Distributional Dynamics*

Christian Bayer†
University of Bonn, CEPR, IZA, and CESifo

Luis Calderon‡
University of Bonn

Moritz Kuhn§
University of Mannheim CEPR, IZA, and CESifo

March 20, 2026

Abstract

We develop a new method for deriving high-frequency synthetic distributions of consumption, income, and wealth. These synthetic data incorporate different sources of microdata, and our method can exploit these sources regardless of their frequency or variable coverage. Core to the method is treating distributional data as a time series of functions, whose underlying factor structure follows a state-space model estimated using Bayesian techniques. The method is generic enough to cover the dynamics of joint distributions in general. Using a wide range of U.S. microdata, we demonstrate that this novel approach yields high-quality, high-frequency distributional data on consumption, income, and wealth that are of interest to modern theories of macroeconomic dynamics.

Keywords: *Consumption, income, and wealth inequality; Macroeconomic dynamics; Dynamic state-space model; Functional time-series data; Bayesian statistics*
JEL Classification: *E21, E32, E37, D31, C32, C55*

*We thank Nazarii Salish for discussion at early stages of this project and Lisa Dähne for research assistance. We thank Frank Schorfheide and we thank seminar participants at the Banca d'Italia, Hausdorff Center for Mathematics, HEC Lausanne, KU Leuven, National University of Singapore, Paris School of Economics, and summer SED 2024 for helpful comments. Christian Bayer gratefully acknowledges funding through the ERC-CoG project Liquid-House-Cycle funded by the European Unions Horizon 2020 Program under grant agreement No. 724204. Bayer and Kuhn gratefully acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2126/1 – 390838866) and through CRC TR 224 (Projects A03 and C05). Replication package and data: The code to reproduce all results, together with processed survey moments and reconstructed distributional time series (posterior summaries), is available at: https://luiscald.github.io/distributional-dynamics/.

†*Corresponding author.* Institute for Macroeconomics and Econometrics—University of Bonn, christian.bayer@uni-bonn.de.

‡Institute for Macroeconomics and Econometrics—University of Bonn, luis.calderon@uni-bonn.de.

§Department of Economics—University of Mannheim, mokuhn@uni-mannheim.de.

1 Introduction

Understanding the dynamics of the joint distribution of consumption, income, and wealth is central to understanding macroeconomic dynamics, the transmission of monetary and fiscal policy, and their cross-sectional effects.[1] The scarcity of high-frequency information on the joint distribution is a significant limitation in this endeavor. We propose a novel and general technique based on functional data analysis and Bayesian time-series methods to obtain high-frequency estimates of this (or other) joint distribution(s). This technique is flexible enough to combine distributional data from different microdata sources with aggregate data, even when these data are of mixed frequency and only one micro-dataset contains all variables of interest and others only a subset.

The challenge is that the joint distributions of consumption, income, and wealth are functional data and hence, in principle, infinite-dimensional objects. Our novel method, however, projects the infinite-dimensional objects onto a *finite* set of basis functions and then exploits standard statistical dimensionality reduction techniques. These techniques reduce the dynamics of the distribution to a factor model that has a state-space representation. This step is motivated by the heterogeneous agent macroeconomics literature which suggests that a small number of factors is sufficient to approximate distributional dynamics. Intuitively, this mirrors the fact that a small set of aggregate prices shapes household decisions and therefore affects the distribution of consumption, income, and wealth (Auclert, Bardóczy, Rognlie, and Straub 2021; Bayer, Born, and Luetticke 2024). In fact, more reduced form inequality research (Di Maggio, Kermani, and Majlesi 2020; Chodorow-Reich, Nenov, and Simsek 2021; Kuhn, Schularick, and Steins 2020) has so far found much support for the price-distribution nexus, too. These prices are closely tied to the aggregate economy, which itself has a low dimensional factor structure (Stock and Watson 2002).

This set of findings from the macroeconomic literature has two implications for the joint evolution of aggregate and distributional data: First, the dynamics of the distributional data can be represented by a medium-size state-space model. Second, the states of this model are driven by a small set of aggregate factors and unobserved distributional shocks. The combination of these two facts is the key innovation to overcome the challenge of dealing with multidimensional functional data. In practice, we use factor analysis to uncover the lower-dimensional state-space representation of the distributional dynamics and its ag-

1. See, e.g., Mian, Straub, and Sufi (2020), Holm, Paul, and Tischbirek (2021), Andersen, Johannesen, Jørgensen, and Peydró (2023), Bhandari, Evans, Golosov, and Sargent (2021), and Bayer, Luetticke, Pham-Dao, and Tjaden (2019).

gregate drivers. Given these factor structures for the aggregate and distributional data, we estimate the time-series behavior of the functional, i.e., distributional, data using Bayesian techniques and link aggregate and distributional data without imposing a fully-structural macroeconomic model. This makes the method general such that it can be applied to a wide array of distributional dynamics.

The state-space representation lends itself naturally to the use of the Kalman-filter for Bayesian estimation of the state-space model. This has several important advantages. It allows us to use and merge numerous micro-datasets that refer to the same economic object but with different operationalized measures, e.g., differences in the sources of income covered. These different operationalizations, alongside sampling uncertainty, are captured by the measurement error in the observation equation of the model. The observation equation also allows for the combination of micro-datasets with different sampling frequencies and also allows us to exploit information from microdata that contain only a subset of the variables of interest (Schorfheide and Song 2015; Durbin and Koopman 2012).

Finally, we overcome the limited availability of high-frequency distributional information and construct estimates of business cycle fluctuations for joint distributions at any point in time, including periods where microdata on distributions are unavailable. The estimated state-space model allows us to construct synthetic high-frequency distributional data by means of the Kalman smoother. The synthetic data itself then incorporate the information of various micro- and aggregate data sources. Although the synthetic distributions are originally functional, they can be expressed approximately in the form of repeated cross sections of microdata with consumption, income, and wealth of synthetic households.

To demonstrate the power of the novel method, we apply the new estimation technique to a rich set of U.S. household microdata from the *Panel Study of Income Dynamics* (PSID), the *Survey of Consumer Finances* (SCF), the *Consumption Expenditure Survey* (CEX), the *Survey of Income and Program Participation* (SIPP), and the *Current Population Survey* (CPS). Using the various microdata jointly, our method overcomes three existing challenges. First, only the PSID contains all three variables of interest: consumption, income, and wealth. In the other micro-datasets, at least one of the variables is absent. At the same time, *all* of the micro-datasets contain information on the joint distribution of consumption, income, and wealth. Second, all of these datasets are available at different frequencies. Third, they differ in sampling approaches and details of measurement concepts. Our method deals with all three challenges. We complement these microdata with a comprehensive set of macroeconomic time series.

From the estimation of the state-space model on these data, we then construct

high-frequency synthetic distributional data represented by groups of households. Each group is defined by a particular combination of quantiles of consumption, income, and wealth. Over time, the conditional expectations for each quantile change and so do the consumption, income, and wealth of each group. The population weight reflects how likely it is to observe combinations of quantiles, and therefore also the weight changes over time. Thus, the dynamics of the population weights induce the dynamics of the cross-sectional correlations in the three variables. We construct the detrended business cycle variations of the joint distribution in consumption, income, and wealth from 1962 to 2024.

We carefully validate each step of the estimation procedure. First, we show that the factor representation of the distributional data imposes almost no loss of information compared to the information provided by the microdata when observed. Second, we show that the model closely predicts the distributional data, even when unobserved, through significant comovement with aggregates. Specifically, we show this for the consumption distributions of the CEX and the wealth distributions of the SCF. Finally, we use simulated data from a theoretical HANK model as a data generating process and show that the method can closely track the true high frequency movements of the distribution given the typical sample sizes and frequencies of micro data. Thus, we conclude that the imputed distributional data stemming from our statistical model are reliable.[2]

Finally, we illustrate the usefulness of the method in two applications. First, we use the estimated distributional state-space model to quantify what drives distributional fluctuations. A forecast error variance decomposition shows that aggregate shocks account for the bulk of the variation in the distributional factors (at least 70% for every factor and typically above 85%), and similarly for distributional moments such as conditional means of consumption, income, and wealth. Then, we show consumption dynamics along the joint distribution of income and wealth across the last three U.S. recessions. This application provides a natural stepping stone for future research to better understand particular distributional channels. Furthermore, it provides novel empirical estimates on various groups of households in the joint distribution of consumption, income, and wealth that remain unobserved in existing data—providing new empirical model targets to guide the modeling of business cycles with heterogeneous agents.

2. In addition, we also validate the choice of priors in Bayesian estimation, particularly with respect to measurement error, and show that the state-space model is consistent with the sampling uncertainty of the observed distributional data at the sampling points. We further compare the prediction for the dynamics of income and wealth distribution with that implied by the distributional flow of funds methodology (DFA, see Batty, Bricker, Briggs, Friedman, Nemschoff, Nielsen, Sommer, and Volz 2020) and that estimated by the World Inequality Database (Alvaredo, Atkinson, Chancel, Piketty, Saez, and Zucman 2016; Piketty, Saez, and Zucman 2018). These results can be found in Appendix G.

The observed dynamics are also economically interesting. We find overwhelming evidence that the impact of macroeconomic shocks on household consumption (relative to the aggregate) is not uniform across the household groups or across recessions. We see this as a complementary work to Bilbiie, Galaasen, Gürkayna, Mæhlum, and Molnar (2025), showing for Norway that automatic stabilizers *on average* render household heterogeneity largely irrelevant for aggregate dynamics. Our analysis for the United States, however, shows a more nuanced, state-dependent role for heterogeneity, where the nature of the prevailing shock in a recession is a critical determinant of how distributional dynamics affect the macroeconomy. As we expand on this later, these findings show that the joint distribution of income and wealth is crucial for understanding consumption dynamics around recessions and policy decisions.

The remainder of this paper is organized as follows: Section 2 provides an overview of the relevant literature. Section 3 develops our estimation method and Section 4 evaluates its quality. Section 5 provides an application of the novel method. Finally, Section 6 concludes the paper. An appendix follows.

2 Literature

Methodological Antecedents. The paper most closely related to ours is Chang, Chen, and Schorfheide (2024), which develops a Bayesian state-space approach to estimate the coevolution of aggregate variables and the marginal distribution of earnings. Our work differs in three key aspects. First, we extend their approach to model the evolution of joint distributions over time, e.g., distributions of consumption, income, and wealth with also a richer set of aggregates. This allows us to capture the dependence structure across variables within the distribution, rather than focusing solely on marginal distributions, and to understand the interplay between the distribution and aggregates. Second, instead of directly estimating the distributional dynamics, we follow Kneip and Utikal (2001) and Tsay (2016) by additionally projecting the functional data on an ideal basis through PCA (see also Meeks and Monti 2023, for a macroeconomic application). We then estimate a state-space model in the lower-dimensional factor space—not possible with the functional data alone. Third, our paper focuses on generating high-frequency synthetic distribution data, addressing challenges related to missing observations and the mixed frequencies of aggregate and microdata.[3]

3. While there is a nascent literature (see, e.g., Ettmeier, Hyun Kim, and Schorfheide 2024) emphasizing the role of unit-level dynamics in properly addressing the mixed-frequency problem, we abstract from this consideration in the present study and follow the standard mixed-frequency approach.

More broadly, our method builds on the literature that formulates temporal disaggregation in a state-space framework (see e.g., Harvey and Chung 2000) and on functional data methods in economics (see e.g., Kneip and Utikal 2001; Diebold and Li 2006; Chang, Kim, and Park 2016). The key advantage over static methods such as Chow and Lin (1971) and Fernandez (1981) is that the state-space representation is inherently dynamic and lends itself to the Kalman filter for estimation and diagnostics.

Macroeconomic motivation. Our contribution relative to the growing body of work linking macroeconomic aggregates with distributional dynamics (see e.g., Baumeister et al. 2025; Koop et al. 2026; Ettmeier, Hyun Kim, and Schorfheide 2024) is to treat the *joint* distribution—marginals and their dependence structure—as the state variable itself, recovering its dynamics from multiple partially overlapping surveys at mixed frequencies. This provides the still-missing descriptions of the short-run dynamics of the consumption, income, and wealth distribution to the theoretical literature on heterogeneity and aggregate dynamics (Kaplan, Moll, and Violante 2018; Bayer et al. 2019; Bayer, Born, and Luetticke 2024) and extends the rapidly growing empirical work on the effects of policy shocks on marginal distributions.[4]

Through our application, we also speak to a central debate on macroeconomic amplification from household inequality. Auclert (2019) and Bilbiie (2020) formalize the idea that amplification depends on the covariance between the marginal propensity to consume and the cyclicality of income. Patterson (2023) finds this covariance to be strongly positive for U.S. labor earnings, creating a powerful "matching multiplier"; recent evidence from Bilbiie et al. (2025) for Norway provides a counterpoint, showing that once disposable income is considered, automatic stabilizers render the covariance negligible over *average* business cycles. Our synthetic data reveal that consumption distribution dynamics are potentially *very* business-cycle-specific, suggesting that neither finding generalizes unconditionally.

Distributional measurement. The paper also contributes to a large empirical literature, beginning with Piketty and Saez (2003), measuring trends and fluctuations in inequality. Most of this work produces high-frequency data for *marginal* distributions—income or wealth separately—but not for their joint distribution.

4. See, for example, Berger, Bocola, and Dovis (2023), Coibion et al. (2017), Cloyne, Ferreira, and Surico (2020), Holm, Paul, and Tischbirek (2021), Chang and Schorfheide (2024), and Bartscher et al. (2022). McKay and Wolf (2023) surveys this literature.

Blanchet, Saez, and Zucman (2022) construct monthly income and wealth distributions for the United States; Smith, Zidar, and Zwick (2023), using administrative data and the capitalization method of Saez and Zucman (2016), provide high-frequency wealth estimates concentrated on the upper tail. The Distributional Financial Accounts (Batty et al. 2020) produces quarterly balance-sheet estimates since 1989 by combining SCF microdata with the Financial Accounts through Chow-Lin/Fernández-type models. Our state-space approach offers three advantages over these methods: it treats the underlying microdata as noisy samples of a time series of distribution functions, explicitly handling sampling uncertainty in a dynamic setting; it can combine multiple microdata sources with different operationalizations of the same economic concepts; and it recovers the *joint* distribution of consumption, income, and wealth—not just individual marginals—going back to the 1960s.

3 Method

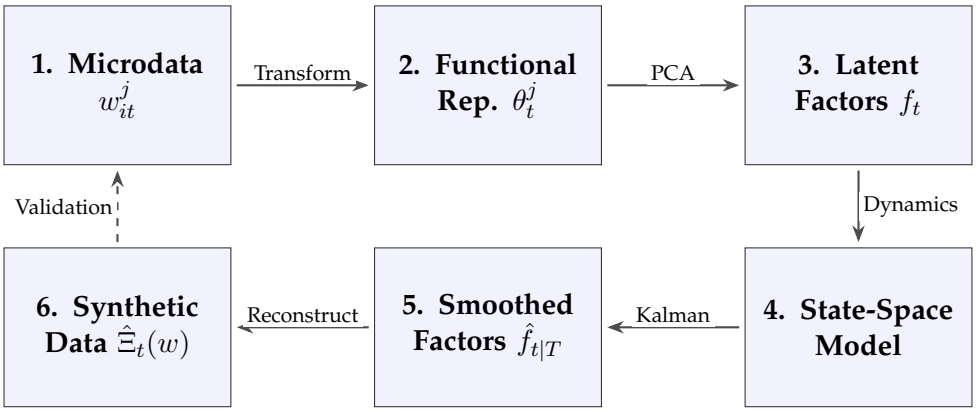


Figure 1: Visual Roadmap of the Methodology

This section describes a general method for generating high-frequency estimates of joint distributions of economic variables of interest over a large number of micro-units. The method proceeds in four main steps, see Figure 1. First, starting from microdata, we estimate and represent joint distributions by their quantile functions coupled with a copula, isolating marginal shapes from their dependence structure (Section 3.1), projecting these functions onto Legendre polynomials (Section 3.2.1). Second, we reduce dimensionality using functional principal component analysis (PCA) (Section 3.2.2). Third, we formulate the dynamics of these latent factors as a state-space model that handles mixed frequencies (Section 3.2.3). Fourth, we estimate the system using Bayesian techniques and re-

cover the full distributional time series via the Kalman smoother (Section 3.3). The reverse provides us with the high-frequency synthetic distributional data.

This method uses microeconomic and aggregate data as inputs. It requires only the joint observation of the microeconomic variables in at least one dataset over several, but potentially infrequent, time periods. The developed method treats the distributional data as functional data in a time-series state-space framework with unobserved states. In the following, we describe the method using the example of the joint distribution of consumption, income, and wealth, an important macroeconomic application.

3.1 Distributions as time series of functional data

We consider a sequence $\Xi_t(w)$ of multidimensional distribution functions (cdfs), indexed by $t \in \mathbb{T} := \{1 \dots T\}$, defined over a d-dimensional *vector* $w \in \mathbb{R}^d$. In the case of our application, we have $d = 3$, where w is a vector of consumption, income, and wealth at the household level. In addition to this sequence of distribution *functions*, there is a sequence of real-valued vectors Y_t of stationary aggregate data. In the following exposition, we assume that Y_t is observed at all times $t \in \mathbb{T}$. The extension to missing observations in Y_t is standard.

From the distributions $\Xi_t(w)$, we observe only randomly drawn samples. We allow these samples to come from different sampling procedures or to have different operationalizations of the underlying theoretical variables. For example, the PSID and the SCF use different sampling procedures and slightly different concepts of wealth and income. We index each of the sampling procedures/datasets by $j = 1 \dots J$. All of these different datasets are typically not observed for all time periods. Instead, dataset j is only observed in a particular subset $\mathcal{T}^j \subset \mathbb{T}$. In addition, not all samples contain all variables of interest, but may contain only a subset $\mathcal{D}^j \subseteq \mathbb{D} := \{1, \dots, d\}$ of variables. For example, the Current Population Survey (CPS) provides only income information but neither wealth nor consumption. However, at least one dataset, j, must contain all variables of interest for $\mathcal{D}^j = \mathbb{D}$. In our application, this dataset is the PSID, which contains information on consumption, income, and wealth (for some years).

Our goal is to obtain estimates of the joint distribution functions, $\hat{\Xi}_t(w)$, $\forall t \in \mathbb{T}$, by efficiently combining information from the various related microdata sources and aggregate information, Y_t. We assume that there is a time-series structure such that the density $d\Xi_t$ evolves according to the functional difference equation

$$d\Xi_{t+1} = G\left(d\Xi_t, Y_{t-\ell:t+1}\right) + \epsilon_{t+1}, \tag{1}$$

where $Y_{t-\ell:t+1}$ are observed contemporaneous and lag aggregate data; G determines the dynamics of the system, and ϵ_t are the corresponding shocks to the functional equation.[5] This time-series structure arises naturally in so-called HANK models (see e.g. Kaplan, Moll, and Violante 2018; Bayer, Born, and Luetticke 2024). It is assumed that Ξ_t is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and thus, $d\Xi_t$ exists and is continuous. Moreover, we require that the operator G is Lipschitz-continuous in its first argument, a condition satisfied mechanically under the proposed linear model. We provide details below.

Viewing the J sampling procedures as capturing the same fundamental object Ξ_t but with some measurement error, ν_t, allows us to combine the data in a systematic way. This means that a dataset gives us an estimate

$$d\tilde{\Xi}_t^j = \int_{\mathbb{D}\setminus\mathcal{D}^j} d\Xi_t + \nu_t \qquad \text{for} \quad t \in \mathcal{T}^j, \tag{2}$$

where the integral $\int_{\mathbb{D}\setminus\mathcal{D}^j}$ reflects that those variables unobserved in dataset j have been integrated out.[6] The measurement error then captures sampling uncertainty, differences in sampling procedures, and differences in operationalization of economic concepts.

3.2 Implementing the Estimation

Estimating (1) directly is not feasible because it is an infinite-dimensional nonlinear functional difference equation and, of course, we only observe samples of the distribution functions, not the functional data itself. Our innovation is to overcome this challenge by making it possible to estimate (1) using traditional Bayesian techniques and a Kalman-filter (Section 3.2.4). This requires transforming (1) into a linearized (infinite-dimensional) state-space model (Section 3.2.3), which is estimable once we render it finite and of reduced dimensionality under an appropriate factor representation (Section 3.2.2). However, first, we need to operationalize the measurement of the distribution functions as they appear in (2) by transforming the microdata samples into estimates of the distribution functions themselves (Section 3.2.1). In doing so, we have to account for changes in the effective support of the distributions and deal with the unobservability of some of the micro-variables in some datasets. In the following, we cover these points, which ultimately define our estimation procedure.

5. To ensure Ξ_{t+1} is a valid distribution: $\int G(d\Xi, \cdot)(w), dw = 1$, $\int \epsilon_t(w), dw = 0$, and $G(d\Xi, Y_{t-\ell:t+1})(w) \geq -\epsilon_t(w)$ (non-negativity). Non-negativity applies to $d\tilde{\Xi}$ and ν in (2).

6. For clarification, the tilde ($d\tilde{\Xi}$) is to denote that this is an estimate from the data vs. an estimate generated from the model, which will be denoted by a hat ($d\hat{\Xi}$).

3.2.1 Transforming the microdata

Handling changes in scale One challenge in working with distributional data is that the magnitude of the variables of interest in w, and thus the support of Ξ, changes over time. We deal with this in two ways. First, to deal with level changes, we rescale the vector w observed for individual i in the micro-dataset j at time t, w_{it}^j, by its dataset- and time-specific mean $\bar{w}_t^j$.[7] Second, to deal with changes in the width of the support, we decompose the distributional data into its marginals and a copula. Copulas, by definition, have a constant support (hypercubes of $[0, 1]$). Representing the marginals by their quantile functions (i.e., the inverse of the marginal cumulative distribution function) again achieves constant support by construction. This quantile-and-copula representation contains the same information as Ξ, but makes the support of all functions time-invariant.

This relies on Sklar's Theorem (Sklar 1973), which states that any multivariate cumulative distribution function $\Xi_t(w)$ can be written in terms of its marginal distributions $\Xi_{mt}(w)$, along the dimension $m \in \mathbb{D}$, and a copula, $C_t : [0, 1]^d \longrightarrow [0, 1]$ with uniform marginals. The copula captures the dependence structure between the random variables in w and is invariant to monotone transformations in w. Intuitively, the copula isolates the dependence structure (e.g., whether high-income households also tend to be high-wealth) from the shape of the individual distributions (e.g., how skewed income is). For our application, the copula is

$$\begin{aligned} C_t(u_1, \ldots, u_d) &= P_t(U_1 \leq u_1, \ldots, U_m \leq u_m, \ldots U_d \leq u_d) \\ &= P_t\left(w_1 \leq \Xi_{1t}^{-1}(u_1), \ldots, w_m \leq \Xi_{mt}^{-1}(u_m), \ldots, w_d \leq \Xi_{dt}^{-1}(u_d)\right) \qquad \forall t \in \mathbb{T} \end{aligned} \tag{3}$$

where $U_m \sim U[0, 1]$ for $m \in \mathbb{D}$ are the uniform marginals generated by taking the probability integral transform of each component m s.t. $U_m = \Xi_m(w_m) \sim U[0, 1]$. The second line highlights the quantile functions or the inverse transform of the univariate CDFs, $\Xi_{mt}^{-1}(w_m)$, where:

$$\Xi_{mt}^{-1}(u_m) = \inf\{w_m \in \mathbb{R} : \Xi_m(w_m) \geq u_m\} \qquad \forall t \in \mathbb{T}, m \in \mathbb{D}. \tag{4}$$

Finally, for C_t to be a copula, the constraint must hold for all $m \in \{1 \ldots d\}$ that when integrating out all but one dimension (the marginal distribution) m, the

7. Estimating relative to dataset-specific time means also lets us flexibly align the synthetic distributions with per-capita or per-household aggregate targets: we simply rescale using the desired aggregate target rather than the dataset mean. A consensus business-cycle component can be obtained by rescaling with average fixed effects and omitting any trend.

copula is identical to the value of the marginal distribution:

$$C_t(1, \dots u_m, \dots 1) = u_m. \tag{5}$$

Dealing with the infinite-dimensionality of the distribution functions The second challenge in our endeavor is dealing with the infinite-dimensionality of the distribution functions, while maintaining a functional approach. To do so, we will rely on a series estimator for both the copula and quantile function. Series estimators project the functions of interest onto some infinite-dimensional space of polynomials, namely the space of squared-integrable functions $\mathcal{L}^2$. In the following, we will treat the copula densities dC_t and (transformed) quantile functions Ξ_{mt}^{-1} as elements of $\mathcal{L}^2$, and project them onto a space of shifted orthonormal Legendre polynomials $\{Q_o : o \in \mathbb{N}_0\}$ for some order o, satisfying

$$\int_{[0,1]} Q_o(x)\, Q_k(x)\; dx = \delta_{ok}, \tag{6}$$

with $\delta_{ok} = 1$ if $o = k$ (reflecting the normalization condition) and 0 otherwise by orthogonality.[8] Classical Legendre polynomials are defined on $[-1, 1]$, so the shift is to the same domain as the copula densities and quantile functions—in the support $[0, 1]$. The orthonormal shifted Legendre polynomials Q_o themselves are defined on $[0, 1]$ by

$$Q_o(x) = \sqrt{2o+1}\, L_o(2x - 1), \tag{7}$$

where L_o are the classical Legendre polynomials defined by

$$L_0 = 1, \qquad L_1(x) = x, \qquad (o+1)L_{o+1}(x) = (2o+1)\, x\, L_o(x) - oL_{o-1}(x). \tag{8}$$

One concern is that the approximation method is poorly chosen, as the underlying population functions are themselves not completely square-integrable. For example, empirical evidence for distributions of consumption, income, and wealth points to tail behaviour that violates the $\mathcal{L}^2$ condition (see e.g., Vermeulen

8. One could alternatively project log-densities, as in Chang, Chen, and Schorfheide (2024), which avoids imposing non-negativity and monotonicity constraints in a linear model, but would forfeit the simple series estimators we use. Instead, we verify these constraints ex post. For quantiles, we check monotonicity by evaluating the Legendre approximation on a fine grid (10,000 points) and find no violations in our application. For the copula, the estimator satisfies uniform margins by construction (Sections 2–3 Bakam and Pommeret 2023); we additionally check positivity ex post. Any negative values are negligible (never below -10^{-6}), so the estimates satisfy copula-density constraints for practical purposes.

2018).[9] For the copula density, the same issue arises—density at the tails may fail to lie in the fulldomain space $\mathcal{L}^2([0,1]^d)$ for copulas exhibiting non-zero tail dependence e.g., the Gumbel copula.

To ensure that these underlying functions lie in $\mathcal{L}^2$, rendering the projection apropos, we first treat the data and apply the inverse-hyperbolic-sine transformation on them. This regularizes the upper tail of the distribution. This treatment is similar in nature to working with log densities, as in Chang, Chen, and Schorfheide (2024).[10] We then estimate our quantile functions on these transformed values. Appendix A provides evidence that by using the transformed percentile functions, the projection performs well at capturing variation in the tails. For the copula, we impose a *local* $\mathcal{L}^2$ condition on a trimmed domain $[\varepsilon, 1-\varepsilon]^d$, for some very small $\varepsilon > 0$. Inside that interior cube, $dC_t(\mathbf{u})$ is bounded and hence square-integrable. Appendix A also addresses the treatment of measures in the presence of non-negligible zero-mass.

Altogether, this means the following representation:

$$\Xi_{mt}^{-1}(u_m) = \sum_{o \in \mathbb{N}_0} \xi_{mot} Q_o(u) \tag{10}$$

$$dC_t(u_1, u_2, \ldots u_d) = \sum_{(o_1,\ldots,o_d)\in\mathbb{N}_0^d} \kappa_{(o_1,\ldots,o_d),t} \prod_{m=1}^{d} Q_{o_m}(u_m), \quad u_m \in [\varepsilon, 1-\varepsilon] \tag{11}$$

where both functions are represented as infinite sums of polynomials over the order of the polynomials $o \in \mathbb{N}_0$. Because of orthonormality (6) and using (10), for some fixed o, we obtain:

$$\int_0^1 \Xi_{mt}^{-1}\, Q_o(u) du = \int_0^1 \sum_{k\in\mathbb{N}_0} \xi_{mkt} Q_k(u)\, Q_o(u) du = \sum_{k\in\mathbb{N}_0} \delta_{ko} \xi_{mkt} = \xi_{mot}. \tag{12}$$

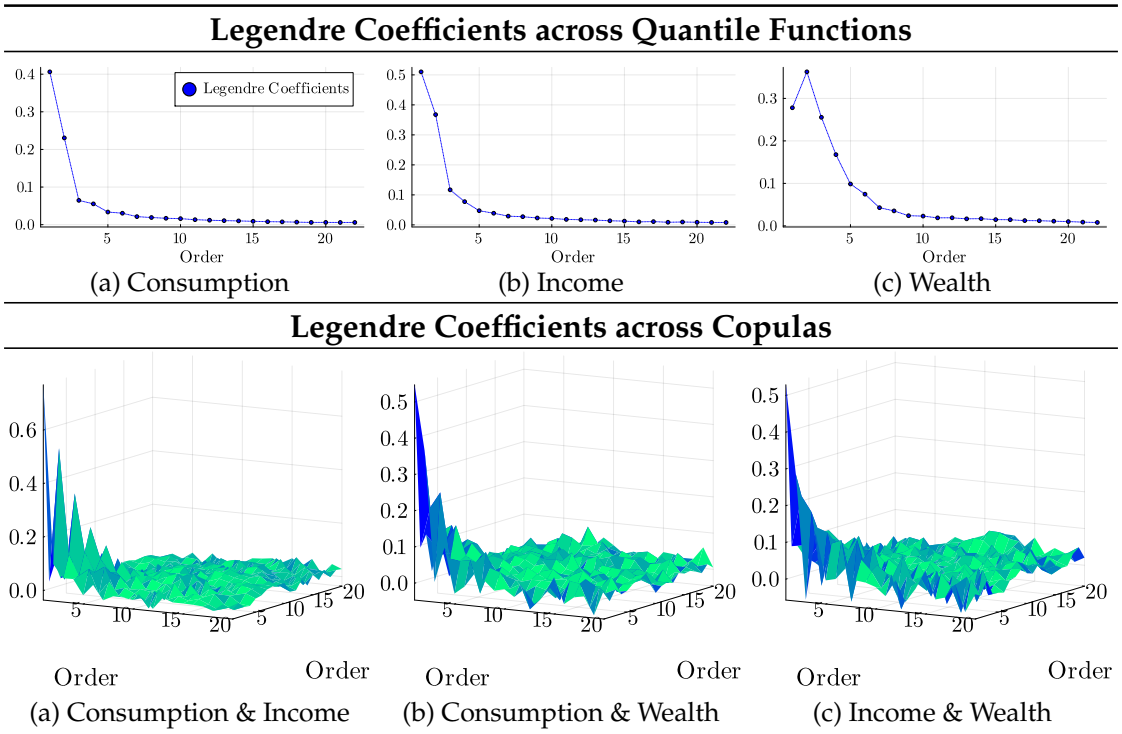
This means that the coefficients can be obtained as the inner products of the functions and the Legendre polynomials of some order o. The same holds true for the copula density:

9.
$$\int_a^b \left[Q(p)\right]^2 dp = \int_a^b \Big[\underbrace{x_m\,(1-p)^{-1/\alpha}}_{\substack{\text{Pareto}\\ \text{quantile}}}\Big]^2 dp = x_m^2 \int_a^b (1-p)^{-2/\alpha}\, dp. \tag{9}$$

In particular, if the uppertail of the income distribution on $p \in [a,b]$ follows a Pareto law with shape α, then its quantile function is $Q(p) = x_m(1-p)^{-1/\alpha}$. This lies in $\mathcal{L}^2[a,b]$ (i.e. $\int_a^b Q(p)^2\, dp < \infty$) if and only if $\alpha > 2$. When $\alpha \le 2$, the integral diverges and the second moment does not exist.

10. For notional convenience, we use $\Xi_{mt}^{-1}(u_m)$ as the treated quantile function.

Figure 2: Legendre Coefficients Across the Distributional Data



Notes: Figure presents two rows of Panels. The first row presents the coefficients (in dots) on the Legendre polynomials from estimating the quantile function, in increasing order. The second row presents the coefficients (as a surface) on the Legendre polynomials from estimating the copula density in lexicographic order. Data are from the 2019 PSID.

$$\int_{[0,1]^d} \prod_{m=1}^{d} Q_{o_m}(u_m)\, dC_t\ du_1, \ldots, du_d = \kappa_{(o_1,\ldots,o_d),t}. \tag{13}$$

For the estimation of these coefficients in practice, we rely on the uniformity of ranks and that ranks are within $[0, 1]$ and replace the inner product by sample averages:

$$\hat{\xi}_{mot} := N_j^{-1} \sum_i w_{mit} Q_o(u_{mit}) \tag{14}$$

$$\hat{\kappa}_{(o_1,\ldots,o_d),t} := N_j^{-1} \sum_i \left(\prod_{m=1}^{d} Q_{o_m}(u_{mit}) \right) \tag{15}$$

where u_{mit} is the data rank of w_{mit}, the sample analogue of Ξ_{mt}^{-1} for observation i.

In the estimation, we truncate the sums in (10) and (11) at a given maximal order O and by orthonormality of the polynomials, the kept coefficients are not affected by the truncation. The coefficients in our case indeed decrease rapidly with each polynomial as we show in Figure 2. The figure plots the absolute coefficient value as a function of the order of the polynomial term's order. At around

order 10, the coefficients all become very small.[11] Evidence from Appendix A supports this, showing a truncation to $O = 11$ captures the variation observed in the data quite well.

Dealing with partial unobservability of the microdata Another difficulty is that the microdata may not always contain the entire vector w, but only a subset. When w is incompletely observed, we cannot estimate the full d-dimensional copula directly. However, we can still estimate copulas with the unobserved dimensions integrated out—that is, lower-dimensional copulas. We show below that these lower-dimensional copulas are *slices* of the higher-dimensional copula of interest, and thus remain informative about the full d-dimensional object.

The representation in the form of (Legendre) polynomials is very useful in this respect. First, we need to show that the density of the higher-dimensional copula must be equal to the lower-dimensional one when we integrate out the "missing" dimension d:

$$\int_0^1 dC(u_1, \ldots, u_d)\, du_d \overset{!}{=} dC(u_1, \ldots, u_{d-1}) \tag{16}$$

In this effort, we write out the integrand using (11) and make use that the first (shifted) Legendre polynomial integrates to one and all others to zero to obtain:

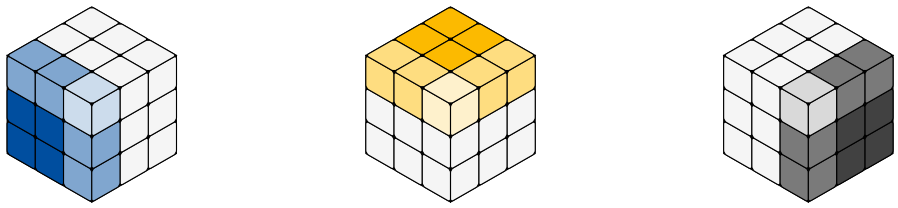
$$\begin{aligned}
\int_0^1 dC(u_1, \ldots, u_d)\, du_d &= \int_0^1 \sum_{o_1} \cdots \sum_{o_d} \kappa_{(o_1,\ldots,o_d)} \prod_{m=1}^{d} Q_{o_m}(u_m)\, du_d \\
&= \sum_{o_1} \cdots \sum_{o_{d-1}} \sum_{\mathbf{o_d}} \kappa_{(o_1,\ldots,o_{d-1},\mathbf{o_d})} \left(\prod_{m=1}^{d-1} Q_{o_m}(u_m) \right) \int_0^1 Q_{\mathbf{o_d}}(u_d)\; du_d \\
&= \sum_{o_1} \cdots \sum_{o_{d-1}} \kappa_{(o_1,\ldots,o_{d-1},1)} \left(\prod_{m=1}^{d-1} Q_{o_m}(u_m) \right).
\end{aligned} \tag{17}$$

In words, the polynomial coefficients of the lower-dimensional copula density are identical to the leading "slice" of the higher-dimensional copula. Therefore, a dataset that contains only two out of the three variables of interest still provides a measurement of a subset of the coefficients; see Figure 3.[12]

11. A small coefficient in absolute value does imply the contribution of the corresponding polynomial to be small in an R^2 sense.

12. By the same line of argument, a copula requires $\kappa_{(1,\ldots,1)} = 1$ and $\kappa_{(1,\ldots,j,\ldots,1)} = 0$.

Figure 3: Geometric Representation of Partially Observed Copula Coefficients



(a) Only Variables 1& 3 (b) Only Variables 2& 3 (c) Only Variables 1& 2

Notes: Figure shows three cubes. A cube can be interpreted as an array of copula coefficients $\kappa^j_{(o_1,\ldots,o_d),t}$ for some dataset j at time t, for $d = 3$. Each cube corresponds to a scenario where one variable is missing in the estimation of the copula density. The light edge denotes the (1,1,1) coordinate. For each scenario, the white boxes are coefficients we cannot estimate. The slightly colored boxes correspond to the immutable coefficients, which have fixed values independent of data. The darker colored boxes are scenario specific and correspond to (time-varying) coefficients that need to be estimated.

3.2.2 Dealing with the Curse of Dimensionality

Vectorizing the coefficients for each time period, t, leaves us with sequences of coefficients, $\theta^j_t = (\xi^j_{mot}, \ldots, \kappa^j_{(o_1,\ldots,o_d),t})$, for each cross-sectional dataset, j. For example, with $d = 3$ dimensions in our application—consumption, income, and wealth and using polynomials up to order eleven ($O = 11$) to organize the data—the copula density would be represented by a vector with $(O+1)^d - (d \times O + 1) = 1694$ variable entries for each time point and $d \times O + 1 = 34$ invariable. These invariable entries are not included in the estimation. In addition, there would be $d \times (O+1) = 36$ coefficients of the polynomials (including a constant) representing the quantile functions, which we collect in θ^j_t as well.

This example makes apparent that even for a modest polynomial order, the dimensionality of $\theta^j_t \in \mathbb{R}^N, N = (O+1)^d + (d-1)$ for a dataset with d variables is large. Equally important is the frequency of missing observations across the N entries; in our application, the data structure is irregular and includes many gaps. Taken together, this makes it impractical to specify and estimate a time-series model directly on the raw θ^j_t coefficients. For this purpose, we postulate (and then estimate) a dynamic factor model for θ. This is where another advantage of the *orthonormal* polynomial representation is advantageous: The variance (over time) of a coefficient is proportional to its contribution to fluctuations of the function (in the $\mathcal{L}^2$ sense). Put simply: The polynomial coefficient provides a useful form of standardization that has a natural metric and allows us to uncover the factor structure behind the time-series changes in the distributions. This factor structure finally allows us to overcome the curse of dimensionality in the distributional data.

For this purpose, all (variable) coefficients of the polynomial representation of dC_t^j (and separately of the quantiles) are horizontally concatenated:

$$\boldsymbol{\theta}^j = \begin{bmatrix} \theta^j_{1,1} & & \theta^j_{1,T} \\ & \ddots & \\ \theta^j_{N,1} & & \theta^j_{N,T} \end{bmatrix}.$$

Define $\boldsymbol{\theta}$ as the matrix obtained by horizontally concatenating the individual blocks $\boldsymbol{\theta}^j$, retaining only those time periods for which every coefficient is available.[13] A PCA (a singular value decomposition) is then performed, which nonparametrically reduces the dimensionality of the data (Breitung and Eickmeier 2006; Chen, Er, and Wu 2005). For the PCA, we define $\tilde{\boldsymbol{\theta}}$, which is then standardized representation of $\boldsymbol{\theta}$. The standardization is done by distribution object $\zeta \in \{1, \ldots, d+1\}$ (d quantile functions and a copula).[14] We standardize by object (and not by coefficient) because the polynomial coefficients represent standardized contributions to the object's structure (e.g., the quantile function or copula). As a result, additional standardization within the same object is unnecessary, since their relative magnitudes are naturally balanced by the properties of the polynomial basis. This leaves us with the standardized observation $\tilde{\theta}^j_{nt}$ of coefficient n in data source j at time t.

The PCA of $\tilde{\boldsymbol{\theta}}$ provides us with a preliminary projection matrix $\tilde{\Gamma} \in \mathbb{R}^{N \times r}$, with full column rank r, that projects $r << N$ factors into the N dimensional distributional data. We then normalize $\tilde{\Gamma}$ by setting $\hat{\Gamma} := \tilde{\Gamma} D^{-1/2}$, where D is the diagonal matrix of eigenvalues of $\tilde{\Gamma}'\tilde{\Gamma}$ in decreasing order. Throughout we work with the normalized matrix $\hat{\Gamma}$.

More specifically, we decompose $\tilde{\boldsymbol{\theta}}$ into latent orthogonal factors $\begin{bmatrix} \hat{f}'_\Gamma & \hat{f}'_\gamma \end{bmatrix}'$ divided into "important" and "unimportant" factors according to their contribution to the total variance (measured by their singular value). Their respective time-invariant loadings are $\begin{bmatrix} \hat{\Gamma} & \hat{\gamma} \end{bmatrix}$, for $\hat{f}_\Gamma \in \mathbb{R}^{r \times T}$ and $\hat{f}_\gamma \in \mathbb{R}^{(N-r) \times T}$. This decomposition is unique up to the scale of each factor, which allows us to normalize

13. Before the concatenation, each coefficient is detrended using an HP-filter. This takes care of dataset-specific effects. We store the information needed to transform $\tilde{\boldsymbol{\theta}}^j$ back to the originally observed objects to obtain source-specific predictions. For example, the income quantiles (that would be one object ζ) in the SCF and the PSID (two sources j) may be permanently different due to differences in sample design and operationalization.

14. The normalization of coefficient n in object ζ in dataset j is given by $\tilde{\theta}^j_{nt} = \left(\frac{\theta^j_{nt} - \mu^j_n}{\sigma^j_{\zeta(n)}} \right)$ where μ^j_n is the coefficient-specific mean and $\sigma^j_{\zeta(n)}$ is the standard deviation of *all* coefficients (rows) pertaining to object ζ. This removes dataset- and coefficient-specific fixed effects, μ^j_n. This can capture, e.g., permanent differences in sampling procedure and operationalization.

the loadings so that all factors have unit variance. Altogether, we have:

$$\tilde{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\Gamma} & \hat{\gamma} \end{bmatrix} \begin{bmatrix} \hat{f}_\Gamma \\ \hat{f}_\gamma \end{bmatrix} \tag{18}$$

where $\hat{f}_\Gamma$ represents the r important factors, which capture almost all of the variation in the data, and $\hat{f}_\gamma$ the $N-r$ less important factors, which can be interpreted as some measurement noise. From this step, we identify an ideal functional basis f_Γ (see Kneip and Utikal 2001; Tsay 2016), which determines the *size* of the model. We then use $\hat{\Gamma}$ in the model as the linear mapping between the high-dimensional $\tilde{\theta}$ and basis f_Γ. The use of $\hat{\Gamma}$ will be a necessary condition for the proposed procedure of estimating our distributional factors. This will be elaborated on in Section 3.2.3.

The estimation of $\hat{\Gamma}$ requires time observations where every coefficient is observed, yet some data sources will not satisfy this criteria *at all*—for example, the SCF does not measure consumption ever. Observations from such incomplete sources are therefore excluded from the PCA that yields $\hat{\Gamma}$. In Appendix B, we consider alternative estimators that accommodate partial unobservability in $\tilde{\boldsymbol{\theta}}^j$, which we detail in Appendix B. Following standard practice, we select among these alternatives the estimator with the highest marginal data density.

3.2.3 Factor State Space Model and Measurement

With this preprocessing of the data, we can turn to specifying the state-space model that captures the evolution of the r latent distributional factors. Due to the possible mixed-frequency nature of the data, we postulate a mixed-frequency state space model as in Schorfheide and Song (2015), with the following augmented state vector

$$F_t \;=\; \begin{bmatrix} f_t' & f_{t-1}' & \cdots & f_{t-p+1}' \end{bmatrix}' \in \mathbb{R}^{rp \times 1}, \tag{19}$$

where F_t collects the r latent distributional factors f_t, in addition to its $p-1$ lags. For the application of our methodology in Section 4, we set $p = 4$, with the high-frequency data of interest being quarterly. To facilitate the reading across sections, we provide the example of this quarterly case herein, though the methodology easily generalizes to other frequencies.

State Equation Stacking the lags directly with the q aggregate factors Y_t[15] in the state gives the following representation

$$\begin{bmatrix} F_t \\ Y_t \end{bmatrix} = \Phi \begin{bmatrix} F_{t-1} \\ Y_{t-1} \end{bmatrix} + \epsilon_t = \begin{bmatrix} \Phi_{FF} & \Phi_{FY} \\ \Phi_{YF} & \Phi_{YY} \end{bmatrix} \begin{bmatrix} F_{t-1} \\ Y_{t-1} \end{bmatrix} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}\big(0,\, \Omega\big) \tag{20}$$

and for the case of quarterly data

$$F_{t-1} = \begin{bmatrix} f_{t-1} \\ f_{t-2} \\ f_{t-3} \\ f_{t-4} \end{bmatrix}, \Phi_{FF} = \begin{bmatrix} A & 0 & 0 & 0 \\ I_r & 0 & 0 & 0 \\ 0 & I_r & 0 & 0 \\ 0 & 0 & I_r & 0 \end{bmatrix}, \Phi_{FY} = \begin{bmatrix} B \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Phi_{YF} = \begin{bmatrix} C \\ 0 \\ 0 \\ 0 \end{bmatrix}', \Phi_{YY} = D,$$

where distributional factors f_t are assumed to follow an AR(1) process, summarized by the law of motion matrix $A \in \mathbb{R}^{r \times r}$. In addition, they are assumed to co-move with aggregate factors Y_t, whose relationship is summarized by loading matrix $B \in \mathbb{R}^{r \times q}$ and $C \in \mathbb{R}^{q \times r}$. To represent the correlation structure between the aggregate factors, we posit $D \in \mathbb{R}^{q \times q}$ and assume it is diagonal, i.e., no cross-correlation between aggregate factors. All other matrices A, B and C are left unrestricted. As in Schorfheide and Song (2015), the remaining rows of Φ are defined to deliver the identities $f_{t-l} = f_{t-l}$ for $l = 1, ..., p-1$—hence the I_r in the example of a quarterly estimation. While it is customary to restrict the innovations, ϵ_t, to be uncorrelated, both serially and between factors, so that Ω is a diagonal matrix, we do estimate both the variances and covariances of Ω. This allows for a shock structure where innovations to the distributional factors can be *contemporaneously* correlated with innovations to the aggregate factors.

Measurement Equation Since the factors are not directly observable, we complement the factor model with an observation equation for each dataset j:

$$\tilde{\theta}_t^j = H_t^j \left(\alpha^j \hat{\Gamma}^{MF} F_t + \nu_{F,t}^j \right), \quad \nu_{F,t}^j \sim \mathcal{N}\left(\mathbf{0}, \hat{\Sigma}_j^{1/2} S^j \hat{\Sigma}_j^{1/2\prime} \right) \tag{21}$$

This observation equation maps factors F_t, through the loading matrix $\alpha^j \hat{\Gamma}^{MF}$, to the standardized coefficients $\tilde{\theta}_t^j$, allowing for a Gaussian measurement error $\nu_{F,t}^j$. The selector matrix $H_t^j \in \{0,1\}^{N \times N}$ indicates whether (parts of) the distribution are observed in data source j at time t, following Durbin and Koopman (2012). The Gaussian error term $\nu_{F,t}^j$ has a variance-covariance matrix decomposed into

15. Recall that the set of aggregate factors at time t depends on an information set that captures both contemporaneous and past aggregate variation.

S^j, a diagonal scaling matrix with diagonal entries s^j_ζ varying by object ζ and $\hat{\Sigma}_j$, a positive semi-definite (covariance) matrix.

The diagonal matrix $\alpha^j \in \mathbb{R}^{N\times N}$ and the matrix $\hat{\Gamma}^{MF}$ are constructed to take time aggregation into account. Specifically, these matrices map the high-frequency factors to the lower-frequency observations. For example, when the frequency of interest is quarterly, as it is in the application, $\alpha^j_{\zeta(n),ii}$ takes one of two values:

$$\alpha^j_{\zeta(n),ii} = \begin{cases} \frac{1}{4} \\ 1 \end{cases} \quad \text{and} \quad \hat{\Gamma}^{MF}_{\zeta(n)} = \begin{cases} \left[\hat{\Gamma}_{\zeta(n)}, \hat{\Gamma}_{\zeta(n)}, \hat{\Gamma}_{\zeta(n)}, \hat{\Gamma}_{\zeta(n)}\right] & \text{annual flow,} \\ \left[\hat{\Gamma}_{\zeta(n)}, \mathbf{0}, \mathbf{0}, \mathbf{0}\right] & \text{quarterly flow or stocks.} \end{cases}$$

First, it can take a value of $\frac{1}{4}$, which means the distributional object corresponds to an *annual flow* (e.g. the PSID income / consumption quantile function). The $\frac{1}{4}$ can be interpreted as averaging over the four quarterly factors $f_t, \ldots, f_{t-3}$ and thus loads on all four quarters i.e., $\hat{\Gamma}^{MF}_{\zeta(n)} = \left[\hat{\Gamma}_{\zeta(n)}, \hat{\Gamma}_{\zeta(n)}, \hat{\Gamma}_{\zeta(n)}, \hat{\Gamma}_{\zeta(n)}\right]$. For *stock variables* (e.g. any wealth object for which the date is known), $\alpha^j_{\zeta(n),ii} = 1$ and loads exclusively on the contemporaneous factor block and the lagged blocks are set to zero i.e., $\hat{\Gamma}^{MF}_{\zeta(n)} = \left[\hat{\Gamma}_{\zeta(n)}, \mathbf{0}, \mathbf{0}, \mathbf{0}\right]$, for $\mathbf{0}$ the same size as $\hat{\Gamma}_{\zeta(n)}$. The same holds for the third case, *quarterly flows* (e.g., SIPP income, CEX), which again loads on the contemporaneous block only. In terms of how α^j varies by object, it is straightforward for the quantile functions, since each quantile function corresponds to one variable at some point in time t; however, for the copula, since multiple variables are necessary for its construction, the different measurements of the variables could be mixed *within* e.g., annual income flows and stock wealth. In these cases, for some point in time t, the lowest frequency is chosen among the set of measurements.

Measurement Error The measurement error for the distributional data, $\nu^j_{F,t}$, is composed of sampling uncertainty and other errors that reflect the fact that a given dataset has its specific operationalizations of the common economic variables being measured. Differences in operationalizations can not only shift the level of a particular measurement (which we capture through fixed effects, see Footnote 14), but can also become differentially important over time.[16] In principle, to fully capture these time-varying differences, we would need to esti-

16. For example, the PSID and SCF differ in the way they ask respondents about their business wealth (Pfeffer, Schoeni, Kennickell, and Andreski 2016). PSID and CPS differ in the sampling unit, making household/family income sensitive to labor supply patterns (Gouskova, Andreski, and Schoeni 2010). Similarly, differences in the propensity to sample business owners between the different datasets make income sensitive to relative changes in business and labor income (Kim and Stafford 2000). Finally, the CEX and the PSID differ in the consumption categories covered in the survey, with the PSID being much coarser (Insolera, Simmert, and Johnson 2021).

mate a measurement error variance for each coefficient and dataset. However, this would be too large a number of parameters to estimate within the time-series model; a common problem when estimating approximate dynamic factor models with maximum likelihood (see e.g., Bai and Li 2016). The standard approach in the literature uses quasi-maximum-likelihood with EM algorithms (Doz, Giannone, and Reichlin 2012; Babura and Modugno 2014; Barigozzi and Luciani 2019), but convergence to the global optimum is not guaranteed in high-dimensional settings (Wu 1983; Balakrishnan, Wainwright, and Yu 2017).

For this reason, we do two things, the second of which will be discussed in Section 3.2.4, when we cover our sampling procedure. First, we assume that the correlation structure of all measurement errors is proportional to the correlation structure for sampling uncertainty. Under this assumption, the matrix Σ_j can be estimated outside the time-series model using bootstraps or the supplied replication weights to estimate the covariance from sampling uncertainty by data source j.[17] This achieves differently the diagonality assumed in the aforementioned literature. This leaves N diagonal elements of S^j to be estimated within the time-series model. To reduce the parameter count further, we constrain these elements to vary only by dataset j and variable ζ, yielding just $d+1=4$ parameters per dataset. Consequently, each diagonal element s^j_ζ indicates by what factor the sampling-based standard error for ζ must be inflated (or deflated) in dataset j. Therefore, the overall measurement error variance in (21) is $\hat{\Sigma}_j^{1/2} S^j (\hat{\Sigma}_j^{1/2})'$.

Thus, similarly to feasible generalized least squares, we transform observations $\tilde{\theta}^j_t$ by pre-multiplying the coefficients with the external estimate $\hat{\Sigma}_j^{-1/2}$

$$\begin{aligned} \hat{\Sigma}_j^{-1/2}\tilde{\theta}^j_t &= H^j_t(\alpha^j \hat{\Sigma}_j^{-1/2}\hat{\Gamma}^{MF} F_t + \tilde{\nu}^j_{F,t}) \qquad \tilde{\nu}^j_{F,t} \sim \mathcal{N}(0, S^j) \\ \tilde{\tilde{\theta}}^j_t &= H^j_t(\alpha^j \tilde{\hat{\Gamma}}^{MF}_j F_t + \tilde{\nu}^j_{F,t}), \qquad \tilde{\hat{\Gamma}}^{MF}_j := \hat{\Sigma}_j^{-1/2}\hat{\Gamma}^{MF}. \end{aligned} \tag{22}$$

With the measurement error treatment in place, denoted by the additional tilde, we can stack the J measurement equations, and with the aggregates, arrive at the final measurement equation

$$\begin{bmatrix} \tilde{\tilde{\theta}}_t \\ \hat{Y}_t \end{bmatrix} = H_t \left(\begin{bmatrix} \alpha \tilde{\hat{\Gamma}}^{MF} & \mathbf{0} \\ \mathbf{0} & I_Y \end{bmatrix} \begin{bmatrix} F_t \\ Y_t \end{bmatrix} + \begin{bmatrix} \tilde{\nu}_{F,t} \\ \nu_{Y,t} \end{bmatrix} \right) \qquad \nu_{Y,t} \sim \mathcal{N}(\mathbf{0}, \Sigma_Y). \tag{23}$$

17. We draw bootstrap samples / use the supplied replication weights for each dataset j, $\{\tilde{\boldsymbol{\theta}}^j_{t,b}\}_{b=1}^B$, for each period t. Then, we demean the bootstrap samples b for each j and t and compute the average within-time variance-covariance matrix $\hat{\Sigma}_j$ pooling the demeaned bootstrap samples of the dataset j. If an object ζ is unobserved in dataset j, we set the covariance terms to zero and the diagonal elements to one to still be able to compute $\Sigma_j^{-1/2}$.

Note that the terms are free of the j index to indicate they have been vertically stacked and note the additional appendage for the aggregates Y_t. We treat the aggregate factors to be observed with some noise Σ_Y, but still precise, since factors are constructed from $N \approx 2000$ series (details in Section 3.2.4).

In sum, the state-space model (20) and (23) forms a linear system of equations that can be estimated using standard Bayesian techniques and a Kalman-filter.

3.2.4 Bayesian Estimation

We need to estimate seven objects. From the state equation, we need to estimate the factor-autoregression matrix A, the loading matrix B from aggregate controls to distributional factors, the loading matrix C from distributional factors to aggregate factors, the vector autocorrelation of the aggregate factors themselves, the diagonal of D, the variance-covariance matrix of the shocks to the factors Ω, the variance-covariance matrices $\Sigma_j^{1/2} S^j \Sigma_j^{1/2'}$ and Σ_Y of the measurement errors. In a first step, the covariance structure $\Sigma^{1/2}$ is estimated outside the time-series model, as noted above, leaving only the scaling matrices S^j to be estimated within the model. Given the size of the A, B, C, D, Ω, S^j, and Σ_Y matrices, we use a Bayesian approach to estimate the system. This imposes regularization via priors on the parameters, as well as hyperpriors on the hyperparameters, to avoid overfitting the data.

The estimation is then one of a fully hierarchical Bayesian state-space model with mixed-frequency data. We collect all parameters and hyperparameters into a single parameter vector $\boldsymbol{\psi} = (\boldsymbol{\psi}_{\text{par}}, \boldsymbol{\psi}_{\text{hyper}})$, estimated jointly. The remainder of this section details the specific prior distributions on $\boldsymbol{\psi}$, first for the state equation parameters and subsequently for those of the measurement equation.

State Equation For the state equation, we define the following prior:

$$\begin{pmatrix} vec(A) \\ vec(B) \\ vec(C) \\ diag(D) \end{pmatrix} \sim \mathcal{MN}(\mu_{\text{Minn}}, V_{\text{Minn}}) \qquad \begin{aligned} L_{chol} &\sim LKJChol(r+q, \eta) \\ T^{-1}(\Omega_{F,ii}) &\sim \mathcal{N}(\mu_{\Omega_F}, \sigma_F^2) \\ T^{-1}(\Omega_{Y,ii}) &\sim \mathcal{N}(\mu_{\Omega_Y}, \sigma_Y^2) \end{aligned} \tag{24}$$

First, we impose a Minnesota prior on A, B, C, and $diag(D)$. This means we set the vector of expected values μ_{Minn} so that all *but* the autocorrelation terms in A and D have an expected value of zero. This implies that B and C are shrunk to zero and so are the off-diagonals of A if not needed to describe the data. The remaining values, the expected values for the autocorrelations (main diagonal of

A and D), are governed by hyperparameters described later in the section. For the prior on Ω, we impose a *separation strategy* (see e.g., Barnard, McCulloch, and Meng 2000) and decompose Ω into a diagonal of standard deviations Ω_σ and a correlation matrix Ω_{corr} i.e., $\Omega = \Omega_\sigma \times \Omega_{corr} \times \Omega_\sigma{}'$. This avoids any apriori correlation between variances and covariances, normally imposed in Inverse-Wishart setups (Barnard, McCulloch, and Meng 2000). For the shocks, Ω_σ is further decomposed into Ω_F and Ω_Y, imposing different priors on the standard deviations of aggregate shocks and those of distributional shocks. It will be on the diagonal elements $\Omega_{\sigma,ii}$ for which we specify the priors.

For the parameters $\Omega_{\sigma,ii}$, we first draw from a set of Normal distributions and map these parameters to $\mathbb{R}^+$ using the soft-plus transformation, denoted as $T(\cdot)$. This provides flexible exploration for the optimizer by avoiding exponential jumps from small perturbations in these parameters and avoids acceptance rates being driven by out of bound draws (Chapter 55, Stan User Guide). For $\Omega_{F,ii}$, the mean μ_{Ω_F} is defined such that the apriori long-run variance of each distributional factor is $1.0 = \frac{\mu_{\Omega_F}}{1-\kappa_3^2}$, consistent with our prior for autocorrelation (in the matrix A) and factor normalization to unit variance. The same is done for the standard deviations of the shocks to the aggregate factors, $\Omega_{Y,ii}$, only using $1.0 = \frac{\mu_{\Omega_Y}}{1-\kappa_4^2}$. Hyperparameters κ_3 and κ_4 are discussed in more detail later.

For the estimation of the correlation matrix, Ω_{corr}, the matrix must satisfy several constraints simultaneously which may be difficult to enforce in any optimization: it must be symmetric, have a unit diagonal, and be positive semidefinite. To circumvent this, we employ a reparameterization strategy. The key insight is to work with the Cholesky factor, L_{chol} where $\Omega_{corr} = L_{chol}\, L'_{chol}$. Following the methodology of Lewandowski, Kurowicka, and Joe (2009, hereafter: LKJ), we construct the Cholesky factor, L_{chol}, from an unconstrained vector of real numbers, $v \in \mathbb{R}^{(r+q)(r+q-1)/2}$ and it is precisely on this Cholesky factor that we evaluate our prior. The LKJ-Choleksy prior will be governed by a single shape parameter, $\eta \geq 1$, which will control the expected correlation strength between aggregate and distributional shocks.

For the state equation, there are $|\boldsymbol{\psi}_{hyper}| = 6$ hyperparameters that are jointly estimated with $\boldsymbol{\psi}_{par}$. One hyperparameter shapes V_{Minn}; two hyperparameters govern the autocorrelation in A and D; and the remaining three hyperparameters govern Ω. See Appendix C for further information on these hyperparameters.

Measurement Equation First, for the distributional measurements, we specify independent priors for the measurement error variances for each dataset and

object, s^j_ζ. Recall that estimating the diagonal of S^j would imply one variance parameter per coefficient n. As such, we impose the restrictions explained in the previous section, assigning one variance parameter s^j_ζ per object ζ and dataset j. This means a maximum (since some measures are unobserved) of $J \times (d+1)$ variance parameters. To ensure strict positivity of s^j_ζ while maintaining efficient sampling, we again parameterize each variance using a softplus transformation and place Normal priors on the unconstrained parameters $T^{-1}(s_\zeta)$. For the distributional data, we assign a weakly informative prior centered such that the expected variance is 1.0, with a standard deviation of 2.0. This says, first, on average, there is only sampling uncertainty and no additional measurement error reflecting conceptual differences, but that, second, $\hat{\Sigma}_j$ is still itself an estimate, additionally perturbing each s^j_ζ away from 1.[18]

$$T^{-1}(s^j_\zeta) \sim \mathcal{N}(0.54, 2), \qquad T^{-1}(\Sigma_{Y,ii}) \sim \mathcal{N}(-7.6, 0.02). \tag{25}$$

For the aggregate factors, we impose highly informative priors on variances $\Sigma_{Y,ii}$, reflecting the fact that these factors have been estimated with high precision. These priors are centered at a near-zero variance with a very narrow standard deviation of 0.02, effectively constraining the model to treat these aggregates as nearly perfectly measured.

Likelihood and sampling With this prior on $\boldsymbol{\psi}$, we obtain the model-likelihood $p(\tilde{\tilde{\boldsymbol{\theta}}}|\boldsymbol{\psi})$ using a Kalman-filter. The posterior log-likelihood is then calculated as the sum of the hyperprior log-probability, the prior log-probability and the data log-likelihood. To sample from the potentially complex, multi-modal, high-dimensional posterior distribution, we employ the Differential-Independence Mixture Ensemble (DIME) sampler from Boehl (2024). Details and convergence results are in Appendix D.

3.3 Reconstructing and Using the Distributional Data

After finding the posterior mode of our model parameters, we apply the Kalman smoother to recover the full path of latent factors $\{\hat{F}_t\}_{t=1}^T$. Given these smoothed factors, we then reconstruct the sequence of joint distributions, $\hat{\Xi}^j_t$, from the implied standardized series of polynomial coefficients. Appendix E outlines the procedure.

18. Identical priors across datasets mean that we do not a priori prioritize a conceptual measure for object ζ in one dataset over another. Setting different priors is generally possible if a particular measurement concept should be prioritized on theoretical grounds.

This estimation approach allows practitioners to use the above data in three principal ways: First, in its factor representation, $\hat{F}_t$, second, in its functional representation, $\hat{\theta}_t^j$, or, third, in its observational representation, e.g., quantile functions and copula—$\left(\{\hat{\Xi}_{jmt}^{-1}\}_{m=1}^d,\ d\hat{C}_t^j\right)$.

For example, if one is interested in replicating the exercise of Chang and Schorfheide (2024), one can use the second representation and use the coefficients related to the marginal distributions of interest in their *distribution*-augmented VAR. If one is only interested in augmenting a VAR with a low-dimensional summary containing virtually all variation in the joint distribution of consumption, income, and wealth—in the style of these factor-augmented VARs—one can use the first representation.

Lastly, one can use the coefficients to generate some data of economic interest and work with this economically interpretable/meaningful data directly. One can generate the quantile functions and copula densities, but also other common statistics of interest for the marginal distributions, such as, gini coefficients of consumption, income, or wealth; levels, quantiles, shares, etc. and objects that arise from the correlational structure (the copula density) for example, correlation measures, conditional probabilities (e.g., population shares), kendall's tau, and tail-dependence.

For the application in Section 5, we adopt (3) and use the quantile functions and copula densities to generate a synthetic micro-dataset X_{it}, which will ultimately be a pseudo-panel. From the quantile functions, we obtain average realizations of all variables m for different (representative) households i, each defined by a range of ranks $u \in U_i^m$ (e.g., deciles). We do this by integrating the quantile functions over U_i^m, i.e., forming conditional expectations. By having also estimated the copula density, we can combine these averages for household i with the household's probability weight by integrating the copula densities over some hyper-rectangular region, defined by ranks $\prod_{m=1}^d U_i^m$. Below is precisely this combination, for the example of the distribution of consumption, income, and wealth, which together forms a single row of the pseudo-panel:

$$X_{it}^j = \begin{pmatrix} c_{it}^j \\ y_{it}^j \\ w_{it}^j \\ \omega_{it}^j \end{pmatrix} = \begin{pmatrix} \frac{1}{|U_i^c|}\int_{u\in U_i^c} \hat{\Xi}_{jct}^{-1}(u)\ du \\ \frac{1}{|U_i^y|}\int_{u\in U_i^y} \hat{\Xi}_{jyt}^{-1}(u)\ du \\ \frac{1}{|U_i^w|}\int_{u\in U_i^w} \hat{\Xi}_{jwt}^{-1}(u)\ du \\ \iiint_{(u^c,u^y,u^w)\in U_i^c\times U_i^y\times U_i^w} d\hat{C}_t^j(u^c,u^y,u^w) \end{pmatrix}. \tag{26}$$

In this example, we have a synthetic (representative) household i, where $(U_i^c \times$

$U_i^y \times U_i^w$) is the quantile combination that defines household i, e.g., the first decile in consumption c, the third decile in income y, and the seventh decile in wealth w. The mass, ω_i, of the households in that cell defines a weight for that synthetic household.[19] The $|U_i^m|$ refers to the Lebesgue measure over the U_i^m to generate the representative (average) household in this interval i.e., $|U_i^m| = b_i^m - a_i^m$ for $a,\ b$ end points of the closed interval U_i^m. This implementation implies that, without the need for sampling, we obtain an output that can be immediately interpreted as synthetic microdata. As illustrated above, X_{it}^j is dataset specific, but one can obtain a consensus estimate across datasets as a simple average of the $d\hat{C}_t^j$ and $\hat{\Xi}_{jmt}^{-1}$ over datasets j or, alternatively, as a weighted average by using the inverse measurement error standard deviations (by object and dataset) as weights.

4 Credibility Checks

Before turning to the application, we evaluate the reliability of our procedure step-by-step using actual (Section 4.1) and simulated data. This addresses two potential concerns: (i) Does the factor representation lose information? (Section 4.2) and (ii) Does the state-space model predict accurately when data are missing? (Section 4.3). We first show that the factor representation preserves the information needed for state-space estimation, validating the approach described in Sections 3.2.1 and 3.2.2. Without this reduction step, we would need to track and propagate a high-dimensional set of distributional coefficients directly in the state-space model (as in Chang, Chen, and Schorfheide 2024), which is computationally infeasible in our setting. The mapping from these noisy coefficients and latent distributional states then defines our measurement equation, whose priors and hyperparameter choices are validated in Appendix F (validating in part Section 3.2.3 and Section 3.2.4), showing that the estimated path of the functional data falls reasonably well into the bounds given by sampling uncertainty.

With the functional state-space model in place, we perform three further experiments to address the second concern. The first two experiments show that the model predicts well omitted survey waves using actual data. The third experiment evaluates the procedure in a controlled environment with a known data-generating process, using the heterogeneous-agent model of Bayer, Born, and Luetticke (2024). Results from these experiments show the resulting synthetic microdata closely match the omitted observations, indicating that the es-

19. The integrals can be calculated very efficiently as time varying linear combinations of the time invariant integrals of the basis functions.

timated state-space model successfully captures time-series fluctuations in the distributional data (thereby validating the procedure in Sections 3.2.3 and 3.2.4) and in particular, when the data-generating process is known. In Appendix G, we further show that our reconstructed distributional data agree well with the cyclical fluctuations (for the coverage of measures) found in the World Inequality Database (WID) and the Distributional Financial Accounts (DFAs).[20]

4.1 Data

To test and apply our method, we estimate the joint distribution of consumption, income, and wealth at the household level for the United States from 1962 to 2024. We rely on commonly used microdata for the U.S.: the *Consumer Expenditure Survey* (CEX), *Current Population Survey* (CPS, Flood et al. (2023)), *Panel Study of Income Dynamics* (PSID), *Survey of Consumer Finances* (SCF); including the historical backfiles (SCF+), and the *Survey of Income and Program Participation* (SIPP).[21]

We abstain from any sample selection in all of these datasets. For the CEX and SIPP income, which are at quarterly frequency, we remove seasonality using X-13 ARIMA-SEATS. We date CPS to quarter 4; the PSID is assumed to reflect quarter 2; and the SCF is dated to quarter 3. SIPP income data are aggregated to quarterly level and then naturally assigned to the respective quarter; wealth assignment depended on the survey vintage.[22] Further details on the microdata and in particular how the respective measures are defined can be found in Appendix H. Table 1 lists the distributional objects from each dataset, together with the sampling period. Note again that we require at least one dataset j that includes all objects, which in our case is the PSID between 1999 to 2021.

In terms of aggregate data, we use a wide range of standard business cycle data (GDP, consumption, employment, etc.) as well as data on household balance sheets, expectations, asset prices, and interest rates from McCracken and

20. Note that all trends will be fitted by construction, see Section 3.2.1.

21. The methodology accommodates permanent and time-varying differences in measurement, but major survey redesigns require intervention. For the CPS (1992) and SIPP (1996, 2013), we treat pre- and post-break data as separate surveysi.e., distinct measurements of the distribution to avoid spurious dynamics from seam bias.

22. Documentation on the timing of the CPS can be found here, pg. 6. For the timing of the SCF, see here, pg. 33. For the PSID, interview dates are provided as variables, with the vast majority of interviews falling in quarter 2. Given the PSID reports annual income and the model is estimated with mixed-frequency, this quarter 2 assignment only affects the timing of wealth, which is always dated to the interview date. For data on consumption and income, which is for the year prior, it needed to be scaled by the growth rate of their respective aggregate to align with the timing of wealth, which facilitates the interpretation of the copula.

Table 1: Micro Data Sources and their Sample Periods

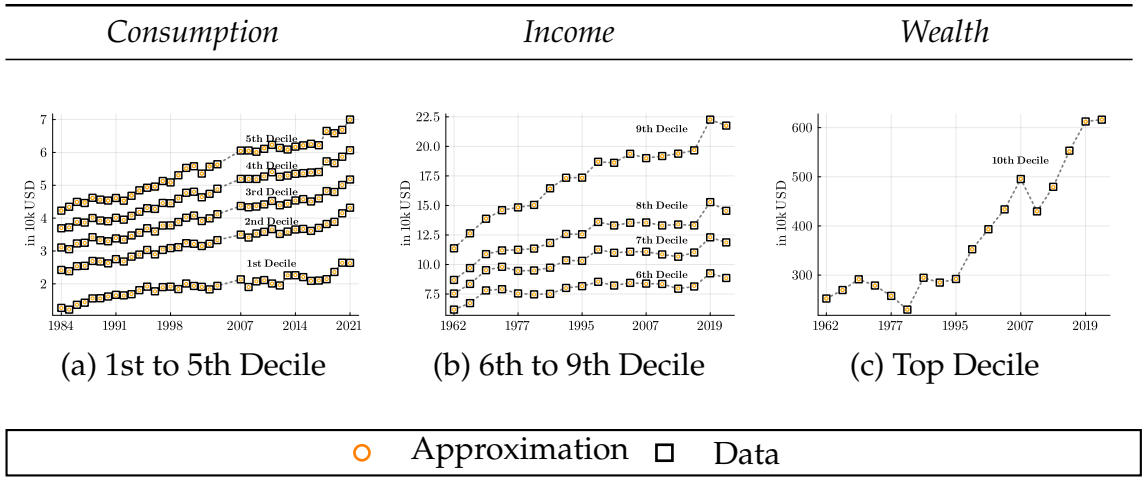
Object	CEX	CPS	SCF	PSID	SIPP
Consumption quantiles	1984Q4 - 2021Q4	-	-	1999Q2-2021Q2	-
Income quantiles	1984Q4 - 2021Q4	1967Q4-2022Q4	1962Q3-2022Q3	1968Q2-2021Q2	1983Q3-2022Q4
Wealth quantiles	-	-	1962Q3-2022Q3	1983Q2-2021Q2	1983Q3-2022Q4
Copula densities	1984Q4 - 2021Q4	-	1962Q3-2022Q3	1983Q2-2021Q2	1983Q3-2022Q4

Notes: The table reports sample periods for different micro datasets across the different objects.

Ng (2021). We include data from 1962Q3 to 2024Q1. The starting point of the aggregate data determines the earliest date for the sample periods of the microdata used. From these aggregate time series, we extract the 11 most important factors. Details on the macro-data can also be found in Appendix H. Further details on factor selection and estimated parameters are available upon request.

4.2 Precision of the Factor Model

Figure 4:
Quantile Functions: Raw vs. Approximation (Important Factors)

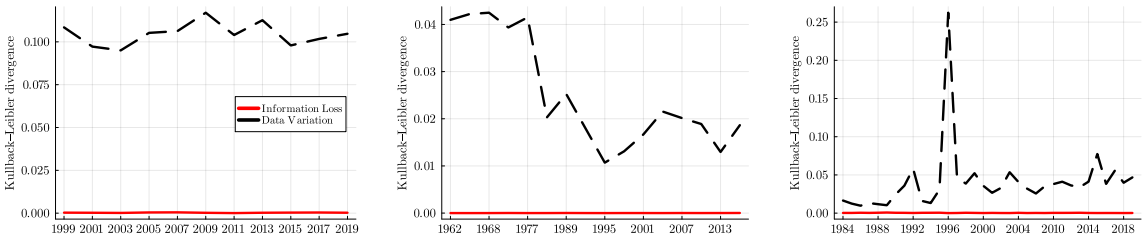


(a) 1st to 5th Decile (b) 6th to 9th Decile (c) Top Decile

Notes: Figure shows the quantile functions (mean within decile) for consumption, income, and wealth deciles from the survey data (squares) and an approximation (circles) using only the fluctuations in the most important factors in (18). Consumption panel shows quantile functions for CEX consumption. Income panel shows quantile functions for CEX income. Wealth panel shows quantile functions for SCF wealth. Dotted lines show linear interpolation between survey waves.

The first step in our procedure is to estimate Legendre polynomial coefficients for the functional representation of the distribution. Then, we estimate the factor structure in these coefficient data. Since we only retain "important" factors, we potentially introduce an approximation error (relative to the data) resulting

Figure 5: Copulas: Raw vs. Approximation (Important Factors)



(a) C, Y, and W (b) Y and W (c) C and Y

Notes: Figure shows the Kullback-Leibler divergence between the raw-data copula and two reference copulas, by survey year. The black dashed line represents the divergence between the time-averaged copula and the raw data copula for each survey year. The red solid line represents the divergence between the copula obtained by allowing only the most important factors in (18) to fluctuate and the raw data copula for each survey year. Panel (a) is from the PSID (Consumption, Income, and Wealth), Panel (b) is from the SCF (Income and Wealth), and Panel (c) is from the CEX (Consumption and Income).

from forcing "unimportant" factors f_t to take time-averaged values. The size of the approximation error can be controlled by choosing how many factors to keep. We choose to retain the eight most important factors, which explain 99% of the (business cycle frequency) variation of the distribution (i.e., of $\tilde{\boldsymbol{\theta}}$ to be precise). The different panels in Figure 4 visualize the approximation error in our application by showing some of the implied deciles. The figure compares the observed conditional decile means for consumption (bottom five), income (next four), and wealth (top) (squares) with their approximated counterparts (circles). The comparison for all deciles and variables looks analogously and is available upon request.

We find that the factor model with its eight main factors is very close to the distributional dynamics over time. The circles are typically entered around the midpoint of the squares. Figure 5 compares the copula over time between the approximation and the raw data. We do this in terms of the Kullback-Leibler divergence. The dashed black line shows how distant the actual distribution is from its long-term average (how much variation is there to capture), and the solid line shows the difference between the actual distribution and the approximation based on the important factors only (how much the factors do not capture). The Kullback-Leibler divergence of the actual copula from its long-run average is between 0.075 and 0.12 (between 1999 and 2019), while the divergence between the approximation and the actual distribution is almost two orders of magnitude smaller. To put this simple: There are significant fluctuations in the copulas over time, but the factors are able to capture them well.

4.3 Predictability of Distributional Data

To validate how well the method can predict distributional dynamics, we perform three experiments. The first experiment fixes the model parameters to that of the baseline and removes some waves of microdata as inputs when running the Kalman smoother. This experiment answers the question how informative are the model structure, aggregate data, and other available microdata for pinning down the distributional series when some survey waves are missing, holding parameters fixed. The second experiment in essence performs the same thing, but *re-estimates* the model. This experiment takes into account that changes in the data also imply changes in the model parameter estimates. In the final experiment, we evaluate our model's ability to recover distributional dynamics generated from a heterogeneous agent model. We show the results of these experiments in Figures 6, 7, and 8.

4.3.1 Predicting Omitted Survey Waves

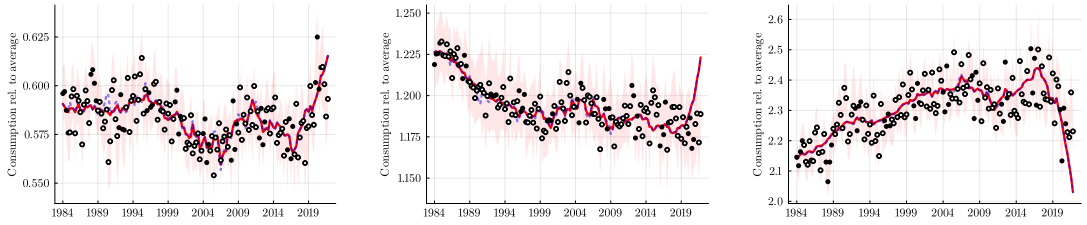
For the first experiment, we first include only every fourth CEX survey year in the estimation, mimicking the fact that countries in Europe only survey consumption every four years and reducing the number of CEX survey years included in the Kalman smoother from 38 to 10. The resulting estimates from this estimation with less data are shown in Figure 6. We show the average of the top 10%, next 40% and bottom 50% for the measures in the CEX: consumption (Panels (a) to (c)) and income (Panels (d) - (f)). Note the large sampling uncertainty around the CEX data, whose 95% confidence intervals are displayed as a red error band. For this reason, even in the data-rich specification (with all CEX data), the smoothed estimate regularly deviates from the raw distributional data, with correlations of the two around 95%. The correlation of the smoothed data using only every fourth survey year with the data using every survey year is very high, meaning that the model can predict the consumption distribution well.

The bottom row of Figure 6, Panels (g) to (i), presents a different exercise. We run 17 mini experiments, each time dropping one SCF wave and generating new smoother estimates. This gives us 17 smoothed distributional data series alongside the one using all SCF waves. The solid lines are estimates from using all the SCF data. The squares represent the smoothed predictions corresponding to the timing of each omitted survey wave. The circle shows the direct empirical estimate of the corresponding survey (with its confidence limits). For example, for Panel (g), the square for 1992 is the prediction for the wealth distribution

Figure 6: Predictability of Distributional Data (given the Baseline Parameters)

Consumption by Consumption Percentile

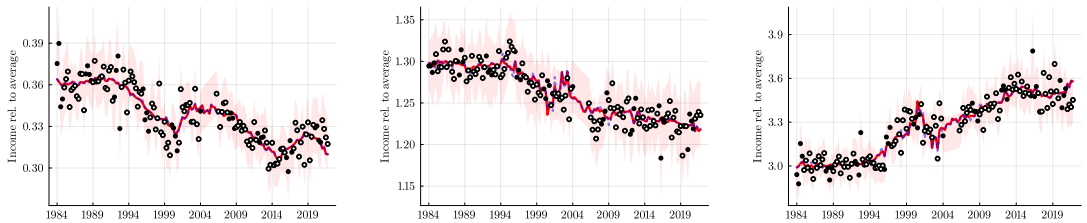
CEX data vs. estimates from using CEX every 4 years



(a) Bottom 50 Percent (b) 50-90 Percent (c) Top 10 Percent

Income by Income Percentile

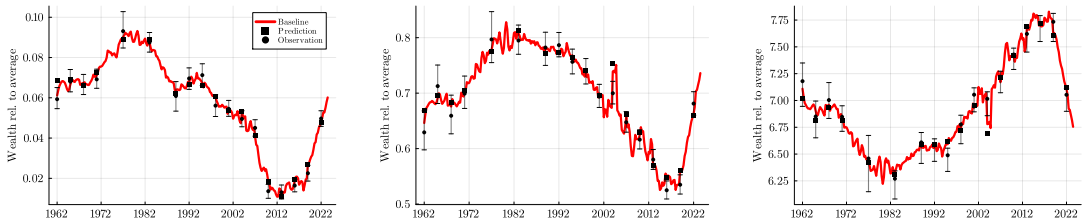
CEX data vs. estimates from using CEX every 4 years



(d) Bottom 50 Percent (e) 50-90 Percent (f) Top 10 Percent

Wealth by Wealth Percentile

Removing one SCF wave at a time



(g) Bottom 50 Percent (h) 50-90 Percent (i) Top 10 Percent

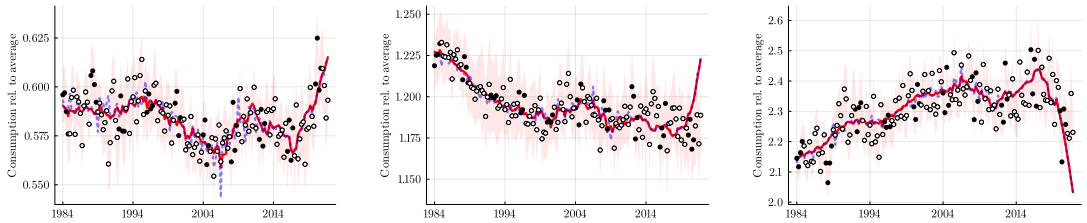
Baseline Less Data

Notes: The figure shows baseline model estimates for consumption (Panels (a) to (c)), income (Panels (d) to (f)), and wealth (Panels (g) to (i)) for different samples. Baseline estimates using all data are always shown as a solid red line. Panel (a) to (f): *less data* (dashed blue line) shows smoothed estimates when CEX microdata enters the smoother only every fourth year (black **solid dots**). **Empty dots** denote the observations removed from the Kalman smoother. Panel (g) to (i): shows smoothed estimates when only a single SCF wave has been dropped in the Kalman smoother. The black squares show the prediction of the dropped data at the survey wave and the dots show the empirical estimate from the survey data in that wave. Error bars/bands in all figures indicate 95% confidence bounds for each individual survey sample.

Figure 7: Predictability of Distributional Data (Re-estimating the Model)

Consumption by Consumption Percentile

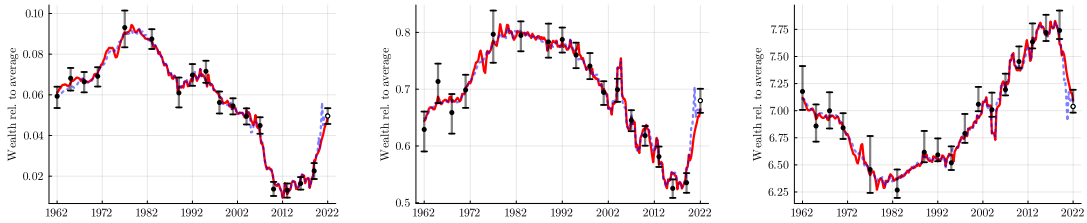
CEX data vs. estimates from using CEX every 4 years



(a) Bottom 50 Percent (b) 50-90 Percent (c) Top 10 Percent

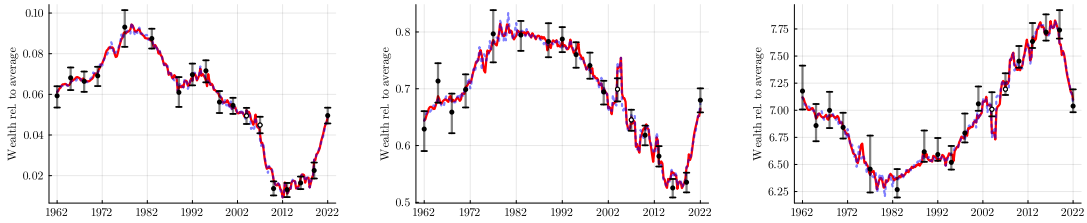
Wealth by Wealth Percentile

Removing microdata from last 4 years of estimation



(d) Bottom 50 Percent (e) 50-90 Percent (f) Top 10 Percent

Removing microdata from the housing cycle



(g) Bottom 50 Percent (h) 50-90 Percent (i) Top 10 Percent

Baseline Less Data

Notes: The figure shows baseline model estimates for consumption (Panels (a) to (c)) and wealth (Panels (d) to (i)) for different samples. Baseline estimates using all data are always shown as a solid red line. Panel (a) to (c): *less data* (dashed blue line) shows the smoothed estimate that results from a re-estimation of the model (and Kalman smoother) when CEX microdata enters only every fourth year (black **solid dots**). Panel (d) to (f): *less data* (dashed blue line) shows smoothed estimates when the last four years of all microdata have been dropped in model estimation (2020Q1 to 2024Q1) and the Kalman smoother. **Empty dots** denote the observations removed from the re-estimation and Kalman smoother. Panel (g) to (i): same exercise as (d) to (f) but dropping the observations over the house price cycle (2004Q1 to 2009Q4). Error bars in all figures indicate 95% confidence bounds for each individual survey sample.

where the 1992 SCF survey *did not* enter the smoother. The fact that the squares are virtually on top of the solid line implies that, conditional on the model, a single observation of the distributional data has little effect on the smoothed series. In other words, the model structure, aggregate factors, and other surrounding micro-data collaborate well in imputing the missing wave. An exception to highlight happens in 2004 during the housing boom, where the model (squares) predicts greater cyclicality than shown in the baseline (red line — with all data).

The first row of Figure 7 repeats the CEX exercise of Figure 6 for the CEX but now re-estimates the entire model. The first row of Figure 7 shows the resulting reconstruction. The difference to Figure 6 is small, reconfirming our claim that the estimation procedure can impute well distributional data. The second and third rows show this in a different vein for wealth, which is observed least frequently. In both rows, we remove a contiguous block of wealth microdata and ask whether the model can still recover distributional dynamics from aggregate data and the remaining (surrounding) microdata. In the second row, we omit the last four years of the sample (2020Q1-2024Q1) to evaluate nowcast performance for the wealth distribution. In the third row, we omit all wealth microdata during the housing cycle (2004Q1-2009Q4), a period with large swings in wealth inequality (Kuhn, Schularick, and Steins 2020).[23]

Across all exercises, the reconstructed series obtained with withheld microdata closely track the baseline estimates, indicating that the model captures wealth distribution dynamics even during periods without microdata-consistent with evidence that aggregate factors account for an important share of distributional fluctuations (Kuhn, Schularick, and Steins 2020; Bayer, Born, and Luetticke 2024), and by information from surrounding microdata, since smoother estimates incorporate the full set of observations.

4.3.2 Predicting Dynamics from a simulated HANK model

Thus far, we have examined the model's ability to predict intentionally omitted data. While informative, these tests do not fully resolve a central concern: in empirical settings, the underlying distributional dynamics are only imperfectly observed and agreement with these sparse microdata (or external benchmarks, see Appendix G) is not definitive proof that the model recovers the true latent process. To assess this, we study the model in a controlled environment with a

23. We also consider an alternative housing-cycle window (2007Q4-2011Q4) to assess predictability during the recovery phase. Appendix I, Table 9 reports correlations between the baseline estimates and all missing-data estimates, including this alternative window.

known data-generating process. For this purpose, we simulate a heterogeneous agent business cycle (HANK) model (concretely, from Bayer, Born, and Luetticke 2024) as a data-generating process. This model is well suited to our setting, as it generates business-cycle fluctuations in general equilibrium in response to a set of aggregate shocks and their interaction with the joint distribution of households, also characterized by a copula and a set of marginals. The simulation produces functional distributional data, which we can use to draw micro-data comparable to ones actually available.

We simulate one economy for $T = 2200$ periods (quarters). For each period t, we buffet the HANK model economy with a series of aggregate shocks, whose effects propagate through linearized policy functions—only first-order responses are retained. From the resulting joint distribution, we generate **four** samples: *A*, *B*, *C*, and *D*. Each sample is drawn from the copula histogram, ensuring that the samples preserve the correlational structure between economic measures. Sampling is stratified along each dimension to ensure coverage of the marginal distributions. Each simulated household defines a unit of observation and is characterized by its consumption, income, and wealth and a sampling weight. We abstract from unit-level measurement error.

Samples *A*, *B*, *C*, and *D* are designed to mimic one dataset from the U.S., each representative, but with differences in their degree of missingness and precision. Dataset *A* can be considered a modern PSID-like dataset: 9,000 household observations per survey, containing consumption, income, and wealth, and released every two years. Dataset *B* can be considered CPS-like: 60,000 observations per survey (very precise), with only income, released every year. Dataset *C* can be considered a CEX-like dataset: 3,000 observations per survey, with income and consumption, released every quarter. Dataset *D* can be considered an SCF-like dataset: 5,000 observations per survey, with only wealth and income observed, released every three years. We abstract from differences in operationalization across the different sample measurements.

The aggregate shocks driving the economy are also recorded at each t, whose factor representation will be used in our state-space system.[24] Before proceeding, the initial 1200 periods are discarded as burn-in, ensuring that the simulated economy has converged to its ergodic distribution independently of initial conditions. The remaining 1000 periods are kept for the exercise and split into 10 chunks, obtaining 10 economies observed for 100 quarters each. Accordingly,

24. The HANK economy is subject to shocks in total factor productivity, investment-specific technology, consumption wedges, labor wedges, technology, monetary policy, government spending, fiscal policy, and household preferences. For a full description of the model, see Bayer, Born, and Luetticke (2024) and the associated `BASEtoolbox.jl` code repository.

we can run 10 separate estimation exercises that are comparable to the actual data also in terms of the time that they span (25 years, similar to the available PSID waves containing consumption). In the end, the procedure suggests five distributional factors and eight aggregate factors.

Table 2: Time-Series Correlations: Model Estimates vs. HANK Simulated Truth

	Consumption		*Income*		*Wealth*	
Percentile Group	DDFM	Sample *C*	DDFM	Sample *A*	DDFM	Sample *D*
			By Consumption Groups			
Bottom < 50	96	84	98	76	–	–
Middle 50-90	97	84	96	55	–	–
Top > 90	93	42	92	55	–	–
			By Income Groups			
Bottom < 50	98	88	98	72	72	46
Middle 50-90	90	78	96	34	69	33
Top > 90	85	42	98	72	81	61
			By Wealth Groups			
Bottom < 50	–	–	95	50	96	96
Middle 50-90	–	–	94	49	81	79
Top > 90	–	–	81	45	87	31

Notes: The table shows correlations between two benchmarks and the HANK simulated truth. Column *DDFM* shows correlations between the HANK simulated truth and our model estimates (*DDFM*). Column *Sample X* shows correlations between the HANK simulated truth and a linearly-interpolated Sample $X \in \{C, A, D\}$. As in the figure above, conditional moments are averages by *Percentile Group* (Bottom 50%, Middle 50-90%, and Top 10%) of the distribution of consumption, income, and wealth and all time series are relative to the economy-wide average. Since Sample *C* does not contain wealth information and Sample *D* does not contain consumption information, correlations of conditional moments involving these variables cannot be computed. Sample *B*, which contains only income, is omitted for brevity. Correlations are averages over the ten simulated economies.

We use the four survey-style samples (*A*–*D*) as microdata inputs to our dynamic factor method alongside the factor representation of the aggregate shocks from the HANK model. The resulting estimates are labeled as *Distributional Dynamics Factor Model (DDFM)* and compared to two benchmarks. The first benchmark are the values we obtain directly from the model simulation, the *Truth*. The second benchmark is a linearly-interpolated series of (conditional) cross-sectional sample moments for some variable and dataset X. The interpolation points would be the survey release dates. This *naive* Sample X benchmark ignores that the samples (*A*–*D*) are, in fact, just samples of an underlying distribu-

tion that has some dynamic structure.

Table 2 compares two benchmarks to the HANK simulated truth. First, column *DDFM* reports correlations between the HANK simulated truth and our DDFM estimates generated from using the incomplete samples. Column *Sample X* reports correlations between the HANK simulated truth and the linearly-interpolated sample $X \in \{C, A, D\}$. These are time correlations, averaged over the ten simulated economies. We find that our DDFM estimates are an improvement over the correlations of the linear-interpolation with the true distributions. In most cases, the correlation between the truth and DDFM implied distributional summary statistics are well above 90%. This suggests that the model approximates well the marginal distributions as well as the correlational structure across variables.

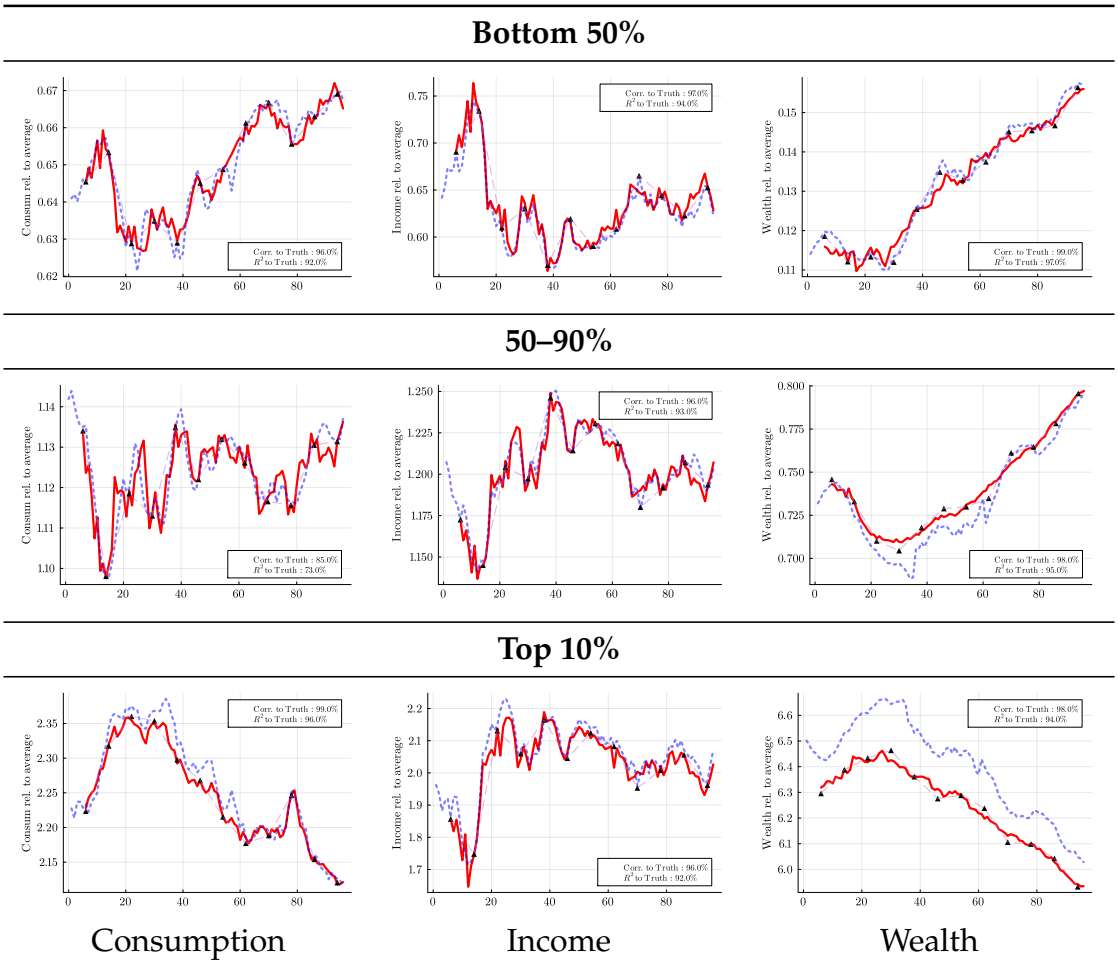
Figure 8 provides the results of this exercise visually (for economy 1), plotting the evolution of consumption, income, and wealth by groups, relative to the economy-wide average. The figure compares the *DDFM* (solid red) estimates to the HANK simulated truth (dotted blue). For comparison, we plot *naive* estimates using Sample A in a light-pink line; simply a linear-interpolation of the black triangles which mark the observations in Sample A. This also visualizes the sparseness of the data that enters our estimation. In general, the movements across the DDFM and the HANK simulated truth are in sync, denoted by the high correlation and around the same magnitudes, denoted by the high R^2 across panels.[25] Using the naive linear-interpolation would miss key distributional movements, especially for consumption and income.

5 Distribution Dynamics over the Business Cycle

The focus of our paper is to provide a method for synthesizing dynamic distributional data from micro and macro data inputs. In the previous section, we have established the reliability of that procedure. Before concluding, we show the usefulness of synthetic data in two application examples. The first example is still closely related to the dynamic distributional factor model that builds the backbone of our method. We ask what role aggregate shocks play in driving the dynamics of distributions. We find that aggregates explain most of the movements in the distribution in line with Bayer, Born, and Luetticke (2024) finding for an estimated HANK model. Second, we show how the method con-

25. The wealth quantile function is steeper in the upper decile than the degree-11 Legendre expansion can resolve and the fat tail leads to a small sample bias of average wealth. Both together a -2 to -3% downward bias in the Top10 mean. Half of this is sampling, half the order of the polynomial being somewhat too small.

Figure 8: Model Estimates (in red) vs. HANK Simulated Truth (in blue)



Notes: The figure reports estimates for consumption, income, and wealth relative to their economy-wide average. Estimates are shown for three groups: the bottom 50%, the next 40%, and the top 10%. Each panel compares model estimates with the HANK simulated truth. Model estimates (*DDFM*) from incomplete samples are shown as a solid red line; *HANK Simulated Truth* is shown as a dotted blue line. The black triangles labeled *HANK Data Used* correspond to the sample observations employed in estimating the red line (calculated based on the polynomial quantile estimator). *Corr. to Truth* reports the correlation between the red and blue lines. R^2 *to Truth* is the coefficient of determination from regressing detrended HANK simulated truth on detrended model estimates.

tributes to ongoing debates on interpreting and understanding macroeconomic dynamics. We document in our synthetic data characteristic differences across recessions in terms of their consumption distribution impact. This speaks to the emerging literature studying how consumption distributions and consumption risk move over the business cycle (see e.g., Berger, Bocola, and Dovis 2023; Bilbiie et al. 2025; Patterson 2023).

Table 3: FEVD on the Distributional Factors, percent contribution of shocks

Distributional Factor No.	1	2	3	4	5	6	7	8
Aggregate shocks	99	99	70	89	93	86	86	77
Distributional shocks	1	1	30	11	7	14	14	23

Notes: Table reports the forecast error variance decomposition (FEVD) for the eight estimated distributional factors. Identification is via a blocked cholesky with aggregates ordered first. The columns show the percentage of variation in each factor explained by distributional shocks versus aggregate shocks. This is over a five-year horizon.

5.1 What drives the distribution dynamics?

Since the dynamic factor model that we estimate is a linear model, it allows us to answer straightforwardly the question whether the dynamics of the distribution is driven primarily by specific distributional shocks, e.g., skills becoming more diverse, asset returns more dispersed, intertemporal preferences more different over time, or whether aggregate shocks are driving the distribution. Bayer, Born, and Luetticke (2024) provides an analysis of this question based on selected distributional information and a structural HANK model. They find that aggregate shocks explain well the actual evolution of wealth inequality even at lower frequencies. In a more reduced form manner, but with the same message, Kuhn, Schularick, and Steins (2020) argue that price changes are essential to understand wealth dynamics.

Here, we revisit this question within our much less structural DDFM. For this purpose, we compute a forecast error variance decomposition (FEVD) for the distributional factors for the five-year horizon. Because reduced-form innovations in the state equation are correlated (i.e., Ω is non-diagonal), the FEVD is based on an identification that orthogonalizes shocks. Specifically, we apply a *blocked* Cholesky factorization with aggregates ordered first. This distinguishes between aggregate shocks as a whole and distributional shocks as a whole but makes no assumption on their causal ordering within group. Table 3 reports the results of this exercise. We find that aggregate shocks explain at least 70 percent of the variation for every factor. For six out of eight factors, the contribution exceeds 85 percent.

While these results for the factors are informative, it is not clear what this means for moments of the data that one would typically look at, like consumption of the poor or the rich. As these moments are linear in the factors, we can directly apply the FEVD.[26] Table 4 reports the contribution of aggregate shocks for these better interpretable moments, decomposing the marginal distributions

26. We refer the reader to Appendix E for the transformations from factors to observables.

Table 4: FEVD of group means and population shares. Percent contribution

	Conditional means of ...					
	Consumption		*Income*		*Wealth*	
Group	D.-shocks	A.-shocks	D.-shocks	A.-shocks	D.-shocks	A.-shocks
Bottom 20%	19	81	17	83	20	80
Bottom 20 – 40%	18	82	10	90	6	94
Middle 40 – 80%	20	80	21	79	15	85
Top 20%	20	80	10	90	6	94
	Population shares with ...					
	Low Consumption		*High Consumption*			
Low Y, Low W	42	58	44	56		
Low Y, High W	40	60	42	58		
High Y, Low W	42	58	40	60		
High Y, High W	45	55	42	58		

Notes: Table reports the forecast error variance decomposition (FEVD, five year horizon) of Table 3 mapped to observables consumption, income, and wealth. The top panel looks at consumption, income, and wealth group means across four household groups given by the bottom 20%, the next 20%, the next 40% and the top 20% of the distribution of the respective variable (consumption (C), income (Y), and wealth (W)). The bottom panel reports the corresponding FEVD for the population shares that have any combination of above/below median consumption, income, and wealth. Identification is via a blocked cholesky with aggregates ordered first. Entries show the share of variation explained by distributional shocks versus aggregate shocks.

of consumption, income, and wealth for four household groups: the bottom 20%, the bottom 20 – 40%, the bottom 40 – 80%, and the top 20% (top panel). We find that, typically, aggregate shocks explain 80 percent of the variation in consumption and even more for income and wealth. Looking at changes of the copula, here represented as the change in the share of four population groups within the joint distribution, for example, all observations with below median consumption, income, and wealth ("Low Consumption, Low Y, Low W"). We now find that distributional shocks are more important, explaining around 40 percent of the fluctuations. At the same time, this means that aggregate shocks still generate the majority of fluctuations.

5.2 Consumption Dynamics over the last three recessions

The relative distribution of consumption losses in recessions has received considerable attention in the macroeconomic literature (e.g., Coibion et al. 2017; Cloyne, Ferreira, and Surico 2020; Bilbiie et al. 2025; Holm, Paul, and Tischbirek 2021; Fagereng, Holm, and Natvik 2021). The motivation is partly normative, in the sense of identifying groups that benefit or are adversely affected, and partly theoretical because a greater cyclicality of incomes of high-MPC households can

destabilize the economy (Bilbiie 2020). Using our synthetic data, we trace consumption dynamics along both the marginal and joint distributions of income and wealth across the last three U.S. recessions. We find that their distributional consequences are state-dependent: wealth losses dominated the dot-com bust, balance-sheet distress the financial crisis, and income disruption COVID-19. In each case, the most affected households are identified by their position in the joint distribution—not by income or wealth alone—challenging research strategies that rely on marginal distributions and unconditional cyclical elasticities.

To begin, for the three recessions, we first consider the consumption dynamics of households along the marginal distributions of income and wealth, creating four groups: (1) the bottom 20%, (2) the 20% to 40%, (3) 40% to 80%, and (4) the top 20%.[27] These are the first two rows of Figure 9. For the final row, we then consider the joint distribution and estimate consumption dynamics for four key household groups. Group (i) consists of liquidity-constrained hand-to-mouth households, defined as the bottom 20% income *and* bottom 20% wealth; group (ii) is a population of asset-rich, low-income households which offers some resemblance to the wealthy hand-to-mouth, defined here as the bottom 40% of income but top 30% of wealth; group (iii) is the middle class, presumably prudent, patient "buffer-stock" households, located between the 40% to 80% in both income and wealth; and group (iv) comprises the richest households in terms of income and wealth, largely unconstrained, presumably Permanent Income Hypothesis households, residing in the top 20% in both income and wealth.[28]

For all groups, absolute consumption falls in a recession and rises in a recovery; therefore, we compare the business cycle dynamics of consumption for the different income and wealth groups, in each period t, *relative* to the household-wide consumption average at period t and then index these relative consumption dynamics to the quarter preceding the recession (the peak).[29] In this sense, our

27. The income measure we use includes transfers, but is not net of taxes.

28. For our analysis, groups are defined by contemporaneous ranks, so membership can change from quarter to quarter. While this implies some *churning* relative to a true panel, it remains informative because it tracks the currently vulnerable population relevant for real-time policy and because compositional changes within our broad bins are unlikely to overturn the qualitative patterns (residual churning is concentrated near bin cutoffs) (cp. Kuhn, Schularick, and Steins (2020)).

This interpretation is consistent with evidence that mobility is limited: positive shocks required for upward mobility are rare at the bottom, while large negative shocks are infrequent at the top (Guvenen et al. 2021); hand-to-mouth status is persistent (Aguiar, Bils, and Boar 2025); wealth positions are especially sticky due to persistent heterogeneity in returns (Fagereng et al. 2020); and earnings mobility is largely confined to adjacent deciles (Ettmeier, Hyun Kim, and Schorfheide 2024). Using coarse groups mitigates these concerns, though they become more salient when conditioning on multiple dimensions.

29. We construct symmetric 3-quarter moving averages and use this moving average for the

goal is to understand how a group's consumption shifts relative to the aggregate over a business cycle. The observable deviations from the aggregate represent the net impact of the distributional channels at work, which in turn amplify or dampen the corresponding macroeconomic shocks. Figure 9 summarizes how these deviations are distributed across households depending on the nature of each recession.

Dot-com recession (Panels a, d, g). This recession resembles a financial wealth shock as income differences played little role: The consumption dynamics are uniform along the income distribution (Panel a), whereas along wealth (Panel d), low-wealth households *gained* relative to average—benefiting from aggressive rate cuts—while the wealthy lost ground. The asset-rich, low-income group (purple, Panel g) suffered heavily, as their equity wealth collapsed.

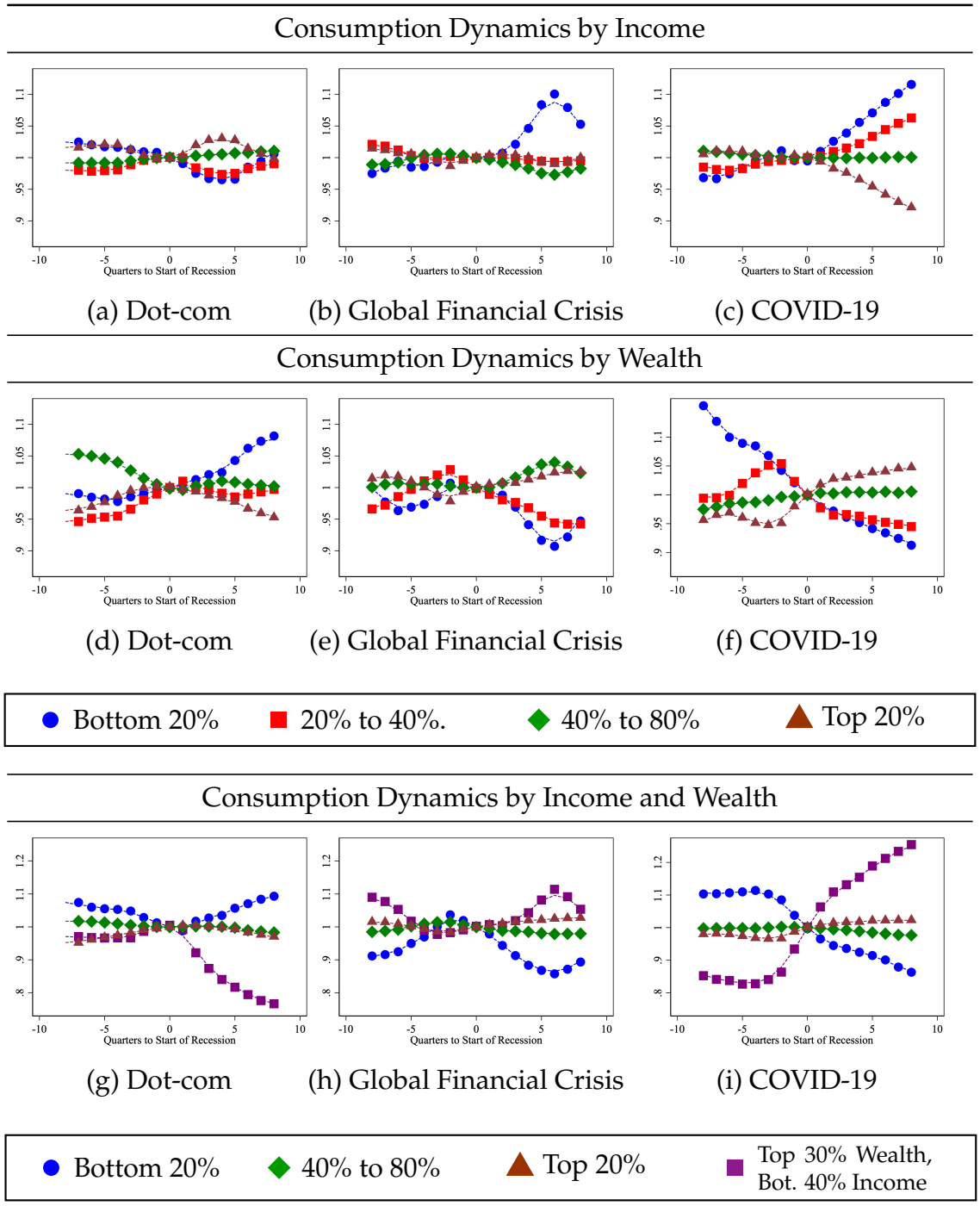
Global Financial Crisis (Panels b, e, h). This recession looks like a balance-sheet recession where now low-wealth, indebted households (Panel e) experienced the sharpest consumption declines—around 10% (evidence for the balance-sheet channel from Mian, Rao, and Sufi (2013)). Low-income households fared relatively well (Panel b), supported by fiscal transfers. The wealthy remained stable or gained.

COVID-19 recession (Panels c, f, i). During this last recession, income becomes the dominant dimension: relative consumption is almost perfectly inversely ordered by income (Panel c). The bottom 20% gain roughly 10% relative to average (due to large government transfers). By wealth (Panel f), the pattern differs: low-wealth households lose persistently with no recovery two years out. The asset-rich, low-income group (purple, Panel i) outperforms all others by over 20%, benefiting from transfers, asset-price rebounds, and insulation from labor-market shocks.

The key finding is *state-dependence*: the channels through which recessions affect households—financial, balance-sheet, or labor-market—vary with the nature of the shock. During the dot-com bust, wealth mattered; during the Global Financial Crisis, debt mattered; during COVID-19, income and fiscal transfers mattered. Neither income nor wealth alone is a sufficient statistic for understanding consumption dynamics. This state-contingency cautions against research strategies that rely on uniform unconditional cyclical elasticities. It underscores that stabilization policy needs to diagnose the primary channel through which a crisis operates and take a holistic view of the sum of all stabilization measures.

normalization to the pre-recession quarter.

Figure 9: Comparison of Consumption Dynamics during Recessions



Notes: Relative consumption dynamics during recessions along the income and wealth distribution. Consumption dynamics of each income/wealth group are shown relative to average household consumption. These relative consumption time series for each group are indexed to the peak. The vertical axis shows changes of consumption over time relative to the change of average consumption over time. The horizontal axis shows the time relative to the start of the recession. The recessions are the dot-com recession in 2001Q1, the Global Financial Crisis in 2007Q4, and the COVID-19 recession in 2019Q4.

6 Conclusion

In this paper, we present a new method to derive synthetic distributional consumption, income, and wealth data. The method contributes to the modern theory of macroeconomic dynamics that has the joint distribution of consumption, income, and wealth as a key determinant of aggregate dynamics. Our method closes a gap as it provides a method for studying the empirical distributional dynamics as a counterpart to the existing theory at business-cycle frequency over time. We have shown that the method can incorporate information from various microdata sources regardless of their frequency and coverage of variables. By forecasting out of sample, we show that our method can generate joint distributional information at high frequency with good precision. Using a HANK model as a laboratory, we validated that the synthetic data from our model closely follow the "true" distributional dynamics when supplied with micro data samples as rich and frequent as in the U.S.

To illustrate the method's usefulness, we further traced consumption dynamics across the last three U.S. recessions along the income and wealth distribution. The application suggests that distributional consequences of recessions are state-dependent, with different channels dominating in different episodes—financial, balance-sheet, or labor-market. Although these findings are only suggestive, they demonstrate how synthetic data can inform debates about the role of heterogeneity in macroeconomic dynamics. Our method also shows that most business cycle fluctuations of the distribution of consumption, income, and wealth are driven by aggregate shocks.

Beyond what we present in this paper, our methodology provides a flexible tool for researchers: it can readily accommodate distributions based on, e.g., liquid and illiquid assets, financial income, or secured and unsecured debt, and even beyond households (e.g., firms and banks) and, in principle, be of higher frequency. Another promising direction is to apply the method to short panels where the joint distribution spans current and lagged values of the same variable. This would allow researchers to estimate high-frequency dynamics of household-level autocorrelation structures as in Almuzara et al. (2025), but incorporating a richer set of variables. We leave these extensions for future work.

References

Aguiar, Mark, Mark Bils, and Corina Boar. 2025. "Who are the Hand-to-Mouth?" *Review of Economic Studies* 92 (3): 1293–1340.

Almuzara, Martn, Manuel Arellano, Richard W Blundell, and Stéphane Bonhomme. 2025. "Nonlinear micro income processes with macro shocks." *FRB of New York Staff Report,* no. 1162.

Alvaredo, Facundo, Anthony Atkinson, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2016. "Distributional National Accounts (DINA) guidelines: Concepts and methods used in WID. world."

Andersen, Asger Lau, Niels Johannesen, Mia Jørgensen, and José-Luis Peydró. 2023. "Monetary policy and inequality." *The Journal of Finance* 78 (5): 2945–2989.

Auclert, Adrien. 2019. "Monetary policy and the redistribution channel." *American Economic Review* 109 (6): 2333–2367.

Auclert, Adrien, Bence Bardóczy, Matthew Rognlie, and Ludwig Straub. 2021. "Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models." *Econometrica* 89 (5): 2375–2408.

Bai, Jushan, and Kunpeng Li. 2016. "Maximum likelihood estimation and inference for approximate factor models of high dimension." *Review of Economics and Statistics.*

Bakam, Yves I Ngounou, and Denys Pommeret. 2023. "Nonparametric estimation of copulas and copula densities by orthogonal projections." *Econometrics and Statistics.*

Balakrishnan, Sivaraman, Martin J Wainwright, and Bin Yu. 2017. "Statistical guarantees for the EM algorithm: From population to sample-based analysis."

Babura, Marta, and Michele Modugno. 2014. "Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data." *Journal of applied econometrics* 29 (1): 133–160.

Barigozzi, Matteo, and Matteo Luciani. 2019. "Quasi maximum likelihood estimation and inference of large approximate dynamic factor models via the EM algorithm." *arXiv preprint arXiv:1910.03821.*

Barnard, John, Robert McCulloch, and Xiao-Li Meng. 2000. "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica,* 1281–1311.

Bartscher, Alina K, Moritz Schularick, Moritz Kuhn, and Paul Wachtel. 2022. "Monetary policy and racial inequality." *Brookings Papers on Economic Activity* 2022 (1): 1–63.

Batty, Michael, Jesse Bricker, Joseph Briggs, Sarah Friedman, Danielle Nemschoff, Eric Nielsen, Kamila Sommer, and Alice Henriques Volz. 2020. "The Distributional Financial Accounts of the United States."

Baumeister, Christiane, Pascal Frank, Florian Huber, and Gary Koop. 2025. "Oil, Inflation Expectations, and Household Characteristics: A Nonlinear Heterogeneous Agent VAR Approach." Working paper, University of Notre Dame.

Bayer, Christian, Benjamin Born, and Ralph Luetticke. 2024. "Shocks, frictions, and inequality in US business cycles." *American Economic Review* 114 (5): 1211–1247.

Bayer, Christian, Ralph Luetticke, Lien Pham-Dao, and Volker Tjaden. 2019. "Precautionary Savings, Illiquid Assets, and the Aggregate Consequences of Shocks to Household Income Risk." *Econometrica* 87 (1): 255–290.

Berger, David, Luigi Bocola, and Alessandro Dovis. 2023. "Imperfect risk sharing and the business cycle." *The Quarterly Journal of Economics* 138 (3): 1765–1815.

Bhandari, Anmol, David Evans, Mikhail Golosov, and Thomas J Sargent. 2021. "Inequality, Business Cycles, and Monetary-Fiscal Policy." *Econometrica* 89 (6): 2559–2599.

Bilbiie, Florin O. 2020. "The new Keynesian cross." *Journal of Monetary Economics.*

Bilbiie, Florin O, Sigurd Molster Galaasen, RS Gürkayna, Mathis Mæhlum, and Krisztina Molnar. 2025. "Hanksson."

Blanchet, Thomas, Emmanuel Saez, and Gabriel Zucman. 2022. *Real-time inequality.* Technical report. National Bureau of Economic Research.

Boehl, Gregor. 2024. "DIME MCMC: A Swiss Army Knife for Bayesian Inference." *Journal of Econometrics.*

Breitung, Jörg, and Sandra Eickmeier. 2006. "Dynamic factor models." *Allgemeines Statistisches Archiv* 90 (1): 27–42.

Chang, Minsu, Xiaohong Chen, and Frank Schorfheide. 2024. "Heterogeneity and aggregate fluctuations." *Journal of Political Economy.*

Chang, Minsu, and Frank Schorfheide. 2024. *On the Effects of Monetary Policy Shocks on Income and Consumption Heterogeneity.* Technical report 32166. NBER.

Chang, Yoosoon, Chang Sik Kim, and Joon Y Park. 2016. "Nonstationarity in time series of state densities." *Journal of Econometrics* 192 (1): 152–167.

Chen, Weilong, Meng Joo Er, and Shiqian Wu. 2005. "PCA and LDA in DCT domain." *Pattern Recognition Letters* 26 (15): 2474–2482.

Chodorow-Reich, Gabriel, Plamen T Nenov, and Alp Simsek. 2021. "Stock market wealth and the real economy: A local labor market approach." *American Economic Review.*

Cloyne, James, Clodomiro Ferreira, and Paolo Surico. 2020. "Monetary policy when households have debt: new evidence on the transmission mechanism." *The Review of Economic Studies* 87 (1): 102–129.

Coibion, Olivier, Yuriy Gorodnichenko, Lorenz Kueng, and John Silvia. 2017. "Innocent Bystanders? Monetary policy and inequality." *Journal of Monetary Economics.*

Di Maggio, Marco, Amir Kermani, and Kaveh Majlesi. 2020. "Stock market returns and consumption." *The Journal of Finance* 75 (6): 3175–3219.

Diebold, Francis X, and Canlin Li. 2006. "Forecasting the term structure of government bond yields." *Journal of Econometrics* 130 (2): 337–364.

Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin. 2012. "A quasi–maximum likelihood approach for large, approximate dynamic factor models." *Review of economics and statistics* 94 (4): 1014–1024.

Durbin, James, and Siem Jan Koopman. 2012. *Time series analysis by state space methods.* Vol. 38. OUP Oxford.

Ettmeier, Stephanie, Chi Hyun Kim, and Frank Schorfheide. 2024. *Measuring the Effects of Aggregate Shocks on Cross-sectional Distributions: Functional vs. Panel Approach.* Technical report.

Fagereng, Andreas, Luigi Guiso, Davide Malacrino, and Luigi Pistaferri. 2020. "Heterogeneity and persistence in returns to wealth." *Econometrica* 88 (1): 115–170.

Fagereng, Andreas, Martin B Holm, and Gisle J Natvik. 2021. "MPC heterogeneity and household balance sheets." *American Economic Journal: Macroeconomics* 13 (4): 1–54.

Flood, Sarah, Miriam King, Renae Rogers, Steven Ruggles, J. Robert Warren, and Michael Westberry. 2023. *IPUMS CPS: Version 11.0 [dataset].* Minneapolis, MN.

Gouskova, Elena, Patricia Andreski, and Robert F Schoeni. 2010. *Comparing estimates of family income in the PSID and the March CPS, 1968-2007.* Survey Research Center, Institute for Social Research.

Guvenen, Fatih, Fatih Karahan, Serdar Ozkan, and Jae Song. 2021. "What do data on millions of US workers reveal about lifecycle earnings dynamics?" *Econometrica.*

Harvey, Andrew, and Chia-Hui Chung. 2000. "Estimating the underlying change in unemployment in the UK." *Journal of the Royal Statistical Society: Series A.*

Holm, Martin Blomhoff, Pascal Paul, and Andreas Tischbirek. 2021. "The transmission of monetary policy under the microscope." *Journal of Political Economy.*

Insolera, Nora E, Beth A Simmert, and David S Johnson. 2021. "An overview of data comparisons between psid and other us household surveys." *Technical Series Paper,* 21–02.

Kaplan, Greg, Benjamin Moll, and Giovanni L Violante. 2018. "Monetary policy according to HANK." *American Economic Review* 108 (3): 697–743.

Kim, Yong-Seong, and Frank P Stafford. 2000. "The quality of PSID income data in the 1990s and beyond." *Technical Series Paper.*

Kneip, Alois, and Klaus J Utikal. 2001. "Inference for density families using functional principal component analysis." *Journal of the American Statistical Association.*

Koop, Gary, Stuart McIntyre, James Mitchell, and Ping Wu. 2026. *Incorporating Micro Data into Macro Models Using Pseudo VARs.* Working Paper 26-04. Federal Reserve Bank of Cleveland.

Kuhn, Moritz, Moritz Schularick, and Ulrike I Steins. 2020. "Income and wealth inequality in America, 1949–2016." *Journal of Political Economy* 128 (9): 3469–3519.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating random correlation matrices based on vines and extended onion method." *Journal of multivariate analysis* 100 (9): 1989–2001.

McCracken, Michael W., and Serena Ng. 2021. "FRED-QD: A Quarterly Database for Macroeconomic Research." *Review* 103 (1): 1–44.

McKay, Alisdair, and Christian K Wolf. 2023. "Monetary policy and inequality." *Journal of Economic Perspectives* 37 (1): 121–144.

Meeks, Roland, and Francesca Monti. 2023. "Heterogeneous beliefs and the Phillips curve." *Journal of Monetary Economics* 139:41–54.

Mian, Atif, Kamalesh Rao, and Amir Sufi. 2013. "Household balance sheets, consumption, and the economic slump." *The Quarterly Journal of Economics* 128 (4): 1687–1726.

Mian, Atif R, Ludwig Straub, and Amir Sufi. 2020. *The Saving Glut of the Rich.* Technical report. National Bureau of Economic Research.

Patterson, Christina. 2023. "The matching multiplier and the amplification of recessions." *American Economic Review* 113 (4): 982–1012.

Pfeffer, Fabian T, Robert F Schoeni, Arthur Kennickell, and Patricia Andreski. 2016. "Measuring wealth and wealth inequality: Comparing two US surveys." *Journal of economic and social measurement* 41 (2): 103–120.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income inequality in the United States, 1913–1998." *The Quarterly journal of economics* 118 (1): 1–41.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018. "Distributional national accounts: methods and estimates for the United States." *The Quarterly Journal of Economics* 133 (2): 553–609.

Saez, Emmanuel, and Gabriel Zucman. 2016. "Wealth inequality in the United States since 1913: Evidence from capitalized income tax data." *The Quarterly Journal of Economics* 131 (2): 519–578.

Schorfheide, Frank, and Dongho Song. 2015. "Real-time forecasting with a mixed-frequency VAR." *Journal of Business & Economic Statistics* 33 (3): 366–380.

Sklar, Abe. 1973. "Random variables, joint distribution functions, and copulas." *Kybernetika* 9 (6): 449–460.

Smith, Matthew, Owen Zidar, and Eric Zwick. 2023. "Top wealth in America: New estimates under heterogeneous returns." *The Quarterly Journal of Economics.*

Stock, James H, and Mark W Watson. 2002. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business & Economic Statistics* 20 (2): 147–162.

Tsay, Ruey S. 2016. "Some methods for analyzing big dependent data." *Journal of Business & Economic Statistics* 34 (4): 673–688.

Vermeulen, Philip. 2018. "How fat is the top tail of the wealth distribution?" *Review of Income and Wealth* 64 (2): 357–387.

Wu, CF Jeff. 1983. "On the convergence properties of the EM algorithm." *The Annals of statistics,* 95–103.

Online Appendix

A Economic Fit

The series estimator fits a single set of coefficients $\theta_t^j = \left(\xi^j_{m,o_1,t}, \ldots, \kappa^j_{(o_1,\ldots,o_d),t}\right)$ by projecting the entire distribution of the variable(s) of interest onto an orthogonal polynomial basis; it is therefore a *global* method in the sense that one parameter set governs all quantile/copula regions simultaneously. If distributions were exactly composed of finitely many orthogonal polynomials, the global estimator would be exact. In practice, however, the truncation we impose may systematically over- or underestimate certain segments, potentially smoothing over economically relevant features.[30] Also, even without truncation, we might fit polynomials to regions with zero mass in the measure—generating, potentially, spurious variation. Both concerns are naturally important in our setting, where our task is to recover the within and across time variation in the entire distribution, *Economic Fit*, not just segments best approximated by our estimator.

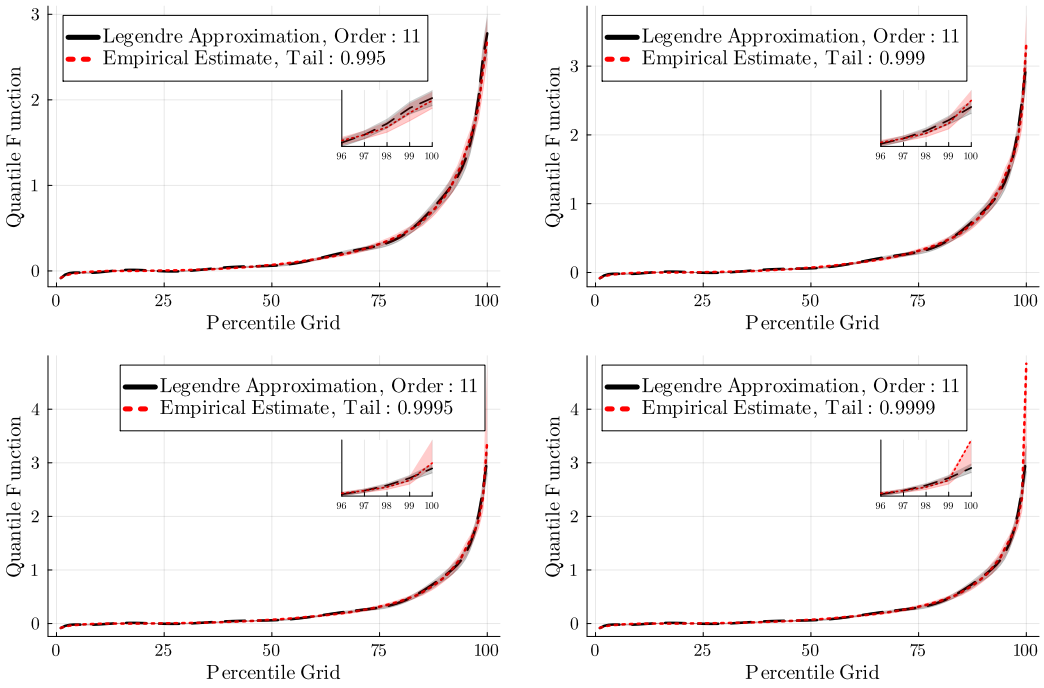
Two natural questions arise: (1) where does our estimator systematically fall short and (2) can these weaknesses be accommodated. For (2), this means (a) can the current (global) polynomial weights be adjusted to better capture specific regions (without compromising other regions), such as the upper tail of the wealth distribution or do they already carry enough flexibility to capture these specific regions (and the others)? and (b) can the estimator accommodate distributions with point-zero mass? In this spirit, the appendix addresses two points. First, it evaluates the degree to which we miss local variations. Second, it shows how an economist with a theoretical prior about which parts of the distribution to emphasize (e.g., the top of the wealth distribution) could adjust the coefficients accordingly. As a byproduct, for a given estimator, it also provides detailed evidence of variation within the distribution via observing weight changes within the distribution. More on this latter point later.

30. Our projection forces parts of the distribution onto the squareintegrable function space $\mathcal{L}^2$, even though those parts themselves do *not* belong to $\mathcal{L}^2$. Good examples are the upper tails of the consumption, income, and wealth, which are often modeled by Pareto laws. Vermeulen (2018) shows that wealth displays extremely heavy tailsso heavy that its variance is undefinedbecause the Pareto "alpha" parameters fall below 2. Likewise, Atkinson, Piketty, and Saez (2011) document that the U.S. incometail Pareto beta rose from 1.82 to 3.42 between 1976 and 2007, implying a shift in the corresponding alpha from 2.22 down to 1.41. As Toda and Walsh (2015) emphasize, analyses involving fat-tailed data require great care: when higherorder moments fail to exist, the Central Limit Theorem breaks down and conventional confidence intervalsfor instance, those on the top10 percent wealth sharelose validity. We are grateful to a referee for highlighting this issue.

A.1 Performance of the Estimator at the Tails

To investigate the first question, we begin with Figure 10, which compares two estimators: (1) the polynomial approximation as in the paper and (2) the empirical percentile function, shown with 99% interval bands. Each panel repeats the same percentile curve but varies its uppertail cutoff, using final grid points $p \in \{0.995, 0.999, 0.9995, 0.9999\}$. This addresses, in one panel, where (1) falls short, with particular focus on the tails, motivated by the footnote. As such, an inset for each panel is provided, zooming into the top 5% of the respective distribution. For wealth, we observe nearcoincidence of the two estimators up to about $p = 0.9995$. Only beyond the 99.95th percentile—the top onetwentieth of one percent—does the series estimate begin to pull away from the empirical percentile function (itself subject to sampling error). Hence, any inference inside the top 0.05 percent of the distribution should be treated with particular caution.[31]

Figure 10: Wealth Tails



Notes: Figure shows four panels plotting the percentile function for consumption, evaluated on a grid $G = \{0.01, \ldots, 0.99,\ p\}$. Consumption, as defined, is scaled by its respective aggregate and transformed via an inverse hyperbolic sine function. The four panels differ in the plotted estimate of the tail, denoted by p, which appears in the legend. The inset of each panel zooms in on the tail, covering the last five percentiles. No estimator is evaluated for $p = 1.0$.

31. Results on consumption and income are readily available and can be provided upon request.

A.2 Local vs. Global Estimation

To answer the second question, we introduce a *local* estimator that computes a set of coefficients within a moving kernel window. By sliding the window along a percentile grid G, we obtain $|G|$ coefficient vectors, each tailored to the local variation within that window. The local estimator is flexible and well-suited for capturing local structure.[32] By comparing the global estimators single set of coefficients with the collection of locally estimated coefficients, we can assess where and how they differ—especially around the tails of the distribution. Below, we describe the comparison procedure in detail and show that, for the economic data analyzed in this paper, our global estimator performs comparably to the local approach with minimal loss.

Data Let $w_i \in \mathbb{R}$ denote the economic variable of interest for household i as described in the main body of the text. The variable, as before, is first scaled by the corresponding nationalaccounts aggregate (quarterly, perhousehold level) and stabilized with an inverse hyperbolic sine transform. Results presented here rely on the 2019 wave of the PSID and are presented for the different measures consumption, income, and wealth. For this reason, we drop all time indices t and measure indices m.

Global Legendreseries estimator To have the appendix section self-contained, we provide again our estimator. Truncating the polynomial order O to 11, the shifted basis of orthogonal Legendre polynomials, evaluated on the empirical cumulative distribution u_i, is $Q_o(u_i)$ for $O = 0, \ldots, 11$. Explicitly, $u_i = \frac{s_i}{\sum_i s_i}$ for s_i the survey weight corresponding to household i after sorting measure m. Projecting w_i on the basis yields coefficients

$$\hat{\xi}_o := N^{-1} \sum_i w_i Q_o(u_i) \qquad \text{for } o = 0, \ldots, 11.$$

The estimated quantile function with these coefficients is

$$\hat{\Xi}^{-1}_{glob}(p) \;=\; \sum_{o=0}^{11} \hat{\xi}_o \, Q_o(p), \qquad p \in [0,1], \tag{27}$$

32. Naturally, the estimator will perform poorly outside the window, but we make clear that when evaluting the percentile function at a certain point, that we use the corresponding set of coefficients for that point i.e., the optimal coefficients in that window in a least squares sense.

where we use *glob* to denote that this is a global estimator and p some grid point in $[0,1]$ for which we can evaluate the percentile function $\hat{\Xi}^{-1}_{glob}(p)$.

Local kernel Legendre estimator For the local estimation, let $p_g \in G$ be some probability grid point and $G = \{0.01, 0.02, \ldots, 0.995\}$. Let $K_h(u)$ denote a Gaussian kernel with bandwidth h[33]

$$K_h(u) \;=\; \frac{1}{h\sqrt{2\pi}}\,\exp\Big\{-\tfrac{1}{2}\big(u/h\big)^2\Big\}.$$

For a given survey weight of the PSID, s_i, we can generate the local weight $s_i^{(g)} = s_i\,K_h(u_i - p_g)$. Let $W_g = \mathrm{diag}(s_1^{(g)}, \ldots, s_n^{(g)})$ and $\Phi_{io} = Q_o(u_i)$. Solving the weighted leastsquares system

$$(\Phi^\top W_g \Phi)\,\boldsymbol{\xi}^{local}(p_g) \;=\; \Phi^\top W_g \mathbf{y}$$

produces the *local* coefficient vector $\boldsymbol{\xi}^{local}(p_g) = (\xi_0(p_g), \ldots, \xi_{11}(p_g))^\top$, which returns anew for every p_g. The estimated *local* quantile function then uses, for each grid point evaluation, a new set of coefficients $\xi_o^{local}(p_g)$, generating

$$\hat{\Xi}^{-1}_{local}(p_g) = \sum_{o=0}^{11} \xi_o^{local}(p_g) Q_o(p_g).$$[34]

Thus, we have the two extremes: the global approach of one set of coefficients to estimate the percentile function and the local approach of many sets of coefficients—localized throughout the entire distribution—to estimate the percentile function.

To see how the coefficients change across the distribution, independently of overall scale, each coefficient vector (from the different estimators) is transformed to unit ℓ^1 length:

$$\tilde{\xi}_o^{local}(p_g) \;=\; \frac{|\xi_o^{local}(p_g)|}{\sum_{o=0}^{11}|\xi_o^{local}(p_g)|}, \qquad \tilde{\xi}_o^{glob} \;=\; \frac{|\widehat{\xi}_o^{glob}|}{\sum_{o=0}^{11}|\xi_o^{glob}(p_g)|}.$$

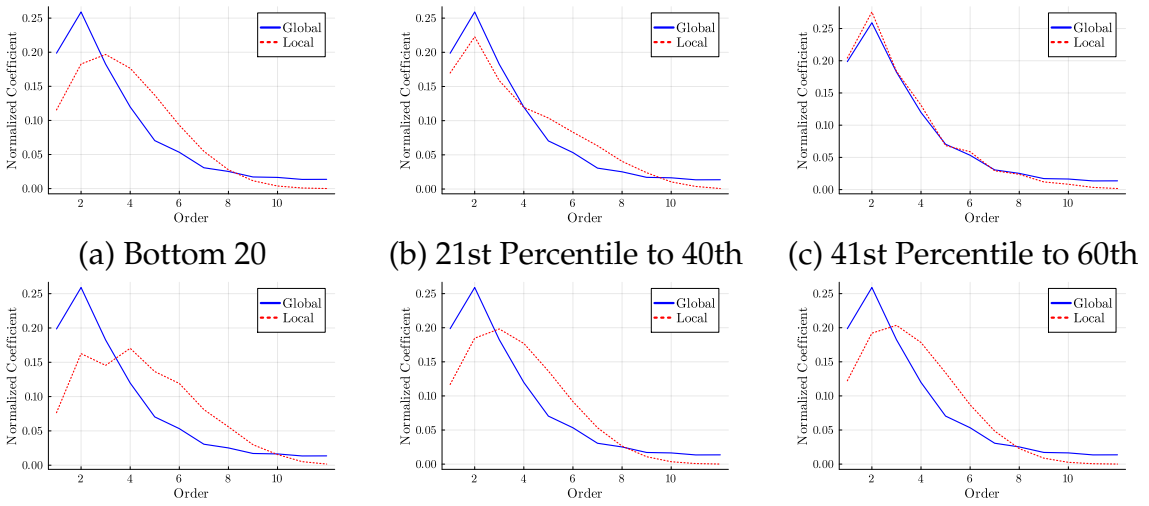
One can then average over the vector of coefficients (over intervals of G) to assess how these weights fluctuate for, for example, the bottom of the wealth distribution versus the top. Figure 11 shows precisely this (results are analogous for consumption and income). In short, our estimator assigns weights identically

33. Results below are unchanged for various smaller bandwidths $h \in \{.005, .01, .05, .10\}$

34. Alluding to our comment before, one can alternatively estimate $\hat{\Xi}^{-1}_{local}(p) = \sum_{o=0}^{11} \xi_o^{local}(p^*) Q_o(p)$, where $\xi_o^{local}(p^*)$ are the weights associated for a selected grid point p^* e.g., the 99th percentile.

to a local estimator that focuses on the 40th to 60th percentile. In every other case, it tends to assign more weight to the polynomials of order 4 to 8 in exchange for less weight on the first 3 polynomials. One might therefore conclude that we should instead use a different set of coefficients. Figure 12 provides evidence to the contrary.

Figure 11: Local vs. Global Estimator: Coefficients

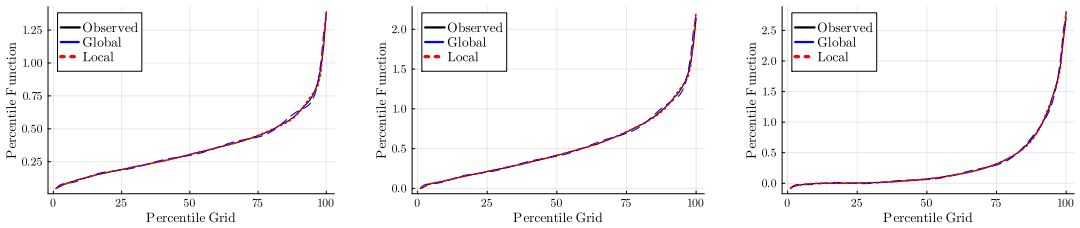


(a) Bottom 20 (b) 21st Percentile to 40th (c) 41st Percentile to 60th

(d) 61st Percentile to 80th (e) 81st Percentile to 100th (f) 91st Percentile to 100th

Notes: Figure shows panels comparing the global estimator to the local estimator for different segments of the wealth distribution. Each panel compares the contribution of each weight across the two estimators. Weights for the global estimator are the same across panels. Weights for the local estimator are calculated by averaging over the set of coefficients in that segment e.g., the Bottom 20 has 20 vectors of coefficients to average over.

Figure 12: Local vs. Global Estimator: Percentile Function



(a) Consumption (b) Income (c) Wealth

Notes: Figure shows three panels of percentile functions, comparing the global estimator to the local estimator, along with the empirical percentile function estimator. Percentile functions are evaluated over a set of grid points $G = \{0.01, \ldots, 0.99,\ 0.995\}$. Data is from the PSID 2019 wave.

A.3 Treatment of non-Differentiability of the Marginal

Some margins (e.g., earnings or financial assets) exhibit an atom at zero in addition to a continuous distribution over $\mathbb{R}_+$. Let Z_{mt} denote marginal m at time t and

$$\pi_{mt} := P_t(Z_{mt} = 0).$$

Write the CDF as a mixture

$$\Xi^{Z}_{mt}(z) = \pi_{mt}\mathbf{1}\{z \geq 0\} + (1 - \pi_{mt})\,\Xi^{Z,c}_{mt}(z), \qquad \Xi^{Z,c}_{mt}(z) := P_t(Z_{mt} \leq z \mid Z_{mt} > 0).$$

The unconditional quantile function is

$$\Xi^{Z,-1}_{mt}(u) = \begin{cases} 0, & u \leq \pi_{mt}, \\ \Xi^{Z,c,-1}_{mt}\Big(\dfrac{u - \pi_{mt}}{1 - \pi_{mt}}\Big), & u > \pi_{mt}, \end{cases} \tag{28}$$

i.e., the atom generates a flat segment of length π_{mt} and the positive part is rescaled to mass $1 - \pi_{mt}$.

We work with a strictly increasing transform $X_{mt} = T(Z_{mt})$ with $T(0) = 0$ (e.g. asinh), so (28) holds verbatim for X_{mt}; quantiles on the original scale follow as $\Xi^{Z,-1}_{mt}(u) = T^{-1}(\Xi^{X,-1}_{mt}(u))$.

Empirically, we estimate the mixed distribution in two steps: (i) estimate the point mass

$$\hat{\pi}_{mt} = \frac{\sum_i s_{it}\mathbf{1}\{Z_{mit} = 0\}}{\sum_i s_{it}},$$

and (ii) using only $Z_{mit} > 0$, estimate the conditional quantile function of $X_{mit} = T(Z_{mit})$ via the Legendre expansion

$$\Xi^{X,c,-1}_{mt}(u) \approx \sum_{o=0}^{O} \xi^{c,X}_{mot} Q_o(u).$$

We treat $\hat{\pi}_{mt}$ as an additional scalar coefficient alongside the Legendre coefficients and include it in the subsequent factor extraction and state-space analysis.

B Estimation of Factor Structures and Marginal Data Densities

A key input to our procedure is the mapping from factors to observables, defined by the projection matrix Γ. To estimate Γ, one method would be based on complete data alone. Complete here will mean there is no missing data in the spatial sense i.e., the data reports consumption, income, and wealth. For our case, this means using exclusively the PSID data to estimate Γ based on the PCA of $\tilde{\boldsymbol{\theta}}^{PSID}$. We consider different scenarios where we retain a different numbers of factors each time. Specifically, for this approach, we estimate the model with 3, 6, 7, and 8 distributional factors, obtained from the 12 PSID waves; however, this approach runs the risk of not fully representing the entire subspace that characterizes the evolution of the joint distributions because of the infrequent releases of the PSID data.

The literature has suggested alternatives for taking into account *incomplete* data in factor models. Here, this would mean surveys that only collect data on some, but not all marginals e.g., only income and wealth as in the SCF. In particular, we explore one alternative estimator for Γ: we consider the Tall-Wide algorithm of Bai and Ng (2021). We compare the quality of the model estimates under the alternative methods using marginal data densities for model comparison.

B.1 Tall-Wide Algorithm of Bai and Ng (2021)

An alternative approach to incorporating an additional block of data to the complete data is provided by Bai and Ng (2021). The paper tackles this problem by identifying two blocks of data within some larger T by N matrix— a tall block (data observed for all periods) and a wide block (time periods for which entire distribution is observed). For our setting, this means we add a tall block to our estimation of Γ, while the wide block would be what we call the complete data.[35] The tall block we add incorporates data on lower dimensional copulas of (income, consumption) and (income, wealth), as well as additional data on the marginals. This would be variation coming from both the CEX and SCF.

This alternative estimator for Γ, since it relies on more data, may improve model fit. For our application, we extract the factors that explain 80% of the

35. It is important to note that, given some data, such an estimator can consistently estimate the common component without making any assumptions on the nature of missingness. We refer the reader to the paper for more details on how the exact procedure is performed.

Table 5: Model Comparison

Model	Harmonic Mean	Bridge
	TW Projection Matrix	
12 distributional factors, 11 agg. factors	-81571.59	-85685.21
10 distributional factors, 11 agg. factors	-80555.15	-83940.35
	Standard Projection Matrix	
3 distributional factors, 3 agg. factors	-84379.20	-86262.30
6 distributional factors, 11 agg. factors	-72985.78	-77007.84
7 distributional factors, 11 agg. factors	-63227.01	-69535.44
8 distributional factors, 3 agg. factors	-54253.22	-56898.68
8 distributional factors, 11 agg. factors	-54137.45	-60434.82
8 distributional factors, 15 agg. factors	-54220.29	-61024.63

Notes: The table reports the log of the marginal data densities (MDD) across different model specifications. The MDD is estimated using two estimators: (1) Harmonic Mean and (2) a Bridge sampler. TW Projection Matrix are models estimated using the Tall-Wide algorithm of Bai and Ng (2021). Standard Projection Matrix is from a PCA estimator. Higher values means most efficient.

summable variation, as well as 85%. This corresponds to 10 and 12 factors; however, as mentioned before, it is unclear the degree of variation shared among this set of factors relative to the original set of factors. Furthermore, keeping more factors from this estimator runs the risk of over-parameterizing the model. Table 5 presents the marginal likelihoods of these *TW* projection matrices. Results suggest these models have marginally inferior fit than models that only rely on complete data.

B.2 Marginal Data Densities and Optimal Factor Structures

We estimate various versions of the state space model for different factor loadings Γ, that differ both in the number of factors and the estimation of Γ itself, and different numbers of aggregate factors. For each estimated model, we calculate the marginal data density (MDD) in order to discriminate between these alternatives. The set of models considered rely on the same data, but differ in the size of the parameter space. The MDD will effectively internalize these two features, selecting the model for which we can expect the best forecasting performance, while penalizing models through a lower density due to larger parameter spaces.

The marginal data densities $p(\tilde{\tilde{\boldsymbol{\theta}}})$ are estimated using Geweke's modified har-

monic mean estimator. Although standard, there may be concerns that, given the dimensionality of the model, such an estimator may not be appropriately approximated by a Gaussian distribution. Second, the estimator may not be numerically stable, as it requires the inversion of matrices—in this case, large ones. In this sense, we also estimate the density using a bridge sampler (Meng and Wong 1996), which is numerically stable (does not require any inversions) and may better internalize the shape of the posterior.

Table 5 covers the span of proposed models that were ultimately assessed, as well as the marginal data densities over the different estimators. First, we find that the MDD is most elastic to the number of distributional factors, as opposed to the number of aggregate factors. The model with the projection matrix from the complete PSID data, tagged *Standard Projector Matrix*, using the maximum number of 8 factors achieves the highest MDD for models with 11 aggregate factors.[36] Using the factor loadings Γ from the Bai and Ng (2021) TW-algorithm also achieves a lower MDD and thus a worse model fit.

C Minnesota Prior

Given the size of the model, we estimate it in a Bayesian framework and regularize the state equation parameters $(A, B, C, \text{diag}(D), \Omega)$ using a Minnesota prior.

Stacking the coefficients as $\theta := (\text{vec}(A)', \text{vec}(B)', \text{vec}(C)', \text{vec}(D)')'$ and setting

$$\theta \sim \mathcal{N}(\mu_{\text{Minn}}, V_{\text{Minn}}),$$

with mean

$$A_{ij} = \begin{cases} \kappa_2, & i = j \text{ (own first lag of distributional factors)}, \\ 0, & i \neq j, \end{cases}$$

$$B = \mathbf{0}, \quad C = \mathbf{0}, \quad D_{ii} = \kappa_3,$$

and diagonal variance

$$V_{\text{Minn},ii} = \begin{cases} \kappa_0 / l^2, & \text{own lags}, \\ (\kappa_0 \kappa_1 / l^2) \cdot (\hat{\sigma}_{ii}^2 / \hat{\sigma}_{jj}^2), & \text{cross-lags}. \end{cases}$$

36. Decreasing the number of aggregate factors increases the MDD, but these models are unfortunately incomparable as they have different measurement vectors. For example, the model with 15 aggregate factors has $10 \times T$ fewer contributions to the likelihood vs. the model with 25 aggregate factors.

we have our Minnesota prior. Here κ_2 and κ_3 control the prior persistence of the distributional and aggregate laws of motion. We work with unconstrained persistence parameters and map them to $(-0.99, 0.99)$ via a scaled inverse-logit; the unconstrained hyperpriors are diffuse, $\kappa_2, \kappa_3 \sim \mathcal{N}(0, 5)$. We fix the overall tightness at $\kappa_0 = 0.05$ motivated by Giannone, Lenza, and Primiceri (2015); κ_1 governs cross-lag shrinkage and is given a wide Normal hyperprior on its unconstrained representation, inducing an approximately uniform prior over $[0.2, 0.99]$ after transformation.

Prior for Ω. For the innovation covariance Ω, we place weakly informative priors on the diagonal elements (with variances governed by κ_4 and κ_5) and an LKJ prior on the correlation matrix. We parameterize the LKJ shape as

$$\eta = 1 + \log(\exp(\kappa_6) + 1),$$

so that $\eta \geq 1$, with $\kappa_4, \kappa_5 \sim \mathcal{N}(0, 1)$ and $\kappa_6 \sim \mathcal{N}(2, 1.5)$. Larger η shrinks correlations toward zero unless strongly supported by the data. We estimate $\{\kappa_i\}_{i=1}^{6}$ jointly with $\boldsymbol{\psi}_{par}$ (except κ_0, fixed at 0.05).

D Details on MCMC

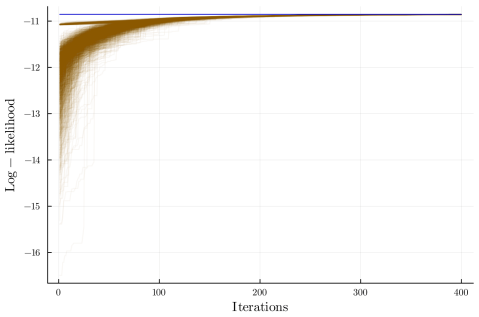
To estimate and sample from the posterior distribution, we employ the DIME sampler from Boehl (2024). The sampler is particularly advantageous for potentially complex, high-dimensional posterior distributions with ex-ante unknown properties.

We run an ensemble of $4n$ chains (for n the size of the parameter vector). To speed up adaptation, 10% of chains are initialized at a tentative mode from a separate optimization procedure; the remainder are initialized from the priors. The ensemble runs for 400 iterations, and we keep the last 25% of draws as posterior samples.[37] The sampler uses a single tuning parameter χ that governs the mixture between the local and global transition kernel; we set $\chi = 0.1$.

Figure 13 shows the traces of the (scaled) log-likelihood for all chains and indicates convergence.[38] Additional diagnostics are reported in the figure notes.

37. For the baseline $|\boldsymbol{\psi}| \approx 400$, this implies $4(400) \times 400 = 640{,}000$ *efficient* runs.
38. Traces of all estimated models can be provided upon request.

Figure 13: Converging Chains



Notes: Figure shows the evolution of the ensemble of chains in terms of (scaled) log-likelihood; the last 25% of draws of each chain are retained. Chains initialized at the tentative mode start at higher likelihoods and generate the "cliff" at initialization. The sampler also reports the log-weight on the history of the proposal distribution and the standard deviation of likelihoods. Early on, log-weights are positive, indicating adaptation; after convergence they are close to zero (around 10^{-7}). Standard deviations are about 1% of the mode log-likelihood. Acceptance rates are in the 20%–40% range, consistent with effective exploration. Refer to the text for model specifications.

E Reconstructing Distributional Data

Following the procedure below, one can obtain the high-frequency distributional data. Given the estimated coefficients $\left(\hat{\xi}_{mot}, \hat{\kappa}_{o_1 \cdots o_d, t}\right)$ and the associated polynomials, the user only needs to specify $\mathbf{u} \in [0,1]^d$, the domain of integration. In our application, we integrate to construct deciles. In practice, the integration bounds exclude the exact endpoints 0 and 1, which are replaced by 10^{-6} and 0.9995, respectively. This is in accordance with our analysis in Appendix A and properties of our estimator.

Procedure:

$$\hat{\tilde{\theta}}_t = \hat{\Gamma}^{MF}\, \hat{F}_t \qquad \text{(Project factors)}$$

$$\hat{\theta}^j_{nt} = \sigma^j_{\zeta(n)} \times \hat{\tilde{\theta}}_{\zeta(n)t} + \mu^j_n + g(t)_n \qquad \text{(Unstandardize and add trend)}$$

$$\hat{\theta}^j_t = (\hat{\xi}^j_{m,o_1,t}, \ldots, \hat{\kappa}^j_{(o_1,\ldots,o_d),t}) \qquad \text{(Decomposition)}$$

$$\hat{\Xi}^{-1}_{jmt}(u_m) = \sum_{o=0}^{O} \hat{\xi}_{mot}\, Q_o(u_m) \qquad \text{(Reconstruct quantile function)}$$

$$d\hat{C}_{jt}(u_1, \ldots, u_d) = \sum_{o_1,\ldots,o_d=0}^{O} \hat{\kappa}_{o_1 \cdots o_d, t} \prod_{m=1}^{d} Q_{o_m}(u_m) \qquad \text{(Reconstruct copula density)}$$

Table 6: Description of Equation Components

Symbol	Description
$\hat{F}_t$	Estimated high-frequency factors
$\hat{\Gamma}^{MF}$	Factor loadings mapping factors to distributional observables
$\hat{\tilde{\theta}}_t$	Projected (standardized) factor representation
$\sigma^j_{\zeta(n)}$	Standard deviation term, varies by object
μ^j_n	Mean adjustment for coefficient n
$g(t)_n$	Non-parametric trend at time t
$\hat{\theta}^j_{nt}$	Unstandardized, trend-adjusted coefficients
$\hat{\xi}^j_{m,o,t}$	Marginal polynomial coefficients
$\hat{\kappa}^j_{(o_1,\ldots,o_d),t}$	Copula polynomial coefficients
$Q_o(u_m)$	Polynomial basis function of order o, O=11
$\hat{\Xi}^{-1}_{jmt}(u_m)$	Reconstructed quantile function (inverse CDF) for margin m
$d\hat{C}_{jt}(u_1,\ldots,u_d)$	Reconstructed copula density (dependence structure)

Notes: Table summarizes the role of each symbol in the above equations, moving from factor projections to reconstruction of marginal quantiles and copula density.

F Validation of Hyperparameter Choices

To evaluate our hyperparameter choices on the Bayesian estimation priors for the measurement error variances, we compare the series resulting from the Kalman smoother after estimation with the actual point estimates and their confidence bounds from the survey data.

Intuitively, if the prior mean for the measurement error variance is too low, it will force the estimator to exactly match each survey estimate of the distribution, despite the fact that each survey estimate is itself subject to measurement error. Thus, we should expect the smoother estimate to fall within the confidence bounds of each sample estimate at most with the corresponding confidence level of the bounds. The fact that the confidence level is an upper bound reflects that the estimated measurement error captures not only the sampling uncertainty that the confidence bounds capture, but also conceptual differences.

Choosing narrow measurement errors would overstate precision and potentially limit comovement with aggregates. As a result, it would drive parameter estimates for B, which captures this comovement, toward zero. Another reason not to be conservative with the measurement errors is that allowing for measurement error also accounts for the fact that we combine data from different sources

to produce a consensus estimate. These different data sources, despite their individual detrending, may produce some temporarily divergent estimates of the distributions. Without sufficient measurement error, the consensus estimate is then forced to oscillate between these different distribution estimates over short time intervals, rather than capturing their co-movements.

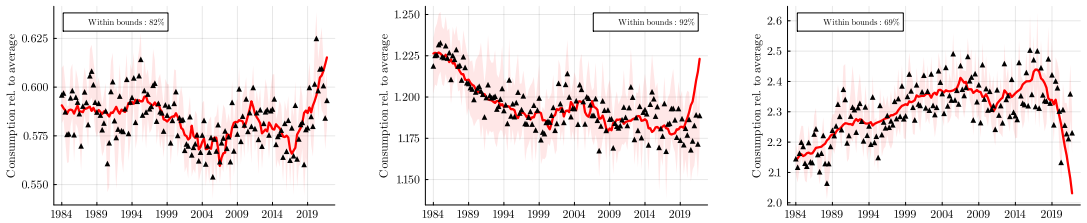
On the other hand, if the prior mean for the measurement error variance is too high, the estimator will treat the data as uninformative, and the smoother will miss the survey estimates more often and to a much greater extent than implied by its confidence bounds. We validate the choice of hyperparameters graphically for income and wealth in the SCF data and provide comprehensive summary statistics across all datasets and estimates. This visual inspection helps confirm that our smoothed series respect the noise inherent in the available microdata.

Figure 14 shows average consumption (top row), income (middle row) and wealth (bottom row) for the bottom 50 percent (first column), the next 40 percent (second column), and the top 10 percent (last column) of the respective distributions. It shows the point estimates from the surveys (black triangles), along with their 95% confidence limits, and the results from the Kalman smoother based on our estimates of the parameters of Equation (20). Overall, the smoothed estimates fall inside their respective confidence bounds in 85 out of the 108 observations (from (c) to (h)).

Table 7 provides a comprehensive summary of this validation approach. For all quantile functions and copulas, we report for each dataset and survey year how often the respective smoother estimate is within the confidence limits. Again, we use a confidence level of 95%. For the quantile functions, we find overall a modest difference (one to three percent) between the confidence level and the fraction of smoothed estimates that fall outside the confidence bounds. Only for the SIPP and to some extent the CPS, our estimator suggests a significant measurement error beyond sampling uncertainty reflecting differences in sample design and income/wealth measurement.

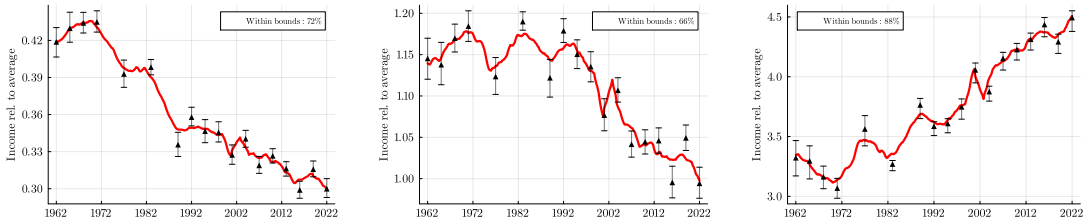
Figure 14: Comparison of Smoothed Distributional Data and Direct Survey Estimates

Consumption by Consumption



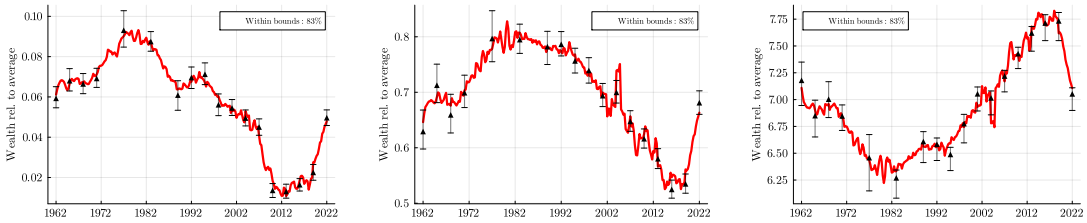
(a) Bottom 50 Percent (b) 50-90 Percent (c) Top 10 Percent

Income by Income



(c) Bottom 50 Percent (d) 50-90 Percent (e) Top 10 Percent

Wealth by Wealth



(f) Bottom 50 Percent (g) 50-90 Percent (h) Top 10 Percent

Model Data

Notes: Figure shows the average consumption, income, and wealth for the bottom 50 percent, 50-90 percent, and top 10 percent of households of the respective distribution. Dots show the estimates from the individual survey waves together with 95% confidence bounds. The solid red line shows the baseline estimate from the Kalman smoother at the posterior mode. Consumption shows CEX data and reconstruction. Income and wealth show SCF data and reconstruction. The legend reports for each Panel the share of smoothed estimates within the confidence bounds of the survey waves.

