right)}\right]^{\frac{\gamma-\sigma}{1-\gamma}}\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}.$$

The first term distorts intertemporal substitution relative to risk aversion, a main goal of recursive utility. Consumption to a power still describes risk aversion, so zero-cost portfolios do not require (or allow) computation of the utility index that makes recursive utility complex and fun.

Why not describe production sets in terms of net output, following this more elegant approach to production-based asset pricing? One answer is that we then lose the connection to standard production theory. A standard intertemporal production function, say

$$y_t = \varepsilon_t f(k_t) \tag{74}$$

$$k_{t+1} = (1-\delta)k_t + (y_t - c_t) \tag{75}$$

does not have a pretty representation in terms of final output $c = y - i$. Derivatives $dc_{t+1}/dc_t = \varepsilon_{t+1} f_k(k_{t+1}) + (1-\delta)$ are well defined, and $dc_{t+1}^2/dc_t^2 < 0$. But the resultant production set is 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 throw out the contact with classic production theory, and in particular with the successes of investment-return-based asset pricing and all existing general equilibrium macroeconomics and asset pricing. Alternatively, we can add productivity choice to standard production theory as I have so far, and express production-based asset pricing as a constraint on productivity rather than final output. That choice 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 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 capital that later produces more output. We think of transformation across states using stories, such as planting in fields with different state sensitivities, investing in machines with different sensitivities, and so on. Perhaps keeping time and risk separate is wise, if inelegant.

5.6 Durability-like dynamics

These extensions to dynamic problems are not as pretty as I would like them to be. Fundamentally, the constraints

$$E\left[(1-\rho)\sum_{t=0}^{\infty} \rho^t (\varepsilon_{t+1}/\theta_{t+1})^{1+\alpha}\right] \leq 1$$

or

$$E\left[\left(\varepsilon_{t+1}/\theta_{t+1}\right)^{1+\alpha}\right] \leq 1$$

allow completely different random variables ε_t at each date t. One would suppose that the distribution of productivity at time t cannot 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 situation is similar in utility theory. Preferences $E\sum_{t=0}^{\infty}\beta^t u(c_t)$ or $E\int e^{\delta t}u(c_t)dt$ allow the consumer to rank consumption processes with completely different distributions at each point in time, a particularly frightful prospect in continuous time.

In both cases, this is not typically a practical worry. If circumstances—the discount factor Λ_t, natural productivity θ_t— evolve as continuous functions of time, so will the choice ε_t. If we invert to find discount factors as a function of choices that vary continuously over time, and whose estimated distributions vary continuously over time, we will find discount factors that vary continuously over time. So the model will not generally produce crazy predictions. Still, the description of production sets is inelegant.

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 (Hindy and Huang 1992). This modification tends not to be used however, because the first-order conditions for durable goods are more complex. While solving models with durable goods is a simple exercise, doing so violates the philosophy of consumption-based pricing, to read discount factors from first-order conditions without solving models.

A similar situation applies to production sets. We would like a productivity choice set in which productivity at nearby dates must have similar distributions. The distributions can then 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 ε_{t+s} 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, just as it does for durable consumption goods.

To explore this idea in a simple environment, suppose we write the choice set as a constraint on the *growth* of productivity:

$$E_\tau\left[\left(\frac{\varepsilon_{t+\Delta}}{\theta_{t+\Delta}}/\frac{\varepsilon_t}{\theta_t}\right)^{1+\alpha}\right] \leq 1. \tag{76}$$

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 includes 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. Specification (76) also leads to a natural continuous-time expression,

$$E_\tau\left[\left(\frac{d(\varepsilon_t/\theta_t)}{\theta_t/\varepsilon_t}\right)^{1+\alpha}\right]=0. \tag{77}$$

Specifications (76) or (77) result in productivity ε_t that wanders further away from its initial value ε_0, and from the underlying shock θ_t, for a longer time horizon.

So far so good, but the first-order conditions become more complicated, because changing ε_t changes the choice set for all subsequent ε_{t+s}. The resultant first-order conditions are difficult to unwind to a discount factor. With a constraint on the growth of productivity (76), the firm's problem is

$$\max E\sum_{t=1}^{\infty}\rho^{t-1}\Lambda_t[\varepsilon_t f(k_t)-i_t]\,\text{s.t.}$$

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

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

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

$$\rho^{t-1}\Lambda_{t+1}f(k_{t+1})=\lambda_t(1+\alpha)\left(\frac{\varepsilon_{t+1}^{\alpha}}{\theta_{t+1}^{1+\alpha}}/\frac{\varepsilon_t^{1+\alpha}}{\theta_t^{1+\alpha}}\right)$$
$$-\lambda_{t+1}(1+\alpha)E_{t+1}\left(\frac{\varepsilon_{t+2}^{1+\alpha}}{\theta_{t+2}^{1+\alpha}}\right)/\frac{\varepsilon_{t+1}^{2+\alpha}}{\theta_{t+1}^{1+\alpha}}. \tag{78}$$

We can use (78) recursively to write the productivity-choice first-order condition as

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

In this form, one can more clearly see 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, one can 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 not difficult to implement. But inferring the discount factor from productivity is not as pretty as in the time-separable cases. A state variable, parallel to the stock of durable goods, would help.

