CAHIER D'ÉTUDES
WORKING PAPER

N° 185

HETEROGENEITY IN MACROECONOMIC MODELS: A REVIEW OF THEORY AND COMPUTATION

JULIEN PASCAL

JUNE 2024

BANQUE CENTRALE DU LUXEMBOURG
EUROSYSTÈME

HETEROGENEITY IN MACROECONOMIC MODELS: A REVIEW OF THEORY AND COMPUTATION

JULIEN PASCAL

Abstract. This paper reviews the literature examining the consequences of heterogeneity in macroeconomic modeling, especially within the context of monetary and fiscal policy transmission. This review reveals that heterogeneity can significantly alter the transmission mechanisms of monetary policy in macroeconomic models and suggests possible advantages from collaboration between fiscal and monetary policies. The paper also provides a critical evaluation of various analytical and numerical methods to solve macroeconomic models with heterogeneity, underscoring the need for a careful methodological choice based on specific circumstances.

Keywords: Literature review, Heterogeneity, Monetary Policy, Fiscal Policy, Macroeconomic Modeling, HANK model.
JEL Codes: E32, E52, E62, D31, C61.

May 2024. Julien Pascal: Banque centrale du Luxembourg, Département Économie et Recherche, 2 boulevard Royal, L-2983 Luxembourg (julien.pascal@bcl.lu). This paper should not be reported as representing the views of the Banque centrale du Luxembourg or the Eurosystem. The views expressed are those of the author and may not be shared by other research staff or policymakers at BCL or the Eurosystem.

Résumé non technique

Ce papier analyse comment les différences entre les acteurs économiques, notamment en termes de revenu ou de patrimoine entre les ménages, peuvent affecter la transmission de la politique monétaire ou fiscale au sein des modèles économiques. Les modèles macroéconomiques néo-keynésiens, souvent utilisés par les banques centrales et les chercheurs académiques pour analyser les cycles économiques, reposent traditionnellement sur l'abstraction d'un ménage représentatif, qui capture le comportement d'un "consommateur moyen". Cette hypothèse de travail simplifie grandement l'analyse des modèles macroéconomiques et permet de définir la réponse "optimale" des politiques monétaires et budgétaires face aux fluctuations économiques.

Ces dernières années, grâce à des avancées théoriques et à une augmentation de la puissance de calcul des ordinateurs, les économistes ont pu étudier des modèles macroéconomiques sans recourir à l'hypothèse d'un ménage représentatif. Ils ont ainsi exploré comment l'hétérogénéité des ménages, notamment en termes de richesse et de revenu, influence l'amplitude des fluctuations économiques. Ce papier présente les principaux résultats établis par cette récente littérature. Notamment, ces modèles à agents hétérogènes prédisent l'existence de ménages avec une plus forte propension à consommer, fait confirmé par des études empiriques sur le comportement des consommateurs. Ces modèles à agents hétérogènes mettent également l'accent sur le rôle des effets indirects des politiques monétaires, définis comme ceux qui se propagent au reste de l'économie via des ajustements sur les marchés de l'emploi, des biens et des actifs financiers (effets d'équilibre général), sans se limiter aux effets directs d'un changement des taux directeurs à court terme. Par exemple, ces effets indirects sont ceux qui se manifestent via un changement sur le revenu des ménages, ou via une relaxation de la contrainte budgétaire du gouvernement. Ce papier synthétise ces différents résultats, tout en offrant une revue des différentes méthodes qui ont été développées pour analyser cette nouvelle classe de modèles économiques. Le choix de la méthode doit être fondé sur une analyse des avantages et des inconvénients associés à chaque approche.

1. Introduction

In recent years, officials and governors of central banks have emphasized the need for a better understanding of how heterogeneity influences aggregate economic outcomes in order to gain insights into fluctuations along the business cycle. This includes former Chair of the US Federal Reserve Janet Yellen, who wondered "whether individual differences within broad groups of actors in the economy can influence aggregate economic outcomes", Haruhiko Kuroda, former Governor of the Bank of Japan, who noted that "central banks should be, and in fact are, open to learning about heterogeneous agent macroeconomics", and Vitor Costancio, former Executive Board member at the European Central Bank, who lamented that "for too long, the distribution of income and wealth was almost ignored by macroeconomics".[1]

This paper contributes to the research agenda examining the impact of heterogeneity on aggregate economic outcomes. It does so by providing a comprehensive literature review of the existing body of work on this topic. For clarity, the literature review contains two parts. First, I review the literature on the mechanisms linking heterogeneity and aggregate outcomes. This literature review starts from a critical review of the Representative Agent (RA) assumption, according to which heterogeneity does not impact aggregate outcomes, which was the predominant paradigm before the Great Financial Crisis. Empirical research and theoretical advances have challenged the Representative Agent (RA) assumption and led to the development of alternative models. Through the use of these more sophisticated models, economists have made two key observations: (i) The transmission mechanism of monetary policy can significantly differ between a Representative Agent (RA) model and a Heterogeneous Agent (HA) model. While the transmission channels in RA models are primarily driven by intertemporal consumption substitution[2], HA models exhibit substantial indirect effects that are absent in RA models. In

[1]See `https://www.bis.org/review/r161017b.pdf`, `https://www.boj.or.jp/en/about/press/koen_2017/data/ko170524a.pdf`, and `https://www.ecb.europa.eu/press/key/date/2017/html/ecb.sp170822.en.html`

[2]When the real interest rate rises, return on savings increases and households have more incentives substitute current consumption with future consumption by saving more today. The elasticity of intertemporal substitution measures the net effect of real interest rates on consumption today, also taking into account the fact that households with positive assets might feel richer because they have more interest income.

standard RA models, a decrease in the real interest rate increases consumption today, and almost all of the effects of monetary policy can be explained by this intertemporal substitution channel. Indirect effects of monetary policy are defined as those causing aggregate demand to change through mechanisms different from the intertemporal substitution channel. For instance, profits may rise, leading firms to produce more and hire more workers, leading to higher labor income. This causes aggregate demand to rise. These types of indirect effects are usually negligible in RA models. (ii) Heterogeneity reinforces the case for collaboration between fiscal and monetary policies, as HA models suggest strong equilibrium effects that challenge the applicability of Ricardian equivalence.[3] When the real interest decreases, this relaxes the budget constraint of the fiscal authority. The government may react by adjusting its total spending, increasing transfers or lowering taxes. Because Ricardian equivalence does not apply in most HA models, even an increase in deficit funded transfers has consequences on aggregate demand, and hence on aggregate supply. First order consequences of changes in aggregate demand may be amplified by the presence of general equilibrium effects. For instance, an increase in output leads to more labor demand, hence more labor income, further stimulating demand for goods, especially from households with high marginal propensity to consume. Other findings relate to estimation and the resolution of some long-standing puzzles.

In a second part, I describe the different methods available to solve macroeconomic models with heterogeneity. Once heterogeneity is allowed to matter at an aggregate level, traditional techniques used to solve Dynamic Stochastic General Equilibrium (DSGE) models are no longer applicable because they are devised to work with low-dimensional models. In DSGE models with heterogeneity, equilibrium prices depend on the distribution of households or firms along one or several dimensions. Since these distributions are infinite-dimensional objects, this represents a challenge for model solution. However, many alternative methodologies have since been developed. This paper critically evaluates analytical and numerical methods for solving macroeconomic models with heterogeneity. Analytical methods rely on simplifying assumptions that allow one to find closed-form solutions or to transform a HA model into a quasi-RA model amenable to

[3]Ricardian equivalence is the theoretical result according to which government spending funded with deficit has no impact on aggregate demand. Forward-looking households anticipate that higher government spending today implies higher taxes in the future. Hence, they save more today, which exactly offsets the current increase in consumption today caused by the government (see Barro (1974)).

traditional techniques. Numerical methods rely on computational approaches to find approximations to RA models, without the need for additional simplifying assumptions. Each approach has its advantages and limitations. For instance, analytical approaches allow for a sharp characterization of equilibria, but such models are often too simple to fit empirical observations. Numerical approaches, such as linearization in the state space or sequence space allow economists to build richer models, but results may be harder to characterize or may involve significant costs in terms of computing time. The key takeaway from this section is that there is no universal approach that suits all situations in this literature. The choice of method should be based on careful consideration of the advantages and disadvantages associated with each approach.

2. Impact of Heterogeneity on Aggregate Economic Outcomes

This section reviews the literature on the role of heterogeneity in macroeconomic models, with a special emphasis on monetary and fiscal policies. It first covers the role that the Representative Agent (RA) assumption played in the micro-foundation of macroeconomics and how it limits the role of heterogeneity in macroeconomic models. By definition, in a RA, heterogeneity does not matter. However, the RA is not immune from critiques and several considerations have led economists to depart from it. Once heterogeneity is introduced in macroeconomic models, the literature finds that (i) the propagation channels of monetary policy are richer, (ii) indirect channels have a leading role in the propagation of monetary policy, (iii) heterogeneity strengthens the case for collaboration between fiscal and monetary policies, (iv) heterogeneity is a key element in solving long-standing theoretical and empirical puzzles.

This literature review is by no means exhaustive. For instance, it does not cover the literature on heterogeneous firms (see for instance Khan and Thomas (2008)), or heterogeneous workers in frictional markets (see for instance Lise and Robin (2017)). This section also does not discuss the technical challenges associated with solving economic models with heterogeneity, which is covered in a subsequent section.

