UNIVERSITY of York

Discussion Papers in Economics

No. 17/07

Housing and Financial Asset Allocations of Heterogeneous Homeowners

Zhechun He

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Housing and Financial Asset Allocations of Heterogeneous Homeowners*

Zhechun He†
University of York

July 4, 2017

Abstract

Market constraints (e.g. borrowing constraint, no-short-selling constraint) are important for household portfolio choices especially for housing. A structural model naturally generates alternative portfolio regimes with different binding constraints. But empirically we cannot observe which households are constrained in safe, risky or housing finance and housing. We use a semiparametric approach on data from Wealth and Asset Survey (WAS) to determine this. We find distinct patterns of housing and financial assets allocation among homeowners by fitting a multivariate Gaussian mixture model via a censored data expectation-maximisation (EM) algorithm. Estimation results reveal that on average about 80% of the households are no-short-selling constrained in risky asset investment and with low net worth. Among other things, we find that households who are younger, less educated with lower income are more likely to be no-short-selling constrained in risky asset investment and with lower net worth. Our predicted regime classification is aligned to those of the structural model.

Key words: household finance, borrowing constraint, no-short-selling constraint, Gaussian mixture model, censored data EM algorithm, semi-parametric

1 Introduction

Recognition of the theoretical and empirical importance of market constraints (especially borrowing and no-short-selling constraints, on household portfolio choices has increased references (Attanasio et al., 2012). The role of housing and housing finance in household portfolios is of special importance given the relative size, illiquidity and transaction costs involved (Campbell, 2006). The study of household finance is challenging because household behaviours are difficult to measure and complicated to model. Compared to corporate finance, household finance has some special features such as planning over long but finite horizons, having nontradable human capital and illiquid housing assets, and facing borrowing constraints. There are some

*I would like to thank my PhD supervisors, Prof. Peter Simmons and Prof. Jo Swaffield, and my thesis committee member Prof. Cheti Nicoletti for their valuable comments and suggestions. I am also grateful to participants at the Thursday workshop in University of York, Royal Economic Society PhD meeting, Scottish Economic Society Annual Conference, and China Meeting of the Econometric Society for their helpful discussion. Any errors are my sole responsibility.

†Department of Economics and Related Studies, University of York, UK. Email: zch501@york.ac.uk.

empirical literatures based on the framework of Merton (1973) where agents plan for the long term with time-varying investment opportunities (Campbell et al., 2003; Kim and Omberg, 1996), which emphasised the distinction between real and nominal returns in the long horizon models. But the Merton framework assumes wealth is liquid and tradable, which is in contradiction with nontradable human capital and illiquid housing. Moreover, as the biggest component of wealth for most households, human capital is nontradable and because much of the labour income risk is idiosyncratic it is unhedgeable (Campbell, 2006). It represents a background risk which could make households more risk averse and invest more cautiously in other risky assets if the correlation between returns on these assets and labour income is positive (Heaton and Lucas, 2000). Or conversely if the correlation is negative. In addition, as an important largely indivisible asset for homeowners, illiquid housing may discourage investment in risky assets by homeowners leading to a crowding-out effect (Cocco, 2005). Adding borrowing constraints makes it even more complicated especially if it is not possible to exactly observe such constraints so that in a household survey we just do not know a priori if a household is constrained or not. Portfolio decisions (and consumption decisions) will generally differ between households which are borrowing constrained or unconstrained. The estimation problem is that from the data we often cannot directly see who is constrained. External evidence suggests that borrowing constraints are typically more important for younger households who have not accumulated sufficient savings and have little or no housing wealth as collateral. Therefore there are some life cycle effects in financial strategies as households age and accumulate wealth.

