Economic Quarterly—Volume 97, Number 3—Third Quarter 2011—Pages 255–326

Macroeconomics with Heterogeneity: A Practical Guide

Fatih Guvenen

> What is the origin of inequality among men and is it authorized by natural law?
>
> —Academy of Dijon, 1754 (Theme for essay competition)

The quest for the origins of inequality has kept philosophers and scientists occupied for centuries. A central question of interest—also highlighted in Academy of Dijon's solicitation for its essay competition[1]—is whether inequality is determined solely through a natural process or through the interaction of innate differences with man-made institutions and policies. And, if it is the latter, what is the precise relationship between these origins and socioeconomic policies?

While many interesting ideas and hypotheses have been put forward over time, the main impediment to progress came from the difficulty of scientifically testing these hypotheses, which would allow researchers to refine ideas that were deemed promising and discard those that were not. Economists, who grapple with the same questions today, have three important advantages that can allow us to make progress. First, modern quantitative economics provides a wide set of powerful tools, which allow researchers to build "laboratories" in which various hypotheses regarding the origins and consequences of

For helpful discussions, the author thanks Dean Corbae, Cristina De Nardi, Per Krusell, Serdar Ozkan, and Tony Smith. Special thanks to Andreas Hornstein and Kartik Athreya for detailed comments on the draft. David Wiczer and Cloe Ortiz de Mendivil provided excellent research assistance. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Chicago, the Federal Reserve Bank of Richmond, or the Federal Reserve System. Guvenen is affiliated with the University of Minnesota, the Federal Reserve Bank of Chicago, and NBER. E-mail: guvenen@umn.edu.

[1] The competition generated broad interest from scholars of the time, including Jean-Jacques Rousseau, who wrote his famous *Discourse on the Origins of Inequality* in response, but failed to win the top prize.

inequality can be studied. Second, the widespread availability of rich micro-data sources—from cross-sectional surveys to panel data sets from administrative records that contain millions of observations—provides fresh input into these laboratories. Third, thanks to Moore's law, the cost of computation has fallen radically in the past decades, making it feasible to numerically solve, simulate, and estimate complex models with rich heterogeneity on a typical desktop workstation available to most economists.

There are two broad sets of economic questions for which economists might want to model heterogeneity. First, and most obviously, these models allow us to study cross-sectional, or distributional, phenomena. The U.S. economy today provides ample motivation for studying distributional issues, with the top 1 percent of households owning almost half of all stocks and one-third of all net worth in the United States, and wage inequality having risen virtually without interruption for the last 40 years. Not surprisingly, many questions of current policy debate are inherently about their distributional consequences. For example, heated disagreements about major budget issues—such as reforming Medicare, Medicaid, and the Social Security system—often revolve around the redistributional effects of such changes. Similarly, a crucial aspect of the current debate on taxation is about "who should pay what?" Answering these questions would begin with a sound understanding of the fundamental determinants of different types of inequality.

A second set of questions for which heterogeneity could matter involves aggregate phenomena. This second use of heterogeneous-agent models is less obvious than the first, because various aggregation theorems as well as numerical results (e.g., Ríos-Rull [1996] and Krusell and Smith [1998]) have established that certain types of heterogeneity do not change (many) implications relative to a representative-agent model.[2]

To understand this result and its ramifications, in Section 1, I start by reviewing some key theoretical results on aggregation (Rubinstein 1974; Constantinides 1982). Our interest in these theorems comes from a practical concern: Basically, a subset of the conditions required by these theorems are often satisfied in heterogeneous-agent models, making the aggregate implications of such models closely mimic those from a representative-agent economy. For example, an important theorem proved by Constantinides (1982) establishes the existence of a representative agent if markets are complete.[3] This central role of complete markets turned the spotlight since the late 1980s onto its testable implications for perfect risk sharing (or "full insurance"). As

[2] These aggregation results do *not* imply that all aspects of a representative-agent model will be the same as those of the underlying individual problem. I discuss important examples to the contrary in Section 6.

[3] (Financial) markets are "complete" when agents have access to a sufficiently rich set of assets that allows them to transfer their wealth/resources across any two dates and/or states of the world.

I review in Section 2, these implications have been tested by an extensive literature using data sets from all around the world—from developed countries such as the United States to village economies in India, Thailand, Uganda, and so on. While this literature delivered a clear *statistical* rejection, it also revealed a surprising amount of "partial" insurance, in the sense that individual consumption growth (or, more generally, marginal utility growth) does not seem to respond to many seemingly large shocks, such as long spells of unemployment, strikes, and involuntary moves (Cochrane [1991] and Townsend [1994], among others).

This raises the more practical question of "how *far* are we from the complete markets benchmark?" To answer this question, researchers have recently turned to directly measuring the degree of partial insurance, defined for our purposes as the degree of consumption smoothing over and above what an individual can achieve on her own via "self-insurance" in a permanent income model (i.e., using a single risk-free asset for borrowing and saving). Although this literature is quite new—and so a definitive answer is still not on hand—it is likely to remain an active area of research in the coming years.

The empirical rejection of the complete markets hypothesis launched an enormous literature on incomplete markets models starting in the early 1990s, which I discuss in Section 3. Starting with Imrohoroglu (1989), Huggett (1993), and Aiyagari (1994), this literature has been addressing issues from a very broad spectrum, covering diverse topics such as the equity premium and other puzzles in finance; important life-cycle choices, such as education, marriage/divorce, housing purchases, fertility choice, etc.; aggregate and distributional effects of a variety of policies ranging from capital and labor income taxation to the overhaul of Social Security, reforming the health care system, among many others. An especially important set of applications concerns trends in wealth, consumption, and earnings inequality. These are discussed in Section 4.

A critical prerequisite for these analyses is the disentangling of "ex ante heterogeneity" from "risk/uncertainty" (also called ex post heterogeneity)—two sides of the same coin, with potentially very different implications for policy and welfare. But this is a challenging task, because inequality often arises from a mixture of heterogeneity and idiosyncratic risk, making the two difficult to disentangle. It requires researchers to carefully combine cross-sectional information with sufficiently long time-series data for analysis. The state-of-the-art methods used in this field increasingly blend the set of tools developed and used by quantitative macroeconomists with those used by structural econometricians. Despite the application of these sophisticated tools, there remains significant uncertainty in the profession regarding the magnitudes of idiosyncratic risks as well as whether or not these risks have increased since the 1970s.

The Imrohoroglu-Huggett-Aiyagari framework sidestepped a difficult issue raised by the lack of aggregation—that aggregates, including prices, depend on the entire wealth distribution. This was accomplished by abstracting from aggregate shocks, which allowed them to focus on stationary equilibria in which prices (the interest rate and the average wage) were simply some constants to be solved for in equilibrium. A far more challenging problem with incomplete markets arises in the presence of aggregate shocks, in which case equilibrium prices become *functions* of the entire wealth distribution, which varies with the aggregate state. Individuals need to know these equilibrium functions so that they can forecast how prices will evolve in the future as the aggregate state evolves in a stochastic manner. Because the wealth distribution is an infinite-dimensional object, an exact solution is typically not feasible. Krusell and Smith (1998) proposed a solution whereby one approximates the wealth distribution with a finite number of its moments (inspired by the idea that a given probability distribution can be represented by its moment-generating function). In a remarkable finding, they showed that the first moment (the mean) of the wealth distribution was all individuals needed to track in this economy for predicting all future prices. This result—generally known as "approximate aggregation"—is a double-edged sword. On the one hand, it makes feasible the solution of a wide range of interesting models with incomplete markets and aggregate shocks. On the other hand, it suggests that ex post heterogeneity does not often generate aggregate implications much different from a representative-agent model. So, the hope that some aggregate phenomena that were puzzling in representative-agent models could be explained in an incomplete markets framework is weakened with this result. While this is an important finding, there are many examples where heterogeneity *does* affect aggregates in a significant way. I discuss a variety of such examples in Section 6.

Finally, I turn to computation and calibration. First, in Section 5, I discuss some details of the Krusell-Smith method. A number of potential pitfalls are discussed and alternative checks of accuracy are studied. Second, an important practical issue that arises with calibrating/estimating large and complex quantitative models is the following. The objective function that we minimize often has lots of jaggedness, small jumps, and/or deep ridges because of a variety of reasons that have to do with approximations, interpolations, binding constraints, etc. Thus, local optimization methods are typically of little help on their own, because they very often get stuck in some local minima. In Section 7, I describe a global optimization algorithm that is simple yet powerful and is fully parallelizable without requiring any knowledge of MPI, OpenMP, and so on. It works on any number of computers that are connected to the Internet and have access to a synchronization service like DropBox. I provide a discussion of ways to customize this algorithm with different options to experiment.

1. AGGREGATION

Even in a simple static model with no uncertainty we need a way to deal with consumer heterogeneity. Adding dynamics and risk into this environment makes things more complex and requires a different set of conditions to be imposed. In this section, I will review some key theoretical results on various forms of aggregation. I begin with a very simple framework and build up to a fully dynamic model with idiosyncratic (i.e., individual-specific) risk and discuss what types of aggregation results one can hope to get and under what conditions.

Our interest in aggregation is not mainly for theoretical reasons. As we shall see, some of the conditions required for aggregation are satisfied (sometimes inadvertently!) by commonly used heterogeneous-agent frameworks, making them behave very much like a representative-agent model. Although this often makes the model easier to solve numerically, at the same time it can make its implications "boring"—i.e., too similar to a representative-agent model. Thus, learning about the assumptions underlying the aggregation theorems can allow model builders to choose the features of their models carefully so as to avoid such outcomes.

A Static Economy

Consider a finite set $\mathcal{I}$ (with cardinality I) of consumers who differ in their preferences (over l types of goods) and wealth in a static environment. Consider a particular good and let $x_i(p, w_i)$ denote the demand function of consumer i for this good, given prices $p \in R^l$ and wealth w_i. Let $(w_1, w_2, ..., w_I)$ be the vector of wealth levels for all I consumers. "Aggregate demand" in this economy can be written as

$$x\,(p, w_1, w_2, ..., w_I) = \sum_{i=1}^{I} x_i(p, w_i).$$

As seen here, the aggregate demand function x depends on the entire wealth distribution, which is a formidable object to deal with. The key question then is, when can we write $x(p, w_1, w_2, ..., w_n) \equiv x(p, \sum w_i)$? For the wealth distribution to not matter, we need aggregate demand to not change for any redistribution of wealth that keeps aggregate wealth constant ($\sum dw_i = 0$). Taking the total derivative of x, and setting it to zero yields

$$\frac{\partial x\left(p, \sum w_i\right)}{\partial w_i} = 0 \Rightarrow \sum_{i=1}^{n} \frac{\partial x_i\,(p, w_i)}{\partial w_i} dw_i = 0$$

for all possible redistributions. This will only be true if

$$\frac{\partial x_i\,(p, w_i)}{\partial w_i} = \frac{\partial x_j\,(p, w_j)}{\partial w_j} \quad \forall i, j \in \mathcal{I}.$$

Thus, the key condition for aggregation is that individuals have the same marginal propensity to consume (MPC) out of wealth (or linear Engel curves). In one of the earliest works on aggregation, Gorman (1961) formalized this idea via restrictions on consumers' indirect utility function, which delivers the required linearity in Engel curves.

Theorem 1 (Gorman 1961) *Consider an economy with $N < \infty$ commodities and a set $\mathcal{I}$ of consumers. Suppose that the preferences of each consumer $i \in I$ can be represented by an indirect utility function*[4] *of the form*

$$v_i\,(p, w_i) = a_i\,(p) + b\,(p)\,w_i,$$

and that each household $i \in \mathcal{I}$ has a positive demand for each commodity, then these preferences can be aggregated and represented by those of a representative household, with indirect utility

$$v\,(p, w) = a\,(p) + b\,(p)\,w,$$

where $a(p) = \sum_i a_i(p)$ and $w = \sum_i w_i$ is aggregate income.

As we shall see later, the importance of linear Engel curves (or constant MPCs) for aggregation is a key insight that carries over to much more general models, all the way up to the infinite-horizon incomplete markets model with aggregate shocks studied in Krusell and Smith (1998).

A Dynamic Economy (*No* Idiosyncratic Risk)

Rubinstein (1974) extends Gorman's result to a dynamic economy where individuals consume out of wealth (no income stream). Linear Engel curves are again central in this context.

Consider a frictionless economy in which each individual solves an intertemporal consumption-savings/portfolio allocation problem. That is, every period current wealth w_t is apportioned between current consumption c_t and a portfolio of a risk-free and a risky security with respective (gross) returns R_t^f and R_t^s.[5] Let α_t denote the portfolio share of the risk-free asset at time t, and δ denote the subjective time discount factor. Individuals solve

[4] Denoting the consumer's utility function over goods with U, the indirect utility function is simply $v_i(p, w_i) \equiv U(x_i(p, w_i))$—that is, the maximum utility of a consumer who has wealth w_i and faces price vector p.

[5] We can easily allow for multiple risky securities at the expense of complicating the notation.

$$\max_{\{c_t,\alpha_t\}} E\left(\sum_{t=1}^{T}\delta^t U\left(c_t\right)\right)$$
$$\text{s.t. } w_{t+1} = \left(w_t - c_t\right)\left(\alpha_t R_t^f + \left(1-\alpha_t\right) R_t^s\right).$$

Furthermore, assume that the period utility function, U, belongs to the hyperbolic absolute risk aversion (HARA) class, which is defined as utility functions that have linear risk tolerance: $T(c) \equiv -U(c)'/U(c)'' = \rho + \gamma c$ and $\gamma < 1$.[6] This class encompasses three utility functions that are well-known in economics: $U(c) = (\gamma-1)^{-1}(\rho+\gamma c)^{1-\gamma^{-1}}$ (generalized power utility; standard constant relative risk aversion [CRRA] form when $\rho \equiv 0$); $U(c) = -\rho \times \exp(-c/\rho)$ if $\gamma \equiv 0$ (exponential utility); and $U(c) = 0.5(\rho - c)^2$ defined for values $c < \rho$ (quadratic utility).

The following theorem gives six sets of conditions under which aggregation obtains.[7]

Theorem 2 (Rubenstein 1974) *Consider the following homogeneity conditions:*

1. All individuals have the same resources w_0, and tastes δ and U.

2. All individuals have the same δ and taste parameters $\gamma \neq 0$.

3. All individuals have the same taste parameters $\gamma = 0$.

4. All individuals have the same resources w_0 and taste parameters $\rho = 0$ and $\gamma = 1$.

5. A complete market exists and all individuals have the same taste parameter $\gamma = 0$.

6. A complete market exists and all individuals have the same resources w_0 and taste δ, $\rho = 0$, and $\gamma = 1$.

Then, all equilibrium rates of return are determined in case (1) as if there exist only composite individuals each with resources w_0 and tastes δ and U; and equilibrium rates of return are determined in cases (2)–(6) as if there exist only composite individuals each with the following economic characteristics: (i) resources: $w_0 = \sum w_0^i/I$; (ii) tastes: $\sigma = \Pi(\sigma^i)^{(\rho_i/\sum\rho_i)}$ (where $\sigma \equiv 1/\delta - 1$) or $\delta = \sum \delta^i/I$; and (iv) preference parameters: $\rho = \sum \rho_i/I$, and γ.

Several remarks are in order.

---

[6] "Risk tolerance" is the reciprocal of the Arrow-Pratt measure of "absolute risk aversion," which measures consumers' willingness to bear a fixed amount of consumption risk. See, e.g., Pratt (1964).

[7] The language of Theorem 2 differs from Rubinstein's original statement by assuming rational expectations and combines results with the extension to a multiperiod setting in his footnote 5.

Demand Aggregation

An important corollary to this theorem is that whenever a composite consumer can be constructed, in equilibrium, rates of return are insensitive to the distribution of resources among individuals. This is because the aggregate demand functions (for both consumption and assets) depend only on total wealth and not on its distribution. Thus, we have "demand aggregation."

Aggregation and Heterogeneity in Relative Risk Aversion

Notice that all six cases that give rise to demand aggregation in the theorem require individuals to have the same curvature parameter, γ. To see why this is important, note that (with HARA preferences) the optimal holdings of the risky asset are a linear function of the consumer's wealth: $\kappa_1 + \kappa_2 w_t/\gamma$, where κ_1 and κ_2 are some constants that depend on the properties of returns. It is easy to see that with identical slopes, $\frac{\kappa_2}{\gamma}$, it does not matter who holds the wealth. In other words, redistributing wealth between any two agents would cause changes in total demand for assets that will cancel out each other, because of linearity and same slopes. Notice also that while identical curvature is a necessary condition, it is not sufficient for demand aggregation: Each of the six cases adds more conditions on top of this identical curvature requirement.[8]

A Dynamic Economy (*With* Idiosyncratic Risk)

While Rubinstein's (1974) theorem delivers a strong aggregation result, it achieves this by abstracting from a key aspect of dynamic economies: uncertainty that evolves over time. Almost every interesting economy that we discuss in the coming sections will feature some kind of idiosyncratic risk that individuals face (coming from labor income fluctuations, shocks to health, shocks to housing prices and asset returns, among others). Rubinstein's (1974) theorem is silent about how the aggregate economy behaves under these scenarios.

This is where Constantinides (1982) comes into play: He shows that if markets are complete, under much weaker conditions (on preferences, beliefs, discount rates, etc.) one can replace heterogeneous consumers with a planner who maximizes a weighted sum of consumers' utilities. In turn, the central planner can be replaced by a composite consumer who maximizes a utility function of aggregate consumption.

To show this, consider a private ownership economy with production as in Debreu (1959), with m consumers, n firms, and l commodities. As in Debreu

[8] Notice also that, because in some cases (such as [2]) heterogeneity in ρ is allowed, individuals will exhibit different relative risk aversions (if they have different w_t), for example in the generalized CRRA case, and still allow aggregation.

(1959), these commodities can be thought of as date-event labelled goods (and concave utility functions, U_i, as being defined over these goods), allowing us to map these results into an intertemporal economy with uncertainty. Consumer i is endowed with wealth $(w_{i1}, w_{i2}, ..., w_{il})$ and shares of firms $(\theta_{i1}, \theta_{i2}, ..., \theta_{in})$ with $\theta_{ij} \geq 0$ and $\sum_m \theta_{ij} = 1$. Let the vectors C_i and Y_j denote, respectively, individual i's consumption set and firm j's production set.

An equilibrium is an $(m+n+1)$-tuple $((\mathbf{c}_i^*)_{i=1}^m, (\mathbf{y}_j^*)_{j=1}^n, p^*)$ such that, as usual, consumers maximize utility, firms maximize their profits, and markets clear. Under standard assumptions, an equilibrium exists and is Pareto optimal.

Optimality implies that there exist positive numbers λ_i, $i = 1, ..., m$, such that the solution to the following problem (P1),

$$\begin{aligned} & \max_{\mathbf{c},\mathbf{y}} \sum_{i=1}^{m} \lambda_i U_i(\mathbf{c}_i) \qquad \text{(P1)} \\ \text{s.t. } \mathbf{y}_j &\in Y_j, \quad j = 1, 2, ...n; \\ \mathbf{c}_i &\in C_i, \quad i = 1, 2, ..., m; \\ \sum_{i=1}^{m} c_{ih} &= \sum_{j=1}^{n} y_{jh} + \sum_{i=1}^{m} w_{ih,} \quad h = 1, 2, ...l, \end{aligned}$$

(where h indexes commodities) is given by $(\mathbf{c}_i) = (\mathbf{c}_i^*)$ and $(\mathbf{y}_j) = (\mathbf{y}_j^*)$. Let aggregate consumption be $\mathbf{z} \equiv (z_1, \cdots, z_l)$, $z_h \equiv \sum_{i=1}^m c_{ih}$. Now, for a given $\mathbf{z}$, consider the problem (P2) of efficiently allocating it across consumers:

$$\begin{aligned} U(\mathbf{z}) &\equiv \max_{\mathbf{c}} \sum_{i=1}^{m} \lambda_i U_i(\mathbf{c}_i) \qquad \text{(P2)} \\ \text{s.t. } \mathbf{c}_i &\in C_i, \quad i = 1, 2, ..., m, \\ \sum_{i=1}^{m} c_{ih} &= z_h, \quad h = 1, 2, ..., l. \end{aligned}$$

Now, given the production sets of each firm and the aggregate endowments of each commodity, consider the optimal production decision (P3):

$$\begin{aligned} & \max_{\mathbf{y},\mathbf{z}} U(\mathbf{z}) \qquad \text{(P3)} \\ \text{s.t. } \mathbf{y}_j &\in Y_j, \forall j; \quad z_h = \sum_j y_{jh} + w_h, \forall_h. \end{aligned}$$

Theorem 3 (Constantinides [1982, Lemma 1]) *(a) The solution to (P3) is* $(\mathbf{y}_j) = (\mathbf{y}_j^*)$ *and* $z_h = \sum_{j=1}^n y_{jh}^* + w_h, \forall_h$.

