Ahn, SeHyoun; Kaplan, Greg; Moll, Benjamin; Winberry, Thomas; Wolf, Christian

Working Paper

When Inequality Matters for Macro and Macro Matters for Inequality

CESifo Working Paper, No. 6581

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Ahn, SeHyoun; Kaplan, Greg; Moll, Benjamin; Winberry, Thomas; Wolf, Christian (2017) : When Inequality Matters for Macro and Macro Matters for Inequality, CESifo Working Paper, No. 6581, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/167567

CESifo WORKING PAPERS

6581
2017

July 2017

When Inequality Matters for Macro and Macro Matters for Inequality

SeHyoun Ahn, Greg Kaplan, Benjamin Moll, Thomas Winberry, Christian Wolf

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editors: Clemens Fuest, Oliver Falck, Jasmin Gröschl
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 6581*
Category 7: Monetary Policy and International Finance

When Inequality Matters for Macro and Macro Matters for Inequality

Abstract

We develop an efficient and easy-to-use computational method for solving a wide class of general equilibrium heterogeneous agent models with aggregate shocks, together with an open source suite of codes that implement our algorithms in an easy-to-use toolbox. Our method extends standard linearization techniques and is designed to work in cases when inequality matters for the dynamics of macroeconomic aggregates. We present two applications that analyze a two-asset incomplete markets model parameterized to match the distribution of income, wealth, and marginal propensities to consume. First, we show that our model is consistent with two key features of aggregate consumption dynamics that are difficult to match with representative agent models: (i) the sensitivity of aggregate consumption to predictable changes in aggregate income and (ii) the relative smoothness of aggregate consumption. Second, we extend the model to feature capital-skill complementarity and show how factor-specific productivity shocks shape dynamics of income and consumption inequality.

JEL-Codes: A000, C000, E000.

SeHyoun Ahn
Princeton University
USA - Princeton, NJ 08544
sehyouna@princeton.edu

Greg Kaplan
University of Chicago
USA - Chicago, IL 60637
gkaplan@uchicago.edu

Benjamin Moll
Princeton University
USA - Princeton, NJ 08544
moll@princeton.edu

Thomas Winberry
Chicago Booth
USA - Chicago, IL 60637
Thomas.Winberry@chicagobooth.edu

Christian Wolf
Princeton University
USA - Princeton, NJ 08544
ckwolf@princeton.edu

June 15, 2017
We thank Chris Carroll, Chris Sims, Jonathan Parker, Bruce Preston, Stephen Terry, and our discussant John Stachurski for useful comments. Paymon Khorrami provided excellent research assistance. The Matlab toolbox referred to in the paper is currently available at https://github.com/gregkaplan/phact.

1 Introduction

Over the last thirty years, tremendous progress has been made in developing models that reproduce salient features of the rich heterogeneity in income, wealth, and consumption behavior across households that is routinely observed in micro data. These heterogeneous agent models often deliver strikingly different implications of monetary and fiscal policies than do representative agent models, and allow us to study the distributional implications of these policies across households.[1] In principle, this class of models can therefore incorporate the rich interaction between inequality and the macroeconomy that characterizes our world: on the one hand, inequality shapes macroeconomic aggregates; on the other hand, macroeconomic shocks and policies also affect inequality.

Despite providing a framework for thinking about these important issues, heterogeneous agent models are not yet part of policy makers' toolbox for evaluating the macroeconomic and distributional consequences of their proposed policies. Instead, most quantitative analyses of the macroeconomy, particularly in central banks and other policy institutions, still employ representative agent models. Applied macroeconomists tend to make two excuses for this abstraction. First, they argue that the computational difficulties involved in solving and analyzing heterogeneous agent models render their use intractable, especially compared to the ease with which they can analyze representative agent models using software packages like `Dynare`. Second, there is a perception among macroeconomists that models which incorporate realistic heterogeneity are unnecessarily complicated because they generate only limited additional explanatory power for aggregate phenomena. Part of this perception stems from the seminal work of Krusell and Smith (1998), who found that the business cycle properties of aggregates in a baseline heterogeneous agent model are virtually indistinguishable from those in the representative agent counterpart.[2,3]

Our paper's main message is that both of these excuses are less valid than commonly

[1] For examples studying fiscal policy, see McKay and Reis (2013) and Kaplan and Violante (2014); for monetary policy, see McKay, Nakamura and Steinsson (2015), Auclert (2014), and Kaplan, Moll and Violante (2016).

[2] More precisely, in Krusell and Smith (1998)'s baseline model, which is a heterogenous agent version of a standard Real Business Cycle (RBC) model with inelastic labor supply, the effects of technology shocks on aggregate output, consumption and investment are indistinguishable from those in the RBC model.

[3] Lucas (2003) succinctly captures many macroeconomists' view when he summarizes Krusell and Smith's findings as follows: "For determining the behavior of aggregates, they discovered, realistically modeled household heterogeneity just does not matter very much. For individual behavior and welfare, of course, heterogeneity is everything." Interestingly, there is a discrepancy between this perception and the results in Krusell and Smith (1998): they show that an extension of their baseline model with preference heterogeneity, thereby implying a more realistic wealth distribution, "features aggregate time series that depart significantly from permanent income behavior."

thought. To this end, we make two contributions. First, we develop an efficient and easy-to-use computational method for solving a wide class of general equilibrium heterogeneous agent macro models with aggregate shocks, thereby invalidating the first excuse. Importantly, our method also applies in environments that violate what Krusell and Smith (1998) have termed "approximate aggregation", i.e. that macroeconomic aggregates can be well described using only the mean of the wealth distribution.

Second, we use the method to analyze the time series behavior of a rich two-asset heterogeneous agent model parameterized to match the distribution of income, wealth, and marginal propensities to consume (MPCs) in the micro data. We show that the model is consistent with two features of the time-series of aggregate consumption that have proven to be a challenge for representative agent models: consumption responds to predictable changes in income but at the same time is substantially less volatile than realized income. We then demonstrate how a quantitatively plausible heterogeneous agent economy such as ours can be useful in understanding the distributional consequences of aggregate shocks, thus paving the way for a complete analysis of the transmission of shocks to inequality. These results invalidate the second excuse: not only does macro matter for inequality, but inequality also matters for macro. We therefore view an important part of the future of macroeconomics as the study of distributions – the representative-agent shortcut may both miss a large part of the story (the distributional implications) and get the small remaining part wrong (the implications for aggregates).

In Section 2, we introduce our computational methodology, which extends standard linearization techniques, routinely used to solve representative agent models, to the heterogeneous agent context.[4] For pedagogical reasons, we describe our methods in the context of the Krusell and Smith (1998) model, but the methods are applicable much more broadly. We first solve for the stationary equilibrium of the model *without* aggregate shocks (but with idiosyncratic shocks) using a global non-linear approximation. We use the finite difference method of Achdou et al. (2015) but, in principle, other methods can be used as well. This approximation gives a discretized representation of the model's stationary equilibrium, which includes a non-degenerate distribution of agents over their individual state variables. We then compute a first-order Taylor expansion of the discretized model *with* aggregate shocks around the stationary equilibrium. This results in a large, but linear, system of stochastic differential equations, which we solve using standard solution techniques. Although our

[4]As we discuss in more detail below, the use of linearization to solve heterogeneous agent economies is not new. Our method builds on the ideas of Dotsey, King and Wolman (1999), Campbell (1998), Veracierto (2002), and Reiter (2009), and is related to Preston and Roca (2007). In contrast to these contributions, we cast our linearization method in continuous time. While discrete time poses no conceptual difficulty, working in continuous time has a number of numerical advantages that we heavily exploit.

solution method relies on linearization with respect to the economy's aggregate state variables, it preserves important non-linearities at the micro level. In particular, the response of macroeconomic aggregates to aggregate shocks may depend on the distribution of households across idiosyncratic states because of heterogeneity in the response to the shock across the distribution.

Our solution method is both faster and more accurate than existing methods. Of the five solution methods for the Krusell and Smith (1998) model included in the *Journal of Economic Dynamics and Control* comparison project (Den Haan (2010)), the fastest takes around 7 minutes to solve. With the same calibration our model takes around a quarter of a second to solve. The most accurate method in the comparison project has a maximum aggregate policy rule error of 0.16% (Den Haan (2010)'s preferred accuracy metric). With a standard deviation of productivity shocks that is comparable to the Den Haan, Judd and Julliard (2010) calibration, the maximum aggregate policy rule error using our method is 0.05%. Since our methodology uses a linear approximation with respect to aggregate shocks, the accuracy worsens as the standard deviation of shocks increases.[5]

However, the most important advantage of our method is not its speed or accuracy for solving the Krusell and Smith (1998) model. Rather, it is the potential for solving much larger models in which approximate aggregation does not hold and existing methods are infeasible. An example is the two-asset model of Kaplan, Moll and Violante (2016), where the presence of three individual state variables renders the resulting linear system so large that it is numerically impossible to solve. In order to be able to handle larger models such as this, in Section 3 we develop a model-free reduction method to reduce the dimensionality of the system of linear stochastic differential equations that characterizes the equilibrium. Our method generalizes Krusell and Smith (1998)'s insight that only a small subset of the information contained in the cross-sectional distribution of agents across idiosyncratic states is required to accurately forecast the variables that agents need to know in order to solve their decision problems. Krusell and Smith (1998)'s procedure posits a set of moments that capture this information based on economic intuition, and verifies its accuracy ex-post using a forecast-error metric; our method instead leverages advances in engineering to allow the computer to identify the necessary information in a completely model-free way.[6]

To make these methods as accessible as possible, and to encourage the use of heteroge-

[5]See Table 16 of Den Haan (2010). See Section 2 for a description of this error metric and how we compare our continuous-state, continuous-time productivity process with the two-state, discrete-time productivity process in Den Haan (2010)

[6]More precisely, we apply tools from the so-called *model reduction* literature, in particular Amsallem and Farhat (2011) and Antoulas (2005). We build on Reiter (2010) who first applied these ideas to reduce the dimensionality of linearized heterogeneous agent models in economics.

neous agent models among researchers and policy-makers, we are publishing an open source suite of codes that implement our algorithms in an easy-to-use toolbox.[7] Users of the codes provide just two inputs: (i) a function that evaluates the discretized equilibrium conditions; and (ii) the solution to the stationary equilibrium *without* aggregate shocks. Our toolbox then solves for the equilibrium of the corresponding economy *with* aggregate shocks – linearizes the model, reduces the dimensionality, solves the system of stochastic differential equations and produces impulse responses.[8]

In Sections 5 and 6 we use our toolbox to solve a two-asset heterogeneous agent economy inspired by Kaplan and Violante (2014) and Kaplan, Moll and Violante (2016), in which households can save in liquid and illiquid assets. In equilibrium, illiquid assets earn a higher return than liquid assets because they are subject to a transaction cost. This economy naturally generates "wealthy hand-to-mouth" households – households who endogenously choose to hold all their wealth as illiquid assets, and to set their consumption equal to their disposable income. Such households have high MPCs, in line with empirical evidence presented in Johnson, Parker and Souleles (2006), Parker et al. (2013) and Fagereng, Holm and Natvik (2016). Because of the two-asset structure and the presence of the wealthy hand-to-mouth, the parameterized model can match key features of the joint distribution of household portfolios and MPCs - properties that one-asset models have difficulty in replicating.[9] Matching these features of the data leads to a failure of approximate aggregation, which together with the model's size, render it an ideal setting to illustrate the power of our methods. To the best of our knowledge, this model cannot be solved using any existing methods.

In our first application (Section 5) we show that inequality can matter for macro aggregates. We demonstrate that the response of aggregate consumption to an aggregate productivity shock is larger and more transitory than in either the corresponding representative

[7]The codes are initially available as a `Matlab` toolbox at https://github.com/gregkaplan/phact, but we hope to make them available in other languages in future releases. Also see the Heterogeneous Agent Resource and toolKit (HARK) by Carroll et al. (2016, available at https://github.com/econ-ark/HARK) for another project that shares our aim of encouraging the use of heterogeneous agent models among researchers and policy-makers by making computations easier and faster.

[8]We describe our methodology in the context of incomplete markets models with heterogeneous households, but the toolbox is applicable for a much broader class of models. Essentially any high dimensional model in which equilibrium objects are a smooth function of aggregate states can be handled with the linearization methods.

[9]One-asset heterogeneous agent models models, in the spirit of Aiyagari (1994) and Krusell and Smith (1998), endogenize the fraction of hand-to-mouth households with a simple borrowing constraint. Standard calibrations of these models which match the aggregate capital-income ratio feature far too few high-MPC households relative to the data. In contrast when these models are calibrated to only *liquid* wealth, they are better able to match the distribution of MPCs in the data. Such economies, however, grossly understate the level of aggregate capital, and so are ill-suited to general equilibrium settings. They also miss almost the entire wealth distribution, and so are of limited use in studying the effects of macro shocks on inequality.

agent or one-asset heterogeneous agent economies, whereas a shock to productivity *growth* is substantially smaller and more persistent in the two-asset economy than in either the corresponding representative agent or one-asset heterogeneous agent economies. Matching the wealth distribution, in particular the consumption-share of hand-to-mouth households, drives these findings – since hand-to-mouth households are limited in their ability to immediately increase consumption in response to higher future income growth, their impact consumption response is weaker, and their lagged consumption response is stronger, than the response of non hand-to-mouth households. An implication of these individual-level consumption dynamics is that the two-asset model outperforms the representative agent models in terms of its ability to match the smoothness and sensitivity of aggregate consumption.[10] Jointly matching these two features of aggregate consumption dynamics has posed a challenge for many benchmark models in the literature (Campbell and Mankiw (1989),Christiano (1989), Ludvigson and Michaelides (2001)).

In our second application (Section 6) we show that macro shocks can additionally matter for inequality, resulting in rich interactions between inequality and the macroeconomy. To clearly highlight how quantitatively realistic heterogeneous agent economies such as ours can be useful in understanding the distributional consequences of aggregate shocks, in Section 6 we relax the typical assumption in incomplete market models that the cross-sectional distribution of labor income is exogenous. We adopt a nested CES production function with capital-skill complementarity as in Krusell et al. (2000), in which high skilled workers are more complementary with capital in production than are low skilled workers. First, we show how a negative shock to the productivity of unskilled labor generates a recession that disproportionately hurts low-skilled workers, thus also leading to an increase in income and consumption inequality. Second, we show how a positive shock to the productivity of capital generates a boom that disproportionately benefits high-skilled workers, thus leading to an increase in income and consumption inequality. The response of aggregate consumption to both of these aggregate shocks differs dramatically from that in the representative agent counterpart, thereby providing a striking counterexample to the main result of Krusell and Smith (1998). These findings illustrate how different aggregate shocks shape the dynamics of inequality and may generate rich interactions between inequality and macroeconomic aggregates.

[10] "Sensitivity" is a term used to describe how aggregate consumption responds more to predictable changes in aggregate income than implied by benchmark representative agent economies. "Smoothness" is a term used to describe how aggregate consumption growth is less volatile, relative to aggregate income growth, than implied by benchmark representative agent economies.

2 Linearizing Heterogeneous Agent Models

We present our computational method in two steps. First, in this section we describe our approach to linearizing heterogeneous agent models. Second, in Section 3 we describe our model-free reduction method for reducing the size of the linearized system. We separate the two steps because the reduction step is only necessary for large models.

We describe our method in the context of the Krusell and Smith (1998) model. This model is a natural expository tool because it is well-known and substantially simpler than the two-asset model in Section 5. As we show in Section 5, our method is applicable to a broad class of models.

Continuous Time We present our method in continuous time. While discrete time poses no conceptual difficulty (in fact, Campbell (1998), Dotsey, King and Wolman (1999), Veracierto (2002), and Reiter (2009) originally proposed this general approach in discrete time), working in continuous time has three key numerical advantages that we heavily exploit.

First, it is easier to capture occasionally binding constraints and inaction in continuous time than in discrete time. For example, the borrowing constraint in the Krusell and Smith (1998) model below is absorbed into a simple boundary condition on the value function and therefore the first-order condition for consumption holds with equality everywhere in the interior of the state space. Occasionally binding constraints and inaction are often included in heterogeneous agent models in order to match features of micro data.

Second, first-order conditions characterizing optimal policy functions typically have a simpler structure than in discrete time and can often be solved by hand.

Third, and most importantly in practice, continuous time naturally generates sparsity in the matrices characterizing the model's equilibrium conditions. Intuitively, continuously moving state variables like wealth only drift an infinitesimal amount in an infinitesimal unit of time, and therefore a typical approximation that discretizes the state space has the feature that households reach only states that directly neighbor the current state. Our two-asset model in Section 5 is so large that sparsity is necessary to store and manipulate these matrices.[11]

[11] As Reiter (2010) notes in his discussion of a related method "For reasons of computational efficiency, the transition matrix [...] should be sparse. With more than 10,000 state variables, a dense [transition matrix] might not even fit into computer memory. Economically this means that, from any given individual state today (a given level of capital, for example), there is only a small set of states tomorrow that the agent can reach with positive probability. The level of sparsity is usually a function of the time period. A model at monthly frequency will probably be sparser, and therefore easier to handle, then a model at

2.1 Model Description

Environment There is a continuum of households with fixed mass indexed by $j \in [0, 1]$ who have preferences represented by the expected utility function

$$\mathbb{E}_0 \int_0^\infty e^{-\rho t} \frac{c_{jt}^{1-\theta}}{1-\theta} dt,$$

where ρ is the rate of time preference and θ is the coefficient of relative risk aversion. At each instant t, a household's idiosyncratic labor productivity is $z_{jt} \in \{z_L, z_H\}$ with $z_L < z_H$. Households switch between the two values for labor productivity according to a Poisson process with arrival rates λ_L and λ_H.[12] The aggregate supply of efficiency units of labor is exogenous and constant and denoted by $\bar{N} = \int_0^1 z_{jt} dj$. A household with labor productivity z_{jt} earns labor income $w_t z_{jt}$. Markets are incomplete; households can only trade in productive capital a_{jt} subject to the borrowing constraint $a_{jt} \geq 0$.

There is a representative firm which has access to the Cobb-Douglas production function

$$Y_t = e^{Z_t} K_t^\alpha N_t^{1-\alpha},$$

where Z_t is (the logarithm of) aggregate productivity, K_t is aggregate capital and N_t is aggregate labor. The logarithm of aggregate productivity follows the Ornstein-Uhlenbeck process

$$\mathrm{d}Z_t = -\eta Z_t \mathrm{d}t + \sigma \mathrm{d}W_t, \tag{1}$$

where $\mathrm{d}W_t$ is the innovation to a standard Brownian motion, η is the rate of mean reversion, and σ captures the size of innovations.[13]

Equilibrium In equilibrium, household decisions depend on individual state variables, specific to a particular household, and aggregate state variables, which are common to all households. The individual state variables are capital holdings a and idiosyncratic labor

annual frequency." We take this logic a step further by working with a continuous-time model. As Reiter's discussion makes clear, discrete-time models can also generate sparsity in particular cases. However, this will happen either in models with very short time periods (as suggested by Reiter) which are known to be difficult to solve because the discount factor of households is close to one; or the resulting matrices will be sparse but with a considerably higher *bandwidth* or *density* than in the matrices generated by a continuous time model. A low bandwidth is important for efficiently solving sparse linear systems.

[12] The assumption that idiosyncratic shocks follow a Poisson process is for simplicity of exposition; the method can also handle diffusion or jump-diffusion shock processes.

[13] This process is the analog of an AR(1) process in discrete time.

productivity z. The aggregate state variables are aggregate productivity Z_t and the cross-sectional distribution of households over their individual state variables, $g_t(a,z)$.

For notational convenience, we denote the dependence of a given equilibrium object on a particular realization of the aggregate state $(g_t(a,z), Z_t)$ with a subscript t. That is, we use time-dependent notation with respect to those aggregate states. In contrast, we use recursive notation with respect to the idiosyncratic states (a,z). This notation anticipates our solution method which linearizes with respect to the aggregate states but not the idiosyncratic states.[14] An equilibrium of the model is characterized by the following equations:

$$\begin{aligned}\rho v_t(a,z) = \max_c \; & u(c) + \partial_a v_t(a,z)(w_t z + r_t a - c) \\ & + \lambda_z(v_t(a,z') - v_t(a,z)) + \frac{1}{\mathrm{d}t}\mathbb{E}_t[\mathrm{d}v_t(a,z)], \quad a \geq 0\end{aligned} \tag{2}$$

$$\frac{\mathrm{d}g_t(a,z)}{\mathrm{d}t} = -\partial_a[s_t(a,z)g_t(a,z)] - \lambda_z g_t(a,z) + \lambda_{z'} g_t(a,z'), \tag{3}$$

$$\mathrm{d}Z_t = -\eta Z_t \mathrm{d}t + \sigma \mathrm{d}W_t, \tag{4}$$

$$w_t = (1-\alpha)e^{Z_t}K_t^{\alpha}\bar{N}^{-\alpha}, \tag{5}$$

$$r_t = \alpha e^{Z_t}K_t^{\alpha-1}\bar{N}^{1-\alpha} - \delta, \tag{6}$$

$$K_t = \int a g_t(a,z)\mathrm{d}a\mathrm{d}z. \tag{7}$$

and where $s_t(a,z) = w_t z + r_t a - c$ is the optimal saving policy function corresponding to the household optimization problem (2).

For detailed derivations of these equations, see Achdou et al. (2015). The household's Hamilton-Jacobi-Bellman equation (2) is the continuous-time analog of the discrete time Bellman equation. The flow value of a household's lifetime utility is given by the sum of four terms: the flow utility of consumption, the marginal value of savings, the expected change due to idiosyncratic productivity shocks, and the expected change due to aggregate productivity shocks. Due to our use of time-dependent notation with respect to aggregate states, $\mathbb{E}_t$ denotes the conditional expectation with respect to aggregate states only.[15] The Kolmogorov Forward Equation (3) describes the evolution of the distribution over time. The flow change in the mass of households at a given point in the state space is determined by their savings behavior and idiosyncratic productivity shocks. Equation (4) describes the

[14] Appendix A.1 writes the equilibrium conditions using fully recursive condition and shows how to obtain the system here by evaluating these conditions "along the characteristic" $(g_t(a,z), Z_t)$.

[15] The borrowing constraint only affects (2) through the boundary condition $u'(w_t z_i) \geq \partial_a v_t(0,z)$ for $i = L, H$. We impose this condition in our numerical computations, but for the ease of exposition suppress the notation here.

evolution of aggregate productivity. Finally, equations (5) to (7) define prices given the aggregate state.

We define a *steady state* as an equilibrium with constant aggregate productivity $Z_t = 0$ and a time-invariant distribution $g(a, z)$. The steady state system is given by

$$\rho v(a,z) = \max_c \; u(c) + \partial_a v(a,z)(wz + ra - c) + \lambda_z (v(a,z') - v(a,z)), \quad a \geq 0 \tag{8}$$

$$0 = -\partial_a [s(a,z) g(a,z)] - \lambda_z g(a,z) + \lambda_{z'} g(a,z'), \tag{9}$$

$$w = (1-\alpha) K^{\alpha} \bar{N}^{-\alpha}, \tag{10}$$

$$r = \alpha K^{\alpha-1} \bar{N}^{1-\alpha} - \delta, \tag{11}$$

$$K = \int a g(a,z) \mathrm{d}a \mathrm{d}z. \tag{12}$$

2.2 Linearization Procedure

Our linearization procedure consists of three steps. First, we solve for the steady state of the model without aggregate shocks but with idiosyncratic shocks. Second, we take a first-order Taylor expansion of the equilibrium conditions around the steady state, yielding a linear system of stochastic differential equations. Third, we solve the linear system using standard techniques. Conceptually, each of these steps is a straightforward extension of standard linearization techniques to the heterogeneous agent context. However, the size of heterogeneous agent models leads to a number of computational challenges which we address.

Step 1: Approximate Steady State Because households face idiosyncratic uncertainty, the steady state value function varies over individual state variables $v(a, z)$, and there is a non-degenerate stationary distribution of households $g(a, z)$. To numerically approximate these functions we must represent them in a finite-dimensional way. We use a non-linear approximation in order to retain the rich non-linearities and heterogeneity at the individual level. In principle, any approximation method can be used in this step; we use the finite difference methods outlined in Achdou et al. (2015) because they are fast, accurate, and robust.

