Distributional Dynamics*

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

Luis Calderon‡
University of Bonn

Moritz Kuhn§
University of Mannheim CEPR, and IZA

October 18, 2024

Abstract

We develop a new method for deriving high-frequency synthetic distributions of consumption, income, and wealth. Modern theories of macroeconomic dynamics identify the joint distribution of consumption, income, and wealth as a key determinant of aggregate dynamics. Our novel method allows us to study their distributional dynamics over time. The method can incorporate different microdata sources, regardless of their frequency and coverage of variables, to generate high-frequency synthetic distributional data. We extend existing methods by allowing for more flexible data inputs. The core of the method is to treat the distributional data as a time series of functions that follow a state-space model, which we estimate using Bayesian techniques. We show that the novel method provides the high-frequency distributional data needed to better understand the dynamics of consumption and its distribution over the business cycle.

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. Christian Bayer gratefully acknowledges funding through the ERC-CoG project Liquid-House-Cycle funded by the European Union's 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).

†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 (Mian, Straub, and Sufi, 2020; Holm, Paul, and Tischbirek, 2021; Andersen et al., 2021; Bhandari et al., 2021; Bayer et al., 2019). The limited availability 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). The proposed method 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 microdata set contains all variables of interest, while others contain only a subset.

The challenge is that the joint distributions of consumption, income, and wealth are infinite-dimensional objects. Our novel method, however, exploits statistical dimensionality reduction techniques by assuming that the distributional dynamics can be captured by a factor model in which the factors themselves have a state-space representation. This assumption builds on insights from the heterogeneous agent macroeconomics literature, which suggests that a few factors should be sufficient to approximate the distributional dynamics, given that a small set of aggregate prices/shocks shape the distribution of consumption, income, and wealth in the short and medium run (Auclert, Bardóczy, Rognlie, and Straub, 2021; Bayer, Born, and Luetticke, 2024). These prices also closely track movements in the aggregate economy. Indeed, the empirical evidence generated by inequality research has so far found much support for this (Di Maggio, Kermani, and Majlesi, 2020; Chodorow-Reich, Nenov, and Simsek, 2021; Kuhn, Schularick, and Steins, 2020).

This set of findings from the macroeconomic literature has three 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 aggregate drivers. Third, 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 the aggregate factors and the distributional data without imposing a structural macroeconomic model.

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 al-

lows us to use and merge numerous microdata sets that refer to the same economic variable but with different operationalized measures, e.g. differences in the sources of income covered. When combining different data sources, it is important that we allow for measurement error in the observation equation of the state-space model. Having an observation equation also allows the combination of microdata sets with different sampling frequencies and also allows us to exploit the information on the evolution of distributions even from microdata that contain only a subset of the variables of interest.

Finally, we overcome the limited availability of high-frequency distributional information and construct estimates of business cycle fluctuations in the joint distributions for any point in time, in particular even for periods for which we do not observe distributions through microdata. The estimated state-space model allows us to construct synthetic high-frequency distributional data by means of the Kalman smoother. The synthetic data itself, while originally functional data, can be expressed approximately in the form of repeated cross-sections of microdata containing consumption, income, and wealth observations of synthetically constructed households. These households represent groups of a granularity that the researcher can flexibly specify.

We demonstrate the power of this novel method by studying the dynamics of the joint distribution of consumption, income, and wealth. 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 of Income and Programme Participation* (SIPP), and the *Current Population Survey* (CPS). We complement the microdata with a comprehensive set of macroeconomic time series. As a first challenge, only the PSID contains all three variables of interest: consumption, income, and wealth. In the other microdata sets at least one of the variables is absent. At the same time, all of the microdata sets contain some information on the joint distribution of consumption, income, and wealth. Second, all of these data sets are available at different frequencies. Third, they differ in sampling approaches and details of their measurement concepts. Our method deals with all three challenges.

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 changes, 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 2021.

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 validate the choice of priors in Bayesian estimation, particularly with respect to measurement error. We show that the state-space model is consistent with the sampling uncertainty of the observed distributional data at the sampling points. Furthermore, 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. Third, we 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). We conclude that our method is capable of producing reliable estimates of distributional data at the business cycle frequency.

Finally, as an application example, we show the dynamics of consumption along the joint distribution of income and wealth, which can provide empirical guidance for model building for macroeconomic models with rich heterogeneity. We show that consumption of the middle class hardly moves in any recession whereas consumption of the poor and the rich show a significant cyclical pattern. Comparing the Dotcom-recession, the Great Recession and the Covid Recession, we however also document, that not all recessions are alike in terms of who in the tails of the distribution wins or looses in terms of consumption. The Great Recession brought consumption losses of the wealth rich, who had even gained in terms of consumption during the Dotcom Recession. The Covid Recession primarily brought about consumption losses of the income rich, but not of the wealth rich.

The remainder of this paper is organized as follows: Section 2 provides an overview of the relevant literature. Section 3 develops our estimation method. Section 4 evaluates the quality of the estimation. Section 5 provides the application examples of the novel method. Finally, section 6 concludes the paper. Appendix A discusses the data sources used in our empirical application.

2 Literature

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. We differ from this paper in three important ways. First,

we develop a method that is suitable for dealing with the evolution of joint distributions over time (distributions of consumption, income, and wealth). Second, we follow Kneip and Utikal (2001) and Tsay (2016) and Ramsay and Silverman (2005, Chapter 8) in not approximating the distribution functions by a fixed set of basis functions, but rather determine the basis functions based on a principal component analysis (see also Meeks and Monti, 2023, for a macroeconomic application of this method). Third, we focus on the production of high-frequency synthetic distribution data, dealing with missing observations and the mixed frequencies of aggregate and microdata.

The latter focus on generating new microdata relates our work to the large body of empirical literature on trends and fluctuations in inequality that took off after the seminal paper by Piketty and Saez (2003): Blanchet, Saez, and Zucman (2022) propose a methodology for producing high-frequency (monthly), timely income and wealth distribution statistics for the United States from 1976 to the present. The paper matches CPS and SCF microdata with individual tax data collected from Piketty, Saez, and Zucman (2018) to produce a harmonized set of monthly microfiles representing synthetic adults, whose income/wealth data are consistent with national accounts totals and whose distribution reflects only publicly observed data. The paper emphasizes the timeliness of its data, which facilitates policymaking and public discourse on social inequality (e.g., age, race, gender) as well as income and wealth inequality. Most of this work focuses primarily on income or wealth separately, and thus often concentrates on marginal distributions, emphasizing specific moments of these distributions, such as top wealth shares. The latter is a widespread feature of the literature. For example, Smith, Zidar, and Zwick (2021), which uses administrative data and the capitalization method of Saez and Zucman (2016), also provides high-frequency estimates, but focuses more on the top of the wealth distribution, albeit at greater wealth granularity.

Similar to what we do, Batty et al. (2020) also construct an extensive synthetic dataset of quarterly estimates since 1989 on the balance sheet of U.S. households; they rely on the granularity of the wealth module of the SCF and aggregate information from the Financial Accounts. They use Chow-Lin/Fernández type models (see Chow and Lin, 1971; Fernandez, 1981; Litterman, 1983) in this endeavor. The advantage of our state-space approach is that it can explicitly deal with sampling uncertainty by treating the underlying microdata as samples of a time series of distribution functions. This means that we can explicitly deal with sampling uncertainty in a dynamic setting and combine different microdata sources that contain the same economic objects with slightly different operationalizations of measurement. Moreover, by combining many microdata sources, we can go a step further and obtain the business cycle fluctuations in the joint distributions of consumption, income, and wealth going back to the 1960s.

With the estimated joint distribution, we can speak to the large macroeconomic literature

that has established the importance of heterogeneity for modeling macroeconomic dynamics (Kaplan, Moll, and Violante, 2018; Bayer et al., 2019; Bayer, Born, and Luetticke, 2024) and complement it with the still missing descriptions of the short-run dynamics of the consumption, income, and wealth distribution. Our new method aims at filling this gap in the literature. The data on the joint distribution allow for an analysis of the dynamics of the joint distribution that provides important information for model building and thus extends the rapidly growing literature that examines the impact of policy shocks on the marginal distributions of consumption, income, or wealth and their aggregate feedbacks (see for example Berger, Bocola, and Dovis, 2023; Coibion, Gorodnichenko, Kueng, and Silvia, 2017; Cloyne, Ferreira, and Surico, 2020; Holm, Paul, and Tischbirek, 2021; Chang and Schorfheide, 2024; Bartscher, Schularick, Kuhn, and Wachtel, 2022). McKay and Wolf (2023) surveys the empirical literature on the effects of monetary policy on inequality.

Methodological Literature. The paper addresses the large methodological and global literature that focuses on estimating high-frequency time series using related series and lower-frequency measures. Common methods for estimating such balanced time series include interpolation (Friedman, 1962; Denton, 1971), regression-based methods with autocorrelated errors (Chow and Lin, 1971; Fernandez, 1981; Litterman, 1983), dynamic Chow-Lin models, and structural multivariate time series models that allow for endogeneity of related series (Salazar and Weale, 1999; Silva and Cardoso, 2001; Gregoir, 2003; Di Fonzo, 2003).

The paper takes a resurfaced approach; specifically, formulating the estimation in a state space framework (Harvey and Pierse, 1984; Harvey, 1990; Harvey and Chung, 2000; Mönch, Uhlig, et al., 2005; Moauro and Savio, 2005; Proietti, 2006) of functional data (see e. g. Kneip and Utikal, 2001; Diebold and Li, 2006; Chang, Kim, and Park, 2016; Inoue and Rossi, 2021; Otto and Salish, 2022, for economic applications). The main advantages of using a functional state-space model lie in its (1) flexibility: with appropriate manipulation, it can encompass the other models, (2) its ability to use the well-studied Kalman filter, its results, and intuitive diagnostics, and (3) its dynamic nature, which is not the case with widely used models such as the Chow-Lin/Fernandez models.

3 Method

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. This method uses microeconomic and aggregate data as inputs. It requires only the joint observation of the microeconomic variables in at least one data set 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 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 $t \in \mathbb{T} := \{1 \dots T\}$. 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 Panel Study of Income Dynamics (PSID) and the Survey of Consumer Finances (SCF) use different sampling procedures and slightly different concepts of wealth and income. We index each of the sampling procedures/data sets by $j = 1 \dots J$. All of these different datasets are typically not observed in all time periods. Instead, data set j is only observed in a particular subset $\mathcal{T}_j \subset \mathbb{T}$. Also, 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 data set, j, must contain all the variables of interest for $\mathcal{D}_j = \mathbb{D}$. In our application, such a dataset is the PSID, which contains information on consumption, income, and wealth (at least 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 the information from the various related microdata sources and the 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

$$d\Xi_{t+1} = G\left(d\Xi_t, Y_t\right) + \epsilon_t, \tag{1}$$

where Y_t are observed aggregate data (including lags), observed controls. G determines the dynamics of the system, and ϵ_t are the corresponding shocks to the functional equation.[1] This structure arises naturally in so-called HANK models (see e.g. Bayer, Born, and Luetticke, 2024). Since we observe Y_t every period and assume that we do so without measurement error, we can disregard any potential endogeneity of Y_t for the purposes of our exercise.

[1] For Ξ_{t+1} to be a distribution, we assume that $\int G(dF, \cdot)(w)dw = 1$, $\int \epsilon_t(w)dw = 0$, and $G(dF, Y_t)(w) \geq -\epsilon_t(w)$.

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 data set gives us an estimate

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

The measurement error then captures time-varying differences in sampling and operationalization of economic concepts. The integral $\int_{\mathbb{D}\setminus\mathcal{D}_j}$ reflects that those variables unobserved in dataset j have been integrated out.

3.2 Implementing the Estimation

Estimating Equation (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 rendering it possible to estimate the state-space model (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 reduce its dimensionality by finding an appropriate factor representation (Section 3.2.2). First, however, we need to operationalize the measurement of the distribution functions as they appear in equation (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 data sets.

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 microdata set j, $w_{i,j,t}$, by its dataset- and time-specific mean $\bar{w}_{j,t}$.[2] 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 $[0,1]$). In addition, we represent the marginals by their quantile functions (i.e., the inverse of

[2]Estimating the model relative to the dataset-specific time means also allows us to flexibly match per capita/household aggregate targets to the synthetic data our estimation produces. This simply requires that the constructed synthetic high-frequency distribution data be scaled back not by the dataset-specific time average, but by the appropriate aggregate target. We can also produce a consensus estimate of the business cycle component across all datasets by using average fixed effects when scaling back and not adding back any trend.

the marginal cumulative distribution function). Again, the quantiles have 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 is based on the fact 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.[3] The copula captures the dependence structure between the random variables in w and is invariant to monotone transformations in w. For our application, the copula is

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

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 of w s.t. $U_m = \Xi_m(w_m) \sim U[0,1]$.[4] 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 $k \in \{1 \ldots d\}$ that when integrating out all but one dimension (the marginal distribution) k, the copula is identical to the value of the marginal distribution.

$$\int C_t(u_1, \ldots u_k, \ldots u_d) du_1 du_{k-1} du_{k+1} \ldots du_d = u_k. \tag{5}$$

Also, $C_t(1, \ldots, 1) = 1$.

In the actual estimation, we work with copula densities $dC_t \in \mathcal{L}^2$, and along with the quantile functions $\Xi_{m,t}^{-1} \in \mathcal{L}^2$, we project them onto a space of orthonormal legendre polynomials $Q \in \mathcal{L}^2$, shifted to fall in the support $[0,1]$.[5]

For the quantile function for variable m, this means the following representation:

$$\Xi_{mt}^{-1}(u_m) = \sum_{o \in \mathbb{N}} \xi_{o,t}^m Q_o(u) \tag{6}$$

[3] The theoretical foundations of splitting distributions into marginals and a copula were laid down by Sklar (1959) and Sklar (1973).

[4] Any distributional transform that produces a uniform cdf will do. See Rüschendorf (2009) for details.

[5] The series estimator also satisfies the uniform margins property for the copula density. See Bakam and Pommeret (2023) for asymptotic properties of the estimator and further details.

and for the copula density

$$dC_t(u_1, u_2, ...u_d) = \sum_{(o_1,...,o_d)\in\mathbb{N}^d} \kappa_{(o_1,...,o_d),t} \prod_{m=1}^{d} Q_{o_m}(u_m) \tag{7}$$

where

$$\xi_{o,t} = \left\langle \Xi_{m,t}^{-1},\, Q_o(u) \right\rangle = \int_0^1 \Xi_{m,t}^{-1}\, Q_o(u) du \tag{8}$$

$$\kappa_{o_1,...,o_d,t} = \left\langle dC_t,\, \prod_{m=1}^{d} Q_{o_m}(u_m) \right\rangle = \int_{[0,1]^d} \prod_{m=1}^{d} Q_{o_m}(u_m)\, dC_t\, du_1, \ldots, du_d \tag{9}$$

the coefficients are the inner products of the functions and the legendre polynomials for some order o. For the estimation of these coefficients, first, we rely on the uniformity of ranks and that ranks are within $[0, 1]$. Second, by orthonormality, these coefficients are identified without impact from other polynomial terms. This implies the inner product can be approximated by a simple sample average representation:

$$\hat{\xi}_{o,t}^{m} := N^{-1} \sum_i w_{m,i,t} Q_o(u_{m,i,t}) \tag{10}$$

$$\hat{\kappa}_{o_1,...,o_d,t} := N^{-1} \sum_i \left(\prod_{m=1}^{d} Q_{o_m}(u_{m,i,t}) \right) \tag{11}$$

where $u_{m,i,t}$ are the data ranks of $w_{m,i,t}$, the sample analogue of $\Xi_{m,t}^{-1}$ for observation i.

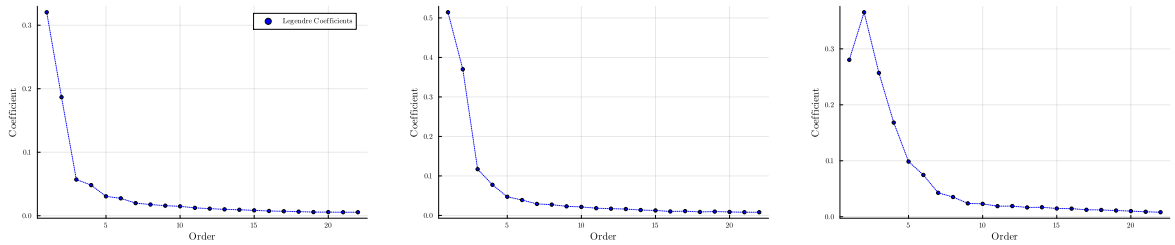
Since the functions are projected onto a space spanned by infinitely many polynomials, this implies estimating infinitely coefficients – just not feasible in any capacity. Thus, we truncate the sums in equations (6) and (7) at a given maximal order $\mathbb{O}$ and by orthonormality of the polynomials, the kept coefficients are unaffected by the truncation. The coefficients in our case indeed decrease rapidly with each polynomial as we will show in Figure 1. Coeffcients beyond order 10 are negligibly small.

Dealing with partial unobservability of the microdata Another difficulty is that not all microdata sets contain the entire vector w as an observable; some contain only a subset. This means that we cannot generate estimates of the full d-dimensional copula at all times and for all data sets. However, we can still estimate copulas with the unobserved dimensions integrated out — lower dimensional copulas.

The representation in the form of Legendre polynomials is very useful in this respect. First, observe that the lower dimensional copula density has to be equal to the higher-dimensional

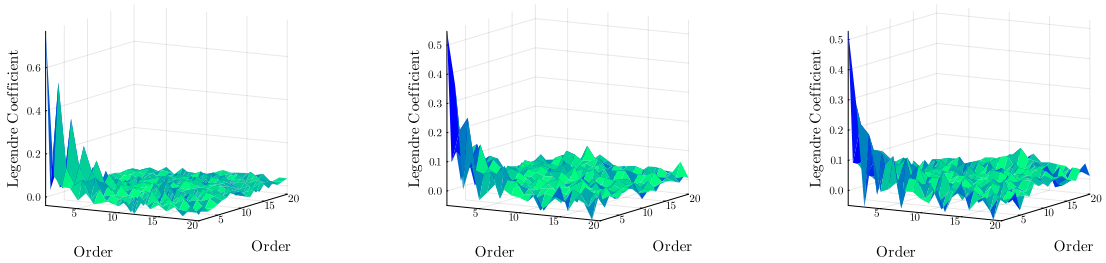
Figure 1: Legendre Coefficients Across the Distributional Data

Legendre Coefficients across Quantile Functions



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

Legendre Coefficients across Copulas



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

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

one when we integrate out the "missing" dimension d:

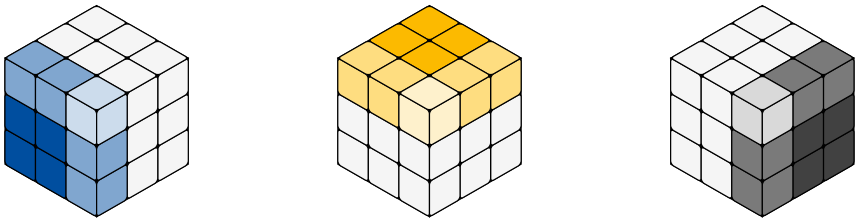
$$dC(u_1,\ldots,u_{d-1}) = \sum_{o_1}\cdots\sum_{o_{d-1}} \kappa_{o_1,\ldots,o_{d-1}} \left(\prod_{m=1}^{d-1} Q_{o_m}(u_m)\right) \stackrel{!}{=} \int_0^1 dC(u_1,\ldots,u_d)\,du_d \tag{12}$$

Next, we write out the integral and make use that the first (shifted) Legendre polynomial integrates to one while all others integrate to zero to obtain:

$$\begin{aligned}\int_0^1 dC(u_1,\ldots,u_d)\,du_d &= \int_0^1 \sum_{o_1}\cdots\sum_{o_{d-1}} \kappa_{(o_1,\ldots,o_d)} \prod_{m=1}^{d} Q_{o_m}(u_m)\,du_d \\ &= \sum_{o_1}\cdots\sum_{o_{d-1}}\sum_{\mathbf{o_d}} \kappa_{(o_1,\ldots,o_{d-1},\mathbf{o_d})} \left(\prod_{m=1}^{d-1} Q_{o_m}(u_m)\right) \int_0^1 Q_{\mathbf{o_d}}(u_d)du_d \\ &= \sum_{o_1}\cdots\sum_{o_{d-1}} \kappa_{(o_1,\ldots,o_{d-1},1)} \left(\prod_{m=1}^{d-1} Q_{o_m}(u_m)\right)\end{aligned} \tag{13}$$

In words, the polynomial coefficients of the lower dimensional copula density are identical to the leading "slice" of the higher dimensional copula. This means that when a dataset does only obtain two out of the three variables of interest, we still obtain a measurement of a subset of the coefficients from this data, see Figure 2.[6]

Figure 2: Geometric representation of partially observed copula density 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 time t. 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. In 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 array of coefficients for each time period, t, leaves us with sequences of coefficients, $\theta^j_t = (\xi^{j,m}_{o_1,t},\ldots\kappa^j_{(o_1,\ldots,o_d),t})$, for each cross-sectional dataset, j. For example, with $d = 3$ dimensions in our application — consumption, income, and wealth and using order

[6] By the same line of argument, a copula requires $\kappa_{(1,\ldots,1)} = 1$ and $\kappa_{(1,\ldots,j,\ldots,1)} = 0$ (i.e., only a single order is not one).

ten polynomials ($O = 10$) to organize the data — the copula density would be represented by a vector with 972 variable entries for each time point and $(d-1) \times O + 1 = 28$ invariable. These 28 invariable entries are redundant due to the constraints imposed by C being a copula. In addition, there would be $d \times O = 3 \times 10$ coefficients of the polynomials representing the quantile functions, which we collect in θ_t^j as well.

From this example it is clear that the dimensionality of $\theta_t^j \in \mathbb{R}^N, N = O^d + 1$ for a dataset with d variables and polynomial order O is too large to formulate and estimate a time series model directly in terms of θ_t^j itself. For this purpose, we postulate (and then estimate) a dynamic factor model for θ. Another advantage of the polynomial representation comes in handy: The variance (over time) of a coefficient is proportional to its contribution to the fluctuations of the function (in the L^2 sense). Put simply: The polynomial coefficient provides a useful form of standardization that provides 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 (free) coefficients of the polynomial representation of dC_t^j (and separately of the quantiles) are horizontally concatenated t:

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

and perform principal component analysis (a singular value decomposition), which nonparametrically reduces the dimensionality of the data (Breitung and Eickmeier, 2006).[7]

Before performing this model reduction, we detrend and standardize the distribution data θ separately by data source j and distribution objects $o \in \{1 \dots d+1\}$ (d quantile functions and a copula) and obtain standardized measures x. [8] This takes care of data source specific effects. We store the information needed to transform x back to the originally observed objects to obtain source-specific predictions. For example, the quantiles of income (that would be one object o) in the SCF and the PSID (two sources j) may be permanently different due to differences in sample design and operationalization. The fact that θ are polynomial coefficients already makes them comparable within the object, so we do not need to standardize them within the object.

[7] Performing principal component analysis on the polynomial coefficient domain or the observed data is equivalent to standardizing the data (Chen, Er, and Wu, 2005).

[8] The normalization of the coefficient n in the object $o(n)$ in the data set j is given by $x_{jnt} = \left(\frac{\theta_{nt}^j - \mu_{jo(n)}}{\sigma_{jo(n)}} \right)$ where $\mu_{jo(n)}$ are the specific means and $\sigma_{jo(n)}$ are the standard deviations of all coefficients of the object $o(n)$ (copula, quantile functions) to which the coefficient n refers.

This leaves us with the standardized observation x_{jnt} of coefficient n in data source j at time t. Note that in some data sources a coefficient n may be impossible to construct, and therefore unobserved, because that data source does not contain information on the corresponding variable. For example, the SCF does not contain information on consumption. For the principal component analysis, we then concatenate all observations that do not have missing coefficients in $\boldsymbol{x}$. The principal component analysis of $\boldsymbol{x}$ provides us with a projection matrix $\Gamma \in \mathbb{R}^{N \times R}$, with full column rank R, that projects $R << N$ factors into the $N = O^d + 1$ dimensional distributional data. More specifically, we decompose $\boldsymbol{x}$ into latent orthogonal factors $\begin{bmatrix} F & f \end{bmatrix}'$ (ordered by importance) and their time constant loadings $\begin{bmatrix} \Gamma & \gamma \end{bmatrix}$. This decomposition is unique up to the scale of each factor, which allows us to normalize the loadings so that all factors have unit variance. The factors obtained are then divided into "important" and "unimportant" factors according to their contribution to the total variance (measured by their singular value):

$$\boldsymbol{x} = \begin{bmatrix} \Gamma & \gamma \end{bmatrix} \begin{bmatrix} F & f \end{bmatrix} \tag{14}$$

where F represents the R important factors, which capture almost all of the variation in the data, and f the $N - R$ less important factors, which can be interpreted as some measurement noise. This step, in a sense, identifies an ideal functional basis (Kneip and Utikal, 2001; Tsay, 2016, see) (the columns of Γ) for approximating the changes in the distribution over time, and reduces the dimensionality of the data entering the state-space model.

We also perform a PCA on the aggregate data (see Appendix A) to further reduce the dimensionality of the controls. The retained important aggregate factors are denoted by Y.

3.2.3 Factor State Space Model and Measurement

With this preprocessing of the data, we can turn to estimating the state-space model that captures the evolution of the distributional factors. Specifically, we postulate the following state space model

$$F_{t+1} = AF_t + BY_{t+1} + \epsilon_{t+1}, \qquad \epsilon_t \sim \mathcal{N}(0, \Omega). \tag{15}$$

Since the factors are orthogonal by construction, we restrict the innovations of the dynamic model ϵ_t to be independent, so that Ω is a diagonal matrix with diagonal entries $\omega_1, \ldots, \omega_R$. The loading matrix B on the aggregate controls Y_t, as well as the law of motion matrix A, are not constrained.

Since the factors are not directly observable, we complement the factor model with an

observation equation for each data set j.

$$\boldsymbol{x}_{t+1}^{j} = H_{t+1}^{j}(\Gamma F_{t+1} + v_{t+1}^{j}), \qquad v_{t+1}^{j} \sim \mathcal{N}(0, \Sigma_{j}^{1/2}\Delta_{j}\Sigma_{j}^{1/2\prime}), \tag{16}$$

where Δ_j is a diagonal matrix with the n-th diagonal entry $\delta_{jo(n)}$ and Σ_j is a positive semidefinite (covariance) matrix. Stacking the data sets j then yields the complete observation equation.

The observation equation (16) translates the factors into observations of the distribution $\boldsymbol{x}_t^j$ for data set j via our estimated projection matrix Γ and the selector matrix H_t^j, which indicates whether (parts of) the distribution are observed in data source j at time t. This logical matrix $H_t^j \in \{0,1\}^{N\times N}$ indicates whether a given variable is observed in a given data set (Durbin and Koopman, 2012, following).

The measurement error, v_t^j, is composed of sampling uncertainty and other errors that reflect the fact that a given data set 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), but can also become differentially important over time.[9] To limit the number of parameters to be estimated within the time series model, we assume that the correlation structure of all measurement errors is the same as 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.[10] The N elements of the diagonal matrix Δ_j are estimated within the time series model with the restriction that its entries vary only by data set and object $o^j(n)$.

[9]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, which makes 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)

[10]To do this, we draw bootstrap samples (or equivalently use the supplied replication weights) for each data set j, $\{\boldsymbol{x}_{t,b}^j\}_{b=1}^B$, 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 data set j. If an object o is unobserved in data set j, we set the covariance terms to zero and the diagonal elements to one to still be able to compute $\Sigma_j^{-1/2}$. In our application, for example, this means that in the PSID, where we observe consumption, income, and wealth and have replication weights, we estimate a full $(N \times N)$ variance covariance matrix Σ_{PSID}. In the CEX, where we only observe consumption and income, we bootstrap the variance covariance matrix for the objects related to these two variables. We set the off-diagonal entries for wealth-related objects to zero and the diagonal elements to one.

Since we have an external estimate of Σ_j, we rewrite the observation equations as

$$\begin{aligned}
\hat{\Sigma}_j^{-1/2}\boldsymbol{x}_{t+1}^j &= \hat{\Sigma}_j^{-1/2}H_{t+1}^j\hat{\Sigma}_j^{1/2}(\hat{\Sigma}_j^{-1/2}\Gamma F_{t+1} + \tilde{v}_{t+1}^j) \\
\hat{\Sigma}_j^{-1/2}\boldsymbol{x}_{t+1}^j &= \hat{\Sigma}_j^{-1/2}H_{t+1}^j\hat{\Sigma}_j^{1/2}(\hat{\Sigma}_j^{-1/2}\Gamma F_{t+1} + \tilde{v}_{t+1}^j) \qquad \tilde{v}_{t+1}^j \sim \mathcal{N}(0, \Delta_j) \\
\tilde{\boldsymbol{x}}_{t+1}^j &= \tilde{H}_{t+1}^j(\tilde{\Gamma}_j F_{t+1} + \tilde{v}_{t+1}^j), \qquad \tilde{H}_{t+1}^j := \hat{\Sigma}_j^{-1/2}H_{t+1}^j\hat{\Sigma}_j^{1/2}, \quad \tilde{\Gamma}_j := \hat{\Sigma}_j^{-1/2}\Gamma.
\end{aligned} \tag{17}$$

The Equations (15) and (17) form a standard system of equations to be estimated by Bayesian techniques using the Kalman filter.

3.2.4 Bayesian Estimation

We need to estimate the (vector) autocorrelation A of the factors, the loading matrix B on the aggregate controls, the variance-covariance matrix of the shocks to the factors Ω, and the variance-covariance matrix $\Sigma_j^{1/2}\Delta_j\Sigma_j^{1/2'}$ of the measurement errors. The covariance structure $\Sigma^{1/2}$ is estimated outside the time series model, as noted above, while the scaling matrices Δ_j are estimated within the model. Given the size of the A, B, Ω, and Δ matrices, we use a Bayesian approach to estimate the system. We do this by shrinking all entries of B and the non-diagonal entries of A to zero if they are not needed to explain the data. For the diagonal entries of Δ, we apply the restrictions described in the last subsection.

The estimation is then a standard Bayesian VAR estimation with mixed frequency data. We collect all parameters in the parameter vector $\boldsymbol{\theta}$ and formulate prior likelihoods $p_{\text{prior}}(\boldsymbol{\theta})$.

Recall that the scaling factors $\delta_{o,j}$ in the matrix Δ define how much larger the actual measurement error standard deviation for each object o (quantile functions and copula) in each data set j (PSID, CPS, CEX, SCF, SIPP) is compared to the corresponding sampling uncertainty that we directly estimate. We assume an inverse gamma prior. We set the mean of this distribution to 5/3, so that a priori we expect additional measurement error from conceptual differences. However, the mode of the distribution is set to one, which would imply only sampling uncertainty and no additional measurement error reflecting conceptual differences. With this prior, two-thirds of the distribution falls between 0.7 and 2.0, with values below 1.0 allowing for the possibility that our estimates $\hat{\Sigma}_j$ are too large, which is important to allow for since they are estimates themselves. Identical priors across datasets mean that we do not a priori prioritize a conceptual measure for object o in one dataset over other datasets. Setting different priors is generally possible if a particular measurement concept should be prioritized on theoretical grounds.

For the matrices A and B, we use Minnesota priors

$$\begin{pmatrix} vec(A) \\ vec(B) \end{pmatrix} \sim \mathcal{MN}(\mu_{Minn}, V_{Minn}). \tag{18}$$

We specify the parameters of the prior distributions, the hyperparameters, as follows: We set the vector of expected values μ_{Minn} so that all but the autocorrelation terms in A have an expected value of zero. The expected values for the autocorrelations (main diagonal of A) are set to 0.90 to reflect the quarterly nature of our data and the typically high persistence in aggregate economic time series. The choice of the variance-covariance hyperparameter V_{Minn} is discussed in detail in the Appendix B using a variant of the original Minnesota prior (Doan, Litterman, and Sims, 1984; Litterman, 1980).

For the variances of shocks to the factors Ω, we specify the prior for each of the diagonal elements as an inverse gamma distribution with mean 0.19, such that the a priori long-run variance of each factor is $1.0 = \frac{0.19}{1-0.90^2}$, consistent with our prior for autocorrelation (in the matrix A) and factor normalization to unit variance, see (15). We set the variance of the prior extremely large to make the prior relatively uninformative.[11]

Likelihood and Sampling With this prior on $\boldsymbol{\theta}$, we obtain the model likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ using a Kalman filter. The posterior log-likelihood is then calculated as the sum of the prior log-likelihood and the model log-likelihood. To sample from the potentially complex, multimodal, 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 Estimating the High-Frequency Fluctuations in the Distributional Data

Given the estimated parameters (the posterior mode), we use the Kalman smoother to estimate a sequence of unobserved factors $\hat{F}_t$. With these generated factors $\hat{F}_t$, we obtain a consensus estimate of the standardized polynomial coefficients of the functional distribution data $\hat{\boldsymbol{x}}_t$ by premultiplying the projection matrix. This gives us, for each data source j, a predicted high-frequency sequence of quantile functions, $\hat{\Xi}^{-1}_{j,m,t}$, and copula densities, $d\hat{C}_{j,t}$. Together they describe the sequence of joint distributions, $\hat{\Xi}_{jt}$, as functional data. In our application, we approximate all functions by polynomials of up to order ten. With the estimated sequences of coefficients at hand, we can then generate arbitrary groups of households formed by a range

[11] The inverse gamma distribution with parameters α and β has a well-defined variance for $\alpha > 2$ and becomes less informative as α decreases. Therefore, we set α slightly above 2 and $\beta = 0.19 \times (\alpha - 1)$.

of ranks and obtain their weight by integrating over the copula densities. Similarly, we can obtain average realizations of variables for these groups by integrating over the quantile functions (i.e., by forming conditional expectations). This implementation implies that, without the need for sampling, we obtain an output that can be immediately interpreted as synthetic microdata. For each cell given by a combination of a consumption quantile, an income quantile, and a wealth quantile, we interpret the vector

$$X_{it} = \begin{pmatrix} c_{it} \\ y_{it} \\ w_{it} \\ \omega_{it} \end{pmatrix} = \begin{pmatrix} \int_{u \in U_i^c} \hat{\Xi}_{j,c,t}^{-1}(u) du \\ \int_{u \in U_i^y} \hat{\Xi}_{j,y,t}^{-1}(u) du \\ \int_{u \in U_i^w} \hat{\Xi}_{j,w,t}^{-1}(u) du \\ \iiint_{(u_c,u_y,u_w) \in U_i^c \times U_i^y \times U_i^w} d\hat{C}_{jt}(u^c, u^y, u^w) \end{pmatrix} \tag{19}$$

as data for a synthetic 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, of the households in that cell defines a weight for that synthetic household.[12] We obtain a consensus estimate across datasets by simply averaging the $d\hat{C}_{j,t}$ and $\hat{\Xi}_{j,m,t}^{-1}$ over datasets j.[13]

4 Application

We apply our method to estimate the joint distribution of consumption, income, and wealth at the household level in the United States from 1962 to 2021. We use microdata from the *Consumer Expenditure Survey* (CEX), *Current Population Survey* (CPS), *Panel Study of Income Dynamics* (PSID), *Survey of Consumer Finances* (SCF); including the historical backfiles (SCF+), and the *Survey of Income and programme participation* (SIPP). We abstain from any sample selection in all of these datasets and pool all CEXs of a given year to remove seasonality. We date the CEX to quarter 4 of a sample year, the CPS date is taken as given. The PSID is assumed to reflect quarter 2, the SCF is dated to quarter 3, and the SIPP data are aggregated to quarterly level and then naturally assigned to the respective quarter.[14] Table 1 lists the distributional objects about which each dataset contains information and the respective sample periods that we use.

In terms of aggregate data, we use a wide range of standard business cycle data (GDP,

[12] The integrals can be calculated very efficiently as time varying linear combinations of the time invariant integrals of the basis functions as, with $\bar{Q}_{o,i} := \int_{u \in U_i^c} Q_o(u) du$, we have $X'_{it} = (\sum_o \xi_{o,t}^c \bar{Q}_{o,i_c}, \quad \sum_o \xi_{o,t}^y \bar{Q}_{o,i_y}, \quad \sum_o \xi_{o,t}^w \bar{Q}_{o,i_w}, \quad \sum_{o_1} \sum_{o_2} \sum_{o_3} \kappa_{(o_1,o_2,o_3),t}^w \bar{Q}_{o_1,i_c} \bar{Q}_{o_2,i_y} \bar{Q}_{o_3,i_w})$

[13] Alternatively, given the estimated measurement error variance for each data source and object type, one could use the inverse measurement error standard deviations as weights in averaging across datasets.

[14] If the survey question refers to the past, such as the PSID question on income, we use aggregate growth rates to impute the current level.

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 the sample periods we use for the different micro datasets.

consumption, employment, etc.) as well as data on household balance sheets, asset prices and interest rates. We include data from 1962:Q3 to 2021:Q4. The starting point of the aggregate data determines the earliest date for the sample periods of the microdata used. From these time series, we extract the 34 most important factors. Details can be found in the Appendix A. Details on the priors can be found in Appendix B. The estimated coefficients of the state space model in the posterior mode can be found in Appendix E.

4.1 Reliability Analysis

In the first pass, we check the reliability of our estimation procedure at each step. First, we show that no relevant information is lost by using the factor model for the distribution (Sections 3.2.1 and 3.2.2). Second, we show that our state-space model typically implies estimates for the distributions that are within the confidence bounds of the microdata with approximately the probability corresponding to the confidence bounds. In other words, the reconstructed time series support our choice of priors for the measurement errors (Section 3.2.4). Third, we rerun the estimation omitting some of the microdata samples at selected points in time. We then show that the reconstructed, synthetic microdata agree well with the omitted microdata. In other words, we show that the estimated state-space model is informative about the time-series fluctuations in the distributional data (i.e., this verifies the steps in Sections 3.2.3 and 3.2.4). Finally, we also show that our reconstructed distributional data agree well with the cyclical fluctuations in the distribution of wealth in the World Inequality Database and the Distributional Financial Accounts.

4.1.1 Precision of the Factor Model

The first step in our procedure is to estimate the decile functions and the copula for each year of observation in each dataset (in terms of Legendre polynomial coefficients). Next, we estimate the factor structure in these data. Since we only retain "important" factors, we potentially introduce an approximation error resulting from forcing "unimportant" factors f_t to take time-averaged values. The size of the approximation error can be controlled by choosing how

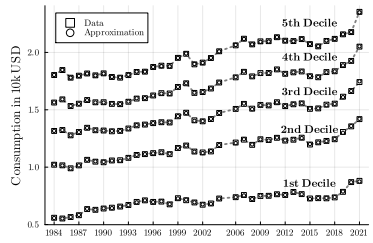
many factors to keep. We choose to retain the seven most important factors, which explain 95% of the (business cycle frequency) variation of the distribution (i.e., of x to be precise).

The different panels in Figure 3 visualize the approximation error in our application. The figure compares the observed conditional decile means for consumption, income, and wealth (squares) with their counterparts from the approximation (circles). We find that the factor model with its seven main factors is very close to the distributional dynamics over time. Figure 4 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 0.09 and 0.12 (between 1999 and 2019),[15] while the divergence between the approximation and the actual distribution is at least an order 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.

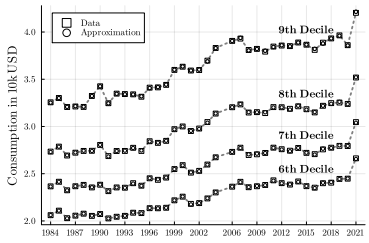
[15]The Kullback-Leibler divergence for the pandemic year 2021 is even 0.2.

Figure 3: Comparison of quantile functions in raw and approximated data (important factors)

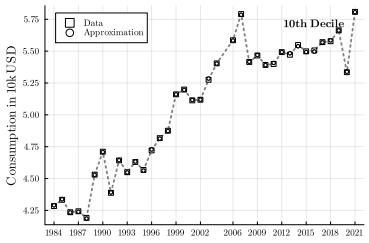
Mean Consumption



(a) 1st to 5th decile

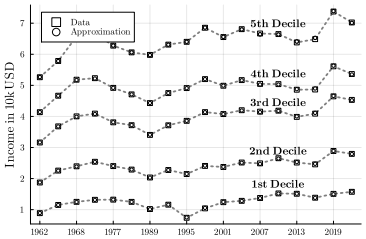


(b) 6th to 9th decile

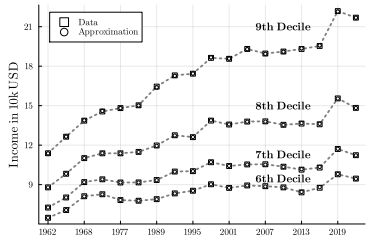


(c) top decile

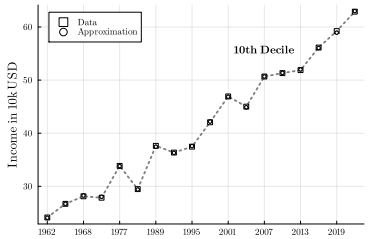
Mean Income



(d) 1st to 5th decile

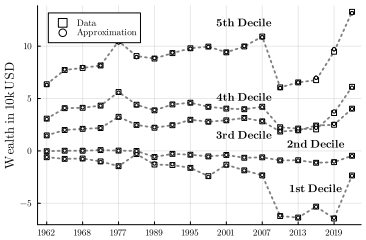


(e) 6th to 9th decile

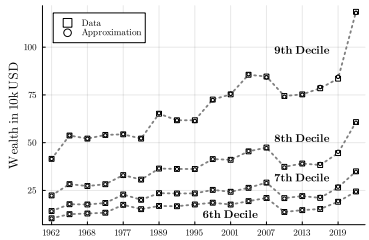


(f) top decile

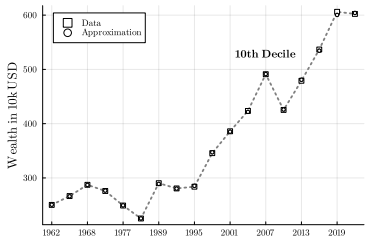
Mean Wealth



(g) 1st to 5th decile



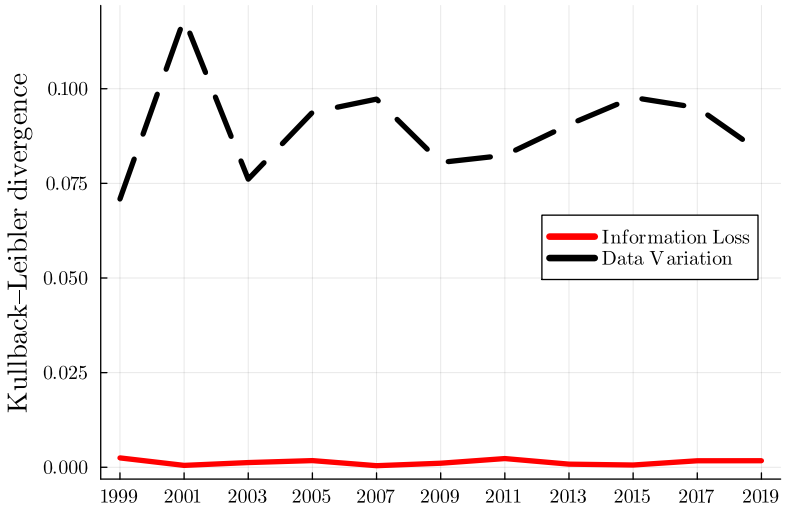
(h) 6th to 9th decile



(i) top decile

Notes: Figure shows the quantile functions (mean within decile) for consumption, income, and wealth deciles from the survey data (squares) and approximation (dots) using only the fluctuations in the most important factors in (14). Top row shows quantile functions for CEX consumption. Middle and bottom row show quantile functions for SCF income and wealth. Dotted lines show linear interpolation between survey waves.

Figure 4: Comparison of the raw data and approximated copula (important factors)



Notes: Figure shows as a black dashed line the distance between the time average copula in the PSID and the raw data copula at every survey year. The red solid line is the distance between the copula that results from letting fluctuate only the most important factors in Equation (14) and the raw data copula at every survey year. Distances are measured in terms of the Kullback-Leibler divergence relative to the raw data copula.

4.1.2 Validation of Hyperparameter Choices

To evaluate our hyperparameter choices, 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 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 probability of the bounds. The fact that the confidence level is an upper bound is because the 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, driving 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 comovements.

Table 2: Deviations of smoothed estimates and microdata: Fraction within confidence bounds

Measure	CEX	CPS	SCF	SIPP	PSID	Overall
Consumption quantiles	87%	—%	—%	—%	100%	89%
Income quantiles	93%	98%	96%	53%	97%	96%
Wealth quantiles	—%	—%	96%	55%	100%	98%
Copula densities	100%	—%	98%	95%	82%	87%

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.

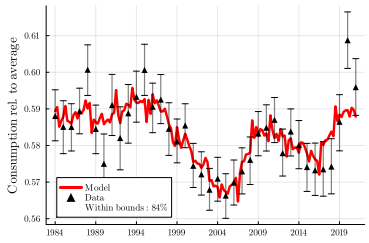
On the other hand, if the prior for the measurement error variance is too high, the estimator will treat the data as uninformative, and the smoother will miss the survey estimates more often and to a much greater extent than implied by its confidence bounds. We validate the choice of hyperparameters graphically for income and wealth in the SCF data and provide comprehensive summary statistics across all datasets and estimates.

Figure 5 shows average income (top row) and wealth (bottom row) for the top 10 percent (first column), the next 40 percent (second column), and the bottom 50 percent (last column) of the respective distributions. It shows the point estimates from the surveys, along with their 95% confidence limits, and the results from the Kalman smoother based on our estimates of the parameters of equation (15). Overall, the smoothed estimates fall outside their respective confidence bounds in only two out of 102 observations. This is a probability slightly smaller than the confidence level.

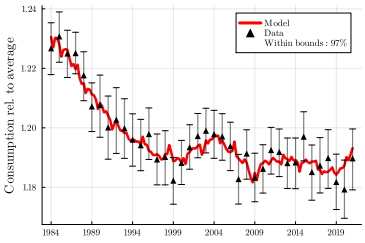
Table 2 provides a comprehensive summary of this validation approach. For all quantile functions and copulas, we report for each data set and survey year how often the respective smoother estimate is within the confidence limits. Again, we use a confidence level of 95%. For the quantile functions, we 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. For no data set and no quantile function does the difference exceed six percentage points. The difference is largest for consumption, where conceptual differences in the surveys are also likely to be the largest. For the copula densities, the smoother shows larger discrepancies, falling in twelve instead of one percent of all observations outside the confidence bounds. Consistent with our findings on consumption quantiles, it is even more difficult to obtain a consensus estimate of the joint rank distribution of consumption and income across PSID and CEX.

Figure 5: Comparison of smoothed distributional data and direct survey estimates

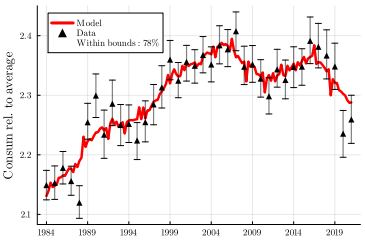
Consumption by consumption



(a) Bottom 50 percent

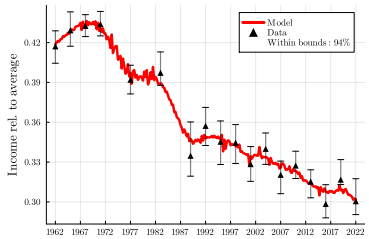


(b) 50-90 percent

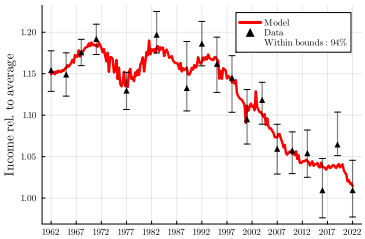


(c) Top 10 percent

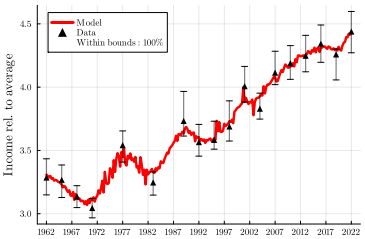
Income by income



(c) Bottom 50 percent

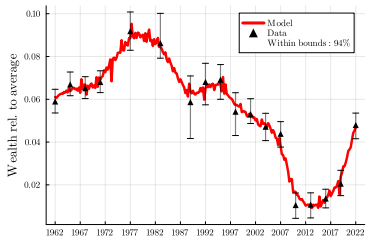


(d) 50-90 percent

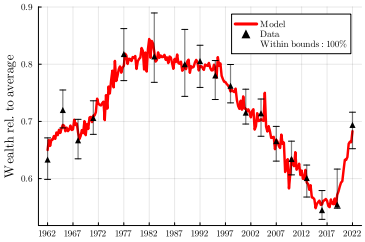


(e) Top 10 percent

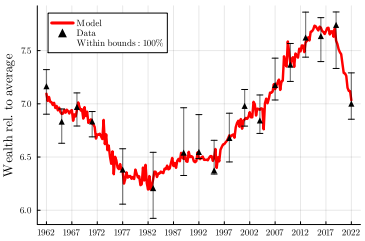
Wealth by wealth



(f) Bottom 50 percent



(g) 50-90 percent



(h) Top 10 percent

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.

4.1.3 Predictability of Distributional Data

To validate how well the method can predict distributional dynamics, we compare how well the model predicts unused microdata. Because we allow for rich dataset-specific fixed effects and trends, we drop a number of observations in the survey year, which we then predict using the model re-estimated on the restricted sample alone (as an out-of-sample prediction). Dropping observations changes the estimated parameters of the factor model (15), but it also changes the estimated measurement errors of the smoothed predictions. Specifically, we perform four experiments shown in the four rows of Figure 6.

First, we include only every fourth CEX survey year in the estimation, reducing the number of CEX survey years included in the estimation from 38 to 10. The focus of this exercise is to predict the distributional dynamics of consumption. Dropping three out of four years is motivated by the fact that most countries in Europe only survey consumption every four years. The first row of Figure 6 shows the average consumption of the top 10%, next 40% and bottom 50% (in terms of consumption) relative to the average across households. There is a large sampling uncertainty around the CEX data, the 95% confidence intervals are displayed as error bars. For this reason, even in the data-rich specification (with annual 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 in the same range, i.e., very high, meaning that the model can predict the consumption distribution well.

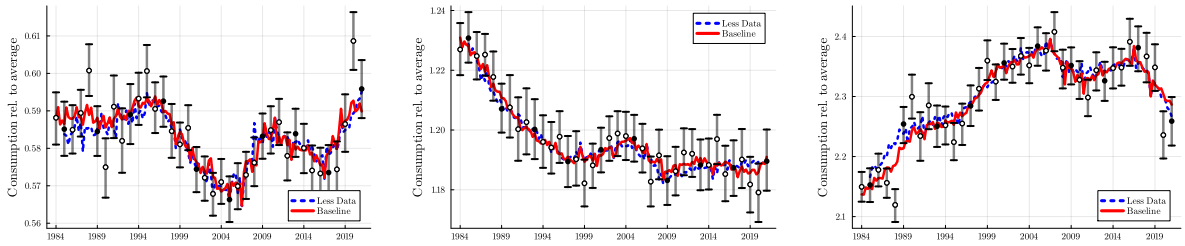
The next three experiments focus on the distributional dynamics of wealth, since wealth is notoriously the least frequently observed data. In the second and third experiments, we remove a series of microdata observations. In the second experiment, this is the last 20 quarters of all microdata. The purpose of this exercise is to assess the predictability of the distributional data using our method in terms of a nowcast of the wealth distribution. The third experiment drops all microdata over the housing cycle of the first decade of the 21^{st} century between 2004Q4 and 2009Q4—arguably a period characterized by large swings in the wealth distribution (Kuhn, Schularick, and Steins, 2020). In both experiments, we let only the aggregate data inform the estimated distributional dynamics for the period in which we remove the microdata. For both experiments, we find that the prediction that uses all microdata and the prediction that omits five years of microdata are very close. Even in the period where we drop the microdata information, the distributional dynamics of wealth are well captured by aggregate factors, in line with previous research (Kuhn, Schularick, and Steins, 2020; Bayer, Born, and Luetticke, 2024).

The fourth experiment rationalizes these results. Rather than re-estimating (15), we drop a single SCF sample at a time when running the Kalman smoother. This gives us 17 smoothed

Figure 6: Predictability of distributional data

Consumption by consumption percentile

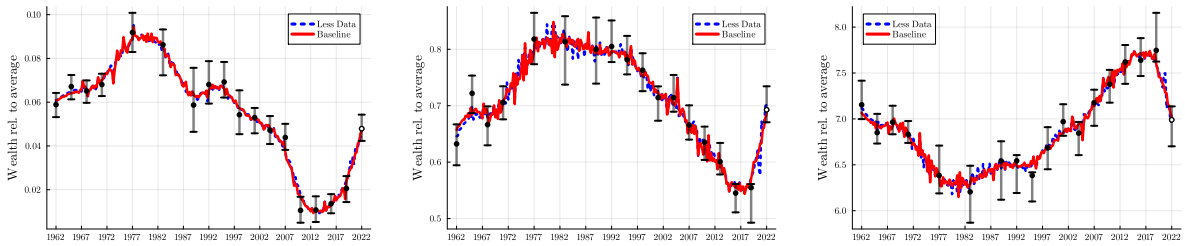
CEX data vs. estimates from using CEX every 4 years



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

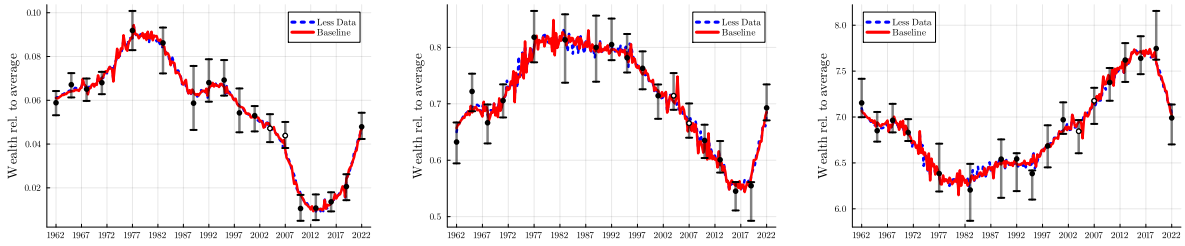
Wealth by wealth percentile

Removing microdata from last 20 quarters of estimation



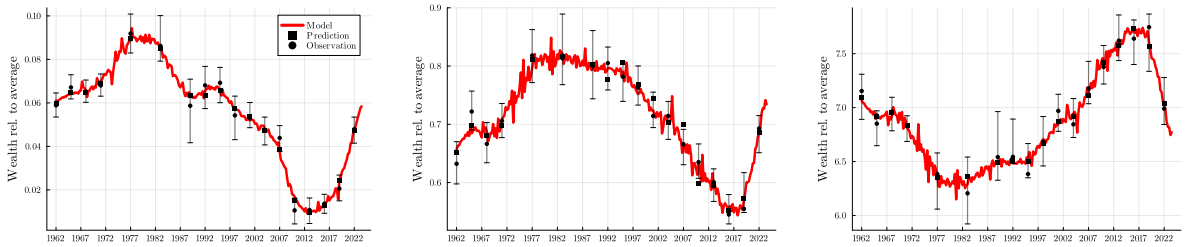
(d) Bottom 50 percent (e) 50-90 percent (f) Top 10 percent

Removing microdata from the housing cycle



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

Removing one SCF wave at a time



(j) Bottom 50 percent (k) 50-90 percent (l) Top 10 percent

Notes: Figure shows baseline model estimate for consumption (panels (a) to (c)) and wealth (panels (d) to (l)) for different samples. Baseline estimates using all data is always shown as 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): dashed blue line shows smoothed estimates when the last 20 quarters of all microdata have been dropped in model estimation and Kalman smoother (empty dots). Panel (g) to (i): Same exercise as (d) to (f) but dropping the observations over the house price cycle (2004Q4 and 2009Q4). Panel (j) to (l): show 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 estimate from the survey data of this wave. Error bars in all figures indicate 95% confidence bounds for each individual survey sample.

distributional data series alongside the one using all SCF waves. The latter is shown as a solid line in the figures in the bottom row of Figure 6. The stars indicate the corresponding smoothed prediction for the time of each of the survey waves that we have omitted. For example, the star in 1992 is the prediction for the wealth distribution where the 1992 SCF survey *not* entered the smoother. The diamond shows the direct estimate from the corresponding survey (with its confidence limits). The fact that the stars 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, aggregate factors must be important.

4.2 Comparison with External Estimates from Other Sources/Methods

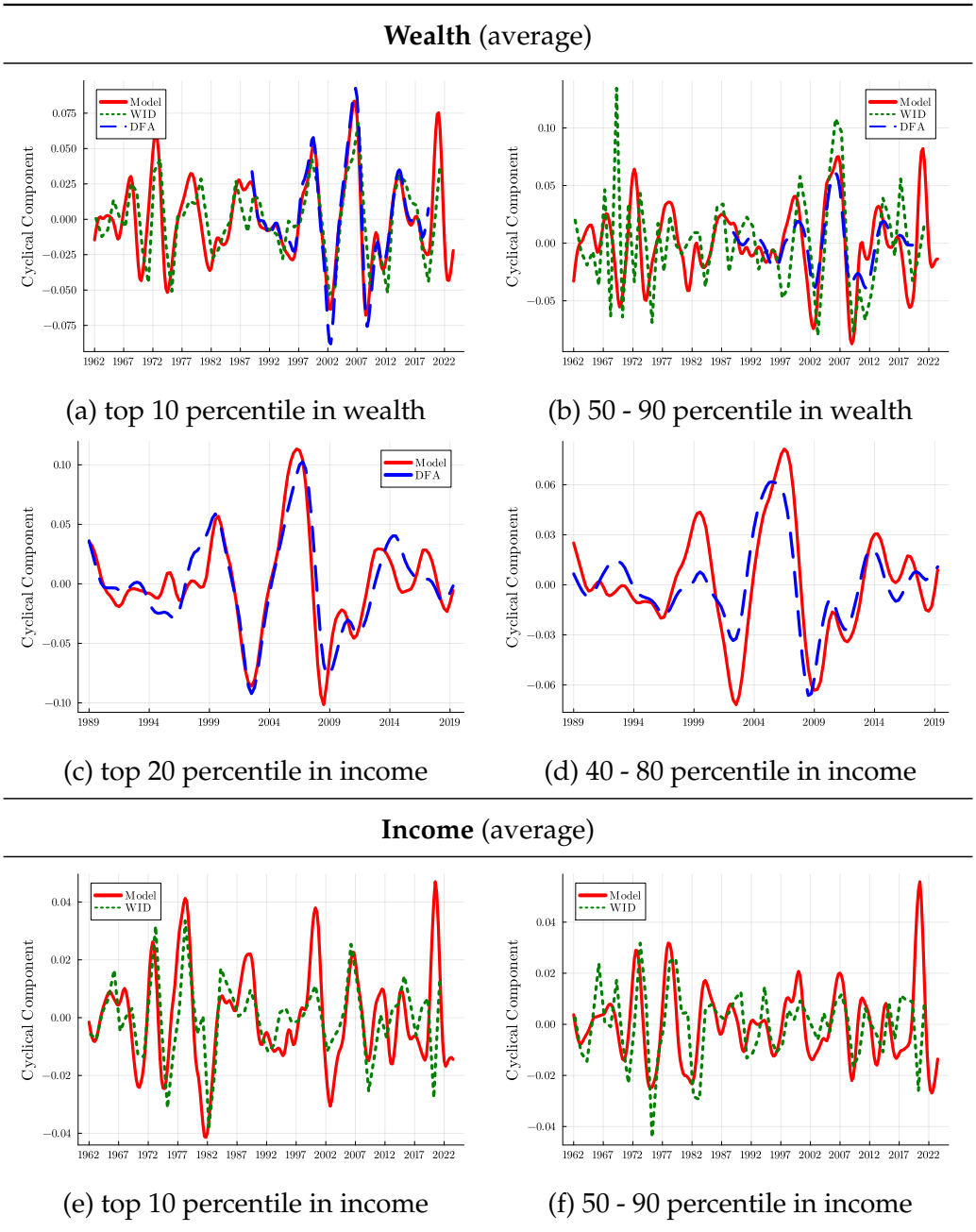
Having established that our method predicts well out of sample but within the surveys used to estimate the distributional data, we compare our high-frequency distributional estimates with estimates for the cyclical component of income and wealth distribution from the *Distributional Financial Accounts (DFA)* (Batty et al., 2020) and the *World Inequality Database (WID)* (Alvaredo et al., 2013). The former is based exclusively on the SCF as microdata and uses a different estimation technique to produce high-frequency distributional estimates. The latter is based on annual tax data, but does not use a time series framework to generate higher frequency data. Thus, both the WID and the DFA are themselves estimates. For comparison, we rely on the DFA estimates of average wealth by income, which allow us to compare the estimates of the joint distribution along this dimension.

Figure 7 shows the results of this comparison for the cyclical fluctuations (in logs).[16] The first two panels compare the wealth by wealth group data from the DFA and the WID. We focus on the wealth of the wealth-richest 10 percent and the wealth of the next 40 percent because the poorest half of the population has wealth very close to zero.

We find that our method produces smoothed estimates that are close to both alternative estimates and well within the range of the DFA and WID. DFA also estimates a time series of wealth by income group, see subfigures (c) and (d). Again, we find a close correlation between our estimates and the DFA estimates where available. Finally, subfigures (e) and (f) look at the income of the richest 10 percent and the next 40 percent, which is only available in the WID. Again, we find a strong comovement between our estimates and the WID estimates, with the WID showing very volatile and non-persistent changes in the income and wealth of the "next 40 percent" households for some years in the 1960s, reflecting the fact that the WID does not use a time series model, so that all estimates are considered independent over time, and it also does not allow for measurement errors resulting from sampling in the underlying microdata.

[16] We define the cyclical component as the difference between the raw series and its HP-filtered counterpart with the smoothness parameter λ set to 1600 for quarterly series and 6 for annual series.

Figure 7: Comparison of cyclical component of distributional data to external sources



(a) top 10 percentile in wealth

(b) 50 - 90 percentile in wealth

(c) top 20 percentile in income

(d) 40 - 80 percentile in income

(e) top 10 percentile in income

(f) 50 - 90 percentile in income

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 $\lambda = 1,600$ for quarterly data.

5 Consumption Dynamics along the Income and Wealth Distribution over the Business Cycle

In estimating consumption dynamics, the literature has traditionally relied on two main approaches: the direct method, which draws on household-level consumption data (e.g., Coibion et al., 2017; Cloyne, Ferreira, and Surico, 2020), and the indirect method, which imputes consumption through a budget identity (e.g., Eika, Mogstad, and Vestad, 2020; Holm, Paul, and Tischbirek, 2021; Fagereng, Holm, and Natvik, 2021).[17] Both methods come with challenges: the direct approach requires rich sampling variation, while the indirect approach is susceptible to errors due to its reliance on assumptions and multiple measurements.

The strength of our method is that it sidesteps both approaches by estimating a joint distribution and with some additional structure, addresses measurement error concerns. On top, the distributional data is of high-frequency and containing income and wealth along with consumption— providing insights that are absent from existing U.S. data. With these high-dimensional data, we can trace the distributional dynamics underlying macroeconomic dynamics, which play an important role in heterogeneous agent business-cycle models (HA) (e.g., Kaplan, Moll, and Violante, 2018; Bayer, Born, and Luetticke, 2024).

5.1 Consumption dynamics over three Recessions

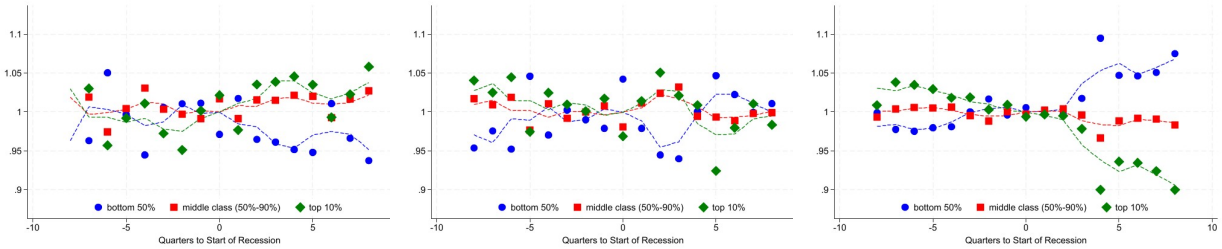
For the application, we understand the key driver of macroeconomic dynamics and a determinant of macroeconomic stabilization policies is the consumption dynamics of households during recessions. In addition, the HA literature has emphasized the importance of heterogeneity in consumption responses along the income and wealth distribution. In this effort, we trace out the consumption dynamics along the income and wealth distribution for the three most recent recessions in the United States: the dot-com recession of the early 2000s, the Financial crisis near the end of the first decade of the 21st century, and most recently, the Covid recession. With this, we contribute to discussions concerning different HA model mechanisms, whose microfoundation still remains constrained by data limitations as high frequency data on consumption dynamics along the income and wealth distribution had remained unavailable.

Exercise. For the three recessions, we consider the consumption dynamics of the bottom 50%, the 50% to 90%, and the top 10% of the income and wealth distribution. To compare the business cycle dynamics of consumption for the different income and wealth groups, we express income relative to the economy-wide average for each period and then index these

[17] The direct approach measures consumption directly from household-level data, while the indirect approach imputes consumption using a household budget identity, which may introduce compounded measurement errors by relying on multiple variables.

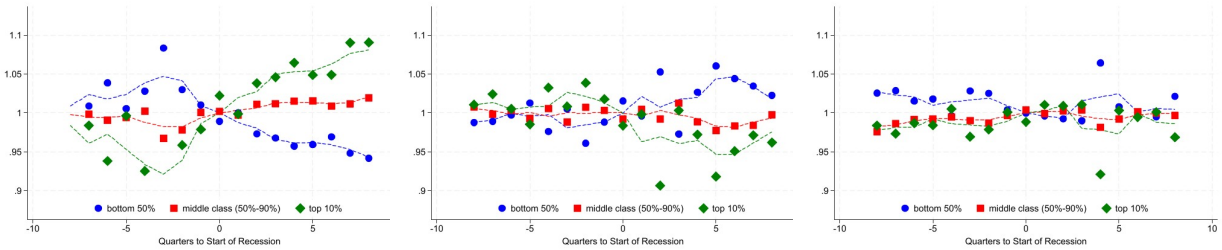
Figure 8: Comparison of consumption dynamics during recessions

Consumption dynamics by income



(a) Dotcom (b) Financial Crisis (c) Covid

Consumption dynamics by wealth



(d) Dotcom (e) Financial Crisis (f) Covid

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 beginning of the recession. The horizontal axes 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 Dotcom recession in 2001q1, Financial Crisis in 2007q4, and Covid in 2019q4. Top (bottom) row shows consumption dynamics by income (wealth) group with the bottom 50%, 50%-90%, and top 10% of the income (wealth) distribution as income (wealth) groups.

relative consumption dynamics to the quarter preceding the recession.[18,19] Given that income is the primary source of consumption financing for most households, Figure 8 first analyzes how consumption evolves across income groups during the last three U.S. recessions.

In Panels (a) to (c), we make several interesting observations. First, outside the Covid recession, the top half of the consumption distribution shows very similar consumption dynamics over recessions. Consumption cyclicality of the top and middle class follow that of the average household before each recession, but the onset of each recession brought different responses: relative increases for the Dotcom bubble, stability during the financial crises, and decreases during Covid. In fact, the top 10% lost 10% relative to average post Covid—a striking pattern for recessions in the 21st century.

Second, we find that the bottom 50% are poorly insured against aggregate risk. The normalization relative to average consumption has already removed aggregate fluctuations in the

[18]We construct symmetric 3-quarter moving averages and use this moving average for the normalization to the pre-recession quarter.

[19]We rely on the NBER business cycle dates with the Dotcom recession starting in 2001q1, Financial Crisis in 2007q4, and Covid in 2019q4.

level of consumption, still we find that relative to the average, the bottom 50% lose typically 5% in terms of consumption growth during the first year of a recession. Again, only the Covid recession is different, where the bottom 50% of the income distribution saw over a 5% increase in their consumption relative to average consumption. In all, only the financial crises saw convergence of households to pre-recession consumption levels — the other recessions exhibit signs of divergence.

Figure 8 also reports consumption dynamics by wealth group in panels (d) to (f). We find that a very different picture emerges. The Covid recession now shows a rather uniform response of consumption across wealth groups. The aftermath of the Financial Crisis and the Dotcom recessions reveals strong diverging dynamics, plausibly driven by large swings in asset prices (Kuhn, Schularick, and Steins, 2020). In the first year of the Financial Crisis, consumption of the wealthiest 10% of U.S. households declined relative to the average by 5% and did not recover in the second year. In the aftermath of the Dotcom recession, we find a flipped picture with the top 10% of households showing much higher consumption growth than the bottom 50% of the wealth distribution. Lottery-like realized gains from short-selling at the top likely fueled these responses (Ofek and Richardson, 2003; Lamont and Stein, 2004) and rising house prices during the early 2000s kept them afloat.

To conclude, two key insights emerge. First, consumption dynamics differ significantly between the income and wealth rich and poor. Whereas some recessions reduce consumption inequality such as the Covid recession, others such as the Dotcom recession increase consumption inequality. Second, asset prices are a likely important driver of the differential consumption dynamics by income and wealth. Comparing the Dotcom recession with large swings in asset prices before and after and the Covid crisis with a strong fiscal response and income support programs, we find very different consumption patterns by income and wealth that a one-dimensional analysis of only consumption by income or by wealth would not have detected. Our novel data allow us to identify such underlying differences in the potential drivers of recession dynamics by jointly studying consumption dynamics by income and wealth. Future work will have to explore in more detail which of the proposed economic mechanisms in the rich class of heterogeneous agent models can account for these new facts that our synthetic distributional data has uncovered.

6 Conclusion

In this paper, we presented 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 deter-

minant of aggregate dynamics. Our method closes a gap as it provides a method to study the empirical distributional dynamics as counterpart to the existing theory over time. We have shown that the method is able to incorporate information from various microdata sources independent 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 a good precision. We show that the derived data can shed new light on the question of how business cycle fluctuations and the distribution of consumption, income, and wealth interact.

References

[1] Seung C Ahn and Alex R Horenstein. "Eigenvalue ratio test for the number of factors". In: *Econometrica* 81.3 (2013), pp. 1203–1227.

[2] Facundo Alvaredo et al. "The top 1 percent in international and historical perspective". In: *Journal of Economic perspectives* 27.3 (2013), pp. 3–20.

[3] Facundo Alvaredo et al. "Distributional National Accounts (DINA) guidelines: Concepts and methods used in WID. world". In: (2016).

[4] Asger Lau Andersen et al. "Monetary policy and inequality". In: (2021).

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

[6] Orazio Attanasio and Luigi Pistaferri. "Consumption inequality over the last half century: some evidence using the new PSID consumption measure". In: *American Economic Review* 104.5 (2014), pp. 122–126.

[7] Adrien Auclert et al. "Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models". In: *Econometrica* 89.5 (2021), pp. 2375–2408.

[8] Jushan Bai and Serena Ng. "Determining the number of factors in approximate factor models". In: *Econometrica* 70.1 (2002), pp. 191–221.

[9] Jushan Bai and Serena Ng. "Rank regularized estimation of approximate factor models". In: *Journal of Econometrics* 212.1 (2019), pp. 78–96.

[10] Yves I Ngounou Bakam and Denys Pommeret. "Nonparametric estimation of copulas and copula densities by orthogonal projections". In: *Econometrics and Statistics* (2023).

[11] Alina K Bartscher et al. "Monetary policy and racial inequality". In: *Brookings Papers on Economic Activity* 2022.1 (2022), pp. 1–63.

[12] Michael Batty et al. "The Distributional Financial Accounts of the United States". In: *Measuring and Understanding the Distribution and Intra/Inter-Generational Mobility of Income and Wealth*. University of Chicago Press, 2020.

[13] Christian Bayer, Benjamin Born, and Ralph Luetticke. "Shocks, frictions, and inequality in US business cycles". In: *American Economic Review* 114.5 (2024), pp. 1211–1247.

[14] Christian Bayer et al. “Precautionary Savings, Illiquid Assets, and the Aggregate Consequences of Shocks to Household Income Risk”. In: *Econometrica* 87.1 (2019), pp. 255–290. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA13601.

[15] David Berger, Luigi Bocola, and Alessandro Dovis. “Imperfect risk sharing and the business cycle”. In: *The Quarterly Journal of Economics* 138.3 (2023), pp. 1765–1815.

[16] Anmol Bhandari et al. “Inequality, Business Cycles, and Monetary-Fiscal Policy”. In: *Econometrica* 89.6 (2021), pp. 2559–2599.

[17] Thomas Blanchet, Emmanuel Saez, and Gabriel Zucman. *Real-time inequality*. Tech. rep. National Bureau of Economic Research, 2022.

[18] Gregor Boehl. “DIME MCMC: A Swiss Army Knife for Bayesian Inference”. In: *Journal of Econometrics* (2024).

[19] Jörg Breitung and Sandra Eickmeier. “Dynamic factor models”. In: *Allgemeines Statistisches Archiv* 90.1 (2006), pp. 27–42.

[20] Minsu Chang, Xiaohong Chen, and Frank Schorfheide. “Heterogeneity and aggregate fluctuations”. In: *Journal of Political Economy* forthcoming (2024).

[21] Minsu Chang and Frank Schorfheide. *On the Effects of Monetary Policy Shocks on Income and Consumption Heterogeneity*. Working Paper 32166. National Bureau of Economic Research, 2024.

[22] Yoosoon Chang, Chang Sik Kim, and Joon Y Park. “Nonstationarity in time series of state densities”. In: *Journal of Econometrics* 192.1 (2016), pp. 152–167.