Putting these thoughts together, a useful way to describe the choice of technology may be to let the firm change its technology shock distribution, but at a cost. For example, let technology follow

$$y_t = e^{\varepsilon_t} f(k_t)$$

$$\frac{dk_t}{dt} = -\delta k_t + i_t.$$

Let there be a vector of shocks $dz_t = [dz_{1t}\ dz_{2t}\ ...dz_{Nt}]'$, and productivity responds by

$$d\varepsilon_t = \mu'_{\varepsilon t} dt + \sigma'_{\varepsilon t} dz_t. \tag{79}$$

The discount factor responds to the same shocks,

$$d\Lambda_t = \mu'_{\Lambda t} dt + \sigma'_{\Lambda t} dz_t.$$

The firm maximizes the contingent claim value of output

$$\max E \int_{t=0}^{\infty} \rho^t \Lambda_t (y_t - i_t - \psi_t,) dt$$

where ψ_t are the costs of adjusting the technology shock distribution. If the firm does nothing, technology will evolve as usual as described by (79). However, the firm can adjust the distribution of technology, at a cost,

$$\psi dt = \left(\frac{d\mu_{\varepsilon t}}{dt}\right)^{1+\alpha} + \sum_{j=1}^{N} \left(\frac{d\sigma_{\varepsilon tj}}{dt} / \theta_j\right)^{1+\alpha}.$$

We can even let the vector of cost weights move over time,

$$d\theta_t = \mu'_{\theta t} dt + \sigma'_{\theta t} dz_t.$$

This direction may give an interesting dynamic model of productivity choice. It generates a marginal rate of transformation for discrete time intervals, as the firm can get to any σ_Λ. But it does not seem to lead to an analog to consumption-based asset pricing, so I do not follow this lead here.

6. General Equilibrium, Identification, and Calibration

The discount factors we have studied focus on powers of productivity growth, for example,

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

and

$$m_{t+1} = \rho \left(\frac{\varepsilon_{t+1}}{\varepsilon_t}\right)^{1+\alpha} \left(\frac{y_{t+1}}{y_t}\right)^{-1} \left(\frac{\theta_{t+1}}{\theta_t}\right)^{-(1+\alpha)} \tag{80}$$

as well as other production-related macroeconomic variables including labor, capital, wages, and investment. Will these or similar specifications of such a production-based model be empirically successful? This section takes up this question, as well as the troublesome question of whether and what kinds of natural productivity shocks θ we need, and how to identify them.

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. 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,

$$E(R^e) = -\frac{cov(m, R^e)}{E(m)}. \tag{81}$$

This relation implies

$$\frac{E(R^e)}{\sigma(R^e)} = -\frac{\sigma(m)}{E(m)}\rho(m, R^e). \tag{82}$$

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 γ are difficult to swallow. It is difficult to generate $\sigma[(c_{t+1}/c_t)^{-\gamma}]$ on the order of 50%.

Output, productivity, and employment 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 $\sigma(\varepsilon^{\alpha})$ will not be difficult, and the classic equity premium puzzle does not cause an obvious problem for production-based asset pricing. More deeply, since marginal rates of transformation don't have any natural relationship to probabilities, the natural force toward risk neutrality is missing. Easy simulation models tend to produce equity premiums that are too *high*.

The discount factor should have a low and fairly stable conditional mean, to generate a low and relatively stable real risk-free rate $E_t(m) = 1/R_t^f$. In all these models, the level of returns and risk-free rate are governed by conventional investment-return economics, separate from risk premiums, so the model really has nothing new to say about the level of the risk-free rate. The sum-constraint discount factor (80) does predict a risk free rate. But in the model, the firm chooses productivity ε so that investment returns are correctly priced so again the model as a whole has nothing new to offer. The data may violate that restriction. However, the growth rates in (80) are stationary and not severely serially correlated, so risk-free rate variability is not an immediate problem. Since productivity is more volatile than consumption, an equity premium is not likely to require an enormous value of α and hence an enormous $E(m)$ via $E\left[(\varepsilon_{t+1}/\varepsilon_t)^{1+a}\right]$. And the framework, with additional right-hand-side variables and especially the free parameter ρ and growth in the θ process beckoning, seems flexible enough to quickly adapt to any problems with $E(m)$.

Relation (82) 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 variation in α_t endogenously.

The most obvious obstacle, however, is the sign of the covariance term in (81). To generate a positive risk premium $E(R^e_{t+1})$, the discount factor m_{t+1} must covary negatively with the ex post excess return R^e_{t+1}. In consumption-based asset pricing, $m_{t+1} = \beta(c_{t+1}/c_t)^{-\gamma}$, 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. That model's failure is one of magnitude, not of sign. In production-based asset pricing we have $m^*_{t+1} = (\varepsilon_{t+1}/\varepsilon_t)^{\alpha}(\theta_{t+1}/\theta_t)^{-(1+\alpha)}$, however, with $\alpha > 0$. If there are no natural productivity shocks θ, productivity growth $\varepsilon_{t+1}/\varepsilon_t$ is positively correlated with the discount factor. A positive correlation of productivity growth with asset returns R^e predicts 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 more in such high-price "bad times" states.

Now, many possibilities avoid this conundrum. The conundrum presumes a one-factor model in which consumption, productivity, and asset returns all move together. Maybe productivity is higher in bad times. Whether productivity is procyclical or countercyclical is debated in macroeconomics. (Measuring productivity is a headache too.) Yes, real business cycle models generate recessions by procyclical productivity shocks, but the rest of macroeconomics in the new Keynesian DSGE tradition is deliberately constructed to avoid that mechanism. In recessions, firms not only produce less but also shed workers and machines, especially unproductive workers and machines (Grigsby 2020). Varying composition, factor utilization, effort, and labor hoarding cloud the productivity picture.

Already, output enters discount factors (58) and (59)) with the "right" negative sign, as do labor input and capital growth. Maybe even if productivity growth is positively correlated with asset returns a rich enough model's discount factor will be negatively correlated with returns, without underlying θ shocks. Perhaps in a dynamic model, shocks θ that vary across time, but not states of nature, are sufficient to produce the observed moments.

Moreover, 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. 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.

Furthermore, 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 choose higher productivity ε in that state. Consumers will consume more in that state, driving down the discount factor or contingent claim price of that state. This lower price causes firms to back off so as to raise productivity ε somewhat less in the high-θ, low-price state, and to lower productivity somewhat less in low-θ, high-price states. But the product $m = (\varepsilon_{t+1}/\varepsilon_t)^{\alpha} (\theta_{t+1}/\theta_t)^{-(1+\alpha)}$ still moves negatively with θ, so the discount factor moves negatively with productivity, consumption, and asset returns, despite the positive coefficient α.

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 toward 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.

6.1 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. I present the model in this subsection, and analyze the central equilibrium conditions in the next subsection.

Add consumers with utility

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

where

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

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.

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

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

The consumers' first-order conditions are

$$m = \lambda_c u'(c) = \lambda_c c^{-\gamma}/\phi^{1-\gamma} \tag{83}$$

so consumption is

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

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

$$c = E(me) \frac{m^{-\frac{1}{\gamma}} \phi^{\frac{\gamma-1}{\gamma}}}{E\left[(m\phi)^{\frac{\gamma-1}{\gamma}}\right]}. \tag{84}$$

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

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

Producers' first-order conditions are

$$m = \lambda \frac{\varepsilon^{\alpha}}{\theta^{1+\alpha}}. \tag{85}$$

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

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

In equilibrium, consumers own the firm, so their endowment equals the firm profit, $e=\varepsilon$, and consumption equals output, $c=\varepsilon$. This equality is an important limitation of this static analysis. In a dynamic model, equilibrium requires $c_t = y_t - i_t$.

We can find the equilibrium from the planning problem

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

The first-order condition is

$$\frac{c^{-\gamma}}{\phi^{1-\gamma}} = \lambda_p \frac{c^{\alpha}}{\theta^{1+\alpha}}. \tag{86}$$

Imposing the productivity choice constraint to eliminate the Lagrange multiplier λ_p, the full solution of the planning problem is[5]

$$\log c = -\frac{1}{1+\alpha} \log\left\{E\left[\left(\frac{\theta}{\phi}\right)^{\frac{(1+\alpha)(1-\gamma)}{\alpha+\gamma}}\right]\right\} + \frac{1+\alpha}{\alpha+\gamma}\log\theta + \frac{\gamma-1}{\alpha+\gamma}\log\phi. \tag{87}$$

[5] From (86),

$$c = \lambda_p^{-\frac{1}{\alpha+\gamma}} \left(\frac{\theta^{1+\alpha}}{\phi^{1-\gamma}}\right)^{\frac{1}{\alpha+\gamma}}$$

The constant is not interesting for us, however, so I suppress it below.

Using either discount factor formula, that is, (83) or (85), the equilibrium discount factor is

$$\log m = \text{const.} - \gamma \frac{1+\alpha}{\alpha+\gamma} \log\theta + \alpha \frac{\gamma-1}{\alpha+\gamma} \log\phi. \tag{88}$$

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,

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

Assuming normal distributions, the risk premium of the consumption claim is

$$\begin{aligned}\frac{E(R^e)}{R^f} = {} & \gamma\sigma^2 \left[\frac{1+\alpha}{\alpha+\gamma}\log\theta\right] - \alpha\sigma^2\left[\frac{\gamma-1}{\alpha+\gamma}\log\phi\right] \\ & + (\gamma-\alpha)cov\left[\frac{1+\alpha}{\alpha+\gamma}\log\theta, \frac{\gamma-1}{\alpha+\gamma}\log\phi\right]. \end{aligned} \tag{89}$$

6.2 Identification and measurement

In sum, consumer and producer first-order conditions are (83) and (85),

$$\log m = \text{const.} - \gamma \log c + (\gamma-1)\log\phi \tag{90}$$

$$\log m = \text{const.} + \alpha \log\varepsilon - (1+\alpha)\log\theta. \tag{91}$$

---

and

$$\left(\frac{c}{\theta}\right)^{1+\alpha} = \lambda_p^{-\frac{1+\alpha}{\alpha+\gamma}} \left(\frac{\theta}{\phi}\right)^{\frac{(1-\gamma)(1+\alpha)}{\alpha+\gamma}}.$$

Imposing the constraint.

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

Substituing out λ_p,

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

To clarify the formulas, let

$$\theta^* \equiv \frac{1+\alpha}{\alpha+\gamma}\log\theta;\ \phi^* \equiv \frac{(\gamma-1)}{\alpha+\gamma}\log\phi.$$

The equilibrium is then given by (87) and (88),

$$\log c = \log\varepsilon = \text{const.} + \theta^* + \phi^* \tag{92}$$

$$\log m = \text{const.} - \gamma\theta^* + \alpha\phi^*. \tag{93}$$

Equation (89) gives the equilibrium risk premium,

$$\frac{E(R^e)}{R^f} = \gamma\sigma^2(\theta^*) - \alpha\sigma^2(\phi^*) + (\gamma-\alpha)\sigma(\theta^*,\phi^*).$$

If we run a regression $\log m = \beta\log\varepsilon + u$, the coefficient is

$$\beta = \frac{\sigma(\log m, \log\varepsilon)}{\sigma^2(\log\varepsilon)} = -\gamma\frac{\sigma^2(\theta^*)}{\sigma^2(\theta^*+\phi^*)} + \alpha\frac{\sigma^2(\phi^*)}{\sigma^2(\theta^*+\phi^*)} + (\alpha-\gamma)\frac{\sigma(\theta^*,\phi^*)}{\sigma^2(\theta^*+\phi^*)}.$$

Now, what do we see? Suppose there are no preference shocks ϕ, and only by natural productivity shocks θ. Consumption and productivity rise with the shock, and the discount factor declines. The equity premium is positive. The data trace out $\log m = \text{const.} - \gamma\log\varepsilon$ with no error. The coefficient of the regression of $\log m$ on $\log\varepsilon$ is $-\gamma$. Data trace out the marginal rate of *substitution* curve and identify risk aversion γ, for any α.