We approximate the value function and distribution over a discretized grid of asset holdings $\mathbf{a} = (a_1 = 0, a_2, ..., a_I)^{\mathrm{T}}$. Denote the value function and distribution along this discrete grid using the vectors $\mathbf{v} = (v(a_1, z_L), ..., v(a_I, z_H))^{\mathrm{T}}$ and $\mathbf{g} = (g(a_1, z_L), ..., g(a_I, z_H))^{\mathrm{T}}$; both $\mathbf{v}$ and $\mathbf{g}$ are of dimension $N \times 1$ where $N = 2I$ is the total number of grid points in the individual state space. We solve the steady state versions of (2) and (3) at each point on

this grid, approximating the partial derivatives using finite differences. Achdou et al. (2015) show that if the finite difference approximation is chosen correctly, the discretized steady state is the solution to the following system of matrix equations:

$$\begin{aligned} \rho \mathbf{v} &= \mathbf{u}(\mathbf{v}) + \mathbf{A}(\mathbf{v};\mathbf{p})\,\mathbf{v} \\ \mathbf{0} &= \mathbf{A}(\mathbf{v};\mathbf{p})^{\mathrm{T}}\,\mathbf{g} \\ \mathbf{p} &= \mathbf{F}(\mathbf{g}). \end{aligned} \tag{13}$$

The first equation is the approximated steady state HJB equation (8) for each point on the discretized grid, expressed in our vector notation. The vector $\mathbf{u}(\mathbf{v})$ is the maximized utility function over the grid and the matrix multiplication $\mathbf{A}(\mathbf{v};\mathbf{p})\,\mathbf{v}$ captures the remaining terms in (8). The second equation is the discretized version of the steady state Kolmogorov Forward equation (9). The transition matrix $\mathbf{A}(\mathbf{v};\mathbf{p})$ is simply the transpose of the matrix from the discretized HJB equation because it encodes how households move around the individual state space. Finally, the third equation defines the prices $\mathbf{p} = (r, w)^{\mathrm{T}}$ as a function of aggregate capital through the distribution $\mathbf{g}$.[16]

Since $\mathbf{v}$ and $\mathbf{g}$ each have N entries, the total system has $2N + 2$ equations in $2N + 2$ unknowns. In simple models like this one, highly accurate solutions can be obtained with as little as $N = 200$ grid points (i.e., $I = 100$ asset grid points together with the two income states); however, in more complicated models, such as the two-asset model in Section 5, N can easily grow into the tens of thousands. Exploiting the sparsity of the transition matrix $\mathbf{A}(\mathbf{v};\mathbf{p})$ is necessary to even represent the steady state of such large models.

Step 2: Linearize Equilibrium Conditions The second step of our method is to compute a first-order Taylor expansion of the model's discretized equilibrium conditions around steady state. With aggregate shocks, the discretized equilibrium is characterized by

$$\begin{aligned} \rho \mathbf{v}_t &= \mathbf{u}(\mathbf{v}_t) + \mathbf{A}(\mathbf{v}_t;\mathbf{p}_t)\,\mathbf{v}_t + \frac{1}{\mathrm{d}t}\mathbb{E}_t \mathrm{d}\mathbf{v}_t \\ \frac{\mathrm{d}\mathbf{g}_t}{\mathrm{d}t} &= \mathbf{A}(\mathbf{v}_t;\mathbf{p}_t)^{\mathrm{T}}\,\mathbf{g}_t \\ \mathrm{d}Z_t &= -\eta Z_t \mathrm{d}t + \sigma \mathrm{d}W_t \\ \mathbf{p}_t &= \mathbf{F}(\mathbf{g_t}; Z_t). \end{aligned} \tag{14}$$

[16] The fact that prices are an explicit function of the distribution is a special feature of the Krusell and Smith (1998) model. In general, market clearing conditions take the form $\mathbf{F}(\mathbf{v}, \mathbf{g}, \mathbf{p}) = \mathbf{0}$. Our solution method also handles this more general case.

The system (14) is a non-linear system of $2N+3$ stochastic differential equations in $2N+3$ variables (the $2N+2$ variables from the steady state, plus aggregate productivity Z_t). Shocks to TFP Z_t induce fluctuations in marginal products and therefore prices $\mathbf{p}_t = \mathbf{F}(\mathbf{g_t}; Z_t)$. Fluctuations in prices in turn induce fluctuations in households' decisions and therefore in $\mathbf{v}_t$ and the transition matrix $\mathbf{A}(\mathbf{v}_t; \mathbf{p}_t)$.[17] Fluctuations in the transition matrix then induce fluctuations in the distribution of households $\mathbf{g}_t$.

The key insight is that this large-dimensional system of stochastic differential equations has exactly the same structure as more standard representative agent models which are normally solved by means of linearization methods. To make this point, Appendix A.2 relates the system (14) to the real business cycle (RBC) model. The discretized value function points $\mathbf{v}_t$ are jump variables, like aggregate consumption C_t in the RBC model. The discretized distribution $\mathbf{g}_t$ points are endogenous state variables, like aggregate capital K_t in the RBC model. TFP Z_t is an exogenous state variable. Finally, the wage and real interest rate are statically defined variables, just as in the Krusell and Smith (1998) model.

As already anticipated, we exploit this analogy and solve the non-linear system (14) by linearizing it around the steady state. Since the dimension of the system is large it is impossible to compute derivatives by hand. We use a recently developed technique called automatic (or algorithmic) differentiation that is fast and accurate up to machine precision. It dominates finite differences in terms of accuracy and symbolic differentiation in terms of speed. Automatic differentiation exploits the fact that the computer represents any function as the composition of various elementary functions, such as addition, multiplication, or exponentiation, which have known derivatives. It builds the derivative of the original function by iteratively applying the chain rule. This allows automatic differentiation to exploit the sparsity of the transition matrix $\mathbf{A}(\mathbf{v}_t;\mathbf{p}_t)$ when taking derivatives, which is essential for numerical feasibility in large models.[18]

[17] We have written the price vector $\mathbf{p}_t$ as a function of the state vector to easily exposit our methodology in a way that directly extends to models with more general market clearing conditions (see footnote 16). However, this approach is not necessary in the Krusell and Smith (1998) model because we can simply substitute the expression for prices directly into the households' budget constraint and hence the matrix $\mathbf{A}(\mathbf{v}_t; \mathbf{p}_t)$.

[18] To the best of our knowledge, there is no existing open-source automatic differentiation package for `Matlab` which exploits sparsity. We therefore wrote our own package for the computational toolbox.

The first-order Taylor expansion of (14) can be written as:[19]

$$\mathbb{E}_t \begin{bmatrix} \mathrm{d}\widehat{\mathbf{v}}_t \\ \mathrm{d}\widehat{\mathbf{g}}_t \\ \mathrm{d}Z_t \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{vv} & \mathbf{0} & 0 & \mathbf{B}_{vp} \\ \mathbf{B}_{gv} & \mathbf{B}_{gg} & 0 & \mathbf{B}_{gp} \\ \mathbf{0} & \mathbf{0} & -\eta & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{pg} & \mathbf{B}_{pZ} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{v}}_t \\ \widehat{\mathbf{g}}_t \\ Z_t \\ \widehat{\mathbf{p}}_t \end{bmatrix} \mathrm{d}t \tag{15}$$

The variables in the system, $\widehat{\mathbf{v}}_t$, $\widehat{\mathbf{g}}_t$, Z_t and $\widehat{\mathbf{p}}_t$, are expressed as deviations from their steady state values, and the matrix is composed of the derivatives of the equilibrium conditions evaluated at steady state. Since the pricing equations are static, the fourth row of this matrix equation only has non-zero entries on the right hand side.[20] It is convenient to plug the pricing equations $\widehat{\mathbf{p}}_t = \mathbf{B}_{pg}\widehat{\mathbf{g}}_t + \mathbf{B}_{pZ}Z_t$ into the remaining equations of the system, yielding

$$\mathbb{E}_t \begin{bmatrix} \mathrm{d}\widehat{\mathbf{v}}_t \\ \mathrm{d}\widehat{\mathbf{g}}_t \\ \mathrm{d}Z_t \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{B}_{vv} & \mathbf{B}_{vp}\mathbf{B}_{pg} & \mathbf{B}_{vp}\mathbf{B}_{pZ} \\ \mathbf{B}_{gv} & \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg} & \mathbf{B}_{gp}\mathbf{B}_{pZ} \\ \mathbf{0} & \mathbf{0} & -\eta \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} \widehat{\mathbf{v}}_t \\ \widehat{\mathbf{g}}_t \\ Z_t \end{bmatrix} \mathrm{d}t. \tag{16}$$

Step 3: Solve Linear System The final step of our method is to solve the linear system of stochastic differential equations (16). Following standard practice, we perform a Schur decomposition of the matrix $\mathbf{B}$ to identify the stable and unstable roots of the system. If the Blanchard and Kahn (1980) condition holds, i.e., the number of stable roots equals the

[19] To arrive at (15), we first rearrange (14) so that all time derivatives are on the left-hand side. We then take the expectation of the entire system and use the fact that the expectation of a Brownian increment is zero $\mathbb{E}_t[\mathrm{d}W_t] = 0$ to write (14) compactly without the stochastic term as

$$\mathbb{E}_t \begin{bmatrix} \mathrm{d}\mathbf{v}_t \\ \mathrm{d}\mathbf{g}_t \\ \mathrm{d}Z_t \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{u}\left(\mathbf{v}_t;\mathbf{p}_t\right) + \mathbf{A}\left(\mathbf{v}_t;\mathbf{p}_t\right)\mathbf{v}_t - \rho\mathbf{v}_t \\ \mathbf{A}\left(\mathbf{v}_t;\mathbf{p}_t\right)^{\mathrm{T}}\mathbf{g}_t \\ -\eta Z_t \\ \mathbf{F}\left(\mathbf{g_t};Z_t\right) - \mathbf{p}_t \end{bmatrix} \mathrm{d}t.$$

Finally, we linearize this system to arrive at (15). Note that this compact notation loses the information contained in the stochastic term $\mathrm{d}W_t$. However, since we linearize the system, this is without loss of generality – as we discuss later linearized systems feature certainty equivalence.

[20] The special structure of the matrix $\mathbf{B}$ involving zeros is particular to the Krusell and Smith (1998) model and can be relaxed. In addition, the fact that we can express prices as a static function of $\widehat{\mathbf{g}}_t$ and Z_t is a special feature of the model; more generally, the equilibrium prices are only defined implicitly by a set of market clearing conditions.

number of state variables $\widehat{\mathbf{g}}_t$ and Z_t, then we can compute the solution:

$$\begin{aligned}
\widehat{\mathbf{v}}_t &= \mathbf{D}_{vg}\widehat{\mathbf{g}}_t + \mathbf{D}_{vZ}Z_t, \\
\frac{\mathrm{d}\widehat{\mathbf{g}}_t}{\mathrm{d}t} &= (\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg} + \mathbf{B}_{gv}\mathbf{D}_{vg})\widehat{\mathbf{g}}_t + (\mathbf{B}_{gp}\mathbf{B}_{pZ} + \mathbf{B}_{gv}\mathbf{D}_{vZ})Z_t, \\
\mathrm{d}Z_t &= -\eta Z_t \mathrm{d}t + \sigma \mathrm{d}W_t, \\
\widehat{\mathbf{p}}_t &= \mathbf{B}_{pg}\widehat{\mathbf{g}}_t + \mathbf{B}_{pZ}Z_t.
\end{aligned} \tag{17}$$

The first line of (17) sets the control variables $\widehat{\mathbf{v}}_t$ as functions of the state variables $\widehat{\mathbf{g}}_t$ and Z_t, i.e. the matrices $\mathbf{D}_{vg}$ and $\mathbf{D}_{vZ}$ characterize the optimal decision rules as a function of aggregate states. The second line plugs that solution into the system (16) to compute the evolution of the distribution. The third line is the stochastic process for the aggregate productivity shock and the fourth line is the definition of prices $\widehat{\mathbf{p}}_t$.

2.3 What Does Linearization Capture and What Does It Lose?

Our method uses a mix of nonlinear approximation with respect to individual state variables and linear approximation with respect to aggregate state variables. Concretely, from the first line of (17), the approximated solution for the value function is of the form

$$v_t(a_i, z_j) = v(a_i, z_j) + \sum_{k=1}^{I}\sum_{\ell=1}^{2} \mathbf{D}_{vg}[i,j;k,l](g_t(a_k, z_\ell) - g(a_k, z_\ell)) + \mathbf{D}_{vZ}[i,j]Z_t, \tag{18}$$

where $\mathbf{D}_{vg}[i,j;k,l]$ and $\mathbf{D}_{vZ}[i,j]$ denote the relevant elements of $\mathbf{D}_{vg}$ and $\mathbf{D}_{vZ}$, and $v(a,z)$ and $g(a,z)$ are the steady state value function and distribution. Given the value function $v_t(a_i, z_j)$, optimal consumption at different points of the income and wealth distribution is then given by

$$c_t(a_i, z_j) = (\partial_a v_t(a_i, z_j))^{-1/\theta}. \tag{19}$$

Certainty Equivalence Expressions (18) and (19) show that our solution features *certainty equivalence* with respect to aggregate shocks; the standard deviation σ of aggregate TFP Z_t does not enter households' decision rules.[21] This is a generic feature of all linearization techniques.

However, our solution does *not* feature certainty equivalence with respect to idiosyncratic shocks, because the distribution of idiosyncratic shocks enters the HJB equation (2) as well

[21] Note that σ does not enter the matrix $\mathbf{B}$ characterizing the linearized system (16) and therefore also does not enter the matrices characterizing the optimal decision rules $\mathbf{D}_{vg}$ and $\mathbf{D}_{vZ}$.

as its linearized counterpart in (16) directly. A corollary of this is that our method *does* capture the effect of aggregate uncertainty to the extent that aggregate shocks affect the distribution of idiosyncratic shocks. For example, Bloom et al. (2014) and Bayer et al. (2015) study the effect of "uncertainty shocks" that result in an increase in the dispersion of idiosyncratic shocks and can be captured by our method.[22]

Our solution method may instead be less suitable for various asset-pricing applications in which the direct effect of aggregate uncertainty on individual decision rules is key. In future work we hope to encompass such applications by extending our first-order perturbation method to higher orders, or by allowing the decision rules to depend non-linearly on relevant low-dimensional aggregate state variables (but not the high-dimensional distribution). Yet another strategy could be to assume that individuals are averse to ambiguity so that risk premia survive linearization (Ilut and Schneider, 2014).

Distributional Dependence of Aggregates A common motivation for studying heterogeneous agent models is that the response of macroeconomic aggregates to aggregate shocks may depend on the distribution of idiosyncratic states. For example, different joint distributions of income and wealth $g(a, z)$ can result in different impulse responses of aggregates to the same aggregate shock. Our solution method preserves such *distributional dependence*.

To fix ideas, consider the impulse response of aggregate consumption C_t to a productivity shock Z_t, starting from the steady-state distribution $g(a, z)$. First consider the response of initial aggregate consumption C_0 only. We compute the impact effect of the shock on the initial value function $v_0(a, z)$ and initial consumption $c_0(a, z)$ from (18) and (19). Integrate this over households to get aggregate consumption

$$C_0 = \int c_0(a, z)g(a, z)dadz \approx \sum_{i=1}^{I}\sum_{j=1}^{2} c_0(a_i, z_j)g(a_i, z_j)\Delta a\Delta z.$$

The impulse response of C_0 depends on the initial distribution $g_0(a, z)$ because the elasticities of individual consumption $c_0(a, z)$ with respect to the aggregate shock Z_0 are different for individuals with different levels of income and/or wealth. These individual elasticities are then aggregated according to the initial distribution. Therefore, the effect of the shock depends on the initial distribution $g_0(a, z)$.

To see this even more clearly, it is useful to briefly work with the continuous rather than

[22] McKay (2017) studies how time-varying idiosyncratic uncertainty on aggregate consumption dynamics. Terry (2017) studies how well discrete-time relatives of our method capture time-variation in the dispersion of productivity shocks in a heterogeneous firm model.

discretized value and consumption policy functions. Analogous to (18), we can write the initial value function response as $\widehat{v}_0(a,z) = \mathbf{D}_{vZ}(a,z)Z_0$ where $\mathbf{D}_{vZ}(a,z)$ are the elements of $\mathbf{D}_{vZ}$ in (17) and where we have used the fact that the initial distribution does not move (i.e. $\widehat{g}_0(a,z) = 0$) by virtue of g being a state variable. We can use this to show that the deviation of initial consumption from steady state satisfies $\widehat{c}_0(a,z) = \mathbf{D}_{cZ}(a,z)Z_0$ where $\mathbf{D}_{cZ}(a,z)$ captures the responsiveness of consumption to the aggregate shock.[23] The impulse response of initial aggregate consumption is then

$$\widehat{C}_0 = \int \mathbf{D}_{cZ}(a,z)g(a,z)dadz \times Z_0. \tag{20}$$

It depends on the steady-state distribution $g(a,z)$ since the responsiveness of individual consumption to the aggregate shock $\mathbf{D}_{cZ}(a,z)$ differs across (a,z).

Size- and Sign-Dependence Another question of interest is whether our economy features size- or sign-dependence, that is, whether it responds non-linearly to aggregate shocks of different sizes or asymmetrically to positive and negative shocks.[24] In contrast to state dependence, our linearization method eliminates any potential sign- and size-dependence. This can again be seen clearly from the impulse response of initial aggregate consumption in (20) which is linear in the aggregate shock Z_0. This immediately rules out size- and sign-dependence in the response of aggregate consumption to the aggregate shock.[25]

In future work we hope to make progress on relaxing this feature of our solution method. Extending our first-order perturbation method to higher orders would again help in this regard. Another idea is to leverage the linear model solution together with parts of the full non-linear model to simulate the model in a way that preserves these nonlinearities. In particular one could use the fully nonlinear Kolmogorov Forward equation in (14) instead of the linearized version in (16) to solve for the path of the distribution for times $t > 0$: $d\mathbf{g}_t/dt = \mathbf{A}\left(\mathbf{v}_t;\mathbf{p}_t\right)^{\mathrm{T}}\mathbf{g}_t$. This procedure allows us to preserve *size-dependence* after the initial impact $t > 0$ because larger shocks potentially induce non-proportional movements in the individual state space, and therefore different distributional dynamics going forward.[26]

[23] In particular $\mathbf{D}_{cZ}(a,z) = (\partial_a v(a,z))^{-\frac{1}{\theta}-1}\partial_a \mathbf{D}_{vZ}(a,z)$. To see this note that $\widehat{c}_0(a,z) = (\partial_a v(a,z))^{-\frac{1}{\theta}-1}\partial_a \widehat{v}_0(a,z) = (\partial_a v(a,z))^{-\frac{1}{\theta}-1}\partial_a \mathbf{D}_{vZ}(a,z)Z_0 := \mathbf{D}_{cZ}(a,z)Z_0$.

[24] Note that this is separate from the state dependence we just discussed which is concerned with how the distribution may affect the *linear* dynamics of the system.

[25] Note that expression (20) only holds at $t = 0$. At times $t > 0$, the distribution also moves $\widehat{g}_t(a,z) \neq 0$. The generalization of (20) to $t > 0$ is $\widehat{C}_t \approx \int \widehat{c}_t(a,z)g(a,z)dadz + \int c(a,z)\widehat{g}_t(a,z)dadz$. Since both $\widehat{c}_t(a,z)$ and $\widehat{g}_t(a,z)$ will be linear in Z_t, so will be $\widehat{C}_t$, again ruling out size- and sign-dependence.

[26] An open question is under what conditions this procedure would be consistent with our use of linear approximations to solve the model. One possible scenario is as follows: even though the time path for the

Small versus Large Aggregate Shocks Another generic feature of linearization techniques is that the linearized solution is expected to be a good approximation to the true non-linear solution for small aggregate shocks and less so for large ones. Section 2.4 below documents that our approximate dynamics of the distribution is accurate for the typical calibration of TFP shocks in the Krusell and Smith (1998) model, but breaks down for very large shocks.[27]

2.4 Performance of Linearization in Krusell-Smith Model

In order to compare the performance of our method to previous work, we solve the model under the parameterization of the JEDC comparison project Den Haan, Judd and Julliard (2010). A unit of time is one quarter. We set the rate of time preference $\rho = 0.01$ and the coefficient of relative risk aversion $\theta = 1$. Capital depreciates at rate $\delta = 0.025$ per quarter and the capital share is $\alpha = 0.36$. We set the levels of idiosyncratic labor productivity z_L and z_H following Den Haan, Judd and Julliard (2010).

One difference between our model and Den Haan, Judd and Julliard (2010) is that we assume aggregate productivity follows the continuous-time, continuous-state Ornstein-Uhlenbeck process (1) rather than the discrete-time, two-state Markov chain in Den Haan, Judd and Julliard (2010). To remain as consistent with Den Haan, Judd and Julliard (2010)'s calibration as possible, we choose the approximate quarterly persistence $\text{corr}(\log Z_{t+1}, \log Z_t) = e^{-\eta} \approx 1 - \eta = 0.75$ and the volatility of innovations $\sigma = 0.007$ to match the standard deviation and autocorrelation of Den Haan, Judd and Julliard (2010)'s two-state process.[28]

In our approximation we set the size of the individual asset grid $I = 100$, ranging from $a_1 = 0$ to $a_I = 100$. Together with the two values for idiosyncratic productivity, the total number of grids is $N = 200$ and the total size of the dynamic system (16) is 400.[29]

distribution might differ substantially when computed using the non-linear Kolmogorov Forward equation, the time path for prices may still be well approximated by the linearized solution. Hence, the error in the HJB equation from using the linearized prices may be small.

[27] Related, our linearization method obviously rules out nonlinear amplification effects that result in a bimodal ergodic distribution of aggregate states as in He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014).

[28] Another difference is that Den Haan, Judd and Julliard (2010) allows the process for idiosyncratic shocks to depend on the aggregate state. We set our idiosyncratic shock process to match the average transition probabilities in Den Haan, Judd and Julliard (2010). We have solved the model with time-varying transition probabilities and obtained quantitatively similar results. Details are available from the authors upon request.

[29] In this calculation, we have dropped one grid point from the distribution using the restriction that the distribution integrates to one. Hence there are $N = 200$ equations for $\widehat{\mathbf{v}}_t$, $N - 1 = 199$ equations for $\widehat{\mathbf{g}}_t$ and one equation for Z_t.

Table 1: Run Time for Solving Krusell-Smith Model

	Full Model
Steady State	0.082 sec
Derivatives	0.021 sec
Linear system	0.14 sec
Simulate IRF	0.024 sec
Total	**0.27 sec**

Notes: Time to solve Krusell-Smith model once on `MacBook Pro 2016` laptop with `3.3 GHz` processor and `16 GB RAM`, using `Matlab R2016b` and our code toolbox. "Steady state" reports time to compute steady state. "Derivatives" reports time to compute derivatives of discretized equilibrium conditions. "Linear system" reports time to solve system of linear differential equations. "Simulate IRF" reports time to simulate impulse responses reported in Figure 1. "Total" is the sum of all these tasks.

Table 1 shows that our linearization method solves the Krusell and Smith (1998) model in approximately one quarter of one second. In contrast, the fastest algorithm documented in the comparison projection by Den Haan (2010) takes over seven minutes to solve the model – more than 1500 times slower than our method (see Table 2 in Den Haan (2010)).[30] In Section 3 we solve the model in approximately 0.1 seconds using our model-free reduction method.

Accuracy of Linearization The key restriction that our method imposes is linearity with respect to the aggregate state variables Z_t and $\widehat{\mathbf{g}}_t$. We evaluate the accuracy of this approximation using the error metric suggested by Den Haan (2010). The Den Haan error metric compares the dynamics of the aggregate capital stock under two simulations of the model for $T = 10,000$ periods. The first simulation computes the path of aggregate capital K_t from our linearized solution (17). The second simulation computes the path of aggregate capital K_t^* from simulating the model using the nonlinear dynamics (3) as discussed in Section 2.3. We then compare the maximum log difference between the two series,

$$\epsilon^{\text{DH}} = 100 \times \max_{t \in [0,T]} \left| \log K_t - \log K_t^* \right|.$$

[30] As discussed by Den Haan (2010), there is one algorithm (Penal) that "is even faster, but this algorithm does not solve the actual [Krusell-Smith] model specified."

Table 2: Maximum den Haan Error in %

St. Dev Productivity Shocks (%)	Maximum den Haan Error (%)
0.01	0.000
0.1	0.001
0.7	**0.049**
1.0	0.118
5.0	3.282

Notes: Maximum percentage error in accuracy check suggested by Den Haan (2010). The error is the percentage difference between the time series of aggregate capital under our linearized solution and a nonlinear simulation of the model, as described in the main text. The bold face row denotes the calibrated value $\sigma = 0.007$.

Den Haan originally proposed this metric to compute the accuracy of the forecasting rule in the Krusell and Smith (1998) algorithm; in our method, the linearized dynamics of the distribution $\mathbf{g}_t$ are analogous to the forecasting rule.

When the standard deviation of productivity shocks is 0.7%, our method gives a maximum percentage error $\epsilon^{\text{DH}} = 0.049\%$, implying that households in our model make small errors in forecasting the distribution. Our method is three times as accurate as the Krusell and Smith (1998) method, which is the most accurate algorithm in Den Haan (2010) and gives $\epsilon^{\text{DH}} = 0.16\%$. Table 2 shows that, since our method is locally accurate, its accuracy decreases in the size of the shocks σ. However, with the size of aggregate shocks in the baseline calibration, it provides exceptional accuracy.

3 Model Reduction

Solving the linear system (16) is extremely efficient because the Krusell and Smith (1998) model is relatively small. However, the required matrix decomposition becomes prohibitively expensive in larger models like the two-asset model that we will study in Section 5. We must therefore reduce the size of the system to solve these more general models. Furthermore, even in smaller models like Krusell and Smith (1998), model reduction makes likelihood-based estimation feasible by reducing the size of the associated filtering problem.[31]

[31] Mongey and Williams (2016) use a discrete-time relative of our method without model reduction to estimate a small heterogeneous firm model. Winberry (2016) provides an alternative parametric approach for reducing the distribution and also uses it to estimate a small heterogeneous firm model.

In this section, we develop a model-free reduction method to reduce the size of the linear system while preserving accuracy. Our approach projects the high-dimensional distribution $\widehat{\mathbf{g}}_t$ and value function $\widehat{\mathbf{v}}_t$ onto low-dimensional subspaces and solves the resulting low-dimensional system. The main challenge is reducing the distribution, which we discuss in Sections 3.1, 3.2, and 3.3. Section 3.4 describes how we reduce the value function. Section 3.5 puts the two together to solve the reduced model and describes the numerical implementation. Finally, Section 3.6 shows that our reduction method performs well in the Krusell and Smith (1998) model.

In order to simplify notation, for the remainder of this section we use $\mathbf{v}_t, \mathbf{g}_t$ and $\mathbf{p}_t$ to denote the *deviations from steady state* in the value function, distribution, and prices. In Section 2, we had denoted these objects using $\widehat{\mathbf{v}}_t, \widehat{\mathbf{g}}_t$ and $\widehat{\mathbf{p}}_t$. This change of notation applies to Section 3 only, and we will remind the reader whenever this change could cause confusion.

3.1 Overview of Distribution Reduction

The basic insight that we exploit is that only a small subset of the information in $\mathbf{g}_t$ is necessary to accurately forecast the path of prices $\mathbf{p}_t$. In fact, in the discrete time version of this model, Krusell and Smith (1998) show that just the mean of the asset distribution $\mathbf{g}_t$ is sufficient to forecast $\mathbf{p}_t$ according to a forecast-error metric. However, the success of their reduction strategy relies on the economic properties of the model, so it is not obvious how to generalize it to other environments. We use a set of tools from the engineering literature known as *model reduction* to generalize Krusell and Smith (1998)'s insight in a model-free way, allowing the computer to compute the features of the distribution that are necessary to accurately forecast $\mathbf{p}_t$.[32]

It is important to note that the vector $\mathbf{p}_t$ does not need to literally consist of prices; it is simply the vector of objects we wish to accurately describe. In practice, we often also include other variables of interest, such as aggregate consumption or output, to ensure that the reduced model accurately describes their dynamics as well.

[32]The following material is based on lecture notes by Amsallem and Farhat (2011), which in turn build on a book by Antoulas (2005). Lectures 3 and 7 by Amsallem and Farhat (2011) and Chapters 1 and 11 in Antoulas (2005) are particularly relevant. All lecture notes for Amsallem and Farhat (2011) are available online at https://web.stanford.edu/group/frg/course_work/CME345/ and the book by Antoulas (2005) is available for free at http://epubs.siam.org/doi/book/10.1137/1.9780898718713. Also see Reiter (2010) who applies related ideas from the model reduction literature in order to reduce the dimensionality of a linearized discrete-time heterogeneous agent model.

3.1.1 The Distribution Reduction Problem

We say that the distribution *exactly reduces* if there exists a $k_{\mathcal{S}}$-dimensional time-invariant subspace $\mathcal{S}$ with $k_{\mathcal{S}} << N$ such that, for all distributions $\mathbf{g}_t$ which occur in equilibrium,

$$\mathbf{g}_t = \gamma_{1t}\mathbf{x}_1 + \gamma_{2t}\mathbf{x}_2 + ... + \gamma_{k_{\mathcal{S}}t}\mathbf{x}_{k_{\mathcal{S}}},$$

where $\mathbf{X}_{\mathcal{S}} = [\mathbf{x}_1, ..., \mathbf{x}_{k_{\mathcal{S}}}] \in \mathbb{R}^{N \times k_{\mathcal{S}}}$ is a basis for the subspace $\mathcal{S}$ and $\gamma_{1t}, ..., \gamma_{k_{\mathcal{S}}t}$ are scalars. If we knew the time-invariant basis $\mathbf{X}_{\mathcal{S}}$, we could decrease the dimensionality of the problem by tracking only the $k_{\mathcal{S}}-$dimensional vector of coefficients γ_t.

Typically exact reduction as described above does not hold, so we instead must estimate a *trial basis* $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_k] \in \mathbb{R}^{N \times k}$ such that the distribution *approximately reduces*, i.e.,

$$\mathbf{g}_t \approx \gamma_{1t}\mathbf{x}_1 + \gamma_{2t}\mathbf{x}_2 + ... + \gamma_{kt}\mathbf{x}_k,$$

or, in matrix form, $\mathbf{g}_t \approx \mathbf{X}\gamma_t$. Denote the resulting approximation of the distribution by $\widetilde{\mathbf{g}}_t = \mathbf{X}\gamma_t$ and the approximate prices by $\widetilde{\mathbf{p}}_t = \mathbf{B}_{pg}\widetilde{\mathbf{g}}_t + \mathbf{B}_{pZ}Z_t$.

Our model maps directly into the prototypical problem considered by the model reduction literature if the decision rules are exogenous, i.e. the matrices $\mathbf{D}_{vg}$ and $\mathbf{D}_{vZ}$ in (17) are exogenously given.[33] This case assumes away a crucial part of the economics we are interested in studying but nevertheless has pedagogical use in connecting to the existing literature. In this case, using the second and fourth equations of (17) and recalling our convention in this section to drop hats from variables, our dynamical system becomes

$$\begin{aligned} \frac{\mathrm{d}\mathbf{g}_t}{\mathrm{d}t} &= \mathbf{C}_{gg}\mathbf{g}_t + \mathbf{C}_{gZ}Z_t \\ \mathbf{p}_t &= \mathbf{B}_{pg}\mathbf{g}_t + \mathbf{B}_{pZ}Z_t, \end{aligned} \tag{21}$$

where $\mathbf{C}_{gg} = \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg} + \mathbf{B}_{gv}\mathbf{D}_{vg}$ and $\mathbf{C}_{gZ} = \mathbf{B}_{gp}\mathbf{B}_{pZ} + \mathbf{B}_{gv}\mathbf{D}_{vZ}$. This system maps a low-dimensional vector of "inputs" (aggregate productivity Z_t) into a low-dimensional vector of "outputs" (prices $\mathbf{p}_t$), intermediated through the high-dimensional distribution $\mathbf{g}_t$.[34] The model reduction literature provides an off-the-shelf set of tools to replace the high-

[33] Exogenous decision rules usually relate the value function to prices, i.e. $\mathbf{v}_t = \mathbf{D}_{vp}\mathbf{p}_t$. But prices $\mathbf{p}_t = \mathbf{B}_{pg}\mathbf{g}_t + \mathbf{B}_{pZ}Z_t$ in turn depend on the distribution $\mathbf{g}_t$ and productivity Z_t. Hence so do the decision rules: $\mathbf{v}_t = \mathbf{D}_{vg}\mathbf{g}_t + \mathbf{D}_{vZ}Z_t$, with $\mathbf{D}_{vg} = \mathbf{D}_{vp}\mathbf{B}_{pg}$ and $\mathbf{D}_{vZ} = \mathbf{D}_{vp}\mathbf{D}_{pZ}$.

[34] The system (21) is called a *linear time invariant (LTI) system.* Z_t is an *input* into the system and $\mathbf{p}_t$ is an *output.* If both inputs and outputs are scalars, the system is called a *single-input-single-output (SISO) system.* If both inputs and outputs are vectors, it is called a *multiple-input-multiple-output (MIMO) system.* Instead of assuming that decision rules are exogenous, we could have assumed that there is no feedback

dimensional "intermediating variable" $\mathbf{g}_t$ with a low-dimensional approximation γ_t while preserving the mapping from inputs to outputs.

Of course, our economic model is more complicated than this special case because the distribution reduction feeds back into agents' decisions through the endogenous value function $\mathbf{v}_t$. It is helpful to restate the system with endogenous $\mathbf{v}_t$ in a form closer to that in the model reduction literature:

$$\begin{aligned}\begin{bmatrix}\mathbb{E}_t[\mathrm{d}\mathbf{v}_t]\\ \mathrm{d}\mathbf{g}_t\end{bmatrix} &= \begin{bmatrix}\mathbf{B}_{vv} & \mathbf{B}_{vp}\mathbf{B}_{pg}\\ \mathbf{B}_{gv} & \mathbf{B}_{gg}+\mathbf{B}_{gp}\mathbf{B}_{pg}\end{bmatrix}\begin{bmatrix}\mathbf{v}_t\\ \mathbf{g}_t\end{bmatrix}\mathrm{d}t + \begin{bmatrix}\mathbf{B}_{vp}\mathbf{B}_{pZ}\\ \mathbf{B}_{gp}\mathbf{B}_{pZ}\end{bmatrix}Z_t\mathrm{d}t\\ \mathbf{p}_t &= \mathbf{B}_{pg}\mathbf{g}_t + \mathbf{B}_{pZ}Z_t,\end{aligned} \tag{22}$$

given the exogenous stochastic process for productivity (4). This system still maps the low-dimensional input Z_t into the low-dimensional output $\mathbf{p}_t$. However, the intermediating variables are now both the distribution $\mathbf{g}_t$ and the forward-looking decisions $\mathbf{v}_t$.

3.1.2 Deriving The Reduced System Given Basis X

Model reduction involves two related tasks: first, given a trial basis $\mathbf{X}$, we must compute the dynamics of the reduced system in terms of the distribution coefficients γ_t; and second, we must choose the basis $\mathbf{X}$ itself. In this subsection, we complete the first step of characterizing the reduced system given a basis $\mathbf{X}$, which is substantially easier than the second step of choosing the basis. Sections 3.2 and 3.3 discuss how we choose the basis.

Mathematically, we project the distribution $\mathbf{g}_t$ onto the subspace spanned by the basis $\mathbf{X} \in \mathbb{R}^{N\times k}$. Write the requirement that $\mathbf{g}_t \approx \mathbf{X}\gamma_t$ as

$$\mathbf{g}_t = \mathbf{X}\gamma_t + \varepsilon_t, \tag{23}$$

where $\varepsilon_t \in \mathbb{R}^N$ is a residual. The formulation (23) is a standard linear regression in which the distribution $\mathbf{g}_t$ is the dependent variable, the basis vectors $\mathbf{X}$ are the independent variables, and the coefficients γ_t are to be estimated.

Just as in ordinary least squares, we can estimate the projection coefficients γ_t by imposing the orthogonality condition $\mathbf{X}^{\mathrm{T}}\varepsilon_t = 0$, giving the familiar formula

$$\gamma_t = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{g}_t. \tag{24}$$

from individuals' decisions to the distribution $\mathbf{B}_{gv} = 0$. In that case the system (16) again becomes a backward-looking system of the LTI form (21), now with $\mathbf{C}_{gg} = \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg}$ and $\mathbf{C}_{gZ} = \mathbf{B}_{gp}\mathbf{B}_{pZ}$.

A sensible basis will be orthonormal, so that $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{I}$, further simplifying (24) to $\gamma_t = \mathbf{X}^{\mathrm{T}}\mathbf{g}_t$.[35] We can compute the evolution of this coefficient vector by differentiating (24) with respect to time and using (23) to get

$$\begin{aligned}\frac{\mathrm{d}\gamma_t}{\mathrm{d}t} = \mathbf{X}^{\mathrm{T}}\frac{\mathrm{d}g_t}{\mathrm{d}t} &= \mathbf{X}^{\mathrm{T}}\mathbf{B}_{gv}\mathbf{v}_t + \mathbf{X}^{\mathrm{T}}\left(\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg}\right)\left(\mathbf{X}\gamma_t + \varepsilon_t\right) + \mathbf{X}^{\mathrm{T}}\mathbf{B}_{gp}\mathbf{B}_{pg}Z_t \\ &\approx \mathbf{X}^{\mathrm{T}}\mathbf{B}_{gv}\mathbf{v}_t + \mathbf{X}^{\mathrm{T}}\left(\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg}\right)\mathbf{X}\gamma_t + \mathbf{X}^{\mathrm{T}}\mathbf{B}_{gp}\mathbf{B}_{pg}Z_t.\end{aligned}$$

The hope is that the residuals ε_t are small and so the last approximation is good. Assuming this is the case, we have the reduced version of (22)

$$\begin{aligned}\begin{bmatrix}\mathbb{E}_t[\mathrm{d}\mathbf{v}_t] \\ \mathrm{d}\gamma_t\end{bmatrix} &= \begin{bmatrix}\mathbf{B}_{vv} & \mathbf{B}_{vp}\mathbf{B}_{pg}\mathbf{X} \\ \mathbf{X}^{\mathrm{T}}\mathbf{B}_{gv} & \mathbf{X}^{\mathrm{T}}\left(\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg}\right)\mathbf{X}\end{bmatrix}\begin{bmatrix}\mathbf{v}_t \\ \gamma_t\end{bmatrix}\mathrm{d}t + \begin{bmatrix}\mathbf{B}_{vp}\mathbf{B}_{pZ} \\ \mathbf{X}^{\mathrm{T}}\mathbf{B}_{gp}\mathbf{B}_{pZ}\end{bmatrix}Z_t\mathrm{d}t, \\ \widetilde{\mathbf{p}}_t &= \mathbf{B}_{pg}\mathbf{X}\gamma_t + \mathbf{B}_{pZ}Z_t.\end{aligned} \tag{25}$$

Summing up, assuming we have the basis $\mathbf{X}$, this projection procedure takes us from the system of differential equations involving the N-dimensional vector $\mathbf{g}_t$ in (22) to a system involving only the k-dimensional vector γ_t in (25).[36]

3.2 Choosing the Basis X with Exogenous Decision Rules

We now turn to choosing a good basis $\mathbf{X}$. In this section we explain how to choose a basis in a model with exogenous decision rules, allowing us to use preexisting tools from the model reduction literature. In Section 3.3 we extend the strategy to the case with endogenous decision rules.

Mechanically increasing the size of the basis $\mathbf{X}$ will improve the approximation of the distribution $\mathbf{g}_t$; in the limit where $\mathbf{X}$ spans $\mathbb{R}^N$, we will not reduce the distribution at all. The goal of the model reduction literature is to provide a good approximation of the mapping from inputs Z_t to outputs $\mathbf{p}_t$ with as small a basis $\mathbf{X}$ as possible. We operationalize the notion of a "good approximation" by matching the impulse response function of $\mathbf{p}_t$ to a

[35] The assumption that $\mathbf{X}$ is orthonormal is not necessary to derive our results but makes the exposition transparent. Appendix A.3 derives our results using non-normalized projection matrices.

[36] The model reduction literature also presents alternatives to our "least squares" approach to computing the coefficients γ_t. In particular, one can also estimate γ_t using what amounts to an instrumental variables strategy: one can define a second subspace spanned by the columns of some matrix $\mathbf{Z}$ and impose the orthogonality condition $\mathbf{Z}^{\mathrm{T}}\varepsilon_t = 0$. This yields an alternative estimate $\gamma_t = (\mathbf{Z}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{g}_t$. Mathematically, this is called an *oblique projection* (as opposed to an orthogonal projection) of $\mathbf{g}_t$ onto the k-dimensional subspace spanned by the columns $\mathbf{X}$ along the kernel of $\mathbf{Z}^{\mathrm{T}}$. See Amsallem and Farhat (2011, Lecture 3) and Antoulas (2005) for more detail on oblique projections.

shock to Z_t up to a specified order k.[37]

3.2.1 Choosing the Basis in a Simplified Deterministic Model

To transparently motivate our choice of basis $\mathbf{X}$, we begin with a simplified version of the system (21). In particular, we make two simplifying assumptions. First, we assume that there are no aggregate shocks, so that $Z_t = 0$ for all t. This allows us to focus on deterministic transition paths starting from an exogenously given initial distribution $\mathbf{g}_0$; because certainty equivalence with respect to aggregate shocks holds in our linear setting, these transition paths are intimately related to impulses responses driven by shocks to Z_t. Our second simplifying assumption is that $\mathbf{p}_t = p_t$ is a scalar. This emphasizes that the price vector we are trying to approximate is a low-dimensional object. Under these assumptions, we obtain the following simplified version of the system (21)

$$\begin{aligned} \frac{\mathrm{d}\mathbf{g}_t}{\mathrm{d}t} &= \mathbf{C}_{gg}\mathbf{g}_t \\ p_t &= \mathbf{b}_{pg}\mathbf{g}_t, \end{aligned} \tag{26}$$

where $\mathbf{b}_{pg}$ is a $1 \times N$ vector. The reduced version of this system is

$$\begin{aligned} \frac{\mathrm{d}\gamma_t}{\mathrm{d}t} &= \mathbf{X}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{X}\gamma_t \\ \widetilde{p}_t &= \mathbf{b}_{pg}\mathbf{X}\gamma_t, \end{aligned} \tag{27}$$

where $\widetilde{p}_t$ denotes the reduced path of prices. Since the system (26) is linear, it has a simple solution. The solution of the first equation is $\mathbf{g}_t = e^{\mathbf{C}_{gg}t}\mathbf{g}_0$ where $e^{\mathbf{C}_{gg}t}$ is a matrix exponential. Hence

$$p_t = \mathbf{b}_{pg}e^{\mathbf{C}_{gg}t}\mathbf{g}_0. \tag{28}$$

Similarly, we can derive an analogous solution for the reduced prices which satisfy (27)

$$\widetilde{p}_t = \mathbf{b}_{pg}\mathbf{X}e^{\mathbf{X}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{X}t}\gamma_0 \tag{29}$$

The goal is then to choose $\mathbf{X}$ such that p_t in (28) is "close" to $\widetilde{p}_t$ in (29). The key idea is to choose $\mathbf{X}$ such that the kth-order Taylor series approximation of p_t in (28) around $t = 0$ *exactly matches* that of $\widetilde{p}_t$ in (29).

[37] Our approach for choosing the basis $\mathbf{X}$ is a simplified version of what the model reduction literature calls "moment matching." See Amsallem and Farhat (2011, Lecture 7) and Antoulas (2005, Chapter 11). It is also the continuous-time analogue of what Reiter (2010) terms "conditional expectation approach" (see his Section 3.2.2).

The Taylor-series approximation of the time path of prices (28) around $t = 0$ is[38]

$$p_t \approx \mathbf{b}_{pg}\left[\mathbf{I} + \mathbf{C}_{gg}t + \frac{1}{2}\mathbf{C}_{gg}^2 t^2 + \ldots + \tfrac{1}{(k-1)!}\mathbf{C}_{gg}^{k-1}t^{k-1}\right]\mathbf{g}_0 \tag{30}$$

where we have used that $e^{\mathbf{C}_{gg}t} \approx \mathbf{I} + \mathbf{C}_{gg}t + \frac{1}{2}\mathbf{C}_{gg}^2 t^2 + ...$ Similarly, the Taylor-series approximation of reduced prices is

$$\widetilde{p}_t \approx \mathbf{b}_{pg}\mathbf{X}\left[\mathbf{I} + (\mathbf{X}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{X})t + \frac{1}{2}(\mathbf{X}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{X})^2 t + \ldots + \tfrac{1}{(k-1)!}(\mathbf{X}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{X})^{k-1}t^{k-1}\right]\gamma_0. \tag{31}$$

We want to choose $\mathbf{X}$ so that the first k terms of the two Taylor series expansions are identical. With $\gamma_0 = \mathbf{X}^{\mathrm{T}}\mathbf{g}_0$, this means that we require $\mathbf{b}_{pg} = \mathbf{b}_{pg}\mathbf{X}\mathbf{X}^{\mathrm{T}}$, $\mathbf{b}_{pg}\mathbf{C}_{gg} = \mathbf{b}_{pg}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{X}\mathbf{X}^{\mathrm{T}}$, and so on. If γ_t has the same dimensionality as $\mathbf{g}_t$ ($k = N$, i.e., we are not reducing the distribution at all), then $\mathbf{X}$ has to be orthogonal, i.e. $\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{I}$, and the conclusion trivially follows. But once we have proper reduction, this equality does not hold, and the problem of Taylor series coefficient matching becomes non-trivial. Fortunately, the model reduction literature gives us a systematic way for choosing $\mathbf{X}$ such that (30) matches (31). This systematic way builds upon what is known as the order-k *observability matrix* of the system (26):[39]

$$\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg}) := \begin{bmatrix} \mathbf{b}_{pg} \\ \mathbf{b}_{pg}\mathbf{C}_{gg} \\ \mathbf{b}_{pg}\mathbf{C}_{gg}^2 \\ \vdots \\ \mathbf{b}_{pg}\mathbf{C}_{gg}^{k-1} \end{bmatrix}. \tag{32}$$

It turns out that if the basis $\mathbf{X}$ spans the subspace generated by the transpose of the observability matrix $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$, then the kth-order Taylor-series approximation of reduced prices (31) exactly matches that of unreduced prices (30), even though it only uses information on the reduced state vector γ_t. Showing this just requires a few lines of algebra, which we present in Appendix A.3.1.

[38] In this simple deterministic model, (30) can also be derived in a simpler fashion: the Taylor-series approximation around $t = 0$ is $p_t \approx p_0 + \dot{p}_0 t + \frac{1}{2}\ddot{p}_0 t^2 + ... + \frac{1}{(k-1)!}p_0^{(k-1)}t^{k-1}$. This is equivalent to (30) because the derivatives are given by $\dot{p}_t = \mathbf{b}_{pg}\dot{\mathbf{g}}_t = \mathbf{b}_{pg}\mathbf{C}_{gg}\mathbf{g}_t$, $\ddot{p}_t = \mathbf{b}_{pg}\mathbf{C}_{gg}^2\mathbf{g}_t$ and so on. This strategy no longer works in the full model with aggregate productivity shocks. In contrast, the derivation in terms of the matrix exponential $e^{\mathbf{C}_{gg}t}$ can be easily extended to the stochastic case.

[39] Observability of a dynamical system is an important concept in control theory introduced by Rudolf Kalman, the inventor of the Kalman filter. It is a measure of how well a system's states (here $\mathbf{g}_t$) can be inferred from knowledge of its outputs (here p_t). For systems like ours observability can be directly inferred from the observability matrix $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$ with $k = N$. Note that some texts refer only to $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$ with $k = N$ as "observability matrix" and to the matrix with $k < N$ as "partial observability matrix."

To gain some intuition why the observability matrix (32) makes an appearance, note that the Taylor-series approximation (30) can be written more compactly using matrix notation as

$$p_t \approx \left[1, t, \tfrac{1}{2}t^2, ..., \tfrac{1}{(k-1)!}t^{k-1}\right] \mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})\mathbf{g}_0$$

Related, $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})\mathbf{g}_t$ is simply the vector of time derivatives of p_t, i.e. $\dot{p}_t, \ddot{p}_t$ and so on (see footnote 38).

3.2.2 Choosing the Basis in The Stochastic Model

The deterministic case makes clear that the observability matrix $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$ plays a key role in model reduction. The logic of this simple case carries through the stochastic model, but the full derivation is more involved and details can be found in Appendix A.3. Because the model is now stochastic, the correct notion of "matching the path of prices" is to match the impulse response function of prices.

***Proposition* 1.** *Consider the stochastic model with exogenous decision rules (21). Let* $\mathbf{X}$ *be a basis which spans the subspace generated by the observability matrix* $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})^{\mathrm{T}}$ *with* $\mathbf{C}_{gg} = \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{b}_{pg} + \mathbf{B}_{gv}\mathbf{D}_{vg}$. *Then the impulse response function of prices* $\widetilde{p}_t$ *to an aggregate productivity shock* Z_t *in the reduced model equals the impulse response function of prices* p_t *in the unreduced model up to order* k.

Proof. See Appendix A.3.2 □

The impulse response function in the stochastic model combines the impact effect of an aggregate shock Z_t together with the transition back to steady state. We do not reduce the exogenous state variable Z_t, so the reduced model captures the impact effect of a shock exactly. The role of the observability matrix is to approximate the transition back to steady state analogously to the deterministic case.

Finally, note that in this section we have assumed p_t is a scalar to emphasize that it is a low-dimensional object. In general $\mathbf{p}_t$ is an $\ell \times 1$ vector. One can extend the argument above to show that the correct basis $\mathbf{X}$ spans the subspace generated by $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{C}_{gg})^{\mathrm{T}}$ for $\mathbf{C}_{gg} = \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg} + \mathbf{B}_{gv}\mathbf{D}_{vg}$ where now $\mathbf{B}_{pg}$ is an $\ell \times N$ matrix. Matching impulse response functions of $\mathbf{p}_t$ up to order k, requires matching ℓk terms in the corresponding Taylor-series approximation and hence the observability matrix $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{C}_{gg})$ is now of dimension $k_g \times N$, where $k_g = \ell k < N$.

3.3 Choosing the Basis X with Endogenous Decision Rules

Section 3.2 shows that if decision rules $\mathbf{D}_{vg}$ are exogenously given, then choosing the basis $\mathbf{X}$ to span the subspace generated by $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{C}_{gg})^{\mathrm{T}}$ guarantees that the impulse response of the reduced price $\widetilde{p}_t$ matches the unreduced model up to a pre-specified order k. However, when decision rules are endogenous, the choice of basis impacts agents' decisions and therefore the evolution of the distribution. In this case, the results of Section 3.2 do not apply.

However, the choice of basis in Section 3.2 was only dictated by the concern of *efficiently* approximating the distribution with as small a basis as possible; it is always possible to improve *accuracy* by adding additional orthogonal basis vectors. In fact, in the finite limit when $k = N$, any linearly independent basis spans all of $\mathbb{R}^N$ so the distribution is not reduced at all and the reduced model is vacuously accurate. Therefore, setting the basis $\mathbf{X}$ to the subspace generated by $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg}+\mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$, i.e. ignoring feedback from individuals' decisions to the distribution by effectively setting $\mathbf{D}_{vg} = 0$, will not be efficient but may still be accurate. In practice, we have found in both the simple Krusell and Smith (1998) model and the two-asset model in Section 5 that this choice leads to accurate solutions for high enough order k of the observability matrix.

In cases where choosing the basis to span the subspace $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$ is not accurate even for as high an order k as numerically feasible, we suggest an iterative procedure. First, we solve the reduced model (25) based on the inaccurate basis choice for the subspace $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$. This yields decision rules $\mathbf{D}_{v\gamma}$ defining a mapping from the reduced distribution γ_t to the value function. We then use these to construct an approximation to the true decision rules $\mathbf{D}_{vg}$ (which map the full distribution to the value function), i.e. $\widetilde{\mathbf{D}}_{vg} = \mathbf{D}_{v\gamma}\mathbf{X}^{\mathrm{T}}$.[40] Next we choose a new basis of the subspace generated by $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg} + \mathbf{B}_{gv}\widetilde{\mathbf{D}}_{vg})^{\mathrm{T}}$ and solve the model again based on the new reduction. If the second reduction gives an accurate solution, we are done; if not, we continue the iteration. Although we have no theoretical guarantee that this iteration will converge, in practice we have found that it does.

Choosing k and Internal Consistency with Endogenous Decision Rules A key practical step in reducing the distribution is choosing the order of the observability matrix k, which determines the size of the basis $\mathbf{X}$. With exogenous decision rules, we showed that a basis of order k implies that the path of reduced prices $\widetilde{p}_t$ matches the k-th order Taylor

[40] Recall from (24) that the projection of $\mathbf{g}_t$ onto $\mathbf{X}$ defines the reduced distribution as $\gamma_t = \mathbf{X}^{\mathrm{T}}\mathbf{g}_t$. Hence the optimal decision rule can be written as $\widetilde{\mathbf{v}}_t = \mathbf{D}_{v\gamma}\gamma_t = \mathbf{D}_{v\gamma}\mathbf{X}^{\mathrm{T}}\mathbf{g}_t = \widetilde{\mathbf{D}}_{vg}\mathbf{g}_t$ where $\widetilde{\mathbf{D}}_{vg} = \mathbf{D}_{v\gamma}\mathbf{X}^{\mathrm{T}}$.

expansion of the path of true prices p_t, providing a natural metric for assessing accuracy.[41] However, this logic does not carry through with endogenous decision rules, leaving unclear what exactly a basis of order k captures.

In the finite limit when $k = N$, any linearly independent basis spans all of $\mathbb{R}^N$ so the distribution is not reduced at all and the reduced model is vacuously accurate. Hence, a natural procedure is to increase k until the dynamics of reduced prices converge. In practice, this convergence is often monotonic. However, we cannot prove convergence is always monotonic, still leaving open the question of what exactly the reduced model captures for a given order k.

We suggest an *internal consistency* metric to assess the extent to which the reduced model satisfies the model's equilibrium conditions. The spirit of our internal consistency check is similar to Krusell and Smith (1998)'s R^2 forecast-error metric and Den Haan (2010)'s accuracy measure discussed in Section 2: if agents make decisions based on the price path implied by the reduced distribution, but we aggregate those decisions against the true full distribution, do the prices generated by the true distribution match the forecasts?

Concretely, our internal consistency check consists of three steps. First, we compute households' decisions based on the reduced distribution, $\widetilde{\mathbf{v}}_t = \mathbf{D}_{v\gamma}\gamma_t$. Second, we use these decisions to simulate the nonlinear dynamics of the full distribution $\mathbf{g}_t^*$ – not the reduced version γ_t – and its implied prices p_t^* for a given path of aggregate shocks Z_t

$$\begin{aligned} \mathbf{p}_t^* &= \mathbf{B}_{pg}\mathbf{g}_t^* + \mathbf{B}_{pZ}Z_t \\ \frac{\mathrm{d}\mathbf{g}_t^*}{\mathrm{d}t} &= \mathbf{A}(\widetilde{\mathbf{v}}_t, \mathbf{p}_t^*)\mathbf{g}_t^*, \end{aligned}$$

where $\mathbf{A}(\widetilde{\mathbf{v}}_t, p_t^*)$ is the nonlinear transition matrix implied by the decision rules $\widetilde{\mathbf{v}}_t$ and price p_t^*. The third step of our internal accuracy check is to assess the extent to which the dynamics of p_t^* matches the dynamics implied by the reduced system $\widetilde{p}_t$. If the two paths are close, households in the reduced model could not significantly improve their forecasts by using additional information about the distribution. Once again, we compare the maximum log deviation of the two paths

$$\epsilon = \max_i \max_{t \geq 0} \left| \log \widetilde{p}_{it} - \log p_{it}^* \right|,$$

where i denotes an entry in the price vector.

[41] Recall that in general p_t includes both prices and other observables of interest to the researcher.

Computing The Basis X Following the discussion above, we choose the basis $\mathbf{X}$ to span the subspace generated by $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$. However, using $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$ directly is numerically unstable due to approximate multicollinearity; as in standard regression, high degree standard polynomials are nearly collinear due to the fact that, for large k, $\mathbf{B}_{pg}(\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{k-2} \approx \mathbf{B}_{pg}(\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{k-1}$, leaving the necessary projection of the distribution onto $\mathbf{X}$ numerically intractable.

We overcome this challenge by relying on a *Krylov subspace method*, an equivalent but more numerically stable class of methods.[42] For any $N \times N$ matrix $\mathbf{A}$ and $N \times 1$ vector $\mathbf{b}$, the order-k Krylov subspace is

$$\mathcal{K}_k(\mathbf{A}, \mathbf{b}) = \text{span}\left(\left\{\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, ..., \mathbf{A}^{k-1}\mathbf{b}\right\}\right).$$

From this definition it can be seen that the subspace spanned by the columns of $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$ is simply the order-k Krylov subspace generated by $(\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$ and $\mathbf{B}_{pg}^{\mathrm{T}}$, i.e. $\mathcal{K}_k(\mathbf{B}_{gg}^{\mathrm{T}} + \mathbf{B}_{gp}^{\mathrm{T}}\mathbf{B}_{pg}^{\mathrm{T}}, \mathbf{B}_{pg}^{\mathrm{T}})$. Therefore, the projection of $\mathbf{g}_t$ on $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$ is equivalent to the projection of $\mathbf{g}_t$ onto this Krylov subspace.

There are many methods for projecting onto Krylov subspaces in the literature. One important feature of all these methods is that they take advantage of the sparsity of the underlying matrices.[43] We have found that one particular method, *deflated block Arnoldi iteration*, is a robust procedure. Deflated block Arnoldi iteration has two advantages for our application. First, it is a stable procedure to orthogonalize the columns of the basis $\mathbf{X}$ and eliminate the approximate multicollinearity. Second, the *deflation* component handles multicollinearity that can arise even with non-deflated block Arnoldi iteration.

3.4 Value Function Reduction

After reducing the dimensionality of the distribution $\mathbf{g}_t$, we are left with a system of dimension $N + k_g$ with $k_g << N$ (recall $k_g = \ell \times k$ where ℓ is the number of prices and k is the order of the approximation according to which the basis $\mathbf{X}$ is chosen). Although this is considerably smaller than the original system which was of size $2N$, it is still large because it contains N equations for the value function – one for each point in the individual state

[42] See Antoulas (2005, Chapter 11) and Amsallem and Farhat (2011, Lecture 7).

[43] Even though $\mathbf{B}_{gg}$ is sparse and $\mathbf{B}_{gp}$ and $\mathbf{B}_{pg}$ are only $\ell \times N$, the matrix $\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg}$ which actually enters the system (22) is $N \times N$ and not sparse (because $\mathbf{B}_{gp}\mathbf{B}_{pg}$ is $N \times N$ not sparse). In the two-asset model in Section 5, $N = 66,000$, and even storing this matrix is not feasible. Fortunately it is never actually necessary to compute this full matrix; instead, it is only necessary to compute $\mathbf{B}_{pg}(\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})$ which involves the action of $\mathbf{B}_{gp}\mathbf{B}_{pg}$ on a thin $\ell \times N$ matrix $\mathbf{B}_{pg}$ and can be computed as $(\mathbf{B}_{pg}\mathbf{B}_{gp})\mathbf{B}_{pg}$.

space. In complex models, this leaves the linear system too large for matrix decomposition methods to be feasible.[44]

We therefore also reduce the dimensionality of the distribution $\mathbf{v}_t$. Just like in our method for reducing the distribution $\mathbf{g}_t$, we project the (deviation from steady state of the) value function $\mathbf{v}_t$ onto a lower-dimensional subspace. As before, an important question is how to choose the basis for this projection. We choose it by appealing to the theory for approximating smooth functions and approximate $\mathbf{v}_t$ using splines. In most models, the value function is sufficiently smooth that a low-dimensional spline provides an accurate approximation. In particular, any spline approximation can be written as the projection

$$\mathbf{v}_t \approx \mathbf{X}_v \nu_t,$$

where $\mathbf{X}_v$ is a $N \times k_v$ matrix defining the spline knot points and ν_t are the k_v coefficients at those knot points.[45] Given this linear projection the coefficients are given by $\nu_t = (\mathbf{X}_v^{\mathrm{T}}\mathbf{X}_v)^{-1}\mathbf{X}_v^{\mathrm{T}}\mathbf{v}_t = \mathbf{X}_v^{\mathrm{T}}\mathbf{v}_t$, where we have used that we typically choose an orthonormal $\mathbf{X}_v$ so that $\mathbf{X}_v^{\mathrm{T}}\mathbf{X}_v = \mathbf{I}$.

It is worth emphasizing the symmetry with our distribution reduction method, the projection (23). In order to do so we add a g-subscript to the basis in the distribution reduction for the remainder of the paper and write (23) as $\mathbf{g}_t \approx \mathbf{X}_g \gamma_t$. Hence from now on $\mathbf{X}_g$ denotes the basis in the reduction of the distribution $\mathbf{g}_t$ and $\mathbf{X}_v$ denotes the basis in the reduction of the value function $\mathbf{v}_t$. It is also important to note that we are approximating the deviation of the value function from its steady state value, not the value function itself (the reader should recall our convention in the present section to drop hat subscripts from variables that are in deviation from steady state for notational simplicity).

We have found that non-uniformly spaced quadratic splines work well for three reasons. First, the non-uniform spacing can be used to place more knots in regions of the state space with high curvature, allowing for an efficient dimensionality reduction. Second, the quadratic spline preserves monotonicity and concavity between knot points, which is important in computing first-order conditions. Third, and related, the local nature of splines implies that they avoid creating spurious oscillations at the edges of the state space (Runge's phenomenon) which often occurs with global approximations like high-degree polynomials.

[44] One way to overcome this challenge is to use sparse matrix methods to find just the k eigenvalues associated with the stable eigenvectors. This is much faster than computing the full matrix decomposition necessary to obtain the full set of eigenvectors. However, it is slower than the approach we pursue in this subsection.

[45] Note that, in general, the number of coefficients is different from the number of knot points.

It is also important to note the difference between approximating the deviations of the value function from steady state using quadratic splines – which we do – versus solving for the steady state value using quadratic splines – which we do not do. The finite difference method we use to compute the steady state does not impose that the value function is everywhere differentiable, which is potentially important for capturing the effects of non-convexities. However, after having computed the steady state value functions, it is typically the case that they have kinks at a finite number of points and are well-approximated by smooth functions between these points. It is then straightforward to fit quadratic splines between the points of non-differentiability.

3.5 Putting It All Together: A Numerical Toolbox

Summarizing the previous sections, we have projected the distribution $\mathbf{g}_t$ onto the subspace spanned by $\mathbf{X}_g$ and the value function $\mathbf{v}_t$ onto the subspace spanned by $\mathbf{X}_v$. Now we simply need to keep track of the $k_v \times 1$ coefficient vector ν_t for the value function and the $k_g \times 1$ coefficient vector γ_t for the distribution. Because knowledge of these coefficients is sufficient to reconstruct the full value function and distribution, we will also sometimes refer to ν_t as the reduced value function and to γ_t as the reduced distribution. Our original system (16) is now reduced to

$$\mathbb{E}_t \begin{bmatrix} \mathrm{d}\nu_t \\ \mathrm{d}\gamma_t \\ \mathrm{d}Z_t \end{bmatrix} = \begin{bmatrix} \mathbf{X}_v^{\mathrm{T}}\mathbf{B}_{vv}\mathbf{X}_v & \mathbf{X}_v^{\mathrm{T}}\mathbf{B}_{vp}\mathbf{B}_{pg}\mathbf{X}_g & \mathbf{X}_v^{\mathrm{T}}\mathbf{B}_{vp}\mathbf{B}_{pZ} \\ \mathbf{X}_g^{\mathrm{T}}\mathbf{B}_{gv}\mathbf{X}_v & \mathbf{X}_g^{\mathrm{T}}(\mathbf{B}_{gg}+\mathbf{B}_{gp}\mathbf{B}_{pg})\mathbf{X}_g & \mathbf{X}_g^{\mathrm{T}}\mathbf{B}_{gp}\mathbf{B}_{pZ} \\ \mathbf{0} & \mathbf{0} & -\eta \end{bmatrix} \begin{bmatrix} \nu_t \\ \gamma_t \\ Z_t \end{bmatrix} \mathrm{d}t. \tag{33}$$

We have provided a numerical toolbox implementing the key steps in our computational method at the `github` page associated with this project.[46] Broadly, the user provides two files: one which solves for the steady state and another which evaluates the model's equilibrium conditions. Our toolbox then implements the following algorithm (we here revert back to denoting deviations from steady state with hat superscripts):

1. Compute the steady state values of $\mathbf{v}, \mathbf{g}$ and $\mathbf{p}$.

2. Compute a first-order Taylor expansion of the equilibrium conditions (14) around steady state using automatic differentiation, yielding the system (16) in terms of deviations from steady state $\widehat{\mathbf{v}}_t, \widehat{\mathbf{g}}_t, \widehat{\mathbf{p}}_t$ and Z_t.

[46] Currently at: https://github.com/gregkaplan/phact.

3. If necessary, reduce the model, yielding the system (33) in terms of (ν_t, γ_t, Z_t).

   (a) Distribution reduction: compute the basis $\mathbf{X}_g = \mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$ using deflated Arnoldi iteration and project $\widehat{\mathbf{g}}_t$ on $\mathbf{X}_g$ to obtain the reduced distribution γ_t.

   (b) Value function reduction: compute the spline basis $\mathbf{X}_v$ and project $\widehat{\mathbf{v}}_t$ on $\mathbf{X}_v$ to obtain the reduced value function ν_t.

4. Solve the system (16) or, if reduced, (33).

5. Simulate the system to compute impulse responses and time-series statistics.

3.6 Model Reduction in Krusell-Smith Model

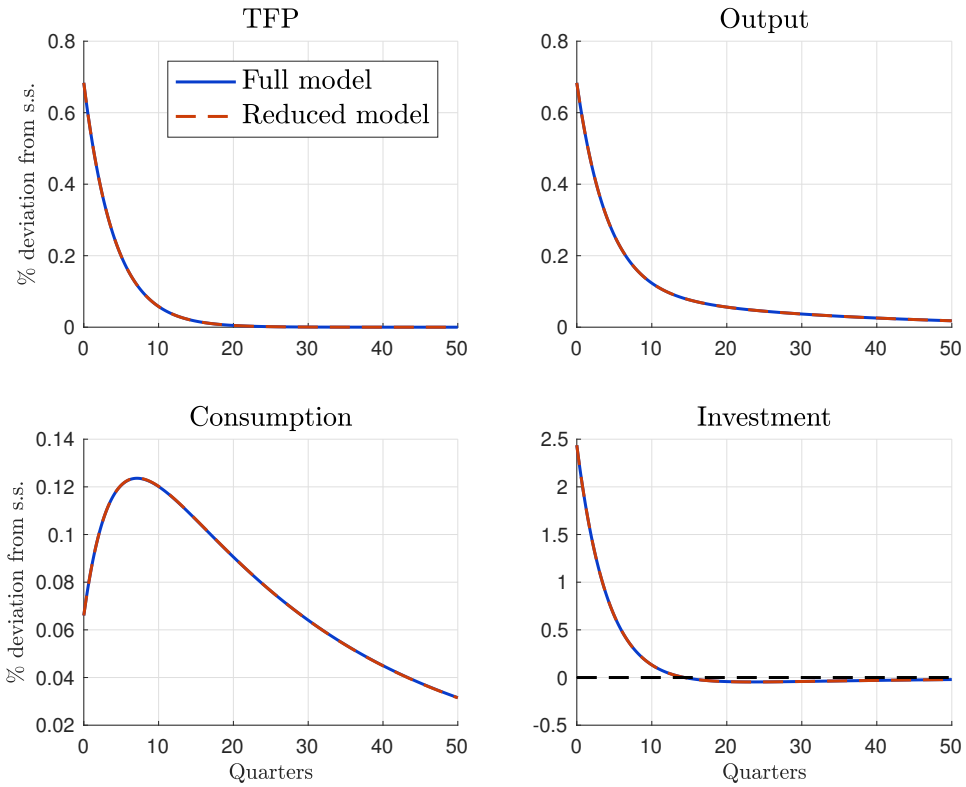
The Krusell and Smith (1998) is a useful environment for evaluating our model reduction methodology because it is possible to solve the full unreduced model as a benchmark. We are able to substantially reduce the size of the system: projecting the distribution on an observability matrix of order $k = 1$ and approximating the value function at 24 spline knot points provides an extremely accurate approximation of the model's dynamics.[47] Figure 1 shows that the impulse responses of key aggregate variables in the reduced model are almost exactly identical to the full, unreduced model, despite approximating the $N = 400$ dimensional dynamic system with a 30-dimensional system.[48]

The fact that we can reduce the distribution with an observability matrix of order $k = 1$ is consistent with Krusell and Smith (1998)'s finding of "approximate aggregation" using a computationally distinct procedure and accuracy measure. In fact, as Figure 2 shows, a $k = 1$ order approximation of the distribution returns precisely the mean. The top left panel of the figure plots the basis vector associated with $k = 1$, split into two 100-dimensional vectors corresponding to the two values for idiosyncratic productivity. It shows that indeed the first basis vector $\mathbf{x}_{g,1} = \left[\begin{smallmatrix}\mathbf{a}\\ \mathbf{a}\end{smallmatrix}\right]$, implying that $\gamma_t = \mathbf{x}_{g,1}^{\mathrm{T}}\mathbf{g}_t = \left[\begin{smallmatrix}\mathbf{a}\\ \mathbf{a}\end{smallmatrix}\right]^{\mathrm{T}}\mathbf{g}_t = \widehat{K}_t$, the (deviation from steady state of the) mean of the distribution. The remaining panels plot the higher-order elements of $\mathbf{X}_g$, which quickly converge to constants that do not add information to

[47] More precisely, we choose the observability matrix so as to forecast $\ell = 5$ equilibrium objects (namely the wage and the interest rate, plus the three equilibrium aggregates we are most interested in: aggregate output, consumption, investment) to order $k = 1$ resulting in a reduced distribution γ_t of dimension $k_g = \ell \times k = 5$, and we approximate the value function at 12 spline knot points in the wealth dimension resulting in a reduced value function ν_t of dimension $k_v = 2 \times 12 = 24$.

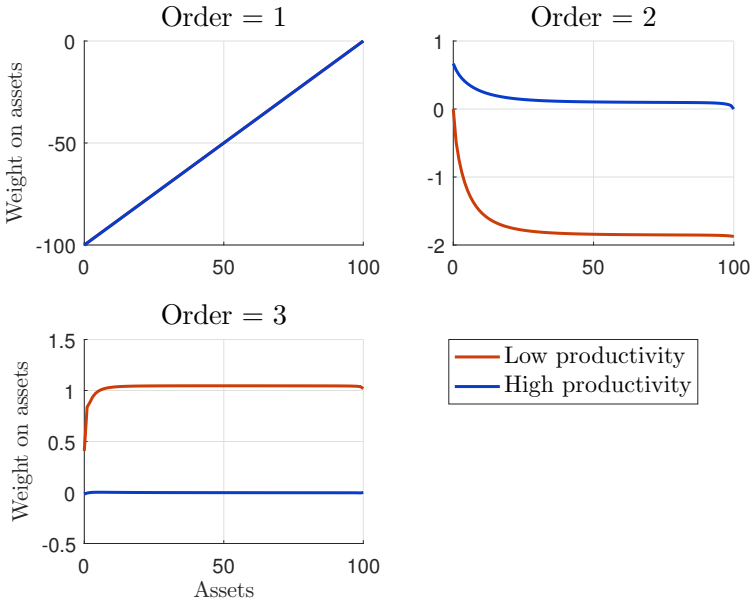
[48] There are $k_v = 12 \times 2 = 24$ points for the value function, $k_g = k \times \ell = 1 \times 5 = 5$ points for the distribution because we are tracking five elements of the $\mathbf{p}_t$ vector, and 1 point for TFP Z_t.

Figure 1: Impulse Responses to TFP Shock in Krusell-Smith Model



Notes: impulse responses to an instantaneous positive unit standard deviation size shock (Dirac delta function) to aggregate TFP. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$. "Full model" refers to model solved without model reduction and "reduced model" with reduction, using $k_g = 2$ (forecasting $\ell = 5$ objects, of which two are linearly independent, with a $k = 1$-order Taylor series approximation) and $k_v = 24$.

Figure 2: Basis Vectors in Distribution Reduction



Notes: The columns of $\mathbf{X}_g = \mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$, here displayed for the capital stock, up to order $k = 4$. These correspond to the basis vectors in the approximated distribution $\mathbf{g}_t \approx \gamma_{1t}\mathbf{x}_{g,1} + ... + \gamma_{4t}\mathbf{x}_{g,4}$.

the approximation. Hence, our model-free reduction method confirms Krusell and Smith (1998)'s approximate aggregation result in this simple model.

With or without dimensionality reduction, our method solves and simulates the model in less than 0.3 seconds. Table 3 reports the running time of using our `Matlab` code suite on a desktop PC. Although reduction is not necessary to solve this simple model, it nevertheless reduces running time by more than 50% and takes approximately 0.1 seconds.[49] In the two-asset model in Section 5, model reduction is necessary to even solve the model.

Our internal consistency check confirms the fact that the distribution reduction is accurate; the maximum log deviation is 0.065%, which is twice as small as the most accurate algorithm in the JEDC comparison Den Haan (2010). Recall that in the unreduced model that the maximum log deviation is 0.049%, capturing the error due to linearization. Hence, the additional error due to our model reduction is extremely small. Figure 3 plots the two series for a random 400-quarter period of simulation and shows that the two series are extremely close to each other.[50]

[49] Recall that the fastest algorithm in the JEDC comparison Den Haan (2010) is more 7 minutes, or 3500 times longer.

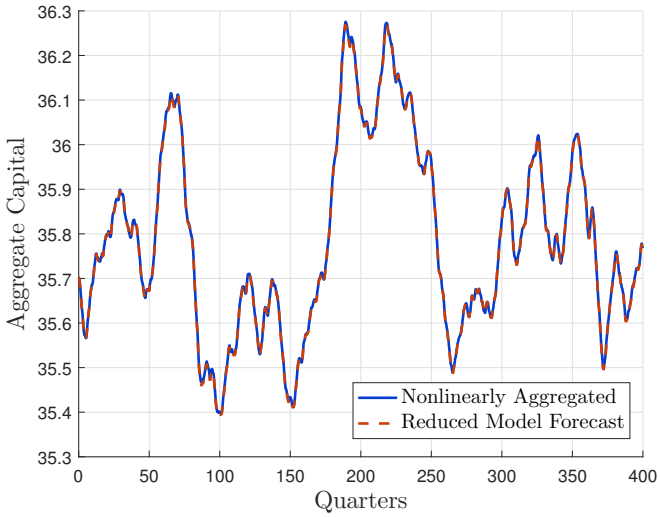
[50] Den Haan (2010) refers to this type of figure as the "fundamental accuracy plot."

Table 3: Run Time for Solving Krusell-Smith Model

	Full Model	**Reduced Model**
Steady State	0.082 sec	0.082 sec
Derivatives	0.021 sec	0.021 sec
Dim reduction	×	0.007 sec
Linear system	0.14 sec	0.002 sec
Simulate IRF	0.024 sec	0.003 sec
Total	**0.267 sec**	**0.116 sec**

Notes: Time to solve Krusell-Smith model once on `MacBook Pro 2016` laptop with `3.3 GHz` processor and `16 GB RAM`, using `Matlab R2016b` and our code toolbox. "Full model" refers to solving model without model reduction and "reduced model" with reduction, using $k_g = 1$ and $k_v = 12$. "Steady state" reports time to compute steady state. "Derivatives" reports time to compute derivatives of discretized equilibrium conditions. "Dim reduction" reports time to compute both the distribution and value function reduction. "Linear system" reports time to solve system of linear differential equations. "Simulate IRF" reports time to simulate impulse responses reported in Figure 1. "Total" is the sum of all these tasks.

Figure 3: Internal Consistency Check



Notes: Two series for aggregate capital that enter the internal consistency check ϵ. "Reduced model forecast" computes the path $\widetilde{K}_t$ implied by the reduced linear model. "Nonlinear model forecast" computes the path K_t^* from updating the distribution according to the nonlinear KFE (3).

4 Two-Asset Incomplete Markets Model

While the Krusell and Smith (1998) model is a useful pedagogical tool for explaining our computational method, it does not reproduce key features of the distribution of household-level income, wealth, and consumption in the micro data. In this section, we apply our method to solve a two-asset incomplete markets model in the spirit of Kaplan and Violante (2014) and Kaplan, Moll and Violante (2016), which is explicitly parameterized to match key features of these distributions. Accurately reproducing these features leads to a failure of approximate aggregation which, together with the model's size, render it an ideal setting to illustrate the power of our method. In Sections 5 and 6, we use the model to illustrate a rich interaction between inequality and macroeconomic dynamics.

4.1 Model

The household side of the model is a simplified version of Kaplan, Moll and Violante (2016), so we refer to the interested reader to that paper for full details. The firm side follows the standard real business cycle model with aggregate productivity shocks.

4.1.1 Environment

Households There is a unit mass of households indexed by $j \in [0,1]$. At each instant of time, households hold liquid assets b_{jt}, illiquid assets a_{jt}, and have labor productivity z_{jt}. Households die with an exogenous Poisson intensity ζ and upon death give birth to an offspring with zero wealth $a_{jt} = b_{jt} = 0$ and labor productivity drawn from its ergodic distribution. There are perfect annuity markets, implying that the wealth of deceased households is distributed to other households in proportion to their asset holdings.[51] Each household has preferences over consumption c_{jt} represented by the expected utility function

$$\mathbb{E}\int_0^\infty e^{-(\rho+\zeta)t}\log c_{jt}\mathrm{d}t.$$

A household with labor productivity z_{jt} earns labor income $w_t z_{jt}$ and pays a linear income tax at rate τ. Each household also receives a constant lump-sum transfer from the government, T. Labor productivity follows a discrete state Poisson process, taking values

[51] We implement perfect annuity markets by making an adjustment to the asset returns faced by households. In order to save on notation, we do not explicitly display these adjustments here, so throughout asset returns should be interpreted as inclusive of annuity payments.

from the set $z_{jt} \in \{z_1, ..., z_J\}$. Households switch from state z to state z' with Poisson intensity $\lambda_{zz'}$.

The liquid asset b_{jt} pays a rate of return r_t^b. Households can borrow in liquid assets up to an exogenous limit $\underline{b}$. The interest rate on borrowing is $r_t^{b-} = r_t^b + \kappa$ where $\kappa > 0$ is a wedge between borrowing and lending rates. Define $r_t^b(b_t)$ to be the interest rate function which takes both of these cases into account.

The illiquid asset a_{jt} pays a rate of return r_t^a. It is illiquid in the sense that households must pay a flow cost $\chi(d_{jt}, a_{jt})$ to transfer assets at rate d_{jt} from the illiquid to liquid account. The transaction cost function is given by[52]

$$\chi(d, a) = \chi_0|d| + \chi_1 \left|\frac{d}{a}\right|^{\chi_2} a.$$

The linear component $\chi_0 > 0$ generates inaction in households' optimal deposit decisions. The convex component ($\chi_1 > 0, \chi_2 > 1$) ensures that deposit rates d/a are finite, so that households' asset holdings never jump. Scaling the convex term by illiquid assets a ensures that marginal transaction costs $\chi_d(d, a)$ are homogenous of degree zero in the deposit rate d/a, which implies that the marginal cost depends on the fraction of illiquid assets transacted rather than the raw size of the transaction.

The laws of motion for liquid and illiquid assets are

$$\begin{aligned}
\frac{\mathrm{d}b_{jt}}{\mathrm{d}t} &= (1-\tau)w_t z_{jt} + T + r_t^b(b_{jt})b_{jt} - \chi(d_{jt}, a_{jt}) - c_{jt} - d_{jt} \\
\frac{\mathrm{d}a_{jt}}{\mathrm{d}t} &= r_t^a a_{jt} + d_{jt}.
\end{aligned}$$

Firms There is a representative firm with the Cobb-Douglas production function

$$Y_t = e^{Z_t} K_t^{\alpha} \overline{L}^{1-\alpha},$$

where as before Z_t is the logarithm of aggregate productivity, K_t is aggregate capital, and $\overline{L}$ is aggregate labor supply which is constant by assumption. The logarithm of aggregate productivity again follows the Ornstein-Uhlenbeck process

$$\mathrm{d}Z_t = -\eta Z_t \mathrm{d}t + \sigma \mathrm{d}W_t,$$

[52]Because the transaction cost at a = 0 is infinite, in computations we replace the term a with $\max\{a, \underline{a}\}$, where the threshold $\underline{a} > 0$ is a small value (2% of quarterly GDP per household, which is around \$500). This guarantees that costs remain finite even for households with $a = 0$.

where $\mathrm{d}W_t$ is the innovation to a standard Brownian motion, η is the rate of mean reversion, and σ captures the size of innovations.

Government There is a government which balances its budget each period. Since the labor tax rate τ and lump-sum transfer rate T are constant, we assume that government spending G_t adjusts each period to satisfy the government budget constraint

$$\int_0^1 \tau w_t z_{jt} \mathrm{d}j = G_t + \int_0^1 T \mathrm{d}j. \tag{34}$$

Government spending G_t is not valued by households.

Asset Market Clearing The aggregate capital stock is the total amount of illiquid assets in the economy,

$$K_t = \int_0^1 a_{jt} \mathrm{d}j.$$

The market for capital is competitive, so the return on the illiquid asset r_t^a is simply the rental rate of capital.

The supply of liquid assets is fixed exogenously at $B_t = B^*$, where B^* is the steady state demand for liquid assets given $r_b^* = 0.005$ (discussed below). For simplicity, we assume that interest payments on the liquid assets come from outside the economy.

4.1.2 Equilibrium

The household-level state variables are illiquid asset holdings a, liquid asset holdings b, and labor productivity z. The aggregate state variables are aggregate productivity Z_t and the cross-sectional distribution of households over their individual state $g_t(a, b, z)$. As in Section 2, we denote an equilibrium object conditional on a particular realization of the aggregate state $(g_t(a, b, z), Z_t)$ with a subscript t.

Households The household's Hamilton-Jacobi-Bellman equation is given by

$$\begin{aligned}(\rho + \zeta) v_t(a, b, z) = &\max_{c,d} \log c + \partial_b v_t(a, b, z)(T + (1 - \tau) w_t e^z + r_t^b(b) b - \chi(d, a) - c - d) \\ &+ \partial_a v_t(a, b, z)(r_t^a a + d) + \sum_{z'} \lambda_{zz'}(v_t(a, b, z') - v_t(a, b, z)) + \frac{1}{\mathrm{d}t} \mathbb{E}_t[\mathrm{d}v_t(a, b, z)].\end{aligned} \tag{35}$$

The cross-sectional distribution $g_t(a,b,z)$ satisfies the Kolmogorov forward equation

$$\begin{aligned}\frac{\mathrm{d}g_t(a,b,z)}{\mathrm{d}t} =& -\partial_a\left(s_t^a(a,b,z)g_t(a,b,z)\right) - \partial_b\left(s_t^b(a,b,z)g_t(a,b,z)\right) \\ &- \sum_{z'}\lambda_{zz'}g_t(a,b,z) + \sum_{z'}\lambda_{z'z}g_t(a,b,z) \\ &- \zeta g_t(a,b,z) + \zeta\delta(a)\delta(b)g^*(z),\end{aligned} \tag{36}$$

where s_t^a and s_t^b are the optimal drifts in illiquid and illiquid assets implied by (35), $g^*(z)$ is the ergodic distribution of z, and δ is the Dirac delta function with $\delta(a)\delta(b)$ capturing birth at $a = b = 0$.

Firms The equilibrium conditions for the production side are the firm optimality conditions, together with the process for aggregate productivity:

$$\begin{aligned}r_t^a &= \alpha e Z_t K_t^{\alpha-1}\overline{L}^{1-\alpha} - \delta \\ w_t &= (1-\alpha)e^{Z_t}K_t^{\alpha}\overline{L}^{-\alpha} \\ \mathrm{d}Z_t &= -\eta Z_t\mathrm{d}t + \sigma\mathrm{d}W_t.\end{aligned}$$

Market Clearing Capital market clearing is given by

$$K_t = \int a g_t(a,b,z)\mathrm{d}a\mathrm{d}b\mathrm{d}z.$$

Liquid asset market clearing is given by

$$B = \int b g_t(a,b,z)\mathrm{d}a\mathrm{d}b\mathrm{d}z.$$

Given these conditions, as well as the government budget constraint (34), the market for output clears by Walras' law.

4.2 Calibration

We calibrate the steady state of the model without aggregate shocks to match key features of the cross-sectional distributions of household income and balance sheets. Our calibration closely follows Kaplan, Moll and Violante (2016).

Exogenously Set Parameters We choose the quarterly death rate $\zeta = 1/180$ so that households live 45 years on average. We set the tax rate $\tau = 30\%$ and set the lump sum transfer T to 10% of steady-state output. Given our labor productivity process, this policy implies that in steady state around 35% of households receive a net transfer from the government, consistent with the Congressional Budget Office (2013). We interpret borrowing in the liquid asset as unsecured credit and therefore set the borrowing limit $\underline{b}$ at one times average quarterly labor income.

We set the capital share in production $\alpha = 0.4$ and the annual depreciation rate on capital $\delta = 0.075$. With an equilibrium steady-state ratio of capital to annual output of 3.0 (see below) this implies an annual return on illiquid assets r^a of 5.8%.

Labor Productivity Shocks Following Kaplan, Moll and Violante (2016), we assume that the discrete-state process for labor productivity is a discretized version of the following continuous-state process. The logarithm of idiosyncratic labor productivity is the sum of two independent components

$$\log z_{jt} = z_{1,jt} + z_{2,jt}, \tag{37}$$

where each process is follows the jump-drift process

$$\mathrm{d}z_{i,jt} = -\beta_i z_{i,jt}\mathrm{d}t + \mathrm{d}J_{i,jt}. \tag{38}$$

Jumps arrive for component i at Poisson arrival rate λ_i. Conditional on a jump, a new log-earnings state $z_{j,it}$ is drawn from a normal distribution with mean zero and variance σ_j^2. Between jumps, the process drifts toward zero at rate β_i.[53] The parameters σ_i govern the size of the shocks, the parameters β_i govern the persistence of the shocks, and the parameters λ_i govern the frequency of arrival of shocks.

Jump-drift processes of this form are closely related to discrete-time AR(1) processes, with the modification that shocks arrive at random, rather than deterministic, dates. Allowing for the random arrival of shocks is important for matching the leptokurtic nature of annual income growth rates, which we discuss below. It is also important for matching observed household portfolio choices of liquid and illiquid assets. If the majority of earnings shocks are transitory and frequent (high β high λ), households would accumulate a buffer stock of liquid assets to self-insure. On the other hand, if earnings shocks are persistent and

[53] See Kaplan, Moll and Violante (2016) for a formal description of these processes.

Table 4: Targeted Labor Income Moments

Moment	Data	Model Estimated	Model Discretized
Variance: annual log earns	0.70	0.70	0.76
Variance: 1yr change	0.23	0.23	0.21
Variance: 5yr change	0.46	0.46	0.46
Kurtosis: 1yr change	17.8	16.5	17.3
Kurtosis: 5yr change	11.6	12.1	10.9
Frac 1yr change $< 10\%$	0.54	0.56	0.64
Frac 1yr change $< 20\%$	0.71	0.67	0.70
Frac 1yr change $< 50\%$	0.86	0.85	0.86

Notes: Moments of the earning process targeted in the calibration. "Data" refers to SSAA data on male earnings from Guvenen et al. (2015). "Model Estimated" refers to the continuous process (37) and (38). "Model Discretized" refers to discrete Poisson approximation of the process used in model computation.

infrequent (low β, low λ), households would prefer to save in high-return illiquid assets and pay the transaction costs to rebalance their portfolio when shocks occur.

Recent work by Guvenen et al. (2015) shows that changes in annual labor income are extremely leptokurtic, meaning that most absolute annual income changes are small but a small number are very large. We use the extent of this leptokurtosis, together with standard moments on the variance of log earnings and log earnings growth rates, to estimate the parameters of the earnings process (37) and (38). The moments we match, together with the fit of the estimated model, are shown in Table 4.

The estimated parameters in Table 5 indicate that the two jump-drift processes can be broadly interpreted as a transitory and a persistent component. The transitory component ($j = 1$) arrives on average once every three years and has a half-life of around one quarter. The persistent component ($j = 2$) arrives on average once every 38 years and has a half-life of around 18 years. In the context of an infinite-horizon model the persistent component can be interpreted as a "career shock." We discretize the continuous process (38) using 10 points for the persistent component and 3 points for the transitory component. The fit of the discretized process for the targeted moments is shown in Table 4.

Table 5: Estimated Labor Income Process

Parameter		Component $j=1$	Component $j=2$
Arrival rate	λ_j	0.080	0.007
Mean reversion	β_j	0.761	0.009
St. Deviation of innovations	σ_j	1.74	1.53

Notes: Parameters of the income process (37) and (38) estimated to match the moments in 4. The $j=1$ component arrives on average once every three years with half-life approximately one quarter. The $j=2$ component arrives once every 38 years with half-life approximately 18 years.

Adjustment Costs and Discount Factor The five remaining parameters on the household side of the model – the discount rate ρ, the borrowing wedge κ, and the parameters of the adjustment cost function χ_0, χ_1, and χ_2 – jointly determine the incentives of households to accumulate liquid and illiquid assets. We choose these parameters to match five moments of household balance sheets from the Survey of Consumer Finances 2004: the mean of the illiquid and liquid wealth distributions, the fraction of poor hand-to-mouth households (with $b=0$ and $a=0$), the fraction of wealthy hand-to-mouth households (with $b=0$ and $a>0$), and the fraction of households with negative assets. We match mean illiquid and liquid wealth so that the model is consistent with the aggregate wealth in the U.S. economy. We match the fraction of hand-to-mouth households because these households have higher than average marginal propensities to consume. See Kaplan, Moll and Violante (2016) for details on the classification of liquid and illiquid assets.

Table 6 shows that our calibrated model matches these five moments well. The implied annual discount rate is 5.8% annually and the annual borrowing wedge is 8.1% annually. Figure 4 plots the calibrated adjustment cost function together with the steady state distribution of quarterly deposits. The transaction cost is less than 1% of the transaction for small transactions and rises to around 10% of the transaction for a quarterly transaction that is 2% of illiquid assets. The function has a kink at $d=0$, which generates a mass of households who neither deposit nor withdraw.

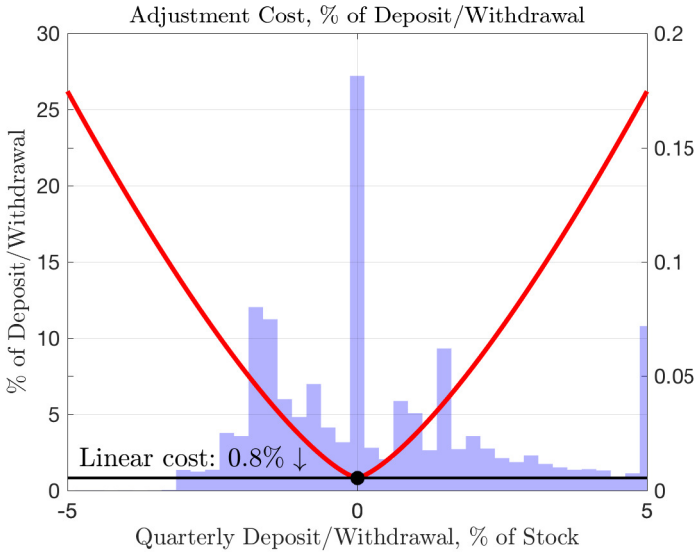
The calibrated distributions of liquid and illiquid wealth are displayed in Figure 5. Approximately 28% of households are hand-to-mouth (i.e. have zero liquid wealth) and another 14% have negative liquid wealth. Roughly two-thirds of the hand-to-mouth households are "wealthy hand-to-mouth," i.e. have positive illiquid assets, while the remaining one-third

Table 6: Targeted Wealth Distribution Moments

	Target	Model
Mean illiquid assets (multiple of annual GDP)	3.000	3.000
Mean liquid assets (multiple of annual GDP)	0.375	0.375
Frac. with $b = 0$ and $a = 0$	0.100	0.105
Frac. with $b = 0$ and $a > 0$	0.200	0.172
Frac. with $b < 0$	0.150	0.135

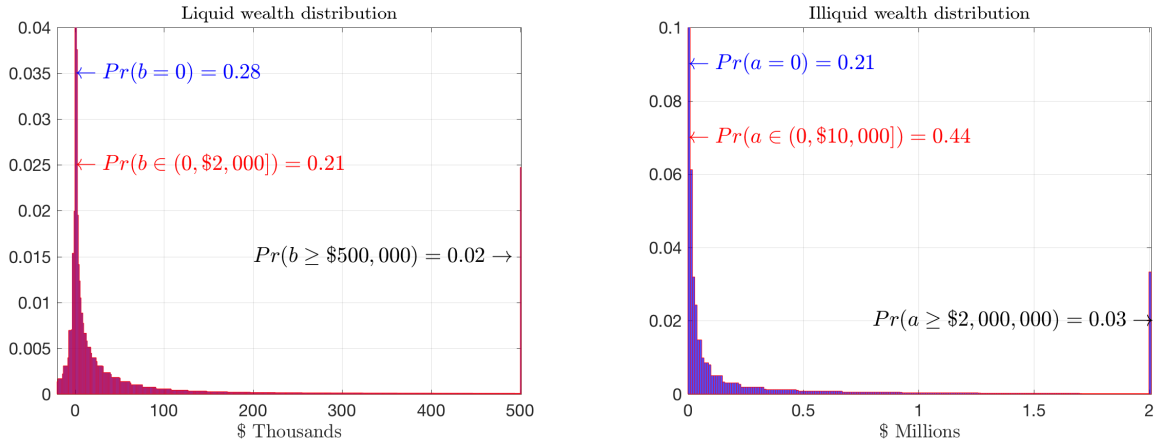
Notes: Moments of asset distribution targeted in calibration. Data source: SCF 2004. Liquid assets are revolving consumer debt, deposits, corporate bonds, and government bonds. Illiquid assets are net housing, net durables, corporate equity, and private equity.

Figure 4: Calibrated Adjustment Cost Function



Notes: Solid line plots adjustment costs as a fraction of the amount being transacted d, $\chi(d,a)/d$, where $\chi(d,a) = \chi_0|d| + \chi_1 \left|\frac{d}{a}\right|^{\chi_2} a$. Histogram displays the steady state distribution of deposit rates d/a.

Figure 5: Liquid and Illiquid Wealth Distribution in Steady State



(a) Liquid assets b (b) Illiquid assets a

Notes: Steady state distributions of liquid and illiquid wealth in the calibrated model.

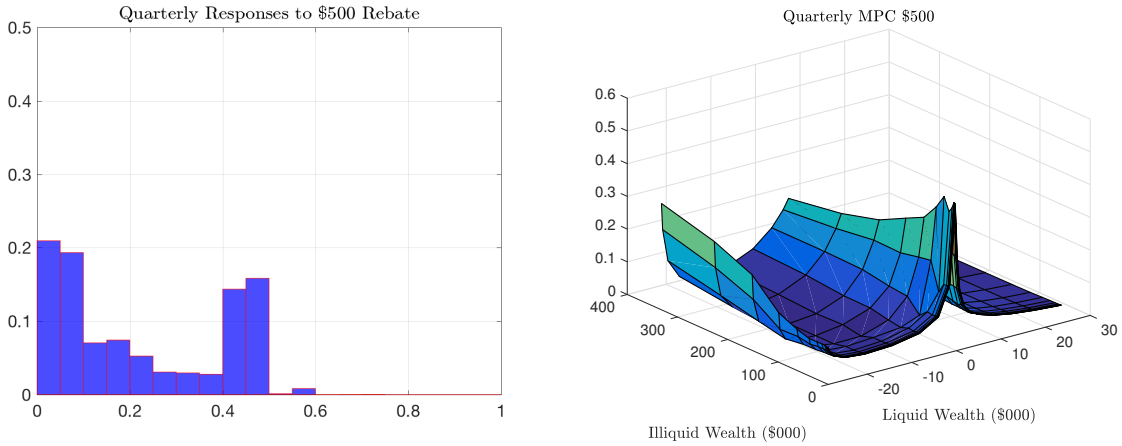
are "poor hand-to-mouth," i.e. have zero illiquid assets. Both distributions are extremely skewed; 3% of households have more than $\$2,000,000$ in illiquid assets and the top 10 percent hold 85% of total illiquid wealth in the economy.

The presence of hand-to-mouth households generates a distribution of marginal propensities to consume in line with empirical evidence. The average quarterly MPC out of a \$500 cash windfall is 22.5%, in line with the empirical estimates of Johnson, Parker and Souleles (2006) and Parker et al. (2013). The average number is composed of high MPCs for hand-to-mouth households (around 0.4) and small MPCs for non-hand-to-mouth households. This bimodality can be seen in Figure 6(a) and is consistent with recent work Fagereng, Holm and Natvik (2016).[54] Figure 6(b) shows that only households with zero (or very negative) liquid wealth have substantial MPCs, even for households with positive illiquid assets.

Aggregate Shocks As in Section 2, we set the rate of mean reversion of aggregate productivity shocks η to ensure that their quarterly autocorrelation $e^{-\eta} \approx 1 - \eta = 0.75$, and we set the volatility of innovations $\sigma = 0.007$.

[54] Fagereng, Holm and Natvik (2016) study consumption responses to lottery winnings using Norwegian administrative data. They find that MPCs are high for households with nearly zero liquid assets, even if the household has positive illiquid assets.

Figure 6: Heterogeneity in MPCs Across Households



(a) Distribution in Steady State

(b) MPC Function

Notes: Quarterly MPCs out of a \$500 windfall in steady state. The MPC over a period τ is $MPC_\tau(a,b,z) = \frac{\partial C_\tau(a,b,z)}{\partial b}$, where $C_\tau(a,b,z) = \mathbb{E}\left[\int_0^\tau c(a_t,b_t,z_t)\mathrm{d}t | a_0 = a, b_0 = b, z_0 = z\right]$.

4.3 Performance of Computational Method

Our discretization of the individual state space (a,b,z) contains $N = 60,000$ points, implying that the total unreduced dynamic system contains more than $120,000$ equations in $120,000$ variables.[55] We reduce the value function $\widehat{\mathbf{v}}_t$ using the spline approximation discussed in Section 3.4, bringing the size of the value function down from $N = 66,000$ gridpoints to $k_v = 2,145$ knot points.

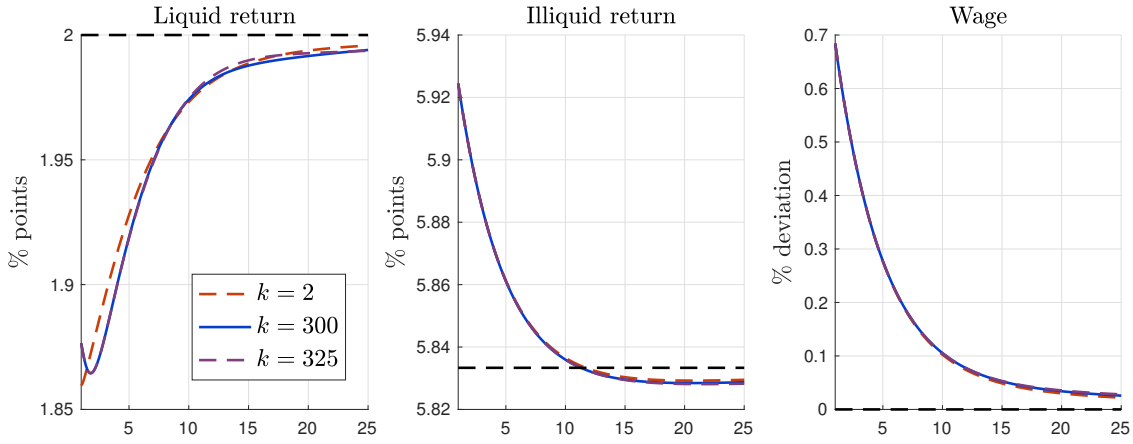
Failure of Approximate Aggregation We reduce the distribution $\widehat{\mathbf{g}}_t$ using a $k = 300$ order observability matrix to form the basis $\mathbf{X}$. In the finite limit where k is equal to the size of the unreduced state space, the reduced model converges to the true unreduced model. Figure 7 shows that the impulse responses of the three prices in the model – liquid return, the illiquid return, and the wage – appear to have converged by $k = 300$.

The fact that the distribution reduction step requires $k > 1$ suggests that "approximate aggregation" does not hold in the two-asset model.[56] Figure 7 shows that using $k = 2$

[55]The two-asset model is so much larger than the simple Krusell and Smith (1998) model because the individual state space is three-dimensional. To ensure an accurate approximation of the steady state, we use 30 grid points for labor productivity, 40 points for the illiquid asset, and 50 points for the liquid asset. The total number of grid points is therefore $N = 30 \times 40 \times 50 = 60,000$.

[56]Recall that $k = 1$ does provide an accurate approximation in the simple Krusell and Smith (1998) model.

Figure 7: Impulse Responses for Different Orders of Distribution Reduction



Notes: impulse responses to an instantaneous positive unit standard deviation size shock (Dirac delta function) to aggregate TFP. "k = 2" corresponds to distribution reduction based on an order 2 observability matrix. "k = 300" corresponds to distribution reduction based on an order 300 observability matrix. "k = 325" corresponds to distribution reduction based on an order 325 observability matrix. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$.

