Monetary Policy and Wealth Inequality over the Great Recession in the UK. An Empirical Analysis

Haroon Mumtaz
Queen Mary, University of London

Angeliki Theophilopoulou
Brunel University London

February 2019

Abstract

The UK has experienced a dramatic increase in wealth and income inequality over the past four decades. By using detailed micro information at household level from the Wealth and Assets Survey we construct monthly historical measures of wealth inequality from 2005 to 2016. We investigate the dynamic relationship between conventional and unconventional monetary policy and whether it played a role in the evolution of wealth inequality measures. Our findings suggest that expansionary monetary policy shocks lead to an increase in wealth inequality in the UK and contribute significantly to its fluctuation. The heterogeneous response of wealth at different quantiles suggests that financial easing has a larger positive effect on high income households and the portfolio channel had a stronger impact than the labour income channel to the low percentiles of the distribution. Our evidence also suggests that the policy of quantitative easing may have contributed to the increase in inequality over the Great Recession.

Preliminary Draft

Keywords: Inequality, Wealth, FAVAR, Monetary Policy Shocks
JEL No. D31, E21, E44, E52

1 Introduction

The aftermath of the Great Recession finds a number of countries with increased income and wealth inequality. Even though during the Great Recession wealthy households experienced loss in their wage and financial income and automatic stabilisation policies were set off to support low income families, a decade after the global financial crisis this trend has been reversed and losses have been more than recovered. The OECD reports (2018) that across the 28 countries the Gini coefficient for disposable income has increased to 0.32 in 2016-17 which is higher than 0.30 on 2006-7 and 10% of households hold 52% of total wealth in 2015.

The study of wealth distribution has attracted renewed interest not only because its importance in income inequalities and social cohesion but it's eminent role in consumption smoothing and maintaining a living standard above the poverty line: households with positive wealth holdings can keep their level of consumption stable when experience loss of income due to loss of employment, health, assets, etc. while households with zero or negative wealth remain vulnerable to swings of the business cycle. On the other hand, income from rents, interest and dividends is on an upward trend in the last few decades and has a significant proportion on total income, especially for households in the right tail of income distribution (Crossley *et al.* (2016)).

According to Piketty *et al.* (2018) wealth inequality in the UK as expressed by the wealth holdings of top 10% was in a downward trend until the end of 1990s which reached its historical lower value. This was because of destruction of physical capital during the WW II, redistribution policies and a large number of state owned companies which played an important role in economic activity. Moreover, higher returns in financial assets observed by the end of that period were largely compensated by repeated advances in property prices which was the main asset of middle class households. In the mid 1980s and towards the end of the decade, higher dispersion in earnings, fall of tax progressivity, higher returns of financial assets and mass privatisation of state owned companies led to a slow but steady increase of wealth held by the top 10%. During the 2000s wealth inequality remained unchanged due to rises in house prices which gave a relief of the middle 40% which are mostly homeowners against the also high rise of returns coming from

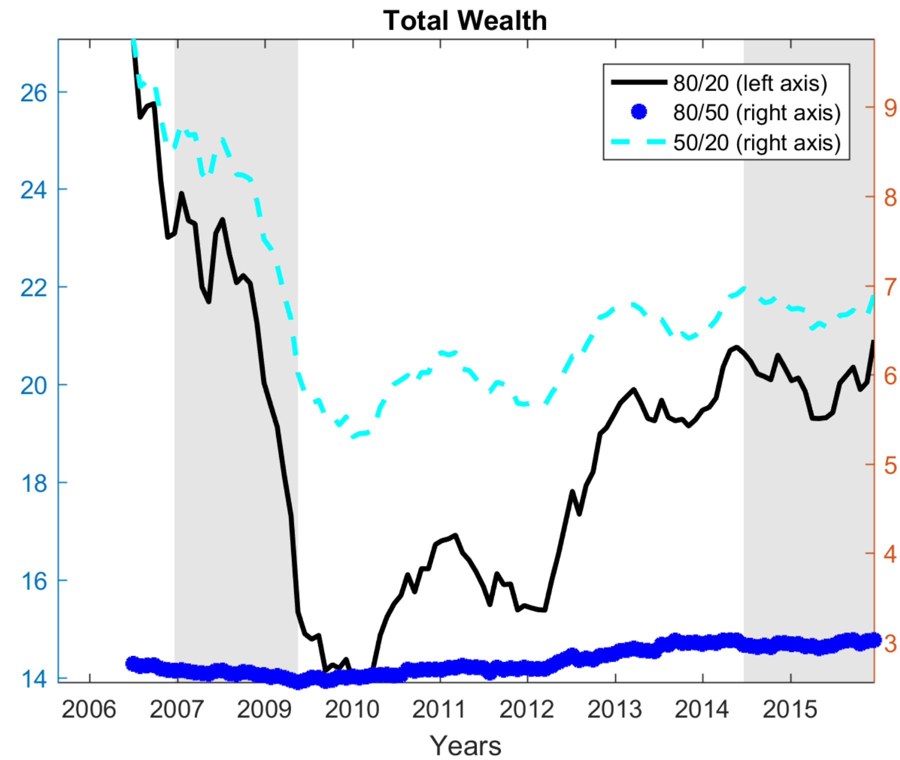

Figure 1: Figure 1: Evolution of Total Wealth in the UK, 2006-2016. The series are constructed by using data from the five waves in Wealth and Assets Survey (WAS).

financial markets. However, the authors of the report note that the rise of property prices generated new types of inequalities between home and no home owners and young earners with no parental gifts which found increasingly difficult to enter the housing market.

