Computing Longitudinal Moments for Heterogeneous Agent Models*

Sergio Ocampo†
University of Western Ontario

Baxter Robinson‡
University of Western Ontario

September 13, 2022

Abstract

Computing population moments for heterogeneous agent models is a necessary step for their estimation and evaluation. Computation based on Monte Carlo methods is usually time- and resource-consuming because it involves simulating a large sample of agents and potentially tracking them over time. We argue in favor of an alternative method for computing both cross-sectional and longitudinal moments that exploits the endogenous Markov transition function that defines the stationary distribution of agents in the model. The method relies on following the distribution of populations of interest by iterating forward the Markov transition function rather than focusing on a simulated sample of agents. Approximations of this function are readily available from standard solution methods of dynamic programming problems. The method provides precise estimates of moments like top-wealth shares, auto-correlations, transition rates, or age-profiles, at lower time- and resource-costs compared to Monte Carlo based methods.

JEL: C6, E2

Keywords: Computational Methods, Heterogeneous Agents, Simulation.

*We thank Emmanuel Murray Leclair and Javier Hernandez for their excellent work as research assistants on this project. We also thank Rory McGee, Emily Moschini, and Thomas Phelan for helpful comments. Replication files for this paper are available at https://github.com/ocamp020/Histogram_Iteration.

†Email: socampod@uwo.ca; Web: https://sites.google.com/site/sergiocampod/.

‡Email: brobin63@uwo.ca; Web: https://sites.google.com/view/baxter-robinson/.

Computing cross-sectional and longitudinal moments is integral to both the estimation and use of heterogeneous agents models that are common in the study of a wide variety of economic phenomena (e.g., Heathcote, Storesletten, and Violante, 2009; De Nardi, French, and Jones, 2016; De Nardi and Fella, 2017). However, calculating these moments frequently poses computational challenges that come from the repeated simulation of the models. These challenges limit how researchers can use these models and the features they are able to include, even as computational power continues to improve.

One key challenge is the cost of calculating longitudinal moments that require following individuals over time (e.g., mobility rates across occupations or the wealth distribution, income persistence, or inter-generational correlations). Standard Monte Carlo based methods used to compute these moments rely on a simulated panel of agents that can fail to be representative of sub-populations like the "very rich" that determine the degree of wealth inequality. So, in order to obtain accurate moments, these panels must be simulated with a large number of agents, often millions of them, which is computationally costly.

We argue in favor of computing longitudinal moments by directly following the distribution of any sub-population over time. We do this by iterating forward the Markov kernel that characterizes how agents transition between different states. This method comes at minimal cost because the Markov kernel is already approximated as part of most solution methods (e.g., Young, 2010; Heer and Maußner, 2005, Ch. 7) and avoids the impreciseness and inefficiencies of Monte Carlo simulation.

We take as given the model's solution in the form of policy functions for agents that, together with the stochastic processes of exogenous states, imply the evolution for the distribution of agents in the economy. This evolution is captured by a Markov kernel, $T\left(s^{'}|s\right)$, that maps the transition of a mass of agents from a current state s into a future state s' in the state space $\mathcal{S}$. The stationary distribution, λ, is the solution to

$$\lambda\left(s^{'}\right)=\int_{s\in\mathcal{S}}T\left(s^{'}|s\right)\lambda\left(s\right)ds, \tag{1}$$

We describe how to use λ and T to directly compute cross-sectional and longitudinal moments, rather than using a simulated panel of agents.

Cross-sectional moments. These moments involve taking expectations over some variable of interest, $x(s)$, for some sub-population characterized by states $s \in S \subseteq \mathcal{S}$,

$$E\left[x|s \in S\right] = \int_{s \in S} x\left(s\right)\lambda_S\left(s\right)ds, \tag{2}$$

where $\lambda_S \equiv {}^{\mathbb{I}_{s\in S}\lambda(s)}/_{\int \mathbb{I}_{s\in S}\lambda(s)ds}$ is the marginal distribution of the sub-population in S, with $\mathbb{I}_{s\in S}$ an indicator variable for whether or not $s \in S$. Equation (2) applies to a wide range of moments. For example, the skewness or kurtosis of the endogenous wealth distribution for the whole population (when $S = \mathcal{S}$) or a subgroup (say among the top income earners).[1] These moments can be computed immediately from the solution of the model's stationary distribution (λ), either by approximating the integral (Judd, 1998, Ch. 7) or by calculating the moment from a discrete approximation of the distribution (Young, 2010).

Longitudinal moments. Many other moments require knowing either the collective outcomes of a group of agents over time (e.g., for computing transition rates across occupations) or the outcomes of individual agents (e.g., for computing the auto-correlation of their consumption).[2] Calculating these moments is hard because of the stochastic nature of the time-paths of individuals. However, we demonstrate that it is both possible and computationally efficient to extend the approach used in computing cross-sectional moments from the model's stationary distribution to longitudinal moments by focusing on the transition of the distribution of agents, taking into account all possible paths an individual can take rather than the realized paths from a Monte Carlo simulation.

[1] Equation (2) can also be used to define percentiles or other descriptors of the distribution. It also applies to moments that depend on both endogenous and exogenous states, for example by making x total income.

[2] Moments that require collective outcomes include mobility rates across the income or wealth distribution, or inter-generational mobility in life cycle models. Moments that require individual outcomes include the distribution of growth rates of income or wealth for individual agents, or the distribution of lifetime earnings.