(b) $U(\mathbf{z})$ *is increasing and concave in* $\mathbf{z}$.

(c) If $z_h = \sum y_{jh}^* + w_h, \forall_h$, *then the solution to (P2) is* $(x_i) = (x_i^*)$.

(d) Given $\lambda_i, i = 1, 2, \cdots, m$*, then if the consumers are replaced by one composite consumer with utility* $U(\mathbf{z})$*, with endowment equal to the sum of* m *consumers' endowments and shares the sum of their shares, then the* $(1 + n + 1)$*-tuple* $(\sum_{i=1}^{m} \mathbf{c}_i^*, (\mathbf{y}_j^*)_{j=1}^{n}, p^*)$ *is an equilibrium.*

Constantinides versus Rubinstein

Constantinides allows for much more generality than Rubinstein by relaxing two important restrictions. First, no conditions are imposed on the homogeneity of preferences, which was a crucial element in every version of Rubinstein's theorem. Second, Constantinides allows for both exogenous endowment as well as production at every date and state. In contrast, recall that, in Rubinstein's environment, individuals start life with a wealth stock and receive no further income or endowment during life. In exchange, Constantinides requires complete markets and does not get demand aggregation. Notice that the existence of a composite consumer does not imply demand aggregation, for at least two reasons. First, composite demand depends on the weights in the planner's problem and, thus, depends on the distribution of endowments. Second, the composite consumer is defined at equilibrium prices and there is no presumption that its demand curve is identical to the aggregate demand function.

Thus, the usefulness of Constantinides's result hinges on (i) the degree to which markets are complete, (ii) whether we want to allow for idiosyncratic risk and heterogeneity in preferences (which are both restricted in Rubinstein's theorem), and (iii) whether or not we need demand aggregation. Below I will address these issues in more detail. We will see that, interestingly, even when markets are not complete, in certain cases, we will not only get close to a composite consumer representation, but we can also get quite close to the much stronger result of demand aggregation! An important reason for this outcome is that many heterogeneous-agent models assume identical preferences, which eliminates an important source of heterogeneity, satisfying Rubinstein's conditions for preferences. While these models do feature idiosyncratic risk, as we shall see, when the planning horizon is long such shocks can often be smoothed effectively using even a simple risk-free asset. More on this in the coming sections.

Completing Markets by Adding Financial Assets

It is useful to distinguish between "physical" assets—those in positive net supply (e.g., equity shares, capital, housing, etc.)—and "financial" assets—those in zero net supply (bonds, insurance contracts, etc.). The latter are simply some contracts written on a piece of paper that specify the conditions under which one agent transfers resources to another. In principle, it can be created with little cost. Now suppose that we live in a world with J physical assets and

that there are $S(> J)$ states of the world. In this general setting, markets are incomplete. However, if consumers have homogenous tastes, endowments, and beliefs, then markets are (effectively) complete by simply adding enough *financial* assets (in zero net supply). There is no loss of optimality and nothing will change by this action, because in equilibrium identical agents will not trade with each other. The bottom line is that the more "homogeneity" we are willing to assume among consumers, the less demanding the complete markets assumption becomes. This point should be kept in mind as we will return to it later.

2. EMPIRICAL EVIDENCE ON INSURANCE

Dynamic economic models with heterogeneity typically feature individual-specific uncertainty that evolves over time—coming from fluctuations in labor earnings, health status, portfolio returns, among others. Although this structure does not fit into Rubinstein's environment, it is covered by Constantinides's theorem, which requires complete markets. Thus, a key empirical question is *the extent to which complete markets can serve as a useful benchmark* and a good approximation to the world we live in. As we shall see in this section, the answer turns out to be more nuanced than a simple yes or no.

To explain the broad variety of evidence that has been brought to bear on this question, this section is structured in the following way. First, I begin by discussing a large empirical literature that has tested a key prediction of complete markets—that marginal utility growth is equated across individuals. This is often called "perfect" or "full" insurance, and it is soundly rejected in the data. Next, I discuss an alternative benchmark, inspired by this rejection. This is the permanent income model, in which individuals have access to only borrowing and saving—or "self-insurance." In a way, this is the other extreme end of the insurance spectrum. Finally, I discuss studies that take an intermediate view—"partial insurance"—and provide some evidence to support it. We now begin with the tests of full insurance.

Benchmark 1: Full Insurance

To develop the theoretical framework underlying the empirical analyses, start with an economy populated by agents who derive utility from consumption c_t as well as some other good(s) d_t : $U^i\left(c^i_{t+1}, d^i_{t+1}\right)$, where i indexes individuals. These other goods can include leisure time (of husband and wife if the unit of analysis is a household), children, lagged consumption (as in habit formation models), and so on.

The key implication of perfect insurance can be derived by following two distinct approaches. The first environment assumes a social planner who pools

all individuals' resources and maximizes a social welfare function that assigns a positive weight to every individual. In the second environment, allocations are determined in a competitive equilibrium of a frictionless economy where individuals are able to trade in a complete set of financial securities. Both of these frameworks make the following strong prediction for the growth rate of individuals' marginal utilities:

$$\delta^i \frac{U_c^i\left(c_{t+1}^i, d_{t+1}^i\right)}{U_c^i\left(c_t^i, d_t^i\right)} = \frac{\Lambda_{t+1}}{\Lambda_t}, \tag{1}$$

where U_c denotes the marginal utility of consumption and Λ_t is the aggregate shock.[9] Thus, this condition says that every individual's marginal utility must grow in locksteps with the aggregate and, hence, with each other. No individual-specific term appears on the right-hand side, such as idiosyncratic income shocks, unemployment, sickness, and so on. All these idiosyncratic events are perfectly insured in this world. From here one can introduce a number of additional assumptions for empirical tractability.

Complete Markets and Cross-Sectional Heterogeneity: A Digression

So far we have focused on what market completeness implies for the study of aggregate phenomena in light of Constantinides's theorem. However, complete markets also imposes restrictions on the evolution of the *cross-sectional distribution*, which can be seen in (1). For a given specification of U, (1) translates into restrictions on the evolutions of c_t and d_t (possibly a vector). Although it is possible to choose U to be sufficiently general and flexible (e.g., include preference shifters, assume non-separability) to generate rich dynamics in cross-sectional distributions, this strategy would attribute all the action to preferences, which are essentially unobservable. Even in that case, models that are not bound by (1)—and therefore have idiosyncratic shocks affect individual allocations—can generate a much richer set of cross-sectional distributions. Whether that extra richness is necessary for explaining salient features of the data is another matter and is not always obvious (see, e.g., Caselli and Ventura [2000], Badel and Huggett [2007], and Guvenen and Kuruscu [2010]).[10]

---

[9] Alternatively stated, Λ_t is the Lagrange multiplier on the aggregate resource constraint at time t in the planner's problem or the state price density in the competitive equilibrium interpretation.

[10] Caselli and Ventura (2000) show that a wide range of distributional dynamics and income mobility patterns can arise in the Cass-Koopmans optimal savings model and in the Arrow-Romer model of productivity spillovers. Badel and Huggett (2007) show that life-cycle inequality patterns (discussed later) that have been viewed as evidence of incomplete markets can in fact be generated using a complete markets model. Guvenen and Kuruscu (2010) show that a human capital model with heterogeneity in learning ability and skill-biased technical change generates rich nonmonotonic

Now I return back to the empirical tests of (1).

In a pioneering article, Altug and Miller (1990) were the first to formally test the implications of (1). They considered households as their unit of analysis and specified a rich Beckerian utility function that included husbands' and wives' leisure times as well as consumption (food expenditures), and adjusted for demographics (children, age, etc.). Using data from the Panel Study of Income Dynamics (PSID), they could not reject full insurance. Hayashi, Altonji, and Kotlikoff (1996) revisited this topic a few years later and, using the same data set, they rejected perfect risk sharing.[11] Given this rejection in the whole population, they investigated if there might be better insurance *within* families, who presumably have closer ties with each other than the population at large and could therefore provide insurance to the members in need. They found that this hypothesis too was statistically rejected.[12]

In a similar vein, Guvenen (2007a) investigates how the extent of risk sharing varies across different wealth groups, such as stockholders and nonstockholders. This question is motivated by the observation that stockholders (who made up less than 20 percent of the population for much of the 20th century) own about 80 percent of net worth and 90 percent of financial wealth in the U.S. economy, and therefore play a disproportionately large role in the determination of macroeconomic aggregates. On the one hand, these wealthy individuals have access to a wide range of financial securities that can presumably allow better risk insurance; on the other hand, they are exposed to different risks not faced by the less-wealthy nonstockholders. Using data from the PSID, he strongly rejects perfect risk sharing among stockholders, but, perhaps surprisingly, does not find evidence against it among nonstockholders. This finding suggests further focus on risk factors that primarily affect the wealthy, such as entrepreneurial income risk that is concentrated at the top of the wealth distribution.

A number of other articles impose further assumptions before testing for risk sharing. A very common assumption is the separability between c_t and d_t (for example, leisure), which leads to an equation that only involves consumption (Cochrane 1991, Nelson 1994, Attanasio and Davis 1996).[13] Assuming power utility in addition to separability, we can take the logs of both sides of

---

dynamics consistent with the U.S. data since the 1970s, despite featuring no idiosyncratic shocks (and thus has complete markets).

[11] Data sets such as the PSID are known to go through regular revisions, which might be able to account for the discrepancy between the two articles' results.

[12] This finding has implications for the modeling of the household decision-making process as a unitary model as opposed to one in which there is bargaining between spouses.

[13] Non-separability, for example between consumption and leisure, can be allowed for *if* the planner is assumed to be able to transfer leisure freely across individuals. While transfers of consumption are easier to implement (through taxes and transfers), the transfer of leisure is harder to defend on empirical grounds.

equation (1) and then time-difference to obtain

$$\Delta C_{i,t} = \Delta \Lambda_t, \tag{2}$$

where $C_t \equiv \log(c_t)$ and $\Delta C_t \equiv C_t - C_{t-1}$. Several articles have tested this prediction by running a regression of the form

$$\Delta C_{i,t} = \Delta \Lambda_t + \Psi' \mathbf{Z}_t^i + \epsilon_{i,t}, \tag{3}$$

where the vector $\mathbf{Z}_t^i$ contains factors that are idiosyncratic to individual/household/group i. Perfect insurance implies that all the elements of the vector Ψ are equal to zero.

Cochrane (1991), Mace (1991), and Nelson (1994) are the early studies that exploit this simple regression structure. Mace (1991) focuses on whether or not consumption responds to idiosyncratic wage shocks, i.e., $\mathbf{Z}_t^i = \Delta W_t^i$.[14] While Mace fails to reject full insurance, Nelson (1994) later points out several issues with the treatment of data (and measurement error in particular) that affect Mace's results. Nelson shows that a more careful treatment of these issues results in strong rejection.

Cochrane (1991) raises a different point. He argues that studies such as Mace's, that test risk sharing by examining the response of consumption growth to income, may have low power if income changes are (at least partly) anticipated by individuals. He instead proposes to use idiosyncratic events that are arguably harder to predict, such as plant closures, long strikes, long illnesses, and so on. Cochrane rejects full insurance for illness or involuntary job loss but not for long spells of unemployment, strikes, or involuntary moves. Notice that a crucial assumption in all of the work of this kind is that none of these shocks can be correlated with unmeasured factors that determine marginal utility growth.

Townsend (1994) tests for risk sharing in village economies of India and concludes that, although the model is statistically rejected, full insurance provides a surprisingly good benchmark. Specifically, he finds that individual consumption co-moves with village-level consumption and is not influenced much by own income, sickness, and unemployment.

Attanasio and Davis (1996) observe that equation (2) must also hold for multiyear changes in consumption and when aggregated across groups of individuals.[15] This implies, for example, that even if one group of individuals experiences faster income growth relative to another group during a 10-year period, their consumption growth must be the same. The substantial rise in the education premium in the United States (i.e., the wages of college graduates

[14] Because individual wages are measured with (often substantial) error in microsurvey data sets, an ordinary least squares estimation of this regression would suffer from attenuation bias, which may lead to a failure to reject full insurance even when it is false. The articles discussed here employ different approaches to deal with this issue (such as using an instrumental variables regression or averaging across groups to average out measurement error).

[15] Hayashi, Altonji, and Kotlikoff (1996) also use multiyear changes to test for risk sharing.

relative to high school graduates) throughout the 1980s provided a key test of perfect risk sharing. Contrary to this hypothesis, Attanasio and Davis (1996) find that the consumption of college graduates grows much faster than that of high school graduates during the same period, violating the premise of perfect risk sharing.

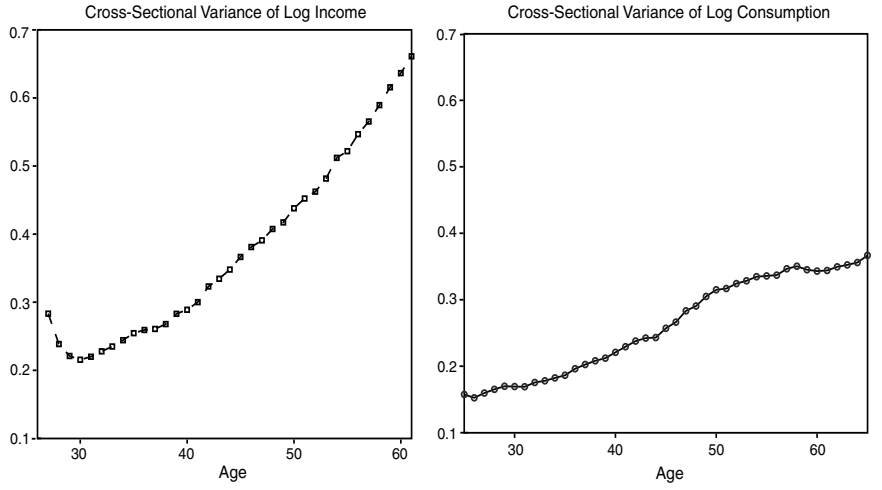
Finally, Schulhofer-Wohl (2011) sheds new light on this question. He argues that if more risk-tolerant individuals self-select into occupations with more (aggregate) income risk, then the regressions in (3) used by Cochrane (1991), Nelson (1994), and others (which incorrectly assume away such correlation) will be biased toward rejecting perfect risk sharing. By using self-reported measures of risk attitudes from the Health and Retirement Survey, Schulhofer-Wohl establishes such a correlation. Then he develops a method to deal with this bias and, applying the corrected regression, he finds that consumption growth responds very weakly to idiosyncratic shocks, implying much larger risk sharing than can be found in these previous articles. He also shows that the coefficients estimated from this regression can be mapped into a measure of "partial insurance."

Taking Stock

As the preceding discussion makes clear, with few exceptions, all empirical studies agree that perfect insurance in the whole population is strongly rejected in a statistical sense. However, this statistical rejection per se is not sufficient to conclude that complete markets is a *poor benchmark* for economic analysis for two reasons. First, there seems to be a fair deal of insurance against certain types of shocks, as documented by Cochrane (1991) and Townsend (1994), and among certain groups of households, such as in some villages in less developed countries (Townsend 1994), or among nonstockholders in the United States (Guvenen 2007a). Second, the reviewed empirical evidence arguably documents statistical tests of an extreme benchmark (equation [1]) that we should not expect to hold precisely—for every household, against every shock. Thus, with a large enough sample, statistical rejection should not be surprising.[16] What these tests do not do is tell us how "far" the economy is from the perfect insurance benchmark. In this sense, analyses such as in Townsend (1994)—that identify the types of shocks that are and are not insured—are somewhat more informative than those in Altug and Miller (1990), Hayashi, Altonji, and Kotlikoff (1996), and Guvenen (2007a), which rely on model misspecification-type tests of risk sharing.

[16] One view is that hypothesis tests without an explicit alternative (such as the ones discussed here) often "degenerate into elaborate rituals designed to measure the sample size (Leamer 1983, 39)."

Figure 1 Within-Cohort Inequality over the Life Cycle



Benchmark 2: Self-Insurance

The rejection of full consumption insurance led economists to search for other benchmark frameworks for studying individual choices under uncertainty. One of the most influential studies of this kind has been Deaton and Paxson (1994), who bring a different kind of evidence to bear. They begin by documenting two empirical facts. Using microdata from the United States, United Kingdom, and Taiwan, they first document that within-cohort inequality of labor income (as measured by the variance of log income) increases substantially and almost linearly over the life cycle. Second, they document that within-cohort consumption inequality shows a very similar pattern and also rises substantially as individuals age. The two empirical facts are replicated in Figure 1 from data in Guvenen (2007b, 2009a).

To understand what these patterns imply for the market structure, first consider a complete markets economy. As we saw in the previous section, if consumption is separable from leisure and other potential determinants of marginal utility, consumption growth will be equalized across individuals, independent of any idiosyncratic shock (equation [2]). Therefore, while consumption level may differ across individuals because of differences in permanent lifetime resources, this dispersion should *not* change as the cohort ages.[17] Therefore, Deaton and Paxson's (1994) evidence has typically been

[17] There are two obvious modifications that preserve complete markets and would be consistent with rising consumption inequality. The first one is to introduce heterogeneity in time

interpreted as contradicting the complete markets framework. I now turn to the details.

The Permanent Income Model

The canonical framework for self-insurance is provided by the permanent income life-cycle model, in which individuals only have access to a risk-free asset for borrowing and saving. Therefore, as opposed to full insurance, there is only "self-insurance" in this framework. Whereas the complete markets framework represents the maximum amount of insurance, the permanent income model arguably provides the lower bound on insurance (to the extent that we believe individuals have access to a savings technology, and borrowing is possible subject to some constraints).

It is instructive to develop this framework in some detail as the resulting equations will come in handy in the subsequent exposition. The framework here closely follows Hall and Mishkin (1982) and Deaton and Paxson (1994). Start with an income process with permanent and transitory shocks:

$$\begin{aligned} y_t &= y_t^P + \varepsilon_t, \\ y_t^P &= y_{t-1}^P + \eta_t. \end{aligned} \tag{4}$$

Suppose that individuals discount the future at the rate of interest and define: $\delta = 1/(1+r)$. Preferences are of quadratic utility form:

$$\begin{aligned} &\max E_0 \left[-\frac{1}{2} \sum_{t=1}^{T} \delta_t \left(c^* - c_t \right)^2 \right] \\ &\text{s.t.} \sum_{t=1}^{T} \delta_t \left(y_t - c_t \right) + A_0 = 0, \end{aligned} \tag{5}$$

where c^* is the bliss level and A_0 is the initial wealth level (which may be zero). This problem can be solved in closed form to obtain a consumption function. First-differencing this consumption rule yields

$$\Delta c_t = \eta_t + \gamma_t \varepsilon_t, \tag{6}$$

where $\gamma_t \equiv 1/\left(\sum_{\tau=0}^{T-t} \delta^\tau \right)$ is the annuitization factor.[18] This term is close to zero when the horizon is long and the interest rate is not too high, the

discounting. This is not very appealing because it "explains" by entirely relying on unobservable preference heterogeneity. Second, one could question the assumption of separability: If leisure is non-separable and wage inequality is rising over the life cycle—which it does—then consumption inequality would also rise to keep marginal utility growth constant (even under complete markets). But this explanation also predicts that hours inequality should also rise over the life cycle, a prediction that does not seem to be borne out in the data—although see Badel and Huggett (2007) for an interesting dissenting take on this point.

[18] Notice that the derivation of (6) requires two more pieces in addition to the Euler equation: It requires us to explicitly specify the budget constraint (5) as well as the stochastic process for income (4).

well-understood implication being that the response of consumption to transitory shocks is very weak given their low annuitized value. More importantly: Consumption responds to permanent shocks one-for-one. Thus, consumption changes reflect permanent income changes.

For the sake of this discussion, assume that the horizon is long enough so that $\gamma_t \approx 0$ and thus $\Delta c_t \cong \eta_t$. If we further assume that $\text{cov}_i\left(c^i_{t-1}, \eta^i_t\right) = 0$ (where i indexes individuals and the covariance is taken cross-sectionally), we get

$$\text{var}_i\left(c^i_t\right) \cong \text{var}_i\left(c^i_{t-1}\right) + \text{var}\left(\eta_t\right).$$

So the rise in consumption inequality from age $t-1$ to t is a measure of the variance of the permanent shock between those two ages. Since, as seen in Figure 1, consumption inequality rises significantly and almost linearly, this figure is consistent with permanent shocks to income that are fully accommodated as predicted by the permanent income model.