The mean net wealth remained unchanged during 2006-11 in the UK while it recovered strongly in 2011-15 with average annual growth of 4% (Balestra and Tonkin (2018)). In Figure 1 we have plot the 80/20 ratio from 2006 to 2016 using data from the ONS Wealth and Assets Survey (WAS). The ratio falls dramatically from 27 in 2006 to 14 in 2010 showing improvement in wealth equality. Then it slowly increases with a second dip in 2012 and after that it depicts a steady recovery. By 2016 it has reached 21 and is in upward trend, having recovered more than 50% of its loss during the financial crisis.

The role of public policies have been proved crucial to wealth distribution. The World Inequality Report (2018) shows that countries with similar growth rates and development have had a different evolution in income and wealth inequality. This supports the importance of public policies to mitigate inequalities and promote a more equitable growth. Even though fiscal policy directly affects income and wealth distributions, the role of monetary policy has largely considered neutral until recently, where a number of researchers find significant distributional effects. Monetary policy can cause heterogeneous effects on households' income and wealth through direct and indirect channels. Change in the policy rate affects directly households' incentive to save or to consume and the return on their net financial assets. The composition of assets and liabilities is crucial for heterogeneous responses. If, for example, liabilities consist of short term or variable rate debt, a decrease of interest rate will benefit more this kind of debt issuers, while debt holders with maturing assets will face reinvestment risk, debt holders in variable rates and savers in current account deposits will be adversely affected. On the other hand, savers in time deposits or bond holders with long term maturities in fixed rates will not be directly affected.

The indirect channel operates through employment and labour income. Expansionary monetary policy aims to boost investment, GDP and employment. A positive impact will depend highly on prices and wages' rigidity. Through an increase in employment, the labour income channel can have heterogeneous implications among different types of workers: High skilled versus low skilled workers or business owners versus employees. The former may have higher returns by an increase in the aggregate demand.(Ampudia *et al.* (2018))

This is what is expected by exercising conventional monetary policy. However, unconventional monetary policy works differently. The recent massive Asset Purchase Programs (APP) by major central banks did not target short term interest rate but asset prices. These programs didn't affect consumption-saving decisions explicitly but changed the prices of financial assets and the valuation of financial portfolios. Studies form the European Central Bank and Bank of England find that the indirect effect dominates and the increase in employment and labour income prevails to loss of interest income for savers (see for example Lenza and Slacalek (2018) and Bunn *et al.* (2018)). Although the value of financial assets and portfolios rises in APPs benefiting more financial markets' participants, these studies find that the increase in employment and labour income in low income percentiles has a dominant effect and thus these policies have managed to reduce income inequality in the long run. For the eurozone, Ampudia et al. (2018) find tentative reduction in wealth inequality through the property channel, where home owners, especially the leveraged ones where benefited by a stimulation in house prices.

The recent OECD report (2018) warns that the latest financial crisis has exacerbated the concentration of wealth on the top part of the distribution in the UK. In this study we examine whether the quantitative monetary policy and the financial easing adopted during the crisis played a role. To our knowledge this is a first attempt to examine the dynamic relationship between monetary policy shocks and wealth inequality measures. Most studies using surveys are constrained by the very low frequency these surveys were conducted, limited number of waves, under-reporting in the high percentiles or measurement errors related to some forms of wealth such as business wealth. Even though we face similar problems, the five waves available in WAS allow us to construct monthly frequency wealth inequality measures and to study in higher detail the dynamic effects of conventional and unconventional monetary policy. We employ a Factor Augment Vector Autoregression (FAVAR) model to take advantage of a rich macroeconomic environment but also to account for measurement errors. Our main findings suggest that wealth inequality increases in the examined period in the UK and monetary policies play an important role to it. We find that the portfolio effect prevails to labour income effect during expansionary monetary policy shocks. In other words the right tail of the wealth distribution seems to be more benefited by higher returns in assets while the labour income effect which can be also expressed through higher employment in the low percentiles of the wealth distribution left tail of the distribution appears to be weaker.

The rest of the paper is structured as follows: Section 2 describes the variables used, their transformations and the construction of the inequality measures. Section 3 describes the estimation of the structural VAR model and the identification scheme. Section 4 presents the main results for earnings, income and consumption, while section 5 concludes.

2 Empirical Analysis

2.1 Empirical model

To estimate the impact of monetary policy shocks, we employ a Factor Augment Vector Autoregression (FAVAR) as our benchmark model. The observation equation of the model is defined as:

$$\begin{pmatrix} R_t \\ X_t \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & \Lambda \end{pmatrix} \begin{pmatrix} R_t \\ F_t \end{pmatrix} + \begin{pmatrix} 0 \\ v_t \end{pmatrix} \tag{1}$$

where R_t denotes the policy interest rate. X_t is $M \times 1$ matrix of variables for the UK covering both aggregate macroeconomic and financial data and the distribution of total wealth, net financial wealth and property wealth. $\underbrace{F_t}_{K\times 1}$ denotes a set K factors that summarise the information in X_t, Λ is a $M \times K$ matrix of factor loadings. Finally, v_t is a $M \times 1$ matrix that holds the idiosyncratic components. We assume that v_t follows an $AR(q)$ process:

$$v_{it} = \sum_{p=1}^{P} \rho_{ip} v_{it-p} + e_{it}, var(e_{it}) = r_i, R = diag\left([r_1, r_2, .., r_M]\right) \tag{2}$$

where $i = 1, 2, .., M$.

Denoting the factors $\begin{pmatrix} R_t \\ F_t \end{pmatrix}$ by the $N \times 1$ vector Y_t, the transition equation can be described as:

$$Y_t = BX_t + u_t \tag{3}$$

where $X_t = [Y'_{t-1}, .., Y'_{t-P}, 1]'$ is $(NP + 1) \times 1$ vector of regressors in each equation and B denotes the $N \times (NP + 1)$ matrix of coefficients $B = [B_1, ..., B_P, c]$. The covariance matrix of the reduced form residuals u_t can be written as:

$$\Sigma = (Aq)(Aq)' \tag{4}$$

where A is the lower triangular Cholesky decomposition of Σ, and q is an element of the family of orthogonal matrices of size N, satisfying $q'q = I_N$.