Consider an outcome of interest $x\left(s, s^{'}\right)$ that depends on some initial and final state of an agent. This outcome could be any function of the agent's initial or final states. For example, it could indicate whether the agent satisfies some condition in the future (e.g., being a top earner or having a certain occupation), or give the difference between the agent's current and future income. The expectation of interest depends on whether we focus on the behavior of the group of agents (as in transition rates) or of individual agents (as in the auto-correlation of income). In the first case, we must follow the group (S) as a whole and compute

$$E\left[x|s\in S\right]=\int_{s\in S}\int_{s^{'}\in\mathcal{S}}x\left(s,s^{'}\right)\lambda_{S}^{'}\left(s^{'}\right)ds^{'}\,\lambda_{S}\left(s\right)ds, \tag{3}$$

where $\lambda_{S}^{'}$ is the future distribution of agents conditional on the initial distribution λ_S. In the second case, we must follow the possible paths of each individual and compute

$$E\left[x|s\in S\right]=\int_{s\in S}\int_{s^{'}\in\mathcal{S}}x\left(s,s^{'}\right)\lambda_{\{s\}}^{'}\left(s^{'}\right)ds^{'}\,\lambda_{S}\left(s\right)ds, \tag{4}$$

where $\lambda_{\{s\}}^{'}$ is the future distribution of the mass of agents that starts in state $s\in S$.

The difficulty in evaluating the expectations in (3) and (4) resides in obtaining $\lambda_{S}^{'}$ and $\lambda_{\{s\}}^{'}$ because doing so requires accounting for the individual paths between the initial and final periods. We directly compute $\lambda_{S}^{'}$ and $\lambda_{\{s\}}^{'}$ by iterating forward the initial distribution of agents using the Markov kernel T,

$$\lambda_{S}^{'}\left(s^{'}\right)=\int_{s\in\mathcal{S}}T\left(s^{'}|s\right)\lambda_{S}\left(s\right)ds;\quad \lambda_{\{s\}}^{'}\left(s^{'}\right)=\int_{s\in\mathcal{S}}T\left(s^{'}|s\right)\delta_{\{s\}}\left(s\right)ds, \tag{5}$$

where $\delta_{\{s\}}$ is the (degenerate) distribution concentrated in state s. The cost of the iteration in (5) is similar those involved in finding the stationary distribution (see equation 1).[3] Moreover, the number of iterations required to compute $\lambda_{S}^{'}$ and $\lambda_{\{s\}}^{'}$ are finite (and known). Once $\lambda_{S}^{'}$ and $\lambda_{\{s\}}^{'}$ are obtained, the moments can be computed from (3) and (4).

[3] The integral in the computation of $\lambda_{\{s\}}^{'}$ is of course superfluous. Nevertheless, integration becomes necessary when iterating more than one period, as $\delta_{\{s\}}$ generically distributes mass across the state space $\mathcal{S}$.

In practice, the stationary distribution of heterogeneous agents models is typically computed using a discrete approximation of the Markov kernel $\hat{T}$ that operates over a discrete state space and induces a discrete distribution $\hat{\lambda}$ in the form of a histogram (see, Young, 2010; Tan, 2020). In this case, the formulas in (1)-(4) replace integrals for sums over the discretized state space. Accordingly, we dub our method the histogram iteration method. We show that it provides fast and precise estimates of moments of interest without involving the computation of new objects, relative to those involved in the model's solution.

We apply our method to two partial equilibrium versions of the standard heterogeneous agent model based on Aiyagari (1994), one with infinitely lived agents and one with overlapping generations. We approximate the stationary distribution and its associated Markov kernel following Young (2010). We calculate moments characterizing the right tail of the wealth distribution and the persistence of consumption and wealth in the infinitely lived agents model. In the overlapping-generations model, we calculate the age-profile of wealth and five- and fifteen-year auto-correlation of wealth. In this way, we use our method to calculate cross-sectional and longitudinal moments for both the entire population of agents and specific sub-populations, like the wealthiest and poorest agents. We compare the results with those from a Monte Carlo simulation, which is common in the literature (e.g., Judd, 1998, Ch. 8).

We find that our histogram iteration method is at least as precise as using large simulated panels while significantly reducing computational time. Time savings come from the difference between computing the histogram relative to simulating a large enough panel of agents. In most cases, these differences more than compensate for the time it takes to compute longitudinal moments by iterating on the histogram, which is considerably higher than the time it takes to compute them from a simulated panel. There are further gains when computing cross-sectional moments because no iteration or simulation is needed when using the histogram.

1 Baseline heterogeneous agent models

We illustrate our method in the context of the baseline Bewley-Hugget-Aiyagari-Imrohoroglu model. The economy is populated by a continuum of agents indexed by $i \in [0,1]$ that differ on their age (h), their labor productivity (ε), their rate of return (ζ), and their endogenous asset holdings (a). Labor productivity and rates of return follow discrete Markov processes with transition matrices P^{ε} and P^{ζ}. Agents are price takers. They receive income from the return on their savings ($r(\zeta)$) and from wages (w) paid for their supply of efficiency units of labor ($\ell(h,\varepsilon)$) that depend on their age and labor productivity.

The dynamic programming problem of an agent is

$$
\begin{aligned}
V_h(\varepsilon,\zeta,a) &= \max_{a',c}\ u(c) + \beta E\left[V_{h+1}\left(\varepsilon',\zeta',a'\right)|\varepsilon,\zeta\right] \\
&\text{s.t. } \ (1+r(\zeta))\,a + w\ell(h,\varepsilon) = c + a';\ a' \geq \underline{a}.
\end{aligned} \tag{6}
$$

The solution to (6) is a savings function ($a_h^{\star}$), such that $a_h^{\star}(\varepsilon,\zeta,a) \geq \underline{a}$ for all (ε,ζ,a) and

$$
V_h(\varepsilon,\zeta,a) = u\left((1+r(\zeta))\,a + w\ell(h,\varepsilon) - a_h^{\star}(\varepsilon,\zeta,a)\right) + \beta E\left[V_{h+1}\left(\varepsilon',\zeta',a_h^{\star}(\varepsilon,a)\right)|\varepsilon,\zeta\right]. \tag{7}
$$

We will focus on a stationary equilibrium with a time-invariant distribution of agents. $\mathcal{S}$ is the state space with typical element $s = (h,\varepsilon,\zeta,a)$. Given a birth/death process for agents, the transition function of labor productivity, and the savings functions, the stationary distribution is a solution to (1), where the Markov kernel $T\left(s'|s\right)$ is constructed using the policy functions and evolution of exogenous states.