2.1. **The Representative Agent assumption: a stepping stone in the micro-foundation of macroeconomics.** The Representative Agent (RA) assumption is an idealized concept in economics according to which a single representative individual or household is assumed to capture the characteristics of the whole population. Similarly, the behavior of all firms in the economy can be summarized by the behavior of a single representative firm. Representative agents can be traced backed to Marshall (1891), who was the first to use the concept of a representative firm when constructing his Producer Theory, writing "we have to consider the conditions of the representative firm rather than a given individual firm". While nowadays it is common to use a representative household, Marshall viewed this idea with a certain amount of skepticism: "I think the notion of 'representative firm' is capable of extension to labour; and I have had some idea of introducing that into my discussion of standard rates of wages. But I don't feel sure I shall: and I almost think I can say what I want to more simply in another way."[4]

The RA assumption was critical in helping economists to develop models that are resilient to Lucas' critique (Lucas, 1976). Lucas pointed out that the large-scale macroeconometric models used at the time could not be used for policy evaluations, because they were not based on forward-looking rational agents. That is, they ignored the endogenous reaction of agents to changes in policy parameters. The RA assumption allowed for the microfoundation of macroeconomic models by drastically reducing the number of variables and by putting aside concerns that may arise from aggregating individual behaviors. It enabled economists to focus on the interaction of one rational and forward-looking representative household with a rational and profit-maximizing representative firm, and the study the aggregate behavior and outcomes for the economy as a whole in models that are resilient to Lucas' critique. The microfoundations movement is best exemplified by the seminal work of Kydland and Prescott (1982), who initiated the real business cycle (RBC) literature. The RBC literature introduced the idea that macroeconomic fluctuations result from the existence of unobserved shocks, which are then propagated through the economy by the optimal reaction of agents (households and firms) following these shocks. These key ideas of the RBC literature are now embedded in the Dynamic Stochastic General Equilibrium (DSGE) models used by academics and central banks to study macroeconomic fluctuations (see Smets and Wouters (2003) and Smets and Wouters (2007)).

In addition to allowing for the microfoundation of macroeconomic models, the use of a representative agent was also initially broadly accepted based on a result known as "approximate aggregation". This is a numerical observation made by Krusell and Smith

[4]See Hartley (1996) for an historical discussion on the birth of representative agents in economics.

(1998), among the first authors to build a macroeconomic model with rational and forward-looking heterogeneous households, that only average variables are required for agents to make forecasts about the future. In a numerical sense, a model with many different agents behaved as if the model was populated by a single representative agent. Because introducing heterogeneity raised computational complexity by several orders of magnitude with the method developed by Krusell and Smith (1998), the "approximate aggregation" result seemed to indicate that, on a cost-benefit basis, adding heterogeneity in macroeconomic models was simply not worth it. Or to put it more bluntly, the "approximate aggregation" result seemed to indicate that heterogeneity does not really matter in macroeconomic models.

2.2. **Limitations of Representative Agent models.** However, the RA assumption is not immune from critiques. First, it can be criticized on the basis of the strong conditions that are required for it to be true. For a representative household to exist in an economic model, at least one of the following three assumptions must hold: (i) all individuals are identical (e.g. identical in terms of preferences, age, wealth), which makes the aggregation of individuals' behavior trivial, (ii) agents in the economy have preferences that can be represented by specific utility functions that have the Gorman polar form[5] (Gorman, 1953), (iii) households have access to a perfect insurance market, implying that household consumption does not exhibit any history dependence and is only a function of aggregate production[6].

Second, the RA assumption has also been criticized because it induces a loss of useful information for economists and policy makers. A first type of information loss occurs when considering the channels at play. Indeed, while it might be true that on aggregate the economy behaves as if there exists a representative agent, the channels through which shocks or policies affect the economy may greatly differ between RA and Heterogeneous Agents (HA) models. A leading example is given by Kaplan, Moll, and Violante (2018), who show that the transmission mechanism from monetary policy to household consumption in a Heterogeneous Agent New Keynesian (HANK) model is completely different from the transmission mechanism in a Representative Agent New Keynesian (RANK)

[5]Examples of utility functions that have the Gorman polar form are the linear, the Leontief and the Cobb-Douglas utility functions.

[6]See chapter 8 of Ljungqvist and Sargent (2018).

model. In that sense, even if "approximate aggregation" holds for a model, using a model that takes into consideration heterogeneity may still be the right approach because a RANK model may lead to false interpretation regarding the mechanisms leading up to the results.

A second type of information loss comes from neglecting the impact of policies on heterogeneity itself. By construction, RA models focus on the effects of policy on the average household. Hence, they cannot be used to analyze issues related to the distributional consequences of certain policies. Yet, there is growing body of literature demonstrating that monetary policy has a direct impact on heterogeneity and inequality. For instance, regarding conventional monetary policy, Coibion, Gorodnichenko, Kueng, and Silvia (2017) and Furceri, Loungani, and Zdzienicka (2018) document that contractionary monetary policy shocks tend to increase income inequality. Regarding unconventional monetary policy, Montecino and Epstein (2015), Mumtaz and Theophilopoulou (2017) and Saiki and Frost (2014) find that quantitative easing (QE) might have widened income inequality in the US, UK and Japan. However, Lenza and Slacalek (2018) find that QE in the euro area compressed the income distribution and had negligible effects on wealth inequality.

A third type of information loss occurs when considering the estimation of macroeconomic models. Because RA models make only predictions on macro series, one is usually restricted to only using aggregated data to estimate the model's parameters. However, because HA models make predictions about the distributions of agents across economic variables, one may complement macro data with micro data. For instance, Challe, Matheron, Ragot, and Rubio-Ramirez (2017), Bayer, Born, and Luetticke (2022) and Papp and Reiter (2020) use both macro and micro data to estimate heterogeneous agent models. Liu and Plagborg-Møller (2023) develop a general method to estimate heterogeneous agent models using both standard macro time series, as well as micro data. They demonstrate that using micro and macro data leads to more accurate estimation of the model's parameters, and that in some instances some parameters can only be identified if micro data are used to complement more traditional macro time series.

For all these reasons, it is now more common for macroeconomic model builders to avoid using a representative household, or a representative firm. The rest of this literature review is devoted to summarizing the main findings of the literature on the links between heterogeneity, monetary and fiscal policies.

2.3. **Heterogeneity and monetary policy: richer direct propagation channels.** In a standard New Keynesian model with a representative agent (see for instance Clarida, Gali, and Gertler (1999) or Galí (2015)), there are two types of distortions that may in principle justify the intervention of a monetary authority. The first distortion results from the presence of market power in the goods markets, implied by the assumption that firms operate in a monopolistic competition setting.[7] This distortion can be eliminated by a fiscal authority, providing an employment subsidy funded by a lump-sum tax. A second type of distortion comes from price setting rigidity among firms, with only a certain fraction of firms able to update their prices within a given period (Calvo, 1983). The monetary authority can maximize social welfare by setting the inflation rate to its target, which in turn closes the output gap. This "divine coincidence" can be implemented with a simple Taylor rule, according to which the central bank reacts to both deviations of inflation from its target and output from its natural level, the output gap.[8] To rule out multiplicity of equilibria, the coefficients of the Taylor rule must respect the "Taylor principle": a one percentage point increase in the inflation rate must be met with more than a one percentage point increase in the nominal interest rate. A key advantage of the representative agent New Keynesian (RANK) model is that its prescriptions for monetary policy are clear, are now well understood, and can guide policymakers.

However, recent empirical findings suggest that RANK models are at odds with the data, casting doubts on the validity of the clear predictions of these models. As pointed out by Kaplan, Moll, and Violante (2018), the transmission mechanism of monetary policy in a RANK model operates almost entirely through a direct substitution effect: when real rates fall, households save less and borrow more, which increases aggregate consumption. Yet, after controlling for income, analyses of time series data have revealed limited sensitivity of consumption to changes in interest rates (see Campbell and Mankiw (1989), Yogo (2004), Canzoneri, Cumby, and Diba (2007)). RANK models also feature a representative household, who behaves as permanent income consumer who barely reacts to transitory income changes. This feature is also contradicted by the data.

[7]Allowing for market power is necessary to introduce sticky prices in New Keynesian model.

[8]The natural level of output is defined as the level of output that would be reached with flexible prices.

Empirical analyses of marginal propensities to consume (MPC) have demonstrated that consumers react strongly to transitory income changes (Johnson, Parker, and Souleles (2006), Parker, Souleles, Johnson, and McClelland (2013)). Empirical works have also highlighted that there exists a wide range of MPCs among consumers, with differences mainly driven by the level of illiquid assets they hold and the composition of their balance sheets (see Mian, Rao, and Sufi (2013), Misra and Surico (2014)). In a representative agent setting, there is by definition a single aggregate MPC and a single aggregate balance sheet.

These inadequacies have led economists to formulate heterogeneous agents New Keynesian (HANK) models, which generate predictions more in line with the data and that contain propagation mechanisms that go beyond the simple direct effects of the intertemporal substitution of consumption (see Figure 1 for a visual summary of the main mechanisms in RANK and HANK models). For instance, Kaplan, Moll, and Violante (2018) design a HANK model in which households can hold both liquid and illiquid assets. The authors show that the transmission of monetary policy to household consumption mainly operates through a general equilibrium increase in labor demand, which is in sharp contrast with the traditional RANK model, in which almost all the effects of monetary policy can be attributed to intertemporal substitution effects. In their model, the impacts of monetary policy can differ from a standard RANK model. For example, they document that a large but transitory nominal rate cut can be more effective in stimulating aggregate consumption than a small but persistent rate cut. In a RANK model, the two policies have similar impacts on the economy as long as the cumulative rate deviations they imply are equal.