2.1.1 Identification of shocks

The structural shocks of the FAVAR model ε_t are defined as

$$\varepsilon_t = A_0^{-1} u_t, \varepsilon_t \sim \mathcal{N}(0, I_N) \tag{5}$$

where $A_0 = Aq$. The shock of interest is the first shock ε_{1t} in the $N \times 1$ vector of shocks $\varepsilon_t = [\varepsilon_{1t}, \varepsilon_{\cdot t}]$, where $\varepsilon_{\cdot t}$ contains the remaining $N - 1$ elements in ε_t. To identify the effect of ε_{1t}, we employ an instrument m_t described by the following equation:

$$m_t = \beta\varepsilon_{1t} + \sigma v_t, \quad v_t \sim \mathcal{N}(0, 1) \tag{6}$$

where $\mathbb{E}(v_t \varepsilon_t) = 0$. The instrument is assumed to relevant (i.e. $\mathbb{E}(v_t \varepsilon_{1t}) = \alpha \neq 0$) and exogenous (i.e. $\mathbb{E}(v_t \varepsilon_{\cdot t}) = 0$).

In our empirical application, the instrument to identify the monetary policy shock is taken from Gerko and Rey (2017) Gerko and Rey (2017) use high frequency data on short-sterling (SS) futures to construct a proxy for a monetary policy shock. In particular, Gerko and Rey (2017) consider changes in SS futures during a tight window around monetary policy events. They argue, that changes in SS futures around the release of the minutes of the monetary policy committee meetings contain information regarding the future stance of conventional and unconventional monetary policy and provide evidence that suggests that this measure is a strong instrument for the policy shock.

The structure of FAVAR model implies that the series in X_t are driven by aggregate shocks ε_t and idiosyncratic shocks e_{it}. When the survey-based wealth series in X_t are considered, our model captures the impact of aggregate shocks *net* of the effect of idiosyncratic disturbances that might proxy measurement error or differences in characteristics specific to the particular percentile group (see **?**).

2.1.2 Model estimation and specification

Following Bruns (2019) and Miescu and Mumtaz (2019), the FAVAR is estimated using a Gibbs sampling algorithm that is an extension of the algorithm proposed by Caldara and Herbst (2016) for proxy VARs. Details of the algorithm and the priors are presented in the technical appendix. As discussed in Caldara and Herbst (2016), the priors for β and σ^2 play an important role as the influence the reliability of the instrument. Mertens and Ravn (2013) define the reliability statistic as the squared correlation between m_t and ε_{1t}:

$$\rho^2 = \frac{\beta^2}{\beta^2 + \sigma^2} \tag{7}$$

In our benchmark model, the priors for β and σ^2 are set to reflect the strong belief that the instruments are relevant and imply that $\rho \approx 0.5$. This prior belief is based on the evidence regarding the high relevance of the instrument presented in Gerko and Rey (2017).

The choice of the number of factors is a key issue with regards to specification of the model. We follow the general approach used in **?**: i.e. the benchmark model is estimated using $K = 4$. We then show that the main results do not change substantially if the number of factors is increased. The lag length P is set to 6.