We solve the model in partial equilibrium taking the wage rate, w, and the average return on savings, $\bar{r}$, as exogenous. We do this to focus on the computation of moments for any given solution of the agents' problem. Our results apply in a general equilibrium setting when computing the moments after finding the market clearing prices.

We solve for two versions of the model that differ in the birth/death process of agents. In both models, we adopt the following functional form for agents' utility:

$$u(c) = \frac{c^{1-\sigma} - 1}{1 - \sigma}. \tag{8}$$

We set σ equal to 2 which is in the range of values used in Aiyagari (1994). We take $\bar{r}$ to be 3.2 in line with historical values for the U.S. and we set w so that labour income in our model matches average labor income for the U.S. in 2019, which is \$53,624.[4] We set $\underline{a} = 0$, preventing borrowing. Below, we outline the differences between the two different versions of the model and their parametrization.

Infinitely lived heterogeneous agent model We consider a version of the model where agents are infinitely lived and their labor efficiency depends only on their labor productivity. In particular, $\ell(h, \varepsilon) = \exp(\varepsilon)$. We focus on the age-invariant solutions to (6) and (7), a value function $V(\varepsilon, \zeta, a)$ and a savings function $a^{\star}(\varepsilon, \zeta, a)$. Accordingly, we drop age from the state vector when referring to the infinitely lived agents model.

We assume that labor productivity follows a discrete Markov process with $n_{\varepsilon} = 11$. We obtain P^{ε} by discretizing an AR(1) process with persistence $\rho_{\varepsilon} = 0.963$ and innovation variance $\sigma_{\varepsilon}^2 = 0.162$ using Rouwenhorst (1995)'s method. The values of ρ_{ε} and σ_{ε} come from Storesletten, Telmer, and Yaron (2004).

We include heterogeneous returns on savings, a key ingredient for generating high levels of wealth inequality (Benhabib, Bisin, and Zhu, 2011; Stachurski and Toda, 2019), by setting an agent's returns to be $r(\zeta) = \bar{r}\exp(\zeta)$. The state ζ follows a discrete Markov process with $n_{\zeta} = 7$ states. We obtain P^{ζ} by discretizing an AR(1) process with persistence $\rho_{\zeta} = 0.70$ and innovation variance $\sigma_{\zeta}^2 = 1.3$ using Tauchen (1986)'s method.

[4] We construct this value from FRED Data (U.S. Bureau of Economic Analysis, 2022) as Total Wages and Salaries (BA06RC1A027NBEA) divided by the 12-month average of Civilian Labor Force Level (CLF16OV).

Overlapping generations heterogeneous agent model In the second version of the model, agents live for $H > 0$ periods and have a final value of $V_{H+1} = 0$. We add mortality risk so that an agent's conditional survival probability going into age h is ϕ_h. We set $\phi_{H+1} = 1$ so that agents die with certainty after age H. We set the survival probabilities following Bell and Miller (2002) projections for the U.S., with each model period corresponding to a single year. Agents are born at age 20 ($h = 1$) and can live to a maximum age of 100 ($H = 81$). Upon death, agents are replaced by a newborn who starts life with $a_1^\star = \$1,000$ of assets.

Efficiency units of labor are $\ell(h, \varepsilon) = \exp(\xi(h) + \varepsilon)$, where $\xi(h)$ is a quadratic polynomial that generates a 50 percent rise in average labor income from age 21 to its peak at age 51 as in Guvenen, Kambourov, Kuruscu, Ocampo, and Chen (2019).[5] We use the same process for labor productivity (ε) as in the infinitely lived agent model. Finally, we eliminate rate of return heterogeneity, so that all agents earn $r_i = \bar{r}$. Accordingly, we drop ζ from the state vector when referring to the overlapping generations model.

2 Solving the models

We solve for the policy functions in (6) using readily available solution methods that exploit the optimality conditions of the savings choice (i.e., Carroll, 2006). Having computed the policy functions, we approximate the Markov kernel, T, of the distribution of agents by discretizing it over assets on a grid $\vec{a}_{n_a}$ following Young (2010). The result is a transition matrix $\hat{T}$, whose elements $\hat{T}(s, s')$ gives the probability that an agent with current state s transitions to state s'. This probability depends on the birth/death process (for instance, agents of age $h = H$ transition to age $h = 1$ with certainty), the transition matrix of the labor productivity process (ε), the transition matrix of the return heterogeneity process (ζ), and the approximation of the transition of assets on the fixed grid $\vec{a}_{n_a}$.[6] Finally, we compute

[5] $\xi(h) = \left(60(h-1) - (h-1)^2\right)/1800$.

[6] An agent with state s transitions with certainty to having assets $a' = a^\star_{h(s)}(\varepsilon(s), \zeta(s), a(s)) \in [\vec{a}_j, \vec{a}_{j+1}]$, for some j. In the discrete approximation the agent transitions to either $\vec{a}_j$ with probability ${}^{\vec{a}_{j+1}-a'}/_{\vec{a}_{j+1}-\vec{a}_j}$ or $\vec{a}_{j+1}$ with probability ${}^{a'-\vec{a}_j}/_{\vec{a}_{j+1}-\vec{a}_j}$.

the stationary distribution of agents on the discrete grid by iterating over

$$\hat{\lambda}^{n+1}\left(s^{'}\right)=\sum_{s}\hat{T}\left(s,s^{'}\right)\hat{\lambda}^{n}\left(s\right), \tag{9}$$

for some initial $\hat{\lambda}^0$. The stationary distribution $\left(\hat{\lambda}\right)$ is the limit of $\hat{\lambda}^n$ as n grows large.

We use the approximated distribution $(\hat{\lambda})$ and Markov kernel $(\hat{T})$ to compute moments for both models in Section 3. The Markov kernel plays an important role in computing moments because it describes the evolution of states given any initial distribution. This is crucial for computing longitudinal moments where it is necessary to know the joint distribution of agents across time. We explore results with grids of different sizes in the following section. All grids are curved so that they are denser for low wealth values. In particular, the n^{th} node of an asset grid with N satisfies $\vec{a}_n = \underline{a} + (\overline{a} - \underline{a})\left({}^{n-1}/_{N-1}\right)^{\theta_a}$, where $\theta_a > 1$ measures the curvature. We use a curvature of $\theta_a = 3.5$ and solve for the policy functions on a grid with 250 nodes before approximating the Markov kernel and the stationary distribution.