Deaton and Paxson's Striking Conclusion

Based on this evidence, Deaton and Paxson (1994) argue that the permanent income model is a better benchmark for studying individual allocations than is complete markets. Storesletten, Telmer, and Yaron (2004a) go one step further and show that a calibrated life-cycle model with incomplete markets can be quantitatively consistent with the rise in consumption inequality as long as income shocks are sufficiently persistent ($\rho \gtrsim 0.90$). In his presidential address to the American Economic Association, Robert Lucas (2003, 10) succinctly summarized this view: "The fanning out over time of the earnings and consumption distributions within a cohort that Deaton and Paxson [1994] document is striking evidence of a sizable, uninsurable random walk component in earnings." This conclusion was shared by the bulk of the profession in the 1990s and 2000s, giving a strong impetus to the development of incomplete markets models featuring large and persistent shocks that are uninsurable. I review many of these models in Sections 3 and 4. However, a number of recent articles have revisited the original Deaton-Paxson finding and have reached a different conclusion.

Reassessing the Facts: An Opposite Conclusion

Four of these articles, by and large, follow the same methodology as described and implemented by Deaton and Paxson (1994), but each uses a data set that extends the original Consumer Expenditure Survey (CE) sample used by these authors (that covered 1980–1990) and differ somewhat in their sample selection strategy. Specifically, Primiceri and van Rens (2009, Figure 2) use data from 1980–2000; Heathcote, Perri, and Violante (2010, Figure 14) use the 1980–1998 sample; Guvenen and Smith (2009, Figure 11) use the 1980–1992

sample and augment it with the 1972–73 sample; and Kaplan (2010, Figure 2) uses data from 1980–2003. Whereas Deaton and Paxson (1994, Figures 4 and 8) and Storesletten, Telmer, and Yaron (2004a, Figure 1) document a rise in consumption inequality of about 30 log points (between ages 25 and 65), these four articles find a much smaller rise of about 5–7 log points.

Taking Stock

Taken together, these re-analyses of CE data reveal that Deaton and Paxson's (1994) earlier conclusion is not robust to small changes in the sample period studied. Although more work on this topic certainly seems warranted,[19] these recent studies raise substantial concerns on one of the key pieces of empirical evidence on the extent of market incompleteness. A small rise in consumption inequality is hard to reconcile with the combination of large permanent shocks *and* self-insurance. Hence, if this latter view is correct, either income shocks are not as permanent as we thought or there is insurance above and beyond self-insurance. Both of these possibilities are discussed next.

An Intermediate Case: Partial Insurance

A natural intermediate case to consider is an environment between the two extremes of full insurance and self-insurance. That is, perhaps individuals have access to various sources of insurance (e.g., through charities, help from family and relatives, etc.) *in addition* to borrowing and saving, but these forms of insurance still fall short of full insurance. If this is the case, is there a way to properly measure the degree of this "partial insurance?"

To address this question, Blundell, Pistaferri, and Preston (2008) examine the response of consumption to innovations in income. They start with equation (6) derived by Hall and Mishkin (1982) that links consumption change to income innovations, and modify it by introducing two parameters—θ and ϕ—to encompass a variety of different scenarios:

$$\Delta c_t = \theta \eta_t + \phi \gamma_t \varepsilon_t. \tag{7}$$

Now, at one extreme is the self-insurance model (i.e., the permanent income model): $\theta = \phi = 1$; at the other extreme is a model with full insurance: $\theta = \phi = 0$. Values of θ and ϕ between zero and one can be interpreted as the degree of partial insurance—the lower the value, the more insurance there

[19] For example, as Attanasio, Battistin, and Ichimura (2007) show, the facts regarding the rise in consumption inequality over time are sensitive to whether one uses the "recall survey" or the "diary survey" in the CE data set. All the articles discussed in this section (on consumption inequality over the life cycle, including Deaton and Paxson [1994]) use the recall survey data. It would be interesting to see if the diary survey alters the conclusions regarding consumption inequality over the life cycle.

is. In their baseline analysis, Blundell, Pistaferri, and Preston (2008) estimate $\theta \approx \frac{2}{3}$ and find that it does not vary significantly over the sample period.[20] They interpret the estimate of θ to imply that about $\frac{1}{3}$ of permanent shocks are insured above and beyond what can be achieved through self-insurance.[21]

A couple of remarks are in order. First, the derivation of equation (6) that forms the basis of the empirical analysis here requires quadratic preferences. Indeed, this was the maintained assumption in Hall and Mishkin (1982) and Deaton and Paxson (1994). Blundell, Pistaferri, and Preston (2008) show that one can derive, as an approximation, an analogous equation (7) with CRRA utility and self-insurance, but now $\theta = \phi \approx \pi_{i,t}$, where $\pi_{i,t}$ is the ratio of human wealth to total wealth. In other words, the coefficients θ and ϕ are both equal to one under self-insurance only if preferences are of quadratic form; generalizing to CRRA predicts that even with self-insurance the response to permanent shocks, given by $\pi_{i,t}$, will be less than one-for-one if non-human wealth is positive. Thus, accumulation of wealth because of precautionary savings or retirement can dampen the response of consumption to permanent shocks and give the appearance of partial insurance. Blundell, Pistaferri, and Preston (2008) examine if younger individuals (who have less non-human wealth and thus have a higher $\pi_{i,t}$ than older individuals) have a higher response coefficient to permanent shocks. They do find this to be the case.

Insurance or Advance Information?

Primiceri and van Rens (2009) conduct an analysis similar to Blundell, Pistaferri, and Preston (2008) and also find a small response of consumption to permanent income movements. However, they adopt a different interpretation for this finding—that income movements are largely "anticipated" by the individuals as opposed to being genuine permanent "shocks." As has been observed as far back as Hall and Mishkin (1982), this alternative interpretation illustrates a fundamental challenge with this kind of analysis: Advance information and partial insurance are difficult to disentangle by simply examining the response of consumption to income.

Insurance or Less Persistent Shocks?

Kaplan and Violante (2010) raise two more issues regarding the interpretation of θ. First, they ask, what if income shocks are persistent but not permanent?

---

[20] They also find $\phi\gamma_t = 0.0533$ (0.0435), indicating very small transmission of transitory shocks to consumption. This is less surprising since it would also be implied by the permanent income model.

[21] The parameter ϕ is of lesser interest given that transitory shocks are known to be smoothed quite well even in the permanent income model and the value of ϕ one estimates depends on what one assumes about γ_t—hence, the interest rates.

This is a relevant question because, as I discuss in the next section, nearly all empirical studies that estimate the persistence coefficient (of an AR(1) or ARMA(1,1)) find it to be 0.95 or lower—sometimes as low as 0.7. To explore this issue, they simulate data from a life-cycle model with self-insurance only, in which income shocks follow an AR(1) process with a first-order autocorrelation of 0.95. They show that when they estimate θ as in Blundell, Pistaferri, and Preston (2008), they find it to be close to the $\frac{2}{3}$ figure reported by these authors.[22] Second, they add a retirement period to the life-cycle model, which has the effect that now even a unit root shock is not permanent, given that its effect does not translate one-for-one into the retirement period. Thus, individuals have even more reason not to respond to permanent shocks, especially when they are closer to retirement. Overall, their findings suggest that the response coefficient of consumption to income can be generated in a model of pure self-insurance to the extent that income shocks are allowed to be slightly less than permanent.[23] One feature this model misses, however, is the age profile of response coefficients, which shows no clear trend in the data according to Blundell, Pistaferri, and Preston (2008), but is upward sloping in Kaplan and Violante's (2010) model.

Taking Stock

Before the early 1990s, economists typically appealed to aggregation theorems to justify the use of representative-agent models. Starting in the 1990s, the widespread rejections of the full insurance hypothesis (necessary for Constantinides's [1982] theorem), combined with the findings of Deaton and Paxson (1994), led economists to adopt versions of the permanent income model as a benchmark to study individual's choices under uncertainty (Hubbard, Skinner, and Zeldes [1995], Carroll [1997], Carroll and Samwick [1997], Blundell and Preston [1998], Attanasio et al. [1999], and Gourinchas and Parker [2002], among many others). The permanent income model has two key assumptions: a single risk-free asset for self-insurance *and* permanent—or very persistent—shocks, typically implying substantial idiosyncratic risk. The more recent evidence, discussed in this subsection, however, suggests that a more appropriate benchmark needs to incorporate either more opportunities for partial insurance or idiosyncratic risk that is smaller than once assumed.

---

[22] The reason is simple. Because the AR(1) shock decays exponentially, this shock loses 5 percent of its value in one year, but $1 - 0.95^{10} \approx 40$ percent in 10 years and 65 percent in 20 years. Thus, the discounted lifetime value of such a shock is significantly lower than a permanent shock, which retains 100 percent of its value at all horizons.

[23] Another situation in which $\theta < 1$ with self-insurance alone is if permanent and transitory shocks are not separately observable and there is estimation risk.

3. INCOMPLETE MARKETS IN GENERAL EQUILIBRIUM

This section and the next discuss incomplete markets models in general equilibrium without aggregate shocks. Bringing in a general equilibrium structure allows researchers to jointly analyze aggregate and distributional issues. As we shall see, the two are often intertwined, making such models very useful. The present section discusses the key ingredients that go into building a general equilibrium incomplete markets model (e.g., types of risks to consider, borrowing limits, modeling individuals versus households, among others). The next section presents three broad questions that these models have been used to address: the cross-sectional distributions of consumption, earnings, and wealth. These are substantively important questions and constitute an entry point into broader literatures. I now begin with a description of the basic framework.

The Aiyagari (1994) Model

In one of the first quantitative models with heterogeneity, Imrohoroglu (1989) constructed a model with liquidity constraints and unemployment risk that varied over the business cycle. She assumed that interest rates were constant to avoid the difficulties with aggregate shocks, which were subsequently solved by Krusell and Smith (1998). She used this framework to re-assess Lucas's (1987) earlier calculation of the welfare cost of business cycles. She found only a slightly higher figure than Lucas, mainly because of her focus on unemployment risk, which typically has a short duration in the United States.[24] Regardless of its empirical conclusions, this article represents an important early effort in this literature.

In what has become an important benchmark model, Aiyagari (1994) studies a version of the deterministic growth framework, with a Neoclassical production function and a large number of infinitely lived consumers (dynasties). Consumers are ex ante identical, but there is ex post heterogeneity because of idiosyncratic shocks to labor productivity, which are not directly insurable (via insurance contracts). However, consumers can accumulate a (conditionally) risk-free asset for self-insurance. They can also borrow in this asset, subject to a limit determined in various ways. At each point in time, consumers may differ in the history of productivities experienced, and hence in accumulated wealth.

[24] There is a large literature on the costs of business cycles following Lucas's original calculation. I do not discuss these articles here for brevity. Lucas's (2003) presidential address to the American Economic Association is an extensive survey of this literature that also discusses how Lucas's views on this issue evolved since the original 1987 article.

More concretely, an individual solves the following problem:

$$\max_{\{c_t\}} E_0 \left[\sum_{t=0}^{\infty} \delta^t U(c_t) \right]$$
$$\text{s.t. } c_t + a_{t+1} = w l_t + (1 + r) a_t,$$
$$a_t \geq -B_{\min}, \tag{8}$$

and l_t follows a finite-state first-order Markov process.[25]

There are (at least) two ways to embed this problem in general equilibrium. Aiyagari (1994) considers a production economy and views the single asset as the capital in the firm, which obviously has a positive net supply. In this case, aggregate production is determined by the savings of individuals, and both r and the wage rate w, must be determined in general equilibrium. Huggett (1993) instead assumes that the single asset is a household bond in zero net supply. In this case, the aggregate amount of goods in the economy is exogenous (exchange economy), and the only aggregate variable to be determined is r.

The borrowing limit $B_{\min}$ can be set to the "natural" limit, which is defined as the loosest possible constraint consistent with certain repayment of debt: $B_{\min} = w l_{\min}/r$. Note that if $l_{\min}$ is zero, this natural limit will be zero. Some authors have used this feature to rule out borrowing (e.g., Carroll [1997] and Gourinchas and Parker [2002]). Alternatively, it can be set to some ad hoc limit stricter than the natural one. More on this later.

The main substantive finding in Aiyagari (1994) is that with incomplete markets, the aggregate capital stock is higher than it is with complete markets, although the difference is not quantitatively very large. Consequently, the interest rate is lower (than the time preference rate), which is also true in Huggett's (1993) exchange economy version. This latter finding initially led economists to conjecture that these models could help explain the equity premium puzzle,[26] which is also generated by a low interest rate. It turns out that while this environment helps, it is neither necessary nor sufficient to generate a low interest rate. I return to this issue later. Aiyagari (1994) also shows that the model generates the right ranking between different types of inequality: Wealth is more dispersed than income, which is more dispersed than consumption.

---

[25] Prior to Aiyagari, the decision problem described here was studied in various forms by, among others, Bewley (undated), Schechtman and Escudero (1977), Flavin (1981), Hall and Mishkin (1982), Clarida (1987, 1990), Carroll (1991), and Deaton (1991). With the exceptions of Bewley (undated) and Clarida (1987, 1990), however, most of these earlier articles did not consider general equilibrium, which is the main focus here.

[26] The equity premium puzzle of Mehra and Prescott (1985) is the observation that, in the historical data, stocks yield a much higher return than bonds over long horizons, which has turned out to be very difficult to explain by a wide range of economic models.

The frameworks analyzed by Huggett (1993) and Aiyagari (1994) contain the bare bones of a canonical general equilibrium incomplete markets model. As such, they abstract from many ingredients that would be essential today for conducting serious empirical/quantitative work, especially given that almost two decades have passed since their publication. In the next three subsections, I review three main directions the framework can be extended. First, the nature of idiosyncratic risk is often crucial for the implications generated by the model. There is a fair bit of controversy about the precise nature and magnitude of such risks, which I discuss in some detail. Second, and as I alluded to above, the treatment of borrowing constraints is very reduced form here. The recent literature has made significant progress in providing useful microfoundations for a richer specification of borrowing limits. Third, the Huggett-Aiyagari model considers an economy populated by bachelor(ette)s as opposed to families—this distinction clearly can have a big impact on economic decisions, which is also discussed.

Nature of Idiosyncratic Income Risk[27]

The rejection of perfect insurance brought to the fore idiosyncratic shocks as important determinants of economic choices. However, after three decades of empirical research (since Lillard and Willis [1978]), a consensus among researchers on the nature of labor income risk still remains elusive. In particular, the literature in the 1980s and 1990s produced two—quite opposite—views on the subject. To provide context, consider this general specification for the wage process:

$$y_t^i = \underbrace{g\,(t, \text{observables}, ...)}_{\text{common systematic component}} + \underbrace{\left[\alpha^i + \beta^i t\right]}_{\text{profile heterogeneity}} + \underbrace{\left[z_t^i + \varepsilon_t^i\right]}_{\text{stochastic component}} \tag{9}$$

$$z_t^i = \rho z_{t-1}^i + \eta_t^i, \tag{10}$$

where η_t^i and ε_t^i are zero mean innovations that are i.i.d. over time and across individuals.

The early articles on income dynamics estimate versions of the process given in (9) from labor income data and find: $0.5 < \rho < 0.7$, and $\sigma_\beta^2 \gg 0$ (Lillard and Weiss 1979; Hause 1980). Thus, according to this first view, which I shall call the "heterogeneous income profiles" (HIP) model, individuals are subject to shocks with modest persistence, while facing life-cycle profiles that

[27] The exposition here draws heavily on Guvenen (2009a).

are individual-specific (and hence vary significantly across the population). As we will see in the next section, one theoretical motivation for this specification is the human capital model, which implies differences in income profiles if, for example, individuals differ in their ability level.

In an important article, MaCurdy (1982) casts doubt on these findings. He tests the null hypothesis of $\sigma_\beta^2 = 0$ and fails to reject it. He then proceeds by imposing $\sigma_\beta^2 \equiv 0$ before estimating the process in (9), and finds $\rho \approx 1$ (see, also, Abowd and Card [1989], Topel [1990], Hubbard, Skinner, and Zeldes [1995], and Storesletten, Telmer, and Yaron [2004b]). Therefore, according to this alternative view, which I shall call the "restricted income profiles" (RIP) model, individuals are subject to extremely persistent—nearly random walk—shocks, while facing similar life-cycle income profiles.

MaCurdy's (1982) Test

More recently, two articles have revived this debate. Baker (1997) and Guvenen (2009a) have shown that MaCurdy's test has low power and therefore the lack of rejection does not contain much information about whether or not there is growth rate heterogeneity. MaCurdy's test was generally regarded as the strongest evidence against the HIP specification, and it was repeated in different forms by several subsequent articles (Abowd and Card 1989; Topel 1990; and Topel and Ward 1992), so it is useful to discuss in some detail.

To understand its logic, notice that, using the specification in (9) and (10), the nth autocovariance of income growth can be shown to be

$$\operatorname{cov}\left(\Delta y_t^i, \Delta y_{t+n}^i\right) = \sigma_\beta^2 - \rho^{n-1}\left(\frac{1-\rho}{1+\rho}\sigma_\eta^2\right), \tag{11}$$

for $n \geq 2$. The idea of the test is that for *sufficiently large* n, the second term will vanish (because of exponential decay in ρ^{n-1}), leaving behind a positive autocovariance equal to σ_β^2. Thus, *if* HIP is indeed important—σ_β^2 is positive—then higher order autocovariances must be positive.

Guvenen (2009a) raises two points. First, he asks how large n must be for the second term to be negligible. He shows that for the value of persistence he estimates with the HIP process ($\rho \cong 0.82$), the autocovariances in (11) do not even turn positive before the 13th lag (because the second term dominates), whereas MaCurdy only studies the first 5 lags. Second, he conducts a Monte Carlo analysis in which he simulates data using equation (9) with substantial heterogeneity in growth rates.[28] The results of this analysis are reproduced here in Table 1. MaCurdy's test does not reject the false null hypothesis of $\sigma_\beta^2 = 0$ for any sample size smaller than 500,000 observations (column 3)!

[28] More concretely, the estimated value of σ_β^2 used in his Monte Carlo analysis implies that at age 55 more than 70 percent of wage inequality is because of profile heterogeneity.

Table 1 How Informative is MaCurdy's (1982) Test?

		Autocovariances			Autocorrelations	
		Data	HIP Process		Data	HIP Process
Lag ↓	$N \longrightarrow$	27,681	27,681	500,00	27,681	27,681
0		.1215	.1136	.1153	1.00	1.00
		(.0023)	(.00088)	(.00016)	(.000)	(.000)
1		−.0385	−.04459	−.04826	−.3174	−.3914
		(.0011)	(.00077)	(.00017)	(.010)	(.0082)
2		−.0031	−.00179	−.00195	−.0261	−.0151
		(.0010)	(.00075)	(.00018)	(.008)	(.0084)
3		−.0023	−.00146	−.00154)	−.0192	−.0128
		(.0008)	(.00079)	(.00020)	(.009)	(.0087)
4		−.0025	−.00093	−.00120	−.0213	−.0080
		(.0007)	(.00074)	(.00019)	(.010)	(.0083)
5		−.0001	−.00080	−.00093	−.0012	−.0071
		(.0008)	(.00081)	(.00020)	(.007)	(.0090)
10		−.0017	−.00003	−.00010	−.0143	−.0003
		(.0006)	(.00072)	(.00019)	(.009)	(.0081)
15		.0053	.00017	.00021	.0438	.0015
		(.0007)	(.00076)	(.00020)	(.008)	(.0086)
18		.0012	.00036	.00030	.0094	.0032
		(.0009)	(.00076)	(.00018)	(.011)	(.0087)

Notes: The table is reproduced from Guvenen (2009a, Table 3). N denotes the sample size (number of individual-years) used to compute the statistics. Standard errors are in parentheses. The statistics in the "data" columns are calculated from a sample of 27,681 males from the PSID as described in that article. The counterparts from simulated data are calculated using the same number of individuals and a HIP process fitted to the covariance matrix of income residuals.

Even in that case, only the 18th autocovariance is barely significant (with a t-statistic of 1.67). For comparison, MaCurdy's (1982) data set included around 5,000 observations. Even the more recent PSID data sets typically contain fewer than 40,000 observations.

In light of these results, imposing the *a priori* restriction of $\sigma_\beta^2 = 0$ on the estimation exercise seems a risky route to follow. Baker (1997), Haider (2001), Haider and Solon (2006), and Guvenen (2009a) estimate the process in (9) without imposing this restriction and find substantial heterogeneity in β^i and a low persistence, confirming the earlier results of Lillard and Weiss (1979) and Hause (1980). Baker and Solon (2003) use a large panel data set drawn from Canadian tax records and allow for both permanent shocks and profile heterogeneity. They find statistically significant evidence of both components.

In an interesting recent article, Browning, Ejrnaes, and Alvarez (2010) estimate an income process that allows for "lots of" heterogeneity. The authors use a simulated method of moments estimator and match a number of

moments whose economic significance is more immediate than the covariance matrix of earnings residuals, which has typically been used as the basis of a generalized method of moments estimation in the bulk of the extant literature. They uncover a lot of interesting heterogeneity, for example, in the innovation variance as well as in the persistence of AR(1) shocks. Moreover, they "find strong evidence against the hypothesis that any worker has a unit root." Gustavsson and Österholm (2010) use a long panel data set (1968–2005) from administrative wage records on Swedish individuals. They employ local-to-unity techniques on individual-specific time series and reject the unit root assumption.