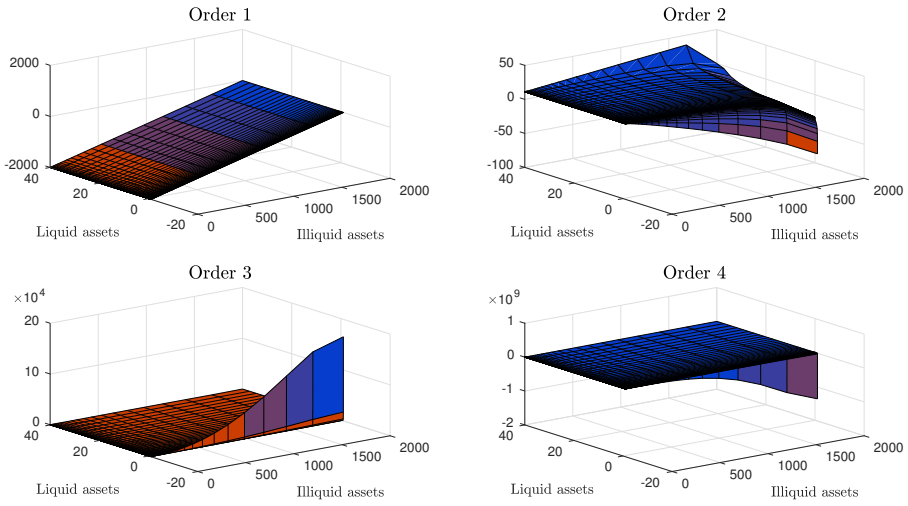
provides a poor approximation of the model's dynamics, particularly for the liquid return r_t^b. This result suggests that approximating the distribution with a small number of moments using Krusell and Smith (1998)'s procedure would be infeasible in this model.[57]

There are two main reasons why approximate aggregation does not hold in the two-asset model. First, recall that the reason for approximate aggregation in the Krusell and Smith (1998) model is that consumption functions are approximately linear in wealth, except for hand-to-mouth households near the borrowing constraint. However, in the one-asset model these households do not contribute very much to the aggregate capital stock (by virtue of holding very little capital) and hence their consumption dynamics are not important for the dynamics of aggregate capital. In contrast, in the two-asset model, there are a substantial number of wealthy hand-to-mouth households who have both highly non-linear consumption functions (by virtue of holding very little liquid wealth) and constitute a non-trivial contribution to the dynamics of aggregate capital (by virtue of holding substantial quantities of illiquid wealth).

Consistent with this intuition, the basis $\mathbf{X}$ of our distribution reduction places weight

[57] As discussed in Section 3.3, with endogenous decision rules our method does not necessarily provide the most efficient choice of basis $\mathbf{X}$. It is possible that by following the iterative procedure outlined in that section, one could obtain an accurate reduced model with $k < 300$.

Figure 8: Basis Vectors for Approximating Distribution in Two-Asset Model



Notes: columns of the observability matrix $\mathcal{O}(\mathbf{B}_{pg}, \mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{B}_{pg})^{\mathrm{T}}$ corresponding to aggregate capital K_t. The four panels plot the first four columns of the observability matrix over liquid and illiquid assets conditional on the median realization of labor productivity z.

on regions of the state space which have a significant fraction of hand-to-mouth households. Figure 8 plots the first four column vectors of the observability matrix associated with forecasting the aggregate capital stock K_t.[58] Each panel plots a given column of the matrix over liquid and illiquid assets, conditional on the median realization of labor productivity z. The first column captures exactly the mean of the illiquid asset distribution, which corresponds to aggregate capital. The next three columns focus on regions of the state space in which households have low liquid assets – and so are hand-to-mouth – as well as high illiquid assets – and so contribute substantially to aggregates capital.

The second reason why approximate aggregation breaks down in the two-asset model is that households must track the liquid return r_t^b in addition to the aggregate capital stock K_t.[59] The dynamics of r_t^b feature stronger distributional dependence than the dynamics of K_t because the liquid asset is in fixed supply B^*; an increase in savings in one region of the state space must be met with a decrease in savings elsewhere in the state space. Indeed, Figure 7 shows that the liquid return is the most poorly approximated variable in a $k = 2$ order approximation.[60]

[58] Recall that our basis $\mathbf{X}$ spans the subspace generated by the columns of the observability matrix.

[59] Note that the aggregate capital stock is sufficient to compute the wage w_t and illiquid return r_t^a.

[60] We have also computed a version of the model in which we drop the liquid asset market clearing condition, and instead assume that the liquid return r_t^b is fixed and that the bond supply adjusts perfectly elastically to

Table 7: Run Time for Solving Two-Asset Model

	$k = 300$	$k = 150$
Steady State	56.64 sec	56.64 sec
Derivatives	13.97 sec	13.97 sec
Dim reduction	199.71 sec	67.48 sec
Linear system	12.89 sec	7.70 sec
Simulate IRF	3.03 sec	2.31 sec
Total	**286.24 sec**	**148.10 sec**

Notes: Time to solve the two-asset model on a `MacBook Pro 2016` laptop with `3.3 GHz` processor and `16 GB RAM`, using `Matlab R2016b` and our code toolbox. k refers to order of the observability matrix used to compute basis $\mathbf{X}$. "Steady state" reports time to compute steady state. "Derivatives" reports time to compute derivatives of discretized equilibrium conditions. "Dim reduction" reports time to compute both the distribution and value function reduction. "Linear system" reports time to solve system of linear differential equations. "Simulate IRF" reports time to simulate impulse responses reported. "Total" is the sum of all these tasks.

Run Time Our numerical toolbox solves and simulates the two-asset model in `4 mins, 46 secs`. Table 7 decomposes the total runtime into various tasks and show that over two-thirds of the time is spent in the model reduction step. In order to illustrate how the method scales with k, Table 7 also reports the runtime for a smaller $k = 150$ order observability matrix. With this smaller approximation of the distribution, the total runtime falls to `2 mins, 28 secs`.

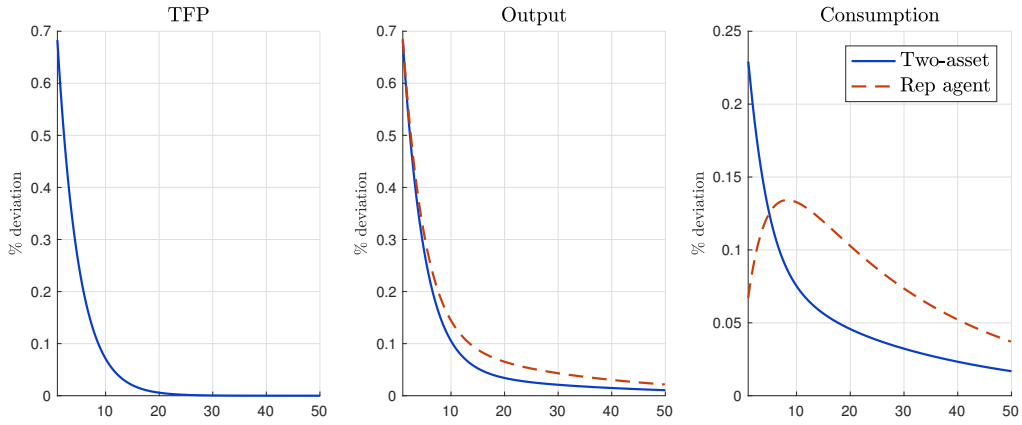
4.4 Impulse Response to TFP Shock Z_t

Figure 9 plots the impulse responses of aggregate output and consumption to a positive aggregate productivity shock Z_t. Higher productivity directly increases output Y_t through the production function. It also increases the return on capital r_t^a, which encourages capital accumulation and further increases output over time. The marginal product of labor also rises, increasing the real wage w_t. Both of these price increases lead to an increase in household income.

The increase in household income has differential effects on the consumption of hand-to-mouth and non-hand-to-mouth households. Non-hand-to-mouth households respond primar-

the meet the demand. In this version of the model, a $k = 100$ order observability matrix appears sufficient to reduce the distribution.

Figure 9: Aggregate Impulse Responses to Aggregate Productivity Shock Z_t



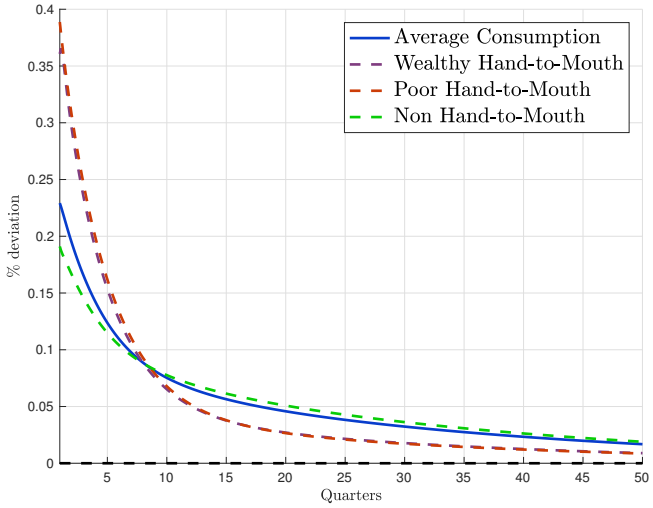
Notes: impulse responses to an instantaneous positive unit standard deviation size shock (Dirac delta function) to aggregate TFP. "Two-asset" refers to the two-asset model developed in Section 4.1. "Representative agent" refers to the representative agent version of the model, in which the households are replaced by a representative household who can only save in aggregate capital; see Appendix A.5 for details. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$ using an implicit updating scheme.

ily to the change in their *permanent income*. The change in permanent income is relatively small (because the productivity shock Z_t is transitory) but is persistent (because of the dynamics of the capital stock). In contrast, hand-to-mouth households respond primarily to the change in their *current income*. The change in current income is larger than the change in permanent income and is less persistent. Consistent with this logic, Figure 10 shows that the consumption of the hand-to-mouth households responds twice as much as average consumption upon impact, but dies out more quickly. Due to the presence of these hand-to-mouth households, the impulse response of aggregate consumption to a productivity shock is very different in the two-asset model compare with a representative agent model.

5 Aggregate Consumption Dynamics in Two-Asset Model

In this section, we use the two-asset model developed in Section 4 to illustrate how inequality shapes the dynamics of macroeconomic aggregates. Specifically, we show that although the model is parameterized to match *household-level* data, it also matches key features of the joint dynamics of *aggregate* consumption and income.

Figure 10: Consumption Response by Hand-to-Mouth Status



Notes: impulse responses to an instantaneous positive unit standard deviation size shock (Dirac delta function) to aggregate TFP. "Wealthy hand-to-mouth" refers to households with $b = 0$ and $a > 0$. "Poor hand-to-mouth" refer to households with $b = 0$ and $a = 0$. "Average consumption" is aggregate consumption. "Non hand to mouth" is computed as the residual. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$ using an implicit updating scheme.

Table 8: Targeted Moments of Real GDP Growth

	Data	Model
$\sigma\left(\Delta \log Y_t\right)$	0.89	0.89
$\text{Corr}(\Delta \log Y_t, \Delta \log Y_{t-2})$	0.21	0.20

Notes: targeted moments of per capita real GDP per capita growth, 1953q1 - 2016q2.

5.1 Model With Growth Rate Shocks

Following a long line of work in the consumption dynamics literature, such as Campbell and Mankiw (1989), we compare the predictions of our model to data on aggregate consumption and aggregate income *growth*. However, the Ornstein-Uhlenbeck process for aggregate productivity we have been working with so far implies that aggregate income growth is negatively autocorrelated, which is at odds with the data. Therefore, we modify the shock process so that aggregate productivity *growth*, rather than the level, follows an Ornstein-Uhlenbeck process. In addition, we assume that the liquid interest rate r_t^b is fixed at its steady state value $r_b^* = 0.005$ and that the liquid asset supply adjusts perfectly elastically at this price, as in a small open econonomy. This simplifying assumption ensures that the only time-varying interest rate is the return on capital, making the comparison with representative agent models more transparent. Both of these modifications apply to this section only.

Production and Aggregate Shock Process The production function with growth rate shocks is

$$Y_t = K_t^{\alpha} \left(Q_t \overline{L}\right)^{1-\alpha},$$

where Q_t is aggregate productivity. Aggregate productivity growth follows the process

$$\begin{aligned} \mathrm{d} \log Q_t &= Z_t \mathrm{d}t \\ \mathrm{d}Z_t &= -\eta Z_t \mathrm{d}t + \sigma \mathrm{d}W_t, \end{aligned}$$

where $\mathrm{d}W_t$ is an innovation to a standard Brownian motion. Hence, aggregate productivity growth is subject to the Ornstein-Uhlenbeck process Z_t.

Given the other calibrated parameters from Section 4, we choose the parameters of the TFP growth process Z_t so that equilibrium dynamics of aggregate income growth $\Delta \log Y_t$

match two key features of the data: the standard deviation of income growth $\sigma(\Delta \log Y_t)$ and the second-order autocorrelation of income growth $\text{Corr}(\Delta \log Y_t, \Delta \log Y_{t-2})$.[61] These moments in the data and the model's fit are reported in Table 8.

Model Computation Many equilibrium objects in the model are nonstationary due to the nonstationarity of aggregate productivity Q_t. We cannot directly apply our computational methodology in this setting, which relies on approximating the model's dynamics around a stationary equilibrium. Therefore, we detrend the model to express the equilibrium in terms of stationary objects; for details, see Appendix A.4.

5.2 Comparison to the Data

We focus our analysis on two sets of facts about the joint dynamics of aggregate consumption and income. The first set of facts, known as *sensitivity*, describe how aggregate consumption growth co-moves with predictable changes in aggregate income growth. The second set of facts, known as *smoothness*, refer to the extent of the time-series variation in aggregate consumption growth.

Sensitivity We present several measures of sensitivity in the top panel of Table 9. These measures all compute how predictable changes in income pass through to changes in consumption, but differ in two key respects. The first two measures of sensitivity are coefficients from ordinary least squares regressions, whereas the second two measures are coefficients from instrumental variables regressions. The second and fourth measures include real interest rates in the conditioning set, whereas the first and third measures do not. We present this range of measures to represent the range of approaches in the existing literature.

We measure aggregate income growth as the quarterly change in log real GDP per capita during the period 1953q1 to 2016q2. We measure aggregate consumption growth as the quarterly change in log real nondurables plus durable services consumption per capita during the same period. Finally, we measure the real interest rate as the real return on 90-day Treasury bills, adjusted for realized inflation.

In the data, all measures of sensitivity indicate that a substantial portion of aggregate income growth passes through to consumption growth. Consistent with the arguments in Campbell and Mankiw (1989) and Ludvigson and Michaelides (2001), among others, the

[61] We match the second-order autocorrelation, rather than the first, due to potential time aggregation issues, as discussed in Campbell and Mankiw (1989).

Table 9: Joint Dynamics of Consumption and Income

Sensitivity to Income				
	Data		**Models**	
		Two-Asset	*Rep Agent*	*Sp-Sa*
$\Delta \log C_t = \beta_0 + \beta_1 \Delta \log Y_{t-2} + \varepsilon_t$	0.12	0.14	0.12	0.16
	(0.03)			
$\Delta \log C_t = \beta_0 + \beta_1 \Delta \log Y_{t-2} + \beta_2 r_{t-2} + \varepsilon_t$	0.12	0.09	0.04	0.11
	(0.03)			
IV($\Delta \log C_t$ on $\Delta \log Y_t$| $\Delta \log Y_{t-2}$)	0.55	0.70	0.54	0.78
	(0.15)			
Campbell-Mankiw IV	0.49	0.40	0.004	0.50
	(0.15)			(calibrated)
Smoothness				
	Data		**Models**	
		Two-Asset	*Rep Agent*	*Sp-Sa*
$\frac{\sigma(\Delta \log C_t)}{\sigma(\Delta \log Y_t)}$	0.52	0.70	0.80	0.70
Corr($\Delta \log C_t, \Delta \log C_{t-2}$)	0.33	0.24	0.16	0.27

Notes: measures of sensitivity of aggregate consumption to income and the smoothness of aggregate consumption. In the data, aggregate consumption C_t is measured as the sum of real nondurable plus durable services, per capita, and aggregate income Y_t is real GDP per capita. Both series are quarterly 1953q1 - 2016q2. "Rep agent" refers to the representative agent model described in Appendix A.5. "Two-asset" refers to the full two-asset model. "Sp-Sa" refers to the spender-saver model described in Appendix A.5. "$\Delta \log C_t = \beta_0 + \beta_1 \Delta \log Y_{t-2} + \varepsilon_t$" refers to β_1 in the regression. "$\Delta \log C_t = \beta_0 + \beta_1 \Delta \log Y_{t-2} + \beta_2 r_{t-2} + \varepsilon_t$" refers to the coefficient β_1 in the regression. "IV($\Delta \log C_t$ on $\Delta \log Y_t | \Delta \log Y_{t-2}$)" refers to β_1 in the instrumental variables regression $\Delta \log C_t = \beta_0 + \beta_1 \Delta \log Y_t + \varepsilon_t$, using $\Delta \log Y_{t-2}$ to instrument for $\Delta \log Y_t$. "Campbell-Mankiw IV" refers to the β_1 in the instrumental variables regression $\Delta \log C_t = \beta_0 + \beta_1 \Delta \log Y_t + \beta_2 r_t + \varepsilon_t$, using $\Delta \log Y_{t-2}$, $\Delta \log Y_{t-3}$, $\Delta \log Y_{t-4}$, r_{t-2}, r_{t-3}, and r_{t-4} to instrument for the right hand side. We time-aggregate our continuous time model to the quarterly frequency by computing the simple average within a quarter.

representative agent model generates too little sensitivity once we condition on the real interest rate.[62] In contrast, the two-asset heterogeneous agent model generates substantial sensitivity of consumption growth to predictable changes in income growth.

Sensitivity in the two-asset model is driven by the presence of hand-to-mouth consumers who do not smooth their consumption over time. In the representative agent model, consumption jumps upon impact of the growth shock Z_t because permanent income immediately jumps to a new level. However, in the two-asset model, consumption of hand-to-mouth households jumps less upon impact – because the change in current income is smaller than the change in permanent income – but is more persistent. The persistence generates autocorrelation in consumption which allows the model to match the fact that consumption responds even to predictable changes in income.

Table 9 also reports the predictions of a simple spender-saver model in the spirit of Campbell and Mankiw (1989). It extends the representative agent model to include an exogenous fraction λ of households who are permanently hand-to-mouth. We calibrate the fraction of spenders λ to match the Campbell-Mankiw IV measure of consumption sensitivity. This reverse engineered model is also consistent with the degree of sensitivity in the data by construction. In contrast, our two-asset model has only been parameterized to match micro-level behavior, not aggregate sensitivity.

Smoothness We present two measures of smoothness in the bottom panel of Table 9. The first is the standard deviation of consumption growth relative to the standard deviation of income growth. In the data, consumption growth is about half as volatile as income growth. The second measure of smoothness is the second-order autocorrelation of consumption growth.

The two-asset heterogeneous agent model, the representative agent model, and the spender-saver model all over-predict the volatility of consumption growth relative to income growth. Consistent with the degree of sensitivity discussed above, both the two-asset model and the spender-saver model generate significant autocorrelation of consumption growth.

[62] In the special case of the representative agent model in which the interest rate is constant and income growth is a random walk, these sensitivity measures are exactly zero. The representative agent version of our model does not satisfy this special case, generating nonzero measures of sensitivity.

6 Business Cycle Dynamics of Inequality

The previous section explored how inequality shapes the joint dynamics of aggregate consumption and income. In this section, we briefly explore how aggregate shocks themselves shape the dynamics of inequality across households. However, with the Cobb-Douglas production function we have used so far, the distribution of labor income is given exogenously by the distribution of labor productivity shocks z. Therefore, we first extend the production side of the economy to include high- and low-skill workers which are not perfect substitutes with each other or with capital. We then explore the effects of shock to the productivity of unskilled labor, and a shock to the productivity of capital. By construction, this shock has differential effects across workers, generating substantial movements in income and consumption inequality. In addition, the resulting dynamics of aggregate variables are different from the representative agent counterpart of the model.

6.1 Model with Imperfect Substitutability Among Workers

Following Krusell et al. (2000), we modify the production function to feature two types of workers and capital-skill complementarity.

Production Structure The production function is

$$Y_t = \left[\mu(Z_t^U U_t)^\sigma + (1-\mu)\left(\lambda(Z_t^K K_t)^\rho + (1-\lambda)S_t^\rho\right)^{\frac{\sigma}{\rho}}\right]^{\frac{1}{\sigma}}, \tag{39}$$

where Z_t^U is an unskilled labor-specific productivity shock, Z_t^K is capital-specific productivity shock, U_t is the amount of unskilled labor, and S_t is the amount of skilled labor (all described in more detail below). The elasticity of substitution between unskilled labor and capital, which is equal to the elasticity between unskilled and skilled labor, is $\frac{1}{1-\sigma}$. The elasticity of substitution between skilled labor and capital is $\frac{1}{1-\rho}$. If, as in our calibration, $\sigma > \rho$, high-skill workers are complementary with capital.[63]

We posit a simple mapping from labor productivity z into skill. Recall that we modeled the logarithm of labor productivity as the sum of two components, $\log z = z_1 + z_2$. We estimated that z_1 is a transitory component and z_2 is a persistent component. With our estimated parameters, shocks to the persistent component arrive on average once every 38

[63] Krusell et al. (2000) assume that only equipment capital features capital-skill complementarity while structures capital has unitary elasticity of substitution. We omit structures capital for simplicity.

years. Hence, a natural interpretation of the persistent component in an infinite-horizon model is a "career shock." We therefore map workers into skills based on the realization of the persistent component – we label the top 50% of workers as high-skill and the bottom 50% as low-skill.

We assume that both aggregate productivity shocks follows Ornstein-Uhlenbeck process

$$\begin{aligned} \mathrm{d}\log Z_t^U &= -\eta_U \log Z_t^U \mathrm{d}t + \sigma_U \mathrm{d}W_t^U \\ \mathrm{d}\log Z_t^K &= -\eta_K \log Z_t^K \mathrm{d}t + \sigma_K \mathrm{d}W_t^K. \end{aligned}$$

where η_U and η_K control the rate of mean reversion and σ_U and σ_K control the size of innovations.

Calibration We set the elasticities of substitution in production to the estimated values in Krusell et al. (2000): $\sigma = 0.401$ and $\rho = -.495$. Since $\sigma > \rho$, the production function features capital-skill complementarity, i.e., capital-specific productivity shocks disproportionately favor skilled labor.

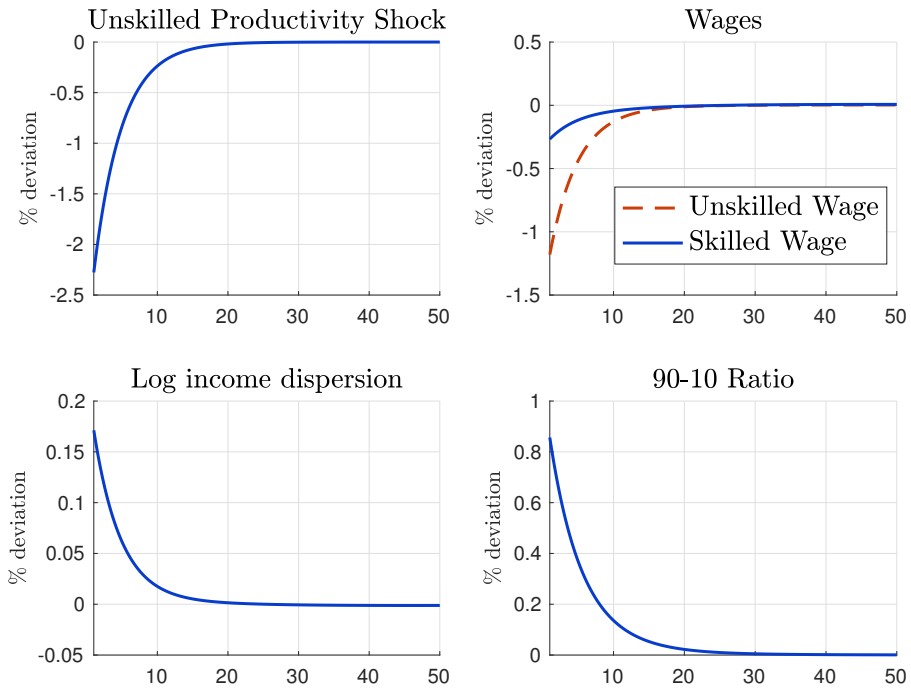
Given the values for these elasticities, and all the other calibrated parameters from Section 4.2, we choose the factor shares μ and λ to match two steady state targets. First, we target a steady state labor share of 60%, as in Section 4.2. Second, we target a steady state skill premium – the ratio of the average skilled worker's earnings to the average unskilled workers' earnings – of 1.97, which is the value of the college skill premium reported in Acemoglu and Autor (2011). This yields $\mu = 0.52$ and $\lambda = 0.86$.

We set the process for the unskilled-labor productivity shock to be equivalent to our factor-neutral productivity shock process in the case of Cobb-Douglas production. Therefore, as in Section 4.2, we set the rate of mean reversion to $\eta_U = 0.25$. We set the standard deviation of innovations σ_U so that they generate the same impact effect on output as the factor-neutral shocks in Section 4.2.