How did we lose the production-based discount factor and α? The production-based discount factor formula (91) is still there. However, ε and θ are perfectly correlated in equilibrium. To use the production-based discount factor in this economy, we would have to account for the movement in θ correlated with productivity ε.

That insight offers 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 that variables such as productivity growth form useful discount factors, but with negative coefficients, not $\alpha > 0$. If underlying productivity shocks dominate, then although the discount factor has a positive and structural coefficient on productivity, in (91), an approximate discount factor that uses productivity but does not (somehow) control for the shock θ, will see a negative coefficient.

Likewise, if there truly were no preference shocks, then ε and θ would be perfectly correlated, and our formulas ignoring θ would perfectly measure the discount factor. The only trouble is that the estimated coefficient would have the wrong sign, relative to the prior that α should be positive. Or, for any α, we could use ε to measure θ.

Suppose instead there are preference shocks ϕ and no underlying technology shocks θ. For the realistic $\gamma > 1$ case, equilibrium consumption in rises with the preference shock, but the discount factor also rises with the preference shock, so the discount factor is high when consumption is high. The equity premium is negative. On the bright side, the production-based formula correctly measures the discount factor, with no correction at all for the underlying preference shock. Data trace out the marginal rate of *transformation* curve and identify production curvature α, for any value of risk aversion γ. The coefficient in the regression of $\log m$ on $\log c$ or $\log \varepsilon$ is α. But the positive equity premium, as well as common sense, suggests at least some underlying productivity shocks.

In reality, then, we likely see a mixture of preference and productivity shocks. The regression coefficient is a mongrel combination of α, γ, and shock variances and covariances.

To identify α, we need to find preference shocks that are orthogonal to the natural productivity shock, or we need to restrict or measure the natural productivity shocks. To construct a production-based discount factor we have somehow to control for or measure the natural productivity shock. Below, I review a clever restriction by Belo 2010 that measures θ, and I discuss identification and other ways to avoid θ problems.

This shock and parameter identification issue is not special to production-based asset pricing. It has important lessons for investor-based asset pricing as well, where in that term I include behavioral and institutional finance.

Traditional consumption-based models just assume away preference shocks. The empirical difficulties of the consumption-based model, and the imprecision, instability, and counterintuitive values of risk aversion γ it reports suggest that preference shocks may indeed be part of the story.

Preference shocks are 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. For example, typical models of the 2008 financial crisis start with a shock to the representative consumer's discount subjective discount factor β. Changing risk aversion is sometimes modeled as a preference shock. And behavioral finance is all about preference shocks. "Sentiment" or irrationally assessed probabilities are equivalent to preference shocks such as ϕ.

All these observations give us comfort that some preference Shocks exist, and that they can be used to identify productivity choice and allow us to see productivity choice respond to preference shocks.

But models with preference shocks will suffer exactly the same identification problem as a production-based model with dominant technology shocks. The shocks and the endogenous variables will be correlated. To identify γ here, we need to find a productivity shock that is orthogonal to preference shocks.

Moreover, the positive equity premium argues 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. Estimates of γ and coefficients of discount factors related to production variables may be unstable mongrels, but they are negative mongrels. This trouble ought to be particularly salient for behavioral finance and intermediary asset pricing, which explicitly posits that preference shocks are a central driving mechanism. As a concrete example, see Albuquerque, Eichenbaum, and Rebelo 2016 for preference shocks in a detailed macro-finance model, and Kruger 2019 for critique that it misses important moments.

This discussion also reminds us that while production and consumption-based asset pricing each exploit the wonderful GMM philosophy of examining one side of the market in isolation, identification requires us to think about general equilibrium and what causes variables to move.

6.3 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 general equilibrium. Figure 4 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 correctly models the equilibrium consumption process *as if* it were an endowment, 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 chosen

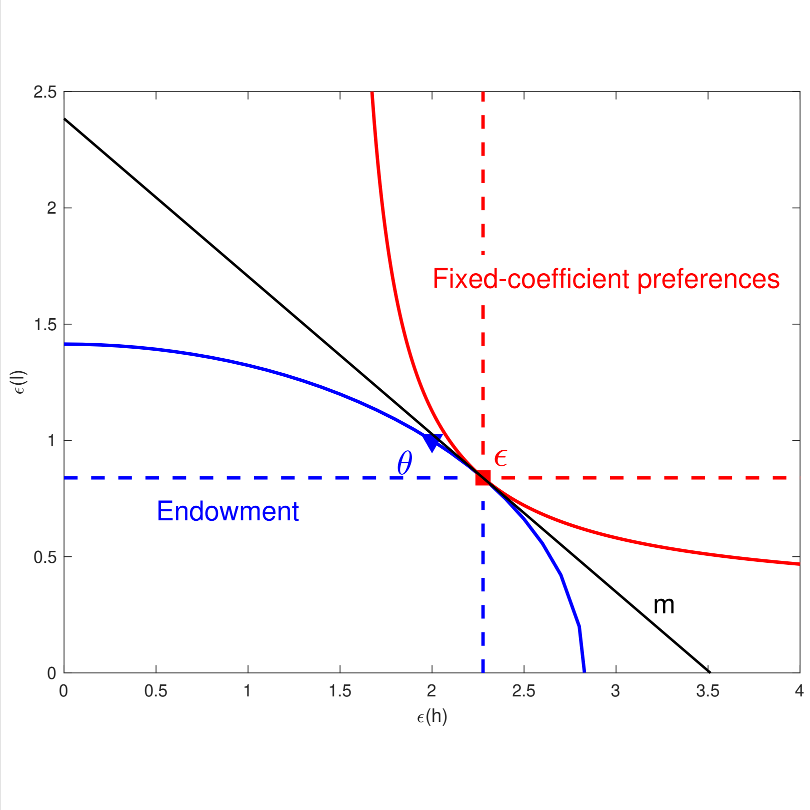

Figure 4
General equilibrium
The outward-bowed curve is a shock choice set $E\left[(\varepsilon/\theta)^{1+\alpha}\right]\leq 1$ with $\alpha=1$, $\theta=[2,1]$. The inward-bowed curve is an indifference curve for a power utility consumer with $u=(c/\theta)^{1-\gamma}$, $\phi=[5,1]$, $\gamma=2$. The dashed lines represent equivalent endowment economies, that is, fixed shocks ε or fixed-coefficient preferences, that deliver the same equilibrium quantities and prices. The symbol m denotes the stochastic discount factor or contingent claim price ratio.

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 4, one can model the production side as the northeast pointing box outlined by the dashed lines, keep the indifference curve, and maintain asset prices.

One can create an analogous production-based asset pricing model. Again, measure or model the consumption or productivity process $\varepsilon^*=\varepsilon$. Leave the production set alone, keeping the same productivity shocks $\theta^* = \theta$ and curvature $\alpha^* = \alpha$. Marry this production process to fixed-coefficient *preferences*. In place of the smooth utility function and preference shocks, let

$$u[c(h),c(l)] = \min\left[\frac{c(h)}{\varepsilon(h)},\frac{c(l)}{\varepsilon(l)}\right]. \tag{94}$$

Then, measure contingent claim prices or the discount factor from the marginal rate of *transformation* alone. This new economy has the same asset prices and quantity implications as the full general equilibrium but spares the researcher from having to model and measure the entire consumption and intermediation side of the economy.

Fixed-coefficient preferences (94) 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. In this way, we can mirror the brilliant simplification that Lucas 1978 brought to consumption-based asset pricing or construct insightful simulation economies in the style of Mehra and Prescott 1985.

7. A Simple Aggregation Model

The main philosophy of 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 total output to weather by varying the amount planted in each of the two fields. 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

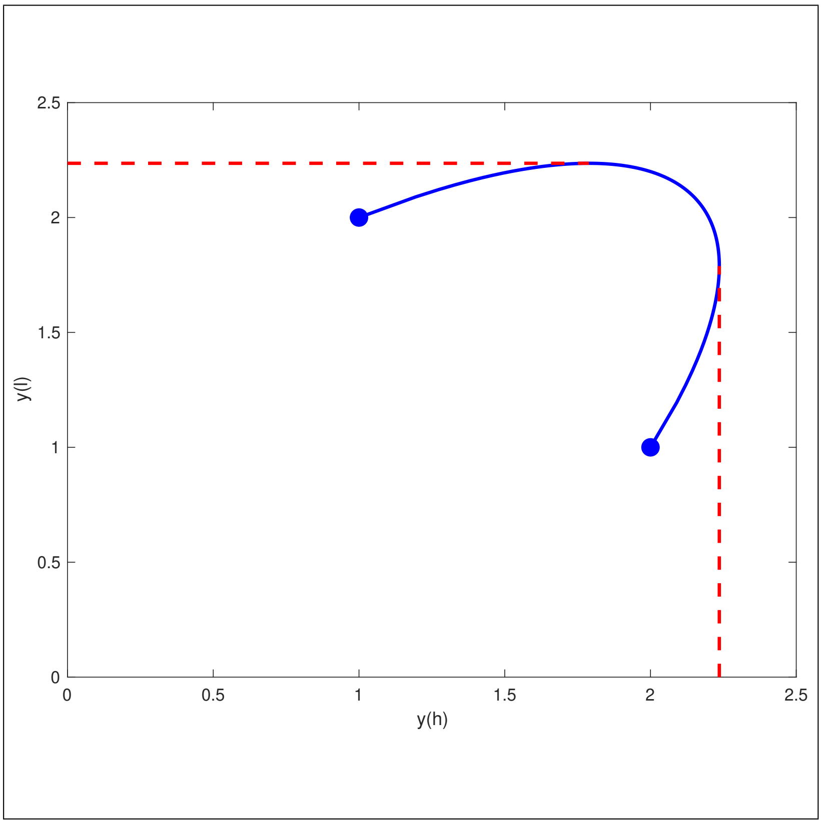

Figure 5
Two-state aggregation
Aggregate production set $\{y(h),y(l)\}$ induced by two technologies, $y(s)=y_1(s)+y_2(s)$; $y_i(s)\leq\theta_i(s)k_i^{0.5}$; $i=1,2$; $s=h,l$; $k_1+k_2=1$.

alone, $y(0)=W-k$ and $y(s)$.) Figure 5 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 represented by dashed lines.

For this construction to work, that is, for the marginal rate of transformation $\partial y(h)/\partial y(l)$ to exist, so we can equate it to contingent claims price ratios $\partial y(h)/\partial y(l)=p(h)/p(l)=\pi(h)m(h)/[\pi(l)m(l)]$ 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 nonsingular. 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 5 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

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

Alternatively, and perhaps more elegantly, 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 aggregation theory or spanning conditions, as the major point of this paper is to write down smooth technologies directly, just as we write down aggregate technologies $y = f(k,n,...)$ that are smooth across inputs rather than derive them from deeper fundamentals. One can see from this discussion where such an aggregation theory would go.

With this basic idea, one can see many potential microfoundations for active trade-offs across states. The firm could invest in capital or R&D to shift its output across states. The firm could buy solar cells or multifuel engines, for example, in order to 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 variables, 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.

8. 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. 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.

This paper is a revision of the first part of Cochrane 1993. That paper introduced the idea of choosing productivity from a set $E\left[(\varepsilon/\theta)^{1+\alpha}\right]$. It sat a long time, as I hoped to complete an empirical counterpart, compute general equilibrium examples, and cleanly solve the θ shock identification and measurement question among other needed improvements. This paper includes a much-improved dynamic extension, but does not achieve those other goals. Bringing these sorts of models to data, or constructing simulation models that may be compared to data, remains an important project, with numerous measurement, specification, and identification issues to face.

Belo 2010 is the first to use this production technology with productivity choice empirically. Belo proposes a clever approach to the identification problem, which could (and should) be generalized to larger and more disaggregated groups of investment returns. Discount factor formulas such as $m^* = \lambda(\varepsilon_{i,t+1}/\varepsilon_{i,t})^{\alpha}(\theta_{i,t+1}/\theta_{i,t})^{-(1+\alpha)}$ hold separately for each firm or industry, just as $m_{t+1} = \beta(c_{i,t+1}/c_{i,t})^{-\gamma}$ holds separately for each individual i. Taking logs,

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

separately for each technology i. (Belo uses α where I use $1+\alpha$.) Belo then assumes that multiple technologies have a factor structure,

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

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

$$\log\left(m^*_{t+1}\right) = \alpha\log\left(\frac{\varepsilon_{1,t+1}}{\varepsilon_{1,t}}\right) - \lambda_1 F_{t+1} \tag{95}$$

$$\log\left(m^*_{t+1}\right) = \alpha\log\left(\frac{\varepsilon_{2,t+1}}{\varepsilon_{2,t}}\right) - \lambda_2 F_{t+1}. \tag{96}$$

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

$$\log\left(m^*_{t+1}\right) = \frac{\alpha}{\lambda_1 - \lambda_2}\left[\lambda_1\log\left(\frac{\varepsilon_{2,t+1}}{\varepsilon_{2,t}}\right) - \lambda_2\log\left(\frac{\varepsilon_{1,t+1}}{\varepsilon_{1,t}}\right)\right]. \tag{97}$$

We observe $\log(\varepsilon_i) = \log(y_i) - \log f(k_i)$. 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 observed productivity shocks reveals the natural productivity shock. Or, solving (95) and (96) for the shocks ε, those 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 is more compelling with more technologies. Across J technologies with productivity ε_j, there are J sources of unobserved movement θ_j and one additional source of movement m. Reducing the dimensionality of the θ_j by only one via a factor structure assumption, we can identify m. To generalize, we need a $J-1$ factor structure of J technology shocks, not a single factor structure. (Belo's online appendix C pursues a $J=3$ factor model.) The essence of business cycles is common movement, and stock market returns display a strong factor structure, so the idea that multiple firm's productivities or other variables follow a reduced factor structure is natural.

Since Belo assumes $y_t = \varepsilon_t f(k_t)$ with k_t predetermined, he uses y_t in place of ε_t in (97). 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, and many related *ad hoc* macro-finance pricing models that use discount factors tied to macroeconomic variables to explain cross-sectional variation in expected returns. 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. That it is similar to existing successful *ad hoc* models says it is robust. Belo also adds a relative price of output and investment goods, which adds a second set of factors, and prices an up-to-date set of asset returns.

Jermann 2013 uses 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 implements the model of Section 7. He creates 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.

8.1 Investment returns

Standard technologies do not allow a general marginal rate of transformation. Firm first-order conditions do, however, give rise to investment returns. Production-based asset pricing has to date largely linked macroeconomics to asset pricing via investment returns.

As outlined in Section 5, producer first-order conditions give rise to a physical return R^I, measurable from investment, capital, output, and labor decisions (55). Any asset returns that can be determined by arbitrage with the investment returns should be so priced. Equivalently, the investment return should be *priced by* any discount factor, $1=E(mR^I)$. When marginal q equals average q, the firm's stock (or stock and bond) return should equal the investment return, ex post and ex ante, a particularly clear instance of this arbitrage pricing result.

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" (capital asset pricing model). As I have emphasized, this paper generalizes investment-based asset pricing, keeping its central prediction $1=E(mR^I)$.

This effort was 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 forecasted by the dividend yield, term spread, investment-to- capital ratio, and other variables matches variation in expected investment returns well. Lamont 2000 shows that measures of investment plans offer even better correlations. When stock prices rise, time is required 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 the same as a 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, e.g.); it often uses cash flow forecasts and other detailed measurements beyond investment and stock prices.

It focuses on failure: theory predicts a 100% R^2, that is, 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 cannot explain, in order to advance a cash flow 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 risk-free 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 more general production-based asset pricing described in this paper. One must make hundreds of implementation decisions. Formal rejections of specific implementations will be easy. Figuring out where theory is most useful will be difficult.

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 a great deal of progress by this approach. 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. Create a portfolio of value firms, and their investment returns should be higher than those of a portfolio of growth firms, matching the average stock returns of those 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 in this way. (The iceberg of which this survey is a tip includes Lyandres, Sun, and Zhang 2008 Li, Livdan, and Zhang 2009, Liu and Zhang 2014, Wu, Zhang, and Zhang 2010, Li and Zhang 2010, Liu, Xiaolei, Whited, and Zhang 2009, and Goncalves, Xue and Zhang 2019. Lots of anomalies and measurement issues must be worked out!)

Whether one can say this approach "explains" the anomalies and if so "rationally" is a contentious question. This literature documents that firms adjust properly in response to expected returns, so investment decisions and expected returns are connected as economics says they should be. There is no arbitrage between investment returns and stock returns. But both investment and stock returns are endogenous variables. Both could be driven by fads and irrationalities on the part of consumers. Still, if expected returns lined up with market, consumption, or factor betas, one could make the same objection to the word "explain," as returns, consumption and its betas are also endogenous variables which might be driven by irrational behavior on the part of producers. So, one can say that the investment CAPM "explains" risk premiums as well as a standard CAPM and consumption CAPM would do, if those models were successful in lining average returns up against covariances with the market return or consumption growth.

I also think the word "investment CAPM" is a bit misleading. The word "CAPM" suggests that expected returns line up with covariances of returns with some variable, 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 in isolation. It just compares a wide range of investment returns to the corresponding wide range of asset returns, in anomaly-sorted portfolios. By contrast, the production-based approach in this paper does offer a "CAPM" representation. But he or she who does the work gets to baptize the results, so just understand how the fundamental structure of an "investment CAPM" remains different from that of a market portfolio CAPM, consumption CAPM, or production-based model, such as this one.

We still desire a general purpose model, a model 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,

$$m = a + b_r R^{I,r}_{t+1} + b_{nr} R^{I,nr}_{t+1}, \tag{98}$$

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 to 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-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 an approximate discount factor of the form (98) for asset returns highly correlated with combinations of the two investment returns. (See Cochrane 2005, chap. 9.4.) Alternatively, Cochrane 1996 (p. 577) speculates, if the investment returns span the investment opportunity set then consumption and marginal utility must be driven by the two investment returns:

> 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 processes, which are the only nondiversifiable payoffs in the economy and add up to aggregate wealth, drive marginal utility growth and hence price assets...

Zhang, Jones and Tüzel 2013, İmrohoroğlu and Tüzel 2014, Belo and Lin 2012 and Belo and Yu 2013 follow a similar approach. Using this logic, they estimate or simulate "production-based" models with discount factors

$$\log m^*_{t+1} = \text{constant} - \gamma_t \varepsilon_{t+1}, \tag{99}$$

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 use consumption-based asset pricing to extend the discount factor from a single investment return to multiple returns. For example, Zhang 2005, (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 $b > 0$.

By contrast, the approach in this paper offers a truly production-based view of where *ad hoc* macro-factor or investment-return models, such as (99) and its generalizations, originate. This paper's approach requires no assumptions about preferences, not even a Sharpe ratio limit, other than the existence of a discount factor or a set of contingent claims prices.

Likewise, we have seen here production-based discount factors with output growth, wages, labor input, labor share growth, and growth in the labor/output ratio along with productivity as pricing factors. These results can provide 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 minus the labor share in the discount factor formula (65), prices a cross-section of returns.

However, theory does not just exist to justify existing *ad hoc* models. This paper links asset prices to production data in a fundamentally different way. New theory ought to inspire new empirical specifications or at least restrict and refine them.

The word "production-based" is also sometimes used to mean "general equilibrium models that include production." An important literature writes models with (interesting and elaborate) preferences, along with detailed (interesting and elaborate) production technologies, and sometimes market frictions as well, calibrated to match asset pricing facts. This general equilibrium literature tackles essential questions, such as: What features of production technology create "growth" and "value" firms in the first place? Where do betas come from? (Croce 2014 is an example that uses the "production-based" label.) As with *ad hoc* models linking discount factors to production data, or models that infer consumption from production data, this literature offers a different meaning than my sense of the word "production-based," that uses only producer first-order conditions.

9. Concluding Comments

This paper is clearly an exploratory step. Lots must be done to create production-based asset pricing models that 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, not just productivity and its underlying shock. More detailed production functions may well change that form. A better handling of dynamics and how firms can slowly change their shock exposures beckons.

I focused on the discount factor question, paralleling consumption-based asset pricing. General equilibrium models, in which one fully solves the productivity choice given external variables, beckon. Such models will likely find it useful to exploit stationarity assumptions, the fact that the same shocks are in some sense repeated. For example, one often starts a general equilibrium model by positing that all uncertainty evolves as a vector-valued Markov process, and looks for solutions as a function of that state variable.

Bringing this production-based approach to data requires many choices. The first is identifying or measuring the underlying productivity shocks θ, or finding 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.

It is possible that this problem is ameliorated with a more detailed production process, and careful measurement of productivity, along with recognition of multiple factors in asset returns and macroeconomic variables, as sketched above. Moreover, discount factors from realistic production functions, including labor, adjustment costs, and other inputs, feature a range of variables that all respond to the same underlying natural productivity θ shocks and should help to identify them. Simply assuming that the natural productivity shock θ is perfectly correlated with actual productivity, excusing negative estimated α, may be enough.

But really this identification problem is no different or worse than the similar identification issues that haunt all of macroeconomics and finance. The example of perfectly correlated natural and chosen productivity 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-side variable (inflation) perfectly correlated with its (monetary policy) shock, so yields exactly the wrong coefficient. VARs are plagued by the question of whether interest rates cause inflation or whether expected inflation causes interest rates. Yet new Keynesian models and VARs are a thriving industry. How? By thinking hard, making identification assumptions, and 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, equations that link variables with no error. Yet a shock in any equation usually means that the equation cannot be directly estimated. Instead, we need a shock somewhere else to do that and an exclusion restriction. Shocks have to be somewhere, and, if we are honest, most likely everywhere. Medium-scale empirical macro models contain shocks in every equation. Increasingly popular preference shocks (risk aversion, discount factor, financial frictions) or their observational equivalents (taste, sentiment, probability) raise exactly the same identification problem for conventional asset pricing. The Belo 2010 factor model approach is a great example of how light and plausible (and clever!) identification assumptions can go a long way.

