Estimating Heterogeneous Economies with Micro Data*

Man Chon (Tommy) Iao† Yatheesan J. Selvakumar‡

July 17, 2026

Abstract

We give sufficient conditions under which dynamic equilibrium models with heterogeneous-agents can be represented by a first-order reduced-rank vector autoregression. We exploit this result to develop an econometric framework that enables the rapid estimation of a rich class of models with macro and repeated cross-sections of micro data. Simulation evidences suggest that our method delivers increased precision of parameter estimates than conventional approaches, particularly for parameters governing cross-sectional dynamics. We apply our method to estimate a medium-scale HANK model with heterogeneous earnings exposures to aggregate fluctuations at the household-level. Our estimates imply that low-income households are more sensitive to changes in aggregate income on average, and that this sensitivity is heightened conditional on monetary policy shocks. Through the lens of the model, our method estimates that heterogeneous earnings exposures amplify the aggregate consumption response to monetary policy shocks by 40%, substantially larger than those implied by traditional estimation methods.

Keywords: structural estimation, singular value decomposition, dynamic mode decomposition, heterogeneous-agent models, earnings heterogeneity, vector autoregressions

JEL Classification Numbers: C13, C32, E1

*We thank Thomas J. Sargent, Jarda Borovicka, Simon Gilchrist, Corina Boar, Virgiliu Midrigan, Juan Rubio-Ramirez, Timothy Cogley, Diego Perez, Mark Gertler, Thomas Philippon, Mikkel Plagborg-Møller, Venky Venkateswaran, Christian Wolf, Gianluca Violante, Joe Hazell, Zejin Shi and Christopher Adjaho for their helpful comments. The views expressed here are those of the authors and do not necessarily reflect those of the Reserve Bank of Australia.

†Reserve Bank of Australia, Email: iaot@rba.gov.au

‡UCL, Email: y4d.s@nyu.edu

1 Introduction

Dynamic stochastic general equilibrium models featuring rich heterogeneity have become a central tool in modern macroeconomics. Their appeal stems from their ability to illuminate how heterogeneity at the micro level shapes aggregate outcomes. To incorporate micro data into the quantitative analysis of these models, the literature has typically followed a two-step procedure: first, calibrating the steady state to match micro data moments, and then estimating the model's dynamics using aggregate time-series data. This approach, however, leaves much of the *dynamic* information contained in micro data untapped.[1] A likely reason is that evaluating the likelihood of datasets containing thousands of cross-sectional observations over several decades is computationally prohibitive for most existing methods.[2]

In this paper, we develop a tractable econometric framework that approximates the joint likelihood of macro and repeated cross-sections of micro data. The outcome is a rapid algorithm that estimates the structural parameters of a dynamic heterogeneous-agent equilibrium model. Our key assumption is that the model has a Gaussian linear state-space representation, where the number of states (r) are small relative to the number of observables ($N \gg r$). While the Gaussian linear state-space representation is natural for many benchmark models such as Bewley-Huggett-Aiyagari models with aggregate shocks, we provide numerical evidence suggesting that these models do indeed exhibit a low-rank structure.[3] Importantly, although our estimation method relies on the presence of such a low-rank structure, it does not require computing an explicit low-rank state-space representation and is therefore fully compatible with the sequence-space solution method (Auclert et al., 2021a).

Our main theoretical result is that such a model is well-approximated by a reduced-rank *first-order* vector autoregression. The logic of the theorem is as follows. In more general settings, Kalman filtering theory dictates that the best predictor of the hidden states (in a least-squares sense) conditions on the infinite history of data. Thus, the state-space model has a reduced-rank VAR(∞) representation, where the rank of the coefficient matrices is no larger than the number of states. In

[1] A notable extension of this approach is to augment the second-step estimation with a small number of time-varying micro moments, such as the top-10 percent wealth share, the standard deviation of sales growth, or the frequency of price adjustments. See Bayer et al. (2024), Mongey and Williams (2017), and Morales-Jimenez and Stevens (2024) for examples.

[2] Liu and Plagborg-Møller (2023) develop a full-information estimation method that evaluates the joint likelihood of macro and micro data. Their approach requires the computation of a state-space representation and the use of Monte Carlo integration at each likelihood evaluation, making it computationally demanding for medium-scale heterogeneous-agent models.

[3] This low-rank structure also exists in micro data, rendering it a realistic feature that should be present in structural models. For instance, Sargent et al. (2026a) find that a three-factor model describes well the cross-section of consumption in the Consumer Expenditure Survey.

our setting, as N grows large, the expectation of the hidden states conditional on the infinite history of data coincides with that conditional on the present data only. In other words, the data at time t is sufficient to precisely estimate the hidden states, effectively substituting for its infinite history. Thus, the VAR(∞) representation converges to a VAR(1), which we show occurs at rate N. Our result is useful because computing the likelihood of a reduced-rank first-order vector autoregression is fast and scales well with N. Its likelihood-based foundation also lends itself naturally to Bayesian inference, which we adopt in our empirical application.

To operationalize our theory, we must take two important steps. The first step is to compute the reduced-rank VAR(1), conditional on a vector of structural parameters. We propose two approaches depending on whether second-moments of the observables are available analytically. If so, the reduced-rank VAR(1) can be computed quickly and accurately using the canonical correlations approach of Anderson and Rubin (1949). This will be the case for our applications below. If not, it can be computed by simulating a long sequence of the observables from the model and computing a Dynamic Mode Decomposition (Brunton et al., 2015; Sargent et al., 2026b) of the simulated data.

The second step is to organize the micro data in a model-consistent way. To make matters concrete, imagine our data consists of repeated cross-sections of consumption. In the recursive representation of the household problem, households are distinguished only by the value of their idiosyncratic states. Every period, we group individuals into bins according to their states[4] and define consumption for each bin as the mean consumption of all the individuals in that bin. Repeating this every period constructs our micro dataset for consumption. The same principles can be applied to micro data on firms or other agents, provided that the model-consistent idiosyncratic state is observed.

What do we gain by using repeated cross-sections of micro data instead of only simple moments? We answer this question using a simulation experiment and an empirical application. Our laboratory is a Krusell-Smith economy with a TFP shock and a income risk shock that generates time-varying income dispersion. We simulate macro and micro consumption data from this economy and estimate the structural parameters. We find that estimating the model using our method that exploits both macro and micro data leads to a sharp increase in the precision of the parameter estimates compared to a MLE benchmark that uses macro data and the time-series of the cross-sectional consumption variance. In particular, our method delivers the biggest improvement in the

[4]Grouping individuals in this way means the same individual may change bins over time. Moreover, the structural model dictates the grouping criteria.

parameter estimates of the income-risk shock process. Overall, the simulation results validate our econometric theory and show that these gains are especially pronounced for parameters governing micro-level dynamics.

In light of these advantages, we apply our method to quantify the amplification of monetary policy arising from heterogeneous earnings exposures—the differential sensitivity of households' earnings to aggregate shocks. Our empirical application is based on a benchmark medium-scale heterogeneous-agent New Keynesian model, similar to Bayer et al. (2024), featuring both price and wage rigidities. We capture heterogeneous earnings exposures through an incidence function that governs the comovement between individual earnings and aggregate income. The incidence function is parameterized flexibly and estimated jointly with the rest of the model. Motivated by the literature on the unequal effects of monetary policy (Holm et al., 2021; Amberg et al., 2022; Andersen et al., 2023; Broer et al., 2026), we allow the incidence function to vary with monetary policy shocks. To the best of our knowledge, this is the first study to estimate the heterogeneous incidence of both monetary policy and aggregate fluctuations by jointly exploiting macroeconomic and household-level data within a fully structural framework.

We estimate the model using two datasets: one with macro data only – aggregate output, consumption, investment, inflation, wages and nominal interest rate between 1959 and 2020 – which we call `Macro`; and one with the preceding macro data *and* repeated cross-sections of consumption and income from the Consumer Expenditure Survey – between 1990 and 2020 – which we call `Macro+Micro`. As is typical, the micro data span a considerably shorter sample period. This further underscores the advantage of our approach over one based solely on aggregate moments, since it enables the researcher to exploit substantially more information within each period.

Comparing the parameter estimates, two differences stand out.

1. `Macro+Micro` implies a "U"-shaped incidence function: low and high income households exhibit similar exposures to aggregate income. This sensitivity is further exacerbated conditional on a monetary policy shock for low income households. On the contrary, `Macro` implies that low income households are less sensitive to changes in aggregate income and that high income households are more exposed to monetary policy shocks, though the parameters are not as precisely estimated.

2. `Macro+Micro` implies larger standard deviations of the structural shocks compared to `Macro`

The reason for these differences can be understood by comparing the volatility of income at the micro level. Income volatility in the data exhibits a "U"-shape with low and high productivity households experiencing similar levels. `Micro+Macro` is able to match both the shape and level of volatility in the data but `Macro` portrays a counterfactually increasing shape and significantly lower volatility for all productivity levels.

These features come at a cost. In comparing the volatility for aggregate variables, we find that `Macro+Micro` overstates the aggregate volatility compared to the data. This is in direct contrast to `Macro` that matches the macro data well along this dimension. The dichotomy borne from this empirical exercise sheds light on an important dimension of misspecification in benchmark HANK models, in that they are unable to generate the appropriate level of volatility at the micro and macro level simultaneously.

Given these directly contrasting parameterization, it appears important to compare them in the spirit of an external validation exercise. We estimate a structural vector autoregression (SVAR) and instrument monetary policy shocks with those identified by Bauer and Swanson (2023). We compare the `Macro` and `Macro+Micro` impulses responses of output and inflation to a 25bps monetary policy shock to those from the SVAR-IV. We find that `Macro+Micro` impulse responses are larger in magnitude and closer to the empirical evidence than `Macro`.

Lastly, we quantify the effect of heterogeneous earning exposures on the amplification of consumption to monetary policy shocks. The exercise is to compare the impulse responses in `Macro` and `Macro+Micro` to counterfactual impulses responses from models *without* heterogeneous earnings exposures. For `Macro+Micro`, this channel leads to a 40% amplification in consumption response. The reason is that in `Macro+Micro`, low-income households are highly sensitive to changes in aggregate labor income. Since they are also high-MPC (marginal propensity to consume) households, they react strongly to large changes in their income, amplifying the shock further.[5] In `Macro`, amplification is much lower – less than 10% – owing to the small sensitivity of low-income households to aggregate income.

The rest of the paper is organized as follows. Section 2 situates our contributions within the context of related work. Section 3 outlines the econometric framework and lays out the theoretical foundations. Section 4 pursues a Monte-Carlo simulation exercise within the context of a Krusell-Smith model with redistributive shocks. Section 5 estimates a medium-scale HANK model to

[5] See Patterson (2023) and Auclert (2019) for a discussion on how the covariance between incidence and MPCs amplifies aggregate shocks.

analyze the effect of earnings heterogeneity on consumption amplification to a monetary policy shock. Section 6 concludes.

2 Related Literature

Our paper contributes to the recent literature on structural estimation of heterogeneous-agent models using both macro and micro data. The paper closest to us is Liu and Plagborg-Møller (2023) who compute a numerically unbiased estimate of the likelihood. Crucial to their framework is a decomposition of the likelihood into a *macro part* and a *micro part, conditional on the macro state variables*, which are treated as unobserved. Thus, the macro data play an important role in determining the aggregate states. Our approach does not require such dependence, but rather infers the states jointly from both macro and micro data. Thus, our approach is still viable were macro data not to be used. Moreover, their approach is better suited to a state-space representation of the model obtained from the Reiter (2009) model solution approach. Our approach is flexible enough such that its convenience is not bound to any particular solution method.

Fernández-Villaverde et al. (2023) suggest a promising avenue for using micro data in estimating a model with financial frictions. However, they assume that the underlying states are observed, while our approach infers them from the macro and micro data jointly. Chang et al. (2021) compute functional vector autoregressions where the microdata are the density of the cross-section. They estimate the feedback loop betwen aggregate time series and micro-data, but do so without the context of fully-specified equilibrium model.

Kase et al. (2022) use neural networks to solve and estimate a non-linear HANK model with the zero-lower bound. They compute the likelihood via the particle filter and focus mainly on aggregate data; and speed up the likelihood computation by effectively treating the parameters as additional inputs into the neural network. In their empirical application, they estimate the model using only macro data and choose to "incorporate information contained in the cross-section through the prior". Unlike them, we estimate linearized versions of HANK models and put a lot of emphasis in computing the joint likelihood of macro *and* micro data.

Traditional estimation methods are not well-suited to scale with the size of the micro data. One approach is Auclert et al. (2021b), who compute the likelihood by exploiting the solution's implied moving average representation. This requires vectorizing the $N \times T$ data matrix, where the corresponding covariance matrix is $NT \times NT$. In conventional situations with only aggregate

data, N is relatively small (typically less than ten), so the covariance matrix is computationally manageable. In using individual-level micro data, N will undoubtedley be large making the inversion of the $NT \times NT$ covariance matrix prohibitively slow.[6] Our paper fills the gap by providing a solution in setting in which the conventional approach becomes infeasible.

Another approach is to use the Whittle likelihood approximation in frequency domain, as in Hansen and Sargent (1981), Christiano and Vigfusson (2003) and Plagborg-Møller (2019). While estimation is feasible for large N, the quality of the approximation relies on large T, which can prove difficult for existing micro-datasets. In contrast, our method relies on large N, which seems to us a far easier requirement to satisfy.

In Appendix C, we describe the above methods and compares the computational efficiency in computing the likelihood, in the context of the Bewley-Huggett-Aiyagari model in Section 4. The results describe how the traditional method is prohibitively slow when N is large. Furthermore while the Whittle likelihood method is feasible, our estimator has more attractive finite-sample properties.

Many papers have estimated structural models using solely macro data, including Winberry (2018), Auclert and Rognlie (2018), Bayer and Luetticke (2020); others have also done so with some moments of the micro-data, including Acharya et al. (2020), Mongey and Williams (2017) and Challe et al. (2017). We employ our method to estimate a benchmark HANK model using individual-level consumption and income from the Consumer Expenditure Survey, as well as the conventional aggregate time-series. We are not aware of other papers that pursue this empirical analysis.[7]

3 Econometric framework

Let $\mathcal{M}_{\theta_0}(\{\mathbf{y}_t\})$ for some $\theta_0 \in \Theta$ be a linearized fully-specified dynamic general equilibrium model that governs the sequence of observables $\mathbf{y}_t \in \mathbb{R}^{N\times 1}$. Similar to the vast literature on estimation of representative-agent models, it has become standard to consider $\mathbf{y}_t$ as a vector of aggregate observables, e.g. GDP, consumption, inflation, and so on. Our goal is to extend this list of observables to include individual-level micro data. To sidestep difficulties of computing the likelihood arising from large N, we propose an econometric framework that well-approximates $\mathcal{M}_{\theta_0}(\mathbf{y}_t)$ in such situations.

[6] In our two examples below, $N = 86$ and $N = 300$; and $T = 120$, for 30 years of quarterly data.
[7] Liu and Plagborg-Møller (2023) show the benefits of including micro data only on simulated data.

3.1 Data organizing principle

A key ingredient in our methodology is the organization of the micro data. Through the lens of the model, individuals differ only in their values of the idiosyncratic state vector, $\mathbf{z} \in \mathcal{Z}$ with cardinality N.[8] To make matters concrete, Aiyagari (1994) distinguish individuals through their asset holdings and productivity levels; Winberry (2021) distinguish firms through their productivity, capital, stock of depreciation allowances and their current draw of fixed costs. Thus, each period, we organize the micro data by placing individuals into one of N bins, according to their states variables. In our baseline theory, we will assume that the econometrician observes the individuals' idiosyncratic state vector. Define $y_t(\mathbf{z})$ the consumption at time t of an individual with idiosyncratic state vector $\mathbf{z}$ and the vector $\mathbf{y}_t = [y_t(\mathbf{z}_1), \ldots, y_t(\mathbf{z}_N)]^\top \in \mathbb{R}^{N\times 1}$ represents the observables at time t of individuals. Importantly, since our notion of an individual is bound by the states, individuals can and will move across rows of $\mathbf{y}_t$ over time.

3.2 State-space representation

We begin our theoretical argument by stating our main assumption.

Assumption 1. *$\mathcal{M}_{\theta_0}(\mathbf{y}_t)$ has a Gaussian linear state-space representation, **and** the number of observables (N) are much larger than the number of states (r)*

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\,\mathbf{x}_t + \mathbf{C}\,\mathbf{w}_{t+1} \\ \mathbf{y}_t &= \mathbf{G}\,\mathbf{x}_t + \mathbf{v}_t, \end{aligned} \tag{1}$$

for states $\mathbf{x}_t \in \mathbb{R}^{r\times 1}$, and observables $\mathbf{y}_t \in \mathbb{R}^{N\times 1}$; where shocks $\mathbf{w}_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r\times r})$, measurement errors $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and $\mathbf{w}_s \perp \mathbf{v}_\tau$ for all s, τ; here $\mathbf{A} \in \mathbb{R}^{r\times r}$, $\mathbf{C} \in \mathbb{R}^{r\times r}$ and $\mathbf{G} \in \mathbb{R}^{N\times r}$ and $\mathbf{R} \in \mathbb{R}^{N\times N}$. Furthermore, the model is stable that all eigenvalues of $\mathbf{A}$ is strictly smaller than 1 in modulus.

That $\mathcal{M}_{\theta_0}(\mathbf{y}_t)$ has a state-space representation is not controversial, it in fact forms the basis for many state-of-the-art solution and estimation methods.[9] The key difference is that the number of states are small relative to the size of the observable vector. At first glance, this might appear restrictive, since an equilibrium dynamic heterogeneous-agent model with aggregate uncertainty typically involves forward looking agents that forecast prices, which depend on the distribution

[8] We assume that any continuous idiosyncratic states, e.g. productivity, have been discretized appropriately, for example via the Rouwenhorst (1995) or Tauchen (1986) method.

[9] Examples include Reiter (2009), Winberry (2018), and Bayer et al. (2024).

of households. The exact solution therefore requires its inclusion into the state vector, making it infinite-dimensional. Krusell and Smith (1998) show that in their heterogeneous-agent , replacing the asset distribution with its first moment results in a "good" approximation for equilibrium of their model. Our assumption is in a similar spirit, though we do not require the state to be moments of the distribution. Moreover, as will be clear, this assumption is instrumental that we do not need to compute a low-dimensional state space representation for the structural model. In Appendix E.1, we show that the workhorse heterogeneous-agent models with aggregate shocks used in the literature today satisfy this assumption.

This assumption is also easily testable. A simple heuristic test is to simulate a long sequence $\{\mathbf{y}_t\}$ and compute its principle components, since Assumption 1 implies that the elements of $\mathbf{y}_t$ loads on a small number of principle components. We discuss other tests within the context of our applications in Sections 4.3 and 5.

We place the following additional restrictions on the Gaussian linear state-space model (1).

Assumption 2. *Gaussian linear state-space system* (1) *also satisfies the following restrictions*

1. $\mathbf{G}, \mathbf{A}$ *has full column rank (i.e.* $rank(\mathbf{G}) = rank(\mathbf{A}) = r$*)*
2. $\frac{1}{N}\mathbf{G}^\top\mathbf{G} \to \mathbf{\Sigma_G}$ *as* $N \to \infty$*, where* $\mathbf{\Sigma_G}$ *is positive definite*
3. $\lambda_{\max}(\mathbf{R}) = o(N)$*, where* $\lambda_{\max}(\mathbf{R})$ *denotes the largest eigenvalue of* $\mathbf{R}$*;* $\liminf \lambda_{\min}(\mathbf{R}) > 0$*, where* $\lambda_{\min}(\mathbf{R})$ *denotes the smallest eigenvalue of* $\mathbf{R}$*.*

The first condition requires that the columns of $\mathbf{G}$ and $\mathbf{A}$ are linearly independent. To gain some intuition, consider an economic model for which the rows of $\mathbf{y}_t$ are the consumption of different households. Columns of $\mathbf{G}$ then represent the response of households' consumption to each aggregate factor. Condition 1 then implies that the consumption responses to each factor are sufficiently different across households. A representative agent model would violate this assumption, for example. This condition highlights our identification strategy that both requires and exploits the rich heterogeneity in the micro-data. The assumption on $\mathbf{A}$ means that there is no redundant state, and allows for higher-order lags for the states.

The second condition concerns the asymptotic property of the model when the number of observables grows and is standard in the factor analysis literature (e.g. Chamberlain and Rothschild 1982, Stock and Watson 2002, Bai and Ng 2006). To continue the example of household consumption, it implies that the cross-sectional variance of consumption is finite.

The third condition requires that the idiosyncratic noise do not grow faster than the number of observables and that a minimal level of noise exists. The former ensures that the high-dimensional observables provide meaningful information for inference and the latter prevents stochastic singularity. Both are standard in factor analysis. As will become important in our empirical section, this assumption is weak enough to allow for heteroskedasticity and correlated measurement errors.

3.3 Theory

The starting point for our theoretical analysis is the vector auto-regression (VAR) representation of system (1).