Auclert (2019) identifies three channels through which monetary policy propagates to consumption in an economy with heterogeneous agents. First, monetary policy operates through an earnings heterogeneity channel: a positive monetary shock increases labor and profit earnings and these gains are not uniformly distributed among the population. Second, monetary policy is transmitted to consumption by the Fisher channel: unexpected inflation shocks transfer wealth from nominal creditors to nominal debtors. The third channel is the interest rate exposure channel: when real rates decrease, the price of both assets and liabilities increase. Hence, households with short-maturity assets and long-maturity liabilities will tend to lose, while households with long-maturity assets and short-maturity liabilities will tend to win. Provided that winners from a monetary policy expansion have higher MPCs than losers, which appears to be supported by empirical evidence, Auclert (2019) demonstrates that the presence of heterogeneity among households amplifies the reaction of aggregate consumption to a change in monetary policy.

A series of works investigate the interplay between agent heterogeneity, monetary policy and the income effect that results from changes in mortgage rates. Eichenbaum, Rebelo, and Wong (2022) find that the distribution of mortgage rates among the population has a direct impact on the optimal conduct of monetary policy. A nominal rate cut makes refinancing mortgages more attractive, which can amplify the transmission of monetary policy. This can also alter some trade-offs: fighting recessions with a prolonged period of low interest rates, which lowers potential gains from the refinancing channel in the future, reduces the potency of monetary policy in the period after interest rates are normalized. While Eichenbaum, Rebelo, and Wong (2022) focus on heterogeneity along the life cycle, Berger, Milbradt, Tourre, and Vavra (2021) find similar results in a model that omits life cycle components but contains heterogeneity in terms of labor income risk. The authors also report that prepayment of mortgages accounts for more than 50% of total monetary transmission in their model. Beraja, Fuster, Hurst, and Vavra (2019) study the refinancing channel of monetary policy in a model with heterogeneous agents and heterogeneous regions. They find that the distribution of housing equity across space and across agents plays a crucial role in the economy's response to interest rate declines. Garriga, Kydland, and Šustek (2017) discuss how the effects of monetary policy shocks differ between fixed-rate mortgage (FRM) owners and adjustable-rate mortgage (ARM) owners. Persistently higher inflation gradually benefits homeowners with FRMs, while it is directly detrimental to homeowners under ARMs. The authors also report that the connection between monetary policy and mortgage rates is such that monetary policy shocks affecting the level of the nominal yield curve have larger real effects than transitory shocks.

2.4. **Heterogeneity and monetary policy: the role of indirect channels.** The literature also emphasizes the role of the indirect effects of monetary policy. That is, monetary policy channels that do not operate through the usual intertemporal substitution effect or direct income effects (e.g. interest or mortgage rates, Fisher effect). For instance, Werning (2015) emphasizes that heightened uncertainty for households leads them to desire precautionary savings for any given current income, which depresses aggregate consumption and income, which should be met with a more accommodating monetary policy stance when shocks occur. Ravn and Sterk (2017) make a similar point, emphasizing that the higher the degree of household risk aversion, the more aggressively the monetary authority should react to a negative shock to ensure that agents' expectations of worsening labor market outcomes and low inflation do not become a self-fulfilling prophecy.

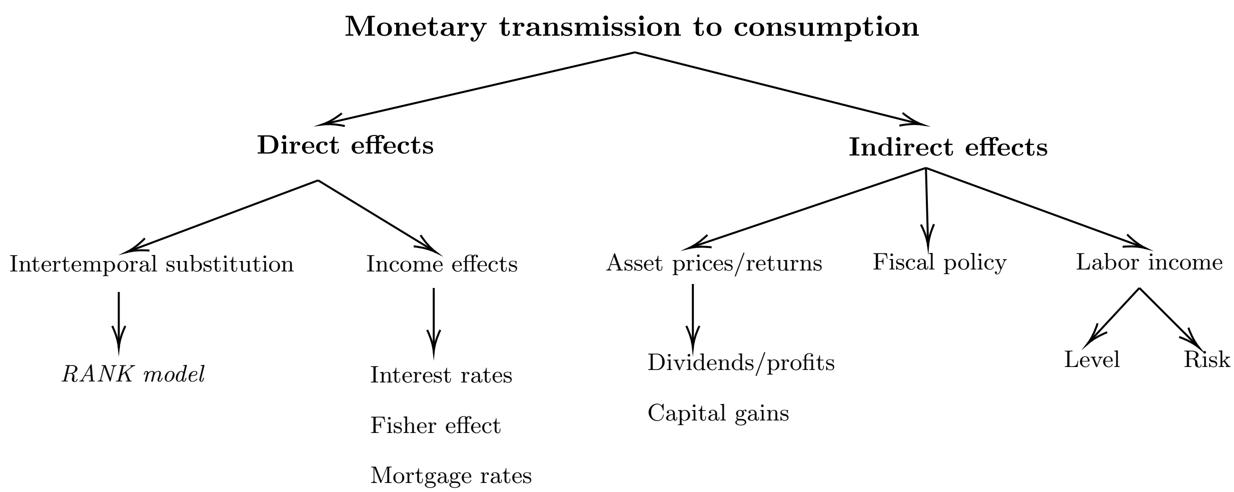
Other indirect effects are related to asset prices or returns. Bilbiie (2008) analyzes the consequences of limited asset market participation for monetary policy. He finds that moderate participation rates in the asset market strengthens the role of monetary policy. Interestingly, he reports that when participation rates are low enough, the "Taylor principle" is inverted, optimal welfare-maximizing discretionary monetary policy requires a passive policy rule. The intuition for his results is that interest rate changes modify the intertemporal consumption and labor supply profiles of asset holders. In a general equilibrium setting, this affects the real wage and consumption of households with no assets. Variations in the real wage lead to variations in profit and hence in the dividend income of asset holders. This feedback effect between households with and without assets is key to understanding the transmission of monetary policy to the economy. Alves, Kaplan, Moll, and Violante (2020) also focus on the uneven distribution of assets among agents. A rate cut drives the price of equity up, benefiting wealthier households who disproportionately own assets. At the same time, less wealthy individuals enjoy an increase in labor income. The question of who benefits the most from a monetary policy cut is ultimately linked to the level of capital adjustment costs in the economy. If capital adjustment costs are low, capital formation strongly reacts and labor income increases, leading to less wealthy households benefiting the most. If capital adjustment costs are high, investment is limited and labor income barely reacts, while equity prices increase, leading to wealthy households benefiting the most.

2.5. **Heterogeneity strengthens the case for collaboration between fiscal and monetary policies.** Another key point made by the HANK literature is that heterogeneity strengthens the case for collaboration between fiscal and monetary policies. In RANK models, Ricardian equivalence applies, so debt-funded government spending has no real effects on consumption or output. Also, because RANK models do not have a distribution of MPCs across agents, any within-period redistribution of income from one group to another has no impact on aggregate variables. In a HANK model, Ricardian equivalence does not hold because of the presence of borrowing constraints and because there exists a diversity of MPCs in the population. An expansionary monetary policy shock reduces interest payments on government debt, which allows for an increase in transfers. High-MPC households, who are close to or at the borrowing constraint react by increasing consumption, while low-MPC households offset to some extent the increase in taxes by working more, which has the overall effect of boosting aggregate demand (Oh and Reis (2012)). Kaplan, Moll, and Violante (2018) use their HANK model calibrated on the US economy to show that the type of fiscal response after a monetary shock matters for the response of the economy. For instance, they report that if the government reacts to reduced interest payment by increasing government spending, the increase in aggregate output is larger than if the government increases transfers while keeping expenditures constant.

Even in frameworks that abstract from monetary policy, the literature has found heterogeneity to be useful to build theories to explain the effects of fiscal policies. For instance, the RA literature is ill-suited to study the type of fiscal response that was undertaken by most governments during the Great Recession of 2007-2009. For example, the U.S. government increased its spending by more than 14% in 2009, with three quarters of this change being explained by an increase in transfers, which include spending related to medical expenses, retirement and disabilities and unemployment insurance. In a RA model, because the behavior of individual agents is similar to the behavior of a representative agent which only considers aggregate variables, non-distorting transfers funded by lump-sum taxes have no aggregate impact since they leave aggregate variables unchanged.

Oh and Reis (2012), with the objective of explaining the transfer-driven fiscal expansion of 2007-2009 in the U.S., build a model in which an increase in targeted non-distorting transfers is expansionary. Their model features households deciding how much to save and whether or not to work. Households face uninsurable income and health risks, as well as nominal price rigidity for firms. Transfers are devised to be redistributed from wealthy and healthy workers to less-wealthy and less-healthy individuals, which is qualitatively what we observe in the data. The expansionary effects of targeted transfers comes from two effects: a neoclassical channel and a Keynesian channel. The neoclassical channel comes from the fact that wealthy workers, who are paying the transfers, react by increasing their hours worked, while less-wealthy and less-healthy recipients increase consumption with almost no effects on hours worked, since they are not likely to already work in the first place. Consequently, both employment and consumption increases. The Keynesian channel results from recipients of transfers having on average higher MPC than the households paying the transfers. Transfer boosts demand and because prices are rigid, firms react by producing more and hiring more workers.

Figure 1. Transmission mechanism of monetary policy to consumption



Notes. This diagram shows the different mechanisms through which monetary policy affects consumption in RANK and HANK models. It is based on the work of Moll (2020).

2.6. **Heterogeneity as a solution to long-standing puzzles.** Heterogeneity has also been invoked to solve long-standing puzzles in the field of monetary economics. For instance, the "forward guidance puzzle" is the observation that forward guidance is too effective in RANK models. This extreme effectiveness is coupled with a curious result. In a RANK model, when making announcements about the future path of the policy rate, the further into the future the cut is supposed to happen, the more effective it is today. McKay, Nakamura, and Steinsson (2016) show that the effects of forward guidance is more reasonable in a HANK model. The intuition is that the forward guidance puzzle is linked to the permanent income consumer behavior of the representative agent, who smooths consumption over time. When agents face uninsurable income risk and borrowing constraints, a precautionary savings effect limits their response to changes in future interest rates. Hence, they behave differently from the permanent income consumer, which reduces the effectiveness of forward guidance.[9]

Heterogeneity has also been used to resolve the "missing deflation puzzle". This puzzle is the observation that traditional RANK models cannot generate a deep recession without simultaneously experiencing deflation.[10] Yet, the Great Recession generated a substantial decline in U.S. GDP of approximately 10 percent relative to its pre-crisis trend, while inflation only dropped by approximately 1.5 percentage point (see Christiano, Eichenbaum, and Trabandt (2015)). In 2010, the president of the New York Fed wrote: "the surprise isn't that inflation has fallen. The surprise is that it's fallen so little, given the depth and duration of the recent downturn" (Williams, 2010). Guerrieri and Lorenzoni (2017) use a model with heterogeneous agents and incomplete markets to study a credit crunch similar to what happened during the Great Recession. After an unexpected permanent tightening in consumers' borrowing capacity, some consumers are forced to repay their debt, while other consumers increase their precautionary savings. Both forces depress aggregate demand, which can lead to a deep recession, even if prices move little because of price or wage rigities. Bilbiie (2018) and Bilbiie (2019) also demonstrate that HANK models have the capacity to generate deep recession without a simultaneous deflation spiral. Increase in perceived risk leads to more precautionary savings, reducing aggregate demand and income. The hand-to-mouth consumers, being at or close to the borrowing constraint, react by reducing consumption, which further depresses aggregate demand. These mechanisms can occur even if prices are fixed.

3. Solving Economic Models with Heterogeneity

Taking into account heterogeneity in macroeconomic models is challenging. This section explains why this is the case and why new methods are required. Nowadays, a vast array of different approaches are available. This section then describes techniques that are "non-numerical", in the sense that they mainly use economic assumptions to

[9]Heterogeneity is not the only way to solve the "forward guidance puzzle". For instance, Gabaix (2020) develops a New Keynesian model with bounded rationality in which forward guidance is much less powerful than in the standard RANK model.

[10]Del Negro, Giannoni, and Schorfheide (2015) challenge the very existence of the "missing deflation puzzle". The authors argue that a standard DSGE model with financial frictions predicts well the behavior of the US economy during the Great Recession.

simplify the additional complexity that heterogeneity introduces. Numerical methods, which directly tackle this additional complexity via numerical techniques, are presented later. Pros and cons of each method are also discussed.

3.1. **General setting.** Most macroeconomic models include a dynamic programming problem in which an agent chooses an infinite sequence of actions $\{u_t\}_{t=0}$ to maximize her expected lifetime utility

$$E_0 \sum_{t=0}^{\infty} \beta^t r(x_t, u_t) \tag{1}$$

subject to the transition equation $x_{t+1} = g(x_t, u_t, \varepsilon_{t+1})$, with x_0 an initial value known in the first period $t = 0$, while ε_t is a sequence of i.i.d. random variables with known distribution. The vector x_t denotes the state vector, which is known at time t. The vector u_t is the control vector, which captures the action taken by the agent at time t. The shock ε_{t+1} is realized in period $t + 1$, after the agent has chosen u_t in period t.

Under standard concavity, smoothness and compactness conditions (see Stokey (1989) or Ljungqvist and Sargent (2018)), the problem of finding an infinite series of actions can be reformulated in recursive form, meaning that the optimal series of actions satisfies the following Bellman equation

$$V(x) = \max_u \left\{ r(x, u) + \beta\, \mathbb{E}_\varepsilon \Big[V(g(x, u, \varepsilon)) | x \Big] \right\} \tag{2}$$

where $V(x)$ denotes the value of being in the state x. Under standard conditions, there exists a time-invariant and continuous and differentiable policy function h, which maps the current state to the current optimal control $u_t = h(x_t)$.

There exists two traditional methods to find V and h. The first one, called value function iteration (VFI), consists in starting with a guess for the value function V_0 and repeatedly solving the next equation until convergence

$$V_{j+1}(x) = \max_u \left\{ r(x, u) + \beta\, \mathbb{E}_\varepsilon \Big[V_j(g(x, u, \varepsilon)) | x \Big] \right\} \tag{3}$$

The second one, called time iteration (TI), consists in directly iterating over the policy function $h(x_t)$. Under additional conditions on the primitives such that the value and the policy functions are differentiable[11], one can show that, if the transition function satisfies $\frac{\partial g(x,u,\varepsilon)}{\partial x} = 0$, the optimal policy function must satisfy the Euler equation

$$\partial_1 r(x,u) = -\beta\, \mathbb{E}_\varepsilon \left[\partial_2 r(g(x,h(x),\varepsilon),h(x))\partial_2 g(x,u,\varepsilon)|x \right] \Big\} \tag{4}$$

where $\partial_i r(x,u)$ denotes the partial derivative of the function r with respect to the i^{th} argument, evaluated at (x,u). The condition that $\frac{\partial g(x,u,\varepsilon)}{\partial x} = 0$ is usually not restrictive in the sense that it can generally be achieved by a judicious choice of what constitutes the state and choice vectors.[12] The TI algorithm starts with an initial guess for the value of the policy function $u_0 = h_0(x)$ and then updates the policy using the updating rule based on (4), which requires using a non-linear solver to find u_{j+1} given u_j

$$\partial_1 r(x,u_{j+1}) = -\beta\, \mathbb{E}_\varepsilon \left[\partial_2 r(g(x,u_j,\varepsilon),u_j)\partial_2 g(x,u_{j+1},\varepsilon)|x \right] \Big\} \tag{5}$$

Both the VFI and TI methods require approximating functions that map a vector from R^n to R, where n is the dimension of the state vector x. In a RANK model, the state vector is usually a low-dimensional object which includes a few aggregate variables, which are sufficient for agents to forecast the behavior of the economy. This is no longer the case in HANK models. Indeed, because of the curvature of the utility function, agents have different marginal propensities to consume and save. A redistribution of capital across agents in the current period changes aggregate demand and aggregate savings. As a result, in an economy with N different agents, the exact wealth of the N agents is an element of the endogenous state vector s. In an economy with a continuum of agents, the entire distribution of wealth, an infinite-dimensional object, is part of s.

VFI and TI become unfeasible when the dimension of s becomes too large. Indeed, the number of sample points required to approximate V or h grows exponentially with the dimension of x_t. This is in essence the "curse of dimensionality" first identified and described by Bellman (1961). To circumvent this obstacle, the economic literature has developed both non-numerical and numerical techniques, which are summarized below.

3.2. **Non-numerical techniques.** The previous section underlined that the key computational and theoretical bottleneck of models with heterogeneity is that the continuous

[11]See Benveniste and Scheinkman (1979) for conditions guaranteeing differentiability of the value function. See Santos (1991) for conditions under which $h(x)$ is differentiable.

[12]See Ljungqvist and Sargent (2018) for a discussion.

distribution of agents along one or several dimensions is part of the relevant state variable. Non-numerical techniques proceed by making relevant assumptions so that the distribution is no longer part of the state variable, or that the distribution of agents has endogenously finite support. These methods are "non-numerical" in the sense that they make HANK models behave as if they were RANK models, using combinations of judicious economic assumptions. Once a HANK model behaves as a RANK model, one may study it using a pen-and-paper approach, or more often using standard numerical algorithms such as VFI or TI. Other approaches, which rely on numerical algorithms to directly solve HANK models, without the intermediate step of transforming them into quasi-RANK models, are presented in a subsequent section.

3.2.1. *TANK models.* RANK models have a single representative household. A natural departure from this benchmark is to posit that the economy is populated not by a single representative household, but by two representative households, leading to the Two-Agent New Keynesian (TANK) model. A first representative household, called the unconstrained household, behaves similarly to the representative household in RANK models. This agent is a permanent income consumer, smoothing consumption over time and strongly reacting to changes in the interest rate, but barely reacting to transitory changes in income. The second representative household, called the constrained household, is a hand-to-mouth agent operating at the borrowing constraint (zero asset holdings), thus consuming all of her income. The share of unconstrained and constrained households in the economy is exogenous and not time-varying. Early explorations of this framework can be found in Galí, López-Salido, and Vallés (2007) and Bilbiie (2008).

This structure fits well with the empirical observation that a sizeable fraction of the population is close to or at the borrowing constraint. One key advantage of TANK models is that, from a computational perspective, they are basically modified RANK models. There is no infinite dimensional distribution with continuous support to track. The distribution of wealth has two mass points: one for the unconstrained representative household and another for the constrained representative household. Debortoli and Galí (2017) show that a TANK model calibrated on US data can reasonably well approximate a full HANK model. Limitations associated with the TANK framework are that one cannot study questions related to the effects of monetary policy on the wealth distribution, or that the share of constrained and unconstrained households is exogenous and must be carefully calibrated.

3.2.2. *Limited-heterogeneity and history-truncated models.* To find the solution of a HANK model, one would need to solve for an economy with an infinite number of agents, with the state of each agent impacting her own decision and the decisions of other agents in the economy. This approach is not feasible in practice. TANK models assume that the economy is populated by a *finite* number of representative agents denoted by N, with $N = 2$ (unconstrained and constrained agents). Challe and Ragot (2016) show that the finiteness of the number of representative agents, assumed by the TANK literature, can be the endogenous outcome of an economy in which unemployment is an “absorbing state”, in the sense that after some periods into unemployment, all unemployed workers behave similarly. The history is effectively “truncated”, as the absorbing state creates a finite number of possible histories, eliminating the need to keep track of the entire history of each agent.

In their model, they consider two types of households. First, patient households who have access to complete insurance markets and behave as permanent income consumers. Hence, the behavior of patient households can be captured by a representative patient household. Second, impatient households who have access to unemployment insurance, which they can complement with self-insurance by holding assets, that are claims on the capital stock. Impatient households can borrow up to an exogenous limit. Assuming a utility function that is concave and then linear above a certain consumption threshold for impatient households, the authors demonstrate that under some conditions, there are only two types of asset holdings for the impatient households. For a sufficiently high discount factor or a sufficiently generous unemployment insurance scheme, unemployed households reach the borrowing constraint in one period and borrow up to the borrowing limit. Employed households hold the same level of assets, due to the linear part of the utility function, to self-insure against possible future unemployment.

Challe, Matheron, Ragot, and Rubio-Ramirez (2017) build upon this idea to show that HANK models can be approximated by a model in which agents’ choices depend only on finite histories. Instead of assuming a concave-linear utility function for impatient households, they assume that households belong to representative families based on their employment histories. Let L denote the number of periods spent in unemployment at the beginning of period t, with $L = 0$ for employed workers. Within each representative family, assets are pooled together at the beginning of the period, and the head of the family makes decisions on how much to consume and save to maximize intertemporal welfare for all family members. Consequently, agents with the same L end up with similar consumption and saving decisions. As previously, employed households belong to a particular family characterized by full risk sharing. As a result, all employed workers make similar decisions regarding investment and consumption. There is no risk-sharing across unemployed families. As previously, one key condition is that unemployment is an absorbing state after a certain number of unemployment periods, in the sense that after L consecutive periods of unemployment, households are at the borrowing constraint and as a result consume the same amount. This condition ensures that there exists a finite number of relevant histories that can arise in equilibrium. As a result, there is also a finite number of possible assets holdings for households. If the borrowing constraint is high enough, the authors show that unemployment is indeed an absorbing state.

One limitation of the construction of Challe, Matheron, Ragot, and Rubio-Ramirez (2017) is that the model does not converge to a fully fledged HANK model because their framework assumes homogeneity across employed households. On the practical side, this implies that the model cannot be estimated using inequality metrics among employed workers. Le Grand and Ragot (2022) provide an alternative construction that asymptotically approaches a fully fledged HANK model. Instead of truncating the histories of unemployed workers only, the authors build truncated histories for unemployed *and* employed workers. The history of an agent i up until period t can be encoded as $s^{i,t} = (..., s^t_{-N+1}, s^t_{-N}, s^t_{-N-1}, ..., s^t_{-1}, s^t_0)$, where s^t_0 denotes the current idiosyncratic status of agent i and s^t_{-k} her idiosyncratic status k periods in the past.[13] In the true HANK model, the entire history of an agent $s^{i,t}$ is the relevant state variable for the agent to make a decision. The assumption of Le Grand and Ragot (2022) is that only the last N elements of a history are relevant, denoted as $S^N \equiv (-s^t_{-N+1}, s^t_{-N}, s^t_{-N-1}, ..., s^t_{-1}, s^t_0)$. Each agent in the economy belongs to a history-specific representative family, in which all resources are shared. Agents in the same history-specific representative family make the same decisions regarding consumption, savings, labor supply or other economic decisions. This implies that the model needs to be solved for a finite number of representative

[13] In the HANK literature, the variable s^t_{-k} is usually the employment status (employed or unemployed) and an idiosyncratic productivity level $s^t_{-k} = \left(e^t_{-k}, y^t_{-k}\right)$.

agents, corresponding to the limited number of truncated histories that can arise in equilibrium. This representation naturally converges to the true model as N goes to infinity. Indeed, when N goes to infinity, each agent is her own representative family, as each agent has a different history, and the assumption of resource pooling inside the family is irrelevant.[14]

Advantages of limited-heterogeneity and history-truncated models include that they can be studied using standard computing tools used for DSGE models and that one can calculate optimal policies. Disadvantages include that these methods introduce new parameters, such as L or N. In Challe, Matheron, Ragot, and Rubio-Ramirez (2017), one must find L by trial and error to ensure that unemployed workers are indeed at the borrowing constraint after L periods. In Le Grand and Ragot (2022), one must try several history lengths N to assess the stability of results.

3.2.3. *No-trade equilibrium models.* HANK models are challenging to solve because the accumulation of idiosyncratic histories of agents across time results in a continuous distribution of agents along one or several dimensions, typically including wealth. Along the business cycle, this infinite-dimensional wealth distribution fluctuates, which prevents economists from using standard numerical techniques. TANK or history-truncated models address this challenge by reducing the number of different histories that matter, thus collapsing the continuous distribution of agents to a distribution with a finite number of mass points, which is amenable to traditional techniques.

No-trade equilibrium models achieve tractability by making the continuous distribution totally irrelevant. As shown in Krusell, Mukoyama, and Smith (2011), this can be achieved by assuming that (i) assets are in zero net supply and production only requires labor, (ii) agents cannot borrow. If agents cannot borrow and because assets are in zero net supply, agents cannot save. The consequence is that all agents in the economy have exactly zero asset holdings every periods. Because the distribution of wealth is summarized by a single mass point (everyone owning exactly zero assets) and is not time-varying, the authors are able to derive closed-form solutions for their economy. Despite the fact that the assumptions used by the authors lead to an "autarky" equilibrium,

[14] This is a simplified representation of Le Grand and Ragot (2022). The authors provide insights on solving the model while incorporating consumption heterogeneity within the same history-specific representative family.

the authors show that their model is able to reproduce many features of observed asset prices. Ravn and Sterk (2017) use a no-trade equilibrium to study the links between job uncertainty and deep recessions in a rich DSGE model. Werning (2015) shows that in such a no-trade equilibrium, one obtains an aggregate Euler equation that is very similar to the one obtained in a RA model. The difference lies in the discount factor used, which captures the role of idiosyncratic uncertainty and heterogeneity.

No-trade equilibrium models are a powerful setup to study analytically the consequences of heterogeneity because they are highly tractable. However, they cannot be used to study the distribution of wealth in the economy, nor can they be used to study the macroeconomic effects of changes in the saving rates along the business cycle.

3.3. **Numerical techniques.** The previous section presented ways to reduce the complexity of HANK models using economic assumptions. This section presents numerical methods to solve HANK models.

3.3.1. *Forecasting rule methods.* Forecasting rule methods are based on the seminal contribution of Krusell and Smith (1998). As discussed previously, the key technical challenge of HANK models is that the relevant state variable is infinite dimensional, rendering the Bellman equation (2) unusable. For example, in the model studied by Krusell and Smith (1998), which adds aggregate uncertainty to the economy of Aiyagari (1994), agents need to know the wage rate and the interest rate next period in order to make an optimal decision on how much to save and consume today. These quantities in turn depend on the aggregate capital stock. But the aggregate capital stock results from the individual decisions of agents, which depend on their savings. Hence, the entire time-varying distribution of wealth is one element of the state variable. More generally, in a HANK model, the state variable can be decomposed as $x = (x_1, x_2)$, with x_1 a finite-dimensional vector and x_2 an infinite-dimensional vector. In the model of Krusell and Smith (1998), x_1 contains individual-specific capital holdings and productivity levels, as well as the aggregate productivity level of the economy. The infinite-dimensional vector x_2 is the distribution of wealth in the economy.

The key insight of Krusell and Smith (1998) is to replace x_2 by a low-dimensional counterpart $\tilde{x}_2$ and to assume that $\tilde{x}_2$ provides enough information for agents to forecast the value of relevant aggregate variable next period. More formally, the Krusell-Smith (KS) algorithm consists of assuming that $A(x_1, x_2) \approx \tilde{A}(x_1, \tilde{x}_2|\theta)$, where A denotes the vector of aggregate quantities that are needed for the agents to optimize their behavior and $\tilde{A}$ is a parametric approximation of A, which uses $\tilde{x_2}$ instead of x_2 and that depends on an unknown parameter θ. That is, the KS algorithm reduces the dimension of x_2, at the cost of introducing a new unknown vector θ. The parameter θ is calculated via Monte Carlo simulation. Hence, the KS algorithm proceeds in three steps:

(1) Conditional on the value of θ, solve the individual problem using conventional techniques, for instance using VFI with equation (3):

$$V_{j+1}(x_1, \tilde{x_2}|\theta) = \max_u \left\{ r(x_1, \tilde{x_2}, u) + \beta\, \mathbb{E}_\varepsilon \left[V_j(g(x_1, \tilde{x_2}, u, \varepsilon))|x, \theta \right] \right\} \tag{6}$$

(2) Based on the solution obtained in the first step, simulate the economy and update the parameter θ such that $\tilde{A}(x_1, \tilde{x_2}|\theta)$ provides a good approximation of aggregate quantities next period.

(3) Repeat steps 1 and 2 until convergence of the forecasting parameter θ, which can be checked using the $L2$.

In the model of Krusell and Smith (1998), A is the aggregate level of capital denoted by K. The authors show that using the first moment of distribution of wealth is enough to build a sufficiently precise forecasting rule $\tilde{A}(x_1, \tilde{x_2}|\theta)$, a result known as "approximate aggregation". For the functional form of the forecasting rule, they use a log-log linear regression of the form $\log(K_{t+1}) = \theta_0 + \theta_1 \log(K_t)$. However, the KS algorithm is not restricted to using only the first moment of the distribution. If the first moment does not contain enough information, one may include higher moments. However, increasing the dimension of $\tilde{x_2}$ makes the first step slower, as more computations are needed in equation (6). Also, the KS algorithm does not restrict the type of tools that can be used to estimate the forecasting rule. For instance, Fernández-Villaverde, Hurtado, and Nuno (2019) use an artificial neural network to find a precise forecasting rule.

Advantages associated with the KS algorithm are that it is an intuitive approach and that it is relatively straightforward to implement. It is also a global method, that does not rely on linearization around a non-stochastic steady state. In terms of disadvantages, the KS algorithm can be computationally intensive and time-consuming. It requires simulating the economy for multiple periods to estimate accurate forecasting rules, and this process needs to be repeated until the forecasting parameter θ converges. Another disadvantage of the KS algorithm is that once a solution has been found, one cannot rule out that other equilibria exist. Indeed, one assumes a given forecasting rule for all agents in the economy and finds a solution conditional on that assumption. However, it might be the case that agents use other equally precise forecasting rules that lead to another equilibrium. In practice, one may use different estimation procedures for the forecasting rule to assess the stability of results, but this can be a time-intensive procedure.

3.3.2. *Linearization in state space.* A popular alternative to the KS algorithm is linearization of models in state space. This approach is based on the recursive representation of a dynamic programming problem using the Bellman equation (2). In the state space representation of an economic model, there exists a state variable that encodes all the information required for an agent to make an optimal decision. This method was first developed by Reiter (2009), realizing that the standard linearization around the non-stochastic steady-state of DSGE models can be extended to models with heterogeneity.

Linearization around the non-stochastic steady-state is another widely used alternative to VFI or TI methods presented in section 3.1. It proceeds by recasting a model in the form of an equation $\mathbb{E}_\varepsilon\left[F(X_{t-1}, X_t, X_{t+1}, \varepsilon_t)\right]$ that must be equal to zero in expectation.[15] Linearization relies on the fact that one can often easily find solutions for the special case with no aggregate uncertainty $\varepsilon_t = 0$, also called the non-stochastic steady-state: $F(\bar{X}, \bar{X}, \bar{X}, 0) = 0$. In simple DSGE models, analytical solutions for $\bar{X}$ are available. Otherwise, root-finding algorithms may be used. Then, in a way analogous to how one may calculate an approximation of a one-dimensional smooth function $f(x)$ in the neighborhood of a using the formula $f(x) \approx f(a) + f'(a)(x - a)$, one can calculate a solution to the economic model in the vicinity of the non-stochastic steady-state. More specifically, note that a linear approximation of the model $\mathbb{E}_\varepsilon\left[F(X_{t-1}, X_t, X_{t+1}, \varepsilon_t)\right] = 0$ is given by

$$A\,\mathbb{E}_\varepsilon\left[\tilde{X}_{t+1}\right] + B\tilde{X}_t + C\tilde{X}_{t-1} + D\varepsilon_t = 0 \tag{7}$$

where $\tilde{X}_t \equiv X_t - \bar{X}$ denotes a vector that measures deviations from the steady state $\bar{X}$ and the matrices A through D are Jacobian matrices evaluated at $\bar{X}$. For instance, A is the Jacobian matrix where entry $A_{i,j}$ is given by $\frac{\partial F_{3,i}}{\partial X_j}$ evaluated at $\bar{X}$, with the function

[15] For instance, based on equation (4), the optimal policy function satisfies $\mathbb{E}_\varepsilon\left[\beta\partial_2 r(g(x, h(x), \varepsilon), h(x))\partial_2 g(x, u, \varepsilon) - \partial_1 r(x, u)|x\right]\Big\} = 0$, which implicitly defines a function F for some elements of X. The function F also results from accounting equations of the model (e.g. the budget constraint) or technological specifications (e.g. production function, or the innovation process).

$F_3(X) \equiv F(\tilde{X}, \tilde{X}, X, 0)$, which fixes the first two and the last input of F to their steady state values. Several methods are available to solve equations of the form (7), including the methods developed by Blanchard and Kahn (1980), Sims (2002) or Rendahl (2017).

In a RANK model, linearization around the non-stochastic steady-state is trivial because X consists of a few aggregate variables, such as aggregate consumption and aggregate capital. In a HANK model, it is not easy to properly define X. First, X contains the distribution of agents across one or several dimensions. Linearization cannot work with infinite-dimensional vectors, so Reiter (2009) uses a histogram represented by the vector h to approximate the distribution as a finite-dimensional object. Second, the decisions of agents in the economy depends on h. Since policy rules are functions, which are also infinite-dimensional vectors, one must also find finite-dimensional counterparts for them. For instance, one can use the values of the policy rules, denoted by a, calculated on a predetermined grid and use linear interpolation to find off-grid values. In that way, one captures how the behavior of agents is impacted by changes in h, while maintaining a finite-dimensional state space.

The algorithm developed by Reiter (2009) can be presented as follows:

(1) Include h, a, as well as the other usual aggregate variables in X.
(2) Solve for the steady-state with no aggregate uncertainty using a solver.
(3) Linearize the system numerically to obtain a system of the form (7).
(4) Find a solution of the linearized system using traditional methods.

One key advantage of the Reiter algorithm is its computational efficiency. It is often two orders of magnitude faster than the KS algorithm. One drawback of Reiter (2009) is that h can be potentially a very large vector. This can be problematic because the larger the dimension of h, the larger the matrices A through D, increasing the computation time required to find a solution of (7). However, Winberry (2018) shows how the algorithm can be improved by using a certain parametric distribution to approximate distributions instead of using histograms. He employs a specific parametric density, parameterized by the vector m. Since the dimension of m is often much smaller than the dimension of h, this usually leads to significant speed improvement.

3.3.3. *Linearization in sequence space.* As discussed in section 3.1, economic models start with a sequential formulation of a dynamic programming problem. Agents must choose a series of actions based on an initial condition and on the history of realized idiosyncratic shocks up to the current period. Under some standard conditions, the Bellman principle applies and the problem of finding an infinite series of actions can be reformulated in the recursive form of equation (2), in which the state variable x encodes all the relevant history of idiosyncratic shocks. When x is low-dimensional, the Bellman principle drastically reduces complexity of the problem. However, when x is high-dimensional, as in models with heterogeneity, the advantages of using x to summarize the state are no longer obvious. That is why methods based on linearization in sequence space avoid using the Bellman principle, but instead use the more primitive sequential form of the dynamic problem. They solve the model in sequential form around a steady-state and they use this information to build an approximation of true dynamics around this steady-state using linearization techniques.

Boppart, Krusell, and Mitman (2018) present a general method that exploits sequence space in an intuitive way. Let x_t denote the value of an endogenous aggregate variable, for instance consumption. In sequence space, the variable x_t can be regarded as a function of the entire history of an exogenous variable z_t, for instance aggregate productivity: $x_t = f(z_t, z_{t-1}, z_{t-2}, ...)$. The authors assume that the model is linear with respect to the exogenous variable z_t, which is equivalent to assuming that the function f is additively separable and that the magnitude of shocks z_t only scales up or down a reference path $(f_0, f_1, ..., f_k, ...)$

$$x_t = z_t f_0 + z_{t-1} f_1 + ... + z_{t-k} f_k + ... \tag{8}$$

where f_k denotes the impact in period t of a one standard deviation of shock that occurred k periods ago.[16] The method assumes that the effects of a shock on today's value is negligible after T periods, so that one has $x_t = z_t f_0 + z_{t-1} f_1 + ... + z_{t-k} f_k$, or more compactly

$$x_t = \sum_{k=0}^{T} z_{t-k} f_k \tag{9}$$

which highlights that the value x_t is a moving average of past shock values. To solve and simulate the model, one simply needs to determine the values of f_k. To accomplish this, Boppart, Krusell, and Mitman (2018) propose calculating the perfect foresight transition path of the economy by starting from the non-stochastic steady-state and applying a one

[16]That is, the path of aggregate consumption after a two standard deviation productivity shock has the same *shape* as the path of aggregate consumption after a one standard deviation productivity shock. While the shape of consumption is preserved, its *magnitude* is multiplied by two.

standard deviation shock. Finding the perfect foresight transition path can be done for instance using a "shooting" algorithm, in which one (i) solves for the policy functions of agents assuming a given path for the economy, (ii) updates the path of the economy using the policy functions obtained in the first step. This procedure is repeated until convergence of the path according to a predetermined convergence criteria. In summary, if one assumes that the model is linear with respect to aggregate shocks, solving for the perfect foresight transition path after a shock provides all the information necessary to solve and simulate the model outside its steady state.