3 Computing moments

Our objective is to compute cross-sectional and longitudinal moments for both the entire population and sub-population of interest, like agents at the top or bottom of the wealth distribution. Cross-sectional moments, like the wealth share of top 1 percent wealth holders can be readily computed by integrating with respect to the stationary distribution (λ) or its discrete approximation $\left(\hat{\lambda}\right)$. However, it is often necessary to follow individuals over time in order to compute longitudinal moments like transition rates across deciles of the wealth distribution. This is often achieved through costly Monte Carlo based simulations of a sample of individuals. Our histogram iteration method relies instead on tracking the distribution of the relevant group of individuals (the sub-population), following its evolution as described by the Markov kernel T. We now describe the method.

The histogram iteration method Consider a moment describing the expectation over some outcome in some future period $x\left(s',s\right)$ for a group of individuals satisfying some condition (the sub-population), say having a certain level of wealth or income. It is possible to determine a subset of the state space $S \subseteq \mathcal{S}$ such that any agent with state $s \in S$ belongs to the sub-population of interest. These moments take the form of the expectations in equations (3) and (4). The objective is to approximate the value of these expectations. We obtain the sub-population's initial distribution, $\hat{\lambda}_S$, from the stationary distribution $\hat{\lambda}$ by restricting its domain to S and normalizing. Tracking the distribution of the sub-population over time involves iterating over $\hat{\lambda}_S$ with the Markov kernel T as in (9).

When the moment requires tracking only the outcomes of the group in S, the expectation of interest is

$$E\left[x|s \in S\right] \approx \sum_{s}\sum_{s'} x\left(s',s\right)\hat{\lambda}'_S\left(s'\right)\hat{\lambda}_S\left(s\right) \tag{10}$$

and when the moment requires tracking the future outcomes of individuals, the expectation of interest is

$$E\left[x|s \in S\right] \approx \sum_{s}\sum_{s'} x\left(s',s\right)\hat{\lambda}'_{\{s\}}\left(s'\right)\hat{\lambda}_S\left(s\right) \tag{11}$$

In this case, $\hat{\lambda}'_{\{s\}}$ is the future distribution of agents that started in state $s \in S$. Below, we apply this method in the models described in Section 1.

3.1 Moments for the infinitely lived agents model

We now compute several moments for the infinitely lived agents model and compare the performance of the histogram iteration method relative to a traditional Monte Carlo simulation. We focus on moments characterizing the wealth distribution and the behaviour of consumption, which are the endogenous outcomes in our setting. In particular, we present results for the tail of the wealth distribution, top wealth shares, the persistence of consumption and wealth, and the ten-year transition rates across wealth deciles.

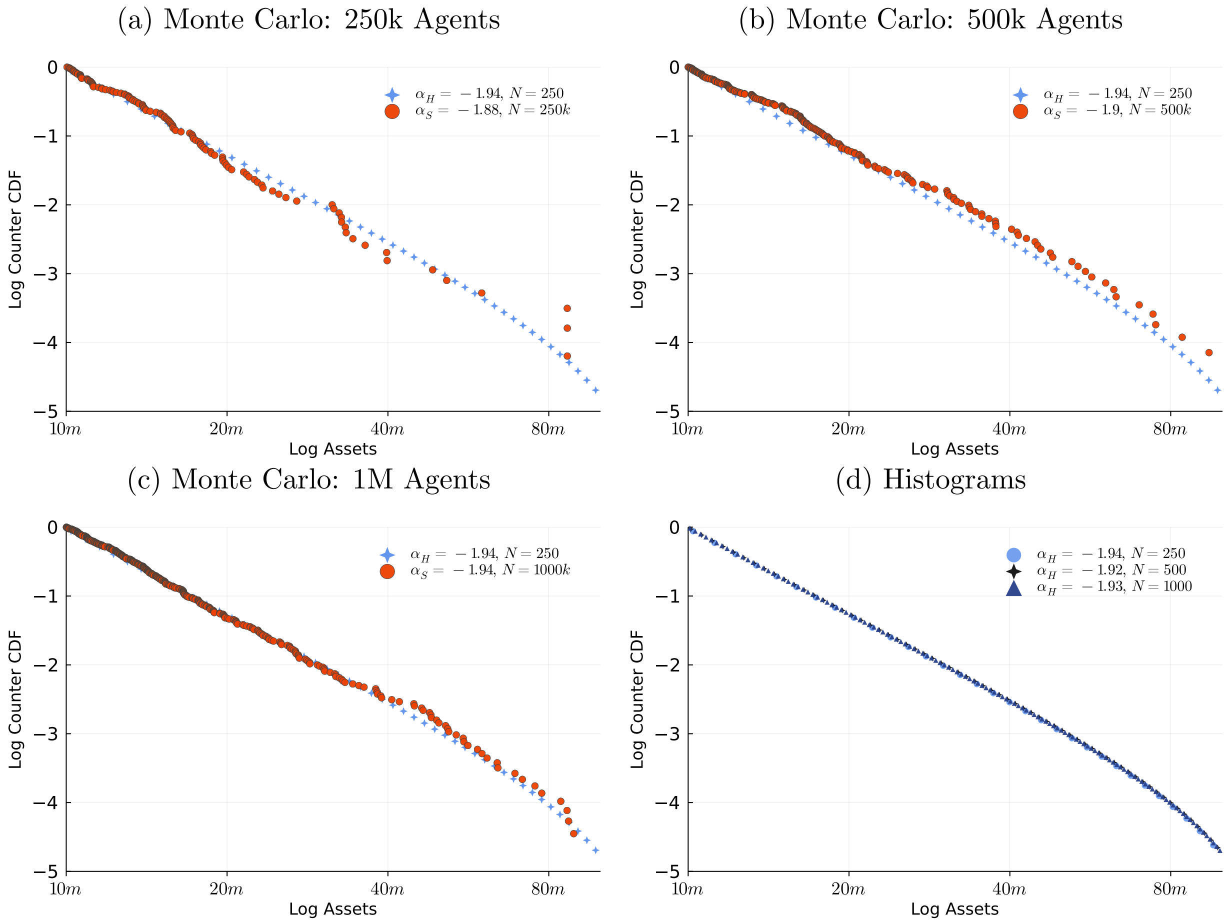
In terms of the accuracy, we find that both methods provide similar estimates for the moments, except for those regarding top-wealth holders: the shape of the tail of the wealth distribution and top wealth shares. The challenge for the Monte Carlo simulation method comes from the large number of agents needed in order to obtain a representative sample of top-wealth holders. The histogram iteration method provides more consistent values of these moments when varying the number of grid nodes in the approximation.