Proposition 1. *There exists an infinite-order VAR representation of* (1) *in* $\mathbf{y}_t$*, given by*

$$\mathbf{y}_t = \sum_{j=1}^{\infty} \mathbf{B}_j^{\infty}\, \mathbf{y}_{t-j} + \mathbf{a}_t \tag{2}$$

$$\mathbb{E}[\mathbf{a}_t\, \mathbf{y}_{t-j}^{\top}] = \mathbf{0} \quad \textit{for all } j \geq 1$$

$$\mathbb{E}[\mathbf{a}_t\, \mathbf{a}_t^{T}] =: \mathbf{\Omega} = \mathbf{G}\, \mathbf{\Sigma}_{\infty}\, \mathbf{G}^{\top} + \mathbf{R}$$

$$\mathbf{B}_j^{\infty} = \mathbf{G}(\mathbf{A} - \mathbf{K}\,\mathbf{G})^{j-1}\, \mathbf{K} \quad \forall j \geq 1 \tag{3}$$

$$rank(\mathbf{B}_j^{\infty}) = r \quad \forall j \geq 1$$

where $\mathbf{K} = \mathbf{A}\, \mathbf{\Sigma}_{\infty}\, \mathbf{G}^{\top}\, \mathbf{\Omega}^{-1}$ *and* $\mathbf{\Sigma}_{\infty} = \mathbf{C}\, \mathbf{C}^{\top} + \mathbf{K}\, \mathbf{R}\, \mathbf{K}^{\top} + (\mathbf{A} - \mathbf{K}\, \mathbf{G}) \mathbf{\Sigma}_{\infty} (\mathbf{A} - \mathbf{K}\, \mathbf{G})^{\top}$.

Proposition 1 states the population formula for computing the the expectations of $\mathbf{y}_{t+1}$, given the information contained in its infinite history. Since $\mathbf{B}_j^{\infty} \neq \mathbf{0} \quad \forall j \geq 1$, it implies that a truncation induces a non-trivial loss in prediction accuracy. Our theoretical analysis studies what happens to this prediction accuracy as $N \to \infty$.

Lemma 1. *Under Assumption 2, as the number of observables grows* ($N \to \infty$)*, the matrix* $\mathbf{A} - \mathbf{K}\, \mathbf{G} \to \mathbf{0}$.

Corollary 1. *When* $\mathbf{A} - \mathbf{K}\, \mathbf{G} = \mathbf{0}$, $\mathbb{E}[\mathbf{x}_t | \, \mathbf{y}^t] = \mathbf{L}\, \mathbf{y}_t$ *and* $\mathbb{E}[\mathbf{y}_{t+1} \, | \, \mathbf{y}^t] = \mathbf{G}\, \mathbf{K}\, \mathbf{y}_t$ *where* $\mathbf{L} = \mathbf{\Sigma}_{\infty}\, \mathbf{G}^{\top}\, \mathbf{\Omega}^{-1}$.

Lemma 1 and Corollary 1 taken together show that as $N \to \infty$, the estimate of $\mathbf{x}_t$ conditional only the contemporaneous data $\mathbf{y}_t$ and that conditional on the full history of data $\mathbf{y}^t$ converge: when $\mathbf{A} - \mathbf{K}\mathbf{G} = 0$, $\mathbb{E}[\mathbf{x}_t \, | \, \mathbf{y}^t] = \mathbb{E}[\mathbf{x}_t \, | \, \mathbf{y}_t]$. Furthermore, the Markovian property of the system implies that next period's forecasts computed conditional on $\mathbf{y}_t$ and $\mathbf{y}^t$ also converge.

Theorem 1. *Suppose Assumption 2 holds. Then as* $N \to \infty$*,*

1. $\mathbf{B}_j^\infty \to \mathbf{0}\ \forall j \geq 2$. *Furthermore,* $\limsup(N^{j-1} \left\|\mathbf{B}_j^\infty\right\|) < \infty\ \forall j \geq 2$.

2. *The infinite-order VAR representation of* (1) *collapses to a first-order VAR representation where*

$$\begin{aligned}
&\mathbf{y}_t = \mathbf{B}_1^\infty \mathbf{y}_{t-1} + \mathbf{a}_t \\
&\mathbb{E}[\mathbf{a}_t \mathbf{y}_{t-j}^\top] = \mathbf{0} \quad \textit{for all } j \geq 1 \\
&\mathbb{E}[\mathbf{a}_t \mathbf{a}_t^T] = \mathbf{\Omega} = \mathbf{G}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top + \mathbf{R} \\
&\mathbf{B}_1^\infty = \mathbf{G}\,\mathbf{K} \\
&rank(\mathbf{B}_1^\infty) = r
\end{aligned} \tag{4}$$

Theorem 1 states that as N increases, all the auto-regressive coefficient matrices in the VAR(∞) representation other than the first lag converge to the zero matrix. Thus, the rank-r VAR(1) coincides with the VAR(∞) representation at the limit.

For the purpose of likelihood approximation using the limit (4), we need to further establish that the approximated likelihood converges to the truth.

Theorem 2. *Let* $\mathcal{P}_\rho(\mathbf{y}^t)$ *be the reduced-rank vector auto-regression* (4) *where* $\rho \in P$ *is the population maximum likelihood estimator. Suppose Assumption* 2 *holds. As* $N \to \infty$, $\mathcal{P}_\rho(\mathbf{y}^t)$ *approximates* $\mathcal{M}_{\theta_0}(\mathbf{y}^t)$ *in the sense that the Kullback-Leibler divergence vanishes*

$$\lim_{N\to\infty} \int \log \frac{\mathcal{M}_{\theta_0}(\mathbf{y}^t)}{\mathcal{P}_\rho(\mathbf{y}^t)} \mathcal{M}_{\theta_0}(\mathbf{y}^t) d\,\mathbf{y}^t = 0. \tag{5}$$

Theorem 2 states that as N becomes large, the structural heterogeneous-agent equilibrium model is well-approximated by a reduced-rank VAR(1). Since computing (4) is fast, this theorem opens the door to rapid computation of the likelihood $\mathcal{M}_{\theta_0}(\mathbf{y}_t)$, intermediated by the reduced-rank VAR(1). Importantly, though the logic justifying our algorithm involves the state-space system with hidden factors, our algorithm conveniently does not require estimating them.

3.4 Reduced-rank first-order VAR

Our theoretical results in the previous subsection justify a fast algorithm for approximating $\mathcal{M}_{\theta_0}(\mathbf{y}_t)$. In this section, we explore two approaches to compute the reduced-rank VAR(1):

$$\begin{aligned} \mathbf{y}_t &= \mathbf{B}\,\mathbf{y}_{t-1} + \mathbf{a}_t \\ rank(\mathbf{B}) &= r \\ \mathbb{E}[\mathbf{a}_t\,\mathbf{a}_t^\top] &= \mathbf{\Omega} \end{aligned} \tag{6}$$

The first approach implements the canonical correlations formulas of Anderson and Rubin (1949) and Anderson (1951). Its main benefit is that it obtains analytic expressions for population $\mathbf{B}$ and $\mathbf{\Omega}$, thus improving on accuracy and speed. This is our preferred approach, which as we show below fits nicely with the sequence-space solution method. The second uses a computationally efficient algorithm that can be used when the structural model does not admit analytic expressions of population covariances but can be simulated from. The approach builds on the Dynamic Mode Decomposition and its connection to linear state-space models made by Sargent et al. (2026b). This algorithm is described in Appendix B.1.

Computing $\mathbf{B}$ and $\mathbf{\Omega}$ Let the population covariance matrices be defined by

$$\mathbf{\Sigma}_0 = \mathbf{E}[\mathbf{y}_t\,\mathbf{y}_t^\top] \quad \mathbf{\Sigma}_1 = \mathbf{E}[\mathbf{y}_t\,\mathbf{y}_{t-1}^\top] \tag{7}$$

The canonical correlation and variates in the population are defined by

$$\begin{pmatrix} -\rho\mathbf{\Sigma}_0 & \mathbf{\Sigma}_1 \\ \mathbf{\Sigma}_1^\top & -\rho\mathbf{\Sigma}_0 \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = 0 \tag{8}$$

where ρ, α, γ satisfy

$$\begin{vmatrix} -\rho\mathbf{\Sigma}_0 & \mathbf{\Sigma}_1 \\ \mathbf{\Sigma}_1^\top & -\rho\mathbf{\Sigma}_0 \end{vmatrix} = 0, \quad \alpha'\mathbf{\Sigma}_0\alpha = 1, \quad \gamma'\mathbf{\Sigma}_0\gamma = 1 \tag{9}$$

Then, the reduced-rank VAR matrix is given by

$$\mathbf{B} = \mathbf{\Sigma}_1\mathbf{\Gamma}\mathbf{\Gamma}^\top \tag{10}$$

where $\mathbf{\Gamma} = [\gamma_1, \ldots, \gamma_r]^\top$. Furthermore, the covariance matrix of the residuals is given by

$$\mathbf{\Omega} = \mathbf{\Sigma}_0 - \mathbf{B}\,\mathbf{\Sigma}_0\,\mathbf{B}^\top \tag{11}$$

We employ the Sequence-Space Jacobian (SSJ) solution method of Auclert et al. (2021a) in both applications below. That the solution provides a moving average representation of the observables means that we can rapidly compute population covariances and therefore reduced-rank population projection coefficients.

After computing $\mathbf{B}$ and $\mathbf{\Omega}$, the likelihood computation is straightforward, given by

$$f(\mathbf{y}_1, \ldots, \mathbf{y}_{T+1}) = \sum_{t=1}^{T} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\det(\mathbf{\Omega}) - \frac{1}{2}(\mathbf{y}_t - \mathbf{B}\,\mathbf{y}_{t-1})^\top \mathbf{\Omega}^{-1}(\mathbf{y}_t - \mathbf{B}\,\mathbf{y}_{t-1}) \tag{12}$$

3.5 The number of factors

A natural question in our strategy is the choice of the number of factors in the linear state-space model representation. In section 3.2 we described a heuristic test of computing principle components on data simulated form the model. In this section, we propose other heuristic and quantitative procedures that offers insights into the appropriate number of factors. In our experience, we have found that our algorithm scales well with r, so one might choose to err on the side of caution and choose a large r rather than a small one.

Check 1. Bai and Ng (2002) show that consistent estimation of the number of factors in the data can be attained by minimizing the information criterion[10]

$$IC(n) = V(n) + n\left(\frac{M+T}{MT}\right)\log\left(\frac{MT}{M+T}\right) \tag{13}$$

where $V(n) = (MT)^{-1}\sum_{i=1}^{M}\sum_{t=1}^{T}(a_{it}^n)^2$.

Check 2. Another approach is to simulate the cross-sectional data from $\mathcal{M}(\theta)$ and study the decay rate of the associated singular values. This approach is a common heuristic used by Dynamic Mode Decomposition (DMD) practitioners[11]. An appropriate choice of r is the number after which the decay rate of the singular values is close to zero.

[10] Though the theoretical analysis in Bai and Ng (2002) is done for principle components estimation of factor models, the same theory applies to any other consistent estimation procedure, as $M, T \to \infty$.

[11] See, for example, Brunton and Kutz (2022, sec. 7.2))

Check 3. An additional implication of Proposition 1 is that the VAR residuals $\widetilde{\mathbf{a}}_t$ are serially uncorrelated. While it follows from the innovations representation of the state-space system (1), there is nothing in the first-order VAR that imposes such a restriction. Simulate the data from the model and compute the reduced-rank VAR residuals

$$\tilde{\mathbf{a}}_t = \widetilde{\mathbf{y}}_t - \mathbf{B}_r \, \widetilde{\mathbf{y}}_{t-1} \tag{14}$$

Construct the sample covariance matrix (15) and choose the r that minimizes the Frobenius norm of the covariance matrix of the residuals.

$$\widehat{E}[\mathbf{a}_{t+1} \, \mathbf{a}_t^\top] = \frac{1}{T} \sum_{t=1}^{T} \widetilde{\mathbf{a}}_{t+1} \widetilde{\mathbf{a}}_t^\top \tag{15}$$

Check 4. The final check is to compute a metric for the "fit" of the VAR to the data generated by the model, akin to a modified R^2. Simulate data from the model. For a fixed r, compute the VAR residuals in equation (14).

Denote $R^2_{i,r}$ as the individual-level R^2 for the VAR regression for $i = 1, \ldots, N$ (i.e. for each row of $\widetilde{\mathbf{y}}_t$), given by

$$R^2_{i,r} = 1 - \frac{\sum_{t=2}^{T} \tilde{a}^2_{i,t}}{\sum_{t=2}^{T} \widetilde{\mathbf{y}}_{i,t} - \frac{1}{T} \sum_{t=2}^{T} \widetilde{\mathbf{y}}_{i,t-1}}$$

where $\widetilde{a}_{i,t}$ is the i-th element of $\widetilde{\mathbf{a}}_t$. Then, calculate the aggregate R^2_r of the approximating model by a weighted sum of the individual R^2_i for $i = 1, \ldots, N$.

$$R^2_r = \tfrac{1}{N} \sum_{i=1}^{N} w(i) R^2_{i,r} \tag{16}$$

where $w(m)$ is a weighting function chosen by the researcher.[12] The appropriate r is the smallest value that maximizes R^2_r.

3.6 Algorithm

The above sections set out the theoretical and computational arguments for our estimation strategy. To recap succinctly, the logic is as follows: Assumption 1 means that $\mathcal{M}_{\theta_0}(\mathbf{y}_t)$ has a linear state-space representation with $r \ll N$ factors. Using Assumption 2, we show that a rank-r first-order VAR

[12] In our example below, we choose an equal weighting scheme ($w(i) = 1 \quad i = 1, \ldots, N$), and sample individuals from the stationary distribution of $\mathcal{M}$.

approximates the linear state-space model $\mathcal{M}_{\theta_0}(\mathbf{y}_t)$ when N is large, and that the approximation error vanishes at rate N. One might compute the likelihood using either the canonical correlations method, or the DMD method, depending on the application. We also provide diagnostics that help with choosing an appropriate value of r.

Algorithm 1 presents pseudo-code for approximating $\mathcal{M}_\theta(\mathbf{y}_1, \ldots, \mathbf{y}_{T+1})$ for any θ.

Algorithm 1 Approximation of $\mathcal{M}_\theta(\mathbf{y}_1, \ldots, \mathbf{y}_{T+1})$

1. Fix some structural parameters θ
2. Choose the rank, r, as discussed in section 3.5
3. If population covariances are available analytically:
   - Compute $\mathbf{B}_r(\theta)$ and $\mathbf{\Omega}_r(\theta)$ using the equations (10) and (11)
4. If not:
   - Simulate time-series $\tilde{\mathbf{y}}_1(\theta), \ldots, \tilde{\mathbf{y}}_{J+1}(\theta)$ from $\mathcal{M}$ for a large J and create data matrices $\widetilde{\mathbf{Y}}(\theta)$ and $\widetilde{\mathbf{Y}}'(\theta)$[13]
   - Calculate $\mathbf{B}_r(\theta)$ and $\mathbf{\Omega}_r(\theta)$ in (B.2) and (B.3)
5. Approximate the log- likelihood $f(\mathbf{y}_1, \ldots, \mathbf{y}_{T+1}\,|\theta)$ implied by $\mathcal{M}_\theta(\mathbf{y}_t)$ by computing (12)

3.7 Bayesian estimation

Our likelihood approximation algorithm can be easily paired with an optimization algorithm for maximum likelihood estimation, or any Monte-Carlo sampling scheme to approximate the posterior distribution of the parameters. Under Assumptions 1 and 2, our theorems imply that our approximation error of posterior also vanishes as the cross-section becomes large ($N \to \infty$). The application in Section 5, estimates the model with a Random Walk Metropolis Hastings algorithm.

Algorithm 2 provides a pseudo-code for one iteration of the RWMH sampler.

3.8 Difficulties with other possible approaches

Kalman filter An obvious question at this point is why not simply evaluate the likelihood of the low-rank linear state-space representation (1) with the Kalman filter? The answer is that evaluating this likelihood via the Kalman filter requires knowing the matrices $\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}$, which itself must be estimated on the data simulated from the structural model. Then for every iteration over θ, one needs to estimate $\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}$ and then use them to compute the likelihood of the data. Even for a

Algorithm 2 One iteration of Random Walk Metropolis Hastings

For iteration n with structural parameter θ^{n-1}:

1. Draw $\theta^* \sim q(\cdot|\theta^{n-1})$
2. Approximate likelihood $f(\{\mathbf{y}_t\}|\theta^*)$ using Algorithm 1
3. Compute $r = \min\left\{1, \frac{f(\mathbf{y}_1,\ldots,\mathbf{y}_{T+1}|\theta^*)p(\theta^*)}{f(\mathbf{Y}|\theta^{n-1})p(\theta^{n-1})}\right\}$
4. Accept θ^* with probability r
5. **if** accept, $\theta^n = \theta^*$, **else** $\theta^n = \theta^{n-1}$

small N, this is computationally inefficient.

What about evaluating the likelihood of the true high-rank linear state-space representation of the structural model? This is also computationally demanding because r is high (even with dimension reduction) and the cost of Kalman filtering scales cubically in r. Moreover, common dimension reduction techniques such as those employed by Bayer et al. (2024) are designed to approximate the model's *aggregate* dynamics, and their accuracy in capturing the underlying *distributional* dynamics remains untested.

Moving average representation The sequence-space Jacobian solution method of Auclert et al. (2021a) obtains a moving average representation of the model. Computing the likelihood requires vectorizing the $N \times T$ data matrix, so the associated covariance matrix is $NT \times NT$. In conventional situations with only aggregate data, N is relatively small (typically less than ten), so the covariance matrix is a computationally manageable. For our relevant case when N is large, the inversion of the covariance matrix is computationally costly.

Whittle Approximation (Frequency domain) An alternative to computing the likelihood of the moving average representation using the above approach of Auclert et al. (2021a) is to compute the likelihood in the frequency-domain using the Whittle approximation, as in Hansen and Sargent (1981), Christiano and Vigfusson (2003), and Plagborg-Møller (2019). The Whittle approximation decomposes the entire $N \times T$-dimensional variance-covariance matrix into the sum of N-dimensional frequency-specific matrices. Using Fast Fourier Transform, the decomposition and associated likelihood evaluation is fast. However, the approximation error depends on T, the number of observations in the time dimension. Given the current availability of micro data, this dependence seems restrictive. This is unlike our method where approximation quality depends on N, which

seems like a far more satisfiable requirement currently. In Appendix C.2, we compare our method with the FD estimation for both bias and efficiency via a Monte-Carlo exercise.

4 Laboratory exercise: Krusell-Smith economy with redistributive shocks

4.1 Model description

The economy is populated by heterogeneous households of mass 1 that maximize their infinite-lifetime utility $\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t\frac{c^{1-\phi}}{1-\phi}\right]$. Each household have two idiosyncratic states: productivity ε and assets a. We assume that productivity follows an exogenous Markov chain, where the probability of moving from ε to ε' is given by $\pi(\varepsilon,\varepsilon')$. The unconditional mean of the productivity is set equal 1. All households work an exogenous amount of hours n, normalized to equal 1. Household labor income is given by $w_t\frac{\varepsilon_{it}^{\xi_t}}{\int\varepsilon_{jt}^{\xi_t}dj}$, where w_t denotes the hourly efficiency wage at time t, common to all households. The aggregate shock ξ_t is a income risk shock that dictates the cross-sectional income dispersion.[14] We assume that $\log\xi_t$ follows an normal AR(1) process with auto-correlation ρ_ξ and standard deviation of innovation σ_ξ. Households can choose to save in assets a, which have a rate of return R. Households also face a borrowing constraint $a\geq 0$.

Output in this economy is produced by a representative firm that takes aggregate capital K_t and labor L_t as inputs using a Cobb-Douglas production function $Y_t=Z_tK_t^{\alpha}L_t^{1-\alpha}$, where Z_t is aggregate total factor productivity. One unit of labor costs the firm w_t, and one unit of capital costs the firm R_t. We assume that the (log) TFP follows an normal AR(1) process with auto-correlation ρ_z and standard deviation of innovation σ_z.

Let $\Lambda_t(a,\varepsilon)$ denote the mass of households with assets a and idiosyncratic productivity ε. Households need to forecast prices R and w that depend on both aggregate productivity and the distribution of assets in the economy. Thus, the household has two idiosyncratic (a,ε) and two aggregate states (Z_t,Λ_t). Let $V(a,\varepsilon;Z_t,\Lambda_t)$ be the value function of household with assets a, productivity ε and that faces aggregate productivity Z_t and the aggregate distribution of wealth Λ_t. The household's problem is given by

[14] Although by construction the shock does not affect the average productivity, it has first-order aggregate effect through the precautionary saving motive of households. Thus, aggregate data alone is informative for this shock.

$$V(a,\varepsilon;Z_t,\Lambda_t) = \max_{\{c,a'\}} \frac{c^{1-\phi}}{1-\phi} + \beta \int V(a',\varepsilon';Z_{t+1},\Lambda_{t+1})\pi(\varepsilon'|\varepsilon)d\varepsilon' dF(Z_{t+1}|Z_t)$$
$$c(a,\varepsilon;Z_t,\Lambda_t) + a'(a,\varepsilon;Z_t,\Lambda_t) = y(\varepsilon_t;Z_t,\Lambda_t) + R(Z_t,\Lambda_t)a$$
$$y(\varepsilon;Z_t,\Lambda_t) = w(Z_t,\Lambda_t)\varepsilon$$
$$\Lambda_{t+1} = \Psi(Z_t,\Lambda_t)$$

where Ψ denotes the law of motion of the endogenous distribution of households, and where we make explicit the dependence on aggregate states.