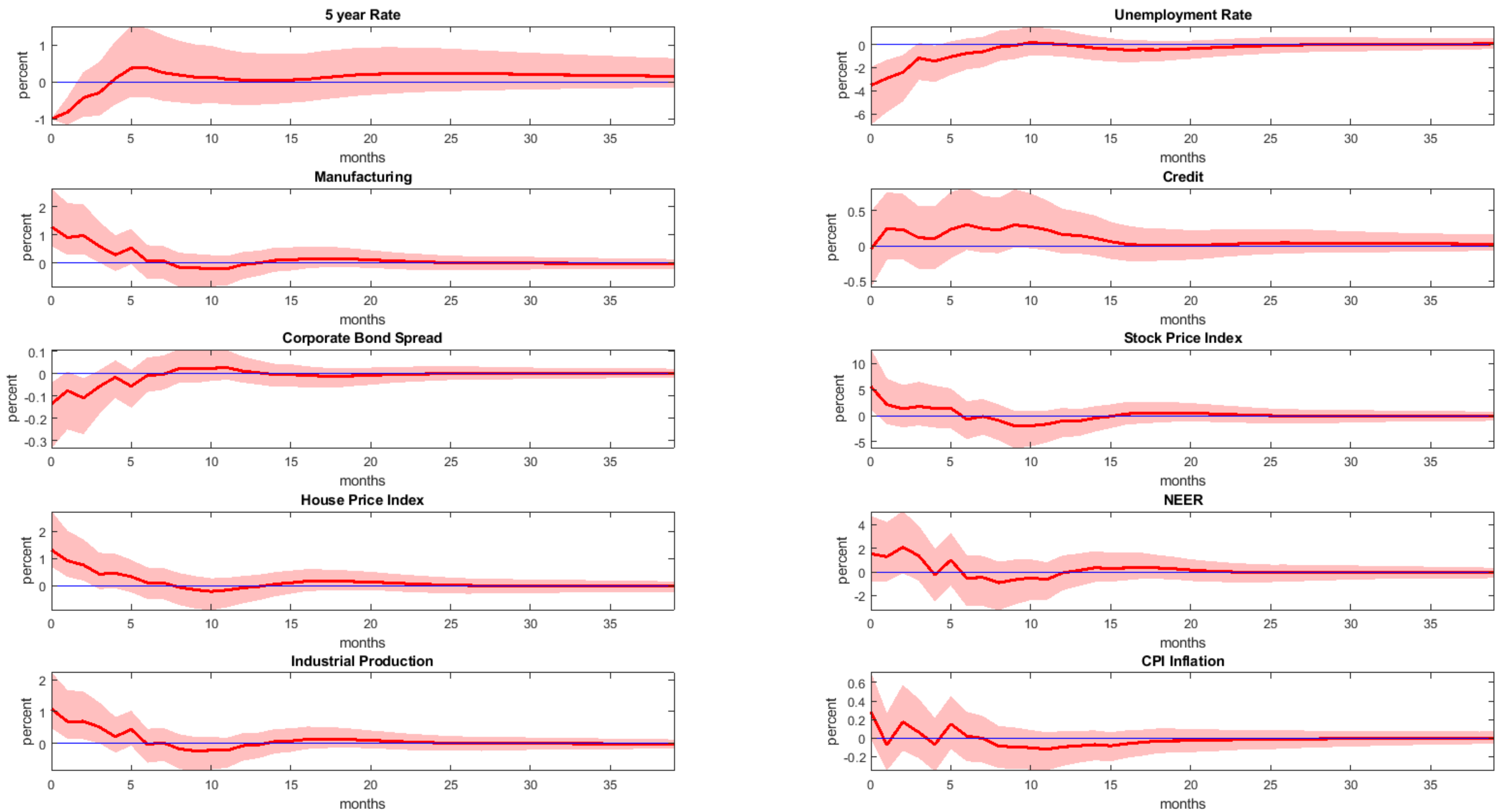

Figure 2: Impulse response of aggregate variables to a monetary policy shock

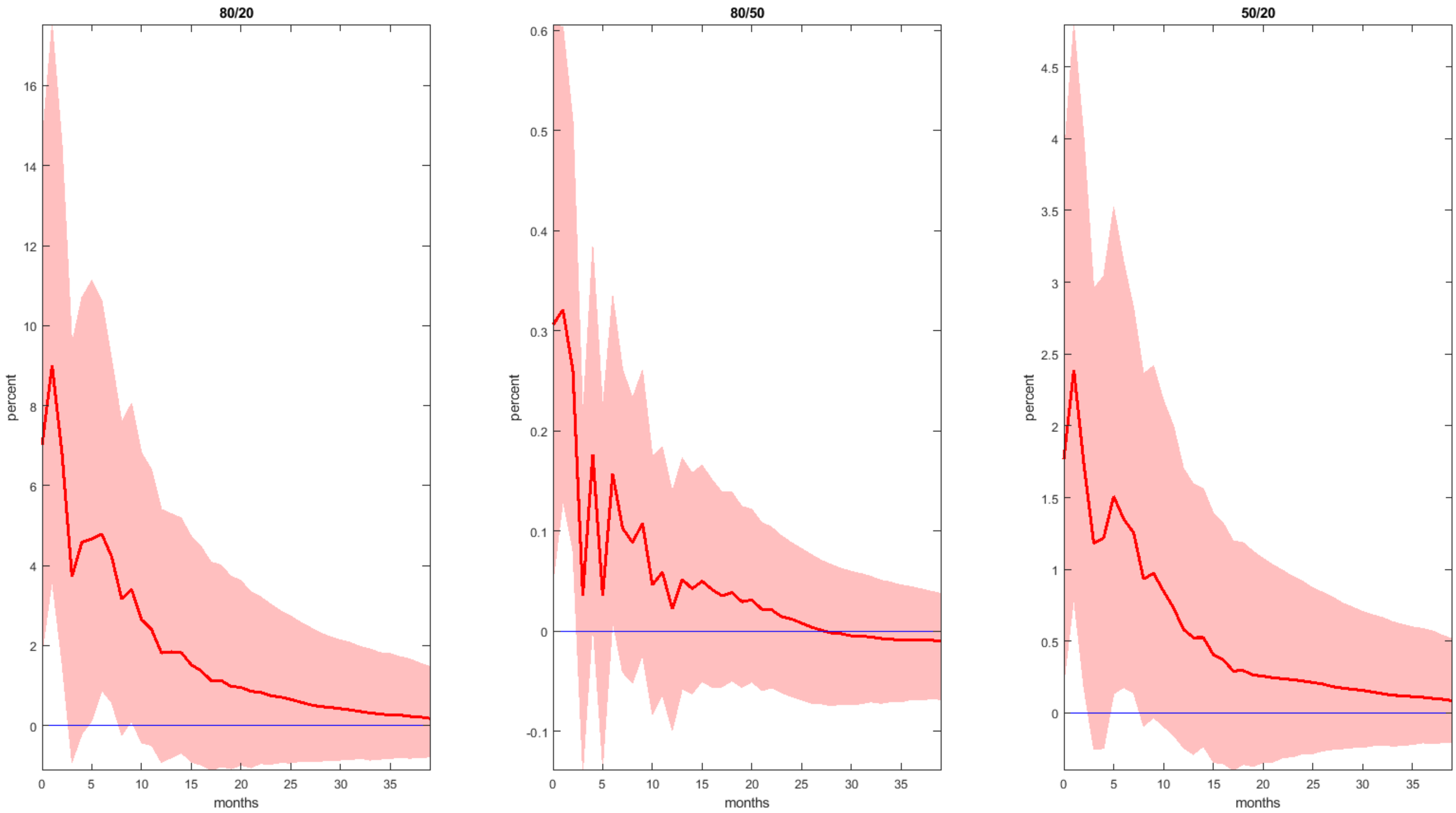

Figure 3: Response of total wealth

2.2 Response to a monetary expansion

Figure 2 shows the response of selected aggregate macroeconomic and financial variables to a monetary policy shock scaled to reduce the five year rate by 100 basis points. The monetary expansion leads to a boost in real economic activity with an increase in manufacturing and industrial production and a decrease in the unemployment rate. There is some indication that CPI inflation rises after the shock. As in Gerko and Rey (2017), the shock is associated with financial easing – the corporate bond spread declines and the stock index rises and the median response of credit is positive. However, unlike Gerko and Rey (2017), we do not find a large response of the exchange rate to the shock. This possibly reflects the smaller sample used in our study.