Existing literature considers the complications of household finance in different ways. One branch of the literature derives numerical solutions to the housing and portfolio decisions in a life cycle framework with such constraints by calibration and simulation (Attanasio et al., 2012; Cocco, 2005; Yao and Zhang, 2005). However, the calibration of state variables is based on some dataset as a whole without considering possible heterogeneity among different groups of household. Here the calibration includes initial wealth as well as the parameters of stochastic processes (e.g. income process, house price process, risky asset return process)(Carroll, 2012). A second branch of the literature estimates the structural parameters (preference parameters) using Euler equations from a theoretical model with and without liquidity constraints (Zeldes, 1989; Whited and Wu, 2006). A third branch of the literature applies reduced form models to find empirical evidence about the impact of individual characteristics (e.g. financial illiteracy (Rooij et al., 2011) and income hedging motives (Bonaparte, et al., 2014)) on household portfolio choice.

A common feature of most of the literature stated above is the modelling of a typical household. That is, they analyse the average behaviour of the population. An exception is Zeldes (1989). Zeldes (1989) a priori selects a set of families that he believes to be not liquidity constrained in terms of wealth to income ratio. In his terms this subsample has all interior solutions to consumption, thus the Euler equation holds as an equality and the preference parameters can be estimated from the Euler equation. However, his analysis strongly relies on the ad hoc criterion used to split the sample into constrained and unconstrained groups: households with wealth to income ratio above a certain threshold are not liquidity constrained and vice versa. In comparison, our paper does not assume any certain criterion to split the sample. Instead, we try to see if the data on the multiple assets holdings give probabilistic splits of the sample and then explore the behaviour of households derived from the probabilistic split. Another exception that considers the heterogeneous intrinsic nature of subsamples is King and Leape (1998), who estimate the joint discrete

and continuous choice of household portfolios by a switching regression model. In their model, both the discrete choice of owning particular combinations of assets and the continuous choice of asset demand system conditional on ownership are parametrised by a set of household characteristics. Besides the different focuses of research, there are two main differences between their model and ours. First, our model studies the asset allocation behaviours at both the extensive and intensive margins at the same time via a censored data EM algorithm, while their model studies the extensive and intensive margins in two steps. Second, as opposed to their fully parametric model, our paper is only semi-parametric in the sense that the classification of households in terms of asset allocations is unconditional, which has the advantage of being more flexible and circumventing possible endogeneity brought by covariates.

Specifically, this paper aims to investigate housing and financial asset allocations decisions (hereafter, asset allocations) by heterogeneous homeowners with a flexible model motivated by economic theory. We find distinct patterns of unconditional housing and financial assets allocation among homeowners by fitting a multivariate Gaussian mixture model via a censored data expectation-maximisation (EM) algorithm. Considering the choices of different assets are made simultaneously, the Gaussian mixture model we fit has a multivariate nature. The existence of no-short-selling constraint on risky asset is considered by the use of the censored data EM algorithm. The assumptions in our mixture model are minimal in the sense that we only assume a multinomial distribution for the component membership indicators and a mixture of multivariate normal distributions for housing and two other assets although we estimate the unknown component density parameters. That is, neither mixing weights nor the mean of each asset is parametrised. This allows much flexibility for the data to talk by avoiding possible spurious inclusion of covariates and subsequent endogeneity bias. The choice of the number of components is based on the economic intuition from a theoretical model presented in Section 2. After finding the chances of a household being in different regimes (mixture components) we want to understand which households are assigned to which mixture components (regimes) and how this aligns with the theoretical regimes. Descriptive statistics are presented to describe the features of each group. Then a linear probability regression is implemented to find the determinants of group membership. The results are encouraging: we use the number of components/regimes suggested by the theory and find quite strong sample separation into these, the no-short-selling constraint on risky financial assets clearly binds in the poorer lower two components but is slack in the two richer components. Within the components where risky financial asset constraints either do or do not bind, there is evidence that the subdivision into two further components (somewhat weakly) supports households who are mortgage borrowing constrained or unconstrained. Our model is within the broader field of latent class models[1] where the discrete and finite latent variables in our model can be interpreted as no-short-selling constraint, mortgage borrowing constraint, heterogeneity in initial wealth and preferences (e.g. different marginal utility, expectation, risk aversion, etc.), and household idiosyncratic shocks.