Market clearing implies that

1. the goods market clears

$$\int c(a,\varepsilon;Z_t,\Lambda_t)\Lambda_t(a,\varepsilon)dad\varepsilon + \int a'(a,\varepsilon;Z_t,\Lambda_t)\Lambda_t(a,\varepsilon)dad\varepsilon = Y_t + (1-\delta)K_t \tag{17}$$

2. the labor market clears

$$L_t = \int \varepsilon\Lambda_t(a,\varepsilon)dad\varepsilon = 1 \tag{18}$$

3. the asset market clears

$$K_t = \int a\Lambda_t(a,\varepsilon)dad\varepsilon \tag{19}$$

Our goal is to estimate the shock parameters of this model, $(\rho_z, \sigma_z, \rho_\xi, \sigma_\xi)$ using both aggregate and data. In the next two sections, we outline how we organize the thousands of micro datapoints, as well as our theoretical results that underpin our algorithm to perform such an estimation.

4.2 Organizing the cross-section data

The household's consumption policy function, similar to the value function, depends on the two idiosyncratic states and the two aggregate states. Since the aggregate states are common for all households, we suppress its notation by writing the consumption policy function at time t, $c_t(a,\varepsilon) \equiv c(a,\varepsilon;Z_t,\Lambda_t)$. By the organizing principle in Section 3.1, we bin individuals according to their idiosyncratic states in each period, which for this model is assets and idiosyncratic productivity. This requires discretizing the state-space into N points of the $a \times \varepsilon$ state space; and grouping

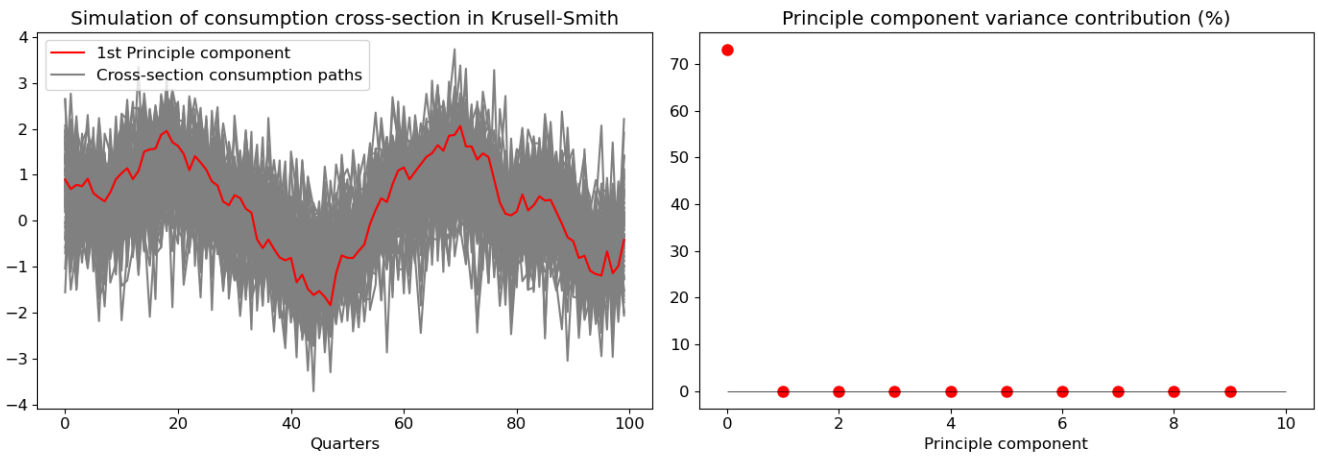

Figure 1: Evolution of cross-section of consumption in a simulated economy

individuals into associated bins. Consumption of the bin (a, z) as the mean consumption of individuals inside that bin. Repeating this for each N bins, we obtain our micro data vector at time $t - \mathbf{y}_t = [c_t(\mathbf{z}_1), c_t(\mathbf{z}_2), \ldots, c_t(\mathbf{z}_N)]^\top$. Finally, we create our data matrix, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T] \in \mathbf{R}^{N \times T}$.

4.3 Connecting our theory

The main assumption underpinning our algorithm is that the equilibrium model has a Gaussian linear state-space representation with relatively few states. Notwithstanding the rank condition, this statement is uncontroversial in the standard heterogeneous-agent models. Figure 1 validates this the low-rank assumption, by plotting simulated paths of consumption for 1000 households over 120 quarters.[15] Although there are idiosyncratic divergences, the evolving cross-section appears to exhibit a common factor. This observation is confirmed by computing principle components of the data. The red line in the left chart plots the first principle component of the data, and the right chart plots the contribution to total variance of the 10 leading principle components. It again suggests the presence of one dominant common factor, $r = 1$. That $N = 1000$ in this setting implies that our $N \gg r$ assumption is satisfied.[16]

Under Assumptions 1 and 2, Corollary 1 states that as the size of the cross-section goes large, $\mathbb{E}[\mathbf{x}_t \,|\, \mathbf{y}^t] \approx \mathbb{E}[\mathbf{x}_t \,|\, \mathbf{y}_t]$. This implies that the hidden factor can be approximately well inferred from contemporaneous data compared to its infinite past. To see the theorem at work in our laboratory, conduct the following experiment. We simulate an economy with 1000 individuals from the structural model above. We imagine an econometrician inferring hidden states by only observing

[15] See Appendix D.2 for the calibration underlying this exercise.
[16] Appendix E.2 performs the battery of checks outlined in Section 3.5, providing further evidence that $r = 1$.

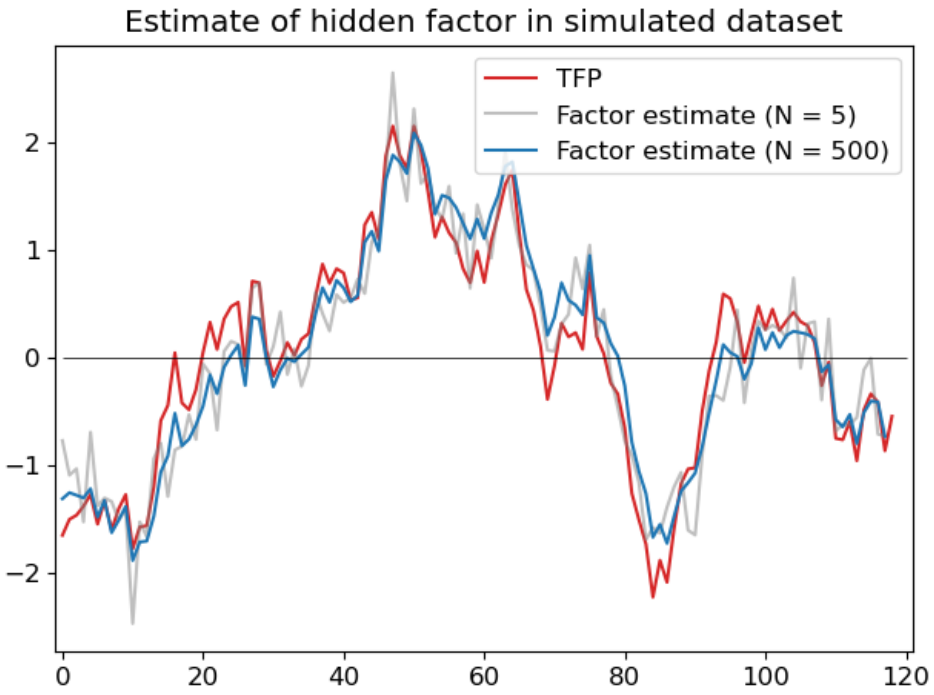

Figure 2: Estimate of hidden factor in simulated dataset

consumption of N individuals. The econometrician then computes principle components of the observed data.[17] Figure 2 shows the estimated hidden factors for $N = 5$ (gray) and $N = 500$ (blue), and the red line is the TFP series generating the data. All series are normalized to unit variance. Indeed the econometrician's estimate of the hidden factor closely resembles TFP when N is large.

4.4 Monte-carlo simulation

We use the above model as a laboratory to test-drive our theory in a controlled environment. Calibrating the model as in Table D.2, we simulate consumption paths of 300 households drawn from the stationary distribution, as well as aggregate output, nominal interest rate and aggregate consumption for 120 quarters. We choose these dimensions to closely replicate what researchers might face in practice. We employ our algorithm to compute maximum likelihood estimates of the parameters of the shock processes in an effort to recover the values that generated the data. We repeat this exercise 500 times to approximate the finite-sample distribution of the maximum likelihood estimator. Table 1 shows the mean and standard deviation of the finite-sample distributions, and compares them against two alternatives common in the literature: 1) using only macro

[17] Importantly for the validity of our exercise, principle components does not assume any particular dynamic process for the data, and so only uses concurrent data to estimate factors.

data – output, consumption, interest rate – called `Macro` and 2) macro data *and* the second moment of the consumption distribution, called `Macro+`. We call the estimation with the macro data *and* micro data `Macro+Micro`. The values colored in red depict mean the estimates that are closest to the true value, and the values in blue denote estimates that have the smallest standard deviation. Apart from one case, the `Macro+Micro` estimates have mean estimates closest to the truth, with the smallest dispersion.

Figure 3 plots the histogram of the distributions with `Macro` in gray, `Macro+` in blue and `Macro+Micro` in red. In the top panel is those for ρ_z and σ_z. The finite-sample distributions look similar across all three estimations, suggesting that there isn't much to be learned from the additional cross-section data in identifying the stochastic process for TFP. The story is much different for the parameters of the redistribution shock (bottom panel). Estimating the model with aggregate data on its own leads to severely biased and dispersed parameter estimates. For example, the estimates for ρ_z ranges from 0.5 to 1.0, with very little mass around the true estimate.[18] Adding the variance in consumption (`Macro+`) improves the bias and dispersion of the estimates. Introducing the cross-section (`Macro+Micro`) dramatically reduces the bias and the dispersion in the estimates, leading to more efficient and accurate inference.[19]

To see how our method exploits the information contained in the micro-data, consider an individual with idiosyncratic states $\mathbf{z}_j$.[20] Through the lens of the model, the change in consumption between t and $t+1$ is caused can only be due to changes in Z_t, ξ_t, Λ_t. The difference in consumption response between individuals j and k, $\forall j \neq k$ provides information about the structural parameters. This is especially ρ_ξ and σ_ξ, which determine the dynamics of post-tax income for each individual.

While the qualitative result – that cross-section data is more informative of parameters that affect the cross-section – is perhaps unsurprising, the magnitude of improvement even in this very simple model is noteworthy. It is possible that in models with much richer interactions between the cross-sections and aggregates, the addition of the cross-section may also help in more accurately estimating other aggregate parameters than would otherwise be the case.

This section studied the impact of including micro data in a controlled environment and found an improvement in bias and efficiency over conventional approaches. This is especially true for

[18] The (gray) histogram bunches at 0.5 and 1.0 for ρ_z, which are the bounds we imposed for for the optimizer. The estimates would be even more dispersed were we to allow for looser bounds.

[19] That marginal propensities to consume (MPCs) differ across households in the economy implies that ξ_t is formally identified in this model – a redistribution of income from high to low productivity households itself generates an aggregate response. In that sense, estimating the model on only aggregate data is not a "strawman".

[20] By our organization of the data, the time-series of their consumption is in the jth row of $\mathbf{Y}$.

			Macro		Macro+		Macro+Micro	
Shock	Parameter	True value	Mean	Std	Mean	Std	Mean	Std
TFP	ρ_z	0.95	0.95	0.005	0.95	0.004	0.95	0.005
	σ_z	0.5	0.52	0.04	0.51	0.04	0.50	0.03
Redistribution	ρ_ξ	0.8	0.72	0.21	0.77	0.1	0.80	0.02
	σ_ξ	0.3	0.07	0.08	0.31	0.05	0.30	0.03

Table 1: Finite-sample parameter distribution with 500 Monte-Carlo samples

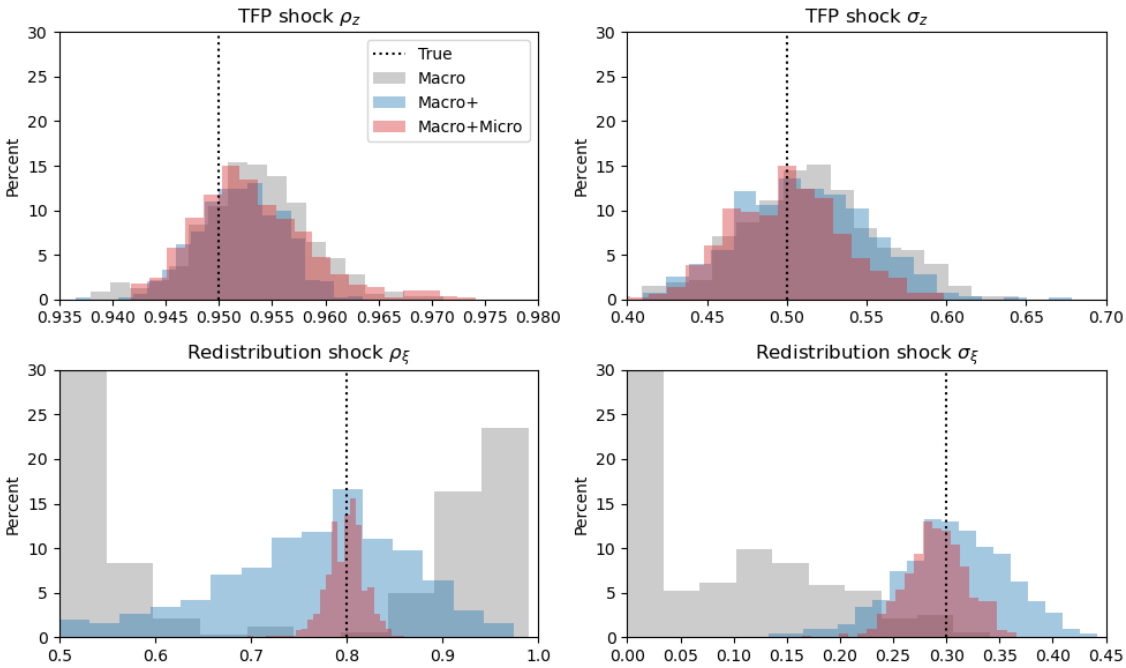

Figure 3: Finite sample parameter distribution across 500 Monte-Carlo samples.
Back to Introduction

parameters associated with redistribution. Next, we turn to an empirical application in which we estimate the parameters of a medium scale HANK model.

5 Application: Medium-scale HANK with earnings exposure heterogeneity

We consider a medium-scale heterogeneous agent model, with aggregate shocks, following closely Bayer et al. (2024). As has been well-documented in the previous literature, heterogeneous exposures to aggregate shocks are a major source of difference between HANK models and RANK model. This model features two important types, in a similar spirit to Alvez et al. (2020), the parameters of which we will estimate using detailed micro data.

5.1 Model

Time is discrete and runs forever, $t = 0, 1, \ldots$.

5.1.1 Households

There is a unit measure of infinitely-lived households in the economy. At time t, a household consumes c_t and supplies hours of labor h_t, chosen by a labor union. They receive per period utility

$$U(c_t, h_t) = \frac{c^{1-\sigma}}{1-\sigma} - \varphi \frac{h_t^{1+\phi}}{1+\phi} \tag{20}$$

where σ is the inverse of the EIS and ϕ is the inverse of the Frisch elasticity of labor supply. A household earns three kinds of income: labor, dividend and asset income. Labor earnings $y_t(z)$ and dividend income $D_t(z)$ depend on the household's stochastic labor productivity z and are given by

$$y_t(z) = w_t h_t \Gamma_t(z) e^{v_t^r} \tag{21}$$

$$D_t(z) = D_t z \tag{22}$$

where w_t is the real wage and D_t is the aggregate dividend in the economy and v_t^r is a monetary policy shock. We normalize productivity to integrate to one. The function $\Gamma_t(z)$, called the "incidence function", will play a central role in our analysis. We consider equation (21) as a generalization of the standard setup in which $\Gamma_t(z) = z$ and $y_t(z) = w_t h_t z$, where $h_t z$ is interpreted as *efficiency hours*. In such settings, z plays a dual role. On the one hand, it encodes the fact that more productive households have higher earnings on average (rationalizing the interpretation of efficiency hours); on the other, it implies that they are more sensitive to changes in $w_t h_t$ (aggregate earnings) and therefore experience more volatile income over the business cycle. We follow Alves et al. (2020) by replacing z with $\Gamma_t(z)$ and ensure it also integrates to one. But unlike Alves et al. (2020) who estimate the parameters of Γ_t using external regressions on micro-data, we will estimate the parameters of Γ_t jointly with the other structural parameters of the model using cross-sectional data, which is made feasible by our likelihood approximation algorithm. See Section 5.2 for the parameterization of Γ_t.

In the standard setup, household income reacts to monetary policy shocks v_t^r through its effect on $w_t h_t$ and the $\Gamma_t(z)$. Here we also allow for the direct effect of monetary policy on income. Though this may seem unnatural at first, we interpret this as a simple way to express sectoral employment and their heterogeneous exposures to interest rates. Coibion et al. (2017) provide

some empirical evidence for the presence of this channel. As described below our specification is flexible enough to allow the data, through the estimation, to reject this importance of this channel. Nevertheless, we stress that the important point for this paper is that our method provides a convenient and fast way to estimate micro parameters such as these.

Post-tax earnings of the household take the Heathcote et al. (2017) form of $(1-\tau_t^y)y_t(z)^{1-\xi}$ where τ_t^y determines the average level of the marginal tax rate and ξ determines its slope. Dividends are also taxed at rate τ_t^D. Finally, households can save and borrow through a risk-free asset, subject to an ad-hoc borrowing constraint $a_t \geq \underline{a}$, which earns the real interest rate r_t. Let $V_t(a,z)$ denote the value function of a household with assets a and productivity z. Given the above setup, its Bellman equation is given by

$$V_t(a,z) = \max_{c,a'} \left\{ \frac{c^{1-\sigma}}{1-\sigma} - \varphi \frac{h_t^{1+\phi}}{1+\phi} + \beta \mathbb{E}_t \left[V_{t+1}(a',z') | z \right] \right\} \tag{23}$$

$$c + a' = (1-\tau_t^y) y_t(z)^{1-\xi} + (1-\tau_t^D) D_t z + (1+r_t) a \tag{24}$$

$$y_t(z) = w_t h_t \Gamma_t^y(z) \tag{25}$$

$$a' \geq \underline{a}$$

5.1.2 Firms

Labor unions We closely follow Auclert et al. (2024) in what has become a standard setup in the New Keynesian sticky wage literature. We assume that labor hours h_{it} are determined by the labor demand of unions, a continuum of which operate in a monopolistically competitive market. Each union k aggregates the efficient hours of its workers to a union-specific task

$$n_{kt} = \int h_{ikt} \Gamma_t^y(z_i) di \tag{26}$$

A competitive labor packer aggregates labor hours across the unions with constant elasticity of substitution ϵ_w

$$N_t = \left(\int n_{kt}^{\frac{\epsilon_w - 1}{\epsilon_w}} dj \right)^{\frac{\epsilon_w}{\epsilon_w - 1}} \tag{27}$$

and then sells it to intermediate goods firms for real wage w_t. Each union k sets a common real wage w_{kt} amongst all its members subject to a quadratic adjustment cost à la Rotemberg (1982) on nominal wages. We restrict the union to a uniform labor allocation rule, i.e. $n_{ikt} = n_{kt}$. This implies that the union sets the real wage w_{kt} to maximize the average utility of its members. The union's

problem is therefore given by

$$\max_{w_{kt+l},n_{kt+l}} \mathbb{E}_t \sum_{k=0}^{\infty} \beta^l \left\{ \left[(C_{t+l})^{-\sigma}(1-\tau_t^y) w_{kt+l} - \varphi N_{t+l}^{\phi} \right] n_{kt+l} - \frac{\epsilon_w}{2\kappa_w} \log\left(\frac{w_{kt+l}}{w_{kt+l-1}} \pi_{t+l} \right)^2 \right\}$$

$$s.t. \quad n_{kt+l} = \left(\frac{w_{kt+l}}{w_{t+l}} \right)^{-\epsilon} N_{t+l}$$

$$n_{kt+l} = \int h_{kt+l,i} \Gamma_t^y(z_i)\, di$$

$$N_{t+k} = \left(\int n_{kt+l}^{\frac{\epsilon_w - 1}{\epsilon_w}}\, dk \right)^{\frac{\epsilon_w}{\epsilon_w - 1}}$$

where the union takes as given the packer demand as a function of its relative real wage $\frac{w_{kt}}{w_t}$. This setup implies that all unions set the same real wage, i.e. $w_{kt} = w_t$ and all households are demanded the same efficient hours, i.e. $n_{kt} = N_t$, and so the first order condition of the labour union leads to the wage Phillips curve

$$\pi_t^w = \kappa_w \left[\varphi N_t^{\phi} - \frac{\epsilon_w - 1}{\epsilon_w} (1-\tau_t^y)(C_t)^{-\sigma} w_t \right] N_t + \beta\, \mathbb{E}_t\, \pi_{t+1}^w + v_t^w \tag{28}$$

where v_t^w is a AR(1) wage-markup shock.