Figure 3 considers the response of the measures of inequality in total wealth. The left panel of the figure shows that after a monetary expansion wealth inequality increases. That is, the gap between the 80th percentile and the 20th percentile increases with the ratio rising by about 8 units. The second and third panels of the figure consider the source of the increase in inequality. The middle panel shows that the increase in inequality does not originate from above the median. While the difference between the 80th percentile and the 50th percentile increases, the magnitude of the increase is negligible. In contrast, as shown in the third panel, the 50/20 measure rise substantially. This suggests that wealth inequality is pulled up a rise in wealth in the middle of the distribution relative to the 20th percentile.

3 Conclusions

This paper considers the evolution of the weal distribution in the UK over the recent two decades. Using data from five waves of the Wealth and Asset Survey we show that wealth inequality was high during the 2006-2007 but declined subsequently. We then consider the impact of monetary policy shocks on the distribution of wealth. This exercise is carried out via a FAVAR where the monetary policy shock is identified via an external instrument. The estimated impulse responses suggest that a monetary expansion increases wealth inequality. The increase is driven by an increase in wealth at the median relative to the left tail of the distribution.

References

Ampudia, Miguel, Dimitris Georgarakos, Jiri Slacalek, Oreste Tristani, Philip Vermeulen and Giovanni L. Violante, 2018, Monetary policy and household inequality, *Working Paper Series 2170*, European Central Bank.

Balestra, Carlotta and Richard Tonkin, 2018, Inequalities in household wealth across OECD countries: Evidence from the OECD Wealth Distribution Database, *OECD Statistics Working Papers 2018/01*, OECD Publishing.

Banbura, Marta, Domenico Giannone and Lucrezia Reichlin, 2010, Large Bayesian vector auto regressions, *Journal of Applied Econometrics* **25**(1), 71–92.

Bruns, Martin, 2019, Combining Factor Models and External Instruments to Identify Uncertainty Shocks, *Mimeo 2017-9*, DIW.

Bunn, Philip, Alice Pugh and Chris Yeates, 2018, The distributional impact of monetary policy easing in the UK between 2008 and 2014, *Bank of England working papers 720*, Bank of England.

Caldara, Dario and Edward Herbst, 2016, Monetary Policy, Real Activity, and Credit Spreads: Evidence from Bayesian Proxy SVARs, *Finance and Economics Discussion Series 2016-049*, Board of Governors of the Federal Reserve System (U.S.).

Carter, C and P Kohn, 1994, On Gibbs sampling for state space models, *Biometrika* **81**, 541–53.

Crossley, Thomas F., Cormac O'Dea, Facundo Alvaredo, Anthony B. Atkinson and Salvatore Morelli, 2016, The Challenge of Measuring UK Wealth Inequality in the 2000s, *Fiscal Studies* **37**, 13–33.

Gerko, Elena and Hélène Rey, 2017, Monetary Policy in the Capitals of Capital, *Journal of the European Economic Association* **15**(4), 721–745.

Lenza, Michele and Jiri Slacalek, 2018, How does monetary policy affect income and wealth inequality? Evidence from quantitative easing in the euro area, *Working Paper Series 2190*, European Central Bank.

Mertens, Karel and Morten O. Ravn, 2013, The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States, *American Economic Review* **103**(4), 1212–47.

Miescu, Mirela and Haroon Mumtaz, 2019, The open economy impact of monetary shocks, *Mimeo 2017-9*, Queen Mary.

Piketty, Thomas, Emmanuel Saez and Gabriel Zucman, 2018, World Inequality Report 2018, *Technical report*.

4 Appendix: Proxy FAVAR model

The observation equation of the FAVAR model is defined as

$$\begin{pmatrix} Z_t \\ X_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \Lambda \end{pmatrix} \begin{pmatrix} Z_t \\ F_t \end{pmatrix} + \begin{pmatrix} 0 \\ v_t \end{pmatrix} \tag{8}$$

where Z_t denotes a set of 'observed' factors. X_t is a $M \times 1$ vector of variables, F_t denotes a $K \times 1$ matrix of unobserved factors while Λ is a $M \times K$ matrix of factor loadings. Finally, v_t is a $M \times 1$ matrix that holds the idiosyncratic components. We assume that v_t follows an $AR\,(q)$ process:

$$v_{it} = \sum_{p=1}^{P} \rho_{ip} v_{it-p} + e_{it}, var(e_{it}) = r_i, R = diag\left([r_1, r_2, .., r_M]\right) \tag{9}$$

where $i = 1, 2, .., M$.

Collecting the factors in the $N \times 1$ vector Y_t, the transition equation can be described as:

$$Y_t = BX_t + u_t \tag{10}$$

where $X_t = [Y'_{t-1}, .., Y'_{t-P}, 1]'$ is $(NP + 1) \times 1$ vector of regressors in each equation and B denotes the $N \times (NP + 1)$ matrix of coefficients $B = [B_1, ..., B_P, c]$. The covariance matrix of the reduced form residuals u_t can be written as:

$$\Sigma = (Aq)\,(Aq)' \tag{11}$$

where A is the lower triangular Cholesky decomposition of Σ, and q is an element of the family of orthogonal matrices of size N, satisfying $q'q = I_N$.

4.1 Identification of shocks

The structural shocks of the FAVAR model ε_t are defined as

$$\varepsilon_t = A_0^{-1} u_t, \varepsilon_t \sim \mathcal{N}\left(0, I_N\right) \tag{12}$$

where $A_0 = Aq$. The shock of interest is the first shock ε_{1t} in the $N \times 1$ vector of shocks $\varepsilon_t = [\varepsilon_{1t}, \varepsilon_{\cdot t}]$, where $\varepsilon_{\cdot t}$ contains the remaining $N - 1$ elements in ε_t. To identify the effect of ε_{1t}, we employ an instrument m_t described by the following equation:

$$m_t = \beta\varepsilon_{1t} + \sigma v_t, \quad v_t \sim \mathcal{N}\left(0, 1\right) \tag{13}$$

where $\mathbb{E}\left(v_t \varepsilon_t\right) = 0$.

4.2 Priors

We assume the following prior distributions:

1. We use a natural conjugate prior for the VAR parameters $b = vec(B'), \Sigma$ implemented via dummy observations (see Banbura *et al.* (2010)):

$$Y_{D,1} = \begin{pmatrix} \frac{diag(\gamma_1\sigma_1...\gamma_N\sigma_N)}{\tau} \\ 0_{N\times(P-1)\times N} \\ .... \\ diag(\sigma_1...\sigma_N) \\ .... \\ 0_{1\times N} \end{pmatrix}, and\ X_{D,1} = \begin{pmatrix} \frac{J_P\otimes diag(\sigma_1...\sigma_N)}{\tau} & 0_{NP\times 1} \\ 0_{N\times NP+1} & \\ .... & \\ 0_{1\times NP} & I_1\times c \end{pmatrix} \quad (14)$$

where γ_1 to γ_N denotes the prior mean for the coefficients on the first lag, τ is the tightness of the prior on the VAR coefficients, c is the tightness of the prior on the constant terms and N is the number of endogenous variables, i.e. the columns of Z_t. In our application, the prior means are chosen as the OLS estimates of the coefficients of an AR(1) regression estimated for each endogenous variable. We use principal component estimates of the factors F_t for this purpose. We set $\tau = 0.2$. The scaling factors σ_i are set using the standard deviation of the error terms from these preliminary AR(1) regressions. Finally we set $c = 1/1000$ in our implementation indicating a flat prior on the constant. We also introduce a prior on the sum of the lagged dependent variables by adding the following dummy observations:

$$Y_{D,2} = \frac{diag(\gamma_1\mu_1...\gamma_N\mu_N)}{\lambda},\ X_{D,2} = \begin{pmatrix} \frac{(1_{1\times P})\otimes diag(\gamma_1\mu_1...\gamma_N\mu_N)}{\lambda} & 0_{N\times 1} \end{pmatrix} \quad (15)$$

where μ_i denotes the sample means of the endogenous variables calculated using F_t.

2. The prior for the factor loadings Λ is normal $\mathcal{N}(\Lambda_{i0}, \Sigma_0)$ where Λ_0 is set to zero and Σ_0 is a diagonal matrix with diagonal elements equal to 100.

3. The prior for $\rho = [\rho_1, \rho_2, ..\rho_P]$ is normal $\mathcal{N}(\rho_0, \Sigma_{\rho 0})$ where $\rho_0 = 0$ and $\Sigma_{\rho 0}$ is an identity matrix. The prior for r_i is inverse Gamma $IG(T_0, D_0)$ where $T_0 = 1$ and $D_0 = 1e-5$.

4. We assume a normal prior for β: $\mathcal{N}(\underline{\beta}, \underline{V})$. The prior for σ^2 is inverse Gamma with mean σ_0 and standard deviation v_0.

5. The initial conditions for the factors are $\mathcal{N}(Y_{0|0}, P_{0|0})$ where $Y_{0|0}$ is set using the initial estimates of the factors obtained via principal components and $P_{0|0}$ is an identity matrix.

4.3 Gibbs algorithm

The Gibbs algorithm samples from the following conditional posterior distributions.

Step 1. $p(b|\Xi_{-b_t}, Y_{1:T}, m_{1:T})$. We write the model in state-space form:

$$\begin{aligned} \begin{pmatrix} Y_t \\ m_t \end{pmatrix} &= \begin{pmatrix} I_N\otimes X_t' \\ 0 \end{pmatrix} b_t + \begin{pmatrix} u_t \\ m_t \end{pmatrix} \text{ observation} \\ b_t &= b_{t-1} \text{ transition} \end{aligned}$$

where $b = vec(B')$. The covariance matrix of the observation equation residuals is:

$$cov\left(\begin{matrix} u_t \\ m_t \end{matrix} \middle| \Xi_{-b_t} \right) = \begin{pmatrix} AA' & Aq_1'\beta \\ \beta q_1 A' & \beta^2+\sigma^2 \end{pmatrix}$$

This system is conditionally linear and Gaussian. Thus the conditional posterior for b is normal: $\mathcal{N}(M, V)$ Following Carter and Kohn (1994), the mean and the variance of this distribution can be calculated by using the last recursion of the Kalman filter. The initial values of the Kalman filter are based on the priors set via dummy observations. Alternatively, one can employ a Metropolis step to sample b as described in Caldara and Herbst (2016).

Step 2. $p\left(\Sigma|\Xi_{-b_t}, Y_{1:T}, m_{1:T}\right)$. We follow Caldara and Herbst (2016) and use a Metropolis step to sample Σ.

(a) Draw a candidate Σ_{new} from the mixture proposal $Q(.) = \chi IW\left(s\left(\Sigma_{old}\right), t\right) + (1-\chi)\, IW\left(u_t' u_t, T\right)$ where $s\left(\Sigma_{old}\right)$ denotes the scale matrix consistent with the previous draw Σ_{old} and χ determines the weight given to each component. We choose χ and t to ensure a reasonable acceptance rate.