Our empirical results contribute to the literature on structural models on household finance by giving a possibly more sensible calibration of state variables for grid search in different structural regimes. On the other hand, it sheds light on empirical work that tries to estimate heterogeneous household behaviours by giving a basis for classifying observations into different latent classes. This paper also makes a preliminary

[1] Latent class models are also referred to as unsupervised learning in the field of machine learning.

attempt to understand the household characteristics that may be required to apply the structural model to capture the observed differences in portfolio allocation among households. The rest of the paper is organised as follows. Section 2 shows the theoretical motivation of this paper. Section 3 describes the Wealth and Asset Survey (WAS) data we use. Section 4 presents the econometric model, estimation method and algorithm we use to estimate the model. Section 5 reports the estimation results. Finally, Section 6 concludes.

2 Theoretical motivation

2.1 A model for stable homeowners

Considering the infrequency of housing purchase observed from the data (Section 3), we focus on the behaviours of stable homeowners who own their main residence only[2] and don't move during the timespan we study. For these stable homeowners, housing consumption is constant across time. For this reason the tenure choice of housing (whether to be a renter or homeowner) and the decision to upsize or downsize the house are beyond the scope of this paper. We formulate the model as follows.

Families are treated as forward looking. In period t, they face uncertainty in general about future income I_{t+1}, house prices p_{t+1}, return on risky asset $r_{f,t+1}$. $I_{t+1}, r_{f,t+1}, p_{t+1}$ are random and only realised at the start of the next period $t+1$ On the other hand, the return on safe asset $r_{a,t+1}$ is time-varying but non-random and known by the households[3]. Family utility in period t depends on a composite consumption c_t and utility derived from their present housing quantity H_t.Figure 1 shows the timeline of the model.

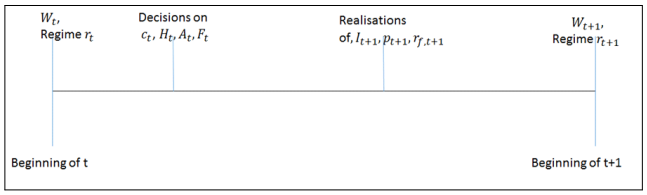

Figure 1: Timeline of the model

Suppose families live for T periods. For $t \leq T$, families have a time additive expected utility life cycle objective

$$\Sigma_t \beta^t E_t U_t(c_t, H_t)$$

where $U_t(.)$ is the per-period utility function at time t, β is the constant rate of time preference, the expectations operator E_t is taken when any of future house prices, asset returns and income flows are uncertain.

[2] We exclude households owning buy-to-let properties. Therefore households cannot get any rental income from the houses they own.

[3] We assume households have rational expectations.

Families can access financial markets. There are three: there is a safe asset with a known one period return on asset r_{at}. The holding of the safe asset at t is X_{st}. There is a risky asset with an return $r_{f,t+1}$ that is only realised at the end of period t (at the beginning of period $t+1$) after the investment in period t is made. Holdings of equities in t are $F_t \geq 0$ since borrowing in equity is infeasible (short selling is not allowed). There is also a housing mortgage market with an interest rate $r_{m,t+1}$ realised after the mortgage of period t has been taken. There are two borrowing constraints associated with the mortgage: the loan-to-value ratio constraint and the loan-to-income ratio constraints:

$$M_t \leq min[\tau_1 p_t H_t, \tau_2 I_t]$$

For simplicity, now suppose we can assume $r_{mt} = r_{at}$ (Attanasio et al., 2012) and borrowing in the safe asset is only possible via a mortgage. Define net safe assets $A_t = X_{st} - M_t$ where $X_{st} \geq 0, M_t \leq min[\tau_1 p_{bt} H_t, \tau_2 I_t]$ which means that $A_t \geq \max[-\tau_1 p_{bt} H_t, -\tau_2 I_t]$[4].

We assume a composite consumption good c_t with the price normalised to 1 in each period t. As a result, in each period, monetary variables including return on assets, labour income, rent and house prices are expressed as a ratio of the consumption price. In other words, every monetary variable is in real term.