Final-good producers A perfectly competitive representative final-good producer aggregates the continuum of retail firms $j \in (0,1)$ with constant elasticity of substitution ϵ

$$Y_t = \left(\int_0^1 y_{jt}^{\frac{\epsilon-1}{\epsilon}} dj \right)^{\frac{\epsilon}{\epsilon-1}}$$

Taking prices as given, the demand for intermediate good j is given by

$$y_{jt} = \left(\frac{p_{jt}}{P_t} \right) Y_t, \quad \text{where} \quad P_t = \left(\int_0^1 p_{jt}^{1-\epsilon} dj \right)^{\frac{1}{1-\epsilon}} \tag{29}$$

Intermediate-good firms We follow closely a standard specification of intermediate-good firms. Intermediate-good firms operate in a monopolistically competitive market. Each differentiated good j is produced with a Cobb-Douglas production function

$$y_{jt} = Z_t k_{jt}^{\alpha} l_{jt}^{1-\alpha} \tag{30}$$

Firms hire capital and labor at prices r_t^k and w_t and pay a fixed cost Ξ. In choosing prices p_{jt}, they take into account their demand by the final good firm and Rotemberg (1982) quadratic price

adjustment costs Ψ_{jt}^p on their price inflation relative to a fraction ι_p of the price changes last period.

The Bellman equation for the retail firms is consequently

$$J_t^R(p_{jt-1}, p_{jt-2}) = \max_{k_{jt}, l_{jt}, y_{jt}, p_{jt}} \left\{ \frac{p_{jt}}{P_t} y_{jt} - w_t n_{jt} - r_t^k k_{jt} - \Psi_{jt}^p - \Xi + \mathbb{E}_t \left[\frac{J_{t+1}^R(p_{jt}, p_{jt-1})}{1 + r_t} \right] \right\}$$

subject to

$$\begin{aligned} y_{jt} &= Z_t k_{jt}^{\alpha} l_{jt}^{1-\alpha} \\ y_{jt} &= \left(\frac{p_{jt}}{P_t} \right)^{-\epsilon} Y_t \\ \Psi_{jt}^p &= \frac{\psi_p}{2} \left[\log \left(\frac{p_{jt}}{p_{jt-1}} \right) - \iota_p \log \left(\frac{p_{jt-1}}{p_{jt-2}} \right) \right]^2 Y_t \end{aligned}$$

The first order condition of the firms imply constant capital-labor ratios

$$\frac{(1-\alpha)}{\alpha} \frac{r_t^k}{w_t} = \frac{N_t}{K_t}$$

In a symmetric equilibrium where all firms choose $y_{jt} = y_t, k_{jt} = k_t = K_t$ and $l_{jt} = l_t = N_t$ implies identical marginal costs given by

$$mc_t = \frac{1}{Z_t} \left(\frac{r_t^k}{\alpha} \right)^{\alpha} \left(\frac{w_t}{1-\alpha} \right)^{1-\alpha}$$

giving rise to the aggregate production function

$$Y_t = Z_t K_t^{\alpha} N_t^{1-\alpha} \tag{31}$$

and the aggregate price Phillips curve with associated price markup shock v_t^p that follows an AR(1) process.

$$\pi_t - \iota_p \pi_{t-1} = \kappa_p \left(mc_t - \frac{\epsilon - 1}{\epsilon} \right) + \frac{1}{1 + r_t} \mathbb{E}_t \left[\frac{Y_{t+1}}{Y_t} (\pi_{t+1} - \iota_p \pi_t) \right] + v_t^p \tag{32}$$

Furthermore, aggregate resources spent of adjustments costs are

$$\Psi_t^p = \frac{\psi_p}{2} \left[\log (\pi_t) - \iota_p \log (\pi_{t-1}) \right]^2 Y_t$$

and profits from the intermediate firms, distributed as dividends are given by:

$$D_t^R = Y_t - w_t N_t - r_t^k K_t - \Psi_t^p$$

Capital good firms A representative firms owns the capital stock and rents it to the intermediate good producers at rate r_t^k. The firm chooses investment I_t and capital tomorrow K_{t+1} subject to the capital law of motion. The Bellman equation is given by

$$J_t^K(K_t, I_{t-1}) = \max_{K_t, I_t} \left\{ r_t^k K_t - I_t + E_t \left[\frac{J_{t+1}^K(K_{t+1}, I_t)}{1 + r_t} \right] \right\} \tag{33}$$

subject to

$$K_{t+1} = (1-\delta)K_t + \mu_t e^{v_t^\mu} \left[1 - S\left(\frac{I_t}{I_{t-1}} \right) \right] I_t \tag{34}$$

where $\delta \in (0,1)$ is the depreciation rate of capital, and μ_t is the marginal efficient of investment as in Justiniano et al. (2011), and $S(\cdot)$ is a convex function that satisfies $S(1) = S'(1) = 0$ and $S''(1) = \psi_i$. We will set $S(x) = \frac{\psi_i}{2}(x-1)^2$. Defining Tobin's Q as the marginal value of capital at time t, $\frac{\partial_K J_{t+1}(K_t, I_t)}{1+r_t}$, the dynamics of investment are characterized by

$$Q_t = \frac{r_{t+1}^k + E_t[Q_{t+1}(1-\delta)]}{1 + r_{t+1}} \tag{35}$$

Finally, the capital good firm makes profits, distributed as dividends

$$D_t^K = r_t^K K_{t-1} - I_t \tag{36}$$

5.1.3 Government

The fiscal authority issues one-period real bonds B_t, conduct government spending G_t and collects taxes T_t. It is bound by the budget constraint

$$B_{t+1} + T_t = (1 + r_t)B_t + G_t \tag{37}$$

Given a particular income tax rate τ_t^y, tax revenues are the sum of labor and dividend taxes

$$T_t = w_t N_t - \int (1 - \tau_t^t) y(z)^{1-\xi} dz + \tau_{ss}^D D_t \tag{38}$$

The tax rate τ_t^y is chosen according to a rule that prevents large swings in the tax rate but ensures that the real government debt remains stationary.

$$T_t = \tau_{ss}^y w_t N_t + \tau_{ss}^D D_t + (1 - \rho_b)(B_t - B_{ss}) \tag{39}$$

Finally, we assume monetary policy follows a smoothed Taylor rule

$$i_t = \rho_r i_{t-1} + (1 - \rho_r)(\phi_\pi \pi_t + r_{ss}) + v_t^r \tag{40}$$

with monetary policy shock v_t^r, while the Fisher equation links the nominal interest rate to the real interest rate

$$(1 + r_t) = \frac{1 + i_{t-1}}{1 + \pi_t} \tag{41}$$

Aggregate shocks There are seven aggregate shocks: TFP shock, monetary policy shock, government spending shock, price markup shock, wage markup shock, labor incidence shock, dividend incidence shock.Each follow an independent AR(1) process.

$$\begin{aligned}
\log Z_t &= \rho_Z \log Z_{t-1} + \sigma_\Theta \epsilon_t^Z \\
v_t^r &= \rho_r v_{t-1}^r + \sigma_r \epsilon_t^r \\
G_t &= \rho_g v_{t-1}^G + \sigma_g \epsilon_t^g \\
v_t^p &= \rho_p v_{t-1}^p + \sigma_p \epsilon_t^p \\
v_t^w &= \rho_w v_{t-1}^w + \sigma_w \epsilon_t^w \\
v_t^\mu &= \rho_\mu v_{t-1}^\mu + \sigma_\mu \epsilon_t^\mu
\end{aligned}$$

Equilibrium The aggregate resource constraint is

$$Y_t - \Psi_t^p - \Xi = C_t + G_t + I_t \tag{42}$$

The asset market clearing is

$$A_t = B_t \tag{43}$$

Labor market clearing requires

$$h_t = N_t \tag{44}$$

A *rational expectation equilibrium* consists of a sequence of policy functions $\{c_t, a_t, h_t\}$, a sequence of value functions $\{V_t\}$, a sequence of prices $\{w_t, r_t^k, \pi_t, \pi_t^w, \tau^y, r_t, i_t\}$, a sequence of aggregate objects $\{Y_t, C_t, K_t, N_t, \Psi_t^p, D_t, I_t, Q_t, B_t, T_t\}$, a sequence of distribution $\{F_t\}$, a sequence of exogenous states $\{\Theta_t, v_t^r, v_t^g, v_t^p, v_t^w, v_t^\mu\}$, and a sequence of beliefs over prices such that

1. Given the sequence of value functions, prices, and policy functions, the household Bellman equation holds.
2. Given the sequence of beliefs over prices, all agents optimize.
3. The evolution of the distribution is consistent with the policy.
4. All markets clear.

5.2 Incidence functions

A differentiating aspect of this model is our specification of the incidence function $\Gamma_t^y(z)$, that encodes differences in sensitivity to aggregate income across the earnings distribution. We define the incidence function

$$\Gamma_t^y(z) = \frac{z\left(\frac{w_t N_t}{w_{ss} N_{ss}}\right)^{\gamma_y(z)} \left(e^{v_t^r}\right)^{-\gamma_r(z)}}{\mathbb{E}_z\left[z\left(\frac{w_t N_t}{w_{ss} N_{ss}}\right)^{\gamma_y(z)} \left(e^{v_t^r}\right)^{-\gamma_r(z)}\right]} \tag{45}$$

with the normalizations $\mathbb{E}[z\gamma_y(z)] = \mathbb{E}[z\gamma_r(z)] = \mathbb{E}[z] = 1$. In steady-state, $\Gamma_t^y(z) = z$. Out of steady state, the normalizations imply that $\gamma_y(z)$ represents the elasticity of idiosyncratic income to aggregate income, i.e.

$$\frac{\partial \log y_t(z)}{\partial \log w_t N_t} = \gamma_y(z) \tag{46}$$

and that the elasticity of idiosyncratic income to monetary policy shock is given by

$$\frac{\partial \log y_t(z)}{\partial v_t^r} = \gamma_y(z) \frac{\partial \log w_t N_t}{\partial v_t^r} - \gamma_r(z) \tag{47}$$

We call $\gamma_r(z)$ the *excess elasticity* since it is additional response to a monetary policy shock over

and above the response through aggregate income.[21] We define $\gamma_i(z) = Beta(F_z(z); \alpha^i, 1)$ for $i \in \{y, r\}$, where $F_z(z)$ is the cumulative distribution function for z and $Beta(F_z(z); \alpha, 1) : [0, 1] \to \mathbb{R}_+$ is the beta distribution with parameters $(\alpha, 1)$. Figure 4 plots the income elasticity for different values of α_y. A value $\alpha^y < 1$ determines a downward sloping elasticity as z increases, implying that low income individuals have a higher income sensitivity to aggregate income than high income individuals; and conversely for $\alpha^y > 1$. The value $\alpha^y = 1$ covers the equal-elasticity case across all households. [22, 23]

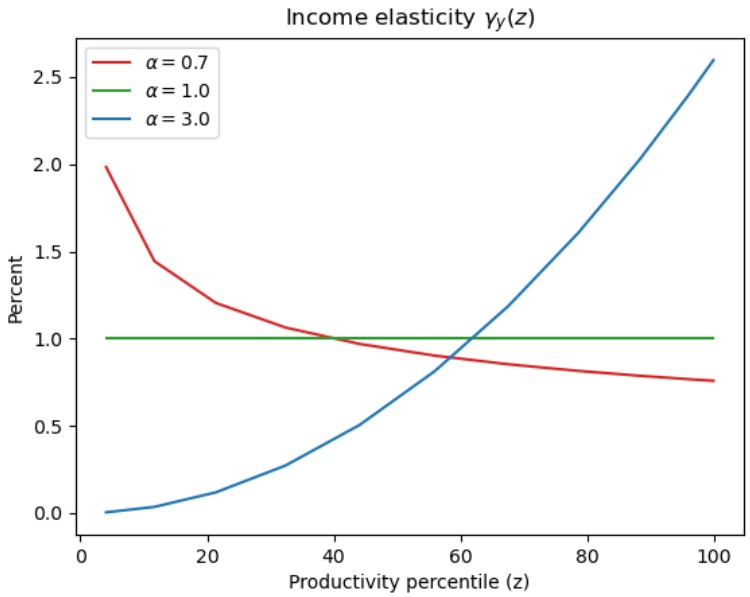

Figure 4: Income elasticity for different values α

5.3 Solution in sequence-space

We employ the sequence-space Jacobian (SSJ) method of Auclert et al. (2021a) to solve for a linearized equilibrium of the model. Although the original SSJ method is developed for computing the aggregate variables, it can be easily extended to obtain a solution for the micro observables. To

[21] See Appendix A.7 for derivation of these relations.

[22] Similarly, $\alpha^r < 1$ implies that the sensitivity to monetary policy is further heightened for low income households, and vice versa for high income households.

[23] A common quantification of this function in the literature is to calibrate $\gamma_y(z)$ by estimating regressions of the form

$$\log y_{it} = \alpha + \beta_i \log Y_t + e_{it} \qquad \forall i \tag{48}$$

See Auclert and Rognlie (2018),Alves et al. (2020) for exampels of this approach. In our model setup however, the contemporaneous relationship between w_t, N_t and v_t^r, implies the above regression will produced biased estimates of β_i. Structural estimation is immune from these criticism since is estimates *all* the parameters jointly.

see this, let $\mathbf{c}_t = (c_{1,t}, \ldots, c_{M,t})^\top$ be the vector of cross-sectional consumption and $\boldsymbol{\epsilon}_t \in \mathbb{R}^S$ be the vector of shocks at time t. We have the following proposition.

Proposition 2. *In the linearized equilibrium, $\mathbf{c}_t$ has a moving average (MA) representation*

$$\mathbf{c}_t = \mathbf{c}_{ss} + \sum_{j=0}^{\infty} \Theta_j^c \boldsymbol{\epsilon}_{t-j} \tag{49}$$

Furthermore, the MA coefficient matrix Θ_j^c is given by

$$\Theta_j^c = \sum_{p \in \mathcal{P}} \mathcal{J}_p^y F^j \mathcal{I}_e^p$$

where $\mathcal{P}$ denotes the set of aggregate inputs that enter the household's problem and

- *$\mathcal{J}_p^c \in \mathbb{R}^{M \times \infty}$ is the cross-section of gradients of consumption wrt. the future path of aggregate input p*
- *$\mathcal{I}_e^p \in \mathbb{R}^{\infty \times S}$ is the impulse response functions of aggregate input p*
- *F is the shift-forward operator*

In light of Proposition 2, simulation of the micro consumption dynamics is straightforward, and computation of the moving average coefficient matrices is trivial because $\mathcal{J}_p^c$ and $\mathcal{I}_e^p$ are products of the SSJ method.[24] To operationalize the algorithm, we truncate the horizon, to $T = 300$. This implies that $\mathcal{J}_p^c \in \mathbb{R}^{M \times T}$, $\mathcal{I}_e^p \in \mathbb{R}^{T \times S}$.

5.4 Population VAR(1) coefficients

Because the SSJ solution method implies a (truncated) moving-average representation of the observables (both aggregate and cross-section), the population covariances are thus available in closed-form, given a θ. We can therefore use the population canonical correlations approach in computing the reduced-rank VAR, as outlined in Section 3.4.

To obtain the VAR(1) covariance matrix, we need the coincident and lagged covariance matrix. This is given by

$$\mathbb{E}[\mathbf{y}_t \mathbf{y}_t^\top] = \sum_{j=0}^{J} \Theta_j^y \boldsymbol{\Sigma} \Theta_j^{y\top} \tag{50}$$

$$\mathbb{E}[\mathbf{y}_t \mathbf{y}_{t-1}^\top] = \sum_{j=0}^{J} \Theta_j^y \boldsymbol{\Sigma}_{-1} \Theta_j^{y\top} \tag{51}$$

[24] The gradients $\mathcal{J}_p^c$ are computed by backward iteration in the first step of the "Fake news algorithm". We refer interested readers to the original paper Auclert et al. (2021b).

where $\mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_S^2)$ and $\mathbf{\Sigma}_{-1}$ is the same matrix but shifted one below the off diagonal

$$\mathbf{\Sigma}_{-1} = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ \sigma_1^2 & 0 & \ldots & 0 & 0 \\ 0 & \sigma_2^2 & \ldots & 0 & \\ \vdots & & \vdots & & \vdots \\ 0 & 0 & \ldots & \sigma_S^2 & 0 \end{bmatrix} \tag{52}$$

Having these population covariances in closed form rids the need for simulating the model, allowing for a much faster and accurate likelihood computation.

5.5 Calibration

The model is in quarterly frequency. Table 2 reports the parameters that we calibrate. In addition, we internally calibrate β to clear the asset market, which has value 0.967. We normalize output in steady state to one, and our calibration implies that labor share is 63% of output.

We estimate other 7 model parameters: $[\kappa_w, \kappa_p, \rho_i, \rho_r, \phi_\pi, \alpha^y, \alpha^{mp}]$, and the AR(1) persistence and innovation variance parameters of the 6 aggregate shocks above. In all, this gives us 19 parameters to estimate.

Parameter	Interpretation	Value	Justification
σ	CRRA	0.5	Standard
ϕ^{-1}	Frisch elasticity	2.0	Standard
α	Capital share	0.33	Standard
τ_d^{ss}	Dividend tax rate	0.2	US tax code
G/Y	Steady-state government spending ratio	0.2	Auclert et al. (2021b)
δ	Capital depreciation rate	0.02	Auclert et al. (2021b)
ψ_i	Capital adjustment cost	1.5	Bardóczy and Guerreiro (2023)
ι_p	Inflation indexation	0.25	Bardóczy and Guerreiro (2023)
$(\epsilon - 1)/\epsilon$	Steady-state price markup	1.06	Auclert et al. (2020)
$(\epsilon_w - 1)/\epsilon_w$	Steady-state wage markup	1.1	Auclert et al. (2021b)
ξ	HSV Tax slope	0.04	Boar and Midrigan (2022)
ι	Lump sum transfer	0.16	Boar and Midrigan (2022)

Table 2: Calibrated parameters

5.6 Data

We estimate our model using individual-level consumption and income data, and macro data, consisting of aggregate output, aggregate consumption, aggregate investment, interest rate, inflation

and nominal wage growth. The organization of the micro data into model-relevant objects is an important step of our methodology, which we describe below.

5.6.1 Micro data

The micro data on consumption and income is sourced from the Interview Survey section of the Consumer Expenditure Survey (CEX), which is a rotating panel of representative households of the U.S. population and provides detailed information about U.S. household expenditures and income. Each household is interviewed for a maximum of four quarters. Expenditure is reported each quarter, income is reported the first and last quarters, and assets is reported only in the last quarter. We use CEX data between 1990 Q1 and 2020 Q1 to avoid the COVID-19 period.

Labor income is defined as wage or salary income (`FSALARYX`) plus self-employed income (`FNONFRMX`). The measure of individual consumption is the sum of expenditures in the current (`TOTEXPCQ`) and past quarter (`TOTEXPPQ`). This is a common adjustment made suggested by the BLS to account for the possibility that even though interviews are conducted once every three months, they might not coincide with calendar quarters. Our notion of assets in the model is liquid assets in the data. Prior to 2013, our measure is equal to the value in checking/brokerage accounts (`CKBKACTX`) plus saving accounts in banks (`SAVACCTX`), credit unions etc. and loads and securities held in mutual funds (`SECESTX`). After 2013, the concept is conveniently aggregated into the total market value in checking, savings, money market accounts, CDs, etc. and other similar accounts (`LIQUIDX`).

We do some preliminary cleaning of the data: we drop households where the principal earner is below 25 or above 65 years old; those where incomes are negative and those whose quarterly consumption are above two times their annual income. Since output in the model is normalized to 1, and the labor share is 63%, we normalize assets per period by dividing by 1.6 times the average labor income to be consistent with the output normalization in the model.

Inferring individual productivity. Since households in our model have two idiosyncratic states, productivity z_{it} and assets a_{it}, we must estimate the unobserved productivity of each individual every period. We follow Floden and Lindé (2001) whose identifying assumption is that permanent income differences can be captured by (observable) individual-specific characteristics. We estimate a pooled Mincer-style regression and regress log labor income on age, age-squared, and dummies

for gender, occupation and education level with time-fixed effects

$$\begin{aligned}\log y_{it} =&\varphi_t + \varphi_1 + \varphi_2 AGE_{it} + \varphi_3 AGE_{it}^2 + \\ &\boldsymbol{\varphi}_M MALE_{it} + \boldsymbol{\varphi}_E \mathbf{EDU}_{it} + \varphi_O \mathbf{OCC}_{it} + z_{it}\end{aligned}$$

where $\mathbf{EDU} = [EDU_1, EDU_2, \ldots, EDU_8]^\top$ is a vector of education dummies and $\mathbf{OCC} = [OCC_1, OCC_2, \ldots, OCC_15]^\top$ is a vector of occupation dummies. Table 3 shows the results of the regression. The adjusted R^2 is around 0.2, inline with the original Floden and Lindé (2001) results. The persistent component of log-productivity analogous to productivity in the model is defined as the residual of the above regression, $z_{it} = \log y_{it} - \widehat{\log y}_{it}$, which we compute each period for all individuals in the dataset. At this stage, for each period, we have data on individuals' consumption, income, their liquid assets and their productivity, which we will now group into bins.

Variable	Estimate	p-value
const	8.038	0.00
AGE	0.081	0.00
AGE^2(/100)	-0.089	0.00
MALE	0.124	0.00
EDUCA	0.024	0.00
R^2	0.24	
TN	341,785	

Table 3: Labor income regression results on U.S data

To construct our micro data matrix each period, we split productivity distribution into deciles and the asset space into 9 equally spaced bins, ranging from \$0 to \$200,000. Taking the cartesian product with the productivity bins gives us 81 bins.[25] In each period, we group individuals into these bins. We compute consumption for each bin as the mean consumption of the individuals in that bin, and analogously for income. In the event that a period has some bins that are empty, we linearly interpolate between the missing data *within* each period.[26] Repeating this for every period and stacking the vectors horizontally gives us a micro-data matrix, one row of which corresponds to the time series of (say) consumption of individuals at the 40th percentile of productivity and \$10,000 in liquid assets. Finally, for each row, we compute growth rates, seasonally adjust, and demean. The result of these operations is a quarterly series of consumption for individuals in each point of the discretized state-space.

One interesting interpretation of our data is that each column of our data matrix is an estimate

[25] Our choices imply that the bins are kept fixed in all periods.

[26] Figure F.3 in the Appendix visualizes the extent of the missing data before this operation.

of the consumption policy function in a given period. Figure 5 plots the time-average (or "steady-state") consumption policy functions as a ratio to average labor income. As our economic model predict, consumption appears to share some familiar characteristics to those in our economic models: they are increasing in productivity and assets, and are highly non-linear at low asset levels and linear at high asset levels.[27]

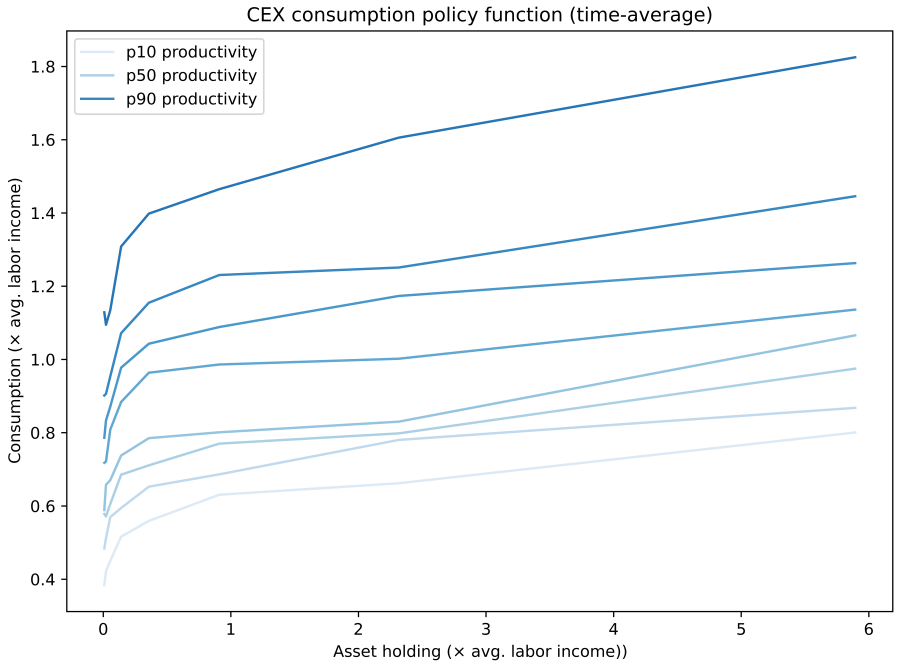

Figure 5: Consumption policy function inferred from CEX data

5.6.2 Macro data

For macro data, we follow Bayer et al. (2024) in terms of data sources. We use growth rate of: real GDP (`GDPC1`) per capita, real personal consumption expenditures per capita (`PCE`), real investment per capita (`FPI`), GDP deflator (`GDPDEF`), and nominal wages (`PRS85006101`). Our measure of the nominal interest rate the fed-funds rate (`FEDFUNDS`), which we splice with the shadow fed-funds rate of Wu and Xia (2016) between 2008-Q4 and 2015-Q4. We use `CNP16OV` in our per capita calculations. Our sample begins in 1959Q1 - 2020Q1.

[27] Arellano et al. (2017) infer more general consumption and income functions from PSID data that are non-parametric functions of many covariates. We do not take that approach here since our goal is estimation of the structural model in Section 5.1

5.7 Estimation

We estimate the model using two datasets: one that includes both micro and macro data as outlined above, which we call `Macro+Micro`; and another that only uses the macro data, which we call `Macro`.

We estimate the remaining parameters using Bayesian sampling methods. First, we perform an extensive mode-finding algorithm, by initializing a bounded `L-BFGS-B` algorithm at 10 random points in the permissible state space. We then take the parameter vector associated with the highest posterior value and perform an additional optimization using the `Nelder-Mead` method. For the `Macro+Micro` estimation, we employ our method to efficiently approximate the likelihood; and for the `Macro` estimation, we use the standard likelihood formulas for a moving average representation as described by Auclert et al. (2021b).

We then approximate the posterior distribution of the parameters using a random-walk Metropolis Hastings algorithm with tuned proposal covariance matrix and adaptive step size (Atchadé and Rosenthal (2005)). We generate 300,000 draws and discard the first 50,000. Though we employ a simple MCMC algorithm here, our method doesn't preclude more sophisticated techniques, such as the Hamiltonian Monte Carlo (Hoffman et al. (2014)) or Sequential Monte Carlo (Herbst and Schorfheide (2014)).

		Prior			Posterior (`Macro+Micro`)		Posterior (`Macro`)	
Parameter	Interpretation	Distribution	Mean	Std	Mean	Std	Mean	Std
κ_p	Phillips curve slope	Gamma	0.05	0.03	0.121	0.018	0.23	0.11
κ_w	Wage Phillips curve slope	Gamma	0.05	0.03	0.165	0.042	0.08	0.02
ϕ_π	Taylor Coefficient	Normal	1.5	0.3	2.228	0.056	1.64	0.18
ϕ_T	Tax smoothing	Beta	0.5	0.2	0.06	0.011	0.93	0.04
ρ_i	Taylor persistence	Beta	0.5	0.2	0.875	0.017	0.17	0.09
α_y	Aggregate income incidence	LogNormal	1.0	0.5	0.078	0.012	0.75	0.13
α_{mp}	Monetary policy incidence	LogNormal	1.0	0.5	0.418	0.083	1.34	0.34
ρ_Z	Persistence TFP shock	Beta	0.5	0.1	0.473	0.046	0.43	0.1
ρ_{v^r}	Persistence Monetary policy shock	Beta	0.5	0.1	0.373	0.063	0.76	0.06
ρ_{v^G}	Persistence Government spending shock	Beta	0.5	0.1	0.429	0.03	0.51	0.09
ρ_{v^w}	Persistence Wage markup shock	Beta	0.5	0.1	0.197	0.047	0.92	0.01
ρ_{v^π}	Persistence Price markup shock	Beta	0.5	0.1	0.205	0.038	0.41	0.09
ρ_{v^μ}	Persistence MEI shock	Beta	0.5	0.1	0.995	0.002	0.44	0.09
σ_Z	Std TFP shock	Inverse Gamma	1.0	0.2	0.509	0.066	0.16	0.04
σ_{v^r}	Std Monetary policy shock	Inverse Gamma	1.0	0.2	0.076	0.011	0.14	0.03
σ_{v^G}	Std Government spending shock	Inverse Gamma	1.0	0.2	0.259	0.028	0.21	0.04
σ_{v^w}	Std wage markup shock	Inverse Gamma	1.0	0.2	1.095	0.096	0.12	0.03
σ_{v^π}	Std price markup shock	Inverse Gamma	1.0	0.2	0.201	0.019	0.10	0.02
σ_{v^μ}	Std MEI shock	Inverse Gamma	1.0	0.2	0.419	0.085	0.18	0.04
σ_Y	Output meas. error	Inverse Gamma	1.0	0.2	0.11	0.01	0.35	0.04
σ_C	Consumption meas. error	Inverse Gamma	1.0	0.2	0.01	0.005	0.43	0.04
σ_i	Interest Rate meas. error	Inverse Gamma	1.0	0.2	0.01	0.004	0.64	0.05
σ_π	Inflation meas. error	Inverse Gamma	1.0	0.2	0.004	0.003	0.12	0.02
σ_w	Wage inflation meas. error	Inverse Gamma	1.0	0.2	0.04	0.003	0.86	0.06
σ_I	Investment Growth meas. error	Inverse Gamma	1.0	0.2	0.03	0.002	0.38	0.12
σ_c	Micro consumption meas. error	Inverse Gamma	3.0	0.4	4.283	0.037	-	-
σ_y	Micro income meas. error	Inverse Gamma	3.0	0.4	0.93	0.021	-	-

Table 4: Prior and posterior parameter estimates. NOTE: All standard deviations are presented $\times 100$

5.8 Results

Table 4 presents the posterior mean and standard deviation of the structural parameters from the estimations. The general observation is that, similar to the Monte-Carlo study in Section 4, `Macro+Micro` estimates are typically more precisely estimated than `Macro` estimates. A few more comparative observations stand out.

First, `Macro+Micro` estimates imply a slope of the price Phillips curve half that of `Macro` estimates, while the slope of the wage Phillips curve is estimated to be almost double. These imply that, all else equal, wages are significantly more volatile in the `Macro+Micro` economy and prices significantly less volatile.

A second important difference between the estimates is the aggressiveness of monetary policy, characterized by lag coefficient in the Taylor rule, ρ_i. `Macro+Micro` estimation has a ρ_i of 0.88, compared to 0.17 for `Macro`. This highly persistent process implies that means that monetary policy is slow to react to inflationary shocks, relative to the `Macro` model.

The third key difference is that almost all the estimated standard deviations of the structural shocks are higher for `Macro+Micro`, particularly the wage markup shock σ_{ε^w}, which is almost ten times higher, and the marginal efficiency of investment shocks σ_{ε^μ} which is around four times higher.

Finally, there are meaningful differences in the estimates of the incidence parameters α_y and α_{mp}. To interpret them more easily, Figure 6 plots the percentage change in idiosyncratic income associated with a one percent increase in aggregate income (left) and 25bps (contractionary) monetary policy shock (right). The swathes denote the 90% credible set, while the solid lines denote the mean.

In the left panel, `Macro+Micro` estimates imply that incomes for both the lowest and highest decile increase by around 1.5% following a 1% increase in aggregate income. This is in contrast with the middle of the productivity distribution, whose income only increases by 0.75%. The U-shape of the elasticities is consistent with empirical studies such as Guvenen et al. (2017), who use US census data to quantify the earnings elasticity to aggregate GDP. They estimate a regression of individual income on aggregate GDP and interpret the regression coefficient as the sensitivity to aggregate fluctuations. On the other hand, the `Macro` estimation obtains an upward sloping shape to the incidence function, with the lowest decile's incomes only increasing by 0.25%, while the income of the highest decile increases by around 2%.

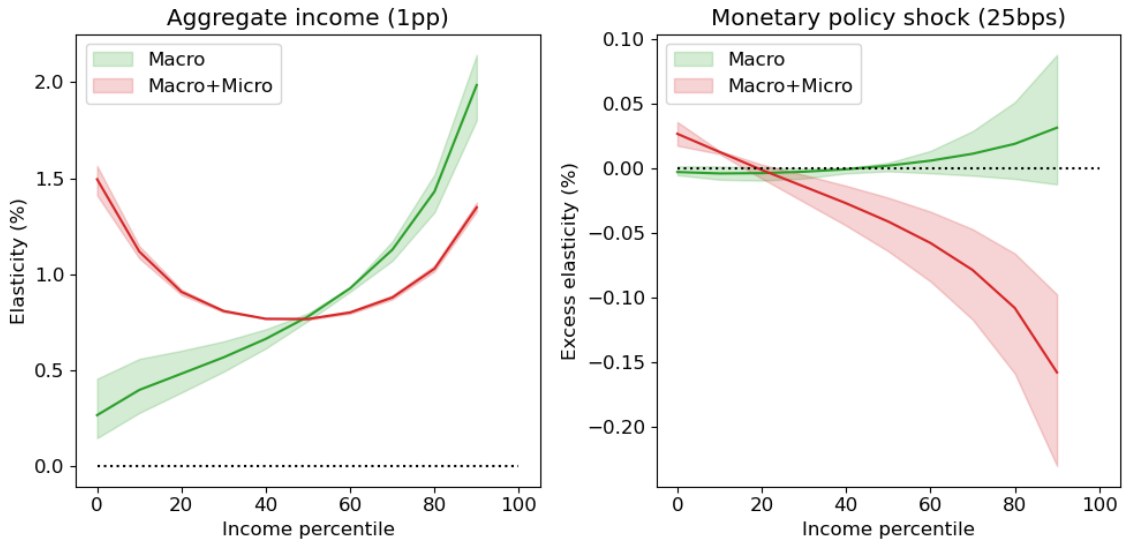

Figure 6: Estimated income elasticities by percentile

The right panel shows the excess elasticity ($\gamma_r(z)$) of idiosyncratic income conditional on a 25bps monetary policy shock. Positive values indicate that incomes react more strongly to monetary policy over and above its indirect effect through aggregate income, and vice versa. `Macro+Micro` estimates dictate that low income households are more exposed to monetary policy shocks, while `Macro` estimates suggest it is high income households that are over exposed. However, the `Macro` estimates are imprecisely estimated, such that the 90% credible band crosses the zero line.

That low income households are more sensitive to monetary policy shocks is consistent with the empirical literature. Coibion et al. (2017) use the CEX data to estimate a local projection of consumption and inequality on identified monetary policy shocks of Romer and Romer (2004). They find that income inequality – defined as either the cross-sectional standard deviation, the Gini coefficient or the p90-p10 – increases following a contractionary monetary policy shock.

5.8.1 Comparison of second moments

Next, we compare our parameter estimates by studying their ability to match the second moments of the data. Figure 7 plots the standard deviation of idiosyncratic income for each decile of productivity for the data, `Macro+Micro` and `Macro`. We compute the model income volatilities analytically, since the sequence-space jacobian solution obtains an MA representation.[28]

The data, in gray, exhibits a U-shaped pattern, with low and high income households sharing

[28]Since `Macro` does not have estimates for the measurement error of the micro variables, we also do not include them in the `Macro+Micro` volatility calculations. Including them makes the `Macro+Micro` income volatilities even closer to the data counterparts.

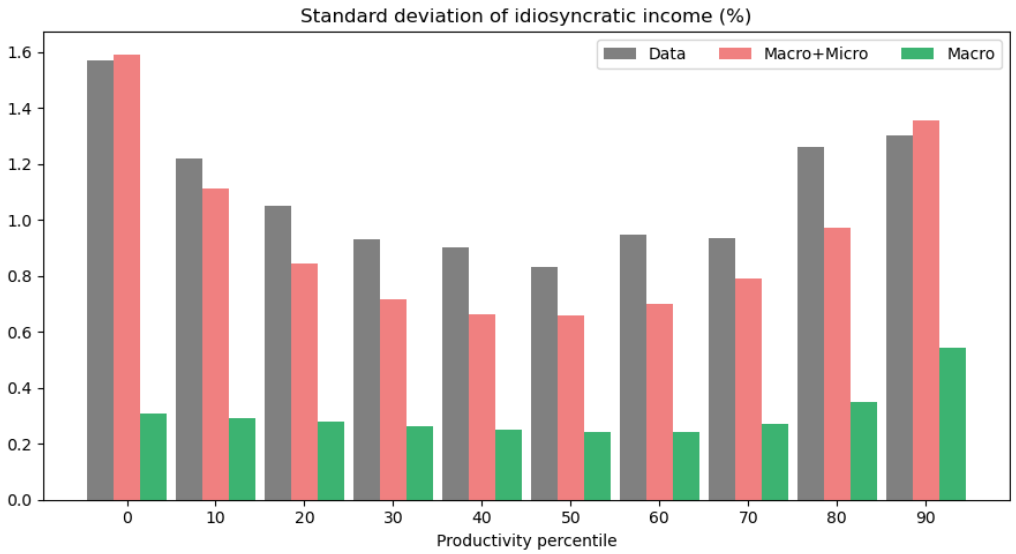

Figure 7: Idiosyncratic income volatility

similar levels of volatility. The `Macro+Micro` estimation is able to match both the level and the shape of the data in this dimension, which is a direct consequence of the elasticities in Figure 6. The `Macro` estimates imply a much smaller level of volatility across the distribution. Moreover, it implies – counterfactually – that high income households have substantially more volatile income than the low income households.

In comparing the models' implications for aggregate quantities, a different picture emerges. Table 5 shows the standard deviation of the aggregate variables used in the estimation and compares that to those implies by both `Macro` and `Macro+Micro`. It shows that `Macro` estimates are able to closely match most of the second moments, and particularly so for output, consumption and investment. On the other hand, `Macro+Micro` estimates generally overstate the volatility of aggregate variables. For example, quarterly consumption volatility in the data is 0.57% versus 0.72% for `Macro+Micro`; investment is 1.87% in the data versus 3.43% for `Macro+Micro`; and wage inflation is 0.89% in the data versus 1.0% in `Macro+Micro`.

Two key driving forces in these differences are the large estimated standard deviations of TFP

	Standard deviation (%)					
	Output	Consumption	Investment	Price inflation	Wage inflation	Interest rate
Data	0.60	0.57	1.87	0.22	0.89	0.69
`Macro`	0.60	0.61	1.79	0.44	0.81	0.74
`Macro+Micro`	0.59	0.72	3.43	0.36	1.00	0.40

Table 5: Standard deviation of aggregate variables

Prod. pctile	Macro+Micro	$\varepsilon^Z = 0$	$\varepsilon^r = 0$	$\varepsilon^G = 0$	$\varepsilon^w = 0$	$\varepsilon^\pi = 0$	$\varepsilon^\mu = 0$
p10	0.18	0.17	0.18	0.12	0.18	0.18	0.15
p20	0.32	0.26	0.32	0.29	0.32	0.32	0.27
p30	0.42	0.32	0.41	0.4	0.41	0.42	0.35
p40	0.48	0.35	0.46	0.46	0.46	0.47	0.39
p50	0.48	0.37	0.46	0.47	0.47	0.47	0.38
p60	0.32	0.29	0.29	0.31	0.31	0.31	0.18
p70	0.37	0.36	0.33	0.37	0.37	0.36	0.19
p80	0.43	0.43	0.4	0.43	0.43	0.43	0.20
p90	0.48	0.48	0.45	0.48	0.48	0.47	0.20
p99	0.52	0.52	0.49	0.52	0.52	0.51	0.20

Table 6: Standard deviation of idiosyncratic consumption

and wage markup shocks in Table 4. Though the large standard deviations in Macro+Micro are useful in being able to match the micro income volatility through the relation $y_t(z) = \Gamma_t(z) w_t N_t e^{\varepsilon_t^r}$, they also imply high volatility at the aggregate level – seen by the larger wage inflation and consumption volatilities.

Another key difference in Table 5 is the standard deviation of investment. Here, the differences come from shocks to TFP and marginal efficiency of investment, which play a crucial role in generating the appropriate amount of volatility in consumption at the micro level. To see this, table 6 presents the average consumption volatility of individuals grouped by productivity decile for various versions of the model. The Macro+Micro column reports the average consumption volatility in the estimated model. The remaining columns report the same object shutting down each structural shock in turn. The cells in red denote the column that corresponds to the largest fall in the volatility from the Macro+Micro value. The table shows that TFP is a dominant driver of consumption volatility for low productivity individuals, followed closely by the investment shock. For the high productivity individuals, the investment shock is the single dominant driver of consumption volatility. Why? The reason comes down to the positive covariance between consumption and investment, conditional on the MEI shock. In the model, following a positive MEI shock, firms increase investment and hire more workers, all else equal. Due wage and price rigidities, aggregate labor income rise too. The estimated "U"-shaped elasticities imply that affects both low- and high- income households more. Thus, larger standard deviations of the MEI shock imply larger standard deviations of income for high-income households. A similar mechanism is at play in Auclert et al. (2023), who document a significant role of investment in the propagation of monetary policy shocks.

5.8.2 External validation

Given such meaningful differences in the models, it is especially useful in this context to compare them against some kind of empirical benchmark. We compare them to the impulse response of output and inflation to a 25bps (contractionary) monetary policy shock.

For our empirical benchmark, we estimate a monthly SVAR-IV with 12 lags between 1973 and 2020 of the following variables, instrumenting the monetary policy shock with the identified monetary policy shock of Bauer and Swanson (2023). We omit the September 2008 from the sample, which coincides with the Lehmann Brothers bankruptcy, by setting the monetary policy shock to zero in that month.

- 1-year nominal Treasury yield (`DGS1`)
- Unemployment rate, defined as the number of unemployed as a percentage of the labor force (16 years or older) (`UNRATE`)
- The change in log CRB Commodity index (`CRBI`)
- PCE inflation, defined as the change in the log PCE Price index (`PCEPI`)
- Detrended log real GDP. We use the monthly real GDP series of Brave-Butters-Kelley, and estimate the trend by an index with growth rate equal to the mean growth rate of the log real GDP. The detrended series is simply the deviations from that linear trend. Implicitly in this transformation, we assume that real GDP was at trend in 1973 Q1.

Figure 8 plots the impulse responses for output (left panel) and inflation (right panel) to a 25bps contractionary monetary policy shock. Each panel plots the theoretical impulse response functions implied by `Macro+Micro` (red) and `Macro` (green). The colored swathes denote the 90% credible sets. The gray line denote the SVAR-IV IRFs and the gray swathe denotes the 68% confidence bands[29].

On impact following the monetary policy shock, output falls by 0.15ppts in the `Macro+Micro` model, compared to 0.04ppts for `Macro`. This is compared to a 0.21ppt fall in output from the empirical IRF. For inflation, `Macro+Micro` implies a 0.09ppt fall in inflation on impact, compared to a 0.05ppts for `Macro`. This is compared to a negative 0.13ppts response on impact in the SVAR-IV IRF. All in all, the larger responses of both output and inflation in `Macro+Micro` model appear

[29]The 60% confidence bands are computed by 10,000 draws of the Wild bootstrap.

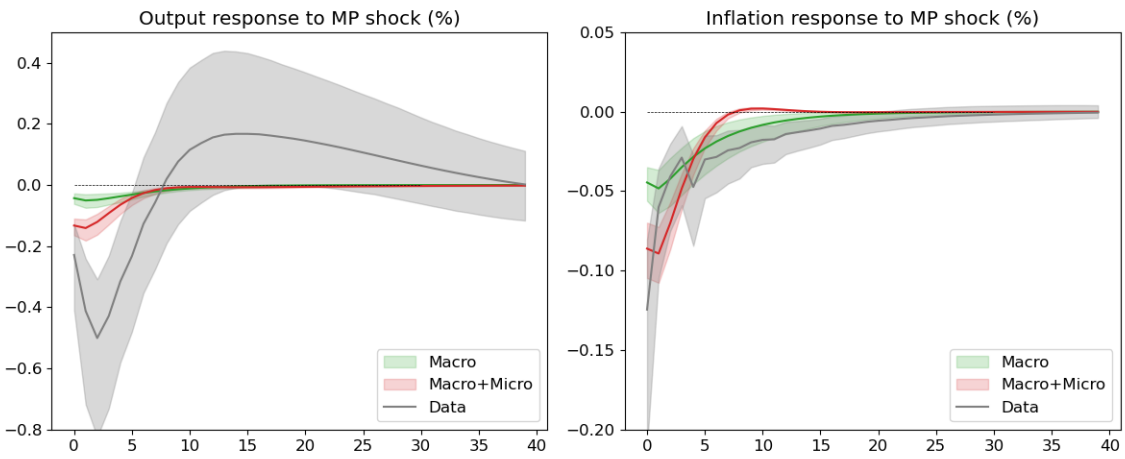

Figure 8: Impulse response to 25bps contractionary monetary policy shock

to be closer to those estimated by a SVAR-IV, which is crucially external (out of sample) to the structural estimation.

The mechanism, through the lens of the model, is as follows. A positive monetary policy shock increases the real rental rate of capital, which firms respond to by scaling back investment, thus reducing output. Since less labor is required, aggregate labor income also falls. Low income households bear the brunt of this decrease, given the U-shaped elasticities that `Macro+Micro` estimates. These households are the ones with the high MPCs, and so cut their individual consumption substantially, which further amplifies the fall in output. The flatter Phillips curve, combined with the highly persistent Taylor rule implies that the central bank does very little to offset the recession in this economy. This is in direct contrast to `Macro` in which i) does not exhibit the "U-shape" in labor incidence and ii) the steep Phillips curve and the low Taylor rule persistence combine to significantly accommodate the shock.

5.8.3 Consumption amplification

We use the model to infer how heterogeneous earnings exposures affect the amplification of monetary policy shocks. For both models, we compute the impulse response of aggregate consumption to a 25bp contractionary monetary policy shock. We compare this to an *equal exposure benchmark* computed by setting $\alpha_y = \alpha_{mp} = 1$. Figure 9 shows the box plots of the amplification factor arising from heterogeneous exposures. A value of one implies no amplification of consumption. The whiskers plot the 90% credible set and the box plots the 68% credible set.

As shown in Figure 9, the unequal incidence results in a mean amplification of around 40%

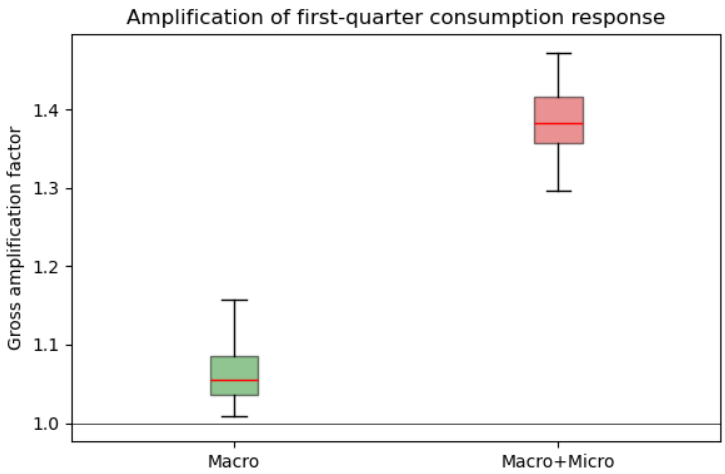

Figure 9: Amplification of first-quarter consumption to monetary policy shock

with a 90% credible set ranging from 30% to 46%. The significant amplification here is due to the "U"-shaped incidence functions. In this model, Low income/high MPC households are highly exposed to the monetary policy shock, and significantly adjust their consumption. This is further amplified by the high slope of the Phillips curve and high Taylor persistence meaning that the monetary policy does little to accommodate the economy. The amplification is much tamed in the `Macro` model - less than 10% - due to the implied mildly procyclical earnings incidence of monetary policy.

6 Concluding remarks

Estimating dynamic heterogeneous-agent models using large individual-level micro datasets would be infeasible for most existing methodologies. Thus, the current literature has resorted to using a limited set of moments, potentially discarding valuable data. This paper presents a tractable econometric framework that approximates the joint likelihood of macro and individual-level micro data. Our theoretical result establishes sufficient conditions under which the structural model can be well-approximated by a reduced-rank VAR(1), which opens the door to rapid estimation of the structural parameters. In estimating a medium-scale HANK model, we find substantial consumption amplification to monetary policy shocks, driven by the fact that high MPC households are also highly sensitive to changes in aggregate income. Our estimation exercise also sheds an important light on misspecification in workhorse HANK models since they are unable to simultaneously match macro and micro second moments.

On the methodology, interesting extensions include extending our algorithm to re-weight the micro and macro contributions to the likelihood. On the modelling side, extensions include suggesting novel mechanisms that enable HANK models to better match macro and micro volatility simultaneously. We leave these to future works.

References

ACHARYA, S., W. CHEN, M. DEL NEGRO, K. DOGRA, E. MATLIN, AND R. SARFATI (2020): "Estimating HANK: macro time series and micro moments," *Manuscript, March*.

AIYAGARI, S. R. (1994): "Uninsured idiosyncratic risk and aggregate saving," *The Quarterly Journal of Economics*, 109, 659–684.

ALVES, F., G. KAPLAN, B. MOLL, AND G. L. VIOLANTE (2020): "A further look at the propagation of monetary policy shocks in HANK," *Journal of Money, Credit and Banking*, 52, 521–559.

ALVEZ, F., G. KAPLAN, B. MOLL, AND G. L. VIOLANTE (2020): "A Further Look at the Propagation of Monetary Policy Shocks in HANK," *Journal of Money, Credit and Banking*, 52, 521–559.

AMBERG, N., T. JANSSON, M. KLEIN, AND A. R. PICCO (2022): "Five facts about the distributional income effects of monetary policy shocks," *American Economic Review: Insights*, 4, 289–304.

ANDERSEN, A. L., N. JOHANNESEN, M. JØRGENSEN, AND J.-L. PEYDRÓ (2023): "Monetary policy and inequality," *The Journal of Finance*, 78, 2945–2989.

ANDERSON, T. W. (1951): "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions," *The Annals of Mathematical Statistics*, 22, 327–351.

ANDERSON, T. W. AND H. RUBIN (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 20, 46–63.

ARELLANO, M., R. BLUNDELL, AND S. BONHOMME (2017): "Earnings and consumption dynamics: a nonlinear panel data framework," *Econometrica*, 85, 693–734.

ATCHADÉ, Y. F. AND J. S. ROSENTHAL (2005): "On adaptive markov chain monte carlo algorithms," *Bernoulli*, 11, 815–828.

AUCLERT, A. (2019): "Monetary policy and the redistribution channel," *American Economic Review*, 109, 2333–2367.

AUCLERT, A., B. BARDÓCZY, AND M. ROGNLIE (2023): "MPCs, MPEs, and multipliers: A trilemma for New Keynesian models," *Review of Economics and Statistics*, 105, 700–712.

AUCLERT, A., B. BARDÓCZY, M. ROGNLIE, AND L. STRAUB (2021a): "Using the sequence-space Jacobian to solve and estimate heterogeneous-agent models," *Econometrica*, 89, 2375–2408.

——— (2021b): "Using the sequence-space Jacobian to solve and estimate heterogeneous-agent models," Tech. rep., Stanford.

AUCLERT, A. AND M. ROGNLIE (2018): "Inequality and aggregate demand," Tech. rep., National Bureau of Economic Research.

AUCLERT, A., M. ROGNLIE, AND S. LUDWIG (2024): "The Intertemporal Keynesian Cross," *Journal of Political Economy*.

AUCLERT, A., M. ROGNLIE, AND L. STRAUB (2020): "Micro jumps, macro humps: Monetary policy and business cycles in an estimated HANK model," Tech. rep., National Bureau of Economic Research.

BAI, J. AND S. NG (2002): "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191–221.

——— (2006): "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Econometrica*, 74, 1133–1150.

BARDÓCZY, B. AND J. GUERREIRO (2023): "Unemployment insurance in macroeconomic stabilization with imperfect expectations," *Manuscript, April*.

BAUER, M. D. AND E. T. SWANSON (2023): "A Reassessment of Monetary Policy Surprises and High-Frequency Identification," *NBER Macroeconomics Annual*, 37, 87–155.

BAYER, C., B. BORN, AND R. LUETTICKE (2024): "Shocks, Frictions, and Inequality in U.S. Business Cycles," *American Economic Review*.

BAYER, C. AND R. LUETTICKE (2020): "Solving heterogeneous agent models in discrete time with many idiosyncratic states by perturbation methods," *Quantitative Economics*, 11, 1253–1288.

BOAR, C. AND V. MIDRIGAN (2022): "Efficient redistribution," *Journal of Monetary Economics*, 131, 78–91.

BROER, T., J. V. KRAMER, AND K. MITMAN (2026): "The curious incidence of monetary policy across the income distribution," .

BRUNTON, S., B. BRUNTON, J. PROCTOR, AND N. KUTZ (2015): "Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control," *arXiv*.

BRUNTON, S. L. AND J. N. KUTZ (2022): *Data-Driven Science and Engineering: Machine Learnings, Dynamical Systems, and Control, second edition*, Cambridge University Press.

CHALLE, E., J. MATHERON, X. RAGOT, AND J. F. RUBIO-RAMIREZ (2017): "Precautionary saving and aggregate demand," *Quantitative Economics*, 8, 435–478.

CHAMBERLAIN, G. AND M. ROTHSCHILD (1982): "Arbitrage, factor structure, and mean-variance analysis on large asset markets," National Bureau of Economic Research Cambridge, Mass., USA.

CHANG, M., X. CHEN, AND F. SCHORFHEIDE (2021): "Heterogeneity and aggregate fluctuations," Tech. rep., National Bureau of Economic Research.

CHRISTIANO, L. J. AND R. J. VIGFUSSON (2003): "Maximum likelihood in the frequency domain: the importance of time-to-plan," *Journal of Monetary Economics*, 50, 789–815.

COIBION, O., Y. GORODNICHENKO, L. KUENG, AND J. SILVIA (2017): "Innocent Bystanders? Monetary policy and inequality," *Journal of Monetary Economics*, 88, 70–89.

FERNÁNDEZ-VILLAVERDE, J., S. HURTADO, AND G. NUNO (2023): "Financial frictions and the wealth distribution," *Econometrica*, 91, 869–901.

FLODEN, M. AND J. LINDÉ (2001): "Idiosyncratic risk in the United States and Sweden: Is there a role for government insurance?" *Review of Economic dynamics*, 4, 406–437.

GUVENEN, F., S. SCHULHOFER-WOHL, J. SONG, AND M. YOGO (2017): "Worker betas: Five facts about systematic earnings risk," *American Economic Review*, 107, 398–403.

HANSEN, L. AND T. SARGENT (1981): "Exact linear rational expectations models: specification and estimation," Tech. rep., Federal Reserve Bank of Minneapolis.

HEATHCOTE, J., K. STORESLETTEN, AND G. L. VIOLANTE (2017): "Optimal tax progressivity: An analytical framework," *The Quarterly Journal of Economics*, 132, 1693–1754.

HERBST, E. AND F. SCHORFHEIDE (2014): "Sequential Monte Carlo sampling for DSGE models," *Journal of Applied Econometrics*, 29, 1073–1098.

HOFFMAN, M. D., A. GELMAN, ET AL. (2014): "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*, 15, 1593–1623.

HOLM, M. B., P. PAUL, AND A. TISCHBIREK (2021): "The transmission of monetary policy under the microscope," *Journal of Political Economy*, 129, 2861–2904.

JUSTINIANO, A., G. E. PRIMICERI, AND A. TAMBALOTTI (2011): "Investment shocks and the relative price of investment," *Review of Economic Dynamics*, 14, 102–121.

KASE, H., L. MELOSI, AND M. ROTTNER (2022): *Estimating nonlinear heterogeneous agents models with neural networks*, Centre for Economic Policy Research.

KRUSELL, P. AND A. A. SMITH, JR (1998): "Income and wealth heterogeneity in the macroeconomy," *Journal of political Economy*, 106, 867–896.

LIU, L. AND M. PLAGBORG-MØLLER (2023): "Full-information estimation of heterogeneous agent models using macro and micro data," *Quantitative Economics*, 14, 1–35.

LUNGQVIST, L. AND T. J. SARGENT (2018): *Recursive Macroeconomic Theory*, The MIT Press.

MONGEY, S. AND J. WILLIAMS (2017): "Firm dispersion and business cycles: Estimating aggregate shocks using panel data," *Manuscript, New York University*.

MORALES-JIMENEZ, C. AND L. STEVENS (2024): "Price Rigidities in U.S. Business Cycles," Tech. rep., University of Maryland, working paper.

PATTERSON, C. (2023): "The matching multiplier and the amplification of recessions," *American Economic Review*, 113, 982–1012.

PLAGBORG-MØLLER, M. (2019): "Bayesian inference on structural impulse response functions," *Quantitative Economics*, 10, 145–184.

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

ROMER, C. D. AND D. H. ROMER (2004): "A new measure of monetary shocks: Derivation and implications," *American economic review*, 94, 1055–1084.

ROTEMBERG, J. J. (1982): "Sticky prices in the United States," *Journal of Political Economy*, 90, 1187–1211.

ROUWENHORST, K. G. (1995): "Asset pricing implications of equilibrium business cycle models," *Frontiers of business cycle research*, 1, 294–330.

SARGENT, T. J., Y. J. SELVAKUMAR, AND Z. YANG (2026a): "Aggregate Shocks and Cross-Section Dynamics: Quantifying Redistribution and Insurance in US Household Data," *Journal of Political Economy Macroeconomics*, forthcoming.

——— (2026b): "Dynamic Mode Decompositions and Vector Autoregressions," *International Economic Review*, n/a.

SCHMIDT, P. AND J. SESTERHENN (2010): "Dynamic Mode Decomposition of numerical and experimental data," *Journal of Fluid Dynamics*, 656, 5–28.

STOCK, J. H. AND M. W. WATSON (2002): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

TAUCHEN, G. (1986): "Finite state markov-chain approximations to univariate and vector autoregressions," *Economics letters*, 20, 177–181.

TU, J. H., C. W. ROWLEY, D. M. LUCHTENBURG, S. L. BRUNTON, AND J. N. KUTZ (2014): "On dynamic mode decomposition: Theory and applications," *Journal of Computational Dynamics*, 1, 391–421.

WINBERRY, T. (2018): "A method for solving and estimating heterogeneous agent macro models," *Quantitative Economics*, 9, 1123–1151.

——— (2021): "Lumpy Investment, Business Cycles, and Stimulus Policy," *American Economic Review*, 111, 364–396.

WU, J. C. AND F. D. XIA (2016): "Measuring the macroeconomic impact of monetary policy at the zero lower bound," *Journal of Money, Credit and Banking*, 48, 253–291.

Appendix A Proofs

A.1 Proof of Proposition 1

Proof. Associated with state-space system (1) is its innovations representation[30]

$$\begin{aligned} \widehat{\mathbf{x}}_{t+1} &= \mathbf{A}\ \widehat{\mathbf{x}}_t + \mathbf{K}\,\mathbf{a}_t \\ \mathbf{y}_t &= \mathbf{G}\ \widehat{\mathbf{x}}_t + \mathbf{a}_t \end{aligned} \tag{A.1}$$

where $\widehat{\mathbf{x}}_t = \mathbb{E}[\mathbf{x}_t \,|\, \mathbf{y}^{t-1}]$, $\mathbf{a}_t = \mathbf{y}_t - \mathbb{E}[\mathbf{y}_t \,|\, \mathbf{y}^{t-1}]$, $\mathbf{a}_t \perp \mathbf{a}_s \,\forall t \neq s$ for $\mathbf{y}^t = \{\mathbf{y}_s\}_{s<t}$ and the Hilbert space $\mathcal{H}(\mathbf{a}^t) = \mathcal{H}(\mathbf{y}^t)$. Furthermore, $\mathbf{\Omega} \equiv \mathbb{E}[\mathbf{a}_t\,\mathbf{a}_t^\top] = \mathbf{G}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top + \mathbf{R}$, where $\mathbf{\Sigma}_\infty$ and $\mathbf{K}$ satisfy

$$\begin{aligned} \mathbf{\Sigma}_\infty &= \mathbb{E}[\mathbf{x}_t - \widehat{\mathbf{x}}_t][\mathbf{x}_t - \widehat{\mathbf{x}}_t]^\top \\ &= \mathbf{C}\,\mathbf{C}^\top + \mathbf{K}\,\mathbf{R}\,\mathbf{K}^\top + (\mathbf{A} - \mathbf{K}\,\mathbf{G})\mathbf{\Sigma}_\infty(\mathbf{A} - \mathbf{K}\,\mathbf{G})^\top \\ \mathbf{K} &= \mathbf{A}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top(\mathbf{G}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top + \mathbf{R})^{-1}. \end{aligned}$$

Notice that $rank(\mathbf{K}) = r$. Rearranging (A.1) gives an expression for $\mathbf{x}_{t+1}$ in terms of $\mathbf{y}_t$ and $\mathbf{x}_t$

$$\begin{aligned} \widehat{\mathbf{x}}_{t+1} &= \mathbf{A}\,\widehat{\mathbf{x}}_t + \mathbf{K}(\mathbf{y}_t - \mathbf{G}\,\widehat{\mathbf{x}}_t) \\ &= (\mathbf{A} - \mathbf{K}\,\mathbf{G})\widehat{\mathbf{x}}_t + \mathbf{K}\,\mathbf{y}_t \end{aligned}$$

Substituting into the measurement equation of (A.1) gives

$$\begin{aligned} \mathbf{y}_t &= \mathbf{G}[(\mathbf{A} - \mathbf{K}\,\mathbf{G})\widehat{\mathbf{x}}_{t-1} + \mathbf{K}\,\mathbf{y}_{t-1}] + \mathbf{a}_t \\ &= \mathbf{G}\,\mathbf{K}\,\mathbf{y}_{t-1} + \mathbf{G}(\mathbf{A} - \mathbf{K}\,\mathbf{G})\widehat{\mathbf{x}}_{t-1} + \mathbf{a}_t \end{aligned}$$

Notice that $\mathbf{B}_1^\infty := \mathbf{G}\,\mathbf{K}$ is a rank-r matrix. Moreover, so is $\mathbf{A} - \mathbf{K}\,\mathbf{G}$. Iterating backward gives us the desired result

$$\mathbf{y}_t = \sum_{j=1}^{\infty} \mathbf{B}_j^\infty\,\mathbf{y}_{t-j} + \mathbf{a}_t \tag{A.2}$$

$$\mathbb{E}[\mathbf{a}_t\,\mathbf{y}_{t-j}^\top] = \mathbf{0} \quad \text{for all } j \geq 1$$

$$\mathbb{E}[\mathbf{a}_t\,\mathbf{a}_t^T] = \mathbf{\Omega} = \mathbf{G}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top + \mathbf{R}$$

$$\mathbf{B}_j^\infty = \mathbf{G}(\mathbf{A} - \mathbf{K}\,\mathbf{G})^{j-1}\,\mathbf{K} \quad \forall j \geq 1 \tag{A.3}$$

where $\text{rank}(\mathbf{B}_j^\infty) = r \quad \forall j \geq 1$. □

[30] A detailed derivation can be found in Lungqvist and Sargent (2018), Ch. 2

A.2 Proof of Lemma 1

Proof. Consider a sequence of models $\{\mathcal{M}_N\}$ indexed by the number of observables $N \in \mathbb{N}$. For each N, the model $\mathcal{M}_N$ is given by

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{A}\,\mathbf{x}_t + \mathbf{C}\,\mathbf{w}_{t+1}\\ \mathbf{y}_t &= \mathbf{G}_N\,\mathbf{x}_t + \mathbf{v}_t,\end{aligned}$$

where shocks $\mathbf{w}_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r\times r})$, measurement errors $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_N)$ and $\mathbf{w}_s \perp \mathbf{v}_\tau$ for all s, τ. Note that the matrices $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{r\times r}$ are fixed across N, meaning that the transition equation of the unobserved state is invariant to the number of observables. In the following, $\|\cdot\|$ denotes the Frobenius norm.

By Proposition 1, we have

$$\mathbf{K}_N = \mathbf{A}\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top(\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top + \mathbf{R}_N)^{-1}$$

where $\mathbf{\Sigma}_{\infty,N} \in \mathrm{GL}(r, \mathbb{R})$ solves the matrix Ricatti equation

$$\begin{aligned}\mathbf{\Sigma}_{\infty,N} &= \mathbf{C}\,\mathbf{C}^\top + \mathbf{K}_N\,\mathbf{R}_N\,\mathbf{K}_N^\top + (\mathbf{A} - \mathbf{K}_N\,\mathbf{G}_N)\mathbf{\Sigma}_{\infty,N}(\mathbf{A} - \mathbf{K}_N\,\mathbf{G}_N)^\top &\text{(A.4)}\\ &= \mathbf{A}\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{A}^\top + \mathbf{C}\,\mathbf{C}^\top - \mathbf{A}\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}^\top(\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top + \mathbf{R}_N)^{-1}\,\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{A}^\top &\text{(A.5)}\end{aligned}$$

Such an invertible solution always exists and is unique under the maintained stability assumption. Furthermore, by construction, $\mathbf{\Sigma}_{\infty,N} = \mathbb{E}[\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{y}^{t-1}]][\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{y}^{t-1}]]^\top$.

We will first show that $\mathbf{\Sigma}_{\infty,N} \to \mathbf{C}\,\mathbf{C}^\top = \mathbb{E}[\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_{t-1}]][\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_{t-1}]]^\top$. By Assumption 2, $rank(\mathbf{G}_N) = r$, so we have the normal equation

$$\mathbf{x}_t = (\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{G}_N^\top\,\mathbf{y}_t - (\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{G}_N^\top\,\mathbf{v}_t$$

Combine with the state transition equation and we obtain

$$\mathbf{x}_t = \mathbf{A}(\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{G}_N^\top\,\mathbf{y}_{t-1} - \mathbf{A}(\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{G}_N^\top\,\mathbf{v}_{t-1} + \mathbf{C}\,\mathbf{w}_t$$

Put $F[\mathbf{x}_t \mid \mathbf{y}_{t-1}] := A(\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{G}_N^\top\,\mathbf{y}_{t-1}$. The forecast variance of this linear predictor is

$$\mathbb{E}[\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]][\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]]^\top = \mathbf{A}(\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{G}_N^\top\,\mathbf{R}_N\,\mathbf{G}_N(\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{A}^\top + \mathbf{C}\,\mathbf{C}^\top$$

Therefore, we have

$$\begin{aligned}\left\|\mathbb{E}[\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]][\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]]^\top - \mathbf{C}\,\mathbf{C}^\top\right\| &= \left\|\mathbf{A}(\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{G}_N^\top\,\mathbf{R}_N\,\mathbf{G}_N(\mathbf{G}_N^\top\,\mathbf{G}_N)^{-1}\,\mathbf{A}^\top\right\|\\ &\le \frac{\lambda_{\max}(\mathbf{R}_N)}{N}\,\|\mathbf{A}\|^2\left\|\left(\frac{1}{N}\,\mathbf{G}_N^\top\,\mathbf{G}_N\right)^{-1}\right\|^2\left\|\frac{1}{N}\,\mathbf{G}_N^\top\,\mathbf{G}_N\right\|\end{aligned}$$

where we have used the fact that $\mathbf{R}_N$ is positive semidefinite and $\lambda_{\max}(\mathbf{R}_N)$ denotes the largest eigenvalue of $\mathbf{R}_N$. By assumption 2 and 3, the RHS converges to 0 as $N \to \infty$. We conclude that $\mathbb{E}[\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]][\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]]^\top \to \mathbf{C}\,\mathbf{C}^\top$ pointwise as $M \to \infty$. Finally, since conditional expectation minimizes mean-square errors, we have

$$\begin{aligned}\mathbf{\Sigma}_{\infty,N} &= \mathbb{E}[\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{y}^{t-1}]][\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{y}^{t-1}]]^\top \\ &\preceq \mathbb{E}[\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]][\mathbf{x}_t - F[\mathbf{x}_t \mid \mathbf{y}_{t-1}]]^\top\end{aligned}$$

where $\preceq$ represents the Loewner order.[31] Clearly, we must have $\mathbf{C}\,\mathbf{C}^\top \preceq \mathbf{\Sigma}_{\infty,N}$ because $\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y}^{t-1}] = \mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_{t-1}]$. Then by the continuity and anti-symmetry of the Loewner order, we conclude that $\mathbf{\Sigma}_{\infty,N} \to \mathbf{C}\,\mathbf{C}^\top$ pointwise as $N \to \infty$.

We are ready to prove that $\|\mathbf{A} - \mathbf{K}_N\,\mathbf{G}_N\| \to 0$. By the matrix Ricatti equation (A.5), we have

$$\left\|(\mathbf{A}\,\mathbf{\Sigma}_{\infty,N})^{-1}(\mathbf{\Sigma}_{\infty,N} - \mathbf{C}\,\mathbf{C}^\top)(\mathbf{A}^\top)^{-1}\right\| = \left\|\mathbf{I}_N - \mathbf{G}_N^\top(\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top + \mathbf{R}_N)^{-1}\,\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\right\|$$

Let $N \to \infty$ and we have the limit

$$\left\|\mathbf{I}_N - \mathbf{G}_N^\top(\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top + \mathbf{R}_N)^{-1}\,\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\right\| \to 0$$

Note that

$$\begin{aligned}\|\mathbf{A} - \mathbf{K}_N\,\mathbf{G}_N\| &= \left\|\mathbf{A} - \mathbf{A}\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top(\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top + \mathbf{R}_N)^{-1}\,\mathbf{G}_N\right\| \\ &\le \|\mathbf{A}\|\left\|\mathbf{I}_N - \mathbf{\Sigma}_{\infty,N}\,\mathbf{G}^\top(\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top + \mathbf{R}_N)^{-1}\,\mathbf{G}_N\right\| \\ &= \|\mathbf{A}\|\left\|\mathbf{I}_N - \mathbf{G}_N^\top(\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\,\mathbf{G}_N^\top + \mathbf{R}_N)^{-1}\,\mathbf{G}_N\,\mathbf{\Sigma}_{\infty,N}\right\|\end{aligned}$$

where the last equality follows from taking transpose and the symmetry of $\mathbf{R}_N$ and $\mathbf{\Sigma}_{\infty,N}$. Let $N \to \infty$ and we have $\|\mathbf{A} - \mathbf{K}_N\,\mathbf{G}_N\| \to 0$, as desired.

We can further compute the convergence rate. Note that

$$\begin{aligned}N\,\|\mathbf{A} - \mathbf{K}_N\,\mathbf{G}_N\| &\le N\,\|\mathbf{A}\|\left\|(\mathbf{A}\,\mathbf{\Sigma}_{\infty,N})^{-1}(\mathbf{\Sigma}_{\infty,N} - \mathbf{C}\,\mathbf{C}^\top)(\mathbf{A}^\top)^{-1}\right\| \\ &\le \|\mathbf{A}\|\left\|(\mathbf{A}\,\mathbf{\Sigma}_{\infty,N})^{-1}\right\|\left\|N(\mathbf{\Sigma}_{\infty,N} - \mathbf{C}\,\mathbf{C}^\top)\right\|\left\|(\mathbf{A}^\top)^{-1}\right\|\end{aligned}$$

It is clear from our proof of $\left\|\mathbf{\Sigma}_{\infty,N} - \mathbf{C}\,\mathbf{C}^\top\right\| \to 0$ that $\limsup_{N\to\infty}\left\|N\mathbf{\Sigma}_{\infty,N} - \mathbf{C}\,\mathbf{C}^\top\right\| < \infty$. Thus, we conclude that $\limsup_{N\to\infty} N\,\|\mathbf{A} - \mathbf{K}_N\,\mathbf{G}_N\| < \infty$.

□

[31] For any pair of positive semidefinite matrices $A, B \in \mathbb{R}^{N\times N}$, $A \preceq B$ iff $B - A$ is positive semidefinite.

A.3 Proof of Corollary 1

Proof. Manipulating the innovations representation from the proof of Proposition 1 gives

$$\widehat{\mathbf{x}}_{t+1} = \mathbf{A}\,\mathbf{x}_t + \mathbf{K}(\mathbf{y}_t - \mathbf{G}\,\widehat{\mathbf{x}}_t) \tag{A.6}$$

$$= (\mathbf{A} - \mathbf{K}\,\mathbf{G})\widehat{\mathbf{x}}_t + \mathbf{K}\,\mathbf{y}_t \tag{A.7}$$

Define $\widetilde{\mathbf{x}}_t := \mathbb{E}[\mathbf{x}_t \,|\, \mathbf{y}^t]$. So, $\widehat{\mathbf{x}}_{t+1} = \mathbf{A}\,\widetilde{\mathbf{x}}_t$. Suppose $\mathbf{A} - \mathbf{K}\,\mathbf{G} = \mathbf{0}$. Then,

$$\widehat{\mathbf{x}}_{t+1} = \mathbf{K}\,\mathbf{y}_t \tag{A.8}$$

$$\mathbf{A}\,\widetilde{\mathbf{x}}_t = \mathbf{A}\,\mathbf{L}\,\mathbf{y}_t \tag{A.9}$$

for $\mathbf{L} = \mathbf{\Sigma}_\infty\,\mathbf{G}\,\mathbf{\Omega}^{-1}$ such that $\mathbf{K} = \mathbf{A}\,\mathbf{L}$. $\mathbf{A}\,\mathbb{E}[\mathbf{x}_t \,|\, \mathbf{y}^t] = \mathbf{A}\,\mathbf{L}\,\mathbf{y}_t$. Assuming an invertible $\mathbf{A}$, we have that

$$\mathbb{E}[\mathbf{x}_t \,|\, \mathbf{y}^t] = \mathbf{L}\,\mathbf{y}_t$$

i.e. that a forecast of $\mathbf{x}_t$ using all past observables $\mathbf{y}^t$ is equivalent to just using the current observables vector $\mathbf{y}_t$. □

A.4 Proof of Theorem 1

Proof. Consider the case $j = 2$. By the definition of Frobenius norm, we have

$$\begin{aligned}
\|\mathbf{B}_2^\infty\| &= \|\mathbf{G}(\mathbf{A} - \mathbf{K}\,\mathbf{G})\,\mathbf{K}\| \\
&= \sqrt{\mathrm{tr}\{\mathbf{K}^\top(\mathbf{A} - \mathbf{K}\,\mathbf{G})^\top\,\mathbf{G}^\top\,\mathbf{G}(\mathbf{A} - \mathbf{K}\,\mathbf{G})\,\mathbf{K}\}} \\
&= \sqrt{\mathrm{tr}\{(\mathbf{A} - \mathbf{K}\,\mathbf{G})^\top(\mathbf{G}^\top\,\mathbf{G})(\mathbf{A} - \mathbf{K}\,\mathbf{G})(\mathbf{K}\,\mathbf{K}^\top)\}} \\
&= \sqrt{\mathrm{tr}\left\{[N(\mathbf{A} - \mathbf{K}\,\mathbf{G})]^\top\left(\frac{1}{N}\,\mathbf{G}^\top\,\mathbf{G}\right)[N(\mathbf{A} - \mathbf{K}\,\mathbf{G})]\left(\frac{1}{N}\,\mathbf{K}\,\mathbf{K}^\top\right)\right\}} \\
&\leq r\,\|N(\mathbf{A} - \mathbf{K}\,\mathbf{G})\|^2\left\|\frac{1}{N}\,\mathbf{K}\,\mathbf{K}^\top\right\|\left\|\frac{1}{N}\,\mathbf{G}^\top\,\mathbf{G}\right\|
\end{aligned} \tag{A.10}$$

By the matrix Ricatti equation (A.4), we have

$$\frac{1}{N}\left\|\mathbf{K}\,\mathbf{K}^\top\right\| \leq \lambda_{\max}(\mathbf{R})\left\|\mathbf{K}\,\mathbf{K}^\top\right\| = \mathbf{\Sigma}_\infty - \mathbf{C}\,\mathbf{C}^\top - (\mathbf{A} - \mathbf{K}\,\mathbf{G})\mathbf{\Sigma}_\infty(\mathbf{A} - \mathbf{K}\,\mathbf{G})^\top$$

By Lemma 1, as $N \to \infty$, the RHS goes to $\mathbf{0}$. Thus, we have $\frac{1}{N}\left\|\mathbf{K}\,\mathbf{K}^\top\right\| \to 0$ as $N \to \infty$. Take $N \to \infty$ of equation (A.10) we obtain $\|\mathbf{B}_2^\infty\| \to 0$. Clearly, the case $j > 2$ can be proved in the same way, as $(\mathbf{A} - \mathbf{K}\,\mathbf{G})^j \to \mathbf{0}$. Inspecting equation (A.10), we can further conclude that for all $j \geq 1$

$$\limsup_{N\to\infty} N^{j-1}\left\|\mathbf{B}_j^\infty\right\| < \infty.$$

Given that $\left\|\mathbf{B}_j^\infty\right\| \to 0$ at rate N^{j-1} for all $j \geq 2$, the Weierstrass M-test asserts that the infinite-order VAR (2) collapses to the first-order VAR (4) as $N \to \infty$, as claimed. □

A.5 Proof of Theorem 2

Proof. Using the infinite-order VAR (2), we can write the DFM likelihood as

$$\begin{aligned}\ell^0(\mathbf{y}^t; \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}) &= \sum_{t=2}^{T} \ell^0(\mathbf{y}_t \mid \mathbf{y}^{t-1}; \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}) \\ &= -\frac{1}{2} \sum_{t=2}^{T} \left\{ \log|\mathbf{\Omega}| + \left(\mathbf{y}_t - \sum_{j=1}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right)^\top \mathbf{\Omega}^{-1} \left(\mathbf{y}_t - \sum_{j=1}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right) \right\}\end{aligned}$$

where $\mathbf{\Omega} = \mathbf{G}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top + \mathbf{R}$ is the variance-covariance matrix of the innovation. Similarly, using the first-order VAR (4), we can write the likelihood as

$$\begin{aligned}\ell^1(\mathbf{y}^t; \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}) &= \sum_{t=2}^{T} \ell^1(\mathbf{y}_t \mid \mathbf{y}_{t-1}; \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}) \\ &= -\frac{1}{2} \sum_{t=2}^{T} \left\{ \log|\mathbf{\Omega}| + \left(\mathbf{y}_t - \mathbf{B}_1^\infty \mathbf{y}_{t-1} \right)^\top \mathbf{\Omega}^{-1} \left(\mathbf{y}_t - \mathbf{B}_1^\infty \mathbf{y}_{t-1} \right) \right\}\end{aligned}$$

Subtract the two expressions and we obtain

$$\begin{aligned}&\ell^0(\mathbf{y}^t; \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}) - \ell^1(\mathbf{y}^t; \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{R}) \\ &= \frac{1}{2} \sum_{t=2}^{T} \left\{ \left(\sum_{j=2}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right)^\top \mathbf{\Omega}^{-1} \left(\sum_{j=2}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right) + 2 \left(\sum_{j=2}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right)^\top \mathbf{\Omega}^{-1} (\mathbf{a}_t) \right\} \\ &\leq \frac{1}{2} \sum_{t=2}^{T} \left\{ \lambda_{\max}(\mathbf{\Omega}^{-1}) \left\| \sum_{j=2}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right\|^2 + 2 \left(\sum_{j=2}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right)^\top \mathbf{\Omega}^{-1} (\mathbf{a}_t) \right\}\end{aligned}$$

where $\lambda_{\max}(\mathbf{\Omega}^{-1})$ denotes the largest eigenvalue of $\mathbf{\Omega}^{-1}$ and $\mathbf{a}_t = \mathbf{y}_t - \sum_{j=1}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j}$.

It suffices to show that

1. $\mathbb{E} \left\| \sum_{j=2}^{t-1} \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right\|^2 \to 0$ for all $t = 1, \ldots, T$
2. $\limsup_{N \to \infty} \lambda_{\max}(\mathbf{\Omega}^{-1}) < \infty$
3. $\mathbb{E}(\mathbf{B}_j^\infty \mathbf{y}_{t-j})^\top \mathbf{\Omega}^{-1} \mathbf{a}_t = 0$ for all $j \geq 2$ and $t = 1, \ldots, T$

Claim 1. Fix $j \geq 2$ and $t \geq 1$. We have

$$\mathbb{E} \left\| \mathbf{B}_j^\infty \mathbf{y}_{t-j} \right\|^2 \leq \left\| \mathbf{B}_j^\infty \right\|^2 \mathbb{E} \left\| \mathbf{y}_{t-j} \right\|^2 = \left\| \mathbf{B}_j^\infty \right\|^2 \operatorname{tr}(\mathbf{G}\,\mathbf{\Sigma}_\mathbf{x}\,\mathbf{G}^\top + \mathbf{R})$$

where $\mathbf{\Sigma_x} = \mathbb{E}\|\mathbf{x}_t\|^2$. By Theorem 1 and Assumption 2, we have $\left\|\mathbf{B}_j^\infty\right\|^2 = O(N^{-2(j-1)})$ and $\text{tr}(\mathbf{G}\,\mathbf{\Sigma_x}\,\mathbf{G}^\top + \mathbf{R}) = O(N)$. Thus, $\mathbb{E}\left\|\mathbf{B}_j^\infty\,\mathbf{y}_{t-j}\right\|^2 = O(N^{3-2j}) \to 0\;\forall j \geq 2$. By Minkowski inequality, we have $\mathbb{E}\left\|\sum_{j=2}^{t-1}\mathbf{B}_j^\infty\,\mathbf{y}_{t-j}\right\|^2 \to 0$ for all $t = 1, \dots, T$.

Claim 2. Fix N. Recall from Theorem 1 that

$$\mathbf{\Omega} = \mathbf{G}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top + \mathbf{R}$$

Since $\mathbf{\Omega}$ is positive definite, we have $\lambda_{\max}(\mathbf{\Omega}^{-1}) = \frac{1}{\lambda_{\min}(\mathbf{\Omega})}$. Note that

$$\lambda_{\min}(\mathbf{\Omega}) \geq \lambda_{\min}(\mathbf{G}\,\mathbf{\Sigma}_\infty\,\mathbf{G}^\top) + \lambda_{\min}(\mathbf{R}) \geq \lambda_{\min}(\mathbf{R})$$

Thus, $\lambda_{\max}(\mathbf{\Omega}^{-1}) \leq \frac{1}{\lambda_{\min}(\mathbf{R})}$. By condition 3 of Assumption 2, we have $\limsup_{N\to\infty}\lambda_{\max}(\mathbf{\Omega}^{-1}) \leq \frac{1}{\liminf_{N\to\infty}\lambda_{\min}(\mathbf{R})} < \infty$

Claim 3. Fix $j \geq 2$. Clearly, we have $\mathbf{B}_j^\infty\,\mathbf{y}_{t-j} \in \mathcal{H}(\mathbf{y}^{t-1})$ and $\mathbf{\Omega}^{-1}\,\mathbf{a}_t \in \mathcal{H}(\mathbf{a}_t)$. Then by the orthogonality condition $\mathbf{a}_t \perp \mathcal{H}(\mathbf{y}^{t-1})$, we have $\mathbb{E}(\mathbf{B}_j^\infty\,\mathbf{y}_{t-j})^\top\mathbf{\Omega}^{-1}\,\mathbf{a}_t = 0$, as desired.

By the three claims, the proof is complete and we conclude that

$$\lim_{N\to\infty}\int \log\frac{\mathcal{M}_{\theta_0}(\mathbf{y}^t)}{\mathcal{P}_\rho(\mathbf{y}^t)}\mathcal{M}_{\theta_0}(\mathbf{y}^t)d\,\mathbf{y}^t = \lim_{N\to\infty}\mathbb{E}\left[\ell^0(\mathbf{Y};\mathbf{A},\mathbf{C},\mathbf{G},\mathbf{R}) - \ell^1(\mathbf{Y};\mathbf{A},\mathbf{C},\mathbf{G},\mathbf{R})\right] = 0$$

□

A.6 Proof of Proposition 2

Proof. WLOG, let $\mathbf{c}_{ss} = \mathbf{0}$. As shown in Auclert et al. (2021a), up to first-order, the household's policy can be written as

$$\mathbf{c}_t = \sum_{j=0}^{\infty}\sum_{p\in\mathcal{P}}\frac{\partial\,\mathbf{c}}{\partial p_j}\,\mathbb{E}_t[\tilde{p}_{t+j}]$$

where $\frac{\partial\,\mathbf{c}}{\partial p_j} \in \mathbb{R}^N$ is the derivative of individual policy wrt. the j-period ahead aggregate input $p \in \mathcal{P}$ and $\tilde{p}_{t+j}$ denotes the deviation of p from its steady-state value.