(b) Accept the draw with probability $\alpha = \min\left[\frac{\frac{P\left(Y_{1:t}|\Sigma_{new},\Xi_{-\Sigma}\right)P\left(m_{1:t}|Y_{1:t},\Sigma_{new},\Xi_{-\Sigma}\right)}{Q(\Sigma_{new})}}{\frac{P\left(Y_{1:t}|\Sigma_{old},\Xi_{-\Sigma}\right)P\left(m_{1:t}|Y_{1:t},\Sigma_{old},\Xi_{-\Sigma}\right)}{Q(\Sigma_{old})}}, 1\right]$. Here $P\left(Y_{1:t}|\Xi\right)$ denotes the likelihood of the VAR, while $P\left(m_{1:t}|Y_{1:t},\Xi\right) = \prod_{t=1}^{T} P\left(m_t|Y_t,\Xi\right)$ with $P\left(m_t|Y_t,\Xi\right) \sim \mathcal{N}\left(\beta q_1' A^{-1} u_t, \sigma^2\right)$.

Step 3. $p\left(q_1|\Xi_{-q_1}, Y_{1:T}, m_{1:T}\right)$. Following Caldara and Herbst (2016) we use a Metropolis step to sample q_1 :

(a) Draw a candidate from as $q_{1,new} = \frac{z}{\|z\|}$ where z is a $N \times 1$ vector from the $\mathcal{N}(0,1)$ distribution

(b) Accept the draw with probability $\alpha = \min\left[\frac{P\left(m_{1:t}|Y_{1:t},q_{1,new},\Xi_{-q_1}\right)}{P\left(m_{1:t}|Y_{1:t},q_{1,old},\Xi_{-q_1}\right)}, 1\right]$

Step 4 $p\left(\beta,\sigma|\Xi_{-[\beta,\sigma]}, Y_{1:T}, m_{1:T}\right)$. The structural shock of interest ε_{1t} can be calculated as $\varepsilon_{1t} = Aq_1u$. Conditional on $\Xi_{-[\beta,\sigma]}$ equation 13 is a standard linear regression, so specifying a conditional Normal-Gamma prior delivers a Normal-Gamma posterior. Particularly, we first draw $p\left(\sigma^2|\Xi_{-[\beta,\sigma]}, Y_{1:T}, m_{1:T}\right)$. Assuming an inverse-Gamma prior, this conditional posterior is also inverse-Gamma. As the prior is parameterised in terms of mean σ_0 and standard deviation v_0, it is convenient to draw the precision $\frac{1}{\sigma^2}$ using Gamma distribution. Note that $\frac{1}{\sigma^2} \sim \mathcal{G}(a,b)$ where $a = \frac{\nu_1}{2}$, $b = \frac{2}{s_1}$. The parameters of this Gamma density are given by $\nu_1 = \nu_0 + T$ and $s_1 = s_0 + \hat{v}_t'\hat{v}_t$ where $\hat{v}_t = m_t - \beta e_{1t}$. s_0 can be calculated as $2\sigma_0\left(1 + \frac{\sigma_0^2}{v_0^2}\right)$ while $\nu_0 = 2\left(2 + \frac{\sigma_0^2}{v_0^2}\right)$. Moreover, assuming a prior for $\beta|\sigma^2, \Xi_{-[\beta,\sigma]} \sim \mathcal{N}(\underline{\beta}, \underline{V}^{-1})$, the posterior is also conditional Normal $p\left(\beta|\Xi_{-[\beta,\sigma]}, \sigma, Y_{1:T}, m_{1:T}\right) \sim \mathcal{N}(\tilde{\beta}, \tilde{V}^{-1})$, where $\tilde{\beta} = \tilde{V}^{-1}\left[\sum_{t=1}^{T} m_t\varepsilon_{1t} + \underline{V}\underline{\beta}\right]$ and $\tilde{V} = \underline{V} + \frac{1}{\sigma^2}\sum_{t=1}^{T}\varepsilon_{1t}^2$.

Step 5 $p\left(m_{-t}|\Xi_{m_{-t}}, Y_{1:T}, m_{1:T}\right)$. (Step is only implemented if observations in m_t are missing; we denote missing observations as m_{-t}). We treat the missing data for the instrument as an unobserved state and write the problem in state-space form. For periods when m_t is unobserved, the observation equation is given by

$$\underbrace{\begin{pmatrix} Y_t \\ m_{-t} \end{pmatrix}}_{M\times 1} = \underbrace{\begin{pmatrix} I_{N\times NS} \\ 0_{1\times NS} \end{pmatrix}}_{M\times NS} \underbrace{\tilde{\beta}_t}_{NS\times 1} + \underbrace{V_t}_{M\times 1}$$

where $M = N + 1$, $\tilde{\beta}_t = \underbrace{\begin{pmatrix} Y_t \\ \hat{m}_t \\ Y_{t-1} \\ \hat{m}_{t-1} \\ . \\ . \\ Y_{t-P+1} \\ \hat{m}_{t-P+1} \end{pmatrix}}_{NS\times 1}$, $\hat{m}_t$ is the estimate of m_t, I denotes an identity matrix, 0 denotes a matrix of zeros, and NS denotes the rows of the state vector, i.e. the number of states. The error term V_t is zero for the equations corresponding to the observed data Y_t. In contrast, the element of V_t corresponding to m_{-t} is assumed to have a very large number.