However, one limitation of Boppart, Krusell, and Mitman (2018) is that calculating a perfect foresight transition path can be slow or unstable. Auclert, Bardóczy, Rognlie, and Straub (2021) offer an alternative method, which also utilizes the sequence space formulation of an economic model. An economic model can generally be expressed as a solution to a non-linear system

$$F(X, Z) = 0 \tag{10}$$

where X is the time path of endogenous variables and Z the path of exogenous shocks. More specifically, X is $n_x \times T$ matrix $X \equiv \left(X_1, X_2, ..., X_T\right)$ where each vector X_t is a $n_x \times 1$ vector of endogenous variables at time t, with n_x denoting the number of endogenous variables and T the number of periods after which the economy is assumed to be back at equilibrium. Similarly, Z is $n_z \times T$ matrix $Z \equiv \left(Z_1, Z_2, ..., Z_T\right)$ where each vector Z_t is a $n_z \times 1$ vector of endogenous variables at time t, with n_z denoting the number of exogenous variables. Equation (10) is a non-linear system of $n_x \times T$ equations in $n_x \times T$ endogenous variables.

Assuming the necessary conditions for the implicit function theorem to apply, one can obtain the path of endogenous variables around the steady state, following a shock dZ, using the formula:

$$dX = -F_X^{-1} F_Z dZ \equiv G dZ \tag{11}$$

where F_X and F_Z represent the Jacobian matrices of F with respect to X and Z evaluated at the steady-state values of the model (X_{SS}, Z_{SS}).[17] One key contribution of

[17] This formula can be illustrated by considering a circle with radius 1, which is defined by the equation $f(x, y) = x^2 + y^2 - 1 = 0$. Let us consider the point on the circle $(x_{ss}, y_{ss}) = (\frac{1}{2}, \frac{\sqrt{3}}{2})$. Formula (11) gives $dx = -\frac{y_{ss}}{x_{ss}} dy = -\sqrt{3} dy$. Let us consider a slight deviation of y from y_{ss}, for example $dy = 0.01$.

Auclert, Bardóczy, Rognlie, and Straub (2021) is the development of an efficient algorithm, referred to as the "fake news" algorithm, for the efficient computation of the Jacobian matrices F_X and F_Z.

Methods utilizing linearization in sequence space generally offer the advantage of computational efficiency. However, as with methods employing linearization in state space, they require the calculation of a steady state, which can be nontrivial at times. Linearization also makes it difficult to study some interesting economic questions because in the linearized model agents are risk neutral.

3.3.4. *Other global methods.* While the KS algorithm is a global method, as it does not rely on linearization around a steady-state, it addresses the dimensionality challenge inherent in models with heterogeneity by introducing a new forecasting rule that needs to be estimated by Monte Carlo. Some authors have used global methods designed to directly confront the high-dimension of the state space, without the need for an intermediate forecasting rule, as in the case of the KS algorithm.

For example, Brumm and Scheidegger (2017) advocate the use of adaptive sparse grids, which are based on sparse grids[18] selectively refined in areas of high curvature and non-differentiability. Using this approach, the number of grid points required to solve (3) or (5) grows gradually with the dimension of x, rather than exponentially. The authors demonstrate successful applications of these methods in solving high-dimensional international real business cycle models with non-linearities. More recently, Maliar, Maliar, and Winant (2021), Azinovic, Gaegauf, and Scheidegger (2022) and Pascal (2024) demonstrate how the combination of neural networks (NN) and Monte Carlo simulation can be used to solve high-dimensional economic models. The key ideas are that (i) neural networks (NN) are largely immune to the curse of dimensionality due to their ability to automatically detect the relevant latent space for a given problem and (ii) utilizing

Using the previous formula, one gets $x = \frac{1}{2} - 0.01\sqrt{3} \approx 0.4826$. The true value for x is given by $x = \sqrt{1 - (\frac{\sqrt{3}}{2} + 0.01)^2} \approx 0.4822$.

[18]Dense grids are based on the Cartesian product of simple univariate grids. For example, if one uses 10 points to approximate one dimension, the two-dimensional dense grid contains 100 grid points. The n-dimensional dense grid contains 10^n grid points. A sparse grid drops selectively some elements of the full Cartesian product grid so that the number of grid points is less than that of the full dense grid.

Monte Carlo simulation to explore the state space instead of predetermined grids leads to significant efficiency gains.

Advantages associated with these global approaches include their independence from linearization around a non-stochastic steady-state. Thus, they can be employed to study highly non-linear models that are not well-suited for linearization. In terms of disadvantages, these methods are generally slower than alternatives based on linearization, and they may require significant coding skills in order to be properly implemented. Slowness of these global methods may prevent them from being estimated using empirical observations, since estimation usually requires solving the same model thousands of times in order to compute simulated moments, likelihood functions, or posterior distributions. As a result, economic models solved using global methods are often calibrated rather than estimated.

4. Conclusion

In a 2016 speech, Janet Yellen listed the role of heterogeneity in macroeconomics as a central topic to analyze for the years to come[19]. She highlighted that "the various linkages between heterogeneity and aggregate demand are not yet well understood, either empirically or theoretically" and that "studying monetary models with heterogeneous agents more closely could help us shed new light on [new] aspects of the monetary transmission mechanism". She also noted that "even though the tools of monetary policy are generally not well suited to achieve distributional objectives, it is important for policymakers to understand and monitor the effects of macroeconomic developments on different groups within society." Five years later, Isabel Schnabel from the ECB stated that "today, heterogeneity in income and wealth is widely considered to be a prime channel of policy transmission"[20]. This shift in the consensus view underlines that economists have invested a considerable amount of time and effort in improving our understanding of this topic.

This paper reviews the body of work that has contributed to our understanding of how heterogeneity can affect aggregate dynamics. The first part of this literature review focuses on the key mechanisms that have been identified, with a special emphasis on

[19]See `https://www.federalreserve.gov/newsevents/speech/yellen20161014a.htm`

[20]See `https://www.bis.org/review/r211123f.htm`

the interactions of heterogeneity and monetary or fiscal policies. The literature review underlines that modern macroeconomics was built around the concept of the Representative Agent (RA). The RA embodies the idea that heterogeneity among households or firms does not matter at the aggregate level. Empirical work, along with theoretical and computational developments, has led economists to question the RA assumption. Economists have since introduced heterogeneity in their models, finding that it significantly affects economic outcomes. The transmission mechanism of monetary policy can differ significantly between a RA model and a Heterogeneous Agent (HA) model. A key finding in the literature is that while the propagation channels of monetary policy in a RA model are almost entirely driven by the intertemporal substitution of consumption, HA models have large indirect effects that are absent from RA models. To paraphrase Kaplan and Violante (2018), heterogeneity restores Keynesian effects in New Keynesian models, which tend to be muted once a RA is assumed. Another important theme in the literature is that, since general equilibrium effects are strong and Ricardian equivalence does not apply in HA models, heterogeneity strengthens the case for collaboration between fiscal and monetary policies. The literature review also makes the point that using HA models instead of RA models can be justified on the basis that more data can be used to estimate HA models. Using microeconomic data to estimate macroeconomic models is a promising area of research, as models become increasingly complicated and additional time series are needed to estimate structural parameter values.

The second part of the literature review focuses on the analytical and computational methods available to solve models with heterogeneity. One could argue that the RA assumption was long used not because of its plausibility, but because no methods were available to solve models with heterogeneity in a reasonable amount of time. This is no longer the case, as a vast array of methods have since been developed. The review considers both analytical and numerical methods. Analytical methods rely on crucial ad-hoc assumptions to simplify models and make them amenable to traditional techniques. For example, Two-Agent New Keynesian (TANK) models assume that there exist two types of households: the constrained and the unconstrained. No-trade equilibrium models assume that assets are in zero net supply and that agents cannot borrow, so that everyone owns exactly zero assets every period. Limited-heterogeneity and history truncated models effectively reduce the number of states that can arise in equilibrium, transforming HA models into quasi-RA models, so that standard numerical techniques used to solve

DSGE models can be utilized. Numerical methods rely on computational approaches to find approximations of HA models. There exists global computational methods (e.g., those based on forecasting rules or neural networks) or methods based on linearization around a steady-state (using the state-space or the sequence space representations). The pros and cons of each method have been discussed.

References

Aiyagari, S. R. (1994): "Uninsured idiosyncratic risk and aggregate saving," *The Quarterly Journal of Economics*, 109(3), 659–684.

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

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

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

Azinovic, M., L. Gaegauf, and S. Scheidegger (2022): "Deep equilibrium nets," *International Economic Review*.

Barro, R. J. (1974): "Are government bonds net wealth?," *Journal of political economy*, 82(6), 1095–1117.

Bayer, C., B. Born, and R. Luetticke (2022): "Shocks, Frictions, and Inequality in US Business Cycles," CEPR Discussion Papers 14364.

Bellman, R. (1961): "Adaptive Control Processes: A Guided Tour," .

Benveniste, L. M., and J. A. Scheinkman (1979): "On the differentiability of the value function in dynamic models of economics," *Econometrica: Journal of the Econometric Society*, pp. 727–732.

Beraja, M., A. Fuster, E. Hurst, and J. Vavra (2019): "Regional heterogeneity and the refinancing channel of monetary policy," *The Quarterly Journal of Economics*, 134(1), 109–183.

Berger, D., K. Milbradt, F. Tourre, and J. Vavra (2021): "Mortgage prepayment and path-dependent effects of monetary policy," *American Economic Review*, 111(9), 2829–78.