In terms of the computational cost of calculating moments, cross-sectional moments come almost from free after solving for the histogram or simulating a panel of agents. Longitudinal moments are significantly more expensive to calculate with the histogram iteration method than from a panel of simulated agents. This is because the histogram iteration method requires iterating forward for all initial states, as all the possible histories of agents are mapped for each initial condition. This makes the computation more expensive than computing the same moment using an already existing panel of agents containing realized histories of consumption and wealth. However, the time required to solve for the histogram is substantially less than the time required to simulate the Monte Carlo panel of agents. As a result, the total time it takes to calculate longitudinal moments is less when using the histogram iteration method than when using the Monte Carlo panel.[7]

Pareto Tail One characteristic of the cross-sectional distribution of wealth that is often difficult to capture in heterogeneous agent models is the behavior of its right tail and the level of wealth concentration. These statistics are crucial when studying inequality, particularly because of their implications for taxation. We report the right tail of the wealth distribution (above ten million dollars) and the corresponding Pareto coefficient for simulations with sample sizes between two hundred and fifty thousand and one million agents in Figure 1 and contrast them with the tail of the stationary distribution of wealth approximated with a

[7] All times are for a Mac Mini with an M1 processor running Julia v1.7.

Figure 1: Pareto Tail - Monte Carlo Simulation and Histogram Method



Notes: The figures plot the log counter CDF of the conditional distribution of wealth above \$10 million. Panels 1a to 1c approximate the CDF using samples of agents from a Monte Carlo simulation and differ in the number of agents being simulated. The blue diamonds correspond to the approximation of the counter CDF using the histogram method with 500 grid nodes. The final panel approximates the CDF using the histogram method with 250, 500, and 1000 grid nodes.

histogram with 500 grid points. We find that simulation-based results require a large number of agents to correctly represent the properties of the right tail of the wealth distribution, and that, by contrast, the histogram provides a more stable picture of the distribution at lower computational cost.[8]

The Monte Carlo simulation agrees in general with the shape of the tail, but has issues populating the top end, even with one million agents. This is apparent in the discrepancies between the tail indexes (α) across simulation samples and also in the wealth shares of the

[8]This is similar to Gouin-Bonenfant and Toda (2022, forthcoming), who propose replacing the grid at the right end of the distribution with an approximation of the continuous distribution using limit results.

Table 1: Cross-sectional and Longitudinal Moments: Infinitely Lived Agents

	Percentage Point Deviations from Reference Value						Ref. Value
	Monte Carlo: Sample Size			Histogram: Grid Size			
	250k	500k	1M	250	500	1000	
Top Wealth Shares							
Top 0.1%	0.12	0.14	−0.52	0.08	−0.04	−0.10	6.29
Top 1%	0.08	0.02	−0.76	0.12	−0.06	0.03	19.00
Pareto Coefficient	−0.04	−0.01	0.03	0.03	0.01	0.02	1.91
Auto-Correlations							
$\rho\left(c_t, c_{t+2}\right)$	0.24	0.00	0.05	0.13	0.01	0.04	82.52
$\rho\left(a_t, a_{t+2}\right)$	0.20	0.97	0.08	1.04	0.07	0.43	49.73
Transition Rates							
$\Pr\left(a_i' \in D_1 \vert a_i \in D_1\right)$	0.17	−0.06	0.05	0.59	0.12	0.31	50.04
$\Pr\left(a_i' \in D_2 \vert a_i \in D_1\right)$	−0.10	0.24	−0.06	−0.19	0.02	−0.03	34.16
Computational Time							
Simulation	689.3	1386.1	2744.3	—	—	—	—
Distribution $\hat{\lambda}$	—	—	—	478.9	881.4	1827.0	—
Top Inequality	0.01	0.02	0.04	1E-4	4E-4	2E-4	—
Auto-Correlation	0.05	0.08	0.18	9.81	21.47	54.76	—
Transition Rates	0.39	0.83	1.58	13.48	26.04	50.48	—

Notes: The table reports moments and computational time in seconds for the infinitely lived agents model. The first block computes the moments approximating the distribution with Monte Carlo simulation on three different samples of 250k, 500k, and 1M agents. The second block computes the moments approximating the stationary distribution with histograms on three different grids with 250, 500, and 1000 nodes. The reference value is obtained from a histogram on a grid with 5000 nodes.

richest agents as shown in Table 1 below. Figure 1d shows that the histogram provides more stable outcomes across grid sizes for both the shape of the distribution and the tail index.

The sensitivity of the right tail to the number of agents being simulated becomes an issue in models that aim to capture the extent of wealth inequality in the data. For instance, Guvenen et al. (2019) pose a model capable of reproducing the shape of the tail of the wealth distribution in the U.S., including the presence of multi-billionaires. In order to generate these very wealthy agents, they use a Monte Carlo simulation with twenty million agents.

Top Wealth Shares We compute the share of wealth owned by the top 1% and top 0.1% of individuals in our model and report them in Table 1. Just as with the shape of the right tail, these measures of top wealth concentration are difficult to measure with the

Monte Carlo simulation because a small number of “very rich” agents play a large role in determining the value of the moments. As a consequence, the top wealth shares are still varying even when the number of simulated agents is increased to one million. The time required to compute the moments is negligible next to the time required to either obtain the stationary distribution of the model or to simulate the agents.