We write the household value function in period t in recursive form as:

$$V_t(W_t) = \max_{\{A_t, F_t\}} U_t(c_t, H_t) + \beta E_t V_{t+1}(W_{t+1})$$

subject to

$$\begin{aligned} W_t &= c_t + A_t + F_t + p_t H_t \\ W_{t+1} &= I_{t+1} + (1 + r_{a,t+1}) A_t + (1 + r_{f,t+1}) F_t + p_{t+1} H_t \\ F_t &\geq 0 \\ A_t &\geq \max[-\tau_1 p_t H_t, -\tau_2 I_t] \end{aligned}$$

Notice the stable homeowners are "locked" in housing consumption in the sense that they are making decisions as if housing consumption H_t is not a choice variable for them. In other words, their decisions for asset allocations are conditional on their unchanged housing consumption $H_t = \overline{H}$. On the other hand, the values of their total housing wealth $p_t H_t$ could well change through time due to the change of house prices p_t.

Forming the KuhnTucker Lagrangian

$$\begin{aligned} L &= U_t[(W_t - A_t - F_t - p_t H_t), H_t] + \\ &\quad \beta E_t V_{t+1}[I_{t+1} + (1 + r_{a,t+1}) A_t + (1 + r_{ft+1}) F_t + p_{t+1} H_t] \\ &\quad + \lambda_{1t}(A_t + \tau_1 p_t H_t) + \lambda_{2t}(A_t + \tau_2 I_t) + \lambda_{3t} F_t \end{aligned}$$

[4] $A_t \geq \max[-\tau_1 p_{bt} H_t, -\tau_2 I_t]$ is the necessary but not sufficient condition of $M_t \leq min[\tau_1 p_{bt} H_t, \tau_2 I_t]$.

Envelope Theorem gives

$$\begin{aligned} V_t'(W_t) &= \frac{\partial U_t}{\partial c_t} \\ V_{t+1}'(W_{t+1}) &= \frac{\partial U_{t+1}}{\partial c_{t+1}} \end{aligned}$$

First order conditions are

$$F_t : -\frac{\partial U_t}{\partial c_t} + \beta E_t(1 + r_{ft+1})\frac{\partial U_{t+1}}{\partial c_{t+1}} + \lambda_{3t} = 0 \tag{1}$$

$$A_t : -\frac{\partial U_t}{\partial c_t} + \beta(1 + r_{a,t+1})E_t\frac{\partial U_{t+1}}{\partial c_{t+1}} + \lambda_{1t} + \lambda_{2t} = 0 \tag{2}$$

2.1.1 Evolution and cross-sectional variation of assets allocations

For notational convenience, hereafter we denote the real value of housing $p_t H_i$ for household i at period t as G_{it}:

$$G_{it} = p_t H_{it}$$

G_{it} is H_{it} scaled by the house price p_t which is assumed to be the same for every household in each period (Law of one price). This assumption is made along with the usual assumption that housing quantity H_i not only represents the physical size of the house but also the quality of the house[5]. In other words, p_t is just a universal conversion factor to translate H_i into the observable monetary value G_{it}.

Notice that the argument of the value function $V_{t+1}(W_{t+1})$ is

$$\begin{aligned} W_{t+1} &= I_{t+1} + (1 + r_{a,t+1})A_t + (1 + r_{f,t+1})F_t + p_{t+1}H_t \\ &= I_{t+1} + (1 + r_{a,t+1})A_t + (1 + r_{f,t+1})F_t + \frac{p_{t+1}}{p_t}G_{it} \end{aligned}$$

which suggests that the direct effect of the current period t on the next period value function V_{t+1} is not from c_t or W_t, but from (G_{it}, A_t, F_t) (Carroll, 2012). That is, for each household i, the vector (G_{it}, A_{it}, F_{it}) is a sufficient statistic which captures all the information from the current period t that is needed to solve the intertemporal maximisation problem in future periods. Thus the evolution of (G_{it}, A_{it}, F_{it})[6] may reflect a combination of the changes in planning horizon[7], updates of random state variables $I_{i,t+1}$, $r_{f,t+1}$, p_{t+1},changes of expectations about the future I_{t+1}, $r_{f,t+1}$, p_{t+1}[8] and even preference parameters such as risk aversion and subsistence level of consