Identifying Macro Shocks From Micro Evidence: A Mixed Autoregressive Approach

Naoya Nagasaka*

October 6, 2025
Click here for the latest version.

Abstract

This paper develops a methodology to identify aggregate shocks by employing heterogeneous direct (partial equilibrium) effects estimated from microeconometric research designs. The total effect of a shock consists of direct and indirect (general equilibrium) effects, but microeconometric tools typically do not capture the latter. Our framework builds on a time-series econometric model that incorporates both aggregate variables and functional observations, such as cross-sectional densities. We show how direct effects can serve as identification restrictions to evaluate the total effect. We illustrate our approach by comparing the effects of lump-sum and targeted stimulus transfer policies on aggregate outcomes and consumption inequality.

JEL Classification Code: C32, C50, E60.
Keywords: Direct and Indirect Effects, Vector Autoregression, Functional Data Analysis, Bayesian Econometrics.

*Indiana University (Email: naonagas@iu.edu). I am grateful to Yoosoon Chang, Bulent Guler, Rupal Kamdar, Christian Matthes, and Joon Park for their continued guidance and support. I also thank Ippei Fujiwara and Tatsuro Senga for insightful comments.

1 Introduction

Among empirical macroeconomists, it has become quite common to employ micro data combined with microeconometric identification strategies in order to draw macroeconomic implications. This popularity is partly owing to difficulty in extracting exogenous variation from aggregate data alone: Researchers have to find reasonable natural experiments, or impose identification restrictions that are often hard to justify. Micro datasets allow them to exploit cross-sectional variation to learn first-round propagation of macroeconomic shocks to cross-sectional units. Such propagation is often heterogeneous; agents are exposed to and respond to shocks differently depending on their own characteristics.

However, the parameters estimated from the microeconometric regressions are not directly informative on aggregate consequence of shocks. The total effect of a macro shock consists of direct (partial equilibrium) and indirect (general equilibrium) effects, where the former captures first-round responses of agents to shocks and the latter reflects feedback from changes in aggregate quantities, such as prices, to agents' decisions. Microeconometric identification strategies are typically meant to capture only direct effects, and indirect effects are omitted by including time-fixed effects. As a conventional way to recover the indirect effect, researchers construct fully-specified general equilibrium models calibrated to match estimates from the regressions. The conclusions drawn from such exercises depend heavily on the specification of the general equilibrium models.

This paper develops a novel approach for investigating total effects of aggregate shocks by exploiting information on direct effects. The key ingredients for our method are (i) heterogeneity in direct effects identified through microeconometric research designs and (ii) observations of cross-sectional densities of micro variables over time. We combine these elements to identify aggregate shocks in a mixed autoregression (MAR), an autoregressive model featuring both aggregate and functional observations. The functional data not only works as one of the primary inputs for identification, but also enables the analysis of the effects of identified shocks on inequality.

We motivate our method by a stylized representation of dynamic heterogeneous-agent problems, accommodating those in most heterogeneous-agent macro models. We consider an aggregate shock perturbing agents' decisions, such as a transfer stimulus payment. We show that the impact of the shock on the current cross-sectional density is defined as a sum of direct and indirect effects. The direct effect reflects the first-round response of agents to the shock, and the indirect effect captures the feedback from aggregate quantities. The indirect effect reflects propagation from current aggregates, and importantly, also from future aggregates

due to agents' forward looking behavior. The individual-level regression with coefficients depending on individual characteristics captures the heterogeneous direct responses to the shock. This information allows us to construct the direct effect on the cross-sectional density, which is a primary source for our identification strategy. On the other hand, the indirect effect is taken up by the time-fixed effect.

Our discussion turns to the autoregressive model with both aggregates and the cross-sectional density as endogenous variables. The prototypical autoregression does not include future aggregates because the autoregressive structure itself captures the dynamic propagation. However, the omission of future aggregates has important consequences in our application: Since the indirect effect depends also on future aggregates, direct and indirect effects are compounded in the autoregressive framework. We cannot disentangle the compound effect by the information from the individual-level regression alone because it is silent on the indirect effect. We solve this problem by imposing a structure on how the indirect effect diminishes over time, and compute an approximated indirect effect based on this structure. We expect that current aggregates matter more than future aggregates for the current density, and we specify how such diminishing occurs. We show, in a quantitative heterogeneous agent New Keynesian (HANK) model featuring several frictions, that even a simple exponential diminishing structure provides a reasonable approximation.

We incorporate our identification scheme into a MAR model consisting of aggregate and functional observations. In particular, we achieve identification by placing prior restrictions on structural parameters rather than fixing them at particular values (Baumeister and Hamilton, 2015). Our identification strategy is imperfect because (i) there is uncertainty due to estimation of direct effects, and (ii) we recover indirect effects based on a certain assumption about how they diminish. We acknowledge possible errors in the identification conditions stemming from these two forces, and reflect them in terms of prior uncertainty. As another advantage of this methodology, we can incorporate additional prior beliefs on how the shock propagates, such as sign restrictions (e.g., a certain aggregate variable increases/ decreases in response to the shock).

Another challenge in taking our methodology to the data is to handle cross-sectional densities. Such functional observations are inherently infinite-dimensional, requiring a dimension-reduction method for estimation. We leverage recent development of functional data analysis, and approximate functional time-series with a finite number of orthonormal basis functions. This approach allows us to represent an approximated MAR as a vector autoregression (VAR) with aggregate variables and coefficients associated with the basis functions so that existing

estimation methods (such as OLS, maximum likelihood, and Bayesian estimation) are applicable. The choice of basis functions is crucial for the performance of the approximation scheme. We follow the recommendation by Y. Chang et al. (2024b) and choose the basis functions in a data-driven way, namely the functional principal component (FPC) basis. The FPC basis is known to explain temporal variation in functional variables better than any other orthonormal basis. This optimal property implies that the FPC summarizes information in functional variables effectively, helping avoid the system getting unnecessarily large.

We evaluate our identification methodology using the medium-scale HANK model with liquid and illiquid assets. The direct effect of the lump-sum transfer shock on the consumption density implied by the model is qualitatively consistent with the empirical counterpart, and thus the model serves as a useful laboratory to check the validity of our methodology before taking it to the data. The model brings the following insights. First, the influence of aggregate variables k periods into the future on the consumption density decays exponentially and monotonically as k grows, across the entire domain of the density. This observation supports the structure used to approximate indirect effects. Second, our identification methodology succeeds in recovering the true impulse responses both qualitatively and quantitatively. Third, information on direct effects reduces the uncertainty surrounding the identification significantly. In particular, the length of the 68% credible interval of the at-impact response of output is shrinks by 36 percentage points once we take into account the information from direct effects. Moreover, the interval becomes more concentrated in the positive region, even though we place a symmetric prior for the output response.

We apply the proposed identification methodology to investigate the macroeconomic and distributional consequences of stimulus transfer policies in the United States. With our framework, we compare two types of policies: (i) a lump-sum cash transfer policy in which each household receives \$100 per family member, and (ii) a targeted cash transfer policy in which households in the bottom 20% of the income distribution receive \$500 per family member. We find that these two policies have only small positive effects on output, an observation consistent with the evidence that cash transfers are not effective tools for stimulating the macroeconomy (e.g., Ramey 2025). On the other hand, we observe a persistent reduction in consumption inequality, measured by the Gini coefficient. This reduction is larger under the targeted policy. Moreover, the total effect on the Gini coefficient is larger in magnitude than the direct effect, suggesting that indirect channels amplify the reduction in consumption inequality. These findings highlight the importance of general equilibrium mechanisms

in assessing the distributional consequences of aggregate shocks.

Literature. This paper contributes to the literature that leverages microeconometric tools to investigate research questions at the aggregate level. Regressions based on microdata cannot be applied directly to derive aggregate implications because time-fixed effects absorb indirect feedback—known as the "missing intercept" problem. A typical response to this problem is to recover indirect effects from fully specified general equilibrium models calibrated with parameters estimated from micro regressions (e.g., Nakamura and Steinsson 2014). The implied total effects depend heavily on the details of the structural models.

Several papers propose methods to address the missing intercept problem without resorting to specific quantitative general equilibrium models (Chodorow-Reich 2019; Herreno, 2023; K. Huber 2023; Matthes et al. 2024; Sarto, 2025; Wolf 2023). Wolf (2023) shows that, under stylized heterogeneous-agent models satisfying certain assumptions, one can construct an approximation of the indirect effect using the aggregate fiscal multiplier. We provide a solution to a similar question, but rather than relying on a general equilibrium model as a theoretical background, we instead exploit heterogeneity in direct effects as a primary source of identification. Interestingly, our framework also implies the relevance of aggregate shocks other than the shock of interest in recovering the indirect effect, the observation echoing his approximation methodology.

Matthes et al. (2024) extend the factor model in Matthes and Schwartzman (Forthcoming) to develop an identification scheme relying on heterogeneous exposure of economic units to aggregate shocks.[1] Their framework requires a sufficiently long panel of units recorded frequently enough. Such data are available at the aggregate level (countries) or semi-aggregate level (regions and sectors), but are rarely available at the micro level, especially for households.[2] We use functional variables rather than panel data, and thus our method is feasible with repeated cross-sectional micro observations. That being said, since their methodology has its own advantages relative to ours (e.g., it can accommodate multiple unit-level variables, such as local output and government spending; it can combine numerous weak restrictions to achieve tight identification), we regard our method as a complement to theirs.

[1] Sarto (2025) also uses a factor model to solve the missing intercept problem. Instead of heterogeneous exposures as in Matthes et al. (2024), his identification conditions involve exclusion restrictions.

[2] The Survey of Consumer Finance contains detailed information on household balance sheet, but the data is collected triennially and does not exhibit a panel structure. The Survey of Consumer Expenditure collects the data each quarter, while it keeps track of the same household for five consecutive quarters at most. The Panel Study of Income Dynamics does have a panel structure and is available from 1968, although the data is collected once every two years, which makes it hard to apply time-series econometric tools.

The empirical framework in this paper treats a time-series of cross-sectional densities as an endogenous variable. There is a growing literature on incorporating such functional observations into econometric models to study the effects of macroeconomic shocks on inequality (Y. Chang et al. 2025; M. Chang et al. 2024; M. Chang and Schorfheide 2024; F. Huber et al. 2024), yield curve (Y. Chang et al. 2023; Inoue and Rossi 2021), heterogeneity in expectations (Y. Chang et al. 2022; Meeks and Monti 2023), and climate change (Y. Chang et al. 2024a). The novelty of this paper relative to those works lies in developing a new scheme for identifying structural shocks from functional observations with the help of microeconometrically identified parameters. To obtain the finite-dimensional approximation of the functional observations, we rely in particular on the functional principal component (FPC) basis, whose theoretical properties have been well studied by, for example, Bosq (2000), Ramsay and Silverman (2005), Mas (2007), and Y. Chang et al. (2024b).

Outline. The rest of this paper is organized as follows. Section 2 outlines our identification scheme. Section 3 introduces the statistical model and discusses the dimension-reduction method and prior specifications. Section 4 validates the identification approach using the quantitative HANK model. We provide an application of our method and investigate the transfer stimulus policy in Section 5. Section 6 concludes.

2 Background for Identification

To illustrate the main idea for identification, we consider a stylized representation of heterogeneous-agent problems. We assume that there is a single aggregate shock ε_t which affects agents' decisions[3]. We are interested in how the shock propagates through the distribution of the endogenously determined idiosyncratic variable c. The measure of agents at time t is denoted by μ_t, and the value function is denoted by v_t. We suppose that the measure and value function evolve according to

$$\mu_{t+1} = \Lambda\left(\mu_t, v_t, X_t\right) \tag{1}$$

$$v_t = V\left(v_{t+1}, X_t, \varepsilon_t\right) \tag{2}$$

[3]We interpret this shock as an MIT shock (i.e., a one-time unexpected disturbance) and explore the perfect-foresight dynamics in response to the shock. Our discussion is built on linearized economies where the certainty equivalence holds. Under the certainty equivalence, the impulse response from the state space representation coincides with the perfect foresight transition following the MIT shock (Boppart et al., 2018).

where $X_t \in \mathbb{R}^k$ is a vector of aggregate inputs for the individual problem (e.g., prices). This specification generalizes the representation of heterogeneous-agent problems in Auclert et al. (2021) by treating μ_t and v_t as functions, rather than vectors of values at discrete grid points. Most dynamic heterogeneous-agent models exhibit this structure.

Let f_t denote the density of c. The future value function v_{t+1} and current inputs X_t determine the current policy function, which in turn, combined with the measure μ_t, forms f_t. Given equations (1) and (2), the density f_t is written as a function of the measure μ_t and the sequences $(\varepsilon_{t+h})_{h\geqslant 0}$ and $(X_{t+h})_{h\geqslant 0}$.

$$f_t = F\left(\mu_t, \varepsilon_t, \varepsilon_{t+1}, \cdots, X_t, X_{t+1}, \cdots\right) \tag{3}$$

Note that in general equilibrium, ε_t influences current and future aggregate inputs $\{X_{t+h}\}_{h\geqslant 0}$. We should account for the propagation through those inputs when considering the effect of ε_t. By contract, μ_t is not affected by ε_t because μ_t is predetermined at period $t-1$. Then equation (3) implies that the response of f_t to the shock ε_t is computed as the sum of direct and indirect effects:

$$\frac{df_t}{d\varepsilon_t} = \underbrace{F_\varepsilon}_{\text{Direct Effect}} + \underbrace{F_0\frac{dX_t}{d\varepsilon_t} + F_1\frac{dX_{t+1}}{d\varepsilon_t} + \cdots}_{\text{Indirect Effect}} \tag{4}$$

where F_ε is the Fréchet derivative of F with respect to ε, and F_h $(h = 0, 1, 2, \cdots)$ is the Fréchet derivative of F with respect to X_{t+h}. The first term, labeled the direct effect, captures the first-round effect (i.e., how the shock itself influences f_t). The indirect effect is captured by the second term and the subsequent ones, reflecting how the shock propagates through current and future aggregate inputs. As a sum of these two, the total effect $\frac{df_t}{d\varepsilon_t}$ is the at-impact (contemporaneous) impulse response of f_t to ε_t.

Example 1. Consider an economy with aggregate and idiosyncratic uncertainty similar to the setup in Krusell and Smith (1998). Household $i \in [0, 1]$ is endowed with asset $a_{i,0}$ and productivity $e_{i,0}$, and solves the intertemporal optimization problem:

$$\max \sum_{t=0}^{\infty} \beta^t u(c_{i,t})$$

subject to

$$\begin{aligned} c_{i,t} + a_{i,t+1} &= (1-\tau_t)w_t e_{i,t} + (1+r_t)a_{i,t} + \eta(a_{i,t}, e_{i,t})\varepsilon_t \\ a_{i,t+1} &\geqslant 0 \\ e_{i,t} &\text{ follows a Markov process } P(e' \mid e) \end{aligned} \tag{5}$$

where ε_t is an exogenous aggregate shock interpreted as an aggregate transfer policy. The coefficient of ε_t is a function $\eta(\cdot)$, reflecting the sensitivity of household income to the shock. This sensitivity function $\eta(\cdot)$ can be determined by the policymaker. The amount of transfer received by household i is given by $\eta(\cdot)\varepsilon_t$. When $\eta(\cdot)$ is a constant function, ε_t is a shock to the lump-sum payment. In this economy, the aggregate inputs for the consumers' decisions are $X_t = (\tau_t, w_t, r_t)'$. The value function $v_t(a, e)$ is determined as

$$v_t(a,e) = \max\left\{u(c) + \beta \int P(e' \mid e) v_{t+1}(a', e') de'\right\}$$

subject to the constraints (5). The right-hand side defines the function $V(\cdot)$ in equation (2). The measure of state variables $\mu_t(a, e)$ evolves as follows:

$$\mu_{t+1}(\mathcal{A}, \mathcal{E}) = \int P(e' \in \mathcal{E} \mid e) \mathbf{1}\left\{a_{t+1}(a,e) \in \mathcal{A}\right\} d\mu_t(a,e)$$

where $\mathcal{A}, \mathcal{E} \in \sigma(\mathbb{R})$ are measurable subsets of the asset and productivity spaces, and $a_{t+1}(\cdot)$ denotes the policy function for assets. The policy function is computed jointly with the value function and thus depends on v_{t+1}, X_t, and ε_t. The right-hand side gives the law of motion $\Lambda(\cdot)$ in equation (1).

The sequence of aggregate inputs $\{X_t\}_{t \geqslant 0}$ is influenced by ε_t through the general equilibrium propagation. For example, households adjust their consumption, savings, and labor supply in response to the shock, which influences the demand for consumption goods and the supply of production inputs. This has an effect on prices, which in turn affects household behavior. In addition, since agents are forward-looking, not only current prices but also future prices matter for current decisions via their effect on v_{t+1}.

2.1 Recovering Direct Effect from Micro Evidence

The central element for shock identification is the direct effect F_ε. Given knowledge of the direct response to the shock for each agent, we can compute how the distribution of c responds

to ε_t abstracting from changes in aggregate inputs. To see this, we consider the following individual-level regression, driven by the linearized policy function.[4]

$$c_{i,t} - c_{i,ss} = \underbrace{\phi(s_{i,ss})}_{\phi_i} \times (\eta_i \varepsilon_t) + \gamma_t + u_{i,t} \tag{6}$$

where γ_t denotes a time fixed effect. If the effect from the sequence $(X_t, X_{t+1}, \cdots)$ (i.e., indirect effect) is common for every individual, it is removed by the time fixed effect γ_t. The explanatory variable $\eta_i \varepsilon_t$ represents how each individual is affected by the shock. In the context of transfer stimulus, it is the payment that agent i receives at time t. The coefficient ϕ is interpreted as the marginal propensity to consume (MPC).

Importantly, we allow ϕ to be individual-dependent by specifying it as a function of pre-shock individual characteristics $s_{i,ss}$. For example, standard incomplete market models imply that households facing borrowing constraints exhibit larger MPCs than unconstrained households. Indeed, the empirical literature shows that financial characteristics (e.g., participation in the credit market and holdings of liquid and illiquid assets) are important sources of MPC heterogeneity, and other characteristics (e.g., impatience) are also relevant.[5] Econometricians typically specify the functional form of $\phi(s)$ (e.g., a parametric function of s or by grouping individuals based on s), while the model (6) is general enough to accommodate nonparametric approaches, including functional coefficient models.[6]

The regression (6) yields estimates of $(\phi_i)_i$, which we then use to construct the perturbed density under a unit shock. This leads directly to the density-based object F_ε. To a first order approximation, F_ε is written as

$$F_\varepsilon \approx F(1, \mu_{ss}, X_{ss}, X_{ss}, \cdots) - F(0, \mu_{ss}, X_{ss}, X_{ss}, \cdots)$$

We take the difference between two consumption densities. First, $F(0, \mu_{ss}, X_{ss}, X_{ss}, \cdots)$ is the steady state density. Second, $F(1, \mu_{ss}, X_{ss}, X_{ss}, \cdots)$ corresponds to the case where

[4] In typical dynamic heterogeneous consumer models, households are still subject to idiosyncratic productivity shocks even at the steady state. Hence, the notation $c_{i,ss}$ should not be interpreted as the deterministic steady-state level of c of individual i. Nevertheless, we adopt this notation to maintain consistency with standard regression practice. The index i can be interpreted as representing a particular combination of idiosyncratic state variables of households (e.g., asset a and productivity e in Example 1).

[5] Just to name a few, see Jappelli and Pistaferri (2014), Fagereng et al. (2021), and Ampudia et al. (2024) for empirical evidence on MPC heterogeneity.

[6] Recently, Lewis et al. (Forthcoming) stress the importance of households' latent characteristics to explain MPC heterogeneity. Equation (6) does accommodate their specification: We can simply incorporate latent factors as an input for $\phi(\cdot)$.

individuals are subject to the unit-sized direct effect but no indirect effect. The latter is equivalent to the density of $(c_{i,ss} + \phi_i \eta_i)_i$, i.e., the steady-state consumption augmented by the individual-specific direct response. This density can be constructed given knowledge of the individual-specific direct effect ϕ_i as well as the policy design η_i. The resulting F_ε is the key object in our identification strategy.

2.2 MAR Model and Identification from Direct Effects

We identify the shock ε_t using information on direct feedback F_ε and investigate the dynamic propagation of the shock in the mixed-autoregressive (MAR) framework, which is an autoregression where both aggregate and functional observations are included as endogenous variables. We typically do not include future aggregates (or expectations of them) in a VAR because the autocorrelation captured by a VAR allows us to analyze these future variables. However, this practice causes a problem in our application because indirect effects partly stem from future aggregates, and thus we fail to capture such feedback.[7]

To see how the omission of future inputs matters in our analysis, consider the following simplified structure, which closely follows the structural MAR model we take to the data.

$$\begin{aligned} X_t &= C_{XX} X_{t-1} + C_{Xf} f_{t-1} + B_X \varepsilon_t \\ f_t &= A_{fX} X_t + C_{fX} X_{t-1} + C_{ff} f_{t-1} + B_f \varepsilon_t \end{aligned}$$

The first equation models the dynamics of X_t. The coefficient C_{XX} is a linear operator such that $C_{XX} : \mathbb{R}^k \to \mathbb{R}^k$, and C_{Xf} is a linear operator such that $C_{Xf} : \mathcal{H} \to \mathbb{R}^k$ where $\mathcal{H}$ denotes a separable Hilbert space of square integrable functions on $\mathbb{R}$. The specification in the second equation follows the decomposition (4) by allowing indirect feedback from X_t to f_t, but terms involving future aggregates are excluded. The objective of this model is to highlight the issues surrounding shock identification. Shocks other than our object of interest, $\varepsilon_t \in \mathbb{R}$, are deliberately omitted in order to focus on the dynamics driven by ε_t. In addition, the lag order is restricted to one for the sake of exposition, but it is straightforward

[7] There is a growing literature on aggregate VAR models incorporating subjective expectations from survey evidence (e.g., Doh and Smith 2022; Adams and Barrett 2025). Incorporating such variables in the system might be helpful for identification in our model. We nevertheless stick to the model without expectations for the following reasons. First, our context requires including expectations at arbitrarily long horizons. Accommodating expectations for all variables and horizons greatly expands the size of the statistical model even with horizon truncation. Second, data on expectations have a limited scope in terms of covered horizons. For example, the Survey of Professional Forecasters (operated by the Federal Reserve Bank of Philadelphia) provides expectations of various macro variables up to one year ahead from the time the survey is taken, although long-term expectations are available for some key variables.

to extend the following discussion to a model with a general lag order.

Combining these two equations yields

$$\begin{bmatrix} I & 0 \\ -A_{fX} & I \end{bmatrix} \begin{bmatrix} X_t \\ f_t \end{bmatrix} = \begin{bmatrix} C_{XX} & C_{Xf} \\ C_{fX} & C_{ff} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ f_{t-1} \end{bmatrix} + \begin{bmatrix} B_X \\ B_f \end{bmatrix} \varepsilon_t$$

Moreover, we can rewrite the model above as the canonical form for a structural autoregressive model.

$$\begin{bmatrix} X_t \\ f_t \end{bmatrix} = \begin{bmatrix} G_{XX} & G_{Xf} \\ G_{fX} & G_{ff} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ f_{t-1} \end{bmatrix} + \begin{bmatrix} H_X \\ H_f \end{bmatrix} \varepsilon_t$$

This representation gives H_X and H_f, at-impact total responses of X_t and f_t respectively, and hence it follows $\frac{dX_t}{d\varepsilon_t} = H_X$ and $\frac{df_t}{d\varepsilon_t} = H_f$.

To evaluate the consequences of omitting future aggregate variables, note that the autoregressive structure implies the relationship between responses at horizon k and at-impact responses $\frac{dX_{t+h}}{d\varepsilon_t} = G^h_{XX}\frac{dX_t}{d\varepsilon_t} + G^h_{Xf}\frac{df_t}{d\varepsilon_t}$ for $h \geqslant 0$ where G^h_{XX} is the upper-left block of G^h and G^h_{Xf} is the upper-right block of G^h. Substituting it into equation (4) implies

$$\begin{aligned} \frac{df_t}{d\varepsilon_t} = & \underbrace{\left(I - \left(F_1 G_{Xf} + F_2 G^2_{Xf} + \cdots\right)\right)^{-1} F_\varepsilon}_{:=(I-\mathcal{M}_f)^{-1}F_\varepsilon = B_f} \\ & + \underbrace{\left(I - \left(F_1 G_{Xf} + F_2 G^2_{Xf} + \cdots\right)\right)^{-1} \left(F_0 + F_1 G_{XX} + F_2 G^2_{XX} + \cdots\right)}_{A_{fX}} \frac{dX_t}{d\varepsilon_t} \end{aligned} \tag{7}$$

This equation characterizes B_f and A_{fX} in the model above. The first term B_f is analogous to direct effect F_ε. However, it is distorted by the inverse of $I - \mathcal{M}_f$: The direct effect is mixed with autoregressive feedback G as well as the indirect effect from future aggregates $(F_1, F_2, \cdots)$. The second term captures the effect through current aggregates X_t, while it is again compounded by G and $(F_1, F_2, \cdots)$. Intuitively, because of the autoregressive structure, responses of future aggregates to the current shock can be expressed as a linear combination of at-impact responses of both f_t and X_t. These responses appear in the coefficients to F_ε and $dX_t/d\varepsilon_t$. We can estimate the reduced form parameters $(G_{Xf}, G^2_{Xf}, \cdots)$ from the data, but $(F_0, F_1, \cdots)$ cannot be estimated since the indirect effect is absorbed by the inclusion of time-fixed effects in the panel regression. In general, identification requires restricting some structural parameters. We would like to restrict B_f using the information on the direct effect F_ε to achieve identification, but it is not possible to recover B_f from F_ε alone without further assumptions.

Our approach is to compute an approximation of the sequence $(F_0, F_1, \cdots)$ from A_{fX}, and combine it with F_ε to approximate B_f. The coefficient A_{fX} is defined by the autoregressive parameters G as well as $(F_0, F_1, \cdots)$, and G is the reduced form parameter we can estimate without any identification assumption. Therefore, if we impose some structure on $(F_0, F_1, \cdots)$, A_{fX} provides information to infer the sequence.

As a baseline, we approximate $(F_0, F_1, \cdots)$ with $(\tilde{F}_0, \tilde{F}_1, \cdots)$ where $\tilde{F}_h = \rho^h \tilde{F}_0$ for a scalar $\rho \in (0, 1)$. Then, we obtain the following approximation.

$$A_{fX} \approx \left(I - \left(\rho \tilde{F}_0 G_{Xf} + \rho^2 \tilde{F}_0 G_{Xf}^2 + \cdots\right)\right)^{-1} \left(\tilde{F}_0 + \rho \tilde{F}_0 G_{XX} + \rho^2 \tilde{F}_0 G_{XX}^2 + \cdots\right)$$

From this, we compute $\tilde{F}_0$ as follows.

$$\tilde{F}_0 = A_{fX} \left(I_k + \left(\rho G_{XX} + \rho^2 G_{XX}^2 + \cdots\right) + \left(\rho G_{Xf} + \rho^2 G_{Xf}^2 + \cdots\right) A_{fX}\right)^{-1} \tag{8}$$

In practice, we can select ρ to minimize the approximation error. With $\tilde{F}_0$ obtained in this way, we approximate B_f as

$$\tilde{B}_f = \left(I - \left(\rho \tilde{F}_0 G_{Xf} + \rho^2 \tilde{F}_0 G_{Xf}^2 + \cdots\right)\right)^{-1} F_\varepsilon \tag{9}$$

which serves as the primary source of identification. That is, we restrict B_f to be $\tilde{B}_f$ so that the shock can be interpreted as the one we are interested in. Since $\tilde{F}_0$ can be solved analytically, this approximation scheme simplifies numerical implementation.

One might worry that introducing a particular decaying structure on the sequence $(F_0, F_1, \cdots)$ is a crucial assumption: It implies that the sequence decays at the same rate across the entire domain of the density and for every variable in X_t. Nevertheless, in Section 4, we see that the approximation $\tilde{B}_f$ closely matches the true B_f in a quantitative HANK model featuring liquid and illiquid assets and standard frictions.

2.3 Identification in Practice

Although we argued that the approximation strategy for B_f performs well in a quantitative model, it is still an approximation; it does not exactly reproduce B_f. Moreover, the direct effect F_ε is subject to uncertainty because it is based on $\phi(\cdot)$ which itself is an estimated object. For these reasons, fixing B_f at $\tilde{B}_f$ for identification is far from the best option. Even if there were sufficient reasons to believe that the approximation strategy works well and one

imposed $B_f = \tilde{B}_f$ dogmatically, this restriction would not be sufficient to just identify the shock in general.

We partially identify the shock by imposing only plausible restrictions rather than imposing strong identification assumptions, such as zero restrictions. In particular, we follow the literature starting from Baumeister and Hamilton (2015) and achieve the shock identification by incorporating prior information on structural parameters. The methodology outlined above provides the information on B_f conditional on A_{fX}. We use this information to specify the conditional prior for B_f given A_{fX} as a distribution centered at the approximation $\tilde{B}_f$.

This method allows researchers to reflect uncertainty in the identification restrictions, helping address the issues discussed above. First, this strategy takes into account the uncertainty surrounding the identification condition. Unlike strategies that restrict certain structural parameters to fixed values (such as Cholesky identification and long-run restrictions), our prior specification allows for ambiguity in the identification conditions. This feature is particularly useful in our framework because, as discussed above, we do not have precise information on B_f; we only know its approximation. Second, we take into account the fact that F_ε itself might be based on estimated objects. We reflect uncertainty pertaining to the estimation of the direct effect $\phi(\cdot)$ by imposing a less restrictive prior.

Another advantage of this identification scheme is its ability to incorporate prior information beyond the direct effect. For example, if we are sure that the transfer policy increases the output, we may incorporate this knowledge by specifying the distribution of the corresponding element of B_X to have support only on positive values (e.g., Gamma or truncated normal distribution). In addition, we may incorporate information regarding A_{fX} to sharpen the identification. We will revisit these points when we introduce the full statistical model.

3 Empirical Framework

This section extends the single-shock illustration in the previous section and presents the baseline empirical model. We also introduce a dimension-reduction method to obtain a finite-dimensional expression of our econometric framework. Our discussion then turns to details on estimation, such as identification, estimation algorithm, and prior specifications.

3.1 MAR Model

We denote the separable Hilbert space of square integrable functions on $\mathbb{R}$ as $\mathcal{H}$. The Hilbert space $\mathcal{H}$ is equipped with the inner product $\langle g, h \rangle = \int g(r)h(r)dr$ $(g, h \in \mathcal{H})$ and the tensor

operator $g \otimes h$ satisfying $(g \otimes h)v = \langle v, h \rangle g$ for $v \in \mathcal{H}$. Let $X_t = [X_{1t}, \cdots, X_{kt}]' \in \mathbb{R}^k$ denote a vector of aggregate variables, $z_t \in \mathbb{R}$ another aggregate variable, and f_t a random function in $\mathcal{H}$ representing a cross-sectional density. Our structural MAR model is given by

$$\underbrace{\begin{bmatrix} I & 0 & 0 \\ -A_{zX} & 1 & 0 \\ -A_{fX} & 0 & I \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} X_t \\ z_t \\ f_t \end{bmatrix}}_{Y_t} = C(L) \underbrace{\begin{bmatrix} X_{t-1} \\ z_{t-1} \\ f_{t-1} \end{bmatrix}}_{Y_{t-1}} + \underbrace{\begin{bmatrix} B_{XX} & B_{Xz} & B_{Xf} \\ 0 & B_{zz} & B_{zf} \\ 0 & B_{fz} & B_{ff} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} \varepsilon_t^X \\ \varepsilon_t^z \\ \varepsilon_t^f \end{bmatrix}}_{\varepsilon_t} \tag{10}$$

The augmented variable Y_t lies in $\mathbb{R}^{k+1} \oplus \mathcal{H}$ where $\oplus$ represents the direct sum of two spaces. The lag polynomial $C(L)$ is defined as $C(L) = C_1 + C_2 L + \cdots + C_p L^{p-1}$. Linear operators A and B map the space $\mathbb{R}^{k+1} \oplus \mathcal{H}$ to itself, and their sub-blocks map the space of the variable indexed by the second subscript to the space of the variable indexed by the first subscript (e.g., $A_{zX} : \mathbb{R}^k \to \mathbb{R}$, and $B_{fz} : \mathbb{R} \to \mathcal{H}$). Every variable is assumed to be demeaned so that constant terms need not be included.[8] The structural shocks ε_t are mutually orthogonal and serially uncorrelated: $\mathbb{E}(\varepsilon_t \otimes \varepsilon_s) = 1\{t = s\} I$.

We are interested in identifying a shock ε_t^z. We partition the aggregate variables into one variable directly tied to the shock, z_t, and others X_t. For example, if we are interested in a transfer shock, z_t would be the aggregate measure for transfers and X_t would be the collection of other relevant aggregate variables. This setup allows us to represent the propagation of the shock ε_t^Z to f_t as the sum of component through B_{fz} and component through X_t via A_{fX}. The counterparts of A_{fX}, B_X, and B_f in the single-shock exposition are A_{fX}, B_{Xz}, and B_{fz} respectively.[9]

This model can be represented as the canonical form for structural autoregressive models by left-multiplying A^{-1}.

$$Y_t = G(L) Y_{t-1} + H \varepsilon_t \tag{11}$$

where $H := A^{-1}B$ is an operator representing the at-impact impulse response and $G(L) := A^{-1}C(L) = G_1 + G_2 L + \cdots + G_p L^{p-1}$ is the lag polynomial. The reduced-form error has variance $\Sigma = HH'$.

[8] For functional observations, demeaning here implies a temporal demeaning $f_t - \frac{1}{T}\sum_t f_t$. This differs from cross-sectional demeaning of micro observations at the same time period, which ensures $\int r f_t(r) dr = 0, \forall t$.

[9] One might worry that we are imposing some identification assumptions by parameterizing (A, B) in the way described in equation (10). Proposition 3 in D shows that it is not the case. This proposition establishes the one-to-one relationship between H and (A, B) under mild invertibility and boundedness conditions. That is, for (almost) any at-impact impulse response H, we can find (A, B) consistent with H, and vice versa. In this sense, our parametrization of (A, B) does not rule out any at-impact impulse responses.

3.2 Reducing Dimensionality

The MAR model (10) cannot be estimated because f_t is infinite-dimensional. We derive the finite-dimensional representation of the model, following Y. Chang et al. (2024b) and Y. Chang et al. (2025). See Appendix C for a more detailed exposition of the approximation approach.

We consider an arbitrary orthonormal basis $(v_i)_{i\geqslant 1}$ spanning the space $\mathbb{R}^{k+1} \oplus \mathcal{H}$. We approximate the functional variable f_t by restricting attention to the finite subset $(v_i)_{i=1}^{k+1+m}$, consisting of the first $(k+1+m)$ basis elements. Define, for any $Y \in \mathbb{R}^{k+1} \oplus \mathcal{H}$,

$$(Y) := \begin{bmatrix} \langle v_1, Y \rangle \\ \vdots \\ \langle v_{k+1+m}, Y \rangle \end{bmatrix} \in \mathbb{R}^{k+1+m}.$$

We also define, for any linear operator P on $\mathbb{R}^{k+1} \oplus \mathcal{H}$,

$$(P) := \left[\langle v_i, P v_j \rangle\right]_{i,j=1,\cdots,k+1+m} \in \mathbb{R}^{(k+1+m)\times(k+1+m)}$$

It can then be shown that (11) can be approximated as

$$(Y_t) = (G(L))(Y_{t-1}) + \underbrace{\begin{bmatrix} H_{XX} & H_{Xz} & (H_{Xf}) \\ H_{zX} & H_{zz} & (H_{zf}) \\ (H_{fX}) & (H_{fz}) & (H_{ff}) \end{bmatrix}}_{(H)} (\varepsilon_t) \tag{12}$$

This expression represents the approximated MAR as a VAR with $(k+1+m)$ endogenous variables. This can be rewritten in a form consistent with (10).

$$\underbrace{\begin{bmatrix} I & 0 & 0 \\ -A_{zX} & 1 & 0 \\ -(A_{fX}) & 0 & I \end{bmatrix}}_{(A)} \underbrace{\begin{bmatrix} X_t \\ z_t \\ (f_t) \end{bmatrix}}_{(Y_t)} = (C(L)) \underbrace{\begin{bmatrix} X_{t-1} \\ z_{t-1} \\ (f_{t-1}) \end{bmatrix}}_{(Y_{t-1})} + \underbrace{\begin{bmatrix} B_{XX} & B_{Xz} & (B_{Xf}) \\ 0 & B_{zz} & (B_{zf}) \\ 0 & (B_{fz}) & (B_{ff}) \end{bmatrix}}_{(B)} \begin{bmatrix} \varepsilon_t^X \\ \varepsilon_t^z \\ (\varepsilon_t^f) \end{bmatrix} \tag{13}$$

where (f_t) is an m-dimensional vector. Estimation can be carried out using the standard VAR methods, such as OLS, maximum likelihood, or Bayesian approaches.

Although this approximation strategy works for any orthonormal basis, its practical per-

formance depends heavily on the choice of basis. The ideal basis is the one that explains the temporal fluctuations of functional observations effectively, thereby improving estimation efficiency. We follow the recommendation by Y. Chang et al. (2024b) and use the functional principal component (FPC) basis as the baseline. One can show that, for a fixed m, the FPC basis captures more functional variation than any other orthonormal basis. Indeed, for most functional observations used in economic empirical analyses, only a few FPC basis functions are sufficient to capture the bulk of the variation. This allows us to maintain parsimony without sacrificing informational content in the estimation.[10]

3.3 Identification, Bayesian Estimation, and Prior

We estimate the VAR representation of the approximate MAR, (12), in a Bayesian framework by assuming that the error is i.i.d. standard Gaussian $\left((\varepsilon_t^X)', \varepsilon_t^z, (\varepsilon_t^f)'\right)' \sim N(0, I_{k+1+m})$.

3.3.1 Identification

The identification problem comes from the fact that there are multiple at-impact impulse responses consistent with the reduced-form variance (Σ). To see this, let L be a lower triangular matrix such that $(\Sigma) = LL'$, and Q_1 and Q_2 be orthogonal matrices. Then, both at-impact responses $(H_1) := LQ_1$ and $(H_2) := LQ_2$ imply the same reduced-form variance (Σ). As demonstrated previously, we identify the model by placing a non-dogmatic prior on Q. In particular, we place a larger weight on a particular region in the domain of Q (i.e., the space of orthogonal matrices) so that the shock of interest has economically meaningful interpretation.

More formally, our goal is to find the posterior distribution of $((G), (\Sigma), Q)$ where $(G) = ((G_1), \cdots (G_p))$ is the autoregressive parameter. Note that these parameters allow us to compute $((A), (B), (C))$ directly. We specify the prior distribution as

$$p((G), (\Sigma), Q) = p((G), (\Sigma))\, p(Q \mid (G), (\Sigma))$$

The first term on the right-hand side gives the prior for reduced-form parameters. We may impose the standard distributional assumption (e.g., normal-inverse-Wishart distribution) for this component to take advantage of well-known posterior samplers for the reduced-form

[10] One of the concerns in applying the FPC basis to densities is that one may not fully enforce the unit-integral and non-negativity constraints. Since we demean the functional observations, the integral constraint is already satisfied in our analysis. Moreover, violations of the non-negativity constraint are mild in practice. See Appendix C for more detailed discussion.

parameters. The second term represents the prior for structural parameters conditional on the reduced-form parameters. Importantly, this component can be expressed in terms of structural parameters of interest. For example, since both (A) and (B) are functions of (Σ) and Q, $p\left(Q \mid (G), (\Sigma)\right)$ can be specified in terms of a prior for (A) and (B). This prior specification separates the computation of the posterior distribution into the estimation part (drawing the reduced-form parameters) and the identification part (drawing the structural parameters conditional on reduced-form parameters).

3.3.2 Estimation Algorithm

The Bayesian estimation can be performed hierarchically: First, we draw the reduced-form parameters $((G), (\Sigma))$ from their posterior distribution. We can employ well known algorithms to make this step done under the standard distributional assumption on the prior.[11] Second, conditional on the reduced-form parameters drawn in the previous step, we draw the orthogonal matrix Q. We rely on the Metropolis-Hastings algorithm for this step: We multiply Q at the previous iteration with the exponential of a random skewed-symmetric matrix to propose a candidate of a new Q. Then we accept or reject the proposal based on the ratio of the posterior kernels. This proposal distribution is centered at the previous Q, and makes sure candidates are orthogonal again. See Appendix B for more details on the Bayesian algorithm.

3.3.3 Prior for Structural Parameters

We discuss the prior specification for structural parameters $p(Q \mid (G), (\Sigma))$, which plays a central role in our shock identification. We represent the conditional prior of Q as

$$p(Q \mid (G), (\Sigma)) \propto 1\{Q \in \mathcal{O}(k+1+m)\} \times \underbrace{p\left(B_{XX}, B_{Xz}, (B_{Xf}) \mid (G), (\Sigma)\right)}_{X_t \text{ block (1st row)}}$$
$$\times \underbrace{p\left(A_{zX}, B_{zz}, (B_{zf}) \mid (G), (\Sigma)\right)}_{z_t \text{ block (2nd row)}} \times \underbrace{p\left((A_{fz}), (B_{fz}), (B_{Xf}) \mid (G), (\Sigma)\right)}_{(f_t) \text{ block (3rd row)}}$$

up to a scaling constant so that it integrates to one with respect to Q, where $\mathcal{O}(k+1+m)$ is a set of $(k+1+m) \times (k+1+m)$ orthogonal matrices. We partition the model into the first, second, and third rows in (13). We assume independence of parameters across blocks, while allowing dependence of them belonging to the same block. We are especially interested in

[11] See, for example, Koop and Korobilis (2010) and Kilian and Lütkepohl (2017).

the third row, which is a collection of the equations that determines $\{f_t\}$. We parameterize this term as

$$\begin{aligned} p\left(\{A_{fz}\},\{B_{fz}\},\{B_{Xf}\} \mid \{G\},\{\Sigma\}\right) =& p\left(\{B_{fz}\} \mid \{A_{fz}\},\{B_{ff}\},\{G\},\{\Sigma\}\right) \\ & \times p\left(\{A_{fz}\} \mid \{B_{Xf}\},\{G\},\{\Sigma\}\right) \\ & \times p\left(\{B_{Xf}\} \mid \{G\},\{\Sigma\}\right) \end{aligned} \tag{14}$$

The first term gives the conditional prior of $\{B_{fz}\}$ (compounded direct effect) given $\{A_{fz}\}$, $\{B_{ff}\}$, and reduced-form parameters. This term is where we can impose our identification strategy. Given the conditioned variables, we can compute the approximation $\widetilde{\{B_{fz}\}}$ based on the finite-dimensional analogue of equation (9).

$$\widetilde{\{B_{fz}\}} = \left(I - \left(\rho\widetilde{\{F_0\}}\{G_{Xf}\} + \rho^2\widetilde{\{F_0\}}\{G_{Xf}^2\} + \cdots\right)\right)^{-1}\{F_\varepsilon\} \tag{15}$$

where $\rho \in (0,1)$, $\{F_\varepsilon\}$ is a finite-dimensional approximation of F_ε given the basis used to derive the approximate MAR, and

$$\widetilde{\{F_0\}} = \{A_{fX}\}\left(I + \left(\rho\{G_{XX}\} + \rho^2\{G_{XX}^2\} + \cdots\right) + \left(\rho\{G_{Xf}\} + \rho^2\{G_{Xf}^2\} + \cdots\right)\{A_{fX}\}\right)^{-1} \tag{16}$$

We specify the conditional prior of $\{B_{fz}\}$ to be the distribution centered at $\widetilde{\{B_{fz}\}}$. We assign a positive prior variance to it so that we reflect the approximation and estimation errors associated with the direct effect.

The prior for parameters in the other blocks can be specified in application-specific ways. We may let a prior for every other parameter uninformative by setting a large prior variance. However, combining informative prior for some of those parameters with the information imposed on $\{B_{fz}\}$ helps to shrink the identification set. Here we give a general guideline on what we can impose, and we will describe our prior choice for simulation exercises and empirical applications later.

Responses of Aggregate Variables to Shock of Interest (B_{Xz} and B_{zz}). Aggregate variables X_t responds to the shock of interest ε_t^z simultaneously, and size of the response is given by B_{Xz}. One can impose additional restrictions for this part if a researcher has any belief about how X_t responds to the shock. We may specify lower- or upper- bounds for those variables such as sign restrictions, or impose best guess of the response as prior mean.

In particular, it is relatively easy to predict B_{zz}, i.e., how much the unit shock increases/ decreases the variable tied to the shock z_t. For example, one can approximate the aggregate scale of the stimulus payment once we know the population size of the targeted individuals and how much each of them receive. Reflecting such information would be helpful to pin down the scale of the unit shock relative to the economy, thereby circumventing the size ambiguity (Stock and Watson 2018).

Responses to Other Aggregate Shocks (B_{XX}). We may sharpen the identification of ε_t^z by identifying other aggregate shocks ε_t^X. To see why, recall that, in equation (16), we construct the approximation of indirect effect from $\{A_{fX}\}$. One can derive $\{A_{fX}\} = \{H_{fX}\}H_{XX}^{-1}$, implying that $\{A_{fX}\}$ is driven from the responses of X_t and f_t to ε_t^X. Identifying aggregate shocks through restrictions on B_{XX} therefore gives a more precise approximation of indirect effects.

The observation that indirect effects are governed by aggregate shocks other than the shock of interest is related to the statement of the demand equivalence theorem in Wolf (2023). The theorem shows that the general equilibrium propagation of private consumption is identical to that of government spending under certain assumptions on the economic environment[12], motivating us to use the empirical evidence on the aggregate fiscal multiplier to construct the indirect effect. Our framework also suggests the usefulness of aggregate shocks other than the shock of interest in approximating the indirect effect, while the shocks that can be utilized here are not limited to the government spending shock, and we do not require the assumptions for the demand equivalence theorem.

To identify those aggregate variables, one may rely on the typical identification methodologies in empirical macroeconomics (e.g., imposing exclusion restrictions, or finding instrumental variables). Note, however, that we do not necessarily assign structural interpretation for those aggregate shocks, as we are not interested in the propagation of them. Instead, we need to distinguish the shocks driving the variations in aggregate variables, ε_t^X, with those driving the variations in functional variables, $\{\varepsilon_t^f\}$. In our applications, we impose agnostic sign restrictions on B_{XX} so that each aggregate shock is interpreted as a main driver of the fluctuation in the corresponding aggregate variable.[13]

[12]Those assumptions are that, (i) Households and government consume the same final good, (ii) borrowing and saving interest rates are identical, and (iii) the wealth effect for labor supply is not present or wages are perfectly sticky.

[13]Alternatively, one may leverage the idea of the "max-share" identification (Uhlig 2004; Angeletos et al. 2020) so that we identify the aggregate structural shocks which account for a large part of business cycle fluctuation.

4 Validating Identification Approach with HANK

This section examine how our methodology performs in a quantitative heterogeneous agent New Keynesian (HANK) model. We overview the environment, evaluate our approximation procedure used to reflect the information on direct effects, and apply our Bayesian sample to the simulated data.

4.1 Environment

To provide a laboratory for our analysis, we construct a two-asset medium-scale heterogeneous agent New Keynesian model. We provide an overview of the model here. See Appendix A for a more detailed description of the model's structure.

There are heterogeneous households, a final good producer, intermediate good producers, a capital good producer, a mutual fund, a labor union, and fiscal and monetary policy authorities. Households can hold illiquid and liquid assets and optimize consumption and saving subject to the borrowing constraint and the law of motion for idiosyncratic productivity. Holdings of liquid assets incur a cost for a liquidity premium, while illiquid assets can be adjusted each period only with an i.i.d. probability. Labor supply is determined by the labor union facing a wage adjustment cost, giving rise to the wage Phillips curve. The capital good producer makes investments subject to adjustment costs and rents capital to intermediate good producers. The optimization behavior of intermediate good producers, subject to price adjustment costs, leads to the price Phillips curve. Finally, the policy rate is determined by the Taylor rule, and the average labor tax rate is set according to the tax rule.

The economy is subject to six aggregate shocks: total factor productivity, government spending, price markup, wage markup, monetary policy, and transfers. The transfer shock $\varepsilon_t^{tr} \sim N(0,1)$ is our object of interest. o illustrate the mechanism of this shock, we present the budget constraint for the household with state (a, b, e), where a and b denote the holdings of illiquid and liquid assets respectively, and e is the idiosyncratic productivity.

$$\begin{aligned} c + a' + b' &= (1 + r_{p,t})a + (1 + r_{p,t} - \omega)b + (1 - \tau_t^y)\left(y_t(e)\right)^{1-\xi} + \left(1 - \tau^{\Pi}\right)\Pi_t(e) + \eta(a,b,e)\sigma^{tr}\varepsilon_t^{tr} \\ y_t(e) &= w_t h_t \Gamma_t(e) \end{aligned} \tag{17}$$

Households receive financial income from the first two terms on the right-hand side, where holdings of liquid assets b carry the liquidity premium ω. Pre-tax labor income y_t is scaled by an incidence factor Γ_t that governs how cross-sectional income responds to aggregate labor

fluctuations. It is then transformed into post-tax income by the progressive tax rule à la Heathcote et al. (2017). The dividend $\Pi_t(e)$ is allocated proportionally to e and is taxed by a fixed tax rate τ^{Π}.

The final term on the right-hand side of the budget constraint represents government transfers, allocated across individuals according to the policy design $\eta(a, b, e)$. In the baseline environment, we study the shock to a lump-sum transfer by setting the sensitivity to be $\eta(\cdot) = 1$ for every state variable. One standard deviation corresponds to 1% of steady-state aggregate output (normalized to be 1), implying $\sigma^{tr} = 0.01$.

4.2 Solution Method and Simulation

We solve the model using the sequence space Jacobian (SSJ) method (Auclert et al. 2021), which provides linearized impulse responses of aggregate variables to each shock. Given those impulse responses, we find the response of the cumulative consumption distribution with backward-forward iterations. The SSJ procedure yields impulse responses over horizons $0, 1, \ldots, T$, where T is chosen to be sufficiently large. These responses allow us to represent both aggregates and distributions as a moving average process.

In numerical computation of heterogeneous agent models, we typically need to discretize the space of idiosyncratic states (productivity e and liquid and illiquid asset holdings a and b in our model) using grids. The combination of discretized idiosyncratic states and the presence of borrowing-constrained households leads to a non-smooth consumption histogram, which contrasts with the smooth consumption distributions observed in the data. To smooth out the consumption distribution from the model, we approximate the consumption cumulative distribution using the I-spline basis functions (Ramsay 1988), which are obtained by integrating the normalized B-spline basis to ensure non-negativity and unit-integral.[14] As such, the I-spline basis functions are monotonically increasing and take values between 0 and 1, which are desirable properties for approximating cumulative distributions. The smoothed density is simply the derivative of the smoothed cumulative distribution. See Appendix A for details on the smoothing procedure.

Our MAR model contains five aggregate variables in X_t: output, inflation, real wage, investment, and government debt.[15] Since we are interested in the effect of transfer policy, we include the aggregate transfer as z_t. The functional observation f_t is the density of

[14] The normalized B-spline is also called M-spline.

[15] We choose those five variables because they serve as inputs of the directed acyclic graph representation of our HANK, implying that dynamics of those variables are sufficient to summarize the propagation of all endogenous variables in the model.

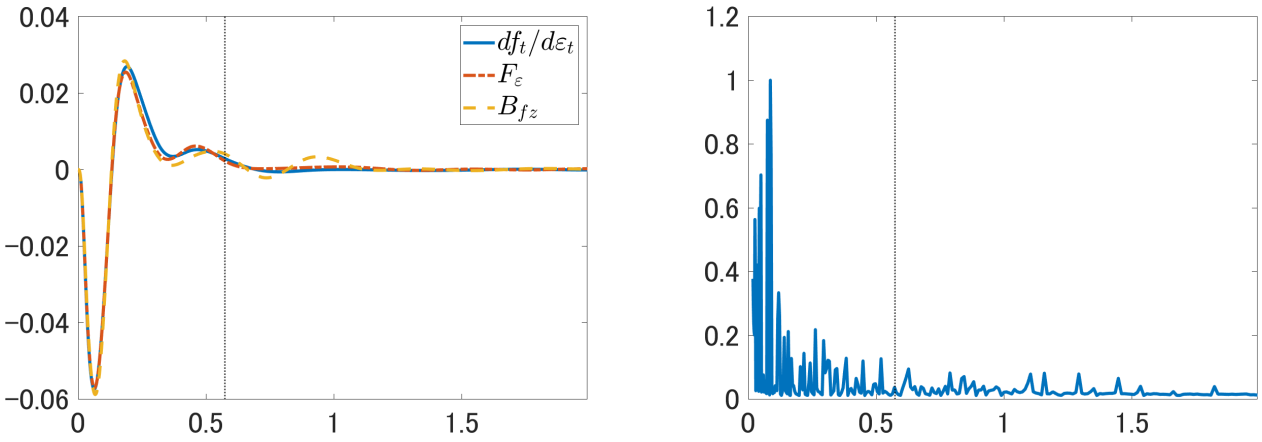

Figure 1: Total Effect, Direct Effect, and B_{fz} (Left) and MPC (Right)

Note: The left panel shows the total effect $df_t/d\varepsilon_t$ (blue solid line), direct effect F_ε (red dashed line), and the compounded direct effect B_{fz} (orange dash dotted line) of the lump-sum transfer shock to the consumption density. The right panel shows the relationship between consumption (x-axis) and MPC (y-axis). For both panels, vertical line shows the average of individual consumption at the steady state.

cross-sectional consumption. These variables are simulated from the moving average representation implied by the SSJ method. All variables are defined in terms of deviation from their steady-state levels, as the MAR requires non-demeaned variables. The autoregressive coefficient associated with the data generating process is computed using the autocovariance of (X_t, z_t, f_t) implied from the moving average representation.

4.3 Total Effect, Direct Effect, and Compounded Direct Effect

Before turning to the simulation exercises, we investigate how the consumption density responds to the lump-sum transfer shock. We first compute the total effect, direct effect, and the compounded direct effect involving the MAR representation. The total effect $\frac{df_t}{d\varepsilon_t}$ is the at-impact impulse response of the consumption density. We compute the direct effect F_ε from the backward and forward iteration of household problems where we impose ε_t while letting the sequence of other aggregate variables stay at the steady-state levels. The compounded direct effect B_{fz} is computed following the definition in equation (7).

The left panel of Figure 1 shows the total effect $\frac{df_t}{d\varepsilon_t}$, direct effect F_ε, and the compounded direct effect B_{fz}. The vertical dotted line represents the cross-sectional mean of consumption at the steady state. These three lines align very well, suggesting that general equilibrium forces do not act significantly on the consumption density.

In response to the transfer, households with the lowest consumption increase their spend-

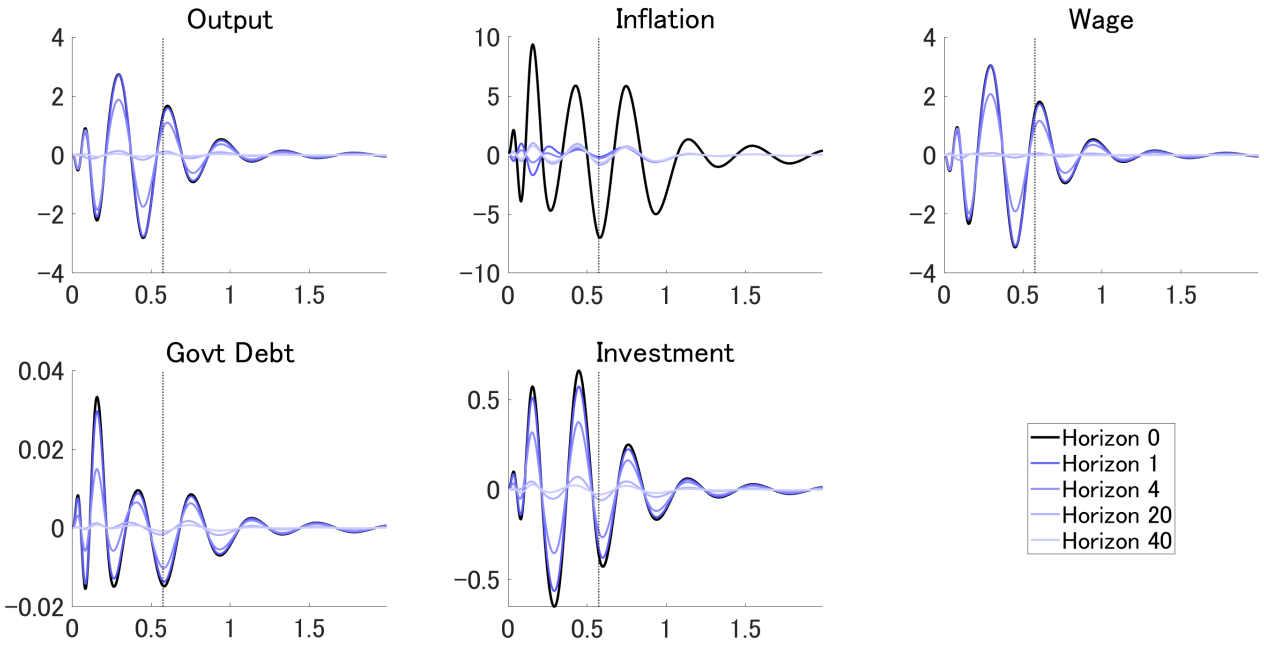

Figure 2: Coefficients in Indirect Effects: $(F_0, F_1, \cdots)$

Note: This figure shows the coefficients associated with impulse response of X_{t+k} in the indirect effect, namely $(F_h)_{h\geqslant 0}$, for output, inflation, wage, government debt, and investment. We plot them for $h = 0, 1, 4, 20, 40$. The vertical dotted line shows the cross-sectional average of consumption at the steady state.

ing. Their consumption changes from around 0.05 to around 0.2, generating a valley and peak in the left-most side of the density. On the other hand, the households that consume above the cross-sectional mean react little to the transfer, and thus we do not see significant variation in the middle-to-right part of the figure. This pattern is consistent with the MPC shown in the right panel of Figure 1. Although it exhibits some fluctuations, households consuming less than others are typically constrained by the borrowing limit and have a larger MPC. The households with large MPCs are concentrated in the lower end of the consumption distribution, to whom the transfer policy is the most effective.

4.4 Approximating B_{fz}

As discussed previously, our methodology involves approximating the compounded direct effect B_{fz} by assuming the decaying structure of $(F_0, F_1, \cdots)$, namely $F_h = \rho^h F_0$ for $\rho \in (0, 1)$. We argue that this approximation procedure performs well in our HANK model.

Figure 2 shows F_h for five aggregate variables which will be included in the MAR (output, inflation, wage, government debt and investment) for $h = 0, 1, 4, 20, 40$. Note that F_h is the

Table 1: Ratio of F_h and F_0 at Cross-Sectional Average Consumption at Steady State

h	Output	Inflation	Wage	Govt Debt	Investment	0.9^h
1	0.935	0.022	0.932	0.923	0.895	0.9
4	0.671	0.039	0.651	0.694	0.643	0.656
20	0.087	0.113	0.038	0.114	0.153	0.122
40	0.033	0.089	-0.008	0.059	0.075	0.015

Note: This table reports the ratio between F_h and F_0 evaluated at the average consumption in the steady state, for horizons $h = 1, 4, 20, 40$. It also shows the power 0.9^h for comparison.

Fréchet derivative of consumption density with respect to X_{t+h}, meaning that F_h itself is interpreted as a function with the same domain as the density. As expected, F_h shrinks toward zero when the horizon k increases. Moreover, F_h monotonically decays over the domain since they are vertically stretched versions of each other, supporting the approach of multiplying a fixed constant to F_0 to approximate F_h. An exception is inflation rate, where F_0 stands out while the other F_h values stay close to zero. One possible explanation is that, real interest rate is (i) positively related to current inflation because a rise in inflation increases nominal interest rate by the Taylor rule, while (ii) negatively related to one-period ahead inflation because real interest rate is computed as the gap between nominal interest rate and inflation expectations through the Fisher equation. Both forces are at work and cancel out each other's effects, except for current inflation ($h = 0$) for which only the first force is present. This creates the difference between F_h at horizon 0 and at other horizons for inflation.

Table 7 takes a closer look at the ratio of F_h ($h = 1, 2, 20, 40$) to F_0 plotted in Figure 2 evaluated at the cross-sectional average of consumption at the steady state (vertical dashed line in the Figure). Except for inflation, these values broadly match the power of 0.9. In Table 7 in Appendix E, we show that this pattern is robust for other choices of points. Overall, these observations suggest that the approximation would work well if inflation rate is not a dominant driver of the indirect effect to consumption density.

With the approximation scheme discussed previously, we compute $\tilde{B}_{fz}$ with different ρ. The grid search over ρ shows that $\rho = 0.898$ minimizes the approximation error in terms of Frobenius norm in equation (16), and thus we treat $\rho = 0.9$ as the baseline. We calculate $\tilde{B}_{fz}$ for $\rho = 0.85, 0.95$ as the robustness check. Figure 3 compares the true B_{fz} (blue solid line) with its approximation $\tilde{B}_{fz}$ (orange dotted line). Although some approximation errors exist for $\rho = 0.95$, $\tilde{B}_{fz}$ aligns well with the true B_{fz}, showing that our approximation methodology

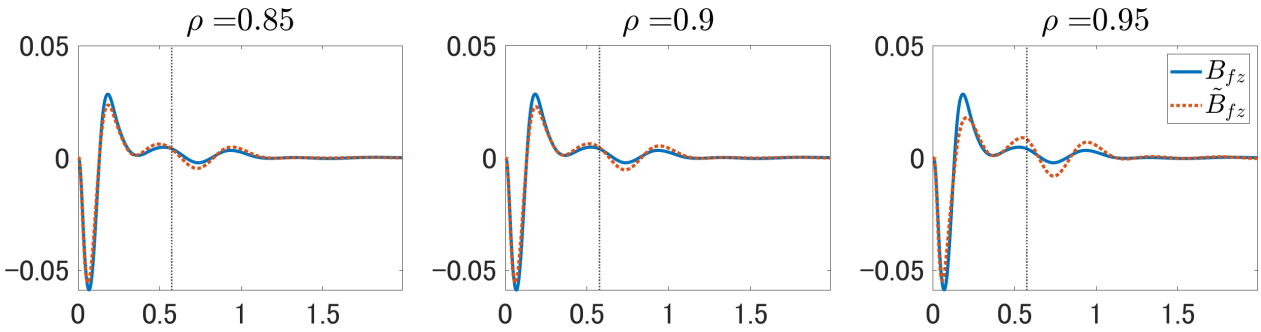

Figure 3: Approximation of B_{fz} for $\rho = 0.85, 0.9, 0.95$

Note: The figure plots the compounded direct effect B_{fz} (blue solid line) and its approximation $\tilde{B}_{fz}$ (orange dotted line) for different assumptions on $\rho = 0.85, 0.9, 0.95$. Vertical line shows the average consumption at the steady state.

works well.

4.5 Simulation Exercises

We now simulate data from the model and apply our identification methodology to assess its performance. We generate the observations from the moving average representation of the model solution for $T = 1000$. The choice of sample size is relatively large compared to actual datasets (typically fewer than 200 observations for quarterly data). We aim to mitigate small-sample issues in estimating the reduced-form model because the objective of this exercise is to illustrate the identification methodology. The variance $\Gamma = \frac{1}{T}\sum_{t=1}^{T}(f_t \otimes f_t)$ is computed from the simulated sequence of (f_t), and we apply functional principal component analysis to obtain the FPC basis. We estimate the reduced form parameters, $([G], [\Sigma])$, by OLS. In the simulation exercise, the reduced-form parameters are fixed at their OLS estimates. The error bands shown below capture the uncertainty due solely to the identification methodology, and do not reflect the uncertainty pertaining to estimation.[16]

As in Y. Chang et al. (2023), we select the number of basis functions m and lag order p jointly to minimize the one step ahead out-of-sample forecast error in rolling-window estimation with each window of size 60 using the reduced-form model with only (f_t) as observables. This criterion suggests $m = 5$ and $p = 1$. These five FPC basis functions account for 99.1%

[16]In terms of the algorithm, we iterate Step (2-ii) of Algorithm 1 for $J + K$ times, and use the last K draws to compute the results shown in the figures.

of variation in (f_t) over time.[17]

4.5.1 Prior

As we fix $([G], [\Sigma])$ at their OLS estimates $\left(\widehat{[G]}, \widehat{[\Sigma]}\right)$, we draw Q from the conditional distribution of it given these reduced-form parameters.

$$p\left(Q \mid \widehat{[G]}, \widehat{[\Sigma]}\right)$$

We discuss our prior choice for Q, which consists of prior for parameters in each block of equations.

Prior on B_{XX}, B_{Xz}, and $[B_{Xf}]$ conditional on $([G], [\Sigma])$. Let X_{ij} denote the (i, j) entry of a matrix X. Note that for $i = 1, \cdots, k$, it follows $\sum_{j=1}^{k}(B_{XX})_{i,j}^2 + (B_{Xz})_{i,1}^2 + \sum_{j=1}^{m}[B_{Xf}]_{i,j}^2 = [\Sigma]_{i,i}$. Therefore, absolute values of all the parameters in the i-th equation are bounded by $[\Sigma]_{i,i}^{0.5}$.

As demonstrated earlier, we impose sign restrictions on B_{XX} so that we provide a more precise view on the approximated indirect effect. Table 2 provides the overview of the sign restrictions. Again, since the goal here is not to give structural interpretation on those aggregate shocks but to distinguish the aggregate and functional shocks, we impose the parsimonious restrictions on them. A demand shock moves the output and inflation to the same direction, while a supply shock moves them to the opposite direction. Wage, government debt, and investment shocks raise the corresponding variables, and are assumed to be the main contributors of the fluctuation of those variables. We require that, for example, the contemporaneous response of wage to the wage shock is larger in absolute value than the responses of it to any other shocks.

We do not impose any information on the size of the responses. Thus, the elements of the i-th row in B_{XX}, B_{Xz}, and $[B_{Xf}]$ have the uniform prior over $\left[-[\Sigma]_{i,i}^{0.5}, [\Sigma]_{i,i}^{0.5}\right]$ if they are not restricted, and over $\left[-[\Sigma]_{i,i}^{0.5}, 0\right]$ or $\left[0, [\Sigma]_{i,i}^{0.5}\right]$ if the sign restriction is imposed depending on whether the parameters are negatively or positively constrained.

[17] We measure variation in (f_t) using the functional R-squared (FR^2), defined as

$$FR_m^2 = \frac{\sum_t \|\Pi_m f_t\|^2}{\sum_t \|f_t\|^2}$$

where Π_m is the projection from the original Hilbert space onto the space spanned by the m basis functions. We find that only one basis function explains 80.4% of the variation.

Table 2: Sign Restrictions on B_{XX}

Shock	Output	Inflation	Wage	Govt Debt	Investment
Demand Shock	+	+			
Supply Shock	−	+			
Wage Shock			++		
Govt Debt Shock				++	
Investment Shock					++

Note: Sign restrictions imposed on B_{XX}. The signs "+" and "−" indicates that the element is restricted to be positive or negative, and "++" indicates the combination of the positivity restriction and that the corresponding element has the largest absolute value than any other elements in the same equation (i.e., the corresponding shock is the dominant contributor of the variable). No restriction is imposed to the elements with blank entry.

Prior on B_{zz}, (B_{zf}), and A_{zX} conditional on $((G), (\Sigma))$. Since the lump-sum transfer shock in our model corresponds to 1% of steady state output, it is fair to guess $B_{zz} = 0.01Y_{ss}$. In reality, this guess can be computed from the knowledge of policy designs, such as the amount of transfer for each household and how many households are included as targets for the policy. We assume that B_{zz} follows a normal distribution with prior mean of $0.01Y_{ss}$ and the prior standard deviation equal to one-tenth of the prior mean.

The prior for the elements of (B_{zf}) is flat again by the same token as above. The prior for entries of A_{zX} corresponding to X_{it} $(i = 1, \cdots, k)$ is normal with mean zero and standard deviation $\lambda \frac{(\Sigma)^0_{k+1,k+1} \cdot 5}{\text{sd}(X_{it})}$ where the denominator is the standard deviation of X_{it} and $\lambda > 0$ is a hyperparameter governing the tightness of the prior. This specification follows the standard practice to determine the prior for regression models with non-standardized data. We set $\lambda = 5$, which is larger than the standard choice so that we apply only a little shrinkage for this part.

Prior on (A_{fX}), (B_{fz}), and (B_{ff}) conditional on $((G), (\Sigma))$. We follow the same procedure above to settle on the prior for (A_{fX}) and (B_{ff}). The prior for (B_{ff}) is flat. The prior for elements of (A_{fX}) corresponding to the coefficient of X_{jt} in the i-th equation $(i = k + 2, \cdots, k + 1 + m,\ j = 1, \cdots, k)$ is normal with mean zero and standard deviation $\lambda \frac{\Sigma^{0.5}_{ii}}{std(X_{jt})}$ where $\lambda = 5$ again.

Conditional on (A_{fX}), we can compute $\widetilde{(B_{fz})}$, a guess for the distributional response to the shock of interest (B_{fz}), from equations (15) and (16). In the baseline, ρ is fixed at 0.9. We specify the prior for (B_{fz}) to be the normal distribution with mean $\widetilde{(B_{fz})}$ and standard

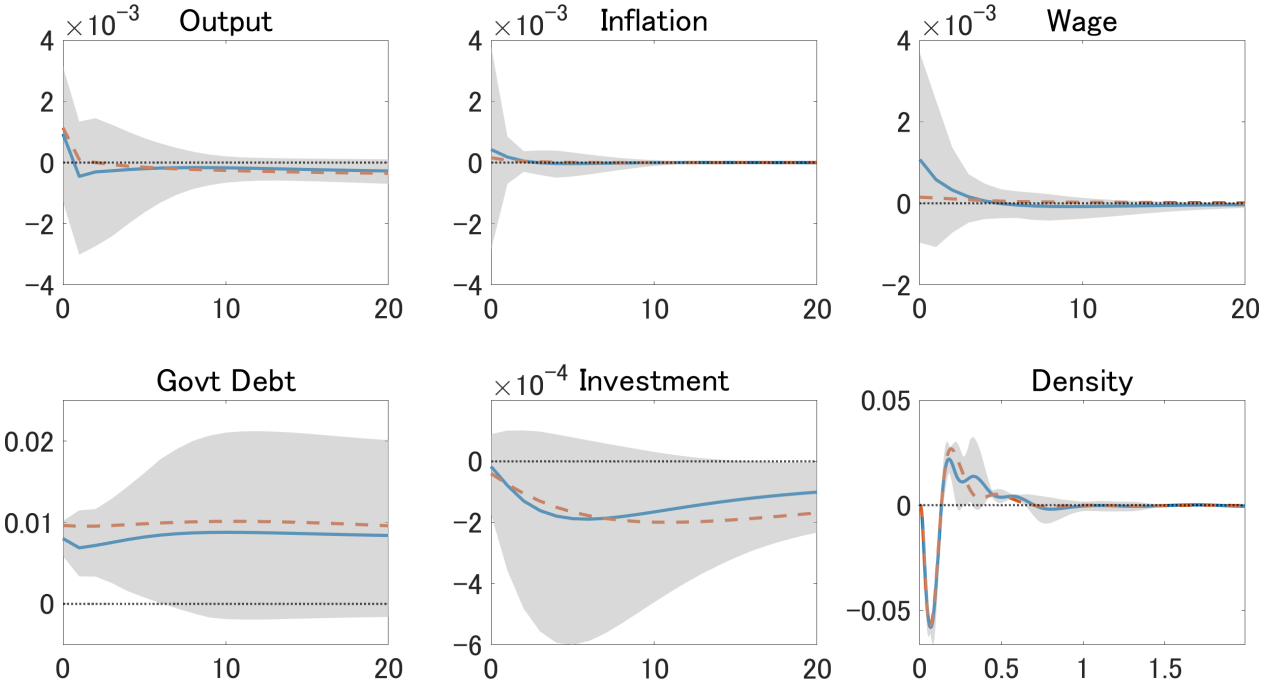

Figure 4: Impulse Responses under $\rho = 0.9$

Note: The bottom-right panel plots the at-impact impulse response of the density. The other panels show the impulse response of aggregate quantities. The blue solid line shows the point-wise mean, along with the 68% interval represented by the shaded area. The orange dashed line shows the corresponding response from the data generating process (i.e., the HANK model). The at-impact response of transfer is normalized to be 0.01.

deviation equal to one-tenth of the prior mean.

4.5.2 Impulse Responses

Figure 4 shows the impulse response functions obtained by applying our identification methodology to the simulated data, where the transfer's at-impact response is normalized to 0.01. The blue solid lines represent the pointwise mean along with the 68% credible intervals (shaded areas). As a baseline, we also plot the true impulse responses implied from the HANK (orange dashed line). The bottom-right panel shows the density response at impact, while the others plot the impulse response of aggregate variables over time. Note that the x-axes differ between the bottom-right panel and others: The x-axis of the bottom-right panel is the consumption level, while the ones for the others are horizons up to 20 quarters.

The bottom-right panel shows that our identification framework captures the overall shape of the density response well, surrounded by the tight credible interval. We impose an informative prior for the compounded direct effect (B_{fz}), while the density response shown

Table 3: 68% Intervals of B_{Xz} ($\times 10^{-3}$)

Variable	Without Direct Effect Information (Uninformed Interval)	With Direct Effect Information (Informed Interval)
Output	$(-3.47, 3.47)$	$(-0.58, 1.93)$
Inflation	$(-2.04, 2.04)$	$(-1.26, 1.61)$
Wage	$(-3.24, 3.24)$	$(-0.46, 1.58)$
Govt Debt	$(-7.27, 7.27)$	$(2.34, 5.83)$
Investment	$(-0.20, 0.20)$	$(-0.07, 0.05)$

Note: This table compares 68% intervals of B_{Xz} corresponding to each variable in the case where we do not incorporate the information on direct effects (left column), and the case where we incorporate the information (right column). Every entry in the table is scaled by 10^{-3}.

here is the total effect, defined as the sum of direct and indirect effects. This result suggests that we can tightly identify the total density response from the information on the direct effect.

Turning to the aggregate variables, we find that the mean of the aggregate impulse responses aligns well with the baseline. Although the size of the wage response produced by our methodology is larger than that by the baseline, our methodology recovers the overall features of the baseline responses. Moreover, our identification methodology helps to shrink the uncertainty in the responses relative to the prior. To see this more closely, Table 3 displays the 68% credible intervals of B_{Xz} for the case where we do not incorporate the information on the direct effect (henceforth, the uninformed interval), and the case where we incorporate the information (the informed interval). All values are reported in units of 10^{-3} for exposition.[18] For instance, the 68% informed interval of the output response ranges from -0.58 to 1.93. This is a huge reduction in the uncertainty given that the uninformed interval covers $(-3.47, 3.47)$. In other words, the length of the informed interval is 36% of that of the uninformed interval. In addition, the informed interval places greater weight on positive responses where the true response lies, contrary to the uninformed interval being symmetric around zero. These observations illustrate the benefit of our identification strategy for recovering the true responses.

[18]The former is derived from the prior of B_{Xz} discussed above. We use the same draws used to depict Figure 4 to compute the latter.

4.6 Robustness

We conduct several exercises with different specifications to see the robustness of our methodology. We overview the results of these exercises here. Appendix E contains figures and more detailed discussions.

Additional Restrictions. As demonstrated above, our framework allows to reflect additional prior belief for responses of aggregate variables to the shock of interest. As an illustration, we restrict the at-impact responses of output and government debt (i.e., the entries of B_{Xz} corresponding to output and government debt) to be positive. That is, we require that output and government debt increases in response to the contemporaneous transfer shock.

Figure 15 plots the responses. The additional sign restrictions greatly shrinks the probability bands relative to the baseline. The impulse responses for output and government debt are associated with much smaller uncertainties over the entire horizon. Interestingly, the interval for the at-impact output response is strictly smaller than the one in Figure 4, i.e., the upper bound of the interval becomes close to the baseline, even though we restrict the response to be positive. Moreover, uncertainties surrounding the responses of non-restricted variables (inflation, wage, and investment) are also reduced. These observations suggest that even agnostic sign restrictions contribute to improving the identification.

Choice of ρ. Figures 16 and 17 repeat the same exercises for alternative assumptions on ρ, namely 0.85 and 0.95. They still capture the baseline well, while we find some discrepancies (e.g., the inflation reacts negatively for $\rho = 0.85$, and overreacts for $\rho = 0.95$). This highlights that the choice of ρ plays an important role in our exercise.

Joint Bayes Estimator. There is a criticism for using point-wise percentiles to summarize uncertainty in impulse responses because they do not take into account dynamics of impulse response functions (i.e., shape of responses) to derive the point estimator as well as the credible intervals (e.g., Inoue and Kilian 2022). This is particularly the case for our exercise because we are interested in the response of cross-sectional density, whose shape does matter for interpretations. Figures 18 and 19 compare the joint posterior distribution under the additive separable loss function with the baseline estimates. Except that the joint Bayes estimator comes with the slightly wider credible interval, the point-wise posterior reported above is very similar to the joint posterior.

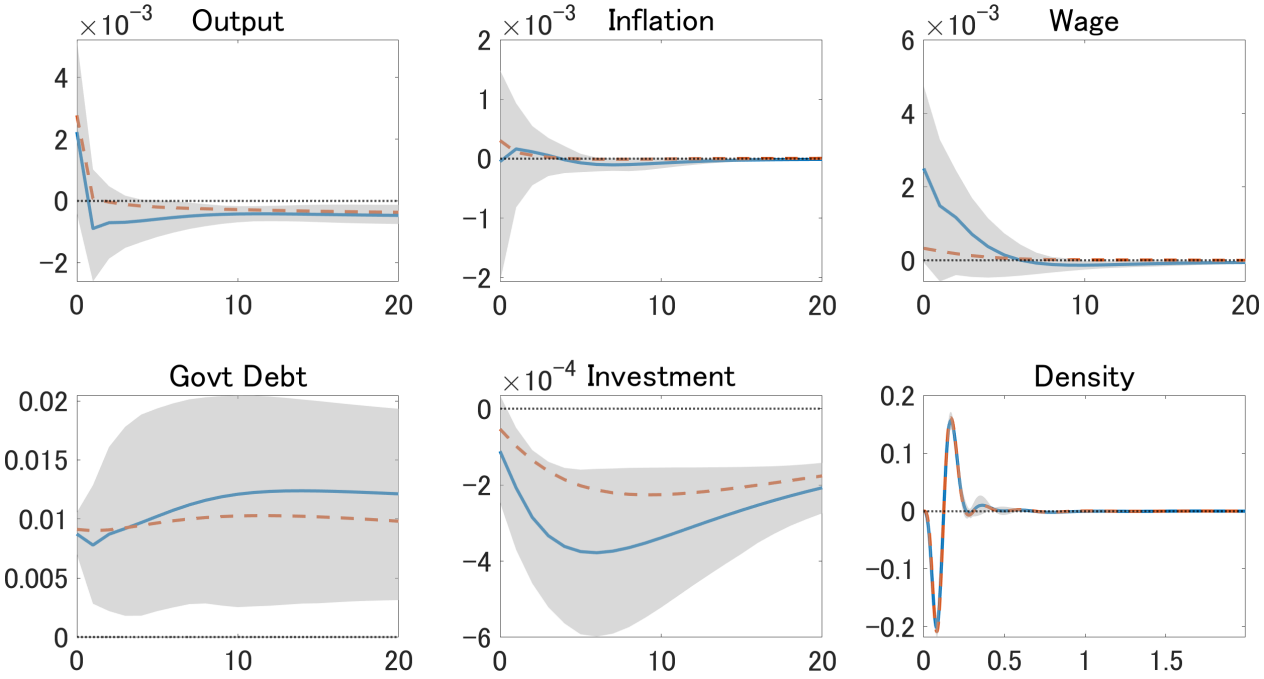

Figure 5: Impulse Responses to Targeted Transfer

Note: See the footnote attached to Figure 4. We set $\rho = 0.9$.

Prior Strength. We weaken the prior of (B_{fz}) and B_{zz} by setting the prior standard deviation to be the absolute value of their prior mean multiplied by 0.5, instead of 0.1 above. Since those parameters are key for identification, we aim to see what happens if these restrictions speak little compared to the baseline. The output from the estimation is shown in Figure 20. Although the posterior mean is consistent with the previous estimation results overall, they come with quite large uncertainties represented by the wide intervals. This suggests that setting the informative prior for these parameters is crucial for identification.

4.7 Extension: Targeted Transfer

The scope for our methodology is not limited to lump-sum transfers. As long as the policy design $\eta(\cdot)$ and heterogeneous direct effects $\phi(\cdot)$ are known, we can compute the direct response of the density associated under an alternative intervention and use it as an input for identification. As an extension, we consider the same economy but analyze the targeted

transfer stimulus policy. Specifically, we change $\eta(\cdot)$ to be

$$\eta(a,b,e) = \begin{cases} \frac{1}{\mathbb{P}(e \leqslant \bar{e})} & \text{if } e \leqslant \bar{e} \\ 0 & \text{otherwise} \end{cases}$$

Households with productivity below the threshold $\bar{e}$ receive the stimulus payment, while others do not receive any transfers under this policy. This specification still yields $\int \eta(\cdot) d\mu_{ss} = 1$, and thus the policy is normalized to the same scale as the lump-sum setting. The MPC $\phi(\cdot)$ remains identical because the household optimization problem, aside from the transfer policy, remains unchanged. We assume that the bottom 21.2% of households ranked by the productivity are eligible for the transfer stimulus (i.e., the threshold $\bar{e}$ is set at the 21.2nd percentile of the productivity distribution).

We compute the direct response F_ε to the targeted transfer stimulus and apply our methodology to see the propagation of the shock. Figure 5 shows the impulse responses of the macroeconomic variables and the consumption density. Again, the orange dashed line represents the true impulse response. The at-impact response of output is almost three times larger than the response to the lump-sum transfer policy even though these two policies exhibit the same scale. The mechanism behind it is quite standard. The constrained households typically face low productivity. They have a higher MPC than others, and thus spend a large share of the stimulus payment immediately. The policy targeted toward the low productivity households increases output more effectively.

The mean of impulse responses from our identification methodology is shown by the blue solid line. It captures the truth well, with the 68% credible interval covering the baseline. This result indicates that the proposed methodology performs well in the alternative setting.

5 Empirical Application: Stimulus Transfer

This section applies the methodology developed so far to investigate the dynamic propagation of stimulus payments to households. We consider two policy scenarios, (i) lump-sum transfer giving \$100 per family member, and (ii) targeted transfer to households in the bottom 20 percent of the income distribution. After describing the dataset, we demonstrate how we compute the household-level MPC. Then we turn to the estimation results.

5.1 Data Description

The dataset consists of quarterly observations of aggregate variables and consumption densities from 1990Q1 to 2019Q4. See Appendix F for details on the construction of the variables. In the baseline specification, we include output, inflation, Wu and Xia (2016) shadow Federal Funds rate, Federal tax revenues, and Federal net transfer payments. Output, tax revenues, and net transfer payments are detrended by the Congressional Budget Office (CBO) estimate of potential output. In the MAR model (10), the net transfer payment is assumed to be directly affected by the shock of interest and thus labeled as z_t, and the other four variables are included in X_t.

We use the Consumer Expenditure Survey (CEX) collected by the Bureau of Labor Statistics as a source for household consumption. CEX consists of quarterly interview surveys meant to capture relatively large expenditure, and weekly diary surveys meant to capture daily expenditure. We mainly use the interview survey because the data collected from the diary survey is summed up across time and appears in the interview survey anyway. The unit of observation is the consumer unit (CU), and CUs are asked about expenditure over the last three months. CEX has a rotating panel structure: Each CU is interviewed for up to five consecutive quarters, while the first interview is preliminary and not used for statistical analysis. They are also asked about their characteristics, including family income over the last 12 months in the first and fifth interview, and information on financial status in the fifth interview.

Our measure for consumption expenditure is the CEX benchmark measure of total expenditure, net of personal insurance, pension, and social security payment. This consumption measure covers food, alcoholic beverages, apparel, housing, transportation, health care, entertainment, personal care, reading, education, tobacco, cash contribution, and miscellaneous. Note that, since the survey is conducted every month and it asks about expenditure for the past three months, reported expenditure does not necessarily align with calendar quarter. We assign each observation to the quarter in which at least two of the reported months fall. For example, observations in 2010Q1 include those who are interviewed in March 2010 (reported expenditure covers December 2009, January 2010, and February 2010), April 2010 (covering January 2010, February 2010, and March 2010), and May 2010 (covering February 2010, March 2010, and April 2010).[19] The expenditure measure is then annualized by multiplying quarterly expenditure by four. We exclude the observations with negative expenditure.

We also use information on CU characteristics, such as annual pre-tax family income

[19] A similar treatment is made in M. Chang and Schorfheide (2024).

and family structure. The consumption measure and income are deflated by the Consumer Price Index. The consumption density for each quarter is estimated by the kernel density estimation method with CU weights provided by the CEX. The expenditure in the CEX is not seasonally adjusted. To remove seasonality, we first extract 20 FPCs and associated loadings from the time-series of demeaned densities. Then, we apply the X-13 seasonally adjustment for each loading. We finally combine the seasonally adjusted loadings with the FPCs to obtain the seasonally adjusted series of densities.

5.2 Computing Household-Level MPC

It is not possible to estimate MPC for every quarter from CEX without further assumptions or information. As MPC is defined as a change of consumption expenditure in response to an unexpected change in income, we need to decompose income changes into unexpected and expected components. This requires us to find a plausible micro identification strategy, or to construct a structural model of consumption behavior in order to extract the unexpected component. In addition, MPCs are heterogeneous across households, but further assumptions are needed about how they relate to household characteristics. To the best of our knowledge, there is no micro data for the joint distribution of consumption and MPC available for a long time span for the United States.

We follow the "reported-preference" approach and leverage the survey evidence on self-reported MPC to impute household-level MPC. Despite the caveat that self-reported responses might be different from actual household behavior, this approach elicits household responses to unexpected income changes in a way that is consistent with economic theory. Another advantage of such survey is in allowing us to relate MPC to various household characteristics.

We take an imputation approach similar to that of Patterson (2023) and Bellifemine et al. (2025), and use the survey evidence by Fuster et al. (2021). The survey was conducted in March 2016, May 2016, January 2017, and March 2017 as an additional module of the Survey of Consumer Expectations (SCE) operated by the Federal Reserve Bank of New York. This survey asks respondents to consider various hypothetical scenarios involving unexpected income changes, and asks how much they would adjust expenditures within a quarter. We focus on the treatment of a \$500 gain for respondents in March 2016 and a \$500 loss in March 2017.[20] We first relate the reported MPC to household characteristics, namely education,

[20]One of the main findings of Fuster et al. (2021) is that there is a substantial difference between the MPCs with respect to income gain and income loss. Since our methodology is built on linearity, we take the

race, sex, home ownership, working status, pre-tax household income, and age. Specifically, we run the following regression twice: The first regression uses the observations getting the gain treatment, and the second regression uses those getting the loss treatment.

$$\begin{aligned} MPC_i =& \alpha + \sum_{j=1}^{3} \beta_{educ,j} \mathbf{1}\{educ_i = j\} + \sum_{j=1}^{3} \beta_{race,j} \mathbf{1}\{race_i = j\} + \beta_{female} \mathbf{1}\{female_i = 1\} \\ &+ \beta_{rent} \mathbf{1}\{rent_i = 1\} + \sum_{j=1}^{2} \beta_{ws,j} \mathbf{1}\{ws_i = j\} + \sum_{j=1}^{10} \beta_{income,j} \mathbf{1}\{income_i = j\} \\ &+ \beta_{age,1} age_i + \beta_{age,2} age_i^2 + \beta_{age,3} age_i^3 + u_i, \quad i = 1, \cdots, n \end{aligned} \tag{18}$$

where $female_i$ equals one if the respondent is female and zero otherwise, $rent_i$ equals one if the respondent rents his/her residence and zero if he/she owns the residence, and age_i is age of the respondent. See footnote for definition of other variables.[21] We then use the estimated coefficient from regression (18) to impute MPC from income gain, $\widehat{MPC}_i^{gain}$, and MPC from income loss, $\widehat{MPC}_i^{loss}$, respectively for the CEX observations.[22] The imputed MPC is defined as the simple average of $\widehat{MPC}_i^{gain}$ and $\widehat{MPC}_i^{loss}$.

An important assumption for the imputation approach is that MPC depends only on variables included in the regression (18). In particular, it is widely recognized that liquid assets are one of the important components to explain MPC. The survey by Fuster et al. (2021) do have contain questions about liquid asset holdings, and CEX asks CUs to report their wealth at the fifth interview. However, it is known that asset information provided in CEX is of low quality due primarily to a high non-response rate. Dropping CUs who do not report their asset holdings may distort the results because such non-response might occur for systematic reasons. Moreover, there was a significant change in the definition of liquid asset in CEX. Until 2013Q1, the measure of liquid asset was typically constructed by summing up balances in checking and savings accounts (e.g., Parker et al. 2013), while a new variable

simple average of the two MPCs. Allowing for such sign asymmetry requires incorporating a certain type of nonlinearities in the framework, which is beyond the scope of this paper.

[21] We construct dummy variables based on the following definition of groups. Education: High school diploma or lower/ Attended college but not BA (including people with associate degrees)/ BA/ Master's or higher. Race: White/ Black/ Asian/ Others. Working Status: Employed (Full-time or part-time)/ Unemployed/ Not in labor force. Pre-tax household income: Less than \$10k/ \$10k-20k/ \$20k-30k/ \$30k-40k/ \$40k-50k/ \$50k-60k/ \$60k-75k/ \$75k-100k/ \$100k-150k/ \$150k-200k/ \$200k or more. We drop dummies for BA, White, Employed, and \$60k-75k because these categories are used as the reference groups.

[22] As described in the main text, CEX asks about family income only in the first and fifth interview. Our assumption is that households stay in the same income category as the one reported in the first interview for the second, third, and fourth interview.

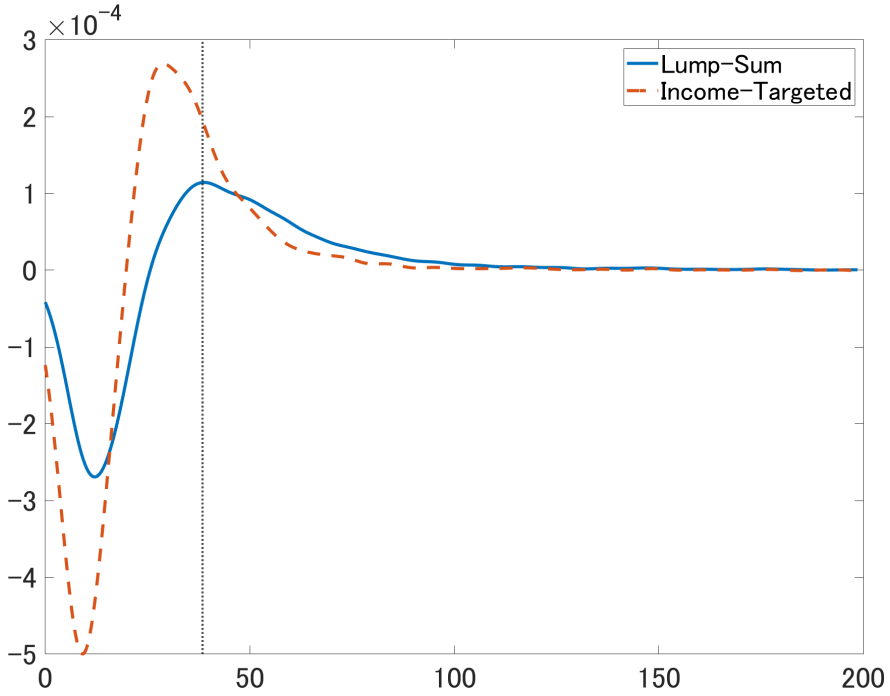

Figure 6: Direct Effect to Consumption Density

Note: This figure shows the direct response of each policy shock to the consumption density. The x-axis is the annualized consumption expenditure in \$1,000. The dotted vertical line shows the mean (first moment) computed from the temporal mean of the consumption densities.

representing liquid asset was added in 2013Q2, which covers values in money market accounts and certificates of deposit in addition to bank balances. This discontinuity creates inconsistencies in the imputation. For this reason, we do not use the financial information explicitly and assume that financial status is captured by the variables included in the regression.

Another assumption is that the distribution of MPC is stable over time. Since we impose linearity, the MPC function $\phi(\cdot)$ in equation (6) should be the one at the steady-state and thus should not be affected by aggregates. The predicted MPC from (18) should be interpreted as the steady-state MPC, although estimated from observations in 2016 and 2017. Since the interview waves are concentrated in the short period of time, it is not possible to investigate the relationship between MPC and the business cycle. Patterson (2023) reports evidence that the contribution of the unemployment rate to the MPC is small both economically and statistically, favoring the view that MPCs do not exhibit substantial variation over time.

5.3 Direct Effect to Consumption Density

We consider two types of policy interventions: lump-sum and income-targeted stimulus transfers. For the lump-sum policy, the government allocates $100 per family member to CUs. The income-targeted stimulus policy is designed as follows. For each quarter, we sort CUs based on the family income divided by the square root of the number of family members, a standard adjustment to account for economies of scale in family expenditures. Recipients of the targeted stimulus are those in the bottom 20% of the population based on adjusted family income, where the cutoff is determined by population shares rather than the number of CUs. The selected CUs receive $500 per family member as the stimulus check. This design ensures that those two policies are of the approximately identical aggregate scale.

The direct effect F_ε is computed as follows. Let $(c_{it})_i$ be the collection of consumption expenditures of CUs at time t. We then compute a hypothetical dataset. In the case of lump-sum payments, each observation of the dataset is $(c_{it} + 100 \times N_{it} \times \widehat{MPC}_{it})_i$ where N_{it} is the number of family members in CU i and $\widehat{MPC}_{it}$ is the imputed MPC based on household characteristics at time t. We construct the hypothetical dataset for the targeted policy analogously. We subtract the time average of densities associated with the original dataset from the time average of densities associated with the hypothetical dataset. The resulting difference is our measure for the direct effect on consumption density.

Figure 6 shows the direct effect computed as described above. The shape of the direct effect resembles the one produced by the quantitative model (Figure 1). The lump-sum transfer shifts the households at the lower-end of the distribution to the right, while it has little effect on other households. Although the mean associated with the time-averaged consumption density is located closer to the peak of the response compared to the model counterpart, the similarity in shape indicates that our quantitative HANK model captures household behavior well, supporting the discussion in the previous section. The direct effect associated with the targeted policy is more concentrated on the low-consumption groups: the density declines more strongly among low-consumption households and rises more steeply just below the mean, reflecting a more concentrated shift in mass toward middle consumption levels.

5.4 Settings

Below we discuss some choices we made concerning basis functions, prior, and algorithm.

Table 4: Sign Restrictions on B_{XX} (Empirics)

Shock	Output	Inflation	Shadow Rate	Tax
Demand Shock	$+^*$	$+^*$		
Supply Shock	$-^*$	$+^*$		
Monetary Policy Shock	$-$	$-$	$+$	
Tax Shock	$-$			$+$

Note: Sign restrictions imposed on B_{XX}. The signs "+" and "−" indicates that the element is restricted to be positive or negative. We impose additional assumptions that, for output and inflation, the sum of squares of at-impact responses to the demand and supply shocks are greater than the squares of responses to other shocks.

Choosing Lag Order and Number of Basis Functions. We continue to use the functional principal component basis to reduce the dimensionality of the functional observations. We again minimize the average of one-step ahead forecast errors from the rolling-window estimations to select m and p jointly, which yields $m = 2$ and $p = 2$. However, as these two basis functions explain only 84.6% of variation in (f_t), we use $m = 4$ in our analysis. This increases the accountability up to 92.3%. The lag order of $p = 2$ is still optimal under $m = 4$.

Prior. Contrary to the simulation exercise where we fixed the reduced form parameters, we also estimate them in the application. We specify the prior for (G) and (Σ) to be the normal-inverse-Wishart distribution. This is one of the typical distributional choices for reduced form VAR models because, as the posterior associated with this type of prior is known analytically, we can make draws without simulation. We parametrize the distribution by incorporating the structure similar to the Minnesota prior where we heavily shrink coefficients associated with higher order lags. See Appendix F for further details.

We follow the same approach as in the simulation exercises to settle on the prior for other structural parameters. The prior mean for B_{zz}, the parameter governing the scale of the policy, is set as follows. For each period, we multiply \$100 with the total population to estimate the total budget spent on the transfer policy, and divide it by the CBO nominal potential output. We set the prior mean of B_{zz} to be the time-average of it.

The prior distribution for B_{XX} is again uniform combined with sign restrictions to distinguish aggregate shocks and functional shocks. We present the sign restrictions on B_{XX} in Table 4. We require that the demand shock changes output and inflation to the same direction and supply shock changes them to the opposite direction. The positive monetary policy shock decreases output and inflation. The tax shock is assumed to be contractionary.

We further impose that demand and supply shocks in sum are primary drivers for output and inflation by assuming that the sum of squares of at-impact responses to those two shocks are larger than squares of at-impact responses to any other shocks. We impose this assumption to mitigate concerns for the "shock masquerading problem" (Wolf 2020). We specify the prior for B_{Xz} to be uniform, implying that we do not inform any beliefs on how aggregate variables respond to the transfer shock.

We choose ρ, the shrinkage parameter for indirect effects, by the following procedure. Fixing the reduced parameters at the posterior mean, we conduct a preliminary estimation under $\rho = 0.9$. We then find ρ which minimizes the approximation error in (???) for the joint Bayes estimator from this preliminary estimation, and use the resulting ρ in the actual estimation. This procedure selects $\rho = 0.86$ under the lump-sum transfer, and $\rho = 0.81$ under the targeted transfer. Unlike the simulation exercise where we used the true direct effect, the direct effect F_ε is based on the estimated MPCs. To account for additional uncertainty owing to the estimation, we multiply the absolute value of $(\widetilde{B_{fz}})$ with 0.25 to determine the prior standard deviation of (B_{fz}) conditional on (A_{fz}), instead of 0.1 in the simulation exercise.

5.5 Impulse Responses

Figure 7 plots the impulse responses to a positive lump-sum transfer shock (the first and second rows) and a positive targeted transfer shock (the third and fourth rows) where the at-impact response of aggregate transfer is normalized to be 0.01 relative to potential output. Overall, the responses of aggregate variables are almost muted, although they are associated with wide credible bands. The finding that the cash transfers are not effective tools to stimulate output is consistent with the evidence discussed by Ramey (2025). The shock generates the persistent increase in transfer payments, but little changes to tax revenues, which suggests that the stimulus policy is mostly deficit-financed.

The at-impact responses of the consumption density imply that these policies cause upward shifts in the consumption at the left end of the distribution. To see the effect of the stimulus transfer policies to the consumption inequality more closely, Figure 8 plots the impulse responses of Gini coefficient, and 10th, 50th, and 90th percentiles. The red markers at horizon zero show the changes in these statistics due only to the direct effect, which can be computed from Figure 6. Both policies increases the 10th and 50th percentiles persistently, but has little effects on the 90th percentile. This leads to the reduction in the consumption inequality represented by the Gini coefficient. The targeted transfer is more effective at reducing the inequality.

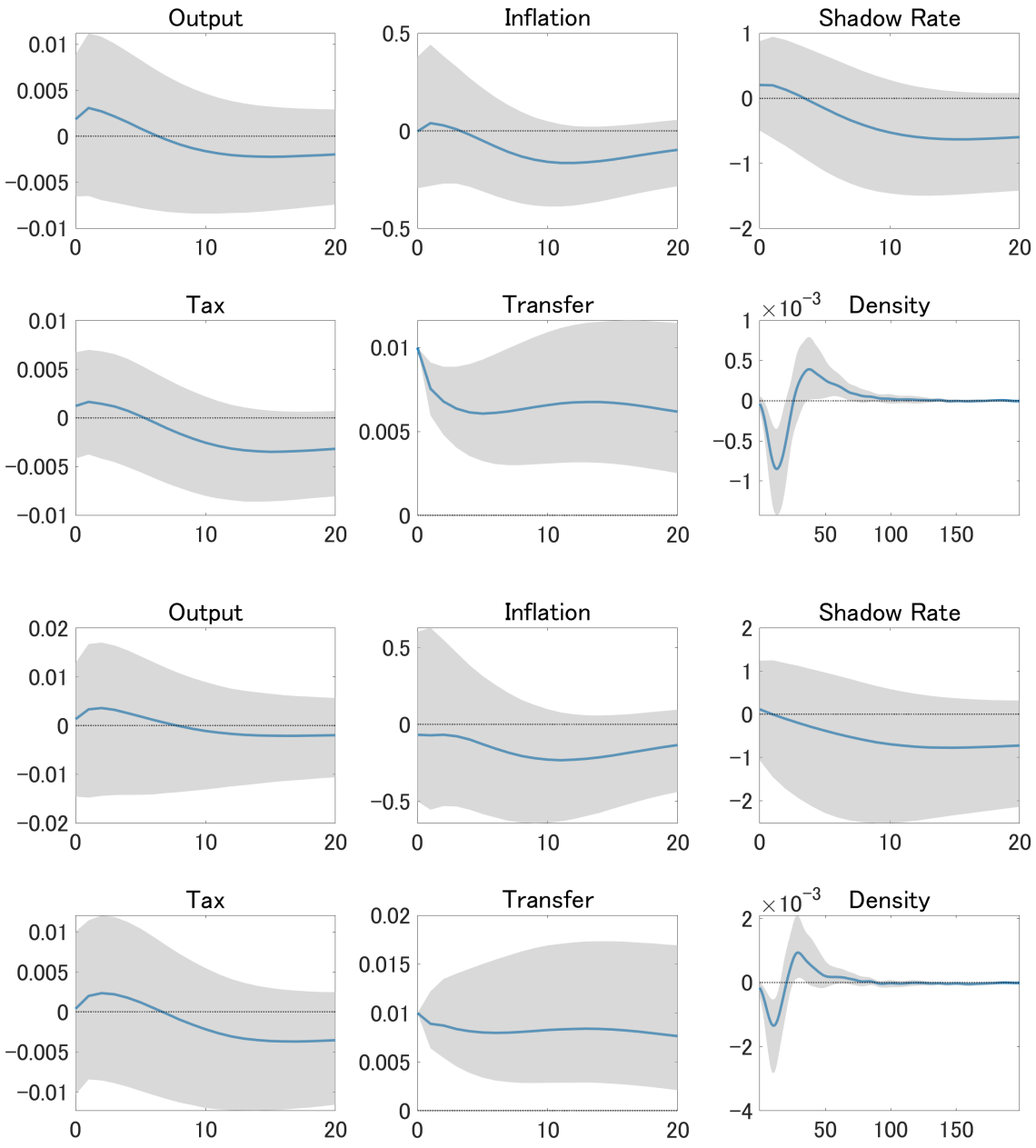

Figure 7: Impulse Responses to a Lump-Sum Transfer Shock (First and Second Rows) and to a Targeted Transfer Shock (Third and Fourth Rows)

Note: The bottom-right panel plots the at-impact impulse response of the density. The other panels show the impulse response of aggregate quantities. The blue solid line shows the point-wise median, along with the 68% credible interval represented by the shaded area. The at-impact response of transfer is normalized to be 0.01.

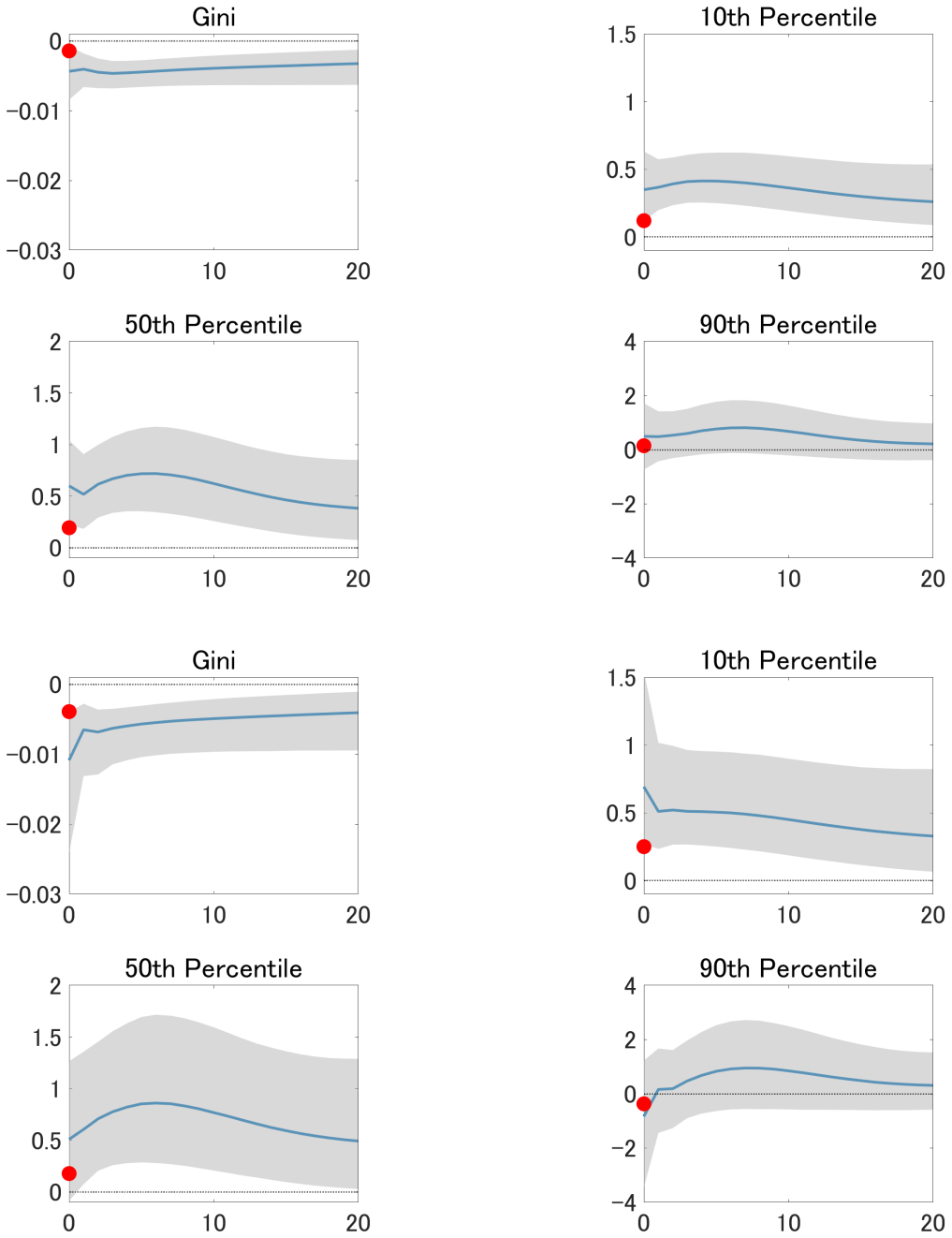

Figure 8: Impulse Responses of Summary Statistics of Consumption Distribution to a Lump-Sum Transfer Shock (First and Second Rows) and to a Targeted Transfer Shock (Third and Fourth Rows)

Note: The blue solid line shows the point-wise mean, along with the 68% credible interval represented by the shaded area. The red markers at horizon zero show the change in these statistics due only to the direct effect. The at-impact response of transfer is normalized to be 0.01.

Relative to the direct effects plotted by the read markers, the at-impact total effects are greater in magnitude for the Gini coefficient as well as 10th and 50th percentiles for both policies. This observation suggests that the indirect effect pushes down the inequality further, and highlights the importance of general equilibrium mechanism in evaluating the distributional consequence of aggregate shocks.

5.5.1 Robustness

discuss robustness

6 Conclusion

This paper develops a novel identification methodology leveraging heterogeneous direct effects. We show how these direct effects can be incorporated in an autoregressive model featuring both aggregate and functional observations. The proposed methodology performs well in the quantitative HANK model. We apply the method to investigate the macroeconomic and distributional effects of stimulus payment policies, and compare the lump-sum and targeted stimulus payments.

Although we presented the framework mainly in the context of stimulus transfer policies, this methodology has a broader range of applications. For example, the direct effects of trade shocks to firms' sales and employment depend on the firms' characteristics such as productivity (e.g., Aghion et al. 2024). Our identification scheme allows us to evaluate the consequence of trade in terms of aggregate quantities and the distribution of firms. Our methodology can be applicable to examples other than the stimulus transfer and trade shocks. These alternative applications are left for future research.

References

Adams, J. J., & Barrett, M. P. (2025). Identifying news shocks from forecasts. Working Paper.

Aghion, P., Bergeaud, A., Lequien, M., Melitz, M. J., & Zuber, T. (2024). Opposing firm-level responses to the china shock: Output competition versus input supply. *American Economic Journal: Economic Policy*, *16*(2), 249–269.

Alves, F., Kaplan, G., Moll, B., & Violante, G. L. (2020). A further look at the propagation of monetary policy shocks in hank. *Journal of Money, Credit and Banking*, *52*(S2), 521–559.

Ampudia, M., Cooper, R., Le Blanc, J., & Zhu, G. (2024). Mpc heterogeneity and the dynamic response of consumption to monetary policy. *American Economic Journal: Macroeconomics*, *16*(3), 343–388.

Angeletos, G.-M., Collard, F., & Dellas, H. (2020). Business-cycle anatomy. *American Economic Review*, *110*(10), 3030–3070.

Arias, J. E., Rubio-Ramírez, J. F., & Waggoner, D. F. (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. *Econometrica*, *86*(2), 685–720.

Auclert, A., Bardóczy, B., Rognlie, M., & Straub, L. (2021). Using the sequence-space jacobian to solve and estimate heterogeneous-agent models. *Econometrica*, *89*(5), 2375–2408.

Auclert, A., Rognlie, M., & Straub, L. (2020). Micro jumps, macro humps: Monetary policy and business cycles in an estimated hank model. Working Paper.

Barnichon, R., & Matthes, C. (2018). Functional approximation of impulse responses. *Journal of Monetary Economics*, *99*, 41–55.

Baumeister, C., & Hamilton, J. D. (2015). Sign restrictions, structural vector autoregressions, and useful prior information. *Econometrica*, *83*(5), 1963–1999.

Bayer, C., Born, B., & Luetticke, R. (2024). Shocks, frictions, and inequality in us business cycles. *American Economic Review*, *114*(5), 1211–1247.

Bellifemine, M., Couturier, A., & Jamilov, R. (2025). The regional keynesian cross. Working Paper.

Blanchard, O., & Perotti, R. (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *Quarterly Journal of Economics*, *117*(4), 1329–1368.

Boppart, T., Krusell, P., & Mitman, K. (2018). Exploiting mit shocks in heterogeneous-agent economies: The impulse response as a numerical derivative. *Journal of Economic Dynamics and Control*, *89*, 68–92.

Bosq, D. (2000). *Linear processes in function spaces: Theory and applications* (Vol. 149). Springer Science & Business Media.

Bruns, M., & Piffer, M. (2023). A new posterior sampler for bayesian structural vector autoregressive models. *Quantitative Economics*, *14*(4), 1221–1250.

Chan, J. (2019). Large bayesian vector autoregressions. In *Macroeconomic forecasting in the era of big data: Theory and practice* (pp. 95–125). Springer.

Chang, M., Chen, X., & Schorfheide, F. (2024). Heterogeneity and aggregate fluctuations. *Journal of Political Economy*, *132*(12), 4021–4067.

Chang, M., & Schorfheide, F. (2024). On the effects of monetary policy shocks on income and consumption heterogeneity. Working Paper.

Chang, Y., Gómez-Rodríguez, F., & Hong, G. H. (2022). The effects of economic shocks on heterogeneous inflation expectations. Working Paper.

Chang, Y., Gómez-Rodríguez, F., & Matthes, C. (2023). The influence of fiscal and monetary policies on the shape of the yield curve. Working Paper.

Chang, Y., Kim, S., & Park, J. (2025). How do macroaggregates and income distribution interact dynamically? a novel structural mixed autoregression with aggregate and functional variables. Working Paper.

Chang, Y., Miller, J. I., & Park, J. (2024a). Shocking climate: Identifying economic damages from anthropogenic and natural climate change. Working Paper.

Chang, Y., Park, J., & Pyun, D. (2024b). From functional autoregressions to vector autoregressions. Working Paper.

Chodorow-Reich, G. (2019). Geographic cross-sectional fiscal spending multipliers: What have we learned? *American Economic Journal: Economic Policy*, *11*(2), 1–34.

Coibion, O., Gorodnichenko, Y., Kueng, L., & Silvia, J. (2017). Innocent bystanders? monetary policy and inequality. *Journal of Monetary Economics*, *88*, 70–89.

Doh, T., & Smith, A. L. (2022). A new approach to integrating expectations into var models. *Journal of Monetary Economics*, *132*, 24–43.

Fagereng, A., Holm, M. B., & Natvik, G. J. (2021). Mpc heterogeneity and household balance sheets. *American Economic Journal: Macroeconomics*, *13*(4), 1–54.

Ferreira, L. N., Miranda-Agrippino, S., & Ricco, G. (2025). Bayesian local projections. *Review of Economics and Statistics*, *107*(5), 1424–1438.

Fuster, A., Kaplan, G., & Zafar, B. (2021). What would you do with $500? spending responses to gains, losses, news, and loans. *Review of Economic Studies*, *88*(4), 1760–1795.

Heathcote, J., Storesletten, K., & Violante, G. L. (2017). Optimal tax progressivity: An analytical framework. *Quarterly Journal of Economics*, *132*(4), 1693–1754.

Herreno, J. (2023). Aggregating the effect of bank credit supply shocks on firms. Working Paper.

Huber, F., Marcellino, M., & Tornese, T. (2024). The distributional effects of economic uncertainty. *arXiv preprint arXiv:2411.12655*.

Huber, K. (2023). Estimating general equilibrium spillovers of large-scale shocks. *The Review of Financial Studies*, *36*(4), 1548–1584.

Iao, M. C., & Selvakumar, Y. J. (2024). Estimating hank with micro data. Working Paper.

Inoue, A., & Kilian, L. (2022). Joint bayesian inference about impulse responses in var models. *Journal of Econometrics*, *231*(2), 457–476.

Inoue, A., & Rossi, B. (2021). A new approach to measuring economic policy shocks, with an application to conventional and unconventional monetary policy. *Quantitative Economics*, *12*(4), 1085–1138.

Jappelli, T., & Pistaferri, L. (2014). Fiscal policy and mpc heterogeneity. *American Economic Journal: Macroeconomics*, *6*(4), 107–136.

Kilian, L., & Lütkepohl, H. (2017). *Structural vector autoregressive analysis*. Cambridge University Press.

Koop, G., & Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, *3*(4), 267–358.

Krusell, P., & Smith, A. A., Jr. (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, *106*(5), 867–896.

Lewis, D., Melcangi, D., & Pilossoph, L. (Forthcoming). Latent heterogeneity in the marginal propensity to consume. *Review of Economic Studies*.

Mas, A. (2007). Weak convergence in the functional autoregressive model. *Journal of Multivariate Analysis*, *98*(6), 1231–1261.

Matthes, C., Nagasaka, N., & Schwartzman, F. (2024). Estimating the missing intercept. Working Paper.

Matthes, C., & Schwartzman, F. (Forthcoming). The consumption origins of business cycles: Lessons from sectoral dynamics. *American Economic Journal: Macroeconomics*.

Meeks, R., & Monti, F. (2023). Heterogeneous beliefs and the phillips curve. *Journal of Monetary Economics*, *139*, 41–54.

Nakamura, E., & Steinsson, J. (2014). Fiscal stimulus in a monetary union: Evidence from us regions. *American Economic Review*, *104*(3), 753–792.

Parker, J. A., Souleles, N. S., Johnson, D. S., & McClelland, R. (2013). Consumer spending and the economic stimulus payments of 2008. *American Economic Review*, *103*(6), 2530–2553.

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

Petersen, A., & Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a hilbert space. *Annals of Statistics*, *44*(1), 183–218.

Plagborg-Møller, M. (2019). Bayesian inference on structural impulse response functions. *Quantitative Economics*, *10*(1), 145–184.

Ramey, V. A. (2025). Do temporary cash transfers stimulate the macroeconomy? evidence from four case studies. *IMF Economic Review*, 1–35.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical science*, 425–441.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. Springer.

Rotemberg, J. J. (1982). Sticky prices in the united states. *Journal of Political Economy*, *90*(6), 1187–1211.

Sarto, A. P. (2025). Recovering macro elasticities from regional data. Working Paper.

Stock, J. H., & Watson, M. W. (2018). Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *The Economic Journal*, *128*(610), 917–948.

Uhlig, H. (2004). What moves gnp?. Working Paper.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics*, *52*(2), 381–419.

Wolf, C. K. (2020). Svar (mis) identification and the real effects of monetary policy shocks. *American Economic Journal: Macroeconomics*, *12*(4), 1–32.

Wolf, C. K. (2023). The missing intercept: A demand equivalence approach. *American Economic Review*, *113*(8), 2232–2269.

Wu, J. C., & Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking*, *48*(2-3), 253–291.

A Details on HANK

We construct a medium-scale HANK model as a laboratory for simulation analysis. Time is discrete and all agents form rational expectations.

A.1 Household

There are infinitely lived households indexed by $i \in [0,1]$ who consume $c_{i,t}$ and supply labor h_t. The labor supply is determined by the labor union and is the same for all households. The log of idiosyncratic productivity e follows an AR(1) process:

$$\log e' = \rho_e \log e + \sigma_e \varepsilon^e$$

and e is normalized so that $\mathbb{E}(e) = 1$. The household can save in liquid asset b and illiquid asset a. They can adjust the illiquid asset with i.i.d. probability p. As a liquidity premium, the households pay ω per one unit of liquid assets. The Bellman equation is characterized as

$$\begin{aligned} V_t^h(1,a,b,e) = &\max_{c,a',b'} \left\{ \frac{c^{1-\gamma}}{1-\gamma} - \varphi \frac{h^{1+\nu}}{1+\nu} \right. \\ &\left. + \beta \left[p \mathbb{E}_t V_{t+1}^h(1,a',b',e') + (1-p) \mathbb{E}_t V_{t+1}^h(0,a',b',e') \right] \right\} \\ V_t^h(0,a,b,e) = &\max_{c,b'} \left\{ \frac{c^{1-\gamma}}{1-\gamma} - \varphi \frac{h^{1+\nu}}{1+\nu} \right. \\ &\left. + \beta \left[p \mathbb{E}_t V_{t+1}^h(1,(1+r_{p,t})a,b',e') + (1-p) \mathbb{E}_t V_{t+1}^h(0,(1+r_{p,t})a,b',e') \right] \right\} \end{aligned}$$

subject to

$$\begin{aligned} c + a' + b' &= (1+r_{p,t})a + (1+r_{p,t}-\omega)b + (1-\tau_t^y) y_t(e)^{1-\xi} + \left(1-\tau^{\Pi}\right) \Pi_t(e) + \eta(a,b,e)\sigma^{tr}\varepsilon_t^{tr} \\ y_t(e) &= w_t h_t \Gamma_t(e) \\ a' &\geqslant 0, \quad b' \geqslant 0 \end{aligned}$$

If the first input in the value function is 1, the household is allowed to adjust its illiquid asset. Otherwise, it keeps the same level of illiquid asset holding. The household receives the gross return from asset holdings, the post-tax labor income $(1-\tau^y)y^{1-\xi}$ and dividend income $(1-\tau^D)\Pi$, and receives transfer payment $\eta(\cdot)\sigma^{tr}\varepsilon_t^{tr}$. The government imposes progressive tax to the pre-tax labor income $y_t(e)$. The specification follows Heathcote et al. (2017), and is known to be a good approximation of the progressive tax system in the US. The curvature parameter ξ governs the progressivity, and, given ξ, time-varying τ_t^y governs the aggregate

scale of labor taxation.

The contribution of idiosyncratic productivity e to the pre-tax income $y_t(e)$ is controlled by the incidence function $\Gamma_t(e)$ (Alves et al., 2020). The incidence function is specified as

$$\Gamma_t(e) = \frac{e\left(\frac{w_t N_t}{w_{ss} N_{ss}}\right)^{\gamma_y(e)}}{\int e'\left(\frac{w_t N_t}{w_{ss} N_{ss}}\right)^{\gamma_y(e')} \mathbb{P}(de')}$$

where $\mathbb{P}$ is the probability measure for e. Following Iao and Selvakumar (2024), $\gamma^y(e)$ is set to be $Beta(F_e(e); \alpha^y, 1)$, the probability density of the Beta distribution with parameters α^y and 1 evaluated at the cumulative probability of $\mathbb{P}$ at e, written as $F_e(e)$. This function governs the sensitivity of cross-sectional income to fluctuation of aggregate labor income. When $\alpha^y < 1$ ($\alpha^y > 1$), the sensitivity to aggregate income is higher for low (high) productivity household, implying that the standard deviation of individual income is countercyclical (procyclical).

We are interested in the propagation of transfer shock ε_t^{tr} through direct and indirect effects. The standard deviation σ_{tr} governs the size of policy. Each household receives a type-specific fraction $\eta(\cdot)$ of the total transfer. It becomes the lump-sum transfer if $\eta_t(\cdot) = \eta$ does not depend on individual characteristics. Negative η represents the lump-sum tax. Since the population size is normalized to be one, we let $\eta = 1$.

A.2 Firms

A.2.1 Final Good Producer

A final good producer combines the intermediate goods produced by a continuum of firms $j \in [0, 1]$ via the CES technology with elasticity parameter η_p.

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

Under the perfect competition, the profit maximization problem gives the demand function for intermediate goods.

$$y_{jt} = \left(\frac{p_{jt}}{P_t}\right)^{-\eta_p} Y_t \tag{19}$$

where $P_t = \left(\int p_{jt}^{1-\eta_p} dj\right)^{\frac{1}{1-\eta_p}}$.

A.2.2 Intermediate Good Producers

The intermediate good firm j produces intermediate goods used as inputs for the final good. As inputs, they use labor services bought from the labor union and their own capital. The production function is specified to be Cobb-Douglas.

$$y_{jt} = Z_t K_{jt-1}^{\alpha} N_{jt}^{1-\alpha} \tag{20}$$

where Z_t is the exogenously given total factor productivity (TFP) common across firms. They face the monopolistic competition, and choose price of their own goods given the demand functions. The Rotemberg (1982) type price adjustment cost is introduced to model price rigidity. The optimization problem firm j solves is

$$\max \mathbb{E}_0 \left\{ \sum_{t=0}^{\infty} \beta^t \left(\frac{p_{jt}}{P_t} Y_{jt} - W_t N_{jt} - r_t^k K_{j,t-1} - \frac{\eta_p}{2\kappa_p} \log \left(\frac{p_{jt}}{p_{jt-1}} \right)^2 Y_{jt} \right) \right\}$$

subject to demand function (19) and technology (20). The equilibrium is symmetric. Aggregation of the first order condition implies the price Phillips curve.

$$\log(1+\pi_t) = \kappa_p \left(mc_t - \frac{\eta_p - 1}{\eta_p} \right) + \beta \mathbb{E}_t \log(1+\pi_{t+1}) + v_t^p$$

where mc_t is the real marginal cost, $\pi_t = P_t/P_{t-1} - 1$ is the aggregate inflation rate, and v_t^p is the price markup shock.

A.2.3 Capital Good Producer

A capital good producer owns capital which is rented to intermediate good producers with price r_t^k. To make investment of an amount I_t, the producer has to pay $1 + S\left(\frac{I_{t+1}}{I_t}\right) I_t$ where $S(x) = \frac{\chi}{2}(x-1)^2$ is the investment adjustment cost. The maximization problem of the firm is given as

$$\max \mathbb{E}_0 \left\{ \sum_{t=0}^{\infty} \left(\prod_{s=0}^{t} \frac{1}{1+r_{s-1}} \right) \left[r_t^k K_t - I_t \left(1 + S\left(\frac{I_t}{I_{t-1}} \right) \right) \right] \right\}$$

subject to the law of motion for capital.

$$K_{t+1} = (1-\delta) K_t + I_t$$

The first order condition for investment implies

$$1 + S\left(\frac{I_{t+1}}{I_t}\right) + \frac{I_{t+1}}{I_t} S'\left(\frac{I_{t+1}}{I_t}\right) = Q_t + \mathbb{E}_t\left[\frac{1}{1+r_{t+1}}\left(\frac{I_{t+2}}{I_{t+1}}\right)^2 S'\left(\frac{I_{t+2}}{I_{t+1}}\right)\right]$$

where Q_t is characterized as

$$Q_t = \mathbb{E}_t\left[\frac{1}{1+r_{t+1}}\left(r^K_{t+2} + (1-\delta)Q_{t+1}\right)\right]$$

The profit of the firm is characterized as

$$\Pi^K_t = r^K_t K_t - I_t\left(1 + S\left(\frac{I_t}{I_{t-1}}\right)\right)$$

A.2.4 Mutual Fund

A mutual fund combines the stocks and government debt and sell the asset to the household. Define the aggregate profit to be the sum of profits from retailing firms and capital good producing firms.

$$\Pi_t = \Pi^R_t + \Pi^K_t = Y_t - w_t N_t - \frac{\eta_p}{2\kappa_p}\left(\log\left(1+\pi_t\right)\right)^2 Y_t - I_t\left(1 + S\left(\frac{I_t}{I_{t-1}}\right)\right),$$

the price of aggregate stock p_t whose quantity is normalized to be 1 is determined recursively as

$$p_t = \frac{\mathbb{E}_t\left[p_{t+1} + (1-\tau^D)\Pi_{t+1}\right]}{1+r_t}$$

A.2.5 Labor Union

A continuum of labor unions determines wage and labor supply under the monopolistic competition. All members in the union is subject to the same level of wage and labor supply. The labor supply chosen by each union $k \in [0,1]$ is aggregated via

$$N_t = \left(\int_0^1 N_{k,t}^{\frac{\eta_w - 1}{\eta_w}}\right)^{\frac{\eta_w}{\eta_w - 1}}$$

which gives the demand $N_{k,t} = \left(\frac{w_{k,t}}{w_t}\right)^{-\eta_w} N_t$. The objective of the unions is to maximize the utility of a hypothetical individual whose consumption and labor supply are equal to their

average. The optimization problem can be formulated as

$$\max \left\{ \sum_{t=0}^{\infty} \left(\prod_{s=0}^{t} \frac{1}{1+r_{s-1}} \right) \left[C_t^{-\gamma}(1-\tau_t^y) w_{k,t} N_{k,t} - \varphi N_{k,t}^{\nu} N_{k,t} - \frac{\varepsilon_w}{2\kappa_w} \log \left(\frac{w_{k,t}}{w_{k,t-1}}(1+\pi_t) \right)^2 \right] \right\}$$

subject to

$$N_{k,t} = \left(\frac{w_{k,t}}{w_t} \right)^{-\eta_w} N_t$$

This optimization problem leads to the standard wage Phillips curve.

$$\log(1+\pi_t^w) = \kappa_w \left(\varphi N_t^{\nu} - \frac{\eta_w - 1}{\eta_w} (1-\tau_t^y) w_t C_t^{-\sigma} \right) N_t + \frac{1}{1+r_t} \mathbb{E}_t \left[\log(1+\pi_{t+1}^w) \right] + v_t^w$$

where $\pi_t^w = \frac{w_t - w_{t-1}}{w_{t-1}}$ is the wage inflation rate and v_t^w is the exogenous wage markup shock.

A.3 Policy

The government collects labor tax, dividend tax, and type-specific tax to operate exogenous government spending and pay back government debt. The government budget constraint is given by

$$B_{t+1}^g + T_t = (1+r_t) B_t^g + G_t$$

where B_t^g is real government debt outstanding and T_t is the total tax defined as the sum of labor, dividend, and type-specific tax revenues.

$$T_t = \underbrace{\left[W_t N_t - \int (1-\tau_t^y) y_t(e)^{1-\xi} \mu_t(dadj, da, db, de) \right]}_{T_t^L} + \tau^{\Pi} \Pi_t + \int \eta(a,b,e) \mu_t(dadj, da, db, de)$$

where μ_t is the measure of household state variables (adj, a, b, e) at time t. We assume that the average tax rate to labor income T_t^L depends on the fluctuation of output and debt outstanding.

$$\frac{T_t^L}{w_t N_t} = \rho_\tau \frac{T_{t-1}^L}{w_{t-1} N_{t-1}} + (1-\rho_\tau) \left(\phi_w \left(w_t N_t - w_{ss} N_{ss} \right) + \phi_B \left(\frac{B_{t-1}^g}{Y_{t-1}} - \frac{B_{ss}^g}{Y_{ss}} \right) \right)$$

The tax scale parameter τ_t^y is adjusted so that the average labor tax rate is equal to the one determined by the aforementioned tax rule.

The nominal interest rate i_t follows the Taylor rule with the smoothing term:

$$i_t = \rho_{mp} i_{t-1} + (1 - \rho_{mp}) (r_{ss} + \phi \pi_t) + v_t^i$$

where v_t^i is the exogenous component in the nominal interest rate. The nominal and real interest rates are linked via the Fisher equation.

$$1 + r_t = \frac{1 + i_t}{1 + \mathbb{E}_t \pi_{t+1}}$$

A.4 Shocks and Market Clearing Conditions

The exogenous variables in the model are TFP Z_t, government spending G_t, price markup v_t^p, wage markup v_t^w, monetary policy surprise v_t^i. Assume that they follow AR(1) processes.

$$\begin{aligned}
\log Z_{t+1} &= (1 - \rho_Z) \log Z_{ss} + \rho_Z \log Z_t + \sigma_Z \varepsilon_t^Z \\
G_t &= (1 - \rho_G) G_{ss} + \rho_G G_{t-1} + \sigma_G \varepsilon_t^G \\
v_t^p &= \rho_p v_{t-1}^p + \sigma_p \varepsilon_t^p \\
v_t^w &= \rho_w v_{t-1}^w + \sigma_w \varepsilon_t^w \\
v_t^i &= \rho_i v_{t-1}^i + \sigma_i \varepsilon_t^i
\end{aligned}$$

There are six standard aggregate shocks: TFP ε_t^Z, government spending ε_t^G, price markup ε_t^p, wage markup ε_t^w, monetary policy ε_t^i, and transfer ε_t^{tr}.

Let $a_t(adj, a, b, e)$, $b_t(adj, a, b, e)$, and $c_t(adj, a, b, e)$ be the policy functions. The asset market clearing condition is

$$\int a_t(adj, a, b, e) d\mu_t + \int b_t(adj, a, b, e) d\mu_t = B_t + p_t$$

Given that labor, capital, and asset markets clear, the final good market also has to clear by the Walras's law.

$$Y_t = \int c_t(adj, a, b, e) d\mu_t + I_t + G_t + \frac{\eta_p}{2\kappa_p} \left(\log (1 + \pi_t)\right)^2 Y_t + \omega \int b_t(adj, a, b, e) d\mu_t$$

A.5 Calibration

Tables 5 and 6 list the calibrated parameters. The parameter values are fairly standard overall. Discount factor is calibrated so that asset market clears at the steady state. Aggregate

Parameter	Definition	Value	Detail
γ	inv elasticity of intertemp substitution	1.0	Standard
ν	inv Frisch elasticity	1.0	Standard
φ	labor disutility	0.5611	$N = 1$
β	discount factor	0.9865	Asset mkt clearing
p	prob for adjusting illiquid asset	0.062	Bayer et al. (2024)
ρ_e	persistency of $\log e$	0.966	Standard
σ_e	std of shock to $\log e$	0.92	Standard
α^y	incidence	0.078	Iao and Selvakumar (2024)
ω	liquidity premium	0.01	4% annually
Z_{ss}	s.s. TFP	0.4843	$Y = 1$
K_{ss}	s.s. capital	9.0	$K/Y = 9.0$
α	capital share	0.33	Standard
δ	depreciation	0.02	Standard
η_p	elasticity of substitution in goods	7.0	s.s. price markup = 1.17
η_w	elasticity of substitution in labor	7.0	s.s. wage markup = 1.17
κ_p	slope of price Phillips curve	0.121	Iao and Selvakumar (2024)
κ_w	slope of wage Phillips curve	0.165	Iao and Selvakumar (2024)
χ	investment adjustment cost	9.639	Auclert et al. (2020)
B^g_{ss}	s.s. government debt	2.8	$B^g/Y = 2.8$
G_{ss}	s.s. government spending	0.2	$G/Y = 0.2$
τ_D	dividend tax rate	0.2	US tax system
ξ	tax progressivity	0.181	Heathcote et al. (2017)
ϕ_b	sensitivity of tax to debt	0.05	Standard
ϕ_w	sensitivity of tax to labor income	0.0	—
ρ_{mp}	interest rate smoothing	0.875	Standard
ρ_τ	tax rate smoothing	0.9	Standard
ϕ_π	Taylor rule coeff to inflation rate	1.5	Standard

Table 5: Calibration Part 1

Parameter	Definition	Value
ρ_Z	persistency of TFP shock	0.373
σ_Z	std of TFP shock	0.00509
ρ_i	persistency of monetary policy shock	0.373
σ_i	std of monetary policy shock	0.000706
ρ_G	persistency of government spending shock	0.429
σ_G	std of government spending shock	0.00259
ρ_p	persistency of price markup shock	0.205
σ_p	std of price markup shock	0.00201
ρ_w	persistency of wage markup shock	0.197
σ_w	std of wage markup shock	0.00201

Table 6: Calibration Part 2

labor supply and output at the steady state are normalized to be one. One period corresponds to a quarter, and thus the steady state capital and government debt correspond to 225% and 70% per annual output. The incidence parameter α^y being less than one implies that poorer households are more sensitive to changes in aggregate labor income. The shock processes listed in Table 6 are from Iao and Selvakumar (2024) who estimate a medium-scale HANK model with both macro and micro data.

A.6 Smoothing Consumption Distribution

The smoothing procedure with the I-spline proceeds as follows. The internal knot points are chosen evenly based on the percentile of the steady state consumption distribution. For example, if we need 4 internal knot points, they are chosen to be 20, 40, 60, and 80 percentiles of the steady state consumption distribution. Those knot points give the set of I-spline basis functions, which are used to approximate the cumulative distribution of consumption. The coefficients of basis functions are restricted to be non-negative, and thus we apply the non-negative least squares to find the best-fitted coefficients. We firstly perform this smoothing for the consumption distribution at the steady state. The smoothed steady state consumption density is computed simply by differentiating the smoothed consumption cumulative distribution.

The backward and forward iteration following the SSJ step gives the sequence of cumulative consumption distributions following an aggregate shock. For each horizon, we smooth the distribution by the I-spline and compute the consumption density. The impulse response of density is simply the difference between the density at each horizon and the consumption

density at the steady state. This will be used to simulate the time-series for consumption densities.

In a nutshell, the flowchart for our simulation is as follows.

(1) The SSJ method gives the impulse response of aggregate variables X_t with respect to each shock. We write them as $(dX_0^j, dX_1^j, \cdots, dX_T^j)$ where j is the index for shock.

(2) For each j, using $(X_{ss} + dX_0^j, X_{ss} + dX_1^j, \cdots, X_{ss} + dX_T^j)$ as inputs, we run a backward and forward iteration to obtain the sequence of cumulative distributions of consumption at each horizon $0, \cdots, T$.

(3) For each j, the cumulative distributions obtained above are smoothed by the I-spline. Differentiating the smoothed cumulative distributions gives the smoothed densities, and impulse response of the density is given by the difference between the density obtained in this way and the steady state consumption density, written as $(df_0^j, df_1^j, \cdots, d_T^j)$.

(4) We can formulate the moving average process from the impulse response $(dX_t^j, df_t^j)_{j,t}$. The time series of aggregates as well as consumption density is simulated from the moving average representation.

A.7 Impulse Responses

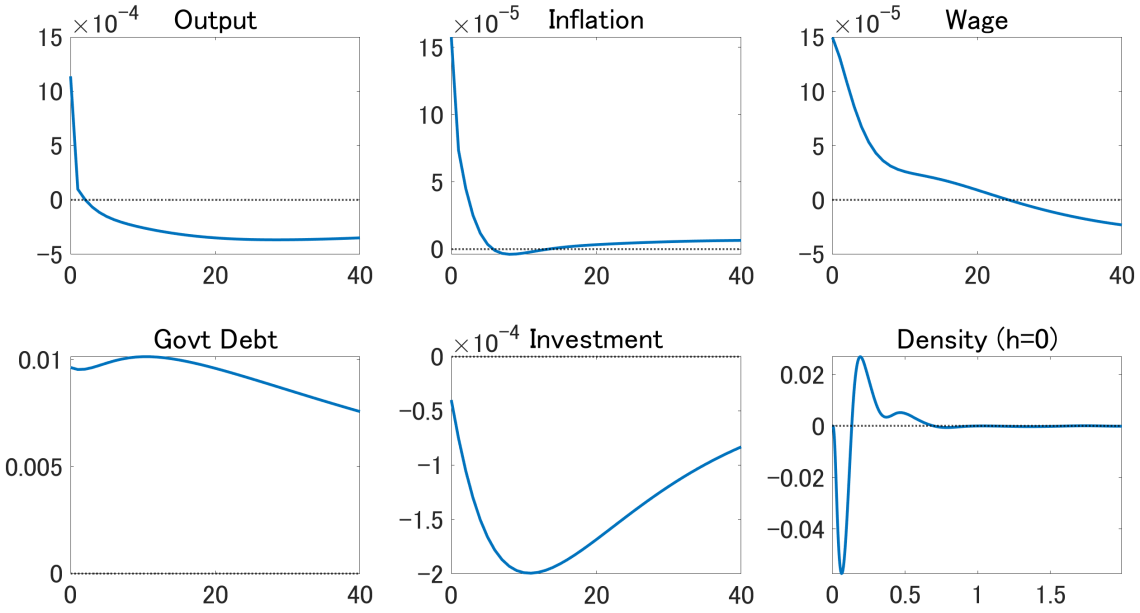

Figure 9: Impulse Response to Transfer Shock

Note: The figure plots the impulse responses with respect to the transfer shock. The right-bottom panel shows the density total response at-impact. Other panels show the impulse responses of aggregate variables (output, inflation rate, wage, government debt, and investment) over time.

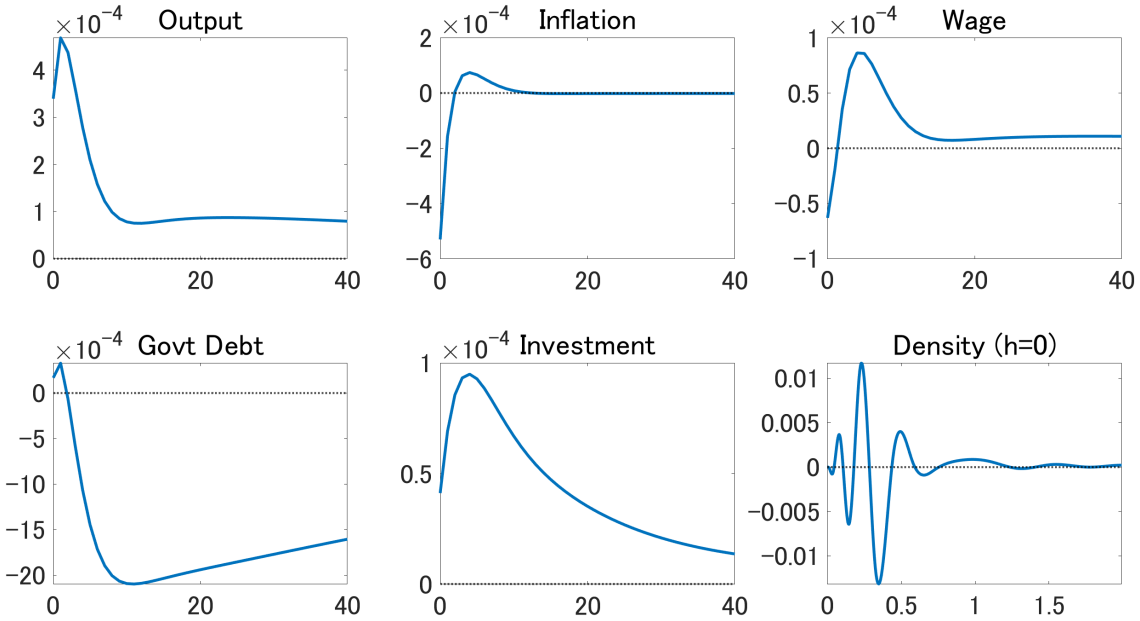

Figure 10: Impulse Response to TFP Shock

Note: The figure plots the impulse responses with respect to the TFP shock. See the note of Figure 14 for detailed description.

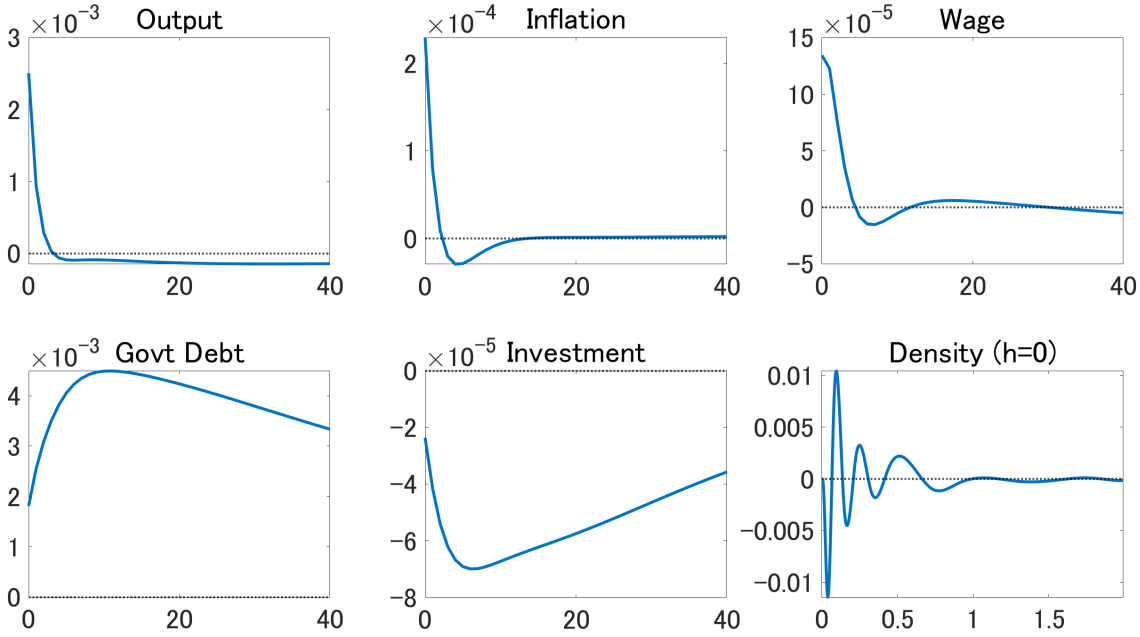

Figure 11: Impulse Response to Government Spending Shock

Note: The figure plots the impulse responses with respect to the government spending shock. See the note of Figure 14 for detailed description.

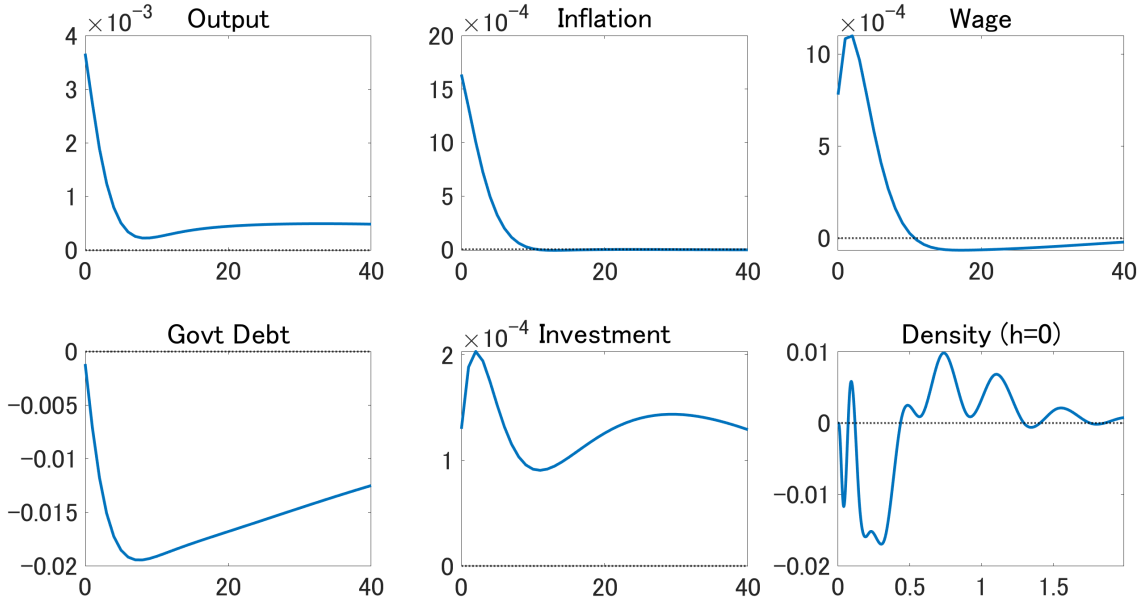

Figure 12: Impulse Response to Monetary Policy Shock

Note: The figure plots the impulse responses with respect to the monetary policy shock. See the note of Figure 14 for detailed description.

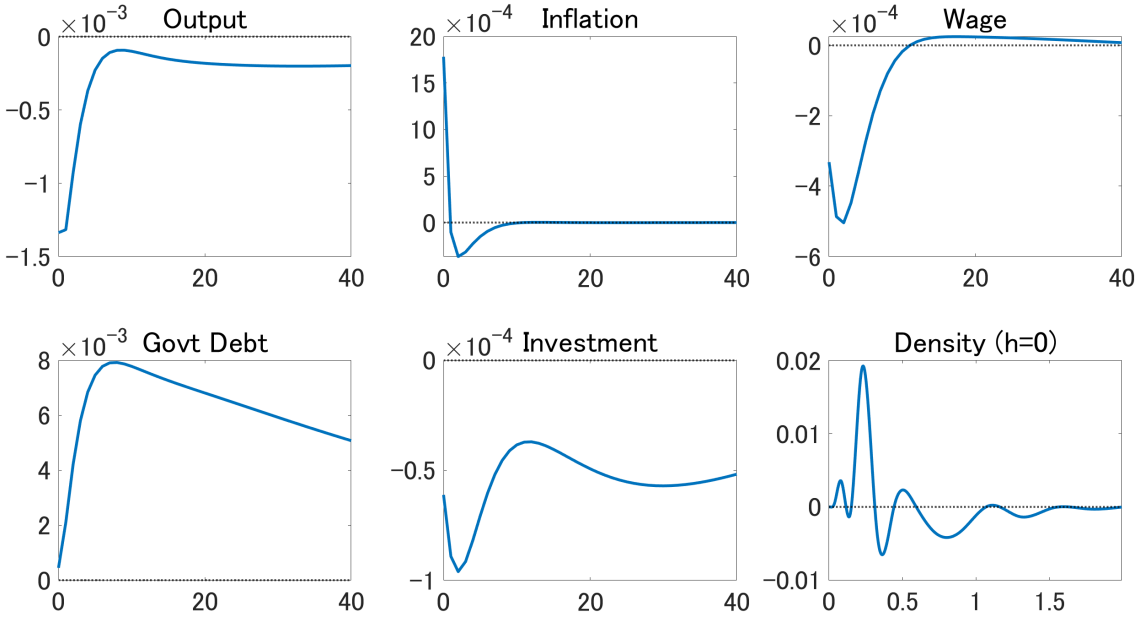

Figure 13: Impulse Response to Price Markup Shock

Note: The figure plots the impulse responses with respect to the price markup shock. See the note of Figure 14 for detailed description.

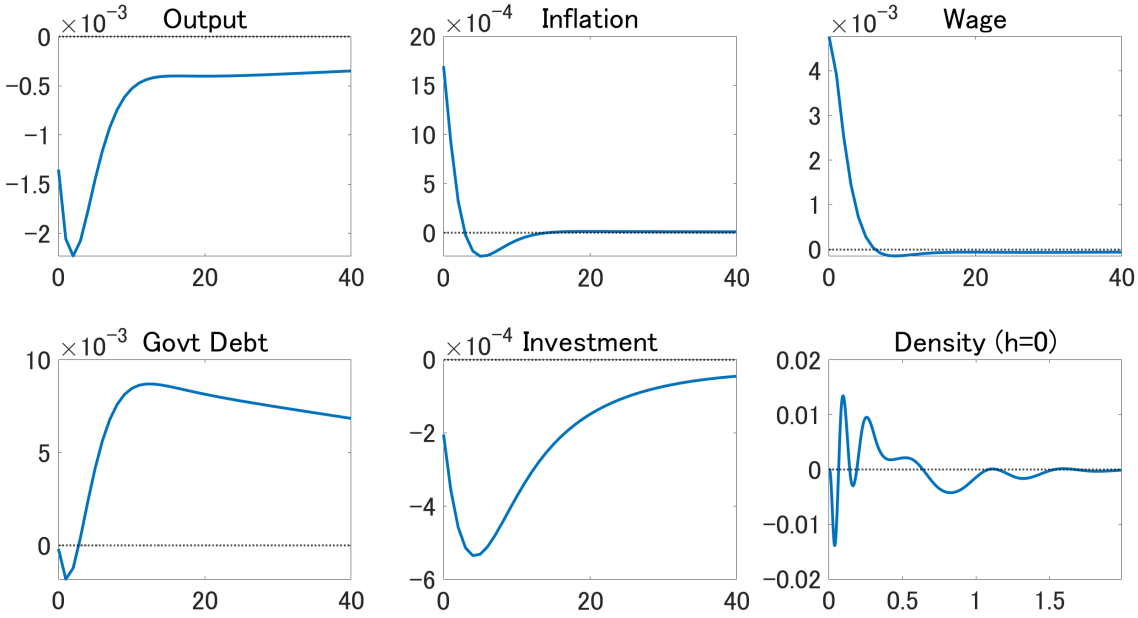

Figure 14: Impulse Response to Wage Markup Shock

Note: The figure plots the impulse responses with respect to the wage markup shock. See the note of Figure 14 for detailed description.

B Bayesian Algorithm

This section introduces a new Bayesian algorithm to estimate structural vector autoregressive models. Although the focus of this paper is entirely on MAR models, the discussion in this section is general enough to accommodate the usual VAR models with only aggregate variables.

Consider an n-variate structural VAR model given by

$$AY_t = C_1 Y_{t-1} + \cdots + C_p Y_{t-p} + B\varepsilon_t$$

where $\varepsilon_t \sim N(0, I_n)$ is independent over time. The history of observations is denoted as $Y = (Y_0, Y_1, \cdots, Y_T)$. An $n \times n$ matrix A shows the contemporaneous relationship between variables, and another $n \times n$ matrix B represents the effects of structural shocks for each equation. Defining $G_l := A^{-1}C_l$ $(l = 1, \cdots, p)$ and $H := A^{-1}B$, it can be written in a canonical form of structural VAR.

$$Y_t = G_1 Y_{t-1} + \cdots + G_p Y_{t-p} + H\varepsilon_t$$

The reduced-form error is represented by $u_t = H\varepsilon_t \sim N(0, \Sigma)$ where $\Sigma = HH'$. As is well known, structural parameters H (or (A, B)) are not identified without additional assumptions because, for any orthogonal matrix Q, we can write $u_t = H\varepsilon_t = (HQ)(Q'\varepsilon_t)$. That is, at-impact structural response of the form HQ for an orthogonal matrix Q is observationally equivalent to H. We (partially) identify the structural parameters by placing prior on Q so that at least one of the elements in ε_t can be economically interpretable.

This framework is general enough to accommodate most structural VAR models. They include the "B-type" where A is set to identity while every element in B is estimated (e.g., Uhlig, 2005), and the "A-type" where the diagonal (off-diagonal) elements of A (B) are set to one (zero) and we estimate off-diagonal elements of A as well as diagonal elements of B (e.g., Baumeister and Hamilton, 2015). Most notably, we can consider more general "AB-type", which is the type of restriction belonging to neither A-type nor B-type.[23] It is clear that our

[23]For example, Blanchard and Perotti (2002) study the effect of government spending and tax to output. The included variables are $Y_t = [\tau_t, g_t, y_t]$ where τ_t is tax rate, g_t is government spending, and y_t is output. The structural parameters are defined as

$$A = \begin{bmatrix} 1 & 0 & a_{13} \\ 0 & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & b_{12} & 0 \\ b_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

structural MAR model also exhibits the AB-type parametrization. Before we describe the algorithm, we discuss how the prior distribution can be formulated.

B.1 Prior

Assume that (A, B) are written as functions of Σ and Q, so that $A = A(\Sigma, Q)$ and $B = B(\Sigma, Q)$. We are interested in the posterior distribution of (G, Σ, Q). We represent the prior for the parameters as

$$p(G, \Sigma, Q) = p(Q \mid G, \Sigma)p(G, \Sigma)$$

The second term gives the prior for the reduced-form parameters (G, Σ). We can use the prior standard in the literature, such as normal-inverse-Wishart distribution. The first term gives the prior for rotation Q given the reduced-form parameters (G, Σ). We represent this prior in terms of (A, B).

$$p(Q \mid G, \Sigma) \propto 1\{Q \in \mathcal{O}(n)\} p\left(A(\Sigma, Q), B(\Sigma, Q) \mid G, \Sigma\right)$$

We specify the prior as the probability distribution of A and B, possibly conditional on (G, Σ). This formulation is consistent with how we build up the prior for our MAR model.

Importantly, $p(Q \mid G, \Sigma)$ can be defined in other ways so that a researcher can reflect the prior information for other objects of interest. For example, if the researcher has prior knowledge on how the impulse response up to horizon h looks like, we can formulate $p(Q \mid G, \Sigma)$ as

$$p(Q \mid G, \Sigma) \propto 1\{Q \in \mathcal{O}(n)\} p\left(IRF_0(Q, \Sigma), IRF_1(Q, G, \Sigma), \cdots, IRF_h(Q, G, \Sigma) \mid G, \Sigma\right)$$

where $IRF_k(\cdot)$ is the impulse response function at horizon $k = 0, \cdots, h$. Note that G is not used as an input for $IRF_0(\cdot)$ because it is a function of only Q and Σ. Also, the forecast error variance decomposition is a function of (Q, G, Σ), which can be reflected as prior if one wishes to.

---

and place the restrictions (i) $a_{23} = 0$ (government spending is not affected by output contemporaneously), (ii) $a_{13} = -2.08$ (based on extraneous evidence on tax elasticity to output), and (iii) either $b_{12} = 0$ or $b_{21} = 0$ (tax policy occurs ahead of government spending or the opposite). Imposing these three restrictions, the VAR is exactly identified. They compute the remaining structural parameters under these assumptions. This structure is of AB-type because A is not identity and we estimate off-diagonal elements of B.

The prior implies the posterior of the form

$$p(G, \Sigma, Q \mid Y) = p(Q \mid G, \Sigma, Y)p(G, \Sigma \mid Y)$$

Note that, the conditional posterior of Q in the first term is proportional to its own prior.

$$p(Q \mid G, \Sigma, Y) \propto \underbrace{p(Y \mid G, \Sigma, Q)}_{=p(Y|G,\Sigma)} p(Q \mid G, \Sigma) \propto p(Q \mid G, \Sigma)$$

where the first transformation is due to the Bayes formula, and the second transformation follows because the likelihood does not depend on Q once we condition on (G, Σ). This implies that Q is not updated by the data given reduced-form parameters, and thus prior plays the sole role for structural identification. Also, given that $p(G, \Sigma)$ belongs to the well known families of distribution, such as the normal-inverse-Wishart, it is quite easy to make draws from $p(G, \Sigma \mid Y)$ since analytical expression for the posterior is available. These observations validate the Bayesian algorithm below.

B.2 Algorithm

We draw parameters from the posterior distribution by the following algorithm. This combines the widely known algorithm for reduced-form VAR (step 1) with the Metropolis Hastings sampler for structural parameters (step 2).

Algorithm 1 (Posterior Sampler to draw $(G_i, \Sigma_i, Q_i)_{i=1,\cdots,I}$)**.**

(1) Draw the reduced-form parameters $(G_i, \Sigma_i)_{i=1,\cdots,I}$ from the posterior $p(G, \Sigma \mid Y)$. If the prior for (G, Σ) belongs to the well-known families of distributions (such as Normal-inverse-Wishart), this step can be done by the existing algorithm for the reduced-form VAR. See, for example, Koop and Korobilis (2010) and Kilian and Lütkepohl (2017).

(2) Run the following steps for $i = 1, \cdots, I$ to draw an orthogonal matrix Q from

$$p(Q \mid G_i, \Sigma_i, Y) \propto p(Q \mid G_i, \Sigma_i) \propto 1\{Q \in \mathcal{O}(n)\} p\left(A(\Sigma_i, Q), B(\Sigma_i, Q) \mid G_i, \Sigma_i\right).$$

(i) Choose an initial value Q_0.

(ii) Iterate the following steps for $j = 1, \cdots, J+1$ times. Let $Q_i := Q_{J+1}$ as a draw from the desired distribution.

(a) Make a proposal Q_p. Let

$$S = \exp\left(c(X - X')\right)$$

where X is a $n \times n$ random matrix following standard matrix normal distribution, c is a scalar tuning parameter which should be adjusted to get the desired MH acceptance rate, and $\exp(\cdot)$ is the operator for matrix exponential. Define Q_p be

$$Q_p = \begin{cases} Q_{j-1}S & \text{with probability } \alpha \\ Q_{j-1}RS & \text{with probability } 1-\alpha \end{cases} \tag{21}$$

where $\alpha \in (0,1)$ is a tuning parameter and $R = \text{diag}(-1, 1, \cdots, 1)$ is $n \times n$.

(b) Let

$$q = \frac{p\left(A\left(\Sigma_i, Q_p\right), B(\Sigma_i, Q_p) \mid G_i, \Sigma_i\right)}{p(A\left(\Sigma_i, Q_{j-1}\right), B(\Sigma_i, Q_{j-1}) \mid G_i, \Sigma_i)}$$

With probability $\min\{1, q\}$, we accept the candidate: $Q_j = Q_p$. Otherwise, we reject the candidate: $Q_j = Q_{j-1}$.

Using the proposal (21) is the important departure from the literature where the proposal distribution is typically the uniform distribution on the Haar measure. The proposal (21) has some desirable properties. First, Q_p is orthogonal given Q_0 being orthogonal.

Proposition 1. Let Q_0 be a n-dimensional orthogonal matrix and X be a $n \times n$ matrix. For any scalar c, $Q_p := Q_0 S = Q_0 \exp(c(X - X'))$ is orthogonal. Also, for a $n \times n$ matrix $R = \text{diag}(-1, 1, \cdots, 1)$, $Q_p := Q_0 RS$ is orthogonal.

Proof. Note that $c(X - X')$ is skewed symmetric: $(c(X - X'))' = -c(X - X')$. This yields $SS' = \exp(c(X - X') + (c(X - X'))') = \exp(O) = I$. Also, $RSS'R' = I$. □

Thus, this proposal scheme generates a new matrix Q_p by giving a perturbation to the original orthogonal matrix Q_0 while ensuring Q_p to be orthogonal. The one close to Q_0 are more likely to be picked up as the candidate because S is concentrated at the identity matrix for small c. The rejection rate of the Metropolis-Hastings step is controlled by a scale parameter c (around 25–40% as a rule of thumb).[24] The lower c implies larger weights to the candidates close to Q_0. This feature is important: When we make a draw from the uniform

[24]In the practical estimation, we make an online adaptation to c. For j-th iteration, we use c_j defined recursively as $\log(c_j) = \log(c_{j-1}) + \gamma_j(a_{j-1} - 0.25)$, where $\gamma_j = \frac{1}{j^\rho}$ with $\rho \in (0.5, 1]$ and a_{j-1} takes one if we accept the candidate at step $j-1$ and zero otherwise.

distribution over Haar measure, the algorithm investigates the region where $p(A(\cdot), B(\cdot) \mid \cdot)$ is low and the candidate is quite unlikely to be accepted, which makes the sampler inefficient. We improve the efficiency by making matrices located in the neighborhood of the original matrix drawn more likely than others.

Second, it is a symmetric proposal.

Proposition 2. Let Q_0 be a n-dimensional orthogonal matrix and X be a $n \times n$ random matrix following the matrix standard normal. For any scalar c, the proposal (21) is symmetric.

Proof. Let $K := X - X'$. Then, diagonal elements of K is zero and off-diagonal elements follow i.i.d. $N(0, 2)$. Thus, K has the same distribution as $-K$. The skewed symmetry of cK implies $S^{-1} = \exp(-cK)$. Since $K =^d -K$, we have $S =^d S^{-1}$.

The proposal kernel relative to Haar measure is written as

$$q(Q_p|Q_0) = \alpha q_1(Q_p|Q_0) + (1-\alpha) q_2(Q_p|Q_0)$$

where the first (second) term reflects the proposal scheme in the first (second) line of (21). We have

$$q_1(Q_p|Q_0) = p_S(Q_0^{-1}Q_p) = p_S(Q_p^{-1}Q_0) = q_1(Q_0|Q_p)$$

which follows from $p_S(s) = p_S(s^{-1})$ as $S =^d S^{-1}$. Hence we have $q_1(Q_p|Q_0) = q_1(Q_0|Q_p)$. We also have

$$\begin{aligned} q_2(Q_p|Q_0) &= p_S(R^{-1}Q_0^{-1}Q_p) = p_S(Q_p^{-1}Q_0 R) = p_S(Q_p^{-1}Q_0 R^{-1}) \\ &= p_S(R^{-1}Q_p^{-1}Q_0) = q_2(Q_0|Q_p) \end{aligned}$$

which follows from Lemma 1. Then we have $q(Q_p|Q_0) = q(Q_0|Q_p)$. □

This leads our algorithm to the random walk Metropolis Hastings, meaning that we do not have to evaluate the ratio of $q(Q_p|Q_0)$ and $q(Q_0|Q_p)$. Indeed, those densities are difficult to evaluate since the density of $S = \exp(c(X - X'))$ is non-standard. Symmetry of the proposal helps us to circumvent this issue.

Third, the proposal is able to visit entire space of orthogonal matrices. Suppose that $\alpha = 1$ and the rotation R is not applied at any time. Then, the sign of determinants of proposal Q_p does not change. If we start from Q_0 with $\det Q_0 = 1$ ($\det Q_0 = -1$), the algorithm searches over the space of orthogonal matrices with positive (negative) determinant, but those with negative (positive) determinant are never investigated. The purpose of mixture α is to let

the algorithm flip the sign of determinant with a certain probability $1-\alpha$ and search over the entire space of orthogonal matrices.

B.3 Comparisons to Other Algorithms

Relative to other algorithms in the literature, this algorithm has several advantages. First, this framework accommodates various form of prior for structural parameters. The estimation algorithm in Baumeister and Hamilton (2015) focuses on the A-type models, where B is assumed to be diagonal and they impose prior on A as well as diagonal elements on B. On the other hand, Bruns and Piffer (2023) restricts A to be identity and place prior on the at-impact response B, i.e., they are interested in the B-type models.[25] Our algorithm allows the models where structural parameters appear both in A and B like our MAR model. Even more generally, identification restrictions can be imposed on other structural parameters of interest, such as dynamic impulse responses as well as forecast error variance decomposition. This generality distinguishes the proposed algorithm than others in the literature.[26]

Second, this algorithm works well when we introduce informative prior on Q. One of the standard ways to make a draw of Q conditional on (G,Σ) is based on the importance sampler: (i) Generating many Q's from the proposal (typically the uniform distribution with respect to the Haar measure), (ii) assigning weights to them based on the value of the density, and (iii) picking one of them based on the weights (e.g., Bruns and Piffer 2023; Arias et al. 2018). However, if the informative prior is involved, it is typical that the uniform proposal does not perform well. That is, the proposal makes draws from the area with low density $p(Q \mid G, \Sigma)$ as likely as other areas, leading the effective sample size being quite small compared to the number of drawn Q's. The proposed Metropolis-Hastings type algorithm sequentially updates the proposal distribution by referring to the previous draw, and thus we can draw Q from the area with high density more effectively. This feature is particularly important in our MAR framework because it restricts (B_{fz}) by making the prior standard deviation of it relatively small.

[25] Bruns and Piffer (2023) mentions that the choice of types "depends on whether the identifying restrictions introduced by the researcher are more naturally expressed on the contemporaneous relation among variables or on the contemporaneous effects of the shocks" (p.1224). Their methodology might be applicable to the other types with some extensions. That being said, the discussion after that point is devoted sorely to the B-type specification.

[26] The idea of imposing prior to dynamics of impulse responses is usually discussed in moving average models (Barnichon and Matthes 2018; Plagborg-Møller 2019) or Bayesian local projections (Ferreira et al. 2025). The novelty here is to show that this is indeed possible in the VAR setting as well.

B.4 Auxiliary Lemma

Lemma 1. Let $S = \exp(c(X - X'))$ where $c > 0$ and X is a $n \times n$ random matrix following matrix standard normal. For any $n \times n$ orthogonal matrix P and $n \times n$ orthogonal matrix s such that $\det s = 1$, we have

$$p_S(s) = p_S(PsP')$$

where p_S is the density for S with respect to Haar measure.

Proof. Let $K = X - X'$. We can write

$$\mathrm{vec}(PXP') = (P \otimes P)\mathrm{vec}(X) =^d \mathrm{vec}(X)$$

as $\mathrm{vec}(X) \sim N(0, I_{n^2})$ and $P \otimes P$ is a $n^2 \times n^2$ orthogonal matrix. It follows that $PXP' =^d X$ and hence $PKP' =^d K$. Thus we obtain

$$\begin{aligned} PSP' = P\exp(cK)P' &= P\left(\sum_j \frac{1}{j!}(cK)^j\right)P' \\ &= \sum_j \frac{1}{j!}(PcKP')^j = \exp(cPKP') =^d \exp(cK) = S \end{aligned}$$

where the second equality follows from the power series expansion $\exp(X) = \sum_j \frac{1}{j!}X^j$, and the third equality is due to $(PcKP')^j = (PcKP')(PcKP')\cdots(PcKP') = P(cK)^jP'$ as P is orthogonal. This is what we desire. □

C An Overview of Mixed Autoregression

We provide a high-level overview of the functional principal component approach to estimate mixed autoregressive (MAR) models. Interested readers are encouraged to refer to Y. Chang et al. (2024b) and Y. Chang et al. (2025) for complete discussion.

C.1 Notation

Let $\mathcal{H}$ be a separable Hilbert space of square integrable functions equipped with inner product $\langle f, g\rangle = \int f(r)g(r)dr$ $(f, g \in \mathcal{H})$. Let $L(\mathcal{H})$ be a space of linear operators on $\mathcal{H}$. The tensor product in $\mathcal{H}$, $f \otimes g$, is defined as the operator satisfying $(f \otimes g)h = \langle h, g\rangle f$ where $h \in \mathcal{H}$. The tensor product is analogous to the outer product in the finite dimensional Euclidean space.

The space $\mathbb{R}^k \oplus \mathcal{H}$ is also Hilbert space with inner product

$$\langle (x, f), (y, g)\rangle = x'y + \langle f, g\rangle \tag{22}$$

where $x, y \in \mathbb{R}^k$ and $f, g \in \mathcal{H}$. The tensor product on $\mathbb{R}^k \oplus \mathcal{H}$ is defined as an operator satisfying

$$((x, f) \otimes (y, g))(w, h) = \langle (w, h) \otimes (y, g)\rangle (x, f) = (w'y + \langle h, g\rangle)(x, f)$$

for any $w \in \mathbb{R}^k$ and $h \in \mathcal{H}$. When $\mathcal{H} = \mathbb{R}^\ell$, we have $f \otimes g = fg'$ and $x \otimes y = xy'$, and the tensor product defined above becomes identical to the sum of outer products of two sets of vectors.

C.2 Mixed Autoregression

Consider a mixed autoregressive model with first-order lag: a MAR(1) model.

$$\underbrace{\begin{bmatrix} X_t \\ f_t \end{bmatrix}}_{Y_t} = G \underbrace{\begin{bmatrix} X_{t-1} \\ f_{t-1} \end{bmatrix}}_{Y_{t-1}} + H\varepsilon_t \tag{23}$$

where $X_t \in \mathbb{R}^k$, $f_t \in \mathcal{H}$, and $G, H \in L(\mathbb{R}^k \oplus \mathcal{H})$ are an autoregressive operator and an operator for at-impact impulse response, respectively. The shock ε_t has mean $\mathbb{E}(\varepsilon_t) = 0$ and variance-covariance operator $\mathbb{E}(\varepsilon_s \otimes \varepsilon_t) = \mathbf{1}\{s = t\}I$. The reduced form variance is defined

as $\Sigma = HH'$ where H' is the adjoint of H. The extension to models with general lag order is trivial. The MAR model (23) is infinite-dimensional and thus cannot be estimated per se. We discuss (i) how to derive the finite-dimensional approximation of the MAR and (ii) how we choose the basis necessary for approximation.

C.2.1 Deriving Finite-Dimensional Representation

Suppose that $\mathbb{R}^k \oplus \mathcal{H}$ is spanned by an orthonormal basis $(v_i)_{i \geqslant 1}$. Then, each element of $\mathbb{R}^k \oplus \mathcal{H}$ can be expressed as

$$Y = \sum_{i=1}^{\infty} \langle v_i, Y \rangle v_i, \quad Y \in \mathbb{R}^k \oplus \mathcal{H}$$

To reduce the dimensionality, we take a subspace $\mathbb{R}^k \oplus \mathcal{V} \subset \mathbb{R}^k \oplus \mathcal{H}$ where $\mathbb{R}^k \oplus \mathcal{V}$ is spanned by a subset of basis $(v_i)_{i=1}^{k+m}$ where m is a finite integer governing the approximation precision. Then, Y is approximated as

$$Y = \Pi Y + (1 - \Pi)Y \approx \Pi Y := \sum_{i=1}^{k+m} \langle v_i, Y \rangle v_i$$

where Π is a projection on a subspace $\mathbb{R}^k \oplus \mathcal{V}$ of $\mathbb{R}^k \oplus \mathcal{H}$ and $(1 - \Pi)$ is an operator for the projection residual satisfying $(1 - \Pi)Y = Y - \Pi Y$. The idea of approximating functions by focusing on sub-basis consisting of several prominent elements is quite common in functional data analysis, and indeed adopted by many works on functional autoregressive models.[27]

Using the projection, the MAR model is approximated as

$$Y_t = G(\Pi Y_{t-1} + (1 - \Pi)Y_{t-1}) + H\varepsilon_t \approx G\Pi Y_{t-1} + H\varepsilon_t \tag{24}$$

Y. Chang et al. (2024b) show that $G(1 - \Pi)Y_{t-1}$ gets negligible asymptotically if we let $m \to \infty$ as $T \to \infty$ with an appropriate rate. We left-multiply Π to the approximated FAR to have

$$\Pi Y_t \approx (\Pi G \Pi)(\Pi Y_{t-1}) + (\Pi H \Pi)(\Pi \varepsilon_t) \tag{25}$$

[27] One of the important exceptions is M. Chang et al. (2024). To derive finite-dimensional expression of the density for micro data, they firstly span the space of log-densities by sets of cubic polynomials. Then, they convert the approximated log-densities back to the (non-log) densities, and choose the density maximizing the likelihood for micro observations, which is the optimization problem with respect to the coefficients of the cubic polynomials.

which is a finite-dimensional representation of the MAR.

We define a $(k+m)$-dimensional vector $\langle Y \rangle$ as

$$\langle Y \rangle = \pi(Y) := \begin{bmatrix} \langle v_1, Y \rangle \\ \vdots \\ \langle v_m, Y \rangle \end{bmatrix}$$

where π is a mapping from $\mathbb{R}^k \oplus \mathcal{H}$ to $\mathbb{R}^k \oplus \mathbb{R}^m$. Once π is restricted on $\mathbb{R}^k \oplus \mathcal{V}$, it is a one-to-one mapping between $\mathbb{R}^k \oplus \mathcal{V}$ and $\mathbb{R}^k \oplus \mathbb{R}^m$ where the inverse mapping is defined as

$$\pi^{-1}(\langle Y \rangle) = \Pi Y$$

We also define a $(k+m) \times (k+m)$ matrix $\langle A \rangle$ as

$$\langle A \rangle = \pi(A) := \begin{bmatrix} \langle v_1, Av_1 \rangle & \cdots & \langle v_1, Av_m \rangle \\ \vdots & \ddots & \vdots \\ \langle v_m, Av_1 \rangle & \cdots & \langle v_m, Av_m \rangle \end{bmatrix}$$

where, with an abuse of notation, π is a mapping from $L(\mathbb{R}^k \oplus \mathcal{H})$ to $\mathbb{R}^{(k+m)\times(k+m)}$. We can see that π restricted on $L(\mathbb{R}^k \oplus \mathcal{V})$ is one-to-one between $L(\mathbb{R}^k \oplus \mathcal{V})$ and $\mathbb{R}^{(k+m)\times(k+m)}$ where the inverse mapping is defined as

$$\pi^{-1}(\langle A \rangle) = \Pi A \Pi$$

Indeed, it is easy to show that those two π's are isometries in the sense that they preserve norms. That is, for any $Y \in \mathbb{R}^k \oplus \mathcal{V}$,

$$\|\langle Y \rangle\|^2 = \sum_{i=1}^{k+m} \langle v_i, Y \rangle^2 = \|Y\|^2$$

where $\|\cdot\|$ is the Euclidean norm for $\langle Y \rangle$ and the norm implied by (22) for Y respectively, and for any Hilbert-Schmidt operator A defined on $\mathbb{R}^k \oplus \mathcal{V}$,

$$|||\langle A \rangle|||^2 = \mathrm{tr}(\langle A \rangle' \langle A \rangle) = \mathrm{tr}(A'A) = |||A|||^2$$

where $|||\cdot|||$ is the Frobenius norm for $\langle A \rangle$ and the Hilbert-Schmidt norm for A.

Using these π's, we can represent the approximated FAR (25) as

$$(Y_t) \approx (G)(Y_{t-1}) + (H)(u_t) \tag{26}$$

which is a $(k+m)$-dimensional VAR. The isometry plays an important role in relating the estimator from (26) with the parameters on the Hilbert space. Indeed, Y. Chang et al. (2024b) show that the estimators

$$\hat{G} := \pi^{-1}\left(\widehat{(G)}\right), \quad \hat{\Sigma} := \pi^{-1}\left(\widehat{(\Sigma)}\right),$$

where $\left(\widehat{(G)}, \widehat{(\Sigma)}\right)$ are the least-square estimator for $((G), (\Sigma))$ from (26), are consistent for G and Σ under some regularity conditions.

C.3 Choice of Basis

The discussion so far is applicable for any orthonormal basis spanning $\mathbb{R}^k \oplus \mathcal{H}$. In practice, however, the performance of our estimator relies heavily on the choice of basis. A good basis represents the fluctuation of functional observations with as small number of leading basis functions as possible, which improves efficiency of the estimators. Although there are well-known bases used to approximate functions,[28] they typically require at least more than 10 functions for preferable approximation precision. We introduce the data-driven methodology to compute basis: the functional principal component analysis.

We assume that the sample mean of $(Y_t)_{t=1}^T$ is zero: $\frac{1}{T}\sum_{t=1}^T Y_t = 0$. It comes without loss of generality since we can simply use the demeaned observations. We let

$$\Gamma = \frac{1}{T}\sum_{t=1}^{T}(f_t \otimes f_t)$$

be the sample variance-covariance operator for (f_t), and define $(u_i^{FPC}, \lambda_i^{FPC})_{i\geqslant 1}$ be the collection of pairs of eigenfunction and eigenvalue with $\lambda_1^{FPC} \geqslant \lambda_2^{FPC} \cdots \geqslant 0$. We call $(u_i^{FPC})_{i=1}^m$ the functional principal component (FPC) basis.

The FPC basis has certain optimality properties: Take an arbitrary orthonormal basis $(u_i)_{i\geqslant 1}$ of $\mathcal{H}$. Let $\mathcal{V}^{FPC}$ be a subspace spanned by $(u_i^{FPC})_{i=1}^m$, and $\mathcal{V}$ be a subspace spanned

[28] They include orthonormalized polynomials, histogram (i.e., splitting the domain of functions and taking local mean for each sub-domain), Fourier series, and Chebyshev polynomials.

by $(u_i)_{i=1}^m$. We can show that, for any m,

$$\sum_{t=1}^{T} \|\Pi^{FPC} f_t\| \geqslant \sum_{t=1}^{T} \|\Pi f_t\|$$

where Π^{FPC} and Π are projections of $\mathcal{H}$ on $\mathcal{V}^{FPC}$ and $\mathcal{V}$ respectively. This implies that, when the number of basis functions m is fixed, the FPC basis explains the temporal variation of (f_t) better than any other basis. Moreover, we have

$$\sum_{t=1}^{T} \left(\Pi^{FPC} f_t \otimes \left(1 - \Pi^{FPC}\right) f_t\right) = 0$$

This shows that the approximation error $(1 - \Pi^{FPC})f_t$ is orthogonal to $\Pi^{FPC} f_t$ in equation (24), suggesting that our estimation is free from the omitted variable bias problem. These properties validate using the FPC basis as a baseline for our exercises.

The benchmark basis used for MAR, $(v_i^*)_{i=1}^{k+m}$ is set as follows.

$$v_i^* = \begin{cases} (e_i, 0) & i = 1, \cdots, k \\ (0, u_{i-k}^{FPC}) & i = k+1, \cdots, k+m \end{cases}$$

where $e_i \in \mathbb{R}^k$ are a vector of zeros except the i-th element being one. This choice incorporates the aggregate variables X_t itself as the first to k-th elements, and the inner product of f_t and FPCs (i.e., functional principal component loadings) as the $(k+1)$-th to $(k+m)$-th elements in the approximate MAR (25).

C.4 Taking Constraints for Densities into Account

One of the concerns to apply the FPC analysis to densities is that one might fail to enforce the unit-integral and non-negativity constraints. Our analysis takes into account the integral constraint because we use the temporally demeaned functional observations all of which are integrated to zero by construction. On the other hand, the non-negativity constraint is not reflected in the FPC analysis.

Petersen and Müller (2016) propose to convert the densities by the log quantile density (LQD) transformation and apply the FPC analysis to the converted densities.[29] The

[29]This methodology is adopted by F. Huber et al. (2024) in the macroeconomic context to approximate the densities of labor earnings.

converted densities are free from these constraints, and thus applying the FPC analysis to them comes without any technical problems. Also, the LQD transformation is invertible, meaning that we can recover the approximated density just by applying the inverse of LQD transformation.

We nevertheless apply the FPC analysis to the original (non-transformed) densities. The approach by Petersen and Müller (2016) find the FPC to the transformed functions, which means that we find the basis approximating the transformed functions well. Because of the non-linearity of the LQD transformation, it does not necessarily imply the good approximation performance for the original densities which are our main objects of interest.[30] Our approximation approach is built on the isometry of π's for which the linearity of transformation plays a crucial role. It is theoretically very challenging to analyze the consequence of such non-linear transformation to approximation quality.

[30]The methodology by M. Chang et al. (2024) faces the similar problem. As discussed in Footnote [27], their basis functions span the space of log-densities and apply the nonlinear transformation to convert them back to the non-log densities.

D Proofs

Proposition 3. If B_{XX} is invertible and both $B_{zX}B_{XX}^{-1}$ and $B_{fX}H_{XX}^{-1}$ are bounded, there is a one-to-one mapping between (A, B) and H.

Proof. The proof is constructive. Computing H from (A, B) is straightforward. We describe how to compute (A, B) from H. Let

$$H = \left[\begin{array}{c|cc} H_{XX} & H_{Xz} & H_{Xf} \\ \hline H_{zX} & H_{zz} & H_{zf} \\ H_{fX} & H_{fz} & H_{ff} \end{array}\right] = \left[\begin{array}{c|c} H_{11} & H_{12} \\ \hline H_{21} & H_{22} \end{array}\right]$$

We define the similar partition for A and B as well. We decompose H to get

$$H = \underbrace{\begin{bmatrix} I & 0 \\ A_{21} & I \end{bmatrix}}_{A^{-1}} \underbrace{\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}}_{B}$$

where $B_{11} = H_{11}$, $B_{12} = H_{12}$, $A_{21} = H_{21}B_{11}^{-1}$, and $B_{22} = H_{22} - A_{21}B_{12}$. Given that A_{21} is bounded, we can write

$$A = \begin{bmatrix} I & 0 \\ -A_{21} & I \end{bmatrix}$$

This constructs the pair (A, B) as desired. □

E Supplementary Figures and Tables for Section 4

E.1 Ratio F_h/F_0 Evaluated at Different Points

Table 7: Ratio of F_h and F_0 at Percentiles

h	Output	Inflation	Wage	Govt Debt	Investment	0.9^h
Panel A: 5th Percentile						
1	0.966	-0.165	0.968	0.877	0.898	0.9
4	0.792	-0.057	0.796	0.479	0.615	0.656
20	0.084	0.097	0.043	0.046	0.100	0.122
40	0.025	0.076	-0.011	0.024	0.045	0.015
Panel B: 16th Percentile						
1	0.935	-0.237	0.936	0.789	0.868	0.9
4	0.541	-0.061	0.533	0.420	0.605	0.656
20	0.076	0.114	0.059	0.007	0.116	0.122
40	0.016	0.093	0.000	0.010	0.031	0.015
Panel C: 50th Percentile						
1	0.986	0.096	0.987	0.911	0.865	0.9
4	0.624	0.120	0.611	0.564	0.564	0.656
20	0.055	0.184	0.027	0.003	0.104	0.122
40	0.017	0.141	-0.007	0.043	0.029	0.015
Panel D: 84th Percentile						
1	0.974	0.037	0.975	0.906	0.871	0.9
4	0.644	0.067	0.626	0.603	0.577	0.656
20	0.055	0.157	0.013	0.023	0.105	0.122
40	0.021	0.121	-0.014	0.040	0.042	0.015
Panel E: 95th Percentile						
1	0.884	0.065	0.877	0.942	0.904	0.9
4	0.664	0.064	0.647	0.751	0.670	0.656
20	0.109	0.113	0.073	0.179	0.187	0.122
40	0.028	0.102	-0.003	0.075	0.098	0.015

Note: This table reports the ratio between F_h and F_0 evaluated at 5th, 16th, 50th, 84th, and 95th percentiles of the steady state consumption distribution for horizons $h = 1, 4, 20, 40$. It also shows the power 0.9^h for comparison.

E.2 More Results

E.2.1 Sign Restrictions on Output and Debt Responses

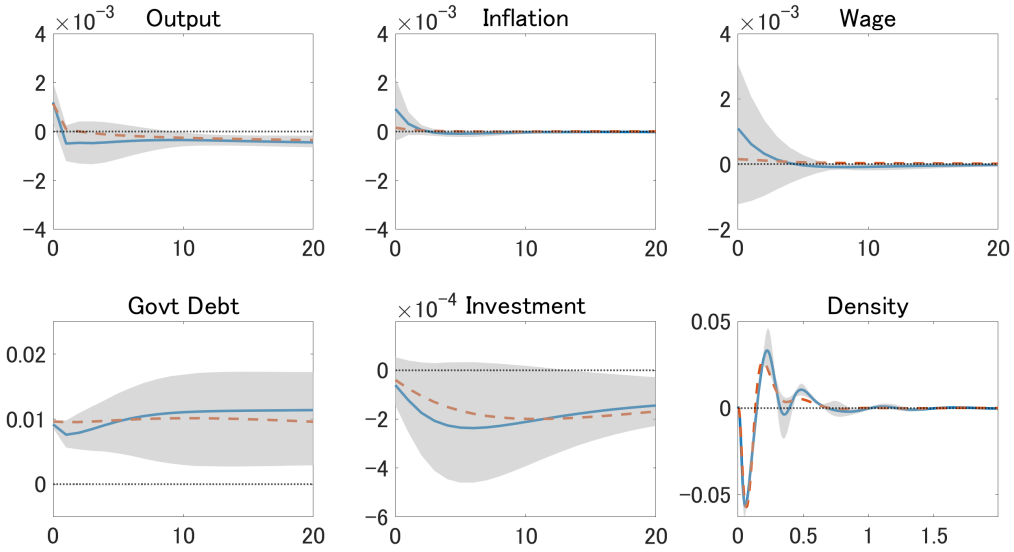

Figure 15: Impulse Responses with Sign Restrictions

Note: See the footnote attached to Figure 4. ρ is set to be 0.9.

E.2.2 Changing ρ

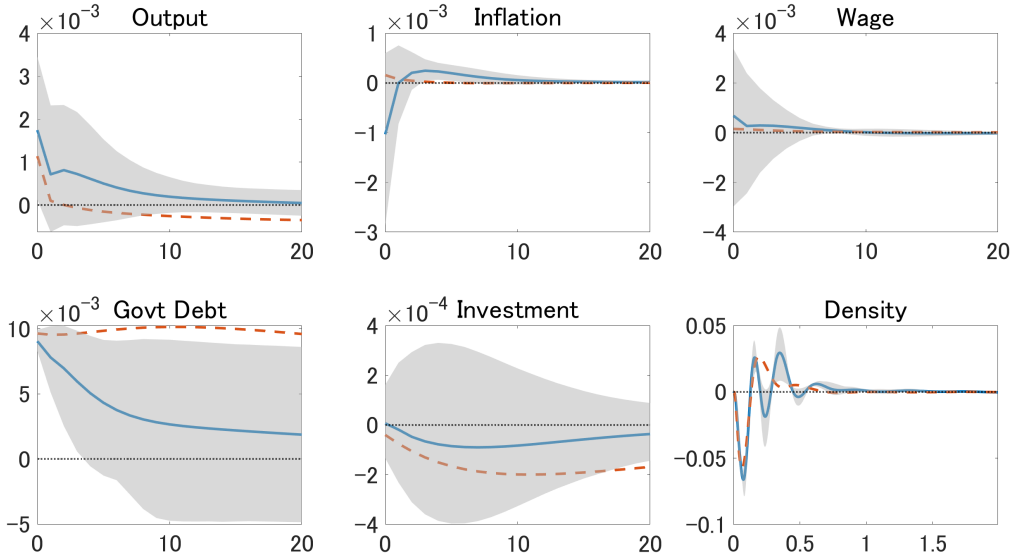

Figure 16: Impulse Responses under $\rho = 0.85$

Note: See the footnote attached to Figure 4. ρ is set to be 0.85.

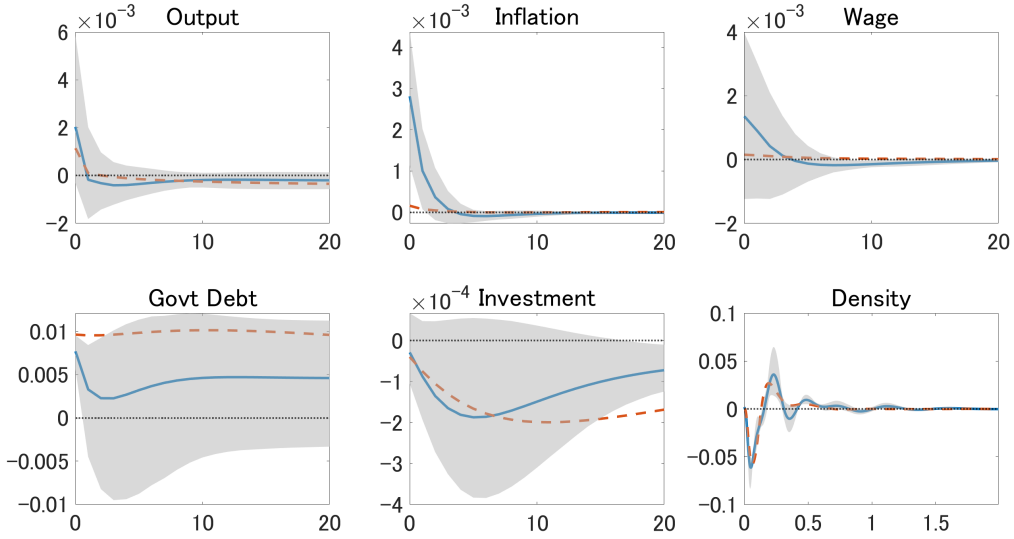

Figure 17: Impulse Responses under $\rho = 0.95$

Note: See the footnote attached to Figure 4. ρ is set to be 0.95.

E.2.3 Joint Bayes Estimator

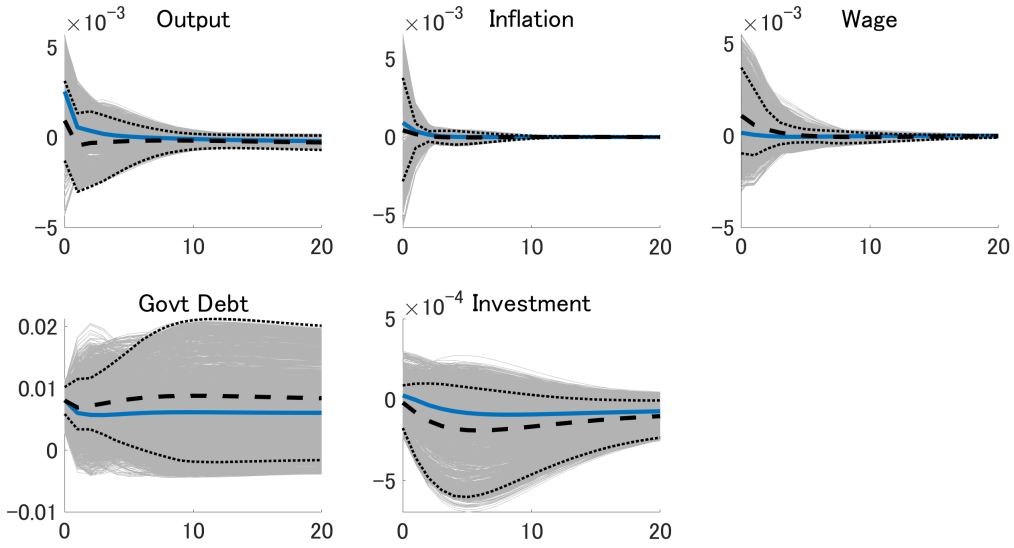

Figure 18: Joint Bayes Estimator

Note: The blue solid line shows the joint Bayes estimator under the absolute additive separable loss function. The gray lines show the 68% credible sets associated with the joint Bayes estimator. The black dashed line shows the point-wise mean along with the point-wise 68% credible intervals with black dotted lines.

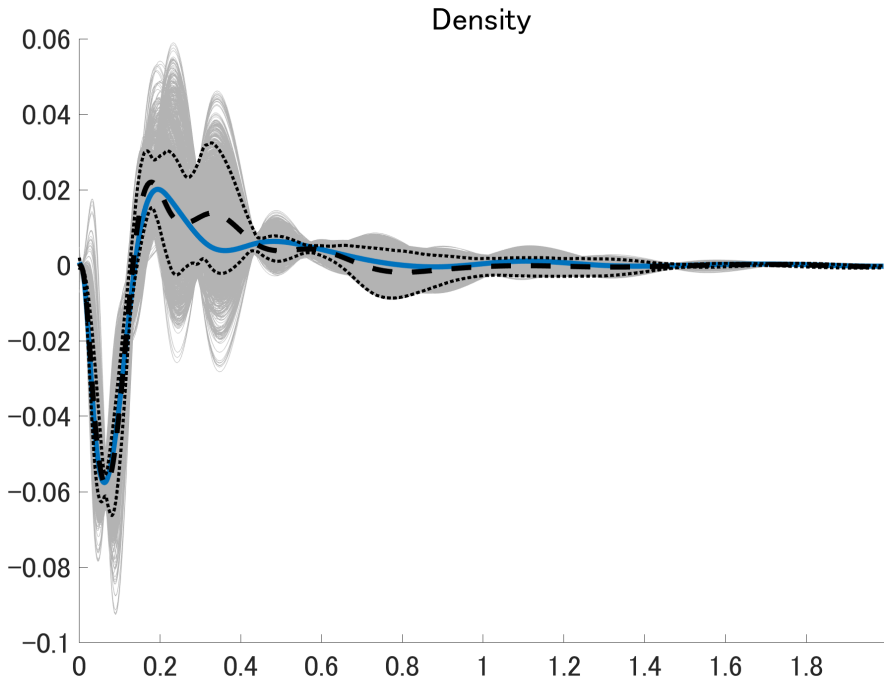

Figure 19: Joint Bayes Estimator for At-Impact Density Response

Note: The blue solid line shows the joint Bayes estimator under the absolute additive separable loss function. The gray lines show the 68% credible sets associated with the joint Bayes estimator. The black dashed line shows the point-wise mean along with the point-wise 68% credible intervals with black dotted lines.

E.2.4 Weaker Prior on $[B_{fz}]$

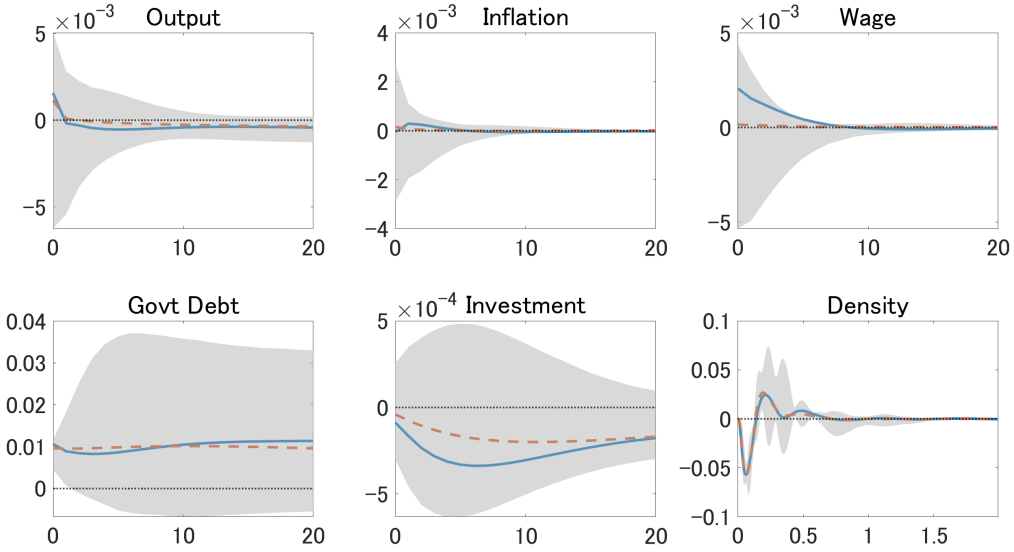

Figure 20: Weaker Prior on $[B_{fz}]$

Note: See the footnote attached to Figure 4. We set $\rho = 0.9$.

F Supplementary Discussion for Section 5

F.1 Data Construction

All the macro variables except the Wu and Xia (2016) shadow rate are constructed from the FRED. Real output corresponds to the FRED mnemonic `GDPC1`. Inflation rate is the log difference of GDP deflator from one year ago (`GDPDEF`). Real Federal net transfer payment is computed as the difference between nominal current transfer payments (`W014RC1Q027SBEA`) and nominal current transfer receipts (`W011RC1Q027SBEA`) deflated by the GDP deflator. Real federal tax revenue is computed as the sum of current tax receipts (`W006RC1Q027SBEA`) and contributions for government social insurance (`W780RC1Q027SBEA`) deflated by the GDP deflator. Real output, net transfer payment, and tax revenue are divided by the CBO estimate of real potential output (`GDPPOT`). The shadow rate is taken from the website of the Federal Reserve Bank of Atlanta.[31]

The measure for micro consumption expenditure is constructed by subtracting personal insurance and pension (sum of CEX variables `perinscq` and `perinspq`) and retirement, pensions, social securities (sum of `retpencq` and `retpenpq`) from total expenditures (sum of `totexpcq` and `totexppq`). It covers food, alcoholic beverages, apparel, housing, transportation, health care, entertainment, personal care, reading, education, tobacco, cash contribution, and miscellaneous expenditures.

The CEX started to report the imputed pre-tax income in 2004Q1 to correct for non-responses as well as responses of zero income. In addition, the raw pre-tax income is not available for 2004 and 2005. As a measure for family income, we use the imputed income from 2004Q1 (`fincbtxm`). Prior to 2004, we impute the pre-tax income by replicating the BLS procedure as closely as possible, following Coibion et al. (2017).

F.2 Prior for Reduced Form Parameters

We omit $(\cdot)$ for exposition. Let $G = [G_1\ G_2\ \cdots\ G_p]$. We set the prior of reduced form parameters (G, Σ) to be normal-inverse-Wishart:

$$p(G, \Sigma) = p(G \mid \Sigma)p(\Sigma)$$

[31] https://www.atlantafed.org/cqer/research/wu-xia-shadow-federal-funds-rate

where

$$vec(G) \mid \Sigma \sim N(g_0, kron(\Sigma, V_g))$$
$$\Sigma \sim IW(\tau, S_0)$$

where $kron(\Sigma, V_g)$ is the Kronecker product between Σ and V_g.[32] The parameters associated with the distributions are chosen following Chan (2019). The conditional prior of $vec(G)$ given Σ is based on the idea of Minnesota prior. The mean is zero except for diagonal elements of G_1, which are one. The variance V_g is diagonal, whose element corresponding to ℓ-th lag of variable j is given as $\frac{\phi^2}{\ell^2 s_{jj}^2}$ where s_{jj}^2 is the (j,j) element in the OLS estimate of Σ. That is, we impose stronger shrinkage for coefficients with respect to distant lags. We set $\phi = 0.2$.

We choose degrees of freedom τ to be 10. The scale parameter S_0 is chosen so that $\mathbb{E}(\Sigma)$ is equal to the OLS estimate of Σ.

[32]We do not use $\otimes$ to denote the Kronecker product intentionally in order to avoid confusion with the tensor product on the Hilbert space.