Bilbiie, F. O. (2008): "Limited asset markets participation, monetary policy and (inverted) aggregate demand logic," *Journal of economic theory*, 140(1), 162–196.

——— (2018): "A Catch 22 for HANK Models: No Puzzles, No Amplification," *CEPR Discussion Paper*, (12601).

——— (2019): "Monetary policy and heterogeneity: An analytical framework," *2019 Meeting Papers 178, Society for Economic Dynamics.*

BLANCHARD, O. J., AND C. M. KAHN (1980): "The solution of linear difference models under rational expectations," *Econometrica: Journal of the Econometric Society*, pp. 1305–1311.

BOPPART, T., P. KRUSELL, AND K. MITMAN (2018): "Exploiting MIT shocks in heterogeneous-agent economies: the impulse response as a numerical derivative," *Journal of Economic Dynamics and Control*, 89, 68–92.

BRUMM, J., AND S. SCHEIDEGGER (2017): "Using adaptive sparse grids to solve high-dimensional dynamic models," *Econometrica*, 85(5), 1575–1612.

CALVO, G. A. (1983): "Staggered prices in a utility-maximizing framework," *Journal of monetary Economics*, 12(3), 383–398.

CAMPBELL, J. Y., AND N. G. MANKIW (1989): "Consumption, income, and interest rates: Reinterpreting the time series evidence," *NBER macroeconomics annual*, 4, 185–216.

CANZONERI, M. B., R. E. CUMBY, AND B. T. DIBA (2007): "Euler equations and money market interest rates: A challenge for monetary policy models," *Journal of monetary economics*, 54(7), 1863–1881.

CHALLE, E., J. MATHERON, X. RAGOT, AND J. F. RUBIO-RAMIREZ (2017): "Precautionary saving and aggregate demand," *Quantitative Economics*, 8(2), 435–478.

CHALLE, E., AND X. RAGOT (2016): "Precautionary saving over the business cycle," *The Economic Journal*, 126(590), 135–164.

CHRISTIANO, L. J., M. S. EICHENBAUM, AND M. TRABANDT (2015): "Understanding the great recession," *American Economic Journal: Macroeconomics*, 7(1), 110–167.

CLARIDA, R., J. GALI, AND M. GERTLER (1999): "The science of monetary policy: a new Keynesian perspective," *Journal of economic literature*, 37(4), 1661–1707.

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

DEBORTOLI, D., AND J. GALÍ (2017): "Monetary policy with heterogeneous agents: Insights from TANK models," *Economics Working Papers, Universitat Pompeu Fabra*, (1686).

DEL NEGRO, M., M. P. GIANNONI, AND F. SCHORFHEIDE (2015): "Inflation in the great recession and new keynesian models," *American Economic Journal: Macroeconomics*, 7(1), 168–196.

Eichenbaum, M., S. Rebelo, and A. Wong (2022): "State-dependent effects of monetary policy: The refinancing channel," *American Economic Review*, 112(3), 721–761.

Fernández-Villaverde, J., S. Hurtado, and G. Nuno (2019): "Financial frictions and the wealth distribution," Discussion paper, National Bureau of Economic Research.

Furceri, D., P. Loungani, and A. Zdzienicka (2018): "The effects of monetary policy shocks on inequality," *Journal of International Money and Finance*, 85, 168–186.

Gabaix, X. (2020): "A behavioral New Keynesian model," *American Economic Review*, 110(8), 2271–2327.

Galí, J., J. D. López-Salido, and J. Vallés (2007): "Understanding the Effects of Government Spending on Consumption," *Journal of the European Economic Association*, 5(1), 227–270.

Galí, J. (2015): *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press.

Garriga, C., F. E. Kydland, and R. Šustek (2017): "Mortgages and monetary policy," *The Review of Financial Studies*, 30(10), 3337–3375.

Gorman, W. M. (1953): "Community preference fields," *Econometrica: journal of the Econometric Society*, pp. 63–80.

Guerrieri, V., and G. Lorenzoni (2017): "Credit crises, precautionary savings, and the liquidity trap," *The Quarterly Journal of Economics*, 132(3), 1427–1467.

Hartley, J. E. (1996): "Retrospectives: The origins of the representative agent," *Journal of Economic Perspectives*, 10(2), 169–177.

Johnson, D. S., J. A. Parker, and N. S. Souleles (2006): "Household expenditure and the income tax rebates of 2001," *American Economic Review*, 96(5), 1589–1610.

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

Kaplan, G., and G. L. Violante (2018): "Microeconomic heterogeneity and macroeconomic shocks," *Journal of Economic Perspectives*, 32(3), 167–94.

Khan, A., and J. K. Thomas (2008): "Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics," *Econometrica*, 76(2), 395–436.

Krusell, P., T. Mukoyama, and A. A. Smith (2011): "Asset prices in a Huggett economy," *Journal of Economic Theory*, 146(3), 812–844.

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

Kydland, F. E., and E. C. Prescott (1982): "Time to build and aggregate fluctuations," *Econometrica: Journal of the Econometric Society*, pp. 1345–1370.

Le Grand, F., and X. Ragot (2022): "Managing inequality over business cycles: Optimal policies with heterogeneous agents and aggregate shocks," *International Economic Review*, 63(1), 511–540.

Lenza, M., and J. Slacalek (2018): "How does monetary policy affect income and wealth inequality? Evidence from quantitative easing in the euro area," .

Lise, J., and J.-M. Robin (2017): "The macrodynamics of sorting between workers and firms," *American Economic Review*, 107(4), 1104–1135.

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

Ljungqvist, L., and T. J. Sargent (2018): *Recursive macroeconomic theory*. MIT press.

Lucas, R. (1976): "Econometric policy evaluation: A critique," in *Carnegie-Rochester conference series on public policy*, vol. 1, pp. 19–46. North-Holland.

Maliar, L., S. Maliar, and P. Winant (2021): "Deep learning for solving dynamic economic models.," *Journal of Monetary Economics*, 122, 76–101.

Marshall, A. (1891): *Principles of Economics: Second Edition*. London: Macmillan and Co.

McKay, A., E. Nakamura, and J. Steinsson (2016): "The power of forward guidance revisited," *American Economic Review*, 106(10), 3133–3158.

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

Misra, K., and P. Surico (2014): "Consumption, income changes, and heterogeneity: Evidence from two fiscal stimulus programs," *American Economic Journal: Macroeconomics*, 6(4), 84–106.

Moll, B. (2020): "The Research Agenda: Ben Moll on the Rich Interactions between Inequality and the Macroeconomy," *Economic Dynamics Research Agenda*, 21(2).

Montecino, J., and G. Epstein (2015): "Did quantitative easing increase income inequality?," *Institute for New Economic Thinking working paper series*, (28).

Mumtaz, H., and A. Theophilopoulou (2017): "The impact of monetary policy on inequality in the UK. An empirical analysis," *European Economic Review*, 98, 410–423.

Oh, H., and R. Reis (2012): "Targeted transfers and the fiscal response to the great recession," *Journal of Monetary Economics*, 59, S50–S64.

Papp, T. K., and M. Reiter (2020): "Estimating linearized heterogeneous agent models using panel data," *Journal of Economic Dynamics and Control*, 115, 103881.

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

Pascal, J. (2024): "Artificial neural networks to solve dynamic programming problems: A bias-corrected Monte Carlo operator," *Journal of Economic Dynamics and Control*, 162, 104853.

Ravn, M. O., and V. Sterk (2017): "Job uncertainty and deep recessions," *Journal of Monetary Economics*, 90, 125–141.

Reiter, M. (2009): "Solving heterogeneous-agent models by projection and perturbation," *Journal of Economic Dynamics and Control*, 33(3), 649–665.

Rendahl, P. (2017): "Linear time iteration," Discussion paper, IHS Economics Series.

Saiki, A., and J. Frost (2014): "Does unconventional monetary policy affect inequality? Evidence from Japan," *Applied Economics*, 46(36), 4445–4454.

Santos, M. S. (1991): "Smoothness of the policy function in discrete time economic models," *Econometrica: Journal of the Econometric Society*, pp. 1365–1382.

Sims, C. A. (2002): "Solving linear rational expectations models," *Computational economics*, 20(1-2), 1.

Smets, F., and R. Wouters (2003): "An estimated dynamic stochastic general equilibrium model of the euro area," *Journal of the European economic association*, 1(5), 1123–1175.

Smets, F., and R. Wouters (2007): "Shocks and frictions in US business cycles: A Bayesian DSGE approach," *American economic review*, 97(3), 586–606.

Stokey, N. L. (1989): *Recursive methods in economic dynamics*. Harvard University Press.

Werning, I. (2015): "Incomplete markets and aggregate demand," Discussion paper, National Bureau of Economic Research.

Williams, J. (2010): "Sailing into headwinds: the uncertain outlook for the US economy," Federal Reserve Bank of San Francisco.

Winberry, T. (2018): "A method for solving and estimating heterogeneous agent macro models," *Quantitative Economics*, 9(3), 1123–1151.

Yogo, M. (2004): "Estimating the elasticity of intertemporal substitution when instruments are weak," *Review of Economics and Statistics*, 86(3), 797–810.

BANQUE CENTRALE DU LUXEMBOURG
EUROSYSTÈME

2, boulevard Royal
L-2983 Luxembourg

Tél.: +352 4774-1
Fax: +352 4774 4910

www.bcl.lu • info@bcl.lu