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, does not identify parameters at all. Other assumptions may produce the same moments. But this process is how we construct such models. Obtaining a model that can match the data is difficult 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 great progress in 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 investment returns to asset returns, we want to use as many investment returns as possible. In the investment return approach, such as Zhang 2017, each firm's investment return is primitive, however. Surely one looks for something more integrative than 3,000 separate investment returns to explain 3,000 stock returns. They likely share a statistical factor structure, but that only ties them together as an empirical observation.

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 should load on all of them, $m = a + b_1 R^I_{1,t+1} + b_2 R^I_{2,t+1} + ... + b_i R^I_{i,t+1}$. Only a second empirical observation, that investment returns obey a factor structure, results in the APT philosophy of a smaller number of pricing factors. 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, 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.

But investment returns and productivity choice are complements as they are parts of the same model, not competitors. 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 7 already points to interesting productivity choice in the aggregate production function that may not exist in firm-level production. Extending that idea to multiple technologies that also have productivity choices should lead to additional insights.

Clearly, the investigation has just begun.

References

Albuquerque, R., M. Eichenbaum, V. X. Luo, and S. Rebelo. 2016. Valuation risk and asset pricing. *Journal of Finance* 71:2861–904.

Belo, Frederico. 2006. A pure production-based asset pricing model. PhD Dissertation, University of Chicago Booth School of Business.

———. 2010. Production-based measures of risk for asset pricing. *Journal of Monetary Economics* 57:146–63.

Belo, F., and X. Lin. 2012. The inventory growth spread. *Review of Financial Studies* 25:278–313.

Belo, F., X. Lin, and S. Bazdresch. 2014. Labor hiring, investment, and stock return predictability in the cross section. *Journal of Political Economy* 122:129–77.

Belo, F., and J. Yu. 2013. Government investment and the stock market. *Journal of Monetary Economics* 60:325–39.

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

Cochrane, J. H. 1988. Production-based asset pricing. Working Paper, Hoover Institution.

———. 1991. Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance* 46:209–37.

———. 1993. Rethinking production under uncertainty. Manuscript, University of Chicago.

———. 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, NJ: Princeton University Press.

———. 2011. Determinacy and identification with Taylor rules. *Journal of Political Economy* 119:565–615.

———. 2017. Macro-finance. *Review of Finance* 21:945–85.

Cochrane, J. H., and L. P. Hansen. 1992. Asset pricing explorations for macroeconomics. *NBER Macroeconomics Annual* 7:115–65.

Constantinides, G. M., and D. Duffie. 1996. Asset pricing with heterogeneous consumers. *Journal of Political Economy* 104:219–40.

Croce, M. M. 2014. Long-run productivity risk: A new hope for production-based asset pricing? *Journal of Monetary Economics* 66:13–31.

Debreu, G. 1959. *The theory of value: an axiomatic analysis of economic equilibrium.* New York: Wiley.

Epstein, L. G., and S. E. Zin. 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica* 57:937–69.

Goncalves, A. S., C. Xue, and L. Zhang. 2020. Aggregation, capital heterogeneity, and the investment CAPM. *Review of Financial Studies* 33:2728–71.

Grigsby, J. 2020. Skill heterogeneity and aggregate labor market dynamics. Manuscript, University of Chicago.

Hansen, L. P., and R. Jagannathan. 1991. Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99:225–62.

Hindy, Ayman and Chi-fu Huang. 1992. Intertemporal preferences for uncertain consumption: A continuous time approach. *Econometrica* 60:781–801.

Houthakker, H. S. 1955. The Pareto distribution and the Cobb-Douglas production function in activity analysis 1. *Review of Economic Studies* 23:27–31.

İmrohoroğlu, A., and Ş. Tüzel. 2014. Firm-level productivity, risk, and return. *Management Science* 60:2073–90.

Jagannathan, R., and Z. Wang. 1996. The conditional CAPM and the cross-section of expected returns. *Journal of Finance* 51:3–53.

Jermann, U. J. 2013. A production-based model for the term structure. *Journal of Financial Economics* 109:293–306.

Jones, C. S., and Ş. Tüzel. 2013. Inventory investment and the cost of capital. *Journal of Financial Economics* 107:557–79.

King, R. G., C. I. Plosser, and S. T. Rebelo. 1988. Production, growth and business cycles: I. The basic neoclassical model. *Journal of Monetary Economics* 21:195–232.

Kruger, S. 2019. Assessing valuation risk: Theory and empirical evidence. Working Paper, University of Texas at Austin.

Kydland, F. E., and E. C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50:1345–70.

Lamont, O. A. 2000. Investment plans and stock returns. *Journal of Finance* 55:2719–45.

Lettau, M., S. Ludvigson, and S. Ma. 2019. Capital share risk in U.S. asset pricing. *Journal of Finance* 74:1753–92.

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

Li, E. X. N., D. Livdan, and L. Zhang. 2009. Anomalies. *Review of Financial Studies* 22:4301–34.

Li, Q., M. Vassalou, and Y. Xing. 2006. Sector investment growth rates and the cross section of equity returns. *Journal of Business* 79:1637–65.

Lin, X., and L Zhang. 2013. The investment manifesto. *Journal of Monetary Economics* 60:351–66.

Liu, L. X., T. M. Whited, and L. Zhang. 2009. Investment-based expected stock returns. *Journal of Political Economy* 117:1105–39.

Liu, L. X., and L. Zhang. 2014. A neoclassical interpretation of momentum. *Journal of Monetary Economics* 67:109–28.

Lucas, R. E., Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46:1429–45.

Lyandres, E., L. Sun, and L. Zhang. 2008. The new issues puzzle: Testing the investment-based explanation. *Review of Financial Studies* 21:2825–55.

Mehra, R., and E. C. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15:145–61.

Wu, J. (G.), L. Zhang, and X. F. Zhang. 2010. The q-theory approach to understanding the accrual anomaly. *Journal of Accounting Research* 48:177–223.

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

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