Table 7: Deviations of Smoothed Estimates and Microdata: Fraction within Confidence Bounds

Measure	CEX	CPS	SCF	SIPP	PSID	Overall
Consumption quantiles	77%	—%	—%	—%	100%	79%
Income quantiles	89%	93%	71%	46%	98%	77%
Wealth quantiles	—%	—%	77%	54%	99%	64%
Copula densities	98%	—%	94%	91%	98%	96%

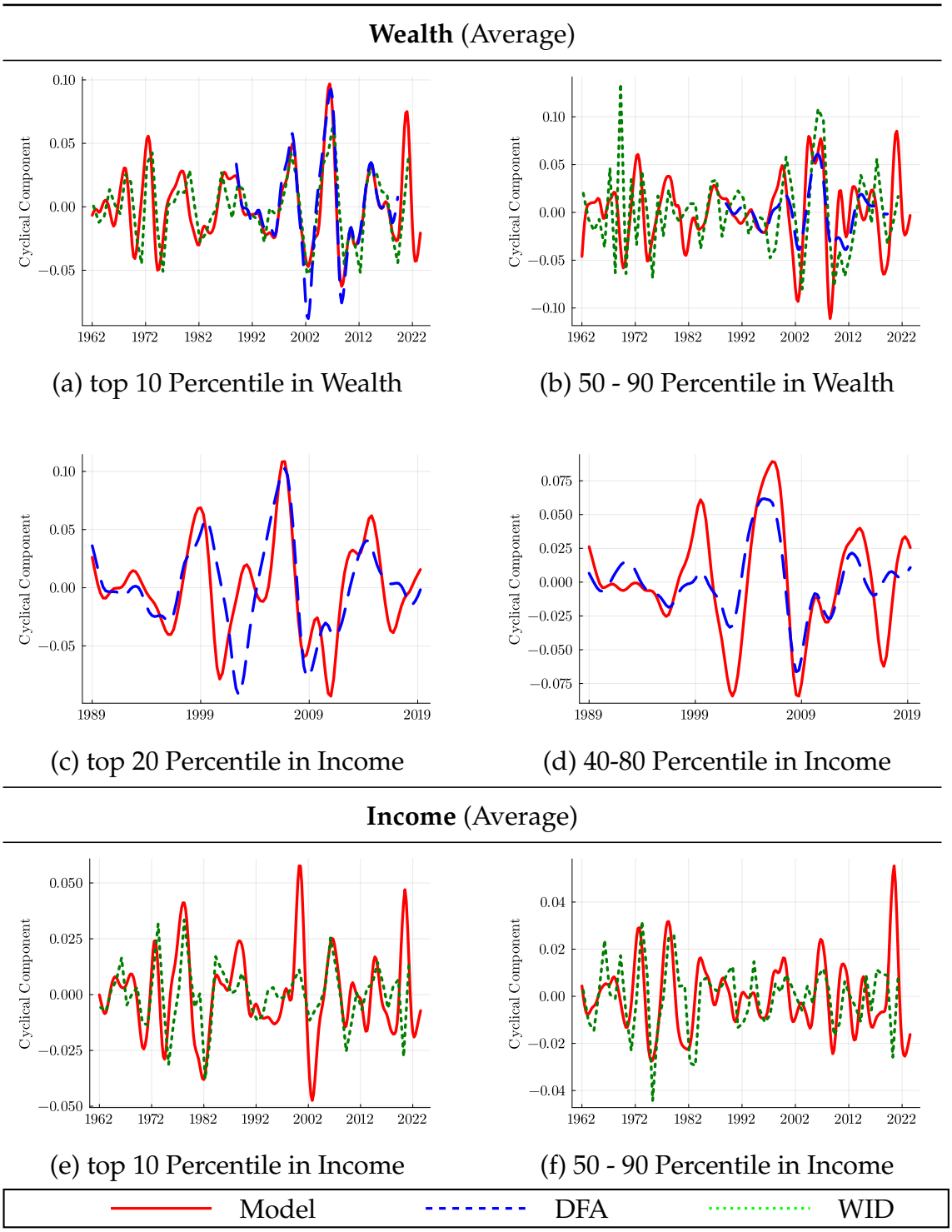
Notes: The table reports, by microdata and object, the fraction of estimates from the Kalman smoother at the posterior mode that fall within the 95% bootstrapped confidence intervals for the respective microdata. Quantile and copula estimates are defined on a decile grid.

G Comparison with External Estimates

The preceding sections established that our method predicts well out of sample and that the estimated distributions lie within sampling variability. We now assess whether the resulting high-frequency income and wealth dynamics align with external benchmarks. Since true distributional dynamics are unobservable, we compare our cyclical estimates to the *Distributional Financial Accounts (DFA)* (Batty et al. 2020) and the *World Inequality Database (WID)* (Piketty, Saez, and Zucman 2018). The DFA produces quarterly wealth distributions from the SCF using a different estimation approach, providing an "SCF flavor" benchmark. The WID provides annual estimates from administrative tax data without a time-series model, offering a robustness check for our assumption of stable and approximately linear distributional dynamics and their link to macro factors. To assess dependence dynamics (copula), we additionally compare our implied wealth-by-income series to the DFA.

Figure 15 reports the cyclical components (in logs). Panels (a)–(b) compare wealth by wealth group, focusing on the top 10% and the next 40% since the bottom half has wealth close to zero (Kuhn and Ríos-Rull 2016). Panels (c)–(d) compare wealth by income groups (informative about the joint distribution). Panels (e)–(f) compare income by income group (available in the WID). Across all cases, our estimates comove closely with DFA and WID and fall within the range of the external series.

Figure 15: Cyclical Component of Distributional Data vs. External Sources



Notes: Figure shows the cyclical component of (log) average wealth of (a) the wealthiest 10 percent, (b) the next wealthiest 40 percent, (c) the 20 percent income-richest households, and (d) the next 40 percent income-richest households. Bottom row shows the cyclical component of (log) average income of (e) the income-richest 10 percent and (f) the next income-richest 40 percent. Red lines show cyclical components from baseline model at quarterly frequency. Dotted green line show annual data from the *World Inequality Database* (WID). Dashed blue lines show quarterly data from the *Distributional Financial Accounts* (DFA). Cyclical components are obtained by an HP-filter with smoothing parameter $\lambda = 6$ for annual data and $\lambda = 1600$ for quarterly data.

H Data

The construction of these estimates relies on a great deal of data. An advantage with our method, however, is that it can incorporate these different microdata and their various differences in generating consensus estimates of the distributional data. Below, we describe the data, all expressed in 2019 dollars, and explain the mappings across data to ensure measures are at some base comparability (Curtin, Juster, and Morgan 1989; Czajka, Jacobson, and Cody 2003; Pfeffer et al. 2016). See Table 1 for information on their availability.

H.1 SIPP

The SIPP panel is a nationally representative, individual-level survey known for providing high-frequency dynamics on employment, earnings, wealth, household composition and program participation in the U.S.. For the data cleaning, the data is aggregated to the household-level, at quarterly frequency. Due to structural breaks induced by changing survey designs, we treat the SIPP as if it were three separate surveys (until 1983Q3-1995Q4, 1996Q1-2012Q4, 2013Q1-2022Q4). This is to minimize spurious fluctuations (e.g., seam bias) that obscure actual economic phenomena.

Income. For the 2014 releases and onward, we use the `THTOTINC` variable for income. For data releases prior, we sum over (1) earnings (`ws1_am, ws1_am`) (2) property/investment income (`tpprpinc`) (3) unemployment (`tuc1amt, tuc2amt, tuc3amt`) and (4) transfers (`tptrninc, tpscininc, twicamt, tfs_am, tssi_amt`) to construct household income.

Wealth. For the 2014 releases and onward, we use the `THNETWORTH` variable for wealth. For data releases prior, wealth is defined as total assets (`hhtwlth`) net total liabilities (`hhusdbt, hhscdbt`).
Note: The SIPP has undergone several major redesigns, and ignoring these seams, even with our prescribed data treatment, would lead to spurious swings in measured inequality—an issue documented by, among others, Moore (2008) and Czajka, Jacobson, and Cody (2003) and confirmed in our own diagnostics. To deal with this issue, we treat the SIPP as if it were three different surveys: (1) SIPP before 1996 (2) SIPP 1996-2012 and (3) SIPP 2013-2023. We found that around these seams, estimates tend to exhibit large jumps, which was at odds with their aggregate counterpart and general business cycle conditions. Ad-

ditional observations were dropped as well ($2003Q4, 1999Q4, 1988Q1, 1985Q4$), since they exhibited odd movements as well, related to the seams noted, as well as older seams. Fortunately, the estimation framework allows each of the three SIPP subsurveys to carry its own measurementerror process, so we retain almost all available information while preventing seam effects from distorting the results.

H.2 SCF+

The Survey of Consumer Finances (SCF), since its inception in 1983, is seen as the data gold mine for household information on income and wealth; however, due to the research excavations of Kuhn, Schularick, and Steins (2020), we are able to combine these triennial cross-sections with historical waves of the SCF; hence the name SCF+. Kuhn, Schularick, and Steins (2020) mention "... the SCF+ is the first dataset that makes it possible to study the joint distributions of income and wealth over the long run". Thus, it goes without saying how requisite this is for our study. The SCF+ is also augmented with the Forbes 400, from Forbes, for the years 2021 to 2024. For the years 1985 to 2020, we use the per capita dataset of Fernholz and Hagler (2023), whose observations originate from families of the Forbes 400.[39] Below we describe the measurements we used from the SCF+.

Income. Our definition of income follows Kuhn, Schularick, and Steins (2020), which consists of the following components: (1) labor income (i.e., earnings) (2) income from public transfers (3) income from professional practice and self-employment (4) income from rents (5) dividend income and (6) business/farm income. A different taxonomy that illustrates these components are taxable and transfer income.

Assets. Total assets include (1) liquid assets such as a household's checking and savings account, CDs, call/money market accounts, short-term government bonds, and mutual funds (2) illiquid assets such as housing and other real estate minus debt on that properties respectively, automobiles (3) defined-contribution retirement plans (4) the cash value of life insurance (5) stocks and (6) business equity.

39. Accessing the data via the usual link (e.g., `http://www.forbes.com/ajax/list/data?year=2005&uri=forbes-400&type=person`) returns Forbes 400 data, but unfortunately incomplete from the 1990s until the late 2000s—with only around 200 of the 400 observations there. Using Fernholz and Hagler (2023), with the same procedure imposed by Bricker, Hansen, and Volz (2019), generates the same result as in Bricker, Hansen, and Volz (2019) (increases the top wealth share by 1.5%), but is more complete (has almost all 400 every year).

Debt. We define debt of a household as the sum of personal (mostly unsecured) debt and housing (mortgage) debt. Housing debt includes debt from all properties and any loans made against the housing e.g., through HELOCs. Personal debt includes car loans, education loans, any loans from relatives, credit card debt, medical debt and legal debt.

Wealth. Wealth is total assets net total debt of a household.

H.3 PSID

The Panel Study of Income Dynamics complements the SCF+ extraordinarily well, as they take our estimations beyond more than half a century. In comparison to the post-1983 SCF, a deeper analysis of their similarity can be found in Pfeffer et al. (2016).

Income. The PSID has collected family income annually from 1968 to 1996 and then biennially from 1997 to 2021. Its measure of income is the sum of taxable income, transfers, and social security for the reference person, the spouse/partner (if any) and other members of the family.[40]

Assets. Data collection on household wealth took place in 1984, 1989, 1994, and then every wave beginning in 1999. The data on assets is split into liquid and illiquid assets. Although minor, the definition of liquid assets will vary between datasets, so careful attention is warranted here. Liquid assets for the PSID include checking and savings accounts, short-term instruments such as money-market accounts, certificates of deposit, and treasury bills. Illiquid assets include business equity, financial assets held in mutual funds, stocks, bond funds, investment funds; real assets held in real estate, vehicles like motor homes, boats, trailers, and cars; and retirement wealth in private annuities or IRAs.

Debt. For the PSID, we achieve the same debt split: personal and mortgage debt. This includes all kinds of real-estate debt, and unsecured debt such as credit card debt, student loans, medical debt, legal debt, and loans from relatives.

Wealth. Wealth is total assets net total debt of a household.

Consumption. Studying papers such as Skinner (1987), Cutler et al. (1991), Flavin

40. In the PSID, a family is a group of people living together who are economically interdependent.

and Yamashita (2002), Attanasio, Hurst, and Pistaferri (2014), and Attanasio and Pistaferri (2014), we define consumption as the sum of these expenditures: food, rent (for renters), housing rental equivalence (for home-owners), utilities, health, public transport, education, and childcare. We set the housing rental equivalence to be 6% of the home market value reported by households in the PSID. Consumption data is only available from 1999 in a biennial interval.

H.4 CPS

We use the Community Population Survey (CPS) Annual Social and Economic Supplement (ASEC). The sample is designed primarily to produce estimates of the labor force characteristics and runs from 1962 to 2022. Similar to the SIPP, we treat the CPS as a combination of two separate surveys to avoid fluctuations driven by survey design: 1967Q4-1993Q4 and 1994Q4-2021Q4.

Income. Income data are collected as part of the ASEC for the months of February, March and April as a supplement to the regular CPS monthly labor force interviews. The ASEC asks each person in the sample who is 15 years old and over about the amount of income received from a list of sources in the previous calendar year. We treat these observations as being observed in quarter four of the previous calendar year. For details on top-coding, see https://cps.ipums.org/cps/topcodes_tables.shtml.

H.5 CEX

The Consumption Expenditure Survey (CEX) is the most comprehensive household survey in the U.S. for recording the consumption habits of households. The CEX has two components: the interview survey (IS) and the diary survey (DS). The interview survey has sufficiently rich data on what we need, so we only use data from this component. Within this component, there are several files, each of which pertain to a topic, from which we can extract information. The following table breaks down each category of consumption, defining which `UCCs` belong to which category and which file it can be found in. All of these categories will combine to make the consumption variable. The table will also define wealth concepts of the CEX we use in our study. Since each household consumption record is with respect to a `UCC`, we find this presentation most apropos.

Item	UCCs / FMLI label	File
	Consumption	
Food	190904, 790220, 190901, 190902, 190903, 790410, 790430, 200900, 790330, 790420, 800700, 790230, 790240	MTBI
Rent	210110, 800710	MTBI
Utilities	250111, 250112, 250113, 250114, 250211, 250212, 250213, 250214, 250221, 250222, 250223, 250224, 250901, 250902, 250903, 250904, 250911, 250912, 250913, 250914, 260111, 260112, 260113, 260114, 260211, 260212, 260213, 260214, 270211, 270212, 270213, 270214, 270310, 270411, 270412, 270413, 270414,270101, 270102, 270104, 270105, 270310, 270311, 690116, 270901, 270902, 270903, 270904	MTBI
Health	570110, 570111, 570210, 570220, 570230, 560110, 560210, 560310, 560330, 560400, 340906, 540000, 550110, 550320, 550330, 550340, 570901, 570903, 570240, 580111, 580112, 580113, 580114, 580311, 580312, 580901, 580903, 580904, 580905, 580906, 580400, 580907	MTBI
Public Transport	520531, 520532, 530311, 530312, 530501, 530902, 530210, 530411, 530412, 520511, 520512, 520521, 520522, 520542, 520902, 520903, 520904, 520905, 520906, 520907, 530110, 530901, 520110, 520310	MTBI
Education	210310, 370903, 390901, 660110, 660210, 660310, 660900, 670110, 670210, 670901, 670902, 800802, 800804, 690111, 690112, 660410, 660902, 670410, 670903, 690114, 690310	MTBI
Child care	340210, 340211, 340212, 670310, 660901	MTBI
Rental Equivalence	910050, 800721 (market value of home), SIMHOUSX, RENTEQVX	FMLI, MTBI

Gas & Vehicle Repairs	470111, 470112, 470113, 470220, 470211, 470212, 480110, 480212, 480213, 480214, 490110, 490211, 490212, 490221, 490231, 490232, 490311, 490312, 490313, 490314, 490318, 490319, 490411, 490412, 490413, 490501, 490502, 490900, 520410, 480215, 620113	MTBI
	Other Concepts	
Housing Debt	QBLNCM1X, QBLNCM2X, QBLNCM3X, QBLNCM1G, PRINAMTX	MOR
Personal Debt	6001, 6002 (1990–2013), 5400, 5500, 5600, CREDITX, STUDNTX, OTHLONX, CREDITX1, CREDITX5, QBALNM1X	MTBI, ITBI, FMLI, FN2
Liquid Assets	SAVACCTX, CKBKACTX, USBNDX, 920010, 920020, 920030, 5100, LIQUIDX	FMLI, ITBI
Financial Assets	5800, 920040, STOCKX, SECESTX, OTHASTX	FMLI, ITBI
Income	FINCBTAX	FMLI

Notes: Table shows, by item, the identifiers necessary to construct each component of consumption, income and wealth for the CEX. The location of these identifiers can be found under the *File* column.

H.6 Aggregates

Together with the microdata, we specify a model component that captures the various aggregate shocks that potentially buffet the joint distribution of consumption, income, and wealth. This is represented in the state equation of the state-space model. The aggregate data we rely on to extract this information comes from the FRED-QD (McCracken and Ng 2021). This has various macrodata on industrial production, employment, housing, inventories, prices, earnings, productivity, household expectations, household balance sheets, interest rates, credit, etc. You can find more information on https://research.stlouisfed.org/econ/mccracken/fred-databases/.

Before performing the PCA on the aggregates, we are careful to check each series for non-stationarity. Recent literature has placed emphasis on the identifiability of orthogonal factors in high-dimensional settings, in particular for

macroeconomic aggregates, and finds non-stationarity to be the culprit of spurious variation (Onatski and Wang 2021; Hamilton and Xi 2022). Running the PCA on the non-stationary data will erroneously find that a large set of aggregates is confined to just a few factors. Taking note, we first remove any variation due to seasonality using the X13-ARIMA and closely follow the transformations (to induce stationarity) proposed by McCracken and Ng (2021). The resulting series satisfy an Augmented-Dickey-Fuller Test with a significance level of $\alpha = 0.05$ and are visually inspected for abnormalities.

The set of now stationary aggregates are concatenated with four of its lags to form a data matrix of quintuple the size and then column-wise standardized. A PCA on this block of data is performed and 11 orthogonal factors are kept. The number of factors chosen is based on Freyaldenhoven (2022). The baseline model estimation includes these 11 factors as inputs Y_t. More on the selection of factors is available upon request.

I Out of Sample Correlations

Table 9: Out of Sample Performance

Condition	Bottom			Middle			Top		
	C	I	W	C	I	W	C	I	W
Excluding Housing Cycle Wealth									
Entire Series	-	0.98	0.96	-	0.98	0.96	-	0.98	0.95
Specific Timeframe	-	0.98	0.94	-	0.96	0.97	-	0.96	0.97
Excluding Last 4 Years									
Entire Series	-	0.95	0.84	-	0.96	0.88	-	0.97	0.85
Specific Timeframe	-	0.65	-0.9	-	0.92	0.00	-	0.90	-0.34
Every 4 Years									
Entire Series	0.93	0.98	-	0.97	0.96	-	0.98	0.97	-
Specific Timeframe	0.94	0.99	-	0.98	0.96	-	0.95	0.97	-

Notes: Table reports correlations between the baseline cyclical estimates and the missing-data models, split by panel. Entire Series reports correlations for all estimates in the estimation timeframe. Specific Timeframe reports correlations of estimates in periods where data was intentionally left out of the estimation of the specific missing-data model. **Bottom**, **Middle**, and **Top** are the bottom 50, next 40, and top 10 of the respective distribution, denoted by **C, I,** and **W**. **C** is for consumption, **I** is for income, and **W** is for wealth. For the first three panels, correlations are made between SCF model estimates (no consumption). The final panel presents correlations from CEX model estimates (no wealth). Specific models (in header) are discussed in Section 4.

Appendix References

Atkinson, Anthony B, Thomas Piketty, and Emmanuel Saez. 2011. "Top incomes in the long run of history." *Journal of economic literature* 49 (1): 3–71.

Attanasio, Orazio, Erik Hurst, and Luigi Pistaferri. 2014. "The evolution of income, consumption, and leisure inequality in the United States, 1980–2010." In *Improving the measurement of consumer expenditures,* 100–140. University of Chicago Press.

Attanasio, Orazio, and Luigi Pistaferri. 2014. "Consumption inequality over the last half century: some evidence using the new PSID consumption measure." *American Economic Review* 104 (5): 122–126.

Bai, Jushan, and Serena Ng. 2021. "Matrix completion, counterfactuals, and factor analysis of missing data." *Journal of the American Statistical Association* 116 (536): 1746–1763.

Bricker, Jesse, Peter Hansen, and Alice Henriques Volz. 2019. "Wealth concentration in the US after augmenting the upper tail of the survey of consumer finances." *Economics Letters* 184:108659.

Curtin, Richard T, Thomas Juster, and James N Morgan. 1989. "Survey estimates of wealth: An assessment of quality." In *The measurement of saving, investment, and wealth,* 473–552. University of Chicago Press, 1989.

Cutler, David M, Lawrence F Katz, David Card, and Robert E Hall. 1991. "Macroeconomic performance and the disadvantaged." *Brookings papers on economic activity* 1991 (2): 1–74.

Czajka, John L, Jonathan E Jacobson, and Scott Cody. 2003. "Survey estimates of wealth: A comparative analysis and review of the Survey of Income and Program Participation." *Soc. Sec. Bull.* 65:63.

Fernholz, Ricardo T, and Kara Hagler. 2023. "Rising inequality and declining mobility in the Forbes 400." *Economics Letters* 230:111235.

Flavin, Marjorie, and Takashi Yamashita. 2002. "Owner-occupied housing and the composition of the household portfolio." *American Economic Review* 92 (1): 345–362.

Freyaldenhoven, Simon. 2022. "Factor models with local factorsdetermining the number of relevant factors." *Journal of Econometrics* 229 (1): 80–102.

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri. 2015. "Prior selection for vector autoregressions." *Review of Economics and Statistics* 97 (2): 436–451.

Hamilton, James D, and Jin Xi. 2022. "Principal Component Analysis for Nonstationary Series."

Kuhn, Moritz, and José-Víctor Ríos-Rull. 2016. "2013 Update on the US earnings, income, and wealth distributional facts: A View from Macroeconomics." *Federal Reserve Bank of Minneapolis Quarterly Review* 37 (1): 2–73.

Meng, Xiao-Li, and Wing Hung Wong. 1996. "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration." *Statistica Sinica*, 831–860.

Moore, Jeffrey C. 2008. "Seam bias in the 2004 SIPP panel: Much improved, but much bias still remains." *US Census Bureau Statistical Research Division Survey Methodology Research Report Series* 3:2008.

Onatski, Alexei, and Chen Wang. 2021. "Spurious factor analysis." *Econometrica* 89 (2): 591–614.

Skinner, Jonathan. 1987. "A superior measure of consumption from the panel study of income dynamics." *Economics Letters* 23 (2): 213–216.

Toda, Alexis Akira, and Kieran Walsh. 2015. "The double power law in consumption and implications for testing Euler equations." *Journal of Political Economy* 123 (5): 1177–1200.