[23] Weilong Chen, Meng Joo Er, and Shiqian Wu. “PCA and LDA in DCT domain”. In: *Pattern Recognition Letters* 26.15 (2005), pp. 2474–2482.

[24] Gabriel Chodorow-Reich, Plamen T Nenov, and Alp Simsek. “Stock market wealth and the real economy: A local labor market approach”. In: *American Economic Review* 111.5 (2021), pp. 1613–57.

[25] Gregory C Chow and An-loh Lin. “Best linear unbiased interpolation, distribution, and extrapolation of time series by related series”. In: *The review of Economics and Statistics* (1971), pp. 372–375.

[26] James Cloyne, Clodomiro Ferreira, and Paolo Surico. "Monetary policy when households have debt: new evidence on the transmission mechanism". In: *The Review of Economic Studies* 87.1 (2020), pp. 102–129.

[27] Olivier Coibion et al. "Innocent Bystanders? Monetary policy and inequality". In: *Journal of Monetary Economics* 88 (2017), pp. 70–89.

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

[29] David M Cutler et al. "Macroeconomic performance and the disadvantaged". In: *Brookings papers on economic activity* 1991.2 (1991), pp. 1–74.

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

[31] Frank T Denton. "Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization". In: *Journal of the american statistical association* 66.333 (1971), pp. 99–102.

[32] Tommaso Di Fonzo. "Constrained retropolation of high-frequency data using related series: A simple dynamic model approach". In: *Statistical Methods and Applications* 12.1 (2003), pp. 109–119.

[33] Marco Di Maggio, Amir Kermani, and Kaveh Majlesi. "Stock market returns and consumption". In: *The Journal of Finance* 75.6 (2020), pp. 3175–3219.

[34] Francis X Diebold and Canlin Li. "Forecasting the term structure of government bond yields". In: *Journal of Econometrics* 130.2 (2006), pp. 337–364.

[35] Thomas Doan, Robert Litterman, and Christopher Sims. "Forecasting and conditional projection using realistic prior distributions". In: *Econometric reviews* 3.1 (1984), pp. 1–100.

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

[37] Lasse Eika, Magne Mogstad, and Ola L Vestad. "What can we learn about household consumption expenditure from data on income and assets?" In: *Journal of Public Economics* 189 (2020), p. 104163.

[38] Andreas Fagereng, Martin B Holm, and Gisle J Natvik. "MPC heterogeneity and household balance sheets". In: *American Economic Journal: Macroeconomics* 13.4 (2021), pp. 1–54.

[39] Roque B Fernandez. "A methodological note on the estimation of time series". In: *The Review of Economics and Statistics* 63.3 (1981), pp. 471–476.

[40] Marjorie Flavin and Takashi Yamashita. "Owner-occupied housing and the composition of the household portfolio". In: *American Economic Review* 92.1 (2002), pp. 345–362.

[41] Simon Freyaldenhoven. "Factor models with local factors—determining the number of relevant factors". In: *Journal of Econometrics* 229.1 (2022), pp. 80–102.

[42] Milton Friedman. "The interpolation of time series by related series". In: *Journal of the American Statistical Association* 57.300 (1962), pp. 729–757.

[43] Patrick Gagliardini, Elisa Ossola, and Olivier Scaillet. "A diagnostic criterion for approximate factor structure". In: *Journal of Econometrics* 212.2 (2019), pp. 503–521.

[44] Elena Gouskova, Patricia Andreski, and Robert F Schoeni. *Comparing estimates of family income in the Panel Study of Income Dynamics and the March Current Population Survey, 1968-2007*. Survey Research Center, Institute for Social Research, University of . . ., 2010.

[45] Stéphane Gregoir. "Propositions pour une désagrégation temporelle basée sur des modèles dynamiques simples". In: *Paris-Bercy* (2003), p. 141.

[46] James D Hamilton and Jin Xi. "Principal Component Analysis for Nonstationary Series". In: (2022).

[47] Andrew Harvey and Chia-Hui Chung. "Estimating the underlying change in unemployment in the UK". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163.3 (2000), pp. 303–309.

[48] Andrew C Harvey. "Forecasting, structural time series models and the Kalman filter". In: (1990).

[49] Andrew C Harvey and Richard G Pierse. "Estimating missing observations in economic time series". In: *Journal of the American statistical Association* 79.385 (1984), pp. 125–131.

[50] Martin Blomhoff Holm, Pascal Paul, and Andreas Tischbirek. "The transmission of monetary policy under the microscope". In: *Journal of Political Economy* 129.10 (2021), pp. 2861–2904.

[51] Atsushi Inoue and Barbara Rossi. "The effects of conventional and unconventional monetary policy: A new approach". In: *Quantitative Economics* 12.4 (2021), pp. 1085–1138.

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

[53] Greg Kaplan, Benjamin Moll, and Giovanni L Violante. "Monetary policy according to HANK". In: *American Economic Review* 108.3 (2018), pp. 697–743.

[54] Yong-Seong Kim and Frank P Stafford. "The quality of PSID income data in the 1990s and beyond". In: *Technical Series Paper* (2000).

[55] Alois Kneip and Klaus J Utikal. "Inference for density families using functional principal component analysis". In: *Journal of the American Statistical Association* 96.454 (2001), pp. 519–542.

[56] Moritz Kuhn, Moritz Schularick, and Ulrike I Steins. "Income and wealth inequality in America, 1949–2016". In: *Journal of Political Economy* 128.9 (2020), pp. 3469–3519.

[57] Owen A Lamont and Jeremy C Stein. "Aggregate short interest and market valuations". In: *American Economic Review* 94.2 (2004), pp. 29–32.

[58] Robert B Litterman. "Techniques for forecasting with vector autoregressions". PhD thesis. Ph. D. thesis, University of Minnesota, 1980.

[59] Robert B Litterman. "A random walk, Markov model for the distribution of time series". In: *Journal of Business & Economic Statistics* 1.2 (1983), pp. 169–173.

[60] Michael W. McCracken and Serena Ng. "FRED-QD: A Quarterly Database for Macroeconomic Research". In: *Review* 103.1 (2021), pp. 1–44.

[61] Alisdair McKay and Christian K Wolf. "Monetary policy and inequality". In: *Journal of Economic Perspectives* 37.1 (2023), pp. 121–144.

[62] Roland Meeks and Francesca Monti. "Heterogeneous beliefs and the Phillips curve". In: *Journal of Monetary Economics* 139 (2023), pp. 41–54.

[63] Atif R Mian, Ludwig Straub, and Amir Sufi. *The Saving Glut of the Rich*. Tech. rep. National Bureau of Economic Research, 2020.

[64] Filippo Moauro and Giovanni Savio. "Temporal disaggregation using multivariate structural time series models". In: *The Econometrics Journal* 8.2 (2005), pp. 214–234.

[65] Emanuel Mönch, Harald Uhlig, et al. "Towards a Monthly Business Cycle Chronology for the Euro Area". In: *Journal of Business Cycle Measurement and Analysis* 2005.1 (2005), pp. 43–69.

[66] Eli Ofek and Matthew Richardson. "Dotcom mania: The rise and fall of internet stock prices". In: *the Journal of Finance* 58.3 (2003), pp. 1113–1137.

[67] Alexei Onatski and Chen Wang. "Spurious factor analysis". In: *Econometrica* 89.2 (2021), pp. 591–614.

[68] Sven Otto and Nazarii Salish. "Approximate Factor Models for Functional Time Series". In: *arXiv preprint arXiv:2201.02532* (2022).

[69] Fabian T Pfeffer et al. "Measuring wealth and wealth inequality: Comparing two US surveys". In: *Journal of economic and social measurement* 41.2 (2016), pp. 103–120.

[70] Thomas Piketty and Emmanuel Saez. "Income inequality in the United States, 1913–1998". In: *The Quarterly journal of economics* 118.1 (2003), pp. 1–41.

[71] Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. "Distributional national accounts: methods and estimates for the United States". In: *The Quarterly Journal of Economics* 133.2 (2018), pp. 553–609.

[72] Tommaso Proietti. "Temporal disaggregation by state space methods: Dynamic regression methods revisited". In: *The Econometrics Journal* 9.3 (2006), pp. 357–372.

[73] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer New York, 2005.

[74] Ludger Rüschendorf. "On the distributional transform, Sklar's theorem, and the empirical copula process". In: *Journal of statistical planning and inference* 139.11 (2009), pp. 3921–3927.

[75] Emmanuel Saez and Gabriel Zucman. "Wealth inequality in the United States since 1913: Evidence from capitalized income tax data". In: *The Quarterly Journal of Economics* 131.2 (2016), pp. 519–578.

[76] Eduardo Salazar and Martin Weale. "Monthly data and short-term forecasting: an assessment of monthly data in a VAR model". In: *Journal of Forecasting* 18.7 (1999), pp. 447–462.

[77] JMC Santos Silva and FN Cardoso. "The Chow-Lin method using dynamic models". In: *Economic modelling* 18.2 (2001), pp. 269–280.

[78] Jonathan Skinner. "A superior measure of consumption from the panel study of income dynamics". In: *Economics Letters* 23.2 (1987), pp. 213–216.

[79] Abe Sklar. "Vol. 8 of Fonctions de Répartition à n Dimensions et Leurs Marges, 229–231". In: *Paris: Publications de l'Institut de statistique de l'Université de Paris* (1959).

[80] Abe Sklar. "Random variables, joint distribution functions, and copulas". In: *Kybernetika* 9.6 (1973), pp. 449–460.

[81] Matthew Smith, Owen M Zidar, and Eric Zwick. *Top wealth in America: New estimates and implications for taxing the rich*. Tech. rep. National Bureau of Economic Research, 2021.

[82] Ruey S Tsay. "Some methods for analyzing big dependent data". In: *Journal of Business & Economic Statistics* 34.4 (2016), pp. 673–688.

A 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.

A.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. For the data cleaning, the data is aggregated to the household-level, at quarterly frequency.

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`).

A.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. Below we describe the concepts in turn.

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.

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.

A.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. Their 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.[20]

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. Albeit minor, the definition of liquid assets will vary between datasets, so careful attention here. Liquid assets for the PSID includes 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

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

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 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.

A.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.

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 4 of the previous calendar year.

A.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
	Comments	

Table 3: 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.

A.6 Aggregates

Together with the microdata, we specify a model component that captures the various aggregate shocks that buffer 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. This has various macro-data on industrial production, employment, housing, inventories, prices, earnings, productivity, household expectations, household balance sheets, interest rates, credit, etc. You can find more information here.

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 via the X13-ARIMA and follow closely the transformations 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 by three lags to form a data matrix of quadruple the size and then column-wise standardized. A PCA on this block of data identifies 21 orthogonal factors. The model estimation includes these factors as inputs Y_t. More on the selection of factors can be found in Appendix G.

B Minnesota Prior

The prior for the bayesian estimation is defined in block 18. In it, is the prior on $\mathbf{\Phi} \subset \boldsymbol{\theta}$, which consists of the parameters governing the state equation. To represent its uncertainty, the following Minnesota prior is proposed:

$$\begin{pmatrix} vec(A) \\ vec(B) \end{pmatrix} \sim \mathcal{MN}(\mu_{Minn}, V_{Minn}). \tag{20}$$

$$A_{ij} = \begin{cases} \kappa_5 & \textbf{for the first lag of the state variable, } i = j \\ 0 & \textbf{for the exogenous terms, } i \neq j \end{cases}$$

$$B = \mathbf{0}$$

$$V_{Minn,ii} = \begin{cases} \frac{\kappa_0}{l^{\kappa_4}} & \textbf{for own lags of the respective state variable } i \\ \frac{\kappa_1 \kappa_0}{l^{\kappa_4}} \frac{\hat{\sigma}_{ii}^2}{\hat{\sigma}_{jj}^2} & \textbf{for lags of the other state variables } j \\ \frac{\kappa_2 \kappa_0}{(l+1)^{\kappa_4}} \frac{\hat{\sigma}_{ii}^2}{\hat{\sigma}_{jj}^2} & \textbf{for exogenous terms} \\ \kappa_3 \kappa_0 \hat{\sigma}_{ii}^2 & \textbf{for deterministic terms} \end{cases} \tag{21}$$

where the prior distribution on A and B is a multivariate normal with μ_{Minn} the mean of the distribution and V_{Minn} the diagonal variance-covariance matrix. Governing the tightness variances are the set of hyperparameters $\{\kappa_i\}_{i=0}^{5}$. Table 4 covers each parameter in necessary detail.[21] $\hat{\sigma}_{ii}^2$ is the estimated variance of the residuals from a least squares estimation of estimated factor i on 4 lags, where factor i is estimated from the PCA and linearly interpolated after. For the deterministic terms, they are sample estimates of their variance.

Hyperparameter	Value	Purpose
κ_0	0.2	controls overall tightness of prior variances and prior variance of endogenous variable i's own lags
κ_1	0.3	the size of the prior variance of state variables, not corresponding to own lags
κ_2	0.001	the size of the prior variance of exogenous terms
κ_3	–	the size of the prior variance of deterministic terms
κ_4	2	the lag decay rate
κ_5	0.90	persistence of state LOM

Table 4: Minnesota Hyperparameters

C Distributional Factors

[21] It should be noted that κ_2 is set low since $\hat{\sigma}_{ii}^2$ is high and $\hat{\sigma}_{jj}^2$ is always equal to 1. κ_3 is not set since we do not have any deterministic/trend exogenous terms in the state equation.

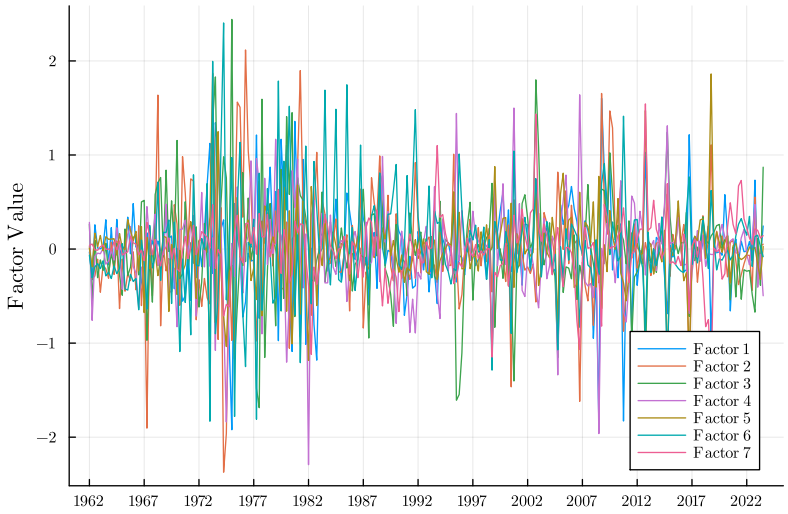


Figure 9: Distributional Factors

D Details on MCMC

To estimate the posterior distribution of the parameters and subsequently sample, we employ the DIME sampler from Boehl (2024). The sampler is particularly advantageous for dealing with potentially complex, high-dimensional, multi-modal posterior distributions, especially when these distributions have ex-ante unknown properties. Traditional MCMC methods often struggle with such distributions due to their reliance on gradient-based optimization or difficulties in converging efficiently. DIME addresses these issues by combining the strengths of global multi-start optimizers with the robustness of Monte Carlo methods, allowing it to explore the typical set of the posterior distribution more quickly and effectively. It also avoids the need for a pre-optimization step.

To initialize the sampler, I let an ensemble of $5n$ chains run for $4000 - 5000$ iterations, for n the size of the parameter vector. The last 500 are kept as the posterior distribution. There is a single tuning parameter χ that dictates (for each iteration and for each chain) the probability of mixture between the local and global transition kernel. We set $\chi = 0.1$, which means with 10% probability, we draw the global transition kernel.

Figure 10 shows the traces of the log-likelihood of all chains over the iterations. The plot clearly shows signs of convergence. The sampler implementation also returns the current log-weight on the history of the proposal distribution. The log-weight measures how much the current ensemble of MCMC samples influences the proposal distribution. Early in the sampling process, the log-weight should be greater than zero, indicating that the current samples strongly influence the proposal distribution, allowing it to adapt to the target distribution. As the sampling progresses and the chains begin to converge, the log-weight should be very close to zero, indicating that the influence of the current samples decreases and the proposal dis-

tribution stabilizes. After the $4000 - 5000$ iterations, the log-weight is always very close to 0 (around 9e-7 to be exact).

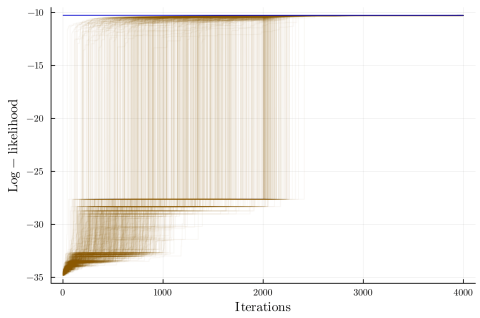
E Estimated Parameters

The parameters of the model are defined by the law of motion of states A, control variables B, the variance covariance matrix of the state process noise Ω and the measurement noise matrix Δ. The resulting parameter vector $\boldsymbol{\theta}$ from the estimated baseline model of consumption, income and wealth is summarized below:

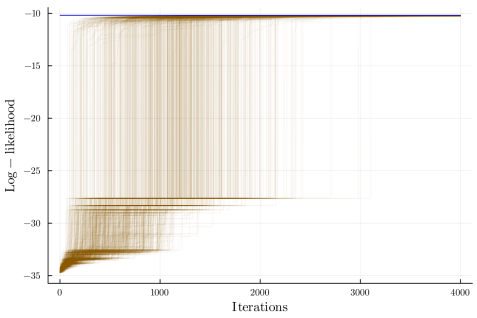
$$A = \begin{bmatrix} -0.172138 & 0.104549 & 0.0375995 & -0.0365716 & 0.0517868 & 0.164601 & 0.169162 \\ -0.198363 & 0.0220448 & 0.0318587 & -0.0755722 & 0.583719 & -0.439327 & 0.179461 \\ -0.232617 & 0.159984 & -0.299377 & -0.312839 & 0.439963 & 0.424342 & 0.405313 \\ 0.122183 & 0.0777094 & 0.0076737 & -0.136233 & 0.0410441 & -0.3757 & -0.469405 \\ 0.090222 & 0.0487302 & -0.035005 & -0.120615 & -0.090295 & -0.296449 & -0.205093 \\ 0.0707263 & -0.0404629 & -0.0218973 & -0.0288991 & 0.13403 & -0.207689 & -0.0808181 \\ -0.0244344 & -0.0121677 & -0.00627248 & 0.0479002 & 0.0182849 & -0.137255 & 0.235568 \end{bmatrix}$$

$$B' = \begin{bmatrix} -0.120962 & -0.0106441 & -0.0429657 & -0.19099 & -0.072416 & 0.127412 & 0.0309303 \\ 0.0416033 & -0.191426 & 0.0890118 & -0.187741 & 0.0113039 & -0.0452964 & -0.0154524 \\ -0.0729248 & 0.0743314 & -0.0351416 & 0.070014 & -0.0580504 & 0.0233263 & 0.00293904 \\ 0.131014 & 0.0758711 & -0.0756035 & 0.0527423 & 0.00081871 & -0.0358998 & 0.00104553 \\ 0.127058 & -0.0273002 & 0.156481 & 0.00843965 & -0.0288098 & -0.0133379 & 0.0103688 \\ 0.0493484 & -0.117778 & 0.01163 & 0.0471963 & -0.0839208 & -0.0387515 & 0.00341004 \\ 0.0231236 & -0.162675 & -0.309534 & -0.140972 & -0.00896298 & -0.0111249 & 0.0111561 \\ -0.133706 & 0.0205727 & -0.0639536 & -0.00693861 & -0.00412038 & -0.0241852 & -0.0270461 \\ 0.0537212 & 0.0539921 & 0.125736 & -0.0690668 & -0.0046014 & 0.0595303 & 0.0261332 \\ 0.177312 & -0.137346 & -0.0582303 & 0.146275 & 0.0111653 & 0.00545539 & 0.0247737 \\ 0.250165 & 0.00894873 & 0.196468 & -0.0737448 & 0.0696563 & -0.051642 & -0.0159294 \\ -0.0669459 & 0.0617581 & 0.176292 & 0.131503 & -0.0221056 & 0.0372529 & -0.0314296 \\ 0.142072 & 0.0840433 & 0.0536954 & -0.122196 & 0.0346887 & -0.0791965 & 0.036459 \\ 0.12649 & 0.00719191 & -0.136866 & 0.0114966 & -0.0452859 & -0.00935734 & -0.0216441 \\ 0.0426566 & 0.0498708 & 0.0383232 & -0.0772039 & -0.0111427 & -0.0114288 & -0.0364814 \\ 0.0721697 & 0.244335 & 0.0769378 & -0.106185 & -0.0292045 & 0.112148 & 0.0281219 \\ -0.0649757 & 0.0215878 & -0.00336058 & -0.140149 & 0.00254169 & -0.0612346 & -0.00957795 \\ 0.0416246 & -0.0187523 & -0.0779292 & 0.154285 & 0.0265518 & 0.105266 & 0.00174256 \\ -0.17362 & -0.286801 & 0.205799 & -0.162271 & 0.0335973 & -0.020848 & 0.0336134 \\ 0.242736 & -0.130083 & 0.116346 & 0.0825682 & 0.136489 & -0.00618597 & -0.0226616 \\ -0.146157 & -0.00539771 & 0.0554425 & 0.00856631 & -0.101038 & 0.0165377 & -0.0332974 \end{bmatrix}$$

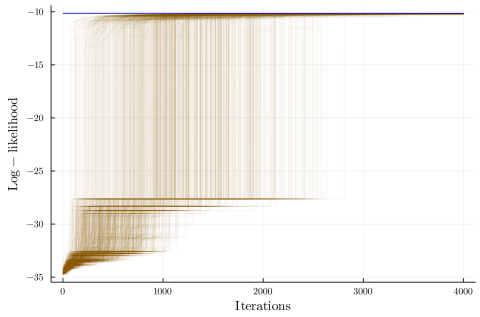
Figure 10: Converging Chains



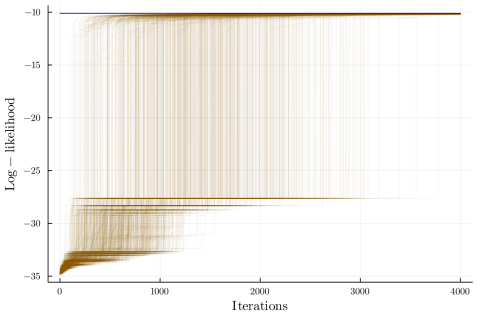
(a) Baseline model



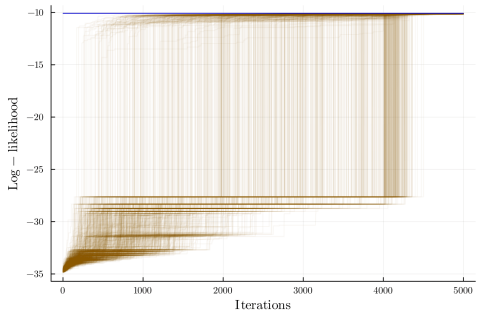
(b) Excluding housing cycle



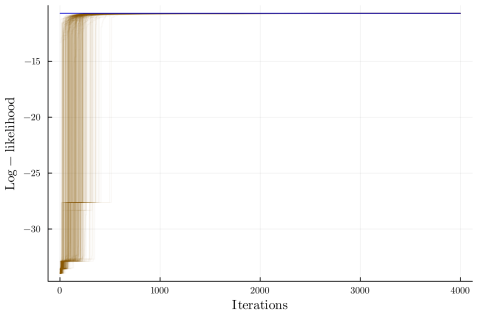
(c) Excluding recent 20 quarters



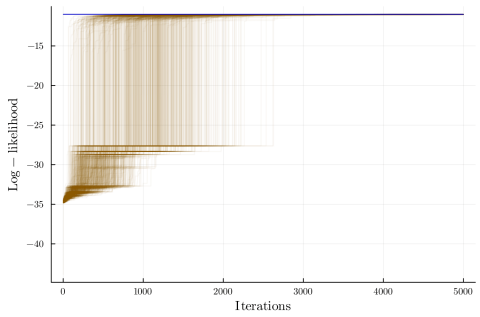
(d) CEX every 4 years



(e) With additional factors



(f) With less factors



(g) Higher order

Notes: Figure shows, for each model, the evolution of the ensemble of chains in terms of log-likelihood. Log-likelihood values are scaled to ensure visualization of the chains' evolution. Last 500 draws are kept as samples of the posterior. Refer to the text for the different model specifications.

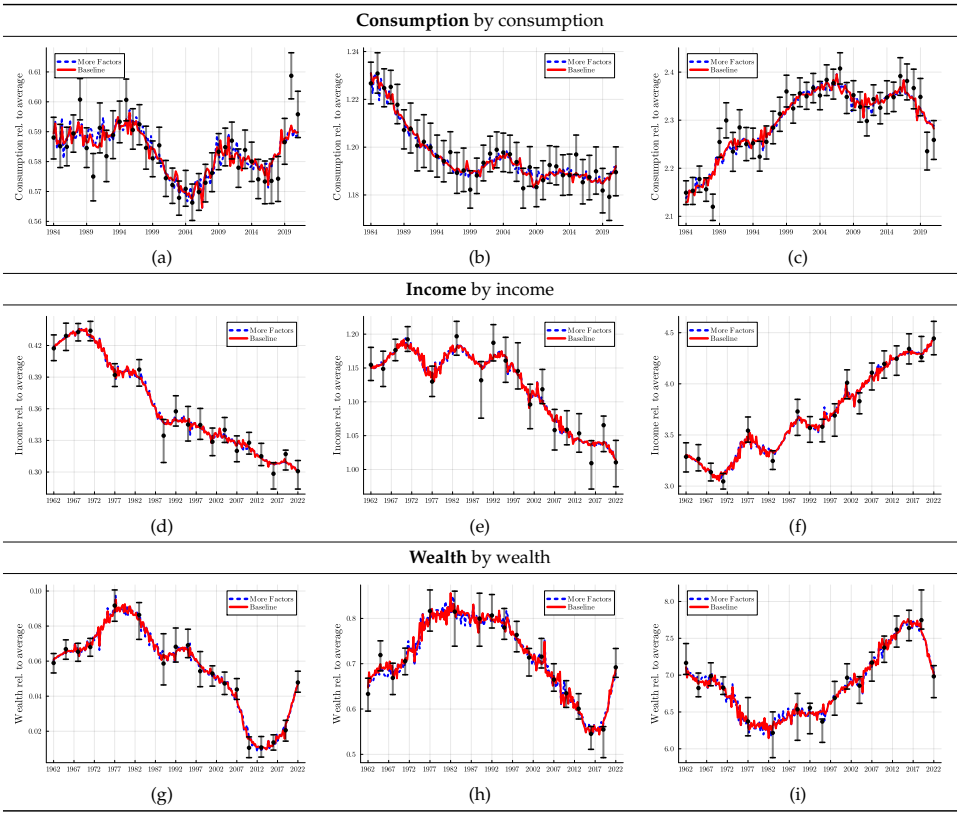
$$diag(\Omega) = \begin{pmatrix} 0.8959743540555566 \\ 1.7705637242549592 \\ 6.3263383206429245 \\ 1.3020308461348842 \\ 0.5291877124871194 \\ 1.5024039261214224 \\ 0.23863667802303593 \end{pmatrix} \qquad diag(\Delta) = \begin{pmatrix} 1.6024590866261839 \\ 4.337758003802917 \\ -0.22295262182858905 \\ 14.715094101556144 \\ 11.405188534807998 \\ -2.8324599266551034 \\ -3.001686737483659 \\ 0.8808642685742953 \\ -2.830542756560333 \\ -0.6929230805702409 \\ 1.9290834527792229 \\ -1.5225502172616414 \\ 1.607191251281862 \\ 30.840680139761012 \\ 2.7987305685544297 \\ 0.7685929868640887 \\ 8.792026736855282 \\ 10.254048620164346 \end{pmatrix}$$

F Additional Factors

There may be concerns that the model is misspecified and that the distribution of consumption, income and wealth are driven by more factors. These additional factors may carry important information on the distributional dynamics *orthogonal* to the other factors and aggre-

gate information. To this, instead of retaining the factors that represent 95% of the microdata variation, we use 99% as our cutoff. Figure 11 compares the baseline estimates to the model estimates with more factors. Clearly, model estimates are unchanged from the addition of more factors.

Figure 11: Estimates unchanged with additional factors



Notes: See Figure 6 for further details.

G Factor Selection

The estimation of the joint distribution of consumption, income, and wealth necessitates samples of this distribution and a comprehensive set of macroeconomic data to inform its dynamics. Macroeconomic theory and empirical validation suggest that both the distribution and business cycle fluctuations are driven by a smaller set of underlying factors. Consequently, a significant area of macro-econometric research has produced several estimators to determine which of these factors to retain and how many factors are necessary.[22] For our setting, it is additionally crucial to motivate the macroeconomic data we project, since, alongside the microdata, this will determine the cyclicality of the distributional data. This interdependency is further elaborated in the following section.

[22]For references, see Bai and Ng (2002), Ahn and Horenstein (2013), Bai and Ng (2019), Gagliardini, Ossola, and Scaillet (2019), and Freyaldenhoven (2022).

G.1 Choice of Factor Representation

The goal of factor decomposition of distributional data is to inform the size of the state-space model and ensure that these factors can accurately reconstruct the cyclical movements observed in the data. As illustrated in Figure 3, the factor decomposition effectively reconstructs the data with minimal to no information loss. Consequently, we remain agnostic about the specific factors retained, opting to retain enough factors to replicate the data on average. Factors may consist of components that explain general or local movements within the distribution (Freyaldenhoven, 2022), possess eigenvalues greater than or less than 1, or induce weak to no cross-correlation in the unretained factors.[23] This observation underscores that solely retaining factors that explain common movements, have eigenvalues greater than 1, or focus on specific subsets of the factor space may inadequately capture the heterogeneity-rich cyclicality of consumption, income, and wealth, which is paramount in this study. Figure XX shows precisely this.

For the estimation of business cycle fluctuations, the selection of macroeconomic data must account for the rich heterogeneity present in the distributional data. The conventional approach to estimating business cycle fluctuations relies on the FRED-QD dataset—a common starting point of 200+ time series for exploratory factor analysis in macroeconomics. Many studies will then estimate the common component of these macroeconomic time series, consisting of a set of factors and their respective loadings, and define it as the most relevant movements in the macroeconomy. For a given estimator, the number of factors from projecting the FRED-QD dataset will vary and explain around $40 - 50\%$ of the (summable) data variation.

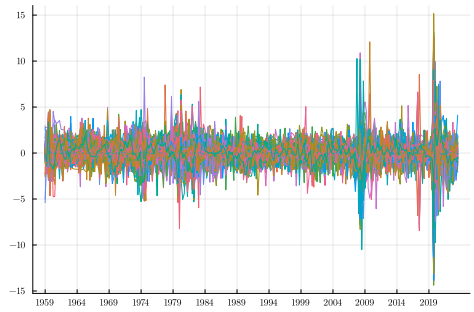
We adopt this approach, but use a more conservative estimator, which augments the normal estimated set of factors with *local* factors (Freyaldenhoven, 2022). These local factors only explain a subset of the data, but carry nonetheless relatively large loadings. This approach would capture the most pervasive business cycle fluctuations as well as the granular movements, both potentially necessary to explain the income and wealth movements relevant for consumption dynamics. Using this estimator, we find that six factors are sufficient to explain the cyclical movement in the aggregate data (Panel (a)).

However, to inform the dynamics of the distributional factors, we specify four quarters of aggregate information. Figure 12, Panel (b) presents the factors from performing the PCA on these four quarters of data (over 1000 time series). Figure 12, Panel (c) plots the unweighted eigenvalues $\hat{\Upsilon}_k^0$ and the eigenvalues accounting for the contribution of the loadings $\hat{\Upsilon}_k^2$. Panel (d) plots $\hat{S}^2$, which measures how concentrated the corresponding eigenvector is on its z largest entries. Taking into account these two plots, we find that around 10 factors are suf-

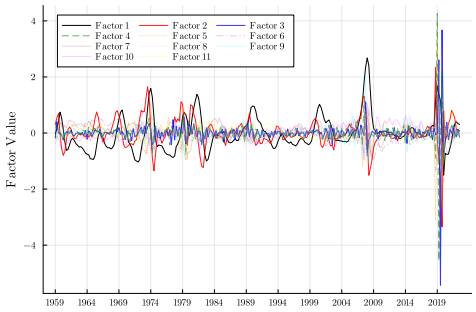
[23] Our decomposition of the joint distribution into its correlational structure and marginal distributions implies the potential existence of local factors. Any local change in a bona fide distribution inherently represents a global change.

ficient to explain the aggregate data. It is around 10 factors that $\hat{\Upsilon}_k^2 < \hat{\Upsilon}_k^0$ for the first time. We ultimately settle on 21 factors, however, since it is around this point that the concentration completely decreases .

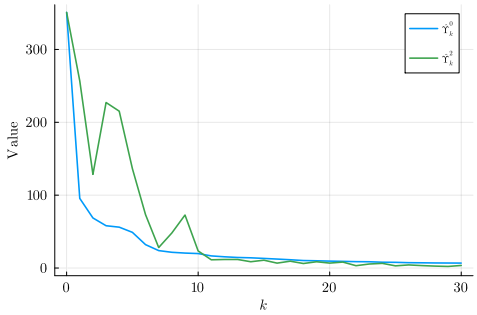
Figure 12: Eigenvalue Analysis



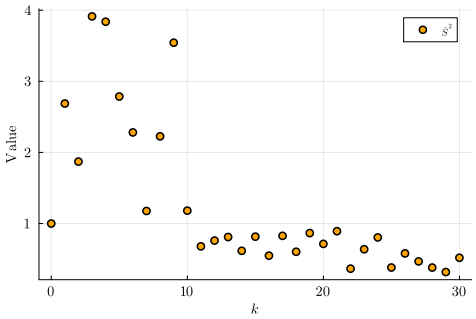
(a) Stationary aggregates



(b) Factors from 4 Lags of Stationary Aggregates



(c) Eigenvalues



(d) Concentration

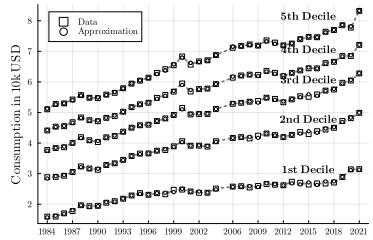
Notes: Figure shows an overview of the aggregate data. Panel (a) shows each aggregate series reduced to its stationary component. Panel (b) are the factors from 4 lags of stationary aggregate data. Panel (c) are the resulting eigenvalues weighted by the respective eigenvector loadings. Panel (d) are the weight contributions to the eigenvalues based on eigenvector loadings. Data used are from FRED-QD 1959Q1 to 2024Q1.

H Less Factors (fix consumption labels)

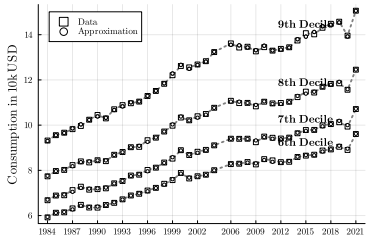
What is lost? Marginals and correlational structure seem to be well represented by fewer factors.

Figure 13: Comparison of quantile functions in raw and approximated data (only 3 factors)

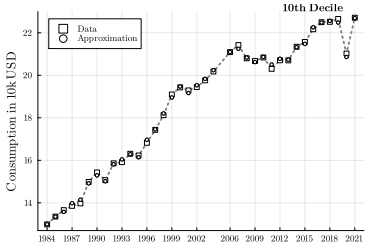
Mean Consumption



(a) 1st to 5th decile

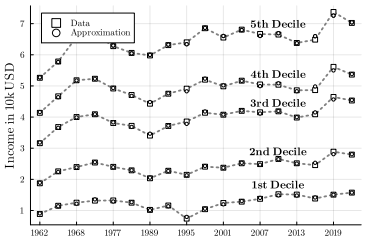


(b) 6th to 9th decile

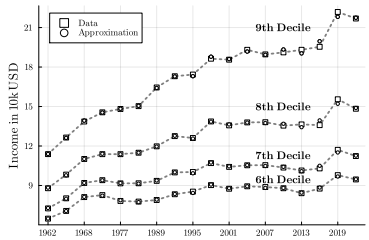


(c) top decile

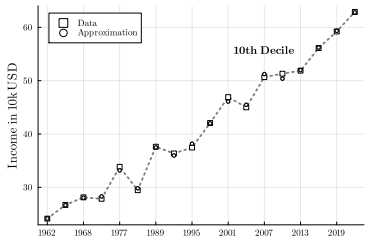
Mean Income



(d) 1st to 5th decile

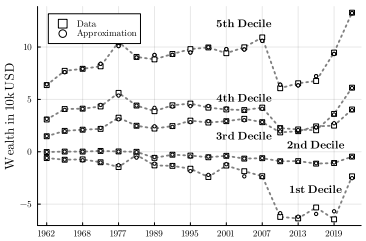


(e) 6th to 9th decile

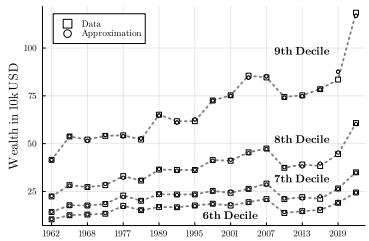


(f) top decile

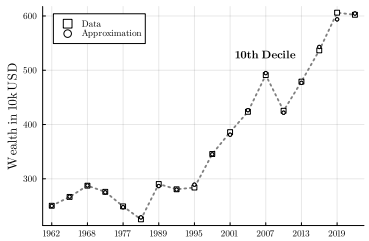
Mean Wealth



(g) 1st to 5th decile



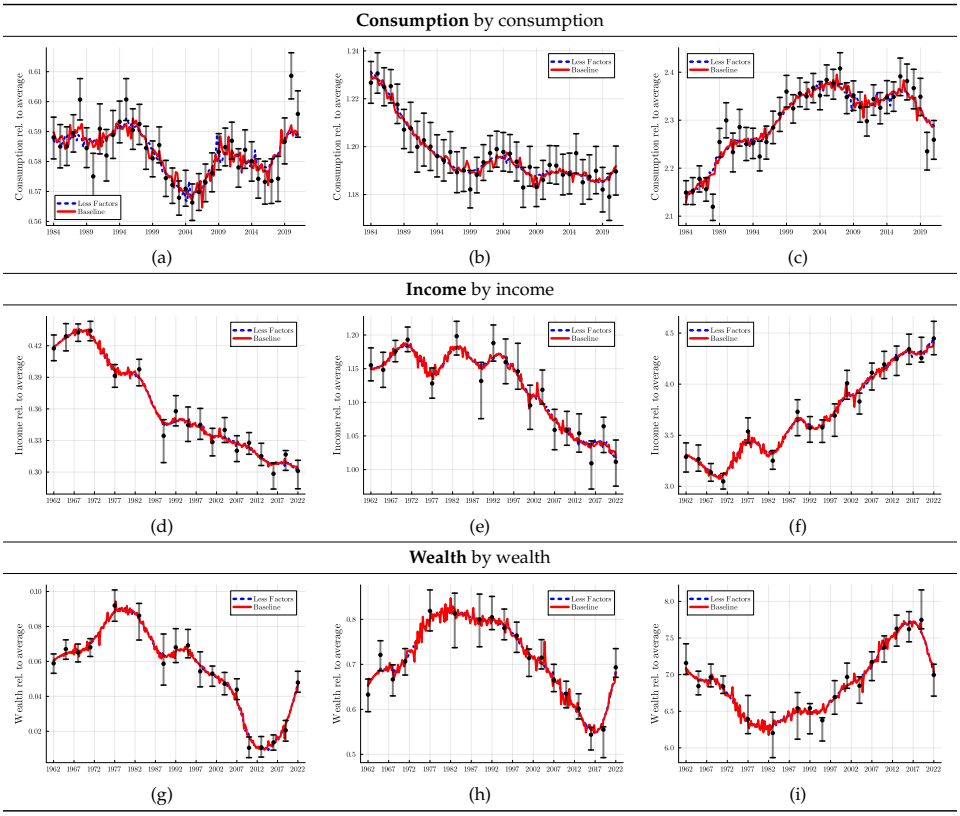
(h) 6th to 9th decile



(i) top decile

Notes: See Figure 3 for details.

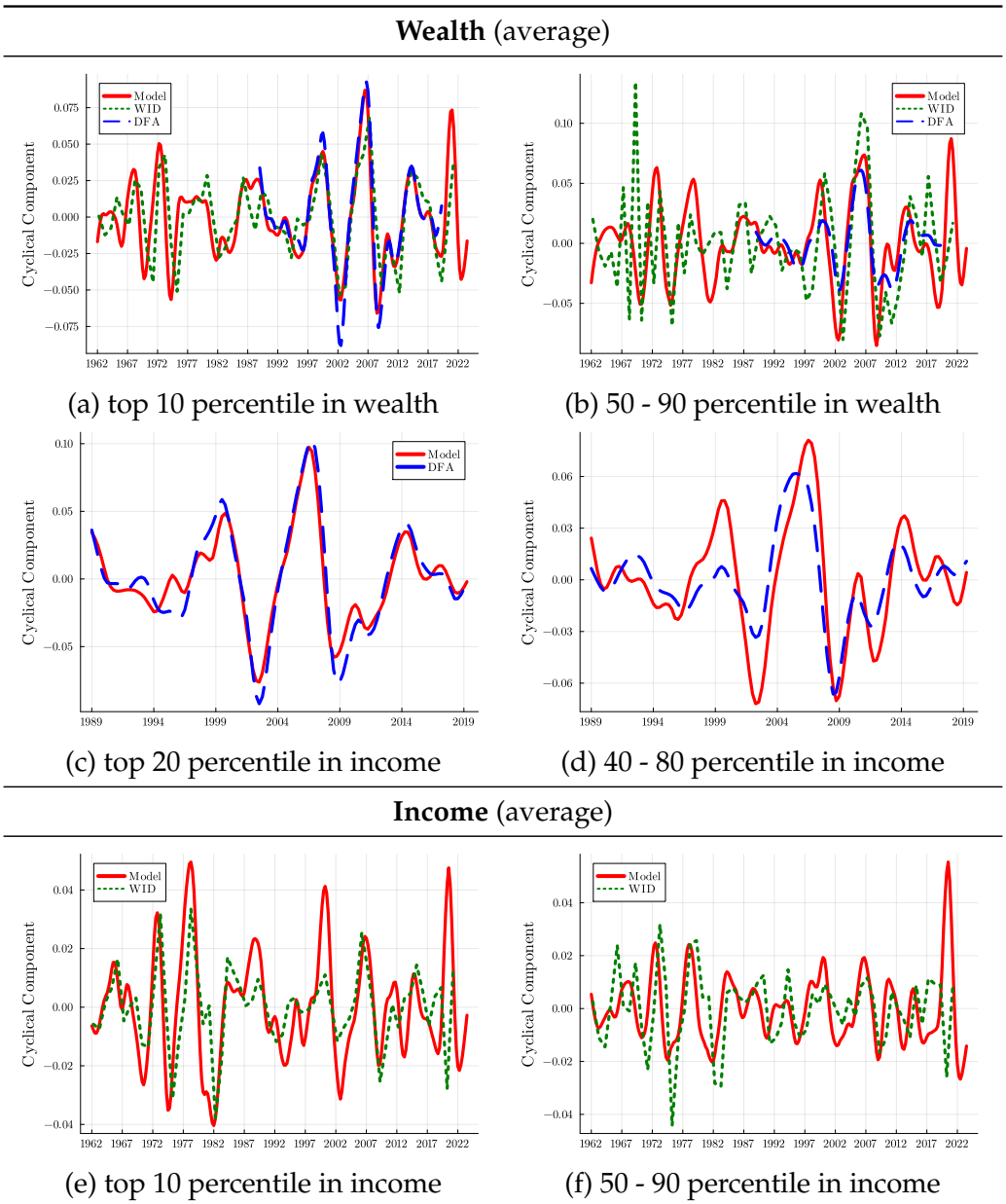
Figure 14: Estimates unchanged with additional factors



Notes: See Figure 6 for further details.

I Do results change with higher order?

Figure 15: Comparison of cyclical component of distributional data to external sources



Notes: Figure presents model implied estimates from a model of higher order. Refer to Figure 7 for remaining notes.