Persistence of Consumption and Wealth We continue by computing the two-year auto-correlations of consumption and wealth, which are informative about the ability of individuals to insure themselves against temporary income fluctuations. These are longitudinal moments that require comparing the level of consumption and wealth for individuals across time. Both the histogram iteration method and Monte Carlo simulation give very similar results for the moments, but they differ markedly on the time it takes to compute the moments. While it is faster to compute moments from an existing panel of agents, this does not take into account the time it takes to generate the panel.

Mobility Finally, we calculate the ten-year transition rates across deciles of the wealth distribution. These rates are commonly used to study the persistence of wealth inequality and the mobility of agents. Unlike the auto-correlation of wealth, computing transition rates does not require following the full path of individuals, rather it is enough to follow a subset of the population satisfying some initial condition. The histogram iteration method takes advantage of this by iterating from the conditional distribution of agents of each decile to obtain their final distribution as in (10). The transition rates are calculated directly as the mass of the final distribution in each decile. As with the auto-correlations, these transition rates take longer to calculate via the histogram method than using an existing panel of agents. However, simulating that panel of agents is costly relative to solving for the histogram.

3.2 Moments for the overlapping generations model

We now conduct similar exercises on the overlapping generations model. We focus on the behavior of agents along their life-cycle. In particular, we present age-profiles of the wealth distribution for agents with above median income at age 45 and the auto-correlation of wealth between the ages of 35 and 40 and 35 and 55. We compute the moments using the histogram iteration method to iterate over the evolution of a cohort and contrast the results with those of Monte Carlo simulations of up to five hundred thousand agents.[9]

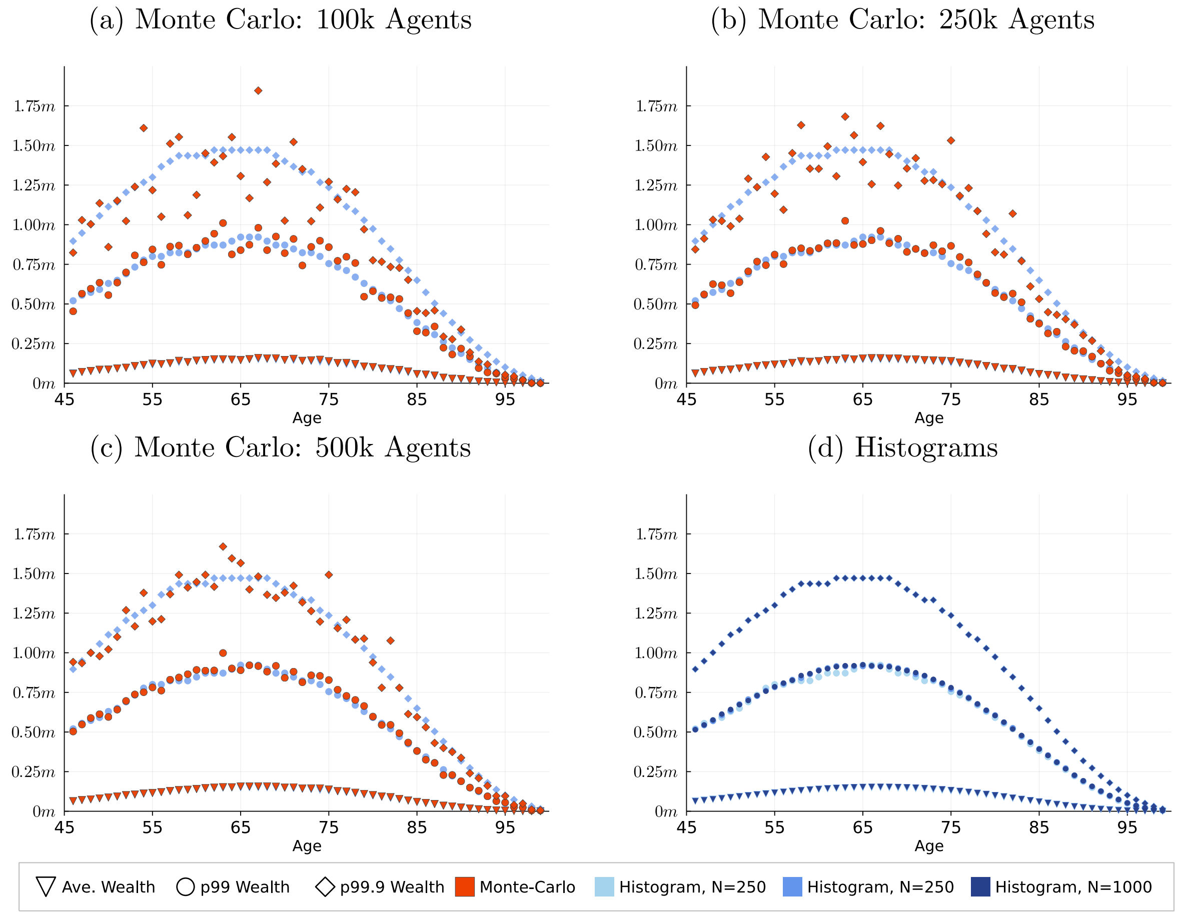
The two methods produce similar moments, with the exception of moments characterizing the top of the wealth distribution. The reason is again that a large number of agents must be simulated in order to have a representative sample of wealthy agents. This is even more so in the context of life-cycle moments because the sample is also conditioned by age, making it more difficult to ensure large sample sizes. In terms of the computational cost, the simulation time is again the main factor making Monte Carlo simulation-based moments more costly.

Wealth Age Profiles We report in Figure 2 the age profile for the wealth of agents with above median income at age 45. The figures show the average wealth along with the 99th and 99.9th percentile of wealth for every age. It is clear that the results obtained from Monte Carlo simulations struggle to capture the top percentiles of the wealth distribution, even though they capture well the average wealth of this sub-population. As before, this is because there are only a small number of "very rich" agents in the Monte Carlo simulation, producing volatile age profiles. This is in contrast with the results obtained from the Histogram that provides stable results even for relatively coarse grids as shown in figure 2d.