6.2 Inequality Dynamics Following Unskilled-Labor Specific Shock

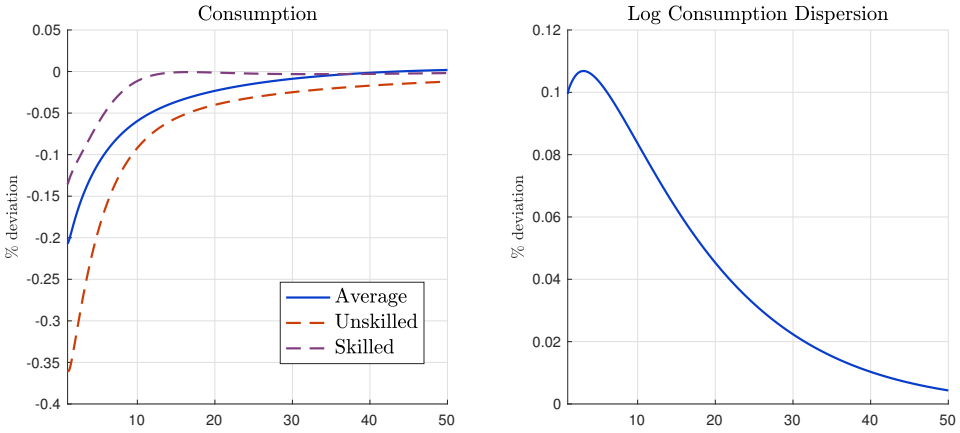
Figure 11 plots the impulse responses of key features of the distribution of income across households. The wage rate of unskilled workers falls fives times more than the wage rate of skilled workers, due to the fact that the shock directly affects the marginal product of unskilled workers and these workers are not perfect substitutes with skilled workers. Hence,

Figure 11: Impulse Responses to Unskilled Labor-Specific Productivity Shock



Notes: impulse responses to an instantaneous positive unit standard deviation size shock (Dirac delta function) to unskilled labor-specific productivity. "Unskilled wage" is the wage rate per efficiency unit of labor for unskilled workers. "Skilled wage" is the wage rate per efficiency unit of labor for skilled workers. "Log income dispersion" is the cross-sectional standard deviation of log pre-tax labor income across households. "90-10 Ratio" is the ratio of the 90th percentile of pre-tax labor income to the 10th percentile. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$ using an implicit updating scheme.

Figure 12: Impulse Responses to Unskilled Labor-Specific Productivity Shock



Notes: impulse responses to an instantaneous negative unit standard deviation size shock (Dirac delta function) to unskilled labor-specific productivity. "Unskilled" is the average consumption of unskilled workers. "Skilled" is the average consumption of skilled workers. "Average" is aggregate consumption. "Log consumption dispersion" is the cross-sectional standard deviation of log consumption across households. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$ using an implicit updating scheme.

the dispersion of pre-tax labor income across households increases by nearly 0.2% and the 90-10 percentile ratio increases by nearly 1%.[64]

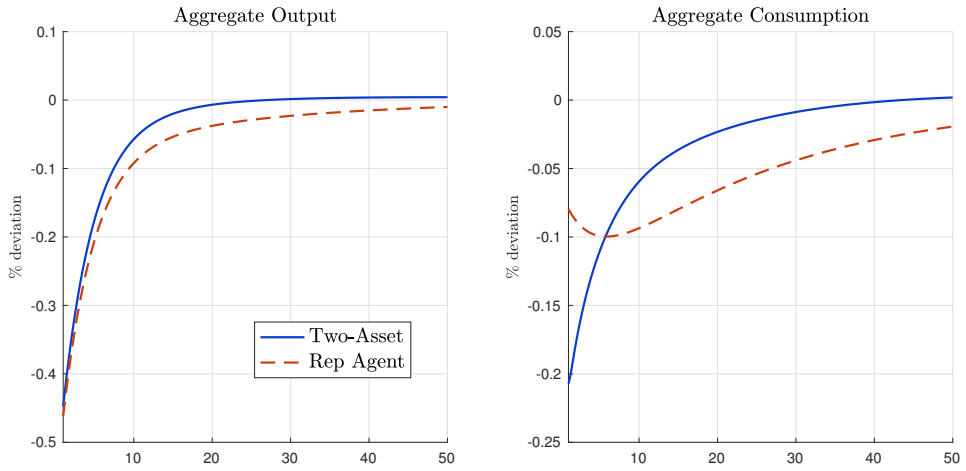
Figure 12 plots the impulse responses of features of the distribution of consumption across workers. The average consumption of low-skill workers falls more than twice the amount of high-skill workers. This differential effect reflects the combination of two forces. First, the shock decreases unskilled workers' wages more than skilled workers, so the permanent income of unskilled workers is lower. Second, unskilled workers are over 30% more likely to be hand-to-mouth, making them more sensitive to changes in income.

6.3 Aggregate Dynamics Following Unskilled-Labor Specific Shock

Figure 13 plots the impulse responses of aggregate output and consumption following the unskilled-specific shock, and compares the responses to the representative agent version of the model. Although the output responses are very similar across the two models, the trough in consumption is more than twice as low in the two-asset model than in the representative agent model.

[64]Recall that with Cobb-Douglas the dispersion of pre-tax labor income is constant.

Figure 13: Impulse Responses to Unskilled Labor-Specific Productivity Shock



Notes: impulse responses to an instantaneous negative unit standard deviation size shock (Dirac delta function) to unskilled labor-specific productivity. "Two-asset" refers to the two-asset model. "Rep agent" refers to the representative agent version of the model, described in Appendix A.5. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$ using an implicit updating scheme.

The severity of the consumption response in the two-asset model reflects the combination of two forces. First, the two-asset model features a substantial fraction of hand-to-mouth households that respond more strongly to income changes than the representative household. The presence of hand-to-mouth households also changed the consumption response to factor neutral shocks, as discussed in Section 4. Second, the unskilled labor-specific shock is concentrated among low-skill workers who are more likely to be hand-to-mouth. This concentration is absent in the factor-neutral shock case, and in that case, the difference between the two-asset and representative agent models is 25% smaller. Hence, the fact that the unskilled labor-specific shock is concentrated among a particular region of the distribution shapes aggregate business cycle dynamics of the model.

6.4 Inequality Dynamics Following Capital-Specific Shock

We close this section with a brief example to show that, due to capital-specific complementarity, a shock to capital-specific productivity Z_t^K can generate dynamics of income inequality. We set the rate of mean reversion $\eta_K = 0.25$ and calibrate the standard deviation of innovations σ_K so that it generates the same impact effect on output as the factor-neutral shocks in Section 4.2.

Figure 14: Impulse Responses to Capital-Specific Productivity Shock

Notes: impulse responses to an instantaneous positive unit standard deviation size shock (Dirac delta function) to capital-specific productivity. "Unskilled wage" is the wage rate per efficiency unit of labor for unskilled workers. "Skilled wage" is the wage rate per efficiency unit of labor for skilled workers. "Log income dispersion" is the cross-sectional standard deviation of log pre-tax labor income across households. We simulate the model by discretizing the time dimension with step size $\mathrm{d}t = 0.1$ using an implicit updating scheme.

Figure 14 shows that the capital shock increases labor income inequality. The left panel shows that high-skill wages increase by more than low-skill wages due to capital-skill complementarity; in response to the capital-specific shock, the representative firm substitutes toward skilled labor. Hence, the dispersion of labor income across households increases as well.

7 Conclusion

Our paper's main message is that two of the most common excuses that macroeconomists make for employing representative agent models are less valid than commonly thought. First, we develop an efficient and easy-to-use computational method for solving a wide class of general equilibrium heterogeneous agent macro models with aggregate shocks, thereby invalidating the excuse that these models are subject to extreme computational difficulties. Second, our results in Sections 5 and 6 show that inequality may matter greatly for the dynamics of standard macroeconomic aggregates. These results invalidate the excuse that heterogeneous agent models are unnecessarily complicated because they generate only limited additional explanatory power for aggregate phenomena.

Due to its speed, our method opens up the door to estimating macroeconomic models in which distributions play an important role with micro data. Existing attempts to bring macroeconomic models to the data, typically use either only aggregate time-series to discipline aggregate dynamics (in the case of representative agent models); or they use cross-sectional micro data at a given point in time to discipline a stationary equilibrium without aggregate shocks (in the case of heterogeneous agent models). Instead, future research should use *micro data capturing distributional dynamics over time*, i.e. panel data or repeated cross-sections. An important hurdle in this endeavour is that micro data, especially from surveys, are often inconsistent with national accounts data on macroeconomic aggregates (see e.g. Deaton, 2005). Attempts to produce time-series on distributional variables that capture 100 percent of national income like the Distributional National Accounts of Piketty, Saez and Zucman (2016) are welcome in this regard.

References

Acemoglu, D., and D. Autor (2011): "Skills, Tasks, and Technologies: Implications for Employment and Earnings," *Handbook of Labor Economics 4*, pp. 1043–1171.

Achdou, Y., J. Han, J.-M. Lasry, P.-L. Lions, and B. Moll (2015): "Heterogeneous Agent Models in Continuous Time," Discussion paper, Princeton University.

Aiyagari, S. R. (1994): "Uninsured Idiosyncratic Risk and Aggregate Saving," *The Quarterly Journal of Economics*, 109(3), 659–684.

Amsallem, D., and C. Farhat (2011): "Lecture Notes for CME 345: Model Reduction," https://web.stanford.edu/group/frg/course_work/CME345/.

Antoulas, A. (2005): *Approximation of Large-Scale Dynamical Systems*. SIAM Advances in Design and Control.

Auclert, A. (2014): "Monetary Policy and the Redistribution Channel," Discussion paper, MIT.

Bayer, C., R. Luetticke, L. Pham-Dao, and V. Tjaden (2015): "Precautionary Savings, Illiquid Assets, and the Aggregate Consequences of Shocks to Household Income Risk," Discussion paper, University of Bonn.

Blanchard, O. J., and C. M. Kahn (1980): "The Solution of Linear Difference Models under Rational Expectations," *Econometrica*, 48(5), 1305–11.

Bloom, N., M. Floetotto, N. Jaimovich, I. Saporta-Eksten, and S. Terry (2014): "Really Uncertain Business Cycles," Discussion paper.

Brunnermeier, M. K., and Y. Sannikov (2014): "A Macroeconomic Model with a Financial Sector," *American Economic Review*, 104(2), 379–421.

Campbell, J. (1998): "Entry, Exit, Embodied Technology, and Business Cycles," *Review of Economic Dynamics*, 1(2), 371–408.

Campbell, J. Y., and N. G. Mankiw (1989): "Consumption, Income and Interest Rates: Reinterpreting the Time Series Evidence," in *NBER Macroeconomics Annual 1989, Volume 4*, NBER Chapters, pp. 185–216. National Bureau of Economic Research.

Carroll, C., M. White, N. Palmer, D. Low, and A. Kaufman (2016): "Heterogenous Agents Resources & toolKit," https://github.com/econ-ark/HARK.

Christiano, L. (1989): "Comment on "Consumption, Income and Interest Rates: Reinterpreting the Time Series Evidence"," in *NBER Macroeconomics Annual 1989, Volume 4*, NBER Chapters, pp. 216–233. National Bureau of Economic Research.

Congressional Budget Office (2013): "The Distribution of Federal Spending and Taxes in 2006," Discussion paper, Congress of the United States.

Deaton, A. (2005): "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)," *The Review of Economics and Statistics*, 87(1), 1–19.

Den Haan, W., K. Judd, and M. Julliard (2010): "Computational Suite of Models with Heterogeneous Agents: Incomplete Markets and Aggregate Uncertainty," *Journal of Economic Dynamics and Control*, 34(1), 1–3.

Den Haan, W. J. (2010): "Comparison of solutions to the incomplete markets model with aggregate uncertainty," *Journal of Economic Dynamics and Control*, 34(1), 4–27.

Dotsey, M., R. King, and A. Wolman (1999): "State-Dependent Pricing and the General Equilibrium Dynamics of Money and Output," *Quarterly Journal of Economics*, pp. 655–690.

Fagereng, A., M. B. Holm, and G. J. Natvik (2016): "MPC Heterogeneity and Household Balance Sheets," Discussion paper, Statistics Norway.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song (2015): "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?," NBER Working Papers 20913, National

Bureau of Economic Research.

He, Z., and A. Krishnamurthy (2013): "Intermediary Asset Pricing," *American Economic Review*, 103(2), 732–770.

Ilut, C. L., and M. Schneider (2014): "Ambiguous Business Cycles," *American Economic Review*, 104(8), 2368–2399.

Johnson, D. S., J. A. Parker, and N. S. Souleles (2006): "Household Expenditure and the Income Tax Rebates of 2001," *American Economic Review*, 96(5), 1589–1610.

Kaplan, G., B. Moll, and G. L. Violante (2016): "Monetary Policy According to HANK," Working Papers 1602, Council on Economic Policies.

Kaplan, G., and G. L. Violante (2014): "A Model of the Consumption Response to Fiscal Stimulus Payments," *Econometrica*, 82(4), 1199–1239.

Krusell, P., L. Ohanian, V. Rios-Rull, and G. Violante (2000): "Capital-Skill Complementarity and Inequality: A Macroeconomic Analysis," *Econometrica*, 68, 1029–1053.

Krusell, P., and A. A. Smith (1998): "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, 106(5), 867–896.

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

Ludvigson, S. C., and A. Michaelides (2001): "Does Buffer-Stock Saving Explain the Smoothness and Excess Sensitivity of Consumption?," *American Economic Review*, 91(3), 631–647.

McKay, A. (2017): "Time-Varying Idiosyncratic Risk and Aggregate Consumption Dynamics," Discussion paper, Boston University.

McKay, A., E. Nakamura, and J. Steinsson (2015): "The Power of Forward Guidance Revisited," NBER Working Papers 20882, National Bureau of Economic Research.

McKay, A., and R. Reis (2013): "The Role of Automatic Stabilizers in the U.S. Business Cycle," NBER Working Papers 19000, National Bureau of Economic Research.

Mongey, S., and J. Williams (2016): "Firm Dispersion and Business Cycles: Estimating Aggregate Shocks Using Panel Data," Working paper, NYU.

Parker, J. A., N. S. Souleles, D. S. Johnson, and R. McClelland (2013): "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, 103(6), 2530–53.

Piketty, T., E. Saez, and G. Zucman (2016): "Distributional National Accounts: Methods and Estimates for the United States," NBER Working Papers 22945, National Bureau of Economic Research, Inc.

Preston, B., and M. Roca (2007): "Incomplete Markets, Heterogeneity and Macroeconomic Dynamics," NBER Working Papers 13260, National Bureau of Economic Research, Inc.

Reiter, M. (2009): "Solving heterogeneous-agent models by projection and perturbation," *Journal of Economic Dynamics and Control*, 33(3), 649–665.

——— (2010): "Approximate and Almost-Exact Aggregation in Dynamic Stochastic Heterogeneous-Agent Models," Economics Series 258, Institute for Advanced Studies.

Terry, S. (2017): "Alternative Methods for Solving Heterogeneous Firm Models," Discussion paper, Boston University.

Veracierto, M. (2002): "Plant Level Irreversible Investment and Equilibrium Business Cycles," *American Economic Review*, 92, 181–197.

Winberry, T. (2016): "A Toolbox for Solving and Estimating Heterogeneous Agent Macro Models," Working paper, University of Chicago.

A Appendix

A.1 Fully Recursive Formulation of Krusell-Smith (1998)

When writing the equilibrium conditions (2) to (7), we used recursive notation with respect to the idiosyncratic states (a, z) but time-dependent notation with respect to the aggregate states (g, Z). For completeness, this Appendix shows how to write the corresponding equations using fully recursive notation, and how to obtain the hybrid-notation conditions in the main text from the equations using fully recursive notation.

To this end, define the wage and interest rate as a function of the state variables (g, Z)

$$w(g, Z) = (1 - \alpha)e^Z K(g)^\alpha, \tag{40}$$

$$r(g, Z) = \alpha e^Z K(g)^{\alpha-1} - \delta, \tag{41}$$

$$\text{where } K(g) = \int ag(a, z)dadz \tag{42}$$

is the aggregate capital stock as a function of the distribution. Furthermore, define the "Kolmogorov Forward operator" $\mathcal{K}_Z$ that operates on distributions g as

$$(\mathcal{K}_Z g)(a, z) := -\partial_a[s(a, z, g, Z)g(a, z)] - \lambda_z g(a, z) + \lambda_{z'} g(a, z')$$

where $s(a, z, g, Z)$ is the optimal saving policy function (determined below). This operator maps distribution functions g into time derivatives of that distribution. Using this tool, one can for example write the Kolmogorov Forward equation (3) compactly as

$$\frac{\mathrm{d}g_t(a, z)}{\mathrm{d}t} = (\mathcal{K}_Z g_t)(a, z).$$

With this machinery in hand, the fully recursive, infinite-dimensional HJB equation is:

$$\begin{aligned}\rho V(a, z, g, Z) = \max_c \; & u(c) + \partial_a V(a, z, g, Z)(w(g, Z)z + r(g, Z)a - c) \\ & + \lambda_z(V(a, z', g, Z) - V(a, z, g, Z)) \\ & + \partial_Z V(a, z, g, Z)(-\eta Z) + \frac{1}{2}\partial_{ZZ} V(a, z, g, Z)\sigma^2 \\ & + \int \frac{\delta V(a, z, g, Z)}{\delta g(a, z)}(\mathcal{K}_Z g)(a, z)\mathrm{d}a\mathrm{d}z\end{aligned} \tag{43}$$

The first and second lines in this infinite-dimensional HJB equation capture the evolution of the *idiosyncratic* states (a, z) (just like in (2)). The third and fourth lines capture the

evolution of the *aggregate* states (g, Z). The third line captures the evolution of aggregate TFP Z with standard "Ito's Formula terms" involving the first and second derivatives of the value function with respect to Z. The fourth line captures the evolution of the distribution g. Since g is a function, it involves the functional derivative of V with respect to g at point (a, z), which we denote by $\delta V/\delta g(a, z)$. The equilibrium in fully recursive notation is then characterized by (43) together with (40), (41) and (42).

To understand the last term in (43), assume momentarily that the distribution is an N-dimensional vector $\mathbf{g} = (g_1, ..., g_N)$ rather than a function (i.e. an infinite-dimensional object). Then the HJB equation would be

$$\begin{aligned}\rho V(a,z,\mathbf{g},Z) = \max_c \; & u(c) + \partial_a V(a,z,\mathbf{g},Z)(w(\mathbf{g},Z)z + r(\mathbf{g},Z)a - c) \\ & + \lambda_z(V(a,z',\mathbf{g},Z) - V(a,z,\mathbf{g},Z)) \\ & + \partial_Z V(a,z,\mathbf{g},Z)(-\eta Z) + \frac{1}{2}\partial_{ZZ} V(a,z,\mathbf{g},Z)\sigma^2 \\ & + \sum_{i=1}^{N} \frac{\partial V(a,z,\mathbf{g},Z)}{\partial g_i}\dot{g}_i\end{aligned}$$

Since a functional derivative $\delta V/\delta g(a, z)$ is the natural generalization of the partial derivative $\partial V/\partial g_i$ to the infinite-dimensional case, if g is a function rather than a vector we get (43).

The equilibrium conditions (2) to (7) in the main text can be obtained from this system by evaluating "along the characteristic" (g_t, Z_t) that satisfies (3) and (4). In particular the value function $v_t(a, z)$ in (2) is obtained from evaluating (43) at (g_t, Z_t), i.e.

$$v_t(a,z) = V(a,z,g_t,Z_t).$$

In particular by Ito's Formula

$$\begin{aligned}\mathrm{d}v_t(a,z) = & \left(\partial_Z V(a,z,g_t,Z_t)(-\eta Z_t) + \frac{1}{2}\partial_{ZZ} V(a,z,g_t,Z_t)\sigma^2\right)\mathrm{d}t + \sigma\partial_Z V(a,z,g_t,Z_t)\mathrm{d}W_t \\ & + \int \frac{\delta V(a,z,g_t,Z_t)}{\delta g_t(a,z)}(\mathcal{K}_Z g_t)(a,z)\mathrm{d}a\mathrm{d}z\;\mathrm{d}t\end{aligned}$$

and hence using that $\mathbb{E}_t[\mathrm{d}W_t] = 0$

$$\frac{1}{\mathrm{d}t}\mathbb{E}_t[\mathrm{d}v_t(a,z)] = \partial_Z V(a,z,g_t,Z_t)(-\eta Z_t) + \frac{1}{2}\partial_{ZZ} V(a,z,g_t,Z_t)\sigma^2 + \int \frac{\delta V(a,z,g_t,Z_t)}{\delta g_t(a,z)}(\mathcal{K}_Z g_t)(a,z)\mathrm{d}a\mathrm{d}z$$

Similarly, the prices and capital stock in (5) to (7) are obtained by evaluating (40) to (42)

at (g_t, Z_t), i.e.

$$w_t = w(g_t, Z_t), \quad r_t = r(g_t, Z_t), \quad K_t = K(g_t).$$

A.2 Connection to Linearization of Representative Agent Models

This Appendix develops the relationship between our linearization of heterogeneous agent models and standard linearization of representative agent business cycle models. For illustration, consider a simple real business cycle model. As in our heterogeneous agent models in the main text, the equilibrium of this representative agent model is characterized by a forward-looking equation for controls, a backward-looking equation for the endogenous state, several static relations and an evolution equation for the exogenous state.

Defining the representative household's marginal utility $\Lambda_t := C_t^{-\gamma}$, the equilibrium conditions can be written as

$$\begin{aligned}
\frac{1}{\mathrm{d}t}\mathbb{E}_t[\mathrm{d}\Lambda_t] &= (\rho - r_t)\Lambda_t \\
\frac{\mathrm{d}K_t}{\mathrm{d}t} &= w_t + r_t K_t - C_t \\
dZ_t &= -\eta Z_t dt + \sigma dW_t \\
r_t &= \alpha e^{Z_t} K_t^{\alpha-1} - \delta \\
w_t &= (1-\alpha) e^{Z_t} K_t^{\alpha}
\end{aligned} \tag{44}$$

and where $C_t = \Lambda_t^{-1/\gamma}$. The first equation is the Euler equation. Marginal utility Λ_t is the single control variable; we could have alternatively written the Euler equation in terms of consumption C_t, but working with marginal utility is more convenient. The second equation is the evolution of the aggregate capital stock, which is the single endogenous state variable. The third equation is the stochastic process for aggregate productivity, which is the exogenous state variable. The last two equations define equilibrium prices.

The equilibrium conditions (14) of the simple Krusell and Smith (1998) model have the same structure as the representative agent model above. The discretized value function $\mathbf{v}_t$ is the endogenous control vector, analogous to marginal utility Λ_t (or aggregate consumption C_t) in the representative agent model. The distribution $\mathbf{g}_t$ is the endogenous state variable, analogous to aggregate capital K_t. Finally, TFP Z_t is the exogeneous state variable, just as in the representative agent model.

The representative agent model's equilibrium conditions can be linearized and the resulting linear system solved exactly as the heterogeneous agent model in the main text. Let

hatted variables denote deviations from steady state. Then we have the control variable $\widehat{\Lambda}_t$, the endogenous state $\widehat{K}_t$, the exogenous state Z_t, and the prices $\widehat{\mathbf{p}}_t = (\widehat{r}_t, \widehat{w}_t)$. We can thus write

$$\mathbb{E}_t \begin{bmatrix} \mathrm{d}\widehat{\Lambda}_t \\ \mathrm{d}\widehat{K}_t \\ \mathrm{d}Z_t \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{\Lambda\Lambda} & 0 & 0 & \mathbf{B}_{\Lambda p} \\ \mathbf{B}_{K\Lambda} & \mathbf{B}_{KK} & 0 & \mathbf{B}_{Kp} \\ 0 & 0 & -\eta & 0 \\ 0 & \mathbf{B}_{pK} & \mathbf{B}_{pZ} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \widehat{\Lambda}_t \\ \widehat{K}_t \\ Z_t \\ \widehat{\mathbf{p}}_t \end{bmatrix} \mathrm{d}t$$

Note that our linearized heterogeneous agent model (15) has the same form as this system.

A.3 Model Reduction and Proof of Proposition 1

This Appendix proves the results cited in the main text regarding our distribution reduction method. We also show that, in discrete time, our approach corresponds to matching the first k periods of the impulse response function.

A.3.1 Deterministic Model

As in the main text consider first the simplified model (26) which we briefly restate here:

$$\begin{aligned} \dot{\mathbf{g}}_t &= \mathbf{C}_{gg}\mathbf{g}_t, \\ p_t &= \mathbf{b}_{pg}\mathbf{g}_t. \end{aligned}$$

Solving this for p_t gives

$$\begin{aligned} p_t &= \mathbf{b}_{pg} e^{\mathbf{C}_{gg}t}\mathbf{g}_0 \\ &= \mathbf{b}_{pg}\left[\mathbf{I} + \mathbf{C}_{gg}t + \frac{1}{2}\mathbf{C}_{gg}^2 t^2 + \frac{1}{6}\mathbf{C}_{gg}^3 t^3 + \ldots\right]\mathbf{g}_0 \end{aligned}$$

We consider a reduced model obtained by means of *projection*. That is, we project the distribution $\mathbf{g}_t$ on a lower-dimensional space, and then analyze the dynamics of the corresponding reduced system. Of course, all that ultimately matters for the dynamics of the reduced system is that projection space itself, and not the particular basis chosen for the purpose of projection. Thus, for ease of presentation, we in the main text consider a semi-orthogonal basis $\mathbf{X}^{\mathrm{T}}$, i.e. a matrix $\mathbf{X}$ that satisfies $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$. Under this assumption, the reduced distribution is given by $\gamma_t = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{g}_t = \mathbf{X}^{\mathrm{T}}\mathbf{g}_t$. For the proofs in this Appendix, however, it will turn out to be more convenient to work with a non-normalized (non-semi-

orthogonal) basis. Specifically, we consider a pair of matrices $\mathbf{V}$, $\mathbf{W}^{\mathrm{T}}$ such that $\mathbf{W}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. This formulation nests our analysis from the main text with $\mathbf{X} = \mathbf{V}$ and $\mathbf{X}^{\mathrm{T}} = \mathbf{W}^{\mathrm{T}}$.