Inferring Risk versus Heterogeneity from Economic Choices

Finally, a number of recent articles examine the response of consumption to income shocks to infer the nature of income risk. In an important article, Cunha, Heckman, and Navarro (2005) measure the fraction of individual-specific returns to education that are predictable by individuals by the time they make their college decision versus the part that represents uncertainty. Assuming a complete markets structure, they find that slightly more than half of the returns to education represent known heterogeneity from the perspective of individuals.

Guvenen and Smith (2009) study the joint dynamics of consumption and labor income (using PSID data) in order to disentangle "known heterogeneity" from income risk (coming from shocks as well as from uncertainty regarding one's own income growth rate). They conclude that a moderately persistent income process ($\rho \approx 0.7$–0.8) is consistent with the joint dynamics of income and consumption. Furthermore, they find that individuals have significant information about their own β^i at the time they enter the labor market and hence face little uncertainty coming from this component. Overall, they conclude that with income shocks of modest persistence and largely predictable income growth rates, the income risk perceived by individuals is substantially smaller than what is typically assumed in calibrating incomplete markets models (many of which borrow their parameter values from MaCurdy [1982], Abowd and Card [1989], and Meghir and Pistaferri [2004], among others). Along the same lines, Krueger and Perri (2009) use rich panel data on Italian households and conclude that the response of consumption to income suggests low persistence for income shocks (or a high degree of partial insurance).[29]

Studying economic choices to disentangle risk from heterogeneity has many advantages. Perhaps most importantly, it allows researchers to bring a

[29] A number of important articles have also studied the response of consumption to income, such as Blundell and Preston (1998) and Blundell, Pistaferri, and Preston (2008). These studies, however, assume the persistence of income shocks to be constant and instead focus on what can be learned about the sizes of income shocks over time.

much broader set of data to bear on the question. For example, many dynamic choices require individuals to carefully weigh the different future risks they perceive against predictable changes before making a commitment. Decisions on home purchases, fertility, college attendance, retirement savings, and so on are all of this sort. At the same time, this line of research also faces important challenges: These analyses need to rely on a fully specified economic model, so the results can be sensitive to assumptions regarding the market structure, specification of preferences, and so on. Therefore, experimenting with different assumptions is essential before a definitive conclusion can be reached with this approach. Overall, this represents a difficult but potentially very fruitful area for future research.

Wealth, Health, and Other Shocks

One source of idiosyncratic risk that has received relatively little attention until recently comes from shocks to wealth holdings, resulting for example from fluctuations in housing prices and stock returns, among others. A large fraction of the fluctuations in housing prices are because of local or regional factors and are substantial (as the latest housing market crash showed once again). So these fluctuations can have profound effects on individuals' economic choices. In one recent example, Krueger and Perri (2009) use panel data on Italian households' income, consumption, and wealth. They study the response of consumption to income and wealth shocks and find the latter to be very important. Similarly, Mian and Sufi (2011) use individual-level data from 1997–2008 and show that housing price boom leads to significant equity extraction—about 25 cents for every dollar increase in prices—which in turn leads to higher leverage and personal default during this time. Their "conservative" estimate is that home equity-based borrowing added $1.25 trillion in household debt and accounted for about 40 percent of new defaults from 2006–2008.

Another source of idiosyncratic shocks is out-of-pocket medical expenditures (hospital bills, nursing home expenses, medications, etc.), which can potentially have significant effects on household decisions. French and Jones (2004) estimate a stochastic process for health expenditures, modeled as a normal distribution adjusted to capture the risk of catastrophic health care costs. Simulating this process, they show that 0.1 percent of households every year receive a health cost shock with a present value exceeding $125,000. Hubbard, Skinner, and Zeldes (1994, 1995) represent the earliest efforts to introduce such shocks into quantitative incomplete markets models. The 1995 article shows that the interaction of such shocks with means-tested social insurance programs is especially important to account for in order to understand the very low savings rate of low-income individuals.

De Nardi, French, and Jones (2010) ask if the risk of large out-of-pocket medical expenditures late in life can explain the savings behavior of the elderly. They examine a new and rich data set called AHEAD, which is part of the Health and Retirement Survey conducted by the University of Michigan, which allows them to characterize medical expenditure risk for the elderly (even for those in their 90s) more precisely than previous studies, such as Hubbard, Skinner, and Zeldes (1995) and Palumbo (1999).[30] De Nardi, French, and Jones (2010) find out-of-pocket expenditures to rise dramatically at very old ages, which (in their estimated model) provides an explanation for the lack of significant dissaving by the elderly. Ozkan (2010) shows that the life-cycle profile of medical costs (inclusive of the costs paid by private and public insurers to providers) differs significantly between rich and poor households. In particular, on average, the medical expenses of the rich are higher than those of the poor until mid-life, after which the expenses of the poor exceed those of the rich—by 25 percent in absolute terms. Further, the expenses of the poor have thick tails—lots of individuals with zero expenses and many with catastrophically high costs. He builds a model in which individuals can invest in their health (i.e., preventive care), which affects the future distribution of health shocks and, consequently, the expected lifetime. High-income individuals do precisely this, which explains their higher spending early on. Low-income individuals do the opposite, which ends up costing more later in life. He concludes that a reform of the health care system that encourages use of health care for low-income individuals has positive welfare gains, even when fully accounting for the increase in taxes required to pay for them.

Endogenizing Credit Constraints

The basic Aiyagari model features a reduced-form specification for borrowing constraints (8), and does not model the lenders' problem that gives rise to such constraints. As such, it is silent about potentially interesting variations in borrowing limits across individuals and states of the economy. A number of recent articles attempt to close this gap.

In one of the earliest studies of this kind, Athreya (2002) constructs a general equilibrium model of unsecured household borrowing to quantify the welfare effects of the Bankruptcy Reform Act of 1999 in the United States. In

[30] Palumbo's estimates of medical expenditures are quite a bit smaller than those in De Nardi, French, and Jones (2010), which are largely responsible for the smaller effects he quantifies. De Nardi, French, and Jones (2010) argue that one reason for the discrepancy could be the fact that Palumbo used data from the National Medical Care Expenditure Survey, which, unlike the AHEAD data set, does not contain direct measures of nursing home expenses. He did imputations from a variety of sources, which may be missing the large actual magnitude of such expenses found in the AHEAD data set.

the *pooling* equilibrium of this model (which is what Athreya focuses on), the competitive lending sector charges a higher borrowing rate than the market lending rate to break even (i.e., zero-profit condition), accounting for the fraction of households that will default. This framework allows him to study different policies, such as changing the stringency of means testing as well as eliminating bankruptcy altogether.

In an important article, Chatterjee et al. (2007) build a model of personal default behavior and endogenous borrowing limits. The model features (i) several types of shocks—to earnings, preferences, and liabilities (e.g., hospital and lawsuit bills, which precede a large fraction of defaults in the United States), (ii) a competitive banking sector, and (iii) post-bankruptcy legal treatment of defaulters that mimics the U.S. Chapter 7 bankruptcy code. The main contribution of Chatterjee et al. (2007) is to show that a *separating* equilibrium exists in which banks offer a *menu* of debt contracts to households whose interest rates vary optimally with the level of borrowing to account for the changing default probability. Using a calibrated version of the model, they quantify the separate contributions of earnings, preferences, and liability shocks to debt and default. Chatterjee and Eyigungor (2011) introduce collateralized debt (i.e., mortgage debt) into this framework to examine the causes of the run-up in foreclosures and crash in housing prices after 2007.

Livshits, MacGee, and Tertilt (2007) study a model similar to Chatterjee et al. (2007) in order to quantify the advantages to a "fresh start" bankruptcy system (e.g., U.S. Chapter 7) against a European style system in which debtors cannot fully discharge their debt in bankruptcy. The key tradeoff is that dischargeable debts add insurance against bad shocks, helping to smooth across states, but the inability to commit to future repayment increases interest rates and limits the ability to smooth across time. Their model is quite similar to Chatterjee et al. (2007), except that they model an explicit overlapping generations structure. They calibrate the model to the age-specific bankruptcy rate and debt-to-earnings ratio. For their baseline parameterization, they find that fresh-start bankruptcy is welfare improving, but that result is sensitive to the process for expenditure and income shocks, the shape of the earnings profile, and household size. Livshits, MacGee, and Tertilt (2010) build on this framework to evaluate several theories for the rise in personal bankruptcies since the 1970s. Finally, Glover and Short (2010) use the model of personal bankruptcy to understand the incorporation of entrepreneurs. Incorporation protects the owners' personal assets and their access to credit markets in case of default, but by increasing their likelihood of default, incorporation also implies a risk premium is built into their borrowing rate.

From Bachelor(ette)s to Families

While the framework described above can shed light on some interesting distributional issues (e.g., inequality in consumption, earnings, and wealth), it is completely silent on a crucial source of heterogeneity—the household structure. In reality, individuals marry, divorce, have kids, and make their decisions regarding consumption, savings, labor supply, and so on jointly with these other life choices. For many economic and policy questions, the interaction between these domestic decisions and economic choices in an incomplete markets world can have a first-order effect on the answers we get. Just to give a few examples, consider these facts: Men and women are well-known to exhibit different labor supply elasticities; the tax treatment of income varies depending on whether an individual is single, married, and whether he/she has kids, etc.; the trends in the labor market participation rate in the United States since the 1960s have been markedly different for single and married women; the fractions of individuals who are married and divorced have changed significantly, again since the 1960s; and so on.

A burgeoning literature works to bring a richer household structure into macroeconomics. For example, in an influential article, Greenwood, Seshadri, and Yorukoglu (2005) study the role of household technologies (the widespread availability of washing machines, vacuum cleaners, refrigerators, etc.) in leading women into the labor market. Greenwood and Guner (2009) extend the analysis to study the marriage and divorce patterns since World War II. Jones, Manuelli, and McGrattan (2003) explore the role of the closing gender wage gap for married women's rising labor supply. Knowles (2007) argues that the working hours of men are too long when viewed through the lens of a unitary model of the household in which the average wage of females rises as in the data. He shows that introducing bargaining between spouses into the model reconciles it with the data. Guner, Kaygusuz, and Ventura (2010) study the effects of potential reforms in the U.S. tax system in a model of families with children and an extensive margin for female labor supply. Guvenen and Rendall (2011) study the insurance role of education for women against divorce risk and the joint evolution of education trends with those in marriage and divorce.

4. INEQUALITY IN CONSUMPTION, WEALTH, AND EARNINGS

A major use of heterogeneous-agent models is to study inequality or dispersion in key economic outcomes, most notably in consumption, earnings, and wealth. The Aiyagari model—as well as its aggregate-shock augmented version, the Krusell-Smith model presented in the next section—takes earnings dispersion to be exogenous and makes predictions about inequality in consumption and wealth. The bulk of the incomplete markets literature follows

this lead in their analysis. Some studies introduce an endogenous labor supply choice and instead specify the wage process to be exogenous, delivering earnings dispersion as an endogenous outcome (Pijoan-Mas [2006], Domeij and Floden [2006], Heathcote, Storesletten, and Violante [2008], among others). While this is a useful step forward, a lot of the dispersion in earnings before age 55 is because of wages and not hours, so the assumption of an exogenous wage process still leaves quite a bit to be understood. Other strands of the literature attempt to close this gap by writing models that also generate wage dispersion as an endogenous outcome in the model—for example, because of human capital accumulation (e.g., Guvenen and Kuruscu [2010, forthcoming], and Huggett, Ventura, and Yaron [2011]) or because of search frictions.[31]

Consumption Inequality

Two different dimensions of consumption inequality have received attention in the literature. The first one concerns how much within-cohort consumption inequality increases over the life cycle. The different views on this question have been summarized in Section 2.[32] The second one concerns whether, and by how much, (overall) consumption inequality has risen in the United States since the 1970s, a question whose urgency was raised by the substantial rise in wage inequality during the same time. In one of the earliest articles on this topic, Cutler and Katz (1992) use data from the 1980s on U.S. households from the CE and find that the evolution of consumption inequality closely tracks the rise in wage inequality during the same time. This finding serves as a rejection of earlier claims in the literature (e.g., Jencks 1984) that the rise of means-tested in-kind transfers starting in the 1970s had improved the material well-being of low-income households relative to what would be judged by their income statistics.

Interest in this question was reignited more recently by a thought-provoking article by Krueger and Perri (2006), who conclude from an analysis of CE data that, from 1980–2003, *within-group* income inequality increased substantially more than within-group consumption inequality (in contrast, they find

[31] The search literature is very large with many interesting models to cover. I do not discuss these models here because I cannot do justice to this extensive body of work in this limited space. For an excellent survey, see Rogerson, Shimer, and Wright (2005). Note, however, that as Hornstein, Krusell, and Violante (2011) show, search models have trouble generating the magnitudes of wage dispersion we observe in the data.

[32] Another recent article of interest is Aguiar and Hurst (2008), who examine the life-cycle mean and variance profiles of the subcomponents of consumption—housing, utility bills, clothing, food at home, food away from home, etc. They show rich patterns that vary across categories, whereby the variance rises monotonically for some categories, while being hump-shaped for others, and yet declining monotonically for some others. The same patterns are observed for the mean profile. These disaggregated facts provide more food for thought to researchers.

that *between-group* income and consumption inequality tracked each other).[33] They then propose an explanation based on the premise that the development of financial services in the U.S. economy has helped households smooth consumption fluctuations relative to income variation.

To investigate this story, they apply a model of endogenous debt constraints as in Kehoe and Levine (1993). In this class of models, what is central is not the ability of households to pay back their debt, but rather it is their incentive or willingness to pay back. To give the right incentives, lenders can punish a borrower that defaults, for example, by banning her from financial markets forever (autarky). However, if the individual borrows too much or if autarky is not sufficiently costly, it may still make sense to default in certain states of the world. Thus, given the parameters of the economic environment, lenders will compute the optimal state-contingent debt limit, which will ensure that the borrower never defaults in equilibrium. Krueger and Perri (2006) notice that if income shocks are really volatile, then autarky is a really bad outcome, giving borrowers less incentive to default. Lenders who know this, in turn, are more willing to lend, which endogenously loosens the borrowing constraints. This view of the last 30 years therefore holds that the rise in the volatility of income shocks gave rise to the development of financial markets (more generous lending), which in turn led to a smaller rise in consumption inequality.[34]

Heathcote, Storesletten, and Violante (2007) argue that the small rise in consumption inequality can be explained simply if the rise in income shocks has been of a more transitory nature, since such shocks are easier to smooth through self-insurance. Indeed, Blundell and Preston (1998) earlier made the same observation and concluded that in the 1980s the rise in income shock variance was mostly permanent in nature (as evidenced by the observation that income and consumption inequality grew together), whereas in the 1990s it was mostly transitory given that the opposite was true. Heathcote, Storesletten, and Violante (2007) calibrate a fully specified model and show that it can go a long way toward explaining the observed trends in consumption inequality. One point to keep in mind is that these articles take as given that the volatility of income shocks rose during this period, a conclusion that is subject to uncertainty in light of the new evidence discussed above.

---

[33] Attanasio, Battistin, and Ichimura (2007) question the use of the CE interview survey and argue that some expenditure items are poorly measured in the survey relative to another component of CE, called the diary survey. They propose an optimal way of combining the two survey data and find that consumption inequality, especially in the 1990s has increased more than what is revealed by the interview survey alone.

[34] Aguiar and Bils (2011) take a different approach and construct a measure of CE consumption by using data on income and (self-reported) savings rate by households. They argue that consumption inequality tracked income inequality closely in the past 30 years. Although this is still preliminary work, the article raises some interesting challenges.

Before concluding, a word of caution about measurement. The appropriate price deflator for consumption may have trended differently for households in different parts of the income distribution (i.e., the "Walmart effect" at the lower end). To the extent that this effect is real, the measured trend in consumption inequality could be overstating the actual rise in the dispersion of material well-being. This issue still deserves a fuller exploration.

Wealth Inequality

The main question about wealth inequality is a cross-sectional one: Why do we observe such enormous disparities in wealth, with a Gini coefficient of about 0.80 for net worth and a Gini exceeding 0.90 for financial wealth?

Economists have developed several models that can generate highly skewed wealth distributions (see, for example, Huggett [1996], Krusell and Smith [1998], Quadrini [2000], Castañeda, Díaz-Giménez, and Ríos-Rull [2003], Guvenen [2006], and Cagetti and De Nardi [2006]). These models typically use one (or more) of three mechanisms to produce this inequality: (1) dispersion in luck in the form of large and persistent shocks to labor productivity: the rich are luckier than the poor; (2) dispersion in patience or thriftiness: the rich save more than the poor; and (3) dispersion in rates of return: the rich face higher asset returns than the poor. This subsection describes a baseline model and variations of it that incorporate various combinations of the three main mechanisms that economists have used to generate substantial inequality in general equilibrium models.[35]

Dispersion in Luck

Huggett (1996) asks how much progress can be made toward understanding wealth inequality using (i) a standard life-cycle model with (ii) Markovian idiosyncratic shocks, (iii) uncertain lifetimes, and (iv) a Social Security system. He finds that although the model can match the Gini coefficient for wealth in the United States, this comes from low-income households holding too little wealth, rather than the extreme concentration of wealth at the top in the U.S. economy. Moreover, whereas in the U.S. data the dispersion of wealth within each cohort is nearly as large as the dispersion across cohorts, the model understates the former significantly.

Castañeda, Díaz-Giménez, and Ríos-Rull (2003) study an enriched model that combines elements of Aiyagari (1994) and Huggett (1996). Specifically, the model (i) has a Social Security system, (ii) has perfectly altruistic bequests,

[35] Some of the models discussed in this section contain aggregate shocks in addition to idiosyncratic ones. While aggregate shocks raise some technical issues that will be addressed in the next section, they pose no problems for the exposition in this section.

(iii) allows for intergenerational correlation of earnings ability, (iv) has a progressive labor and estate tax system as in the United States, and (v) allows a labor supply decision. As for the stochastic process for earnings, they do not calibrate its properties based on microeconometric evidence on income dynamics as is commonly done, but rather they choose its features (the 4×4 transition matrix and four states of a Markov process) so that the model matches the cross-sectional distribution of earnings and wealth. To match the extreme concentration of wealth at the upper tail, this calibration procedure implies that individuals must receive a large positive shock (about 1,060 times the median income level) with a small probability. This high income level is also very fleeting—it lasts for about five years—which leads these high income individuals to save substantially (for consumption smoothing) and results in high wealth inequality.

Dispersion in Patience

Laitner's (1992, 2002) original insight was that wealth inequality could result from a combination of: (1) random heterogeneity in lifetime incomes across generations, and (2) altruistic bequests, which are constrained to be non-negative. Each newly born consumer in Laitner's model receives a permanent shock to his lifetime income and, unlike in the Aiyagari model, faces no further shocks to income during his lifetime. In essence, in Laitner's model only households that earn higher than average lifetime income want to transfer some amount to their offspring, who are not likely to be as fortunate. This altruistic motive makes these households effectively more thrifty (compared to those that earn below average income) since they also care about future utility. Thus, even small differences in lifetime income can result in large differences in savings rates—a fact empirically documented by Carroll (2000)—and hence in wealth accumulation.

The stochastic-beta model of Krusell and Smith (1998) is a variation on this idea in a dynastic framework, where heterogeneity in thrift (i.e., in the time-discount rate) is imposed exogenously.[36] Being more parsimonious, the stochastic-beta model also allows for the introduction of aggregate shocks. Krusell and Smith show that even small differences in the time discount factor that are sufficiently persistent are sufficient to generate the extreme skewness of the U.S. wealth distribution. The intuition for this result will be discussed in a moment.

[36] Notice that in this article I use δ to denote the time discount factor and β was used to denote the income growth rate. I will continue with this convention, except when I specifically refer to the Krusell-Smith model, which has come to be known as a stochastic-beta model.

Dispersion in Rates of Return

Guvenen (2006) introduces return differentials into a standard stochastic growth model (i.e., in which consumers have identical, time-invariant discount factors and idiosyncratic shocks do not exist). He allows all households to trade in a risk-free bond, but restricts one group of agents from accumulating capital. Quadrini (2000) and Cagetti and De Nardi (2006) study models of inequality with entrepreneurs and workers, which can also generate skewed wealth distributions. The mechanisms have similar flavors: Agents who face higher returns end up accumulating a substantial amount of wealth.