The time required to compute the distribution or simulate the agents follows the same pattern described above. As we show in Table 2, the bulk of the computational time is accounted for by computing the histogram $\hat{\lambda}$ or performing the Monte Carlo simulation, with the calculation of the wealth profiles taking just a few seconds at most.

[9] For the cross-sectional moments the sample size refers to the total sample, including agents of all ages. For the auto-correlation we simulate a single cohort of individuals.

Figure 2: Wealth-Age Profiles - Monte Carlo Simulation and Histogram Method



Notes: The figures plot the age profile of wealth starting at age 45. Panels 2a to 2c compute the moments using samples of agents from Monte Carlo simulation and differ in the number of agents being simulated. Triangles correspond to the average wealth at every age, circles to the 99^{th} percentile of the wealth distribution, and diamonds to the 99.9^{th} percentile. Markers in blue correspond to the age profiles using the histogram method. The final panel computes the moments from the conditional distribution of wealth by age using the histogram method with 250, 500, and 1000 grid nodes.

Auto-correlation of Wealth Finally, we compute the five- and fifteen-year auto-correlation of wealth starting at age 35. Both the histogram iteration method and Monte Carlo simulation produce similar results, see Table 2. However, the time required to iterate the histogram increases markedly with the time horizon, making the Monte Carlo simulation faster when computing the fifteen-year auto-correlation.

The reasons for this result is instructive about the relative strengths of the methods. When using Monte Carlo methods, calculating the auto-correlation only requires simulating a single cohort of agents. This cohort simulation is cheaper than a general simulation of

Table 2: Cross-sectional and Longitudinal Moments: Overlapping Generations

	Monte Carlo: Sample Size			Histogram: Grid Size		
	100k	250k	500k	250	500	1000
Wealth Auto-Correlation						
Age 35-40	89.43	89.44	89.46	89.39	89.59	89.66
Age 35-50	62.94	62.85	62.69	63.70	63.84	63.96
Computational Time						
General Simulation	300.5	750.7	1501.1	—	—	—
Cohort Simulation	17.57	43.71	87.33	—	—	—
Distribution $\hat{\lambda}$	—	—	—	113.1	223.9	459.0
Wealth Profiles	0.19	0.71	1.33	0.51	0.99	1.97
Auto-Correlation 35-40	5E-4	4E-3	2E-3	12.45	42.13	145.5
Auto-Correlation 35-50	5E-4	1E-3	2E-3	97.97	355.1	1450.8

Notes: The table reports the auto-correlation of wealth between the ages of 35 and 55. The first block computes the moments approximating the distribution with a Monte Carlo simulation. The second block computes the moments approximating the stationary distribution with histograms on three different grids with 250, 500, and 1000 nodes. The auto-correlation of wealth is computed from the simulation of cohorts between the ages of 35 and 50 of 100k, 250k, and 500k agents, without attrition. The initial distribution is obtained from the histogram with 500 nodes. All times are in seconds.

the whole population and can take advantage of the histogram by using it to obtain the initial distribution of agents at age 35. By contrast, computing the auto-correlation with the histogram iteration method requires solving for the conditional distribution of agents at age 50 $\left(\lambda_s'\right)$ for each initial state s at age 35, see (11). λ_s' describes all the possible paths that a 35 year old can take in their next fifteen years. Computing λ_s' requires iterating forward as in equation (5) multiple times. The complexity of this step increases with the time horizon as the initial mass of agents fans out across the state space.

4 Discussion

We have shown how to use a histogram approximation of the stationary distribution of agents and its associated Markov kernel to efficiently compute cross-sectional and longitudinal moments without having to simulate large samples of agents through Monte Carlo methods. We illustrated the workings of the method in the context of baseline models that abstract from many of the characteristics of applied work. However, the method we propose can also be applied in other scenarios. We therefore end with a short discussion of some of the natural extensions of the histogram iteration method and its main limitations.

Extensions The histogram iteration method can be easily applied to models that allow for additional endogenous choices (e.g., labor supply). In this case the policy functions can be solved with extensions of the endogenous grid method like those in Barillas and Fernández-Villaverde (2007) and Fella (2014). Once the policy functions are obtained, the construction of the Markov kernel and the histogram that approximates the distribution follow as above. Similarly, the method applies to non-stationary problems where the distribution of agents changes over time, or the agents' choices change (therefore making the Markov kernel time-varying). This can happen because of changes in policy variables or prices. The histogram iteration method is already built to capture changes in the distribution, as the iteration in equation (5) shows. The only change comes in by indexing the Markov kernel by time when iterating over an initial distribution of agents.

The histogram iteration method can also be applied with only minor changes to continuous-time heterogeneous agent models (see for instance Herreño and Ocampo, 2020). In particular, the solution of these models by means of the Finite Difference method is constructed from a (sparse) matrix A that characterizes the approximation to the Hamilton-Jacobi-Bellman equation (see, Achdou, Han, Lasry, Lions, and Moll, 2021). The adjoint of this matrix plays the same role as the Markov kernel T described above and characterizes the solution to the Kolmogorov Forward equation that describes the evolution of the distribution of agents. In this way, the solution of the model generates a value function, policy functions, a distribution over states, and an operator to iterate the distribution just as in Section 2.

The differences in the implementation of the histogram iteration method come from the possibility of taking advantages of the sparseness of matrix A (hence of the Markov kernel), the stipulation of a time step (Δt) when iterating forward in time, and the integration with respect to the distribution of continuous states. Alternative solution methods for continuous time models as those in Eslami and Phelan (2022) also allow for a direct implementation of the histogram iteration method.