We then approximate the distribution $\mathbf{g}_t$ through $\gamma_t = (\mathbf{W}^{\mathrm{T}}\mathbf{V})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{g}_t = \mathbf{W}^{\mathrm{T}}\mathbf{g}_t$. Conversely, up to projection error, we have that $\mathbf{g}_t = \mathbf{V}\gamma_t$.[65] Differentiating with respect to time thus gives the reduced-system dynamics

$$\begin{aligned} \dot{\gamma}_t &= \mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\gamma_t \\ \widetilde{p}_t &= \mathbf{b}_{pg}\mathbf{V}\gamma_t \end{aligned}$$

Note that, with $\mathbf{V} = \mathbf{X}$, $\mathbf{W}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}$, this system simply collapses to the formulation in the main text. From here, we then get

$$\begin{aligned} \widetilde{p}_t &= \mathbf{b}_{pg}\mathbf{V}e^{(\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V})t}\mathbf{W}^{\mathrm{T}}\mathbf{g}_0 \\ &= \mathbf{b}_{pg}\mathbf{V}\left[\mathbf{I} + (\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V})t + \frac{1}{2}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V})^2t^2 + \frac{1}{6}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V})^3t^3 + \ldots\right]\mathbf{W}^{\mathrm{T}}\mathbf{g}_0 \end{aligned}$$

We choose the projection matrices $\mathbf{V}$, $\mathbf{W}^{\mathrm{T}}$ to ensure that the dynamics of the reduced $\widetilde{p}_t$ match as closely as possible those of the unreduced p_t. Following insights from the model reduction literature, we take this to mean that Taylor series expansions of p_t and $\widetilde{p}_t$ around $t = 0$ share the first k expansion coefficients. As argued before, the dynamics of the system – and so these expansion coefficients – do not depend on the projection matrices $\mathbf{V}$, $\mathbf{W}^{\mathrm{T}}$ themselves, but only on the subspaces associated with them.[66] It is in this sense that we can first focus on general $\mathbf{V}$, $\mathbf{W}^{\mathrm{T}}$, and then simply conclude that all results will extend to semi-orthogonal matrices $\mathbf{X}$ that span the same subspace of $\mathbb{R}^N$. To match the first k expansion coefficients, it is useful to consider what is known as the order-k observability matrix $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$:

$$\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg}) := \begin{bmatrix} \mathbf{b}_{pg} \\ \mathbf{b}_{pg}\mathbf{C}_{gg} \\ \mathbf{b}_{pg}(\mathbf{C}_{gg})^2 \\ \vdots \\ \mathbf{b}_{pg}(\mathbf{C}_{gg})^{k-1} \end{bmatrix}$$

[65] Formally, $\mathbf{\Pi} := \mathbf{V}\mathbf{W}^{\mathrm{T}}$ is a projection, and we have that $\mathbf{\Pi}\mathbf{g}_t = \mathbf{V}\gamma_t$.

[66] For a detailed discussion of this, see Amsallem and Farhat (2011, Lecture 7). The intuition is that, for the dynamics of a reduced system, only the *space* on which we project the large-dimensional state variable matters. A sketch of the formal argument goes as follows. $\mathbf{V}$ and $\mathbf{X}$ are bases of the same space, so there exists an invertible matrix $\mathbf{Z}$ such that $\mathbf{V}\mathbf{Z} = \mathbf{X}$, so $\mathbf{Z}^{-1} = \mathbf{X}^{\mathrm{T}}\mathbf{V}$ and $\mathbf{Z} = (\mathbf{X}^{\mathrm{T}}\mathbf{V})^{-1}$. Similarly, there exists an invertible matrix $\widetilde{\mathbf{Z}}$ such that $\widetilde{\mathbf{Z}}\mathbf{W}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}$, so $\widetilde{\mathbf{Z}}^{-1} = \mathbf{W}^{\mathrm{T}}\mathbf{X}$ and $\widetilde{\mathbf{Z}} = (\mathbf{W}^{\mathrm{T}}\mathbf{X})^{-1}$. But $\mathbf{W}^{\mathrm{T}}\mathbf{X} = \mathbf{W}^{\mathrm{T}}\mathbf{V}\mathbf{Z} = \mathbf{Z}$, so $\widetilde{\mathbf{Z}} = \mathbf{Z}^{-1}$. Then $\mathbf{V}\mathbf{W}^{\mathrm{T}} = \mathbf{X}\mathbf{Z}^{-1}\mathbf{W}^{\mathrm{T}} = \mathbf{X}\widetilde{\mathbf{Z}}\mathbf{W}^{\mathrm{T}} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ and the projections are identical.

We propose to consider the pair $\mathbf{V}, \mathbf{W}^{\mathrm{T}}$ with $\mathbf{W}^{\mathrm{T}} = \mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$ and $\mathbf{V}$ chosen such that $\mathbf{W}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. To see that this works, let us consider each term separately in the Taylor series expansions derived above. In all of the following, $\mathbf{e}_i$ denotes the ith standard unit vector and $\mathbf{W}_i^{\mathrm{T}}$ denotes the ith submatrix of $\mathbf{W}^{\mathrm{T}}$ (corresponding to $\mathbf{b}_{pg}(\mathbf{C}_{gg})^{i-1}$). First of all we have

$$\begin{aligned} \mathbf{b}_{pg}\mathbf{V}\mathbf{W}^{\mathrm{T}} &= \mathbf{W}_1^{\mathrm{T}}\mathbf{V}\mathbf{W}^{\mathrm{T}} \\ &= \mathbf{e}_1\mathbf{W}^{\mathrm{T}} = \mathbf{W}_1^{\mathrm{T}} = \mathbf{b}_{pg} \end{aligned}$$

where we have used the fact that, by construction, $\mathbf{W}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. Next we have

$$\begin{aligned} \mathbf{b}_{pg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\mathbf{W}^{\mathrm{T}} &= \mathbf{W}_1^{\mathrm{T}}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\mathbf{W}^{\mathrm{T}} \\ &= \mathbf{W}_2^{\mathrm{T}}\mathbf{V}\mathbf{W}^{\mathrm{T}} = \mathbf{e}_2\mathbf{W}^{\mathrm{T}} = \mathbf{W}_2^{\mathrm{T}} = \mathbf{b}_{pg}\mathbf{C}_{gg} \end{aligned}$$

where again we have used that $\mathbf{W}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, together with the definition of $\mathbf{W}^{\mathrm{T}}$. All higher-order terms then follow analogously. Putting things together in the notation of the main text, we see that picking $\mathbf{X}^{\mathrm{T}}$ to be a semi-orthogonal basis of the space spanned by $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$ is sufficient to ensure that the Taylor series expansion coefficients are matched.

A.3.2 Stochastic Model: Proof of Proposition 1

Solving out prices and the decision rules for the controls v_t, we get the system

$$\begin{aligned} \dot{\mathbf{g}}_t &= (\underbrace{\mathbf{B}_{gg} + \mathbf{B}_{gp}\mathbf{b}_{pg} + \mathbf{B}_{gv}\mathbf{D}_{vg}}_{\mathbf{C}_{gg}})\mathbf{g}_t + (\underbrace{\mathbf{B}_{gv}\mathbf{D}_{vZ}}_{\mathbf{C}_{gZ}})Z_t \\ p_t &= \mathbf{b}_{pg}\mathbf{g}_t + \mathbf{b}_{pZ}Z_t \end{aligned}$$

The dynamics of this stochastic system are characterized by the impulse response function

$$h(t) = \mathbf{b}_{pg}e^{\mathbf{C}_{gg}t}\mathbf{C}_{gZ} + \delta(t)\mathbf{b}_{pZ}$$

where $\delta(t)$ is the Dirac delta function. This impulse response function induces the following dynamic behavior:

$$p_t = \underbrace{\mathbf{b}_{pg}e^{\mathbf{C}_{gg}t}\mathbf{g}_0}_{\text{det. part}} + \underbrace{\int_0^t h(t-s)Z_s ds}_{\text{stoch. part}}$$

As before, we consider the projection $\gamma_t = \mathbf{W}^{\mathrm{T}}\mathbf{g}_t$ and (up to projection error) $\mathbf{g}_t = \mathbf{V}\gamma_t$. This gives

$$\begin{aligned}\dot{\gamma}_t &= \mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\gamma_t + \mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ}Z_t \\ \widetilde{p}_t &= \mathbf{b}_{pg}\mathbf{V}\gamma_t + \mathbf{b}_{pZ}Z_t\end{aligned}$$

This model induces the impulse response function

$$\widetilde{h}(t) = \mathbf{b}_{pg}\mathbf{V}e^{(\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V})t}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} + \delta(t)\mathbf{b}_{pZ}$$

and so the dynamics

$$\widetilde{p}_t = \mathbf{b}_{pg}\mathbf{V}e^{(\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V})t}\mathbf{W}^{\mathrm{T}}\mathbf{g}_0 + \int_0^t \widetilde{h}(t-s)Z_s ds$$

We now proceed exactly as before and consider the order-k observability matrix $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$:

$$\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg}) := \begin{bmatrix} \mathbf{b}_{pg} \\ \mathbf{b}_{pg}\mathbf{C}_{gg} \\ \mathbf{b}_{pg}(\mathbf{C}_{gg})^2 \\ \vdots \\ \mathbf{b}_{pg}(\mathbf{C}_{gg})^{k-1} \end{bmatrix}$$

We again set $\mathbf{W}^{\mathrm{T}}$ and $\mathbf{V}$ such that $\mathbf{W}^{\mathrm{T}} = \mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$ and $\mathbf{W}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. Showing that all terms in the deterministic part are matched is exactly analogous to the argument given above. For the stochastic part, we also do not need to change much. The impact impulse response $\mathbf{b}_{pZ}$ is matched irrespective of the choice of projection matrix. Next we have

$$\begin{aligned}\mathbf{b}_{pg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} &= \mathbf{W}_1^{\mathrm{T}}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} \\ &= \mathbf{e}_1\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} = \mathbf{W}_1^{\mathrm{T}}\mathbf{C}_{gZ} = \mathbf{b}_{pg}\mathbf{C}_{gZ}\end{aligned}$$

As before we exploit the definition of $\mathbf{W}^{\mathrm{T}}$ as well as the fact that $\mathbf{W}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. And finally

$$\begin{aligned}\mathbf{b}_{pg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} &= \mathbf{W}_1^{\mathrm{T}}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} \\ &= \mathbf{W}_2^{\mathrm{T}}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} = \mathbf{e}_2\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ} = \mathbf{W}_2^{\mathrm{T}}\mathbf{C}_{gZ} = \mathbf{b}_{pg}\mathbf{C}_{gg}\mathbf{C}_{gZ}\end{aligned}$$

again exactly analogous to the derivation for the deterministic part above. We are thus matching both the deterministic and the stochastic part of the dynamics up to order k in

a Taylor series expansion around time $t = 0$. Finally returning to the notation of the main body of the text, we see that letting $\mathbf{X}^{\mathrm{T}}$ be a semi-orthogonal basis of the space spanned by $\mathcal{O}(\mathbf{b}_{pg}, \mathbf{C}_{gg})$ is again sufficient for the impulse response matching.

A.3.3 Discrete Time

As we have seen, in continuous time, our model reduction procedure ensures that the coefficients of a Taylor series expansion around $t = 0$ are matched. In discrete time, this procedure guarantees that we match the first k periods of the impulse response functions. The stochastic discrete-time model is

$$\begin{aligned} \mathbf{g}_t &= \mathbf{C}_{gg}\mathbf{g}_{t-1} + \mathbf{C}_{gZ}Z_t \\ p_t &= \mathbf{b}_{pg}\mathbf{g}_{t-1} + \mathbf{b}_{pZ}Z_t \end{aligned}$$

The impulse responses of this system are $\mathbf{b}_{pZ}$ on impact and $\mathbf{b}_{pg}\mathbf{C}_{gg}^{h-1}\mathbf{C}_{gZ}$ for horizons $h = 1, 2, \ldots$. As before, we consider the reduced system

$$\begin{aligned} \gamma_t &= \mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\gamma_{t-1} + \mathbf{W}^{\mathrm{T}}\mathbf{C}_{gZ}Z_t \\ p_t &= \mathbf{b}_{pg}\mathbf{V}\gamma_{t-1} + \mathbf{b}_{pZ}Z_t \end{aligned}$$

Equality of the induced impulse responses for impact $h = 0$ is immediate. For all higher horizons, we proceed exactly as before and show that

$$\mathbf{b}_{pg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gz} = \mathbf{b}_{pg}\mathbf{C}_{gz}$$

as well as

$$\mathbf{b}_{pg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gg}\mathbf{V}\mathbf{W}^{\mathrm{T}}\mathbf{C}_{gz} = \mathbf{b}_{pg}\mathbf{C}_{gg}\mathbf{C}_{gz}$$

A.4 Detrending the Nonstationary Model

Many equilibrium objects in the version of the model described in Section 5 are nonstationary. In this appendix, we develop a normalized version of the equilibrium involving only stationary objects. In addition to the production side of the model described in the main text, we make three modifications to the two-asset model in the presence of nonstationary shocks. First, the borrowing constraint for liquid assets is $b > \underline{b}Q_t$, where Q_t is the level of aggregate productivity. Second, the transaction cost for accessing the illiquid account is now $\chi(d, a)Q_t$. Third, the lump-sum transfer from the government is now TQ_t.

The equilibrium of this model can be equivalently represented by a set of normalized objects $\hat{v}_t(\hat{a},\hat{b},z)$, $g_t(\hat{a},\hat{b},z)$, $\hat{K}_t$, r_t^a, $\hat{w}_t$, r_t^b, and Z_t such that

1. *Transformed HJB*: $\hat{v}_t(\hat{a},\hat{b},z)$ solves

$$\begin{aligned}(\rho+\zeta-(1-\theta)Z_t)\hat{v}_t(\hat{a},\hat{b},z) = &\max_{\hat{c},\hat{d}} \frac{\hat{c}^{1-\theta}}{1-\theta} \\ &+ \partial_{\hat{b}}\hat{v}_t(\hat{a},\hat{b},z)(T+(1-\tau)\hat{w}_t e^z + r_t^b(\hat{b})\hat{b} - \chi(\hat{d},\hat{a}) - \hat{c} - \hat{d} - \hat{b}Z_t) \\ &+ \partial_{\hat{a}}\hat{v}_t(\hat{a},\hat{b},z)(r_t^a\hat{a} + \hat{d} - \hat{a}Z_t) + \sum_{z'}\lambda_{zz'}(\hat{v}_t(\hat{a},\hat{b},z') - \hat{v}_t(\hat{a},\hat{b},z)) + \frac{1}{\mathrm{d}t}\mathbb{E}_t[\mathrm{d}\hat{v}_t(\hat{a},\hat{b},z)].\end{aligned} \tag{45}$$

   The fact that TFP growth is permanent changes the effective discount factor in the households' HJB equation.

2. *Transformed KFE*: $g_t(\hat{a},\hat{b},z)$ evolves according to

$$\begin{aligned}\frac{\mathrm{d}g_t(\hat{a},\hat{b},z)}{\mathrm{d}t} =& -\partial_{\hat{a}}\left(s_t^a(\hat{a},\hat{b},z)g_t(\hat{a},\hat{b},z)\right) - \partial_{\hat{b}}\left(s_t^b(\hat{a},\hat{b},z)g_t(\hat{a},\hat{b},z)\right) \\ &- \sum_{z'}\lambda_{zz'}g_t(\hat{a},\hat{b},z) + \sum_{z'}\lambda_{z'z}g_t(\hat{a},\hat{b},z) \\ &- \zeta g_t(\hat{a},\hat{b},z) + \zeta\delta(\hat{a})\delta(\hat{b})g^*(z), \text{ where} \\ s_t^b(\hat{a},\hat{b},z) =& T + (1-\tau)\hat{w}_t e^z + r_t^b(\hat{b})\hat{b} - \chi(\hat{d},\hat{a}) - \hat{c} - \hat{d} - \hat{b}Z_t \text{ and} \\ s_t^a(\hat{a},\hat{b},z) =& r_t^a\hat{a} + \hat{d} - \hat{a}Z_t.\end{aligned} \tag{46}$$

   Permanent TFP shocks change the effective depreciation rate of assets.

3. *Transformed firm conditions*: r_t^a, $\hat{w}_t$, and Z_t satisfy

$$\begin{aligned}r_t^a =& \alpha\hat{K}_t^{\alpha-1}\overline{L}^{1-\alpha} - \delta \\ \hat{w}_t =& (1-\alpha)\hat{K}_t^{\alpha}\overline{L}^{-\alpha} \\ dZ_t =& -\nu Z_t dt + \sigma dW_t.\end{aligned}$$

4. *Transformed asset market clearing conditions*

$$\begin{aligned}B^* &= \int \hat{b}g_t(\hat{a},\hat{b},z)\mathrm{d}\hat{b}\mathrm{d}\hat{a}\mathrm{d}z \\ \hat{K}_t &= \int \hat{a}g_t(\hat{a},\hat{b},z)\mathrm{d}\hat{b}\mathrm{d}\hat{a}\mathrm{d}z\end{aligned}$$

To derive this normalized equilibrium we detrend the model's original equilibrium objects by aggregate productivity Q_t. Almost all variables in the model naturally scale with the level of productivity Q_t; for any such variable x_t, let $\hat{x}_t = \frac{x_t}{Q_t}$ denote its detrended version. The one exception to this scheme is the households' value function $v_t(a, b, z)$, which scales with $Q_t^{1-\theta}$.

HJB Equation First define an intermediate detrended value function $\widetilde{v}_t(a, b, z) = \frac{v_t(a,b,z)}{Q_t^{1-\theta}}$. Divide both sides of the HJB (35) by $Q_t^{1-\theta}$ to get

$$\begin{aligned}(\rho + \zeta)\widetilde{v}_t(a, b, z) =& \max_{c,d} \frac{\hat{c}^{1-\theta}}{1-\theta} + \partial_b \widetilde{v}_t(a, b, z)\left(TQ_t + (1-\tau)w_t e^z + r_t^b(b)b - \chi(d, a)Q_t - c - d\right) \\ &+ \partial_a \widetilde{v}_t(a, b, z)(r_t^a a + d) + \sum_{z'} \lambda_{zz'}(\widetilde{v}_t(a, b, z') - \widetilde{v}_t(a, b, z)) \\ &+ \frac{1}{Q_t^{1-\theta}} \times \frac{1}{dt}\mathbb{E}_t[dv_t(a, b, z)]. \end{aligned} \tag{47}$$

Next, to replace the $\frac{1}{dt}\mathbb{E}_t[dv_t(a, b, z)]$ term, note that by the chain rule

$$\frac{d}{dt}\widetilde{v}_t(a, b, z) = \frac{\frac{d}{dt}v_t(a, b, z)}{Q_t^{1-\theta}} + (\theta - 1)\frac{d \log Q_t}{dt}\widetilde{v}_t(a, b, z),$$

which implies that

$$\frac{1}{Q_t^{1-\theta}} \times \frac{1}{dt}\mathbb{E}_t[dv_t(a, b, z)] = \frac{1}{dt}\mathbb{E}_t[d\hat{v}_t(a, b, z)] + (1-\theta)\frac{d \log Q_t}{dt}\hat{v}_t(a, b, z).$$

Plug this back into (47) and rearrange to get

$$\begin{aligned}(\rho + \zeta + (\theta - 1)\frac{d \log Q_t}{dt})\widetilde{v}_t(a, b, z) =& \max_{c,d} \frac{\hat{c}^{1-\theta}}{1-\theta} + \partial_b \widetilde{v}_t(a, b, z)(TQ_t + (1-\tau)w_t e^z + r_t^b(b)b - \chi(d, a)Q_t \\ &- c - d) + \partial_a \widetilde{v}_t(a, b, z)(r_t^a a + d) + \sum_{z'} \lambda_{zz'}(\widetilde{v}_t(a, b, z') - \widetilde{v}_t(a, b, z)) \\ &+ \frac{1}{dt}\mathbb{E}_t[d\widetilde{v}_t(a, b, z)]. \end{aligned} \tag{48}$$

The formulation in (48) is still not stationary because there are permanent changes in the state variables a and b, the wage w_t, and transaction cost on the right hand side. To address this we characterize the value function in terms of $\hat{a}$ and $\hat{b}$, rather than a and b themselves. Define the final detrended value function $\hat{v}_t(\hat{a}, \hat{b}, z)$ as $\hat{v}_t(\hat{a}, \hat{b}, z) = \widetilde{v}_t(a, b, z)$.

We guess that this function $\hat{v}_t(\hat{a}, \hat{b}, z)$ does not depend on the nonstationary variable Q_t and now verify that indeed it does not. Note that

$$\begin{aligned}
\partial_b \widetilde{v}_t(a,b,z) =& \partial_b \hat{v}_t\left(\frac{a}{Q_t}, \frac{b}{Q_t}, z\right) = \frac{1}{Q_t}\partial_{\hat{b}}\hat{v}_t(\hat{a},\hat{b},z), \\
\partial_a \widetilde{v}_t(a,b,z) =& \partial_a \hat{v}_t\left(\frac{a}{Q_t}, \frac{b}{Q_t}, z\right) = \frac{1}{Q_t}\partial_{\hat{a}}\hat{v}_t(\hat{a},\hat{b},z), \\
\frac{1}{dt}\mathbb{E}_t[d\widetilde{v}_t(a,b,z)] =& \frac{1}{dt}\mathbb{E}_t\left[d\hat{v}_t\left(\frac{a}{Q_t}, \frac{b}{Q_t}, z\right)\right] \\
=& \frac{1}{dt}\mathbb{E}_t\left[d\hat{v}_t\left(\hat{a},\hat{b},z\right)\right] - \partial_{\hat{a}}\hat{v}_t(\hat{a},\hat{b},z)\hat{a}\frac{d\log Q_t}{dt} - \partial_{\hat{b}}\hat{v}_t(\hat{a},\hat{b},z)\hat{b}\frac{d\log Q_t}{dt}.
\end{aligned}$$

These equations then imply

$$\begin{aligned}
&\partial_b \widetilde{v}_t(a,b,z)(TQ_t + (1-\tau)w_t e^z + r_t^b(b)b - \chi(d,a)Q_t - c - d) \\
&\quad = \partial_{\hat{b}}\hat{v}_t(\hat{a},\hat{b},z)(T + (1-\tau)\hat{w}_t e^z + r_t^{\hat{b}}(\hat{b})\hat{b} - \chi(\hat{d},\hat{a}) - \hat{c} - \hat{d})
\end{aligned}$$

and that

$$\partial_a \widetilde{v}_t(a,b,z)(r_t^a a + d) = \partial_{\hat{a}}\hat{v}_t(\hat{a},\hat{b},z)(r_t^a \hat{a} + \hat{d}).$$

Putting all these results together, and using the definition $\frac{d\log Q_t}{dt} = Z_t$, we get the final detrended HJB equation (45).

KFE The cross-sectional distribution of households over $\hat{a}, \hat{b}, z$ is stationary. We will directly construct the KFE for the distribution over this space. Analogously to (36), this is given by

$$\begin{aligned}
\frac{dg_t(\hat{a},\hat{b},z)}{dt} =& -\partial_{\hat{a}}\left(\dot{\hat{a}}_t(a,b,z)g_t(\hat{a},\hat{b},z)\right) - \partial_{\hat{b}}\left(\dot{\hat{b}}_t(\hat{a},b,z)g_t(\hat{a},\hat{b},z)\right) \\
&- \sum_{z'}\lambda_{zz'}g_t(\hat{a},\hat{b},z) + \sum_{z'}\lambda_{z'z}g_t(\hat{a},\hat{b},z) \\
&- \zeta g_t(\hat{a},\hat{b},z) + \zeta\delta(\hat{a})\delta(\hat{b})g^*(z).
\end{aligned}$$

The modified HJB (45) gives the evolution $\frac{\dot{a}}{Q_t} = r_t^a\hat{a} + \hat{d}$. Note that by the product rule,

$$\dot{\hat{a}}_t = \frac{\dot{a}_t}{Q_t} - \frac{d\log Q_t}{dt}\hat{a}_t,$$

so that $\dot{\hat{a}}_t = r_t^a \hat{a} + \hat{d} - \frac{d \log Q_t}{dt} \hat{a}$. Using this result, and the analogous one for $\dot{\hat{b}}_t$, we get the detrended KFE (46).

Other Equilibrium Conditions Detrending the remaining equilibrium conditions is simple:

$$\begin{aligned} r_t^a =& \alpha \hat{K}_t^{\alpha-1} \overline{L}^{1-\alpha} - \delta \\ \hat{w}_t =& (1-\alpha) \hat{K}_t^{\alpha} \overline{L}^{-\alpha} \\ dZ_t =& -\nu Z_t dt + \sigma dW_t. \end{aligned}$$

A.5 Representative Agent and Spender-Saver Models

Representative Agent The representative agent model is identical to the RBC model described in Appendix A.2.

Spender-Saver The spender-saver model extends the household side of the representative agent model above to two types of households. First, there is a fraction λ of hand-to-mouth households who simply consume their income each period. Second, the remaining fraction $1-\lambda$ of households make an optimal consumption-savings decision like in the representative agent model.