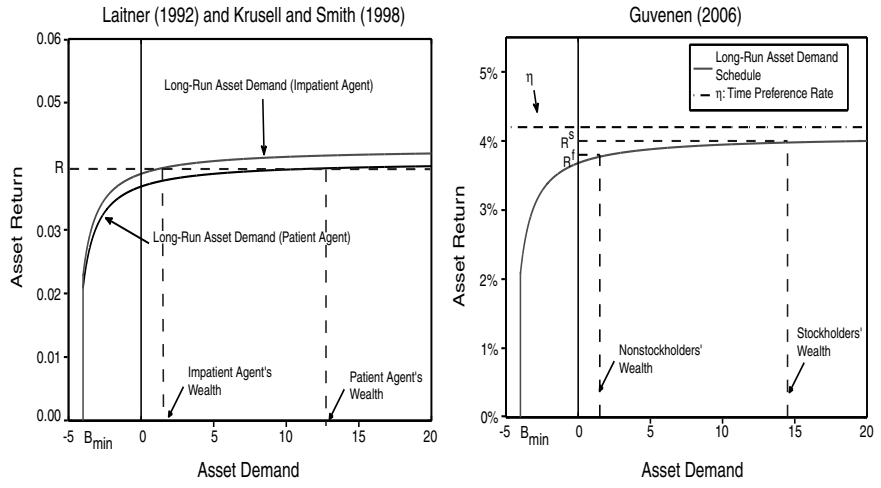
The basic mechanism in Guvenen (2006) can be described as follows. Nonstockholders have a precautionary demand for wealth (bonds), but the only way they can save is if stockholders are willing to borrow. In contrast, stockholders have access to capital accumulation, so they could smooth consumption even if the bond market was completely shut down. Furthermore, nonstockholders' asset demand is even more inelastic because they are assumed to have a lower elasticity of intertemporal substitution (consistent with empirical evidence) and therefore have a strong desire for consumption smoothing. Therefore, trading bonds for consumption smoothing is more important for nonstockholders than it is for stockholders. As a result, stockholders will only trade in the bond market if they can borrow at a low interest rate. This low interest rate in turn dampens nonstockholders' demand for savings further, and they end up with little wealth in equilibrium (and stockholders end up borrowing very little). Guvenen (2009b) shows that a calibrated version of this model easily generates the extremely skewed distribution of the relative wealth of stockholders to nonstockholders in the U.S. data.

Can We Tell Them Apart?

The determination of wealth inequality in the three models discussed so far can be explained using variations of a diagram used by Aiyagari (1994). The left panel of Figure 2 shows how wealth inequality is determined in Laitner's model and, given their close relationship, in the Krusell-Smith model. The top solid curve originating from "$-B_{\min}$" plots the long-run asset demand schedule for the impatient agent; the bottom curve is for the patient agent. A well-known feature of incomplete markets models is that the asset demand schedule is very flat for values of returns that are close to the time preference rate, η (so $\delta \equiv 1/(1+\eta)$). Thus, both types of individuals' demand schedules asymptote to their respective time preference rates (with $\eta_{\text{patient}} < \eta_{\text{impatient}}$).[37] If the equilibrium return (which must be lower than η_{patient} for an equilibrium to exist) is sufficiently close to η_{patient}, the high sensitivity of asset demands

[37] See, for example, Aiyagari (1994) and references therein. This result also holds when asset returns are stochastic (Chamberlain and Wilson 2000).

Figure 2 Determination of Wealth Inequality in Various Models



to interest rates will generate substantial wealth inequality between the two types of agents.

Similarly, the right panel shows the mechanism in the limited participation model, which has a similar flavor. For simplicity, let us focus on the case where stockholders and nonstockholders have the same preferences and face the same portfolio constraints. We have $\eta > R^S > R^f$. Again, given the sensitivity of asset demand to returns near η, even a small equity premium generates substantial wealth inequality. It should be stressed, however, that a large wealth inequality is not a foregone conclusion in any of these models. If returns were too low relative to η, individuals would be on the steeper part of their demand curves, which could result in smaller differences in wealth holdings.

While the mechanics described here may appear quite similar for the three models, their substantive implications differ in crucial ways. For example, consider the effect of eliminating aggregate shocks from all three models. In Guvenen (2006), there will be no equity premium without aggregate shocks and, consequently, no wealth inequality. In Krusell and Smith (1998), wealth inequality will increase as the patient agent holds more of the aggregate wealth (and would own all the wealth if there were no idiosyncratic shocks). In Laitner (1992), wealth inequality will remain unchanged, since it is created by idiosyncratic lifetime income risk. These dramatically different implications suggest that one can devise methods to bring empirical evidence to bear on the relevance of these different mechanisms.

Cagetti and De Nardi (2006)

Cagetti and De Nardi (2006) introduce heterogeneity across individuals in both work and entrepreneurial ability. Entrepreneurial firms operate decreasing returns to scale production functions, and higher entrepreneurial ability implies a higher optimal scale. Because debt contracts are not perfectly enforceable due to limited commitment, business owners need to put up some of their assets as collateral, a portion of which would be confiscated in case of default. Thus, entrepreneurs with very promising projects have more to lose from default, which induces them to save more for collateral, borrow more against it, and reach their larger optimal scale. The model is able to generate the extreme concentration of wealth at the top of the distribution (among households, many of whom are entrepreneurs).

Although this model differs from the limited participation framework in many important ways, the differential returns to saving is a critical element for generating wealth inequality in both models. This link could be important because many individuals in the top 1 percent and 5 percent of the U.S. wealth distribution hold significant amounts of stocks but are not entrepreneurs (hold no managerial roles), which the Cagetti/De Nardi model misses. The opposite is also true: Many very rich entrepreneurs are not stockholders (outside of their own company), which does not fit well with Guvenen's model (see Heaton and Lucas [2000] on the empirical facts about wealthy entrepreneurs and stockholders). The view that perhaps the very high wealth holdings of these individuals is driven by the higher returns that they enjoy—either as a stockholder or as an entrepreneur—can offer a unified theory of savings rate differences.

Wage and Earnings Inequality

Because the consumption-savings decision is the cornerstone of the incomplete markets literature, virtually every model has implications for consumption and wealth inequality. The same is not true for earnings inequality. Many models assume that labor supply is inelastic and the stochastic process for wages is exogenous, making the implications for wage and earnings inequality to be mechanical reflections of the assumptions of the model. Even if labor supply is assumed to be endogenous, many properties of the earnings distribution (exceptions noted below) mimic those of the wage distribution. For things to get more interesting, it is the latter that needs to be endogenized.

In this subsection, I first review the empirical facts regarding wage inequality—both over the life cycle and over time. These facts are useful for practitioners since they are commonly used as exogenous inputs into incomplete markets models. Unless specifically mentioned, all the facts discussed here pertain to male workers, because the bulk of the existing work is

available consistently for this group.[38] Second, I discuss models that attempt to endogenize wages and explain the reported facts and trends.

Inequality Over the Life Cycle

The main facts about the evolution of (within-cohort) earnings inequality over the life cycle were first documented by Deaton and Paxson (1994) and shown in the left panel of Figure 1. The same exercise has been repeated by numerous authors using different data sets or time periods (among others, Storesletten, Telmer, and Yaron [2004a], Guvenen [2009a], Heathcote, Perri, and Violante [2010], and Kaplan [2010]). While the magnitudes differ somewhat, the basic fact that wage and earnings inequality rise substantially over the life cycle is well-established.

One view is that this fact does not require an elaborate explanation, because wages follow a very persistent, perhaps permanent, stochastic process as implied by the RIP model. Thus, the rising life-cycle inequality is simply a reflection of the accumulation of such shocks, which drive up the variance of log wages in a linear fashion (in the case of permanent shocks). I will continue to refer to this view as the RIP model because of its emphasis on persistent "shocks."[39]

An alternative perspective, which has received attention more recently, emphasizes systematic factors—heterogeneity as opposed to random shocks. This view is essentially in the same spirit as the HIP model of the previous section. But it goes one step further by *endogenizing* the wage distribution based on the human capital framework of Becker (1964) and Ben-Porath (1967), among others. In an influential article, Huggett, Ventura, and Yaron (2006) study the distributional implications of the standard Ben-Porath (1967) model by asking about the types of heterogeneity that one needs to introduce to generate patterns consistent with the U.S. data. They find that too much heterogeneity in initial human capital levels results in the counterfactual implication that wage inequality *should fall* over the life cycle. In contrast, heterogeneity in learning ability generates a rise in wage inequality consistent with the data. A key implication of this finding is that the rise in wage inequality can be generated without appealing to idiosyncratic shocks of any kind. Instead, it is the systematic fanning out of wage profiles, resulting from different investment rates, that generates rising inequality over the life cycle. Guvenen, Kuruscu, and Ozkan (2009) and Huggett, Ventura, and Yaron (2011) introduce

[38] Some of the empirical trends discussed also apply to women, while others do not. See Autor, Katz, and Kearney (2008) for a comparative look at wage trends for males and females during the period.

[39] Of course, one can write deeper economic models that generate the observation that wages follow a random walk process, such as the learning model of Jovanovic (1979) in a search and matching environment, or the optimal contracts in the limited commitment model of Harris and Holmstrom (1982).

idiosyncratic shocks into the Ben-Porath framework. Both articles find that heterogeneous growth rates continue to play the dominant role for the rise in wage inequality. The Ben-Porath formulation is also central for wage determination in Heckman, Lochner, and Taber (1998) and Kitao, Ljungqvist, and Sargent (2008).

Inequality Trends Over Time

A well-documented empirical trend since the 1970s is the rise in wage inequality among male workers in the United States. This trend has been especially prominent above the median of the wage distribution: For example, the log wage differential between the 90th and the 50th percentiles has been expanding in a secular fashion for the past four decades. The changes at the bottom have been more episodic, with the log 50-10 wage differential strongly expanding until the late 1980s and then closing subsequently (see Autor, Katz, and Kearney [2008] for a detailed review of the evidence). Acemoglu (2002) contains an extensive summary of several related wage trends, as well as a review of proposed explanations. Here I only discuss the subset of articles that are more closely relevant for the incomplete markets macroliterature.

Larger Shocks or Increasing Heterogeneity?

Economists' interpretations of the rise in wage inequality over the life cycle and over time are intimately related. The RIP view that was motivated by analyses of life-cycle wages was dominant in the 1980s and 1990s, so it was natural for economists to interpret the rise in wage inequality *over time*, through the same lens. Starting with Gottschalk and Moffitt (1994) and Moffitt and Gottschalk (1995), this trend has been broadly interpreted as reflecting a rise in the variances of idiosyncratic shocks, either permanent or transitory (Meghir and Pistaferri 2004; Heathcote, Storesletten, and Violante 2008; etc.). This approach remains the dominant way to calibrate economic models that investigate changes in economic outcomes from the 1970s to date.

However, some recent articles have documented new evidence that seems hard to reconcile with the RIP view. The first group of articles revisits the econometric analyses of wage and earnings data. Among these, Sabelhaus and Song (2009, 2010) use panel data from Social Security records covering millions of American workers, in contrast to the long list of previous studies that use survey data (e.g., the PSID).[40] While this data set has the potential drawback of under-reporting (because it is based on income reported to the Internal Revenue Service), it has three important advantages: (i) a much larger sample size (on the order of 50+ million observations, compared to at

[40] These include, among others, Meghir and Pistaferri (2004), Dynan, Elmendorf, and Sichel (2007), Heathcote, Storesletten, and Violante (2008), and Shin and Solon (2011).

most 50,000 in the PSID), (ii) no survey response error, and (iii) no attrition. Sabelhaus and Song find that the volatility of annual earnings growth increased during the 1970s, but that it *declined* monotonically during the 1980s and 1990s. Furthermore, applying the standard permanent-transitory decomposition as in Moffitt and Gottschalk (1995) and Meghir and Pistaferri (2004) reveals that permanent shock variances were stable and transitory shocks became *smaller* from 1980 into the 2000s. A separate study conducted by the Congressional Budget Office (2008), also using wage earnings from Social Security records from 1980–2003, reached the same conclusion.[41] Finally, Kopczuk, Saez, and Song (2010) document (also using Social Security data) that both long-run and short-run mobility have stayed remarkably stable from the 1960s into the 2000s. But this finding seems difficult to reconcile with Moffitt and Gottschalk (1995) and the subsequent literature that found permanent and transitory shock variances to have risen in different subperiods from 1970 to the 2000s. If true, the latter would result in fluctuations in mobility patterns over these subperiods, which is not borne out in Kopczuk, Saez, and Song's (2010) analysis.

Another piece of evidence from income data is offered by Haider (2001), who estimates a stochastic process for wages similar to the one in Moffitt and Gottschalk (1995) and others, but with one key difference. He allows for individual-specific wage growth rates (HIP) and he also allows for the dispersion of growth rates to vary over time. The stochastic component is specified as an ARMA(1,1). With this more flexible specification, he finds no evidence of a rise in the variance of income shocks after the 1970s, but instead finds a large increase in the dispersion of systematic wage growth rates.

A second strand of the literature studies the trends in labor market flows in the United States. These articles do not find any evidence of rising job instability or churning, which one might expect to see in conjunction with larger idiosyncratic shocks. In contrast, these studies document an across-the-board moderation in labor market flows. For example, Gottschalk and Moffitt (1999) focus on male workers between the ages of 20 and 62 and conclude their analysis as follows:

> [W]e believe that a consistent picture is emerging on changes in job stability and job security in the 1980s and 1990s. Job instability does

[41] Sabelhaus-Song attribute the reason why some earlier studies found rising variances of wage shocks (e.g., Moffitt and Gottschalk 2008) to the inclusion of individuals with self-employment income and those who earn less than the Social Security minimum. Even though there are few of these households, Sabelhaus and Song show that they make a big difference in the computed statistics. Similarly, Shin and Solon (2011, 978–80) use PSID data and also do not find a trend in the volatility of wage earnings changes during the 1980s and 1990s. They argue that the increasing volatility found in earlier studies, such as Dynan, Elmendorf, and Sichel (2007), seems to be coming from the inclusion of some auxiliary types of income (from business, farming, etc.) whose treatment has been inconsistent in the PSID over the years.

> not seem to have increased, and the consequences of separating from an employer do not seem to have worsened.[42]

Shimer (2005, 2007) and Davis et al. (2010) extend this analysis to cover the 2000s and use a variety of data sets to reach the same conclusion. Further, both articles show that expanding the sample of individuals to include women and younger workers shows a *declining* trend in labor market flows and an *increase* in job security.

Taking Stock

To summarize, the seeming consensus of the 1990s—that rising wage inequality was driven by an increase in idiosyncratic shock variances—is being challenged by a variety of new evidence, some of which comes from data sets many orders of magnitude larger than the surveys used in previous analyses. In addition, the evidence from labor market flows described above—while perhaps more indirect—raises questions about the sources of larger idiosyncratic shocks in a period where labor market transitions seem to have moderated. Although, it would be premature to conclude that the alternative view is the correct one—more evidence is needed to reach a definitive conclusion. Having said that, if this alternative view *is* true, and income shock variances have not increased, this new "fact" would require economists to rethink a variety of explanations put forward for various trends, which assumed a rise in shock variances during this period.

5. HETEROGENEITY WITH AGGREGATE SHOCKS

Krusell and Smith (1998) add two elements to the basic Aiyagari framework. First, they introduce aggregate technology shocks. Second, they assume that the cumulative discount factor at time t (which was assumed to be δ^t before), now follows the stochastic process $\delta_t = \tilde{\delta}\delta_{t-1}$, where $\tilde{\delta}$ is a finite-state Markov chain. The stochastic evolution of the discount factors within a dynasty captures some elements of an explicit overlapping-generations structure with altruism and less-than-perfect correlation in genes between parents and children, as in Laitner (1992). With this interpretation in mind, the evolution of $\tilde{\delta}$ is calibrated so that the average duration of any particular value of the discount factor is equal to the lifetime of a generation. (Krusell and Smith [1997] study a version of this model where consumers are allowed to hold a risk-free bond in addition to capital.)

The specifics of the model are as follows. There are two types of shocks: (i) idiosyncratic employment status: $(\epsilon_e, \epsilon_u) \equiv$ (employed, unemployed);

[42] They also say, "Almost all studies based on the various Current Population Surveys (CPS) supplements...show little change in the overall separation rates through the early 1990s."

and (ii) aggregate productivity: $(z_g, z_b) \equiv$ (expansion, recession). Employment status and aggregate productivity jointly evolve as a first-order Markov process. Assume that ϵ is i.i.d. conditional on z, so the fraction of employed workers (and hence l) only depends on z. Competitive markets imply

$$w(K, L, z) = (1-\alpha) z (K/L)^{-\alpha}, \text{ and } r(K, L, z) = \alpha z (K/L)^{\alpha-1}. \quad (12)$$

Finally, the entire wealth distribution, which I denote with Γ is a state variable for this model, and let $\Gamma' = H(\Gamma, z; z')$ denote its endogenous transition function (or law of motion).

Krusell-Smith Algorithm

A key equilibrium object in this class of models is the law of motion, H. In principle, computing this object is a formidable task since the distribution of wealth is infinite-dimensional. Krusell and Smith (1997, 1998) show, however, that this class of models, when reasonably parameterized, exhibits "approximate aggregation": Loosely speaking, to predict future prices consumers need to forecast only a small set of statistics of the wealth distribution rather than the entire distribution itself. This result makes it possible to use numerical methods to analyze this class of models. Another key feature of the Krusell-Smith algorithm is that it solves the model by simulating it. Specifically, the basic version of the algorithm works as follows:

1. Approximate Γ with a finite number (I) of moments. (H reduces to a function mapping the I moments today into the I moments tomorrow depending on z today.)

   (a) We will start by selecting one moment—the mean—so $I = 1$.[43]

2. Select a family of functions for H. I will choose a log-linear function following Krusell and Smith.

$$\begin{aligned} V(k, \epsilon; \Gamma, z) &= \max_{c, k'} \left[U(c) + \delta E \left[V\left(k', \epsilon'; \Gamma', z'\right) \mid z, \epsilon \right] \right] \\ c + k' &= w(K, L, z) \times l \times \epsilon + r(K, L, z) \times k, \quad k' \geq 0 \\ \log K' &= a_0 + a_1 \log K \quad \text{for } z = z_b \\ \log K' &= b_0 + b_1 \log K \quad \text{for } z = z_g \end{aligned}$$

3. Make an (educated) initial guess about $(a_0, a_1, b_0, b_1) \Longrightarrow$ yields initial guess for H_0. Make also an initial guess for Γ_0.

[43] When we add more moments, we do not have to proceed as mean, variance, skewness, and so on. We can include, say, the wealth holdings of the top 10 percent of population, mean-to-median wealth ratio, etc.

4. Solve the consumer's dynamic program. Using only the resulting decision rules, simulate $\{k_{n,t}\}_{n=1,t=1}^{N,T}$ for (N, T) large.

5. Update H by estimating (where $\tilde{K} = \frac{1}{N}\sum_{n=1}^{N} k_n$):

$$\begin{aligned} \log \tilde{K}' &= a_0 + a_1 \log \tilde{K} \quad \text{for } z = z_b \\ \log \tilde{K}' &= b_0 + b_1 \log \tilde{K} \quad \text{for } z = z_g \end{aligned}$$

6. Iterate on 4–5, until the R^2 of this regression is "sufficiently high" and the forecast variance is "small."

   (a) If accuracy remains insufficient, go back to step 1 and increase I.

Details

Educated Initial Guess

As with many numerical methods, a good initial guess is critical. More often than not, the success or failure of a given algorithm will depend on the initial guess. One idea (used by Krusell and Smith) is to first solve a standard representative-agent real business cycle (RBC) model with the same parameterization. Then estimate the coefficients (a_0, a_1, b_0, b_1) using capital series simulated from this model to obtain an initial guess for step 1 above.[44] More generally, a good initial guess can often be obtained by solving a simplified version of the full model. Sometimes this simplification involves ignoring certain constraints, sometimes by shutting down certain shocks, and so on.

Discretizing an AR(1) Process

Oftentimes, the exogenous driving force in incomplete markets models is assumed to be generated by an AR(1) process, which is discretized and converted into a Markov chain. One popular method for discretization is described in Aiyagari (1993) and has been used extensively in the literature. However, an alternative method by Rouwenhorst (1995) (and which received far less attention until recently) is far superior in the quality of the approximation that it provides, especially when the process is very persistent, which is often the case. Moreover, it is very easy to implement. Kopecky and Suen (2010) and Galindev and Lkhagvasuren (2010) provide comprehensive comparisons of different discretization methods, which reveal the general superiority of

[44] Can't we update H without simulating? Yes, we can. Den Haan and Rendahl (2009) propose a method where they use the policy functions for capital holdings and integrate them over distribution $\Lambda(k, \epsilon)$ of households across capital and employment status: $K' = H_j(K, z) = \int k'_j(k, \epsilon; \Gamma, z) d\Lambda(k, \epsilon)$. This works well when the decision rules are parameterized in a particular way. See den Haan and Rendahl (2009).

Rouwenhorst's (1995) method. They also show how this method can be extended to discretize more general processes.