Limitations The histogram iteration method is generally an efficient way for calculating a wide array of cross-sectional and longitudinal moments. However, longitudinal moments that involve individual outcomes of a large subset of the population, or that involve long periods of time, can be expensive to calculate. As we discussed in Section 3, this is because the full history of individuals' paths must be mapped in order to compare the individuals' initial and final outcomes, unlike for other moments that focus on group outcomes like transition rates. This leads to cases where Monte Carlo methods can be more efficient, as was the case with the computation of the twenty-year auto-correlation of wealth discussed in Section 3.2.[10]

The histogram iteration method takes advantage of the histogram approximation of the distribution of agents and the associated Markov kernel, which are often already computed as part of solving the model. Because of this, the histogram method will usually generate time-savings even when the computation of specific moments is costlier than the computation from a simulated Monte Carlo panel, as the simulation has to be conducted on top of the model solution. This makes the key computational trade-off for computing moments clear: the complexity of the moment is weighed against the complexity of simulating a representative sample of agents. We have shown that this trade-off will usually land in favor of using the model's own stationary distribution and Markov kernel, allowing researchers to avoid both coding and running computationally-costly Monte Carlo simulations.

[10] The same principle applies to moments that involve the outcomes of agents in intervening periods, rather than just the initial and final outcomes. For example, computing the distribution of lifetime earnings in our OLG model proves to be unfeasible. Doing so would require us to compute the time paths of each possible income realization over the 81 year lifespan of agents. With 11 income states, there are $11^{81} \approx 6.8 \times 10^{17}$ possible histories of lifetime income.

References

Achdou, Yves, Jiequn Han, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. 2021. "Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach." *The Review of Economic Studies* 89 (1):45–86. URL https://doi.org/10.1093/restud/rdab002.

Aiyagari, S. Rao. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *The Quarterly Journal of Economics* 109 (3):659–684. URL http://www.jstor.org/stable/2118417.

Barillas, Francisco and Jesús Fernández-Villaverde. 2007. "A generalization of the endogenous grid method." *Journal of Economic Dynamics and Control* 31 (8):2698–2712. URL https://www.sciencedirect.com/science/article/pii/S0165188906001783.

Bell, Felicite C. and Michael L. Miller. 2002. "Life Tables for the United States Social Security Area: 1900-2100"." Actuarial Study 116, Office of the Actuary, Social Security Administration. URL https://www.ssa.gov/oact/NOTES/as120/LifeTables_Body.html.

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu. 2011. "The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents." *Econometrica* 79 (1):123–157. URL https://doi.org/10.3982/ECTA8416.

Carroll, Christopher D. 2006. "The method of endogenous gridpoints for solving dynamic stochastic optimization problems." *Economics Letters* 91 (3):312–320. URL https://www.sciencedirect.com/science/article/pii/S0165176505003368.

De Nardi, Mariacristina and Giulio Fella. 2017. "Saving and wealth inequality." *Review of Economic Dynamics* 26:280–300. URL https://www.sciencedirect.com/science/article/pii/S1094202517300546.

De Nardi, Mariacristina, Eric French, and John Bailey Jones. 2016. "Savings After Retirement: A Survey." *Annual Review of Economics* 8 (1):177–204. URL https://doi.org/10.1146/annurev-economics-080315-015127.

Eslami, Keyvan and Thomas Phelan. 2022. "Applications of Markov Chain Approximation Methods to Optimal Control Problems in Economics." Working Paper 21-04R, Federal Reserve Bank of Cleveland. URL https://doi.org/10.26509/frbc-wp-202104r.

Fella, Giulio. 2014. "A generalized endogenous grid method for non-smooth and non-concave problems." *Review of Economic Dynamics* 17 (2):329–344. URL https://www.sciencedirect.com/science/article/pii/S1094202513000392.

Gouin-Bonenfant, Émilien and Alexis Akira Toda. 2022, forthcoming. "Pareto Extrapolation: An Analytical Framework for Studying Tail Inequality." *Quantitative Economics* URL https://qeconomics.org/ojs/forth/1817/1817-2.pdf.

Guvenen, Fatih, Gueorgui Kambourov, Burhanettin Kuruscu, Sergio Ocampo, and Daphne Chen. 2019. "Use It or Lose It: Efficiency Gains from Wealth Taxation." Working Paper 26284, National Bureau of Economic Research. URL http://www.nber.org/papers/w26284.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2009. "Quantitative Macroeconomics with Heterogeneous Households." *Annual Review of Economics* 1 (1):319–354. URL https://doi.org/10.1146/annurev.economics.050708.142922.

Heer, Burkhard and Alfred Maußner. 2005. *Dynamic General Equilibrium Modelling*. Springer.

Herreño, Juan and Sergio Ocampo. 2020. "elf-Employment and Development." Working Paper 2020-9, Western University. URL https://ir.lib.uwo.ca/economicscibc/158/.

Judd, Kenneth L. 1998. *Numerical Methods in Economics*. MIT Press.

Rouwenhorst, K. Geert. 1995. "Asset Pricing Implications of Equilibrium Business Cycle Models." In *Frontiers of Business Cycle Research*, edited by Thomas Cooley, chap. 10. Princeton University Press, 394–330.

Stachurski, John and Alexis Akira Toda. 2019. "An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity." *Journal of Economic Theory* 182:1–24. URL https://www.sciencedirect.com/science/article/pii/S0022053119300353.

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron. 2004. "Cyclical Dynamics in Idiosyncratic Labor Market Risk." *Journal of Political Economy* 112 (3):695–717. URL https://doi.org/10.1086/383105.

Tan, Eugene. 2020. "A fast and low computational memory algorithm for non-stochastic simulations in heterogeneous agent models." *Economics Letters* 193:109285. URL https://www.sciencedirect.com/science/article/pii/S0165176520301907.

Tauchen, George. 1986. "Finite state markov-chain approximations to univariate and vector autoregressions." *Economics Letters* 20 (2):177–181. URL https://www.sciencedirect.com/science/article/pii/0165176586901680.

U.S. Bureau of Economic Analysis. 2022. "Economic Data." FRED, Federal Reserve Bank of St. Louis; April 9, 2022., https://fred.stlouisfed.org.

Young, Eric R. 2010. "Solving the incomplete markets model with aggregate uncertainty using the Krusell–Smith algorithm and non-stochastic simulations." *Journal of Economic Dynamics and Control* 34 (1):36–41. URL https://www.sciencedirect.com/science/article/pii/S0165188909001316.