Put $\tilde{p}_{t:} := (\tilde{p}_t, \tilde{p}_{t+1}, \dots)^\top$. Using the impulse response functions, we can write

$$\begin{aligned}\mathbb{E}_t[\tilde{p}_{t:}] &= \mathbb{E}_{t-1}[\tilde{p}_{t:}] + \mathcal{I}_e^p\boldsymbol{\epsilon}_t \\ &= F\,\mathbb{E}_{t-1}[\tilde{p}_{t-1:}] + \mathcal{I}_e^p\boldsymbol{\epsilon}_t\end{aligned}$$

where F is the shift forward operator. Iterate backward and we obtain the MA representation

$$\mathbb{E}_t[\tilde{p}_{t:}] = \sum_{j=0}^{\infty} F^j \mathcal{I}_e^p \boldsymbol{\epsilon}_{t-j}$$

Let $\mathcal{J}_p^c$ be the infinite-dimensional matrix of which the j column is $\frac{\partial \mathbf{c}}{\partial p_j}$. Substitute back into the policy function:

$$\mathbf{c}_t = \sum_{p\in\mathcal{P}} \mathcal{J}_p^c \, \mathbb{E}_t[\tilde{p}_{t:}] = \sum_{p\in\mathcal{P}} \mathcal{J}_p^c \sum_{j=0}^{\infty} F^j \mathcal{I}_e^p \boldsymbol{\epsilon}_{t-j} = \sum_{j=0}^{\infty} \underbrace{\sum_{p\in\mathcal{P}} \mathcal{J}_p^c F^j \mathcal{I}_e^p}_{\Psi_j^c} \boldsymbol{\epsilon}_{t-j}$$

□

A.7 Other derivations

Claim. For idiosyncratic income given by

$$y_t(z) = \Gamma_t(z) w_t N_t e^{v_t^r} \tag{A.11}$$

$$\Gamma_t(z) = \frac{z\left(\frac{w_t N_t}{w_{ss} N_{ss}}\right)^{\gamma_y(z)} \left(e^{v_t^r}\right)^{-\gamma_r(z)}}{\mathbb{E}_z\left[z\left(\frac{w_t N_t}{w_{ss} N_{ss}}\right)^{\gamma_y(z)} \left(e^{v_t^r}\right)^{-\gamma_r(z)}\right]} \tag{A.12}$$

subject to $\mathbb{E}[z] = \mathbb{E}[z\gamma_y(z)] = \mathbb{E}[z\gamma_r(z)] = 1$. Then the elasticity of idiosyncratic income to aggregate income is

$$\frac{\partial \log y_t(z)}{\partial \log w_t N_t} = \gamma_y(z)$$

Taking logarithms and taking the partial obtains

$$\log y_t = \log \Gamma_t(z) + \log w_t N_t + v_t^r \tag{A.13}$$

$$\frac{\partial \log y_t(z)}{\partial \log w_t N_t} = \frac{\partial \log \Gamma_t(z)}{\partial \log w_t N_t} + 1 \tag{A.14}$$

Moroever

$$\log \Gamma_t(z) = \log z + \gamma_y(z)(\log w_t N_t - \log w_{ss} N_{ss}) - \gamma_r(z) v_t^r - \log \mathbb{E}_z \left[z \left(\frac{w_t N_t}{w_{ss} N_{ss}} \right)^{\gamma_y(z)} \left(e^{v_t^r} \right)^{\gamma_r(z)} \right] \tag{A.15}$$

$$= \gamma_y(z) - \frac{\mathbb{E}_z \left[z \gamma_y(z) \left(\frac{w_t N_t}{w_{ss} N_{ss}} \right)^{\gamma_y(z)-1} \left(e^{v_t^r} \right)^{\gamma_r(z)} \right]}{\mathbb{E}_z \left[z \left(\frac{w_t N_t}{w_{ss} N_{ss}} \right)^{\gamma_y(z)} \left(e^{v_t^r} \right)^{\gamma_r(z)} \right]} \tag{A.16}$$

and evaluating $w_t N_t = w_{ss} N_{ss}$ and $v_t^r = 0$ and imposing the normalizations $\mathbb{E}[z] = \mathbb{E}[z\gamma_y(z)] = 1$ implies

$$\frac{\partial \log y_t(z)}{\partial \log w_t N_t} = \gamma_y(z) - 1 + 1 = \gamma_y(z) \tag{A.17}$$

Claim. Given the above setting, the elasticity of idiosyncratic income to monetary policy shock is

$$\frac{\partial \log y_t(z)}{\partial v_t^r} = \gamma_y(z) \frac{\partial \log w_t N_t}{\partial v_t^r} - \gamma_r(z) \tag{A.18}$$

Equation (A.13) implies

$$\frac{\partial \log y_t(z)}{\partial v_t^r} = \frac{\partial \log \Gamma_t(z)}{\partial v_t^r} + \frac{\partial \log w_t N_t}{\partial v_t^r} - 1 \tag{A.19}$$

Moreover, (A.15) implies

$$\frac{\partial \log \Gamma_t(z)}{\partial v_t^r} = \gamma_y(z) \frac{\partial \log w_t N_t}{\partial v_t^r} - \gamma_r(z) - \tag{A.20}$$

$$\frac{\partial w_t N_t}{\partial v_t^r} \frac{\mathbb{E}_z \left[z \gamma_y(z) \left(\frac{w_t N_t}{w_{ss} N_{ss}} \right)^{\gamma_y(z)-1} \left(e^{v_t^r} \right)^{-\gamma_r(z)} \right]}{\mathbb{E}_z \left[z \left(\frac{w_t N_t}{w_{ss} N_{ss}} \right)^{\gamma_y(z)} \left(e^{v_t^r} \right)^{-\gamma_r(z)} \right]} + \frac{\mathbb{E}_z \left[z \gamma_r(z) \left(\frac{w_t N_t}{w_{ss} N_{ss}} \right)^{\gamma_y(z)-1} \left(e^{v_t^r} \right)^{-\gamma_r(z)} \right]}{\mathbb{E}_z \left[z \left(\frac{w_t N_t}{w_{ss} N_{ss}} \right)^{\gamma_y(z)} \left(e^{v_t^r} \right)^{-\gamma_r(z)} \right]} \tag{A.21}$$

Evaluating $w_t N_t = w_{ss} N_{ss}$ and $v_t^r = 0$ and imposing the normalizations $\mathbb{E}[z] = \mathbb{E}[z\gamma_y(z)] = \mathbb{E}[z\gamma_r(z)] = 1$ obtains

$$\frac{\partial \log y_t(z)}{\partial v_t^r} = \gamma_y(z) \frac{\partial \log w_t N_t}{\partial v_t^r} - \gamma_r(z) - \frac{\partial \log w_t N_t}{\partial v_t^r} + 1 + \frac{\partial \log w_t N_t}{\partial v_t^r} - 1 \tag{A.22}$$

$$= \gamma_y(z) \frac{\partial \log w_t N_t}{\partial v_t^r} - \gamma_r(z) \tag{A.23}$$

Appendix B Algorithms

B.1 Dynamic Mode Decomposition

Another option for computing the reduced-rank VAR(1) matrix $\mathbf{B}_r$ is to employ the Dynamic Mode Decomposition (DMD) algorithm. The DMD has become a workhorse tool in the fluid dynamics literature, originally introduced by Schmidt and Sesterhenn (2010) and later developed by Tu et al. (2014). Existing applications of the DMD also include epidemiology, neuroscience and video processing (see Brunton and Kutz (2022)).

Given simulated data matrices $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{Y}}'$, the DMD estimates the reduced-rank VAR associated with the simulated data by solving

$$\mathbf{B}_r = \underset{\text{rank}(B)=r}{\arg\min} \left\| \widetilde{\mathbf{Y}}' - B\widetilde{\mathbf{Y}} \right\| \tag{B.1}$$

where $\|\cdot\|$ denotes the Frobenius norm. To compute $\mathbf{B}_r$, represent $\mathbf{Y}$ with a reduced Singular Value Decomposition (SVD)

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^\top$$

where $\widetilde{\mathbf{U}}$ is $N \times N$, $\widetilde{\mathbf{\Sigma}}$ is $N \times N$ and $\widetilde{\mathbf{V}}$ is $T \times N$. We compress $\widetilde{\mathbf{Y}}$ by using its r largest singular values:

$$\widetilde{\mathbf{Y}} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U} = \widetilde{\mathbf{U}}[:, :\ r]$, $\mathbf{\Sigma} = \widetilde{\mathbf{\Sigma}}[:\ r, :\ r]$ has r singular values as its only non-zero entries, and $\mathbf{V}^\top = \widetilde{\mathbf{V}}^\top[:\ r, :]$. Here $\mathbf{U}$ is $N \times T$, $\mathbf{V}$ is $T \times r$, $\mathbf{\Sigma}$ is $r \times r$, and $\mathbf{V}^\top$ is $r \times T$.[32]

We use this reduced-order SVD approximation of $\widetilde{\mathbf{Y}}$ to compute

$$\mathbf{B}_r = \widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}^+, \tag{B.2}$$

where by construction $\mathbf{B}_r$ is rank r.[33] The covariance matrix of the residuals, $\widetilde{\mathbf{a}}_t = \widetilde{\mathbf{y}}_t - \mathbf{B}_r\,\widetilde{\mathbf{y}}_{t-1}$, is computed via

$$\mathbf{\Omega}_r = \tfrac{1}{T-1}\sum_{t=1}^{T} \widetilde{\mathbf{a}}_t\widetilde{\mathbf{a}}_t^\top \tag{B.3}$$

[32] Note that all we need here is a truncated SVD, which can be very efficiently computed using existing machine-learning packages (e.g. scikit-learn).

[33] The DMD also provides estimates of the underlying factors $\mathbf{x}_t$ as well as estimates of $\mathbf{G}$ and $\mathbf{A}$. Though we do not use them here, see **?**, sec. 2.1 for full details of the DMD algorithm and connections with linear state-space models.

Appendix C Comparison with other estimation methods

C.1 Conventional approach for likelihood computation

In this section, we compare the speed of our method against two alternatives. The first is the conventional likelihood computation of the moving-average representation, employed by Auclert et al. (2021b). By Proposition 2, the data has a MA representation

$$\mathbf{c}_t = \sum_{j=0}^{\infty} \Theta_j \boldsymbol{\epsilon}_{t-j} + \mathbf{v}_t, \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \Sigma_e)$$

where $\mathbf{v}_t \sim N(\mathbf{0}, \mathbf{R})$ is measurement error.

For data matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$, the likelihood computation involves vectorizing the data matrix into a $NT \times 1$ vectors and computing the $NT \times NT$ covariance matrix $\mathbf{V}$. Then the likelihood is proportional to

$$\mathcal{L}(\mathbf{y}_1, \ldots, \mathbf{y}_T; \theta) = \frac{1}{2} \det \log \mathbf{V} - \frac{1}{2} \boldsymbol{\Theta}^\top \mathbf{V}^{-1} \boldsymbol{\Theta}$$

where $\boldsymbol{\Theta} = [\Theta_1, \ldots, \Theta_J]^\top$. The computational cost is solely to evaluate $\mathbf{V}^{-1}$, which Auclert et al. (2021b) state requires time proportional to N^3T^3 and thus scales poorly with N or T.

C.2 Whittle likelihood approximation

An approach that scales better with N and T is the Whittle approximation to the likellihood, as in Hansen and Sargent (1981) and Plagborg-Møller (2019). The likelihood is approximated by

$$\mathcal{L}(\mathbf{y}_1, \ldots, \mathbf{y}_T; \theta) = -\frac{1}{2} \sum_{j=0}^{T-1} \left[\log 2\pi + \log(\det S(\omega_j; \theta)) + \text{tr}(S(\omega_j; \theta)^{-1} I(\mathbf{c}; \omega_j)) \right] \tag{C.1}$$

where $\omega_j := \frac{2\pi j}{T}$, $S(\omega_j; \theta)$ is the spectral density of $\mathbf{c}$ at frequency ω_j, and $I(\mathbf{c}; \omega_j)$ is the periodogram of the data at frequency ω_j. By definition, the periodogram is given by

$$I(\mathbf{c}; \omega_j) := \frac{1}{T} \left(\sum_{t=1}^{T} \mathbf{c}_t \exp(-i\omega_j t) \right) \left(\sum_{t=1}^{T} \mathbf{c}_t \exp(i\omega_j t) \right)' \tag{C.2}$$

By the MA representation, the spectral density is given by

$$S(\omega_j; \theta) = \left(\sum_{j=0}^{\infty} \Psi_j^c(\theta) \exp(-i\omega_j j) \right) \Sigma_e \left(\sum_{j=0}^{\infty} \Psi_j^c(\theta) \exp(i\omega_j j) \right)' + \mathbf{R}$$

Note that both $I(\mathbf{c}; \omega_j)$ and $S(\omega_j; \theta)$ are M-dimensional matrices and can be computed by applying the Discrete Fourier Transform to the data matrix $\mathbf{c}$ and MA coefficient array $\{\Psi_j^c : j = 0, \ldots, T\}$.

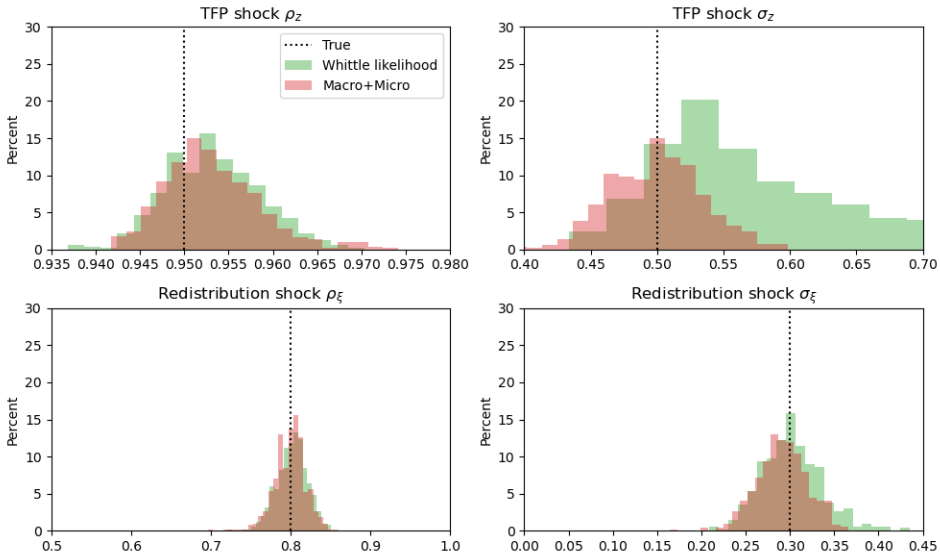

Figure C.1: Finite sample parameter distribution across 500 Monte-Carlo samples

Also, the symmetry of Fourier transform implies that we only need to evaluate the summands in (C.1) for $j = 0, \ldots, \lfloor \frac{T-1}{2} \rfloor$

C.2.1 Comparison of finite-sample parameter distribution

In this section, we implement the Monte-Carlo exercise in Section 4.4 for the Whittle approximation of the likelihood. Figure C.1 plots the distribution of the estimated parameters from 500 Monte-Carlo samples each of length $T = 120$, comparing it our method. The Whittle approximation delivers a distribution close to ours, for ρ_z and ρ_ξ. The distribution of σ_ξ is slightly to the right and has a mean further away from the true value; and the distribution for σ_ϵ is much more dispersed and has a bias compared to our method.

One reason for these differences is that the periodogram and associated approximation quality depends on large T, as seen in equation (C.2). This is unlike our method, where the approximation quality depends on large N.

C.3 Comparison of computing times

Figure C.1 compares the computing times for 100 likelihood evaluations for our method ("Low-rank approximation"), Whittle approximation (Section C.2, and the MA likelihood computation (Section C.1).

For small N, all three methods have comparable computing times, around 11 seconds.

Computing times for our method increases slowly as N increases, rising to 12.6 seconds for $N = 100$. Computing times are similar for the Whittle approximation, which rises to 13.2 seconds for $N = 100$, around 1.08 times larger.

On the other hand, computing times for the exact method increases exponentially: the

	Number of observables, N					
	2	5	10	20	50	100
Low-rank approximation	11.2	11.3	11.4	11.5	11.7	12.6
Whittle approximation	11.2	11.3	11.4	11.6	11.6	13.2
Exact	11.3	11.3	12.2	15.7	53.7	240.8

Table C.1: Summary of computing times for 100 likelihood evaluations

computation takes 4 minutes for 100 evaluations of the likelihood. This implies that a full-blown Monte-Carlo simulations with 100,000 draws will take around 70 hours, or 3 days. In comparison, our method takes around 3.5 hours and Whittle approximation takes around 3.6 hours.

Appendix D Calibration of Krusell-Smith model in Section 4

Table D.2 presents the calibrated parameters used in the Monte-Carlo exercise in Section 4.

Parameter	Interpretation	Value
σ	Inverse EIS	1.00
δ	Depreciation	0.025
α	Labor share	0.11
ρ_e	Persistence of idiosyncratic productivity	0.967
σ_e	Std of idiosyncratic productivity	0.50
ρ_z	Persistence of TFP	0.95
ρ_ξ	Persistence of income risk shock	0.80
σ_z	Std of TFP shock	0.50
σ_ξ	Std of income risk shock	0.30

Table D.2: Calibrated parameters

Appendix E Verifying rank condition

E.1 Low-rank assumption in other benchmark models

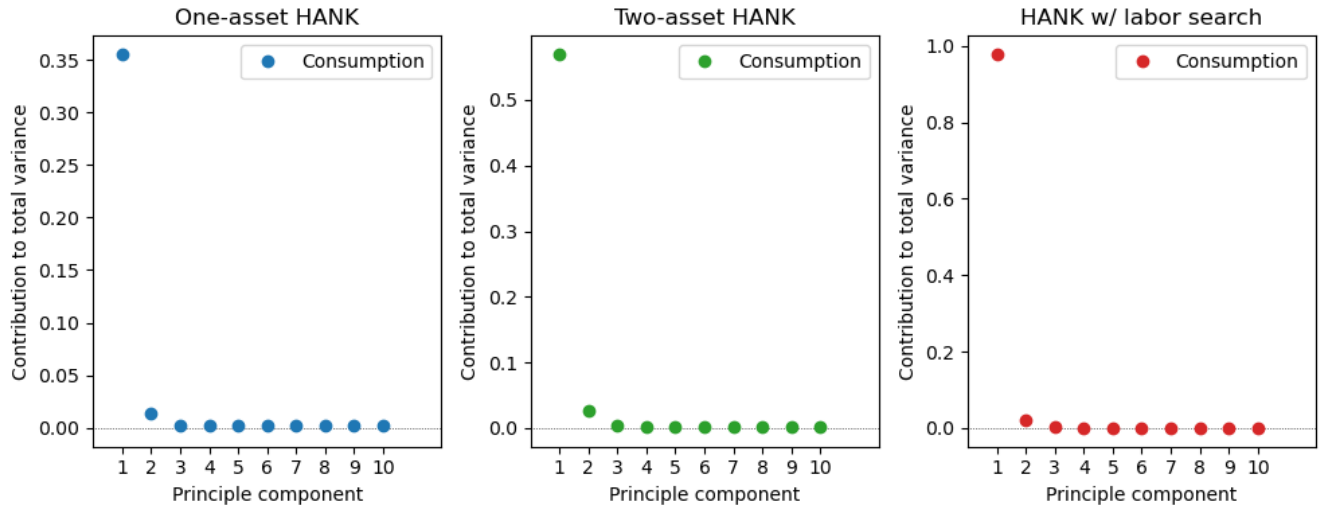

Figure E.2: Contribution to total variance of 10 leading principle components

E.2 Section 4: Krusell-Smith model with redistributive shocks

		Number of factors, N							
		1	2	3	4	5	6	7	8
Check 1.	$IC(N)$	304.88	304.03	303.13	303.11	303.10	303.10	303.10	303.10
Check 2.	Singular Values	334.8	17.7	12.6	11.1	10.9	10.6	10.6	10.5
Check 3.	$\max \lvert \hat{E}[\mathbf{a}_{t+1}\ \mathbf{a}_t] \rvert$	0.79	0.27	0.27	0.27	0.27	0.27	0.27	0.27
Check 4.	R^2_r	0.86	0.892	0.895	0.895	0.895	0.895	0.895	0.895

Table E.3: Rank estimation of model

E.3 Section 5: Heterogeneous-agent New Keynesian model

		Number of factors, N									
		1	2	3	4	5	6	7	8	9	10
Check 1.	$IC(N)$	-9.14	-9.13	-9.26	-9.36	-9.43	-9.37	-9.35	-9.30	-9.24	-9.17
Check 2.	Singular Values	9.77	3.28	1.73	0.43	0.36	0.21	0.19	0.12	0.10	0.10
Check 3.	$\max \lvert \hat{E}[\mathbf{a}_{t+1}\ \mathbf{a}_t] \rvert$	0.097	0.072	0.037	0.016	0.010	0.010	0.003	0.006	0.006	0.006
Check 4.	R^2_r	0.33	0.35	0.47	0.55	0.62	0.62	0.63	0.63	0.63	0.63

Table E.4: Rank estimation of model

Appendix F Additional charts and tables

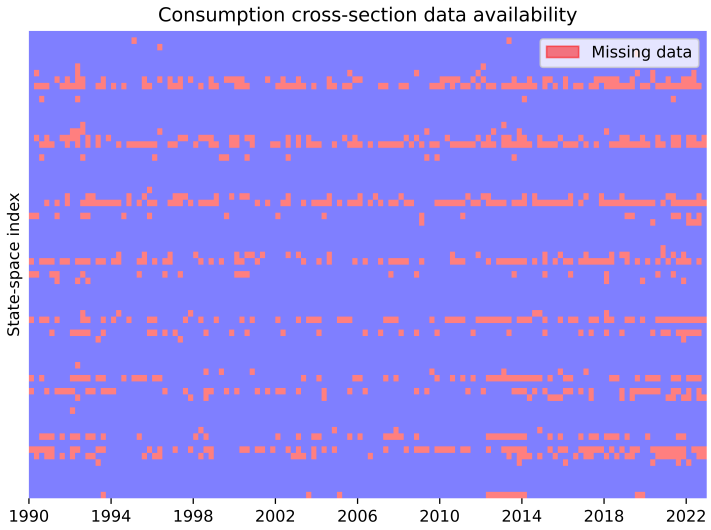

Figure F.3: Heatmap of missing data for micro data