Non-Trivial Equilibrium Pricing Function

One simplifying feature of Krusell and Smith (1998) is that equilibrium prices (wages and interest rates) are determined trivially by the marginal product conditions (12). Thus, they depend only on the aggregate capital stock and not on its distribution. Some models do not have this structure—instead pricing functions must be determined by equilibrium conditions—such as market-clearing or zero-profit conditions—that explicitly depend on the wealth distribution. This would be the case, for example, if a household bond is traded in the economy. Its price must be solved for using a market-clearing condition, which is a challenging task. Moreover, if there is an additional asset, such as a stock, two prices must be determined simultaneously, and this must be done in such a way that avoids providing arbitrage opportunities—along the iterations of the solution algorithm. Otherwise, individuals' portfolio choices will go haywire (in an attempt to take advantage of perceived arbitrage), wreaking havoc with the solution algorithm. Krusell and Smith (1997) solve such a model and propose an algorithm to tackle these issues. I refer the interested reader to that article for details.

Checking for Accuracy of Solution

Many studies with aggregate fluctuations and heterogeneity use two simple criteria to asses the accuracy of the law of motion in their limited information approximation. If agents are using the law of motion

$$\log K' = \alpha_1 + \alpha_2 z + \alpha_3 \log K + u, \tag{13}$$

a perfectly solved model should find $u = 0$. Thus, practitioners will continue to solve the model until either the R^2 of this regression is larger than some minimum or σ_u falls below some threshold (step 6 in the algorithm above).

However, one should *not* rely solely on R^2 and σ_u. There are at least three reasons for this (den Haan 2010). First, both measures average over all periods of the simulation. Thus, infrequent but large deviations from the forecast rule can be hidden in the average. These errors may be very important to agents and their decision rule. For example, the threat of a very large recession may increase buffer stock saving, but the approximation may understate the movement of capital in such a case. Second, and more importantly, these statistics only measure one-step-ahead forecasts. The regression only considers the dynamics from one period to the next, so errors are only the deviations between the actual next-period capital and the expected amount in the next period. This misses potentially large deviations between the *long-term* forecast for capital and its actual level. Aware of this possibility, Krusell and Smith (1998) also

check the R^2 for forecasting prices 25 years ahead (100 model periods) and find it to be extremely high as well! (They also check the maximum error in long-term forecasts, which is very small.) Third, R^2 is scaled by the left-hand side of the regression. An alternative is to check the R^2 of

$$\log K' - \log K = \alpha_1 + \alpha_2 z + (\alpha_3 - 1) \log K.$$

As a particularly dramatic demonstration, den Haan (2010) uses a savings decision rule that solves the Krusell and Smith (1998) model, simulates it for T periods, and estimates a law of motion in the form of equation (13). He then manipulates α_1, α_3 such that $T^{-1} \sum u_t = 0$ but the R^2 falls from 0.9999 to 0.999 and then 0.99. This has economically meaningful consequences: The time series standard deviation of the capital stock simulated from the perturbed versions of equation (13) falls to 70 percent and then 46 percent of the true figure.

Finally, den Haan (2010) proposes a useful test that begins with the approximated law of motion,

$$\log K' = \hat{\alpha}_1 + \hat{\alpha}_2 z + \hat{\alpha}_3 \log K + u, \tag{14}$$

to generate a sequence of realizations of $\left\{\tilde{K}_{t+1}\right\}_{t=0}^{T}$ and then compares these to the sequence generated by aggregating from decision rules, the true law of motion. Because $\left\{\tilde{K}_{t+1}\right\}_{t=0}^{T}$ is obtained by repeatedly applying equation (14) starting from $\tilde{K}_0$, errors can accumulate. This is important because, in the true model, today's choices depend on expectations about the future state, which in turn depends on the future's future expectations and so errors cascade. To systematically compare $\tilde{K}_t$ to K_t, den Haan proposes an "essential accuracy plot." For a sequence of shocks (*not* those originally used when estimating α to solve the model), generate a sequence of $\tilde{K}_t$ and K_t. One can then compare moments of the two simulated sequences. The "main focus" of the accuracy test is the errors calculated by $\tilde{u}_t = \left|\log \tilde{K}_t - \log K_t\right|$, whose maximum should be made close to zero.

Prices are More Sensitive than Quantities

The accuracy of the numerical solution becomes an even more critical issue if the main focus of analysis is (asset) prices rather than quantities. This is because prices are much more sensitive to approximation errors (see, e.g., Christiano and Fisher [2000] and Judd and Guu [2001]). The results in Christiano and Fisher (2000) are especially striking. These authors compare a variety of different implementations of the "parameterized expectations" method and report the approximation errors resulting from each. For the standard deviation of output, consumption, and investment (i.e., "quantities"), the approximation errors range from less than 0.1 percent of the true value to 1

percent to 2 percent in some cases. For the stock and bond return and the equity premium, the errors regularly exceed 50 percent and are greater than 100 percent in several cases. The bottom line is that the computation of asset prices requires *extra* care.

Pros and Cons

An important feature of the Krusell-Smith method is that it is a "local" solution around the stationary recursive equilibrium. In other words, this method relies on simulating a very long time series of data (e.g., capital series) and making sure that after this path has converged to the ergodic set, the predictions of agents are accurate for behavior inside that set. This has some advantages and some disadvantages. One advantage is the efficiency gain compared to solving a full recursive equilibrium, which enforces the equilibrium conditions at every point of a somewhat arbitrary grid, regardless of whether or not a given state is ever visited in the stationary equilibrium.

One disadvantage is that if you take a larger deviation—say by setting K_0 to a value well below the steady-state value—your "equilibrium functions" are likely to be inaccurate, and the behavior of the solution may differ significantly from the true solution. Why should we care about this? Suppose you solve your model and then want to study a policy experiment where you eliminate taxes on savings. You would need to write a separate program from the "transition" between the two stationary equilibria. Instead, if you solve for the full recursive equilibrium over a grid that contains both the initial and final steady states, you would not need to do this. However, solving for the full equilibrium is often much harder and, therefore, is often overkill.

An Alternative to Krusell-Smith: Tracking History of Shocks

Some models have few exogenous state variables, but a large number of endogenous states. In such cases, using a formulation that keeps track of all these state variables can make the numerical solution extremely costly or even infeasible. An alternative method begins with the straightforward observation that all current endogenous state variables are nothing more than functions of the infinite history of exogenous shocks. So, one could replace these endogenous states with the infinite history of exogenous states. Moreover, many models turn out to have "limited memory" in the sense that only the recent history of shocks matters in a quantitatively significant way, allowing us to only track a truncated history of exogenous states. The first implementation of this idea I have been able to find is in Veracierto (1997), who studied a model with plant-level investment irreversibilities, which give rise to S-s type policies. He showed that it is more practical to track a short history of the S-s thresholds instead of the current-period endogenous state variables.

As another example, consider an equilibrium model of the housing market where the only exogenous state is the interest rate, which evolves as a Markov process. Depending on the precise model structure, the individual endogenous state variables can include the mortgage debt outstanding, the time-to-maturity of the mortgage contract, etc., and the aggregate endogenous state can include the entire distribution of agents over the individual states. This is potentially an enormously large state space! Arslan (2011) successfully solves a model of this sort with realistic fixed-rate mortgage contracts, a life-cycle structure, and stochastic interest rates, using four lags of interest rates. Other recent examples that employ this basic approach include Chien and Lustig (2010), who solve an asset pricing model with aggregate and idiosyncratic risk in which agents are subject to collateral constraints arising from limited commitment, and Lorenzoni (2009), who solves a business cycle model with shocks to individuals' expectations. In all these cases, tracking a truncated history turns out to provide computational advantages over choosing endogenous state variables that evolve in a Markovian fashion. One advantage of this approach is that it is often less costly to add an extra endogenous state variable relative to the standard approach, because the same number of lags may still be sufficient. One drawback is that if the Markov process has many states or the model has long memory, the method may not work as well.

6. WHEN DOES HETEROGENEITY MATTER FOR AGGREGATES?

As noted earlier, Krusell and Smith (1998) report that the time series behavior of the aggregated incomplete markets model, by and large, looks very similar to the corresponding representative-agent model. A similar result was reported in Ríos-Rull (1996). However, it is important to interpret these findings correctly and to not overgeneralize them. For example, even if a model aggregates *exactly*, modeling heterogeneity can be very important for aggregate problems. This is because the problem solved by the representative agent can look dramatically different from the problem solved by individuals (for example, have very different preferences). Here, I discuss some important problems in macroeconomics where introducing heterogeneity yields conclusions quite different from a representative-agent model.

The Curse of Long Horizon

It is useful to start by discussing why incomplete markets do not matter in many models. Loosely speaking, this outcome follows from the fact that a long horizon makes individuals' savings function approximately linear in wealth (i.e., constant MPC out of wealth). As we saw in Section 1, the exact linearity of savings rule delivers exact demand aggregation in both Gorman (1961)

and Rubinstein's (1974) theorems. As it turns out, even with idiosyncratic shocks, this near-linearity holds for wealth levels that are not immediately near borrowing constraints. Thus, even though markets are incomplete, redistributing wealth would matter little, and we have something that looks like demand aggregation!

Is there a way to get around this result? It is instructive to look at a concrete example. Mankiw (1986) was one of the first articles in the literature on the equity premium puzzle and one that gave prominence to the role of heterogeneity. Mankiw (1986) shows that, in a two-period model with incomplete markets and idiosyncratic risk of the right form, one can generate an equity premium as large as desired. However, researchers who followed up on this promising lead (Telmer 1993, Heaton and Lucas 1996, and Krusell and Smith 1997) quickly came to a disappointing conclusion: Once agents in these models are allowed to live for multiple periods, trading a single risk-free asset yields sufficient consumption insurance, which in turn results in a tiny equity premium.

This result—that a sufficiently long horizon can dramatically weaken the effects of incomplete markets—is quite general. In fact, Levine and Zame (2002) prove that if, in a single good economy with no aggregate shocks, (i) idiosyncratic income shocks follow a Markov process, (ii) marginal utility is convex, and (iii) all agents have access to a single risk-free asset, then, as individuals' subjective time discount factor (δ) approaches unity, incomplete markets allocations (and utilities) converge to those from a complete markets economy with the same aggregate resources. Although Levine and Zame's result is theoretical for the limit of such economies (as $\delta \longrightarrow 1$), it still sounds a cautionary note to researchers building incomplete markets models: Unless shocks are extremely persistent and/or individuals are very impatient, these models are unlikely to generate results much different from a representative-agent model.

Constantinides and Duffie (1996) show one way to get around the problem of a long horizon, which is also consistent with the message of Levine-Zame's theorem. Essentially, they assume that individuals face permanent shocks, which eliminate the incentives to smooth such shocks. Therefore, they behave as if they live in a static world and choose not to trade. Constantinides and Duffie also revive another feature of Mankiw's (1986) model: Idiosyncratic shocks must have larger variance in recessions (i.e., countercyclical variances) to generate a large equity premium. With these two features, they show that Mankiw's original insight can be made to work once again in an infinite horizon model. Storesletten, Telmer, and Yaron (2007) find that a calibrated model along the lines suggested by Constantinides and Duffie can generate about $\frac{1}{4}$ of the equity premium observed in the U.S. data.

The bottom line is that, loosely speaking, if incomplete markets matter in a model mainly through its effect on the consumption-saving decision, a

long horizon can significantly weaken the bite of incomplete markets. With a long enough horizon, agents accumulate sufficient wealth and end up on the nearly linear portion of their savings function, delivering results not far from a complete markets model. This is also the upshot of Krusell and Smith's (1998) analysis.

Examples Where Heterogeneity Does Matter

There are many examples in which heterogeneity does matter for aggregate phenomena. Here, I review some examples.

First, aggregating heterogeneous-agent models can give rise to preferences for the representative agent that may have nothing to do with the preferences in the underlying model. A well-known example of such a transformation is present in the early works of Hansen (1985) and Rogerson (1988), who show that in a model in which individuals have no intensive margin of labor supply (i.e., zero Frisch labor supply elasticity), one can aggregate the model to find that the representative agent has linear preferences in leisure (i.e., infinite Frisch elasticity!). This conclusion challenges one of the early justifications for building models with microfoundations, which was to bring evidence from microdata to bear on the calibration of macroeconomic models. In an excellent survey article, Browning, Hansen, and Heckman (1999) issue an early warning, giving several examples where this approach is fraught with danger.

Building on the earlier insights of Hansen and Rogerson, Chang and Kim (2006) construct a model in which aggregate labor-supply elasticity depends on the reservation-wage distribution in the population. The economy is populated by households (husband and wife) that each supply labor only along the extensive margin: they either work full time or stay home. Workers are hit by idiosyncratic productivity shocks, causing them to move in and out of the labor market. The aggregate labor-supply elasticity of such an economy is around one, greater than a typical microestimate and much greater than the Frisch elasticity one would measure at the intensive margin (which is zero) in this model. The model thus provides a reconciliation between the micro- and macro-labor-supply elasticities. In a similar vein, Chang, Kim, and Schorfheide (2010) show that preference shifters that play an important role in discussions of aggregate policy are not invariant to policies if they are generated from the aggregation of a heterogeneous-agent model. Such a model also generates "wedges" at the aggregate level that do not translate into any well-defined notion of preference shifters at the microlevel.[45] Finally, Erosa, Fuster, and Kambourov (2009) build a life-cycle model of labor supply, by combining and extending the ideas introduced in earlier articles, such

[45] See Chari, Kehoe, and McGrattan (2007) for the definition of business-cycle wedges.

as Chang and Kim (2006) and Rogerson and Wallenius (2009). Their goal is to build a model with empirically plausible patterns of hours over the life cycle and examine the response elasticities of labor supply to various policy experiments.

In another context, Guvenen (2006) asks why macroeconomic models in the RBC tradition typically need to use a high elasticity of intertemporal substitution (EIS) to explain output and investment fluctuations, whereas Euler equation regressions (such as in Hall [1988] and Campbell and Mankiw [1990]) that use aggregate consumption data estimate a much smaller EIS (close to zero) to fit the data. He builds a model with two types of agents who differ in their EIS. The model generates substantial wealth inequality and much smaller consumption inequality, both in line with the U.S. data. Consequently, capital and investment fluctuations are mainly driven by the rich (who hold almost all the wealth in the economy) and thus reflect the high EIS of this group. Consumption fluctuations, on the other hand, reflect an average that puts much more weight on the EIS of the poor, who contribute significantly to aggregate consumption. Thus, a heterogeneous-agent model is able to explain aggregate evidence that a single representative-agent model has trouble fitting.

In an asset pricing context, Constantinides and Duffie (1996), Chan and Kogan (2002), and Guvenen (2011) show a similar result for risk aversion. Constantinides and Duffie (1996) show theoretically how the cross-sectional distribution of consumption in a heterogeneous-agent model gets translated into a higher risk aversion for the representative agent. Guvenen (2011) shows that, in a calibrated model with limited stock market participation that matches several asset pricing facts, the aggregate risk aversion is measured to be as high as 80, when the individuals' risk aversion is only two. These results, as well as the articles discussed above, confirm and amplify the concerns originally highlighted by Browning, Hansen, and Heckman (1999). The conclusion is that researchers must be very careful when using microeconomic evidence to calibrate representative-agent models.

7. COMPUTATION AND CALIBRATION

Because of their complexity, the overwhelming majority of models in this literature are solved on a computer using numerical methods.[46] Thus, I now turn to a discussion of computational issues that researchers often have to confront when solving models with heterogeneity.

[46] There are a few examples of analytical solutions and theoretical results established with heterogeneous-agent models. See, e.g., Constantinides and Duffie (1996), Heathcote, Storesletten, and Violante (2007), Rossi-Hansberg and Wright (2007), Wang (2009), and Guvenen and Kuruscu (forthcoming).

Calibration and Estimation: Avoid Local Search

Economists often need to minimize an objective function of multiple variables that has lots of kinks, jaggedness, and deep ridges. Consequently, the global minimum is often surrounded by a large number of local minima. A typical example of such a problem arises when a researcher tries to calibrate several structural parameters of an economic model by matching some data moments. Algorithms based on local optimization methods (e.g., Newton-Raphson style derivative-based methods or Nelder-Mead simplex style routines) very often get stuck in local minima because the objective surface is typically very rough (non-smooth).

It is useful to understand some of the sources of this roughness. For example, linear interpolation that is often used in approximating value functions or decision rules generates an interpolated function that is non-differentiable (i.e., has kinks) at every knot point. Similarly, problems with (borrowing, portfolio, etc.) constraints can create significant kinks. Because researchers use a finite number of individuals to simulate data from the model (to compute moments), a small change in the parameter value (during the minimization of the objective) can move some individuals across the threshold—from being constrained to unconstrained or vice versa—which can cause small jumps in the objective value. And sometimes, the moments that the researcher decides to match would be inherently discontinuous in the underlying parameters (with a finite number of individuals), such as the median of a distribution (e.g., wealth holdings). Further compounding the problems, if the moments are not jointly sufficiently informative about the parameters to be calibrated, the objective function would be flat in certain directions. As can be expected, trying to minimize a relatively flat function with lots of kinks, jaggedness, and even small jumps can be a very difficult task indeed.[47]

While the algorithm described here can be applied to the calibration of any model, it is especially useful in models with heterogeneous agents—since such models are time consuming to solve even once, an exhaustive search of the parameter space becomes prohibitively costly (which could be feasible in simpler models).

[47] One simple, but sometimes overlooked, point is that when minimizing an objective function of moments to calibrate a model, one should use the same "seeds" for the random elements of the model that are used to simulate the model in successive evaluations of the objective function. Otherwise, some of the change in objective value will be because of the inherent randomness in different draws of random variables. This can create significant problems with the minimization procedure.

A Simple Fully Parallelizable Global Optimization Algorithm

Here, I describe a global optimization algorithm that I regularly use for calibrating models and I have found it to be very practical and powerful. It is relatively straightforward to implement, yet allows full parallelization across any number of central processing unit (CPU) cores as well as across any number of computers that are connected to the Internet. It requires no knowledge of MPI, OpenMP, or related tools, and no knowledge of computer networking other than using some commercially available synchronization tools (such as DropBox, SugarSync, etc.).

A broad outline of the algorithm is as follows. As with many global algorithms, this procedure combines a global stage with a local search stage that is restarted at various locations in the parameter space. First, we would like to search the parameter space as thoroughly as possible, but do so in as efficient a way as possible. Thoroughness is essential because we want to be sure that we found the true global minimum, so we are willing to sacrifice some speed to ensure this. The algorithm proceeds by taking an initial starting point (chosen in a manner described momentarily) and conducting a local search from that point on until the minimization routine converges as specified by some tolerance. For local search, I typically rely on the Nelder-Mead's downhill simplex algorithm because it does not require derivative information (that may be inaccurate given the approximation errors in the model's solution algorithm).[48] The minimum function value as well as the parameter combination that attained that minimum are recorded in a file saved on the computer's hard disk. The algorithm then picks the next "random" starting point and repeats the previous step of local minimization. The results are then added to the previous file, which records all the local minima found up to that point.

Of course, the most obvious algorithm would be to keep doing a very large number of restarts of this sort and take the minimum of all the minima found in the process. But this would be very time consuming and would not be particularly efficient. Moreover, in many cases, the neighborhood of the global minimum can feature many deep ridges and kinks nearby, which requires more extensive searching near the global minimum, whereas the proposed approach would devote more time to points far away from the true global minimum and to the points near it. Further, if the starting points are chosen literally randomly, this would also create potentially large efficiency losses, because these points have a non-negligible chance of falling near points previously tried. Because those areas have been previously searched, devoting more time is not optimal.

[48] An alternative that can be much faster but requires a bit more tweaking for best performance is the trust region method of Zhang, Conn, and Scheinberg (2010) that builds on Powell's (2009) BOBYQA algorithm.

A better approach is to use "quasi-random" numbers to generate the starting points. Quasi-random numbers (also called low-discrepancy sequences) are sequences of deterministic numbers that spread to any space in the maximally separated way. They avoid the pitfall of random draws that may end up being too close to each other. Each draw in the sequence "knows" the location of previous points drawn and attempts to fill the gaps as evenly as possible.[49] Among a variety of sequences proposed in the literature, the Sobol' sequence is generally viewed to be superior in most practical applications, having a very uniform filling of the space (i.e., maximally separated) even when a small number of points is drawn, as well as a very fast algorithm that generates the sequence.[50]

Next, how do we use the accumulated information from previous restarts? As suggested by genetic algorithm heuristics, I combine information from previous best runs to adaptively direct the new restarts to areas that appear more promising. This is explained further below. Now for the specifics of the algorithm.

Algorithm 1 *Let* $\boldsymbol{p}$ *be a* J*-dimensional parameter vector with generic element* p^j, $j = 1, ..., J$.

- *Step 0.* ***Initialization****:*
  - *Determine bounds for each parameter, outside of which the objective function should be set to a high value.*
  - *Generate a sequence of Sobol' numbers with a sequence length of* I_{max} *(the maximum anticipated number of restarts in the global stage). Set the global iteration number* $i = 1$.
- *Step 1.* ***Global Stage****:*
  - *Draw the* i^{th} *(vector) value in the Sobol' sequence:* $\boldsymbol{s}_i$.
  - *If* $i > 1$*, open and read from the text file* "saved_parameters.dat" *the function values (and corresponding parameter vectors) of previously found local minima. Denote the lowest function value found as of iteration* $i-1$ *as* f_{i-1}^{low} *and the corresponding parameter vector as* $\boldsymbol{p}_{i-1}^{low}$.
  - *Generate a starting point for the local stage as follows:*

[49] Another common application of low-discrepancy sequences is in quasi-Monte Carlo integration, where they have been found to improve time-to-accuracy by several orders of magnitude.

[50] In a wide range of optimization problems, Kucherenko and Sytsko (2005) and Liberti and Kucherenko (2005) find that Sobol' sequences outperform Holton sequences, both in terms of computation time and probability of finding the global optimum. The Holton sequence is particularly weak in high dimensional applications.

  - *If $i < I_{\min}(< I_{\max})$, then use $\mathbf{s}_i$ as the initial guess: $\mathbf{S}_i = \mathbf{s}_i$. Here, $I_{\min}$ is the threshold below which we use fully quasi-random starting points in the global stage.*
  - *If $i \geq I_{\min}$, take the initial guess to be a convex combination of $\mathbf{s}_i$ and the parameter value that generated the best local minima so far: $\mathbf{S}_i = (1-\theta_i)\mathbf{s}_i + \theta \mathbf{p}_{i-1}^{low}$. The parameter $\theta_i \in [0, \bar{\theta}]$ with $\bar{\theta} < 1$, and increases with i. For example, I found that a convex increasing function, such as $\theta_i = \min[\bar{\theta}, (i/I_{\max})^2]$, works well in some applications. An alternative heuristic is given later.*
  - *As θ_i is increased, local searches are restarted from a narrower part of the parameter space that yielded the lowest local minima before.*

- *Step 2:* ***Local Stage****:*
  - *Using $\mathbf{S}_i$ as a starting point, use the downhill simplex algorithm to search for a local minimum. (For the other vertices of the simplex, randomly draw starting points within the bounds of the parameter space.)*
  - *Stop when either (i) a certain tolerance has been achieved, (ii) function values do not improve more than a certain amount, or (iii) the maximum iteration number is reached.*
  - *Open* saved_parameters.dat *and record the local minimum found (function value and parameters).*

- *Step 3.* ***Stopping Rule****:*
  - *Stop if the termination criterion described below is satisfied. If not go to Step 1.*

Termination Criterion

One useful heuristic criterion relies on a Bayesian procedure that estimates the probability that the next local search will find a new local minimum based on the rate at which new local minima have been located in past searches. More concretely, if W different local minima have been found after K local searches started from a set of uniformly distributed points, then the expectation of the number of local minima is

$$W_{\exp} = W\,(K-1)\,/\,(K-W-2)\,,$$

provided that $K > W+2$. The searching procedure is terminated if $W_{\exp} < W+0.5$. The idea is that, after a while of searching, if subsequent restarts keep

finding one of the same local minima found before, the chances of improvement in subsequent searches is not worth the additional time cost. Although this is generally viewed as one of the most reliable heuristics, care must be applied as with any heuristic.

Notice also that W_{exp} can be used to adaptively increase the value of θ_i in the global stage (Step 1 [3] above). The idea is that, as subsequent global restarts do not yield a new local minimum with a high enough probability, it is time to narrow the search and further explore areas of promising local minima. Because jaggedness and deep ridges cause local search methods to often get stuck, we want to explore promising areas more thoroughly.

One can improve on this basic algorithm in various ways. I am going to mention a few that seem worth exploring.

Refinements: Clustering and Pre-Testing

First, suppose that in iteration k, the proposed starting point $\mathbf{S}_k$ ends up being "close" to one of the previous minima, say $\mathbf{p}_n^{low}$, for $n < k$. Then it is likely that the search starting from $\mathbf{S}_k$ will end up converging to $\mathbf{p}_n^{low}$. But then we have wasted an entire cycle of local search without gaining anything. To prevent this, one heuristic (called "clustering methods") proceeds by defining a "region of attraction" (which is essentially a J-dimensional ball centered) around each one of the local minima found so far.[51] Then the algorithm would discard a proposed restarting point if it falls into the region attraction of any previous local minima. Because the local minimization stage is the most computationally intensive step, this refinement of restarting the local search only once in a given region of attraction can result in significant computational gains. Extensive surveys of clustering methods can be found in Rinnooy Kan and Timmer (1987a, 1987b).

Second, one can add a "pre-test" stage where N points from the Sobol' sequence are evaluated before any local search (i.e., in Step 0 above), and only a subset of $N^* < N$ points with lowest objective values are used as seeds in the local search. The remaining points, as well as regions of attraction around them are ignored as not promising. Notice that while this stage can improve speed, it trades off reliability in the process.

Narrowing Down the Search Area

The file `saved_parameters.dat` contains a lot of useful information gathered in each iteration to the global stage, which can be used more efficiently as follows. As noted, the Nelder-Mead algorithm requires $J + 1$ candidate

[51] While different formulas have been proposed for determining the optimal radius, these formulas contain some undetermined coefficients, making the formulas less than useful in real life applications.

points as inputs (the vertices of the J-dimensional simplex). One of these points is given by $\mathbf{S}_i$, chosen as described above; the other vertices were drawn randomly. But as we accumulate more information with every iteration on the global stage, if we keep finding local minima that seem to concentrate in certain regions, it makes sense to narrow the range of values from which we pick the vertices. One way to do this is as follows: After a sufficiently large number of restarts have been completed, rank all the function values and take the lowest x percent of values (e.g., 10 percent or 20 percent). Then for each dimension, pick the minimum (p^j_{min}) and maximum parameter value (p^j_{max}) within this set of minima. Then, to generate vertices, take randomly sampled points between p^j_{min} and p^j_{max} in each dimension j. This allows the simplex algorithm to search more intensively in a narrower area, which can improve results very quickly when there are ridges or jaggedness in the objective function that make the algorithm get stuck.

Parallelizing the Algorithm

The algorithm can be parallelized in a relatively straightforward manner.[52] The basic idea is to let each CPU core perform a separate local search in a different part of the parameter space, which is a time-consuming process. If we can do many such searches simultaneously, we can speed up the solution dramatically. One factor that makes parallelization simple is the fact that the CPU cores do not need to communicate with each other during the local search stage. In between the local stages, each CPU core will contribute its findings (the last local minimum it found along with the corresponding parameter vector) to the collective wisdom recorded in `saved_parameters.dat` and also get the latest updated information about the best local minimum found so far from the same file. Thus, as long as all CPU cores have access to the same copy of the file `saved_parameters.dat`, parallelization requires no more than a few lines for housekeeping across CPUs. Here are some more specifics.

Suppose that we have a workstation with N CPU cores (for example, $N = 4$, 6, or 12). The first modification we need to make is to change the program to distinguish between the different "copies" of the code, running on different CPU cores. This can be done by simply having the program ask the user (only once, upon starting the code) to input an integer value, n, between 1 and N, which uniquely identifies the "sequence number" of the particular instance of the program running. Then open N terminal windows and launch a copy of the program in each window. Then for each one, enter a unique sequence number $n = 1, 2, ..., N$.

[52] I am assuming here that a compiled language, such as Fortran or C, is used to write the program. So multiple parallel copies of the same code can be run in different terminal windows.

Upon starting, each program will first simulate the same quasi-random sequence regardless of n, but each run will pick a different element of this sequence as its own seed. For simplicity, suppose run n chooses the nth element of the sequence as its seed and launches a local search from that point. After completion, each run will open the same file `saved_parameters.dat` and record the local minimum and parameter value it finds.[53]

Now suppose that all copies of the program complete their respective first local searches, so there are N lines, each written by a different CPU core, in the file `saved_parameters.dat`. Then each run will start its second iteration and pick as its next seed the $(N+n)$th element of the quasi-random sequence. When the total number of iterations across all CPUs exceed some threshold $I_{\min}$, then we would like to combine the quasi-random draw with the previous best local minima as described in Step 1 (3) above. This is simple since all runs have access to the same copy of `saved_parameters.dat`.[54]

Notice that this parallelization method is completely agnostic about whether the CPU cores are on the same personal computer (PC) or distributed across many PCs *as long as* all computers keep synchronized copies of `saved_parameters.dat`. This can be achieved by using a synchronization service like DropBox. This feature easily allows one to harness the computational power of many idle PCs distributed geographically with varying speeds and CPU cores.

8. FUTURE DIRECTIONS AND FURTHER READING

This article surveys the current state of the heterogeneous-agent models literature and draws several conclusions. First, two key ingredients in such models are (i) the magnitudes and types of risk that the model builder feeds into the model and (ii) the insurance opportunities allowed in the economy. In many cases, it is difficult, if not impossible, to measure each component separately. In other words, the assumptions a researcher makes regarding insurance opportunities will typically affect the inference drawn about the magnitudes of risks and vice versa. Further complicating the problem is the measurement of risk: Individuals often have more information than the econometrician about

[53] Because this opening and writing stage takes a fraction of a second, the likelihood that two or more programs access the file simultaneously and create a run-time error is negligible.

[54] It is often useful for each run to keep track of the *total* number of local searches completed by all CPUs—call this N_{Last}. For example, sometimes the increase in θ_i can be linked to N_{Last}. This number can be read as the total number of lines recorded up to that point in `saved_parameters.dat`. Another use of this index is for determining which point in the sequence to select as the next seed point. So as opposed to running n by selecting the $(kN+n)$th point in the sequence where k is the number of local searches completed by run n, it could just pick the $(N_{Last}+1)$th number in the sequence. This avoids leaving gaps in the sequence for seeds, in case some CPUs are much faster than others and hence finish many more local searches than others.

future changes in their lives. So, for example, a rise or fall in income that the econometrician may view as a "shock" may in fact be partially or completely anticipated by the individual. This suggests that equating income movements observed in the data with risk (as is often done in the literature) is likely to overstate the true magnitude. This entanglement of "risk," "anticipated changes," and "insurance" presents a difficult challenge to researchers in this area. Although some recent progress has been made, more work remains.

A number of surveys contain very valuable material that are complementary to this article. First, Heathcote, Storesletten, and Violante (2009) is a recent survey of quantitative macroeconomics with heterogeneous households that is complementary to this article. Second, Browning, Hansen, and Heckman (1999) contains an extensive review of microeconomic models that are often used as the foundations of heterogeneous-agent models. It highlights several pitfalls in trying to calibrate macroeconomic models using microevidence. Third, Meghir and Pistaferri (2011) provides a comprehensive treatment of how earnings dynamics affect life-cycle consumption choice, which is closely related to the issues discussed in Section 3 of this survey. Finally, because heterogeneous-agent models use microeconomic survey data in increasingly sophisticated ways, a solid understanding of issues related to measurement error (which is pervasive in microdata) is essential. Failure to understand such problems can wreak havoc with the empirical analysis. Bound, Brown, and Mathiowetz (2001) is an extensive and authoritative survey of the subject.

The introduction of families into incomplete markets models represents an exciting area of current research. For many questions of empirical relevance, the interactions taking place within a household (implicit insurance, bargaining, etc.) can have first-order effects on how individuals respond to idiosyncratic changes. To give a few examples, Gallipoli and Turner (2011) document that the labor supply responses to disability shocks of single workers are larger and more persistent than those of married workers. They argue that an important part of this difference has to do with the fact that couples are able to optimally change their time (and task) allocation within households in response to disability, an option not available to singles. This finding suggests that modeling households would be important for understanding the design of disability insurance policies. Similarly, Guner, Kaygusuz, and Ventura (2010) show that to quantify the effects of alternative tax reforms, it is important to take into account the joint nature of household labor supply. In fact, it is hard to imagine any distributional issue for which the household structure does not figure in an important way.

Another promising area is the richer modeling of household finances in an era of ever-increasing sophistication in financial services. The Great Recession, which was accompanied by a housing market crash and soaring personal bankruptcies, home foreclosures, and so on, has created a renewed sense of

urgency for understanding household balance sheets. Developments on two fronts—advances in theoretical modeling as discussed in Section 3, combined with richer data sources on credit histories and mortgages that are increasingly becoming available to researchers—will make faster progress feasible in this area.

REFERENCES

Abowd, John M., and David E. Card. 1989. "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* 57 (March): 411–45.

Acemoglu, Daron. 2002. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature* 40 (March): 7–72.

Aguiar, Mark, and Erik Hurst. 2008. "Deconstructing Lifecycle Expenditures." Working Paper, University of Rochester.

Aguiar, Mark, and Mark Bils. 2011. "Has Consumption Inequality Mirrored Income Inequality?" Working Paper, University of Rochester.

Aiyagari, S. Rao. 1993. "Uninsured Idiosyncratic Risk and Aggregate Saving." Federal Reserve Bank of Minneapolis Working Paper 502.

Aiyagari, S. Rao. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *The Quarterly Journal of Economics* 109 (August): 659–84.

Altug, Sumru, and Robert A. Miller. 1990. "Household Choices in Equilibrium." *Econometrica* 58 (May): 543–70.

Arslan, Yavuz. 2011. "Interest Rate Fluctuations and Equilibrium in the Housing Market." Working Paper, Central Bank of the Republic of Turkey.

Athreya, Kartik B. 2002. "Welfare Implications of the Bankruptcy Reform Act of 1999." *Journal of Monetary Economics* 49 (November): 1,567–95.

Attanasio, Orazio, and Steven J. Davis. 1996. "Relative Wage Movements and the Distribution of Consumption." *Journal of Political Economy* 104 (December): 1,227–62.

Attanasio, Orazio, Erich Battistin, and Hidehiko Ichimura. 2007. "What Really Happened to Consumption Inequality in the United States?" In *Hard-to-Measure Goods and Services: Essays in Honour of Zvi Griliches*, edited by E. Berndt and C. Hulten. Chicago: University of Chicago Press.

Attanasio, Orazio P., James Banks, Costas Meghir, and Guglielmo Weber. 1999. "Humps and Bumps in Lifetime Consumption." *Journal of Business & Economic Statistics* 17 (January): 22–35.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2008. "Trends in U.S. Wage Inequality: Revising the Revisionists." *The Review of Economics and Statistics* 90 (2): 300–23.

Badel, Alejandro, and Mark Huggett. 2007. "Interpreting Life-Cycle Inequality Patterns as an Efficient Allocation: Mission Impossible?" Working Paper, Georgetown University.

Baker, Michael. 1997. "Growth-Rate Heterogeneity and the Covariance Structure of Life-Cycle Earnings." *Journal of Labor Economics* 15 (April): 338–75.

Baker, Michael, and Gary Solon. 2003. "Earnings Dynamics and Inequality among Canadian Men, 1976–1992: Evidence from Longitudinal Income Tax Records." *Journal of Labor Economics* 21 (April): 267–88.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: University of Chicago Press.

Ben-Porath, Yoram. 1967. "The Production of Human Capital and the Life Cycle of Earnings." *Journal of Political Economy* 75 (4): 352–65.

Bewley, Truman F. Undated. "Interest Bearing Money and the Equilibrium Stock of Capital." Working Paper.

Blundell, Richard, and Ian Preston. 1998. "Consumption Inequality And Income Uncertainty." *The Quarterly Journal of Economics* 113 (May): 603–40.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. "Consumption Inequality and Partial Insurance." *American Economic Review* 98 (December): 1,887–921.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, edited by J. J. Heckman and E. E. Leamer. Amsterdam: Elsevier; 3,705–843.

Browning, Martin, Lars Peter Hansen, and James J. Heckman. 1999. "Micro Data and General Equilibrium Models." In *Handbook of Macroeconomics*, edited by J. B. Taylor and M. Woodford. Amsterdam: Elsevier; 543–633.

Browning, Martin, Mette Ejrnaes, and Javier Alvarez. 2010. "Modelling Income Processes with Lots of Heterogeneity." *Review of Economic Studies* 77 (October): 1,353–81.

Cagetti, Marco, and Mariacristina De Nardi. 2006. "Entrepreneurship, Frictions, and Wealth." *Journal of Political Economy* 114 (October): 835–70.

Campbell, John Y., and N. Gregory Mankiw. 1990. "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence." Working Paper 2924. Cambridge, Mass.: National Bureau of Economic Research (May).

Carroll, Christopher. 2000. "Why Do the Rich Save So Much?" In *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, edited by Joel Slemrod. Boston: Harvard University Press, 466–84.

Carroll, Christopher D. 1991. "Buffer Stock Saving and the Permanent Income Hypothesis." Board of Governors of the Federal Reserve System Working Paper Series/Economic Activity Section 114.

Carroll, Christopher D. 1997. "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis." *The Quarterly Journal of Economics* 112 (February): 1–55.

Carroll, Christopher D., and Andrew A. Samwick. 1997. "The Nature of Precautionary Wealth." *Journal of Monetary Economics* 40 (September): 41–71.

Caselli, Francesco, and Jaume Ventura. 2000. "A Representative Consumer Theory of Distribution." *American Economic Review* 90 (September): 909–26.

Castañeda, Ana, Javier Díaz-Giménez, and José-Víctor Ríos-Rull. 2003. "Accounting for the U.S. Earnings and Wealth Inequality." *The Journal of Political Economy* 111 (August): 818–57.

Chamberlain, Gary, and Charles A. Wilson. 2000. "Optimal Intertemporal Consumption Under Uncertainty." *Review of Economic Dynamics* 3 (July): 365–95.

Chan, Yeung Lewis, and Leonid Kogan. 2002. "Catching up with the Joneses: Heterogeneous Preferences and the Dynamics of Asset Prices." *Journal of Political Economy* 110 (December): 1,255–85.

Chang, Yongsung, and Sun-Bin Kim. 2006. "From Individual to Aggregate Labor Supply: A Quantitative Analysis based on a Heterogeneous-Agent Macroeconomy." *International Economic Review* 47 (1): 1–27.

Chang, Yongsung, Sun-Bin Kim, and Frank Schorfheide. 2010. "Labor Market Heterogeneity, Aggregation, and the Lucas Critique." Working Paper 16401. Cambridge, Mass.: National Bureau of Economic Research.

Chari, V. V., Patrick J. Kehoe, and Ellen R. McGrattan. 2007. "Business Cycle Accounting." *Econometrica* 75 (3): 781–836.

Chatterjee, Satyajit, and Burcu Eyigungor. 2011. "A Quantitative Analysis of the U.S. Housing and Mortgage Markets and the Foreclosure Crisis." Federal Reserve Bank of Philadelphia Working Paper 11-26 (July).

Chatterjee, Satyajit, Dean Corbae, Makoto Nakajima, and José-Víctor Ríos-Rull. 2007. "A Quantitative Theory of Unsecured Consumer Credit with Risk of Default." *Econometrica* 75 (November): 1,525–89.

Chien, YiLi, and Hanno Lustig. 2010. "The Market Price of Aggregate Risk and the Wealth Distribution." *Review of Financial Studies* 23 (April): 1,596–650.

Christiano, Lawrence J., and Jonas D. M. Fisher. 2000. "Algorithms for Solving Dynamic Models with Occasionally Binding Constraints." *Journal of Economic Dynamics and Control* 24 (July): 1,179–232.

Clarida, Richard H. 1987. "Consumption, Liquidity Constraints, and Asset Accumulation in the Presence of Random Income Fluctuations." *International Economic Review* 28 (June): 339–51.

Clarida, Richard H. 1990. "International Lending and Borrowing in a Stochastic, Stationary Equilibrium." *International Economic Review* 31 (August): 543–58.

Cochrane, John H. 1991. "A Simple Test of Consumption Insurance." *Journal of Political Economy* 99 (October): 957–76.

Congressional Budget Office. 2008. "Recent Trends in the Variability of Individual Earnings and Family Income." Washington, D.C.: CBO (June).

Constantinides, George M. 1982. "Intertemporal Asset Pricing with Heterogeneous Consumers and Without Demand Aggregation." *Journal of Business* 55 (April): 253–67.

Constantinides, George M., and Darrell Duffie. 1996. "Asset Pricing with Heterogeneous Consumers." *Journal of Political Economy* 104 (April): 219–40.

Cunha, Flavio, James Heckman, and Salvador Navarro. 2005. "Separating Uncertainty from Heterogeneity in Life Cycle Earnings." *Oxford Economic Papers* 57 (2): 191–261.

Cutler, David M., and Lawrence F. Katz. 1992. "Rising Inequality? Changes in the Distribution of Income and Consumption in the 1980's." *American Economic Review* 82 (May): 546–51.

Davis, Steven J., R. Jason Faberman, John Haltiwanger, Ron Jarmin, and Javier Miranda. 2010. "Business Volatility, Job Destruction, and Unemployment." Working Paper 14300. Cambridge, Mass.: National Bureau of Economic Research (September).