When data on m_t is observed, the observation equation changes to:

$$\begin{pmatrix} Y_t \\ m_t \end{pmatrix} = (I_{M\times NS})\,\tilde{\beta}_t$$

The transition equation is:

$$\tilde{\beta}_t = \tilde{\mu}_t + \tilde{F}_t\tilde{\beta}_{t-1} + w_t,\ var(w_t) = \breve{Q}_t$$

To describe the matrices $\tilde{\mu}_t, \tilde{F}_t, \breve{Q}_t$ consider the following representation of the VAR model:

$$z_t = \zeta_t x_t + \kappa_t$$

where $z_t = \begin{pmatrix} Y_t \\ \hat{m}_t \end{pmatrix}$, x_t collects P lags of z_t, $\kappa_t = \begin{pmatrix} u_t \\ \hat{m}_t \end{pmatrix}$ and ζ_t denotes the matrix of time-varying coefficients B_t augmented with zeros in the equation with m_t as the dependent variable. The constants and autoregressive coefficients of this VAR in companion form are denoted by $\tilde{\mu}_t$ and $\tilde{F}_t$. Similarly, $\breve{Q}_t$ denotes the error covariance $var\begin{pmatrix} u_t \\ \hat{m}_t \end{pmatrix}$ in companion form. With the model in state space form, the Carter and Kohn (1994) algorithm is then used to draw $\tilde{\beta}_t$.

Step 6 $H(\Lambda|\Xi_{-\Lambda}, Y_{1:T}, m_{1:T})$. Given the factors F_t, the observation equation is set of M independent linear regressions with serial correlation

$$X_{it} = F_t\Lambda_i' + v_{it}$$

where Λ_i denotes the ith row of the factor loading matrix. The serial correlation can be dealt with via a GLS transformation of the variables:

$$\tilde{X}_{it} = \tilde{F}_t\Lambda_i' + e_{it}$$

where $\tilde{X}_{it} = X_{it} - \sum_{p=1}^{P}\rho_p X_{it-p}$ and $\tilde{F}_{kt} = F_{kt} - \sum_{p=1}^{P}\rho_p F_{kt-p}$. The conditional posterior is normal $\mathcal{N}(M, V)$:

$$\begin{aligned} V &= \left(\Sigma_0^{-1} + \frac{1}{r_i}\tilde{F}_t'\tilde{F}_t\right)^{-1} \\ M &= V\left(\Sigma_0^{-1}\Lambda_{i0} + \frac{1}{r_i}\tilde{F}_t'\tilde{X}_{it}\right) \end{aligned}$$

To account for rotational indeterminacy the top $K\times K$ block of Λ is set to an identity matrix.

1. $H(r_i|\Xi_{-r}, Y_{1:T}, m_{1:T})$. The conditional posterior for r_i is $IG(T_0 + T, e_{it}'e_{it} + D_0)$ where T is the sample size.

2. $H(\rho|\Xi_{-r}, Y_{1:T}, m_{1:T})$. Given a draw of the factors, the AR coefficients are drawn for each i independently. The conditional posterior is normal $\mathcal{N}(m, v)$

$$\begin{aligned} v &= \left(\Sigma_{\rho 0}^{-1} + \frac{1}{r_i}x_{it}'x_{it}\right)^{-1} \\ m &= V\left(\Sigma_{\rho 0}^{-1}\rho_0 + \frac{1}{r_i}x_{it}'y_{it}\right) \end{aligned}$$

   where $y_{it} = v_{it}$ and $x_{it} = [v_{it-1}, .., v_{it-P}]$

3. $H(F_t|\Xi_{-F_t}, Y_{1:T}, m_{1:T})$. To draw the factors, we write the model in state-space form taking into account the covariance between m_t and u_t and the serial correlation in the idiosyncratic components.

The observation equation is defined as:

$$\underbrace{\begin{pmatrix} Z_t \\ \tilde{X}_t \\ m_t \end{pmatrix}}_{x_t} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 \\ 0 & \Lambda & 0 & , 0 & \tilde{\Lambda}_1 & 0 & , \cdots, & 0 & \tilde{\Lambda}_P & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & & 0 & 0 & 0 \end{pmatrix}}_{H} \underbrace{\begin{pmatrix} Z_t \\ F_t \\ m_t \\ \cdot \\ \cdot \\ \cdot \\ Z_{t-P} \\ F_{t-P} \\ m_{t-P} \end{pmatrix}}_{f_t} + \underbrace{\begin{pmatrix} 0 \\ v_t \\ 0 \end{pmatrix}}_{V_t}$$

where $\tilde{X}_t = \begin{pmatrix} X_{1t} - \sum_{p=1}^{P} \rho_{1p} X_{1t-p} \\ \cdot \\ \cdot \\ X_{Mt} - \sum_{p=1}^{P} \rho_{Mp} X_{Mt-p} \end{pmatrix}$. The blocks of the H matrix contain the factor loadings multiplied by the negative of the corresponding serial correlation coefficient. For example $\tilde{\Lambda}_1 = \begin{pmatrix} -\Lambda_1 \rho_{11} \\ \cdot \\ \cdot \\ -\Lambda_M \rho_{M1} \end{pmatrix}$ where Λ_i denotes the factor loadings for the ith variable X_{it}. Finally, the variance of V_t is $R = diag\left([0, r_1, .., r_M, 0]\right)$. The transition equation is defined as

$$f_t = \mu + \tilde{B} f_{t-1} + U_t$$

where $\tilde{B} = \begin{pmatrix} B_1 & 0_{N\times 1} & \cdot & \cdot & B_P & 0_{N\times 1} & 0_{N\times N+1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ I_{P(N+1)\times(P+1)(N+1)} & & & & & & \end{pmatrix}$, $\mu = \begin{pmatrix} c \\ 0 \\ 0_{P(N+1)} \end{pmatrix}$, $U_t = \begin{pmatrix} u_t \\ m_t \\ 0_{P(N+1)} \end{pmatrix}$. The non-zero block of $cov\left(U_t\right)$ is given by $\begin{pmatrix} AA' & Aq_1'\beta \\ \beta q_1 A' & \beta^2 + \sigma^2 \end{pmatrix}$. In other words, the structure of the transition equation accounts for the relationship between the instrument and the reduced form residuals. Given this Gaussian linear state-space, the state vector can be drawn from the normal distribution using the Carter and Kohn (1994) algorithm.