De Nardi, Mariacristina, Eric French, and John B. Jones. 2010. "Why Do the Elderly Save? The Role of Medical Expenses." *Journal of Political Economy* 118 (1): 39–75.

Deaton, Angus. 1991. "Saving and Liquidity Constraints." *Econometrica* 59 (September): 1,221–48.

Deaton, Angus, and Christina Paxson. 1994. "Intertemporal Choice and Inequality." *Journal of Political Economy* 102 (June): 437–67.

Debreu, Gerard. 1959. *Theory of Value*. New York: John Wiley and Sons.

den Haan, Wouter J. 2010. "Assessing the Accuracy of the Aggregate Law of Motion in Models with Heterogeneous Agents." *Journal of Economic Dynamics and Control* 34 (January): 79–99.

den Haan, Wouter J., and Pontus Rendahl. 2009. "Solving the Incomplete Markets Model with Aggregate Uncertainty Using Explicit Aggregation." Working Paper, University of Amsterdam.

Domeij, David, and Martin Floden. 2006. "The Labor-Supply Elasticity and Borrowing Constraints: Why Estimates are Biased." *Review of Economic Dynamics* 9 (April): 242–62.

Dynan, Karen E., Douglas W. Elmendorf, and Daniel E. Sichel. 2007. "The Evolution of Household Income Volatility." Federal Reserve Board Working Paper 2007-61.

Erosa, Andrés, Luisa Fuster, and Gueorgui Kambourov. 2009. "The Heterogeneity and Dynamics of Individual Labor Supply over the Life Cycle: Facts and Theory." Working Paper, University of Toronto.

Flavin, Marjorie A. 1981. "The Adjustment of Consumption to Changing Expectations About Future Income." *Journal of Political Economy* 89 (October): 974–1,009.

French, Eric, and John Bailey Jones. 2004. "On the Distribution and Dynamics of Health Care Costs." *Journal of Applied Econometrics* 19 (6): 705–21.

Galindev, Ragchaasuren, and Damba Lkhagvasuren. 2010. "Discretization of Highly Persistent Correlated AR(1) Shocks." *Journal of Economic Dynamics and Control* 34 (July): 1,260–76.

Gallipoli, Giovanni, and Laura Turner. 2011. "Household Responses to Individual Shocks: Disability and Labour Supply." Working Paper, University of British Columbia.

Glover, Andrew, and Jacob Short. 2010. "Bankruptcy, Incorporation, and the Nature of Entrepreneurial Risk." Working Paper, University of Western Ontario.

Gorman, William M. 1961. "On a Class of Preference Fields." *Metroeconomica* 13 (June): 53–6.

Gottschalk, Peter, and Robert Moffitt. 1994. "The Growth of Earnings Instability in the U.S. Labor Market." *Brookings Papers on Economic Activity* 25 (2): 217–72.

Gottschalk, Peter, and Robert Moffitt. 1999. "Changes in Job Instability and Insecurity Using Monthly Survey Data." *Journal of Labor Economics* 17 (October): S91–126.

Gourinchas, Pierre-Olivier, and Jonathan A. Parker. 2002. "Consumption over the Life Cycle." *Econometrica* 70 (January): 47–89.

Greenwood, Jeremy, Ananth Seshadri, and Mehmet Yorukoglu. 2005. "Engines of Liberation." *Review of Economic Studies* 72 (1): 109–33.

Greenwood, Jeremy, and Nezih Guner. 2009. "Marriage and Divorce since World War II: Analyzing the Role of Technological Progress on the Formation of Households." In *NBER Macroeconomics Annual*, Vol. 23. Cambridge, Mass.: National Bureau of Economic Research, 231–76.

Guner, Nezih, Remzi Kaygusuz, and Gustavo Ventura. 2010. "Taxation and Household Labor Supply." Working Paper, Arizona State University.

Gustavsson, Magnus, and Pär Österholm. 2010. "Does the Labor-Income Process Contain a Unit Root? Evidence from Individual-Specific Time Series." Working Paper, Uppsala University.

Guvenen, Fatih. 2006. "Reconciling Conflicting Evidence on the Elasticity of Intertemporal Substitution: A Macroeconomic Perspective." *Journal of Monetary Economics* 53 (October): 1,451–72.

Guvenen, Fatih. 2007a. "Do Stockholders Share Risk More Effectively than Nonstockholders?" *The Review of Economics and Statistics* 89 (2): 275–88.

Guvenen, Fatih. 2007b. "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" *American Economic Review* 97 (June): 687–712.

Guvenen, Fatih. 2009a. "An Empirical Investigation of Labor Income Processes." *Review of Economic Dynamics* 12 (January): 58–79.

Guvenen, Fatih. 2009b. "A Parsimonious Macroeconomic Model for Asset Pricing." *Econometrica* 77 (November): 1,711–50.

Guvenen, Fatih. 2011. "Limited Stock Market Participation Versus External Habit: An Intimate Link." University of Minnesota Working Paper 450.

Guvenen, Fatih, and Anthony A. Smith. 2009. "Inferring Labor Income Risk from Economic Choices: An Indirect Inference Approach." Working Paper, University of Minnesota.

Guvenen, Fatih, and Burhanettin Kuruscu. 2010. "A Quantitative Analysis of the Evolution of the U.S. Wage Distribution, 1970–2000." In *NBER Macroeconomics Annual 2009* 24 (1): 227–76.

Guvenen, Fatih, and Burhanettin Kuruscu. Forthcoming. "Understanding the Evolution of the U.S. Wage Distribution: A Theoretical Analysis." *Journal of the European Economic Association*.

Guvenen, Fatih, and Michelle Rendall. 2011. "Emancipation Through Education." Working Paper, University of Minnesota.

Guvenen, Fatih, Burhanettin Kuruscu, and Serdar Ozkan. 2009. "Taxation of Human Capital and Wage Inequality: A Cross-Country Analysis." Working Paper 15526. Cambridge, Mass.: National Bureau of Economic Research (November).

Haider, Steven J. 2001. "Earnings Instability and Earnings Inequality of Males in the United States: 1967–1991." *Journal of Labor Economics* 19 (October): 799–836.

Haider, Steven, and Gary Solon. 2006. "Life-Cycle Variation in the Association between Current and Lifetime Earnings." *American Economic Review* 96 (September): 1,308–20.

Hall, Robert E. 1988. "Intertemporal Substitution in Consumption." *Journal of Political Economy* 96 (April): 339–57.

Hall, Robert E., and Frederic S. Mishkin. 1982. "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households." *Econometrica* 50 (March): 461–81.

Hansen, Gary D. 1985. "Indivisible Labor and the Business Cycle." *Journal of Monetary Economics* 16 (November): 309–27.

Harris, Milton, and Bengt Holmstrom. 1982. "A Theory of Wage Dynamics." *Review of Economic Studies* 49 (July): 315–33.

Hause, John C. 1980. "The Fine Structure of Earnings and the On-the-Job Training Hypothesis." *Econometrica* 48 (May): 1,013–29.

Hayashi, Fumio, Joseph Altonji, and Laurence Kotlikoff. 1996. "Risk-Sharing between and within Families." *Econometrica* 64 (March): 261–94.

Heathcote, Jonathan, Fabrizio Perri, and Giovanni L. Violante. 2010. "Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States, 1967–2006." *Review of Economic Dynamics* 13 (January): 15–51.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2007. "Consumption and Labour Supply with Partial Insurance: An Analytical Framework." CEPR Discussion Papers 6280 (May).

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2008. "The Macroeconomic Implications of Rising Wage Inequality in the United States." Working Paper 14052. Cambridge, Mass.: National Bureau of Economic Research (June).

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2009. "Quantitative Macroeconomics with Heterogeneous Households." *Annual Review of Economics* 1 (1): 319–54.

Heaton, John, and Deborah J. Lucas. 1996. "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing." *Journal of Political Economy* 104 (June): 443–87.

Heaton, John, and Deborah Lucas. 2000. "Portfolio Choice and Asset Prices: The Importance of Entrepreneurial Risk." *Journal of Finance* 55 (June): 1,163–98.

Heckman, James, Lance Lochner, and Christopher Taber. 1998. "Explaining Rising Wage Inequality: Explanations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents." *Review of Economic Dynamics* 1 (January): 1–58.

Hornstein, Andreas, Per Krusell, and Giovanni L. Violante. 2011. "Frictional Wage Dispersion in Search Models: A Quantitative Assessment." *American Economic Review* 101 (December): 2,873–98.

Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes. 1994. "The Importance of Precautionary Motives in Explaining Individual and Aggregate Saving." *Carnegie-Rochester Conference Series on Public Policy* 40 (June): 59–125.

Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes. 1995. "Precautionary Saving and Social Insurance." *Journal of Political Economy* 103 (April): 360–99.

Huggett, Mark. 1993. "The Risk-Free Rate in Heterogeneous-Agent Incomplete-Insurance Economies." *Journal of Economic Dynamics and Control* 17: 953–69.

Huggett, Mark. 1996. "Wealth Distribution in Life-cycle Economies." *Journal of Monetary Economics* 38 (December): 469–94.

Huggett, Mark, Gustavo Ventura, and Amir Yaron. 2006. "Human Capital and Earnings Distribution Dynamics." *Journal of Monetary Economics* 53 (March): 265–90.

Huggett, Mark, Gustavo Ventura, and Amir Yaron. 2011. "Sources of Lifetime Inequality." *American Economic Review* 101 (December): 2,923–54.

Imrohoroglu, Ayse. 1989. "Cost of Business Cycles with Indivisibilities and Liquidity Constraints." *Journal of Political Economy* 97 (December): 1,364–83.

Jencks, Christopher. 1984. "The Hidden Prosperity of the 1970s." *Public Interest* 77: 37–61.

Jones, Larry E., Rodolfo E. Manuelli, and Ellen R. McGrattan. 2003. "Why are Married Women Working So Much?" Federal Reserve Bank of Minneapolis Staff Report 317 (June).

Jovanovic, Boyan. 1979. "Job Matching and the Theory of Turnover." *Journal of Political Economy* 87 (October): 972–90.

Judd, Kenneth L., and Sy-Ming Guu. 2001. "Asymptotic Methods for Asset Market Equilibrium Analysis." *Economic Theory* 18 (1): 127–57.

Kaplan, Greg. 2010. "Inequality and the Life Cycle." Working Paper, University of Pennsylvania.

Kaplan, Greg, and Giovanni L. Violante. 2010. "How Much Consumption Insurance Beyond Self-Insurance?" *American Economic Journal: Macroeconomics* 2 (October): 53–87.

Kehoe, Timothy J., and David K. Levine. 1993. "Debt-Constrained Asset Markets." *Review of Economic Studies* 60 (October): 865–88.

Kitao, Sagiri, Lars Ljungqvist, and Thomas J. Sargent. 2008. "A Life Cycle Model of Trans-Atlantic Employment Experiences." Working Paper, University of Southern California and New York University.

Knowles, John. 2007. "Why Are Married Men Working So Much? The Macroeconomics of Bargaining Between Spouses." Working Paper, University of Pennsylvania.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2010. "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937." *Quarterly Journal of Economics* 125 (February): 91–128.

Kopecky, Karen A., and Richard M. H. Suen. 2010. "Finite State Markov-chain Approximations to Highly Persistent Processes." *Review of Economic Dynamics* 13 (July): 701–14.

Krueger, Dirk, and Fabrizio Perri. 2006. "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory." *Review of Economic Studies* 73 (1): 163–93.

Krueger, Dirk, and Fabrizio Perri. 2009. "How Do Households Respond to Income Shocks?" Working Paper, University of Minnesota.

Krusell, Per, and Anthony A. Smith. 1997. "Income and Wealth Heterogeneity, Portfolio Choice, and Equilibrium Asset Returns." *Macroeconomic Dynamics* 1 (June): 387–422.

Krusell, Per, and Anthony A. Smith, Jr.. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *Journal of Political Economy* 106 (October): 867–96.

Kucherenko, Sergei, and Yury Sytsko. 2005. "Application of Deterministic Low-Discrepancy Sequences in Global Optimization." *Computational Optimization and Applications* 30: 297–318.

Laitner, John. 1992. "Random Earnings Differences, Lifetime Liquidity Constraints, and Altruistic Intergenerational Transfers." *Journal of Economic Theory* 58 (December): 135–70.

Laitner, John. 2002. "Wealth Inequality and Altruistic Bequests." *American Economic Review* 92 (May): 270–3.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (March): 31–43.

Levine, David K., and William R. Zame. 2002. "Does Market Incompleteness Matter?" *Econometrica* 70 (September): 1,805–39.

Liberti, Leo, and Sergei Kucherenko. 2005. "Comparison of Deterministic and Stochastic Approaches to Global Optimization." *International Transactions in Operations Research* 12: 263–85.

Lillard, Lee A., and Robert J. Willis. 1978. "Dynamic Aspects of Earnings Mobility." *Econometrica* 46 (September): 985–1,012.

Lillard, Lee A., and Yoram Weiss. 1979. "Components of Variation in Panel Earnings Data: American Scientists 1960–70." *Econometrica* 47 (March): 437–54.

Livshits, Igor, James MacGee, and Michèle Tertilt. 2007. "Consumer Bankruptcy—A Fresh Start." *American Economic Review* 97 (March): 402–18.

Livshits, Igor, James MacGee, and Michèle Tertilt. 2010. "Accounting for the Rise in Consumer Bankruptcies." *American Economic Journal: Macroeconomics* 2 (April): 165–93.

Lorenzoni, Guido. 2009. "A Theory of Demand Shocks." *American Economic Review* 99 (December): 2,050–84.

Lucas, Jr., Robert E. 1987. *Models of Business Cycles*. New York: Basil Blackwell.

Lucas, Jr., Robert E. 2003. "Macroeconomic Priorities." *American Economic Review* 93 (March): 1–14.

Mace, Barbara J. 1991. "Full Insurance in the Presence of Aggregate Uncertainty." *Journal of Political Economy* 99 (October): 928–56.

MaCurdy, Thomas E. 1982. "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis." *Journal of Econometrics* 18 (January) 83–114.

Mankiw, N. Gregory. 1986. "The Equity Premium and the Concentration of Aggregate Shocks." *Journal of Financial Economics* 17 (September): 211–9.

Meghir, Costas, and Luigi Pistaferri. 2004. "Income Variance Dynamics and Heterogeneity." *Econometrica* 72 (1): 1–32.

Meghir, Costas, and Luigi Pistaferri. 2011. "Earnings, Consumption, and Life Cycle Choices." In *Handbook of Labor Economics*, Vol 4B, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier, 773–854.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15 (March): 145–61.

Mian, Atif, and Amir Sufi. 2011. "House Prices, Home Equity-Based Borrowing, and the U.S. Household Leverage Crisis." *American Economic Review* 101 (August): 2,132–56.

Moffitt, Robert, and Peter Gottschalk. 1995. "Trends in the Covariance Structure of Earnings in the United States: 1969–1987." Institute for Research on Poverty Discussion Papers 1001-93, University of Wisconsin Institute for Research and Poverty.

Moffitt, Robert, and Peter Gottschalk. 2008. "Trends in the Transitory Variance of Male Earnings in the U.S., 1970–2004." Working Paper, Johns Hopkins University.

Nelson, Julie A. 1994. "On Testing for Full Insurance Using Consumer Expenditure Survey Data: Comment." *Journal of Political Economy* 102 (April): 384–94.

Ozkan, Serdar. 2010. "Income Differences and Health Care Expenditures over the Life Cycle." Working Paper, University of Pennsylvania.

Palumbo, Michael G. 1999. "Uncertain Medical Expenses and Precautionary Saving Near the End of the Life Cycle." *Review of Economic Studies* 66 (April): 395–421.

Pijoan-Mas, Josep. 2006. "Precautionary Savings or Working Longer Hours?" *Review of Economic Dynamics* 9 (April): 326–52.

Powell, Michael J. D. 2009. "The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives." Numerical Analysis Papers NA06, Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, Cambridge (August).

Pratt, John W. 1964. "Risk Aversion in the Small and in the Large." *Econometrica* 32 (1/2): 122–36.

Primiceri, Giorgio E., and Thijs van Rens. 2009. "Heterogeneous Life-Cycle Profiles, Income Risk and Consumption Inequality." *Journal of Monetary Economics* 56 (January): 20–39.

Quadrini, Vincenzo. 2000. "Entrepreneurship, Saving, and Social Mobility." *Review of Economic Dynamics* 3 (January): 1–40.

Rinnooy Kan, Alexander, and G. T. Timmer. 1987a. "Stochastic Global Optimization Methods Part I: Clustering Methods." *Mathematical Programming* 39: 27–56.

Rinnooy Kan, Alexander, and G. T. Timmer. 1987b. "Stochastic Global Optimization Methods Part II: Multilevel Methods." *Mathematical Programming* 39: 57–78.

Ríos-Rull, José Victor. 1996. "Life-Cycle Economies and Aggregate Fluctuations." *Review of Economic Studies* 63 (July): 465–89.

Rogerson, Richard. 1988. "Indivisible Labor, Lotteries and Equilibrium." *Journal of Monetary Economics* 21 (January): 3–16.

Rogerson, Richard, and Johanna Wallenius. 2009. "Micro and Macro Elasticities in a Life Cycle Model With Taxes." *Journal of Economic Theory* 144 (November): 2,277–92.

Rogerson, Richard, Robert Shimer, and Randall Wright. 2005. "Search-Theoretic Models of the Labor Market: A Survey." *Journal of Economic Literature* 43 (December): 959–88.

Rossi-Hansberg, Esteban, and Mark L. J. Wright. 2007. "Establishment Size Dynamics in the Aggregate Economy." *American Economic Review* 97 (December): 1,639–66.

Rouwenhorst, K. Geert. 1995. "Asset Pricing Implications of Equilibrium Business Cycle Models." In *Frontiers of Business Cycle Research*. Princeton, N.J.: Princeton University Press, 294–330.

Rubinstein, Mark. 1974. "An Aggregation Theorem for Securities Markets." *Journal of Financial Economics* 1 (September): 225–44.

Sabelhaus, John, and Jae Song. 2009. "Earnings Volatility Across Groups and Time." *National Tax Journal* 62 (June): 347–64.

Sabelhaus, John, and Jae Song. 2010. "The Great Moderation in Micro Labor Earnings." *Journal of Monetary Economics* 57 (May): 391–403.

Schechtman, Jack, and Vera L. S. Escudero. 1977. "Some Results on an 'Income Fluctuation Problem.'" *Journal of Economic Theory* 16 (December): 151–66.

Schulhofer-Wohl, Sam. 2011. "Heterogeneity and Tests of Risk Sharing." Federal Reserve Bank of Minneapolis Staff Report 462 (September).

Shimer, Robert. 2005. "The Cyclicality of Hires, Separations, and Job-to-Job Transitions." Federal Reserve Bank of St. Louis *Review* 87 (4): 493–507.

Shimer, Robert. 2007. "Reassessing the Ins and Outs of Unemployment." Working Paper 13421. Cambridge, Mass.: National Bureau of Economic Research (September).

Shin, Donggyun, and Gary Solon. 2011. "Trends in Men's Earnings Volatility: What Does the Panel Study of Income Dynamics Show?" *Journal of Public Economics* 95 (August): 973–82.

Storesletten, Kjetil, Christopher I. Telmer, and Amir Yaron. 2004a. "Consumption and Risk Sharing Over the Life Cycle." *Journal of Monetary Economics* 51 (April): 609–33.

Storesletten, Kjetil, Christopher I. Telmer, and Amir Yaron. 2004b. "Cyclical Dynamics in Idiosyncratic Labor Market Risk." *Journal of Political Economy* 112 (June): 695–717.

Storesletten, Kjetil, Christopher I. Telmer, and Amir Yaron. 2007. "Asset Pricing with Idiosyncratic Risk and Overlapping Generations." *Review of Economic Dynamics* 10 (October): 519–48.

Telmer, Christopher I. 1993. "Asset Pricing Puzzles and Incomplete Markets." *Journal of Finance* 48 (December): 1,803–32.

Topel, Robert H. 1990. "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority." Working Paper 3294. Cambridge, Mass.: National Bureau of Economic Research (March).

Topel, Robert H., and Michael P. Ward. 1992. "Job Mobility and the Careers of Young Men." *Quarterly Journal of Economics* 107 (May): 439–79.

Townsend, Robert M. 1994. "Risk and Insurance in Village India." *Econometrica* 62 (May): 539–91.

Veracierto, Marcelo. 1997. "Plant-Level Irreversible Investment and Equilibrium Business Cycles." Federal Reserve Bank of Minneapolis Discussion Paper 115 (March).

Wang, Neng. 2009. "Optimal Consumption and Asset Allocation with Unknown Income Growth." *Journal of Monetary Economics* 56 (May): 524–34.

Zhang, Hongchao, Andrew R. Conn, and Katya Scheinberg. 2010. "A Derivative-Free Algorithm for Least-Squares Minimization." *SIAM Journal on Optimization* 20 (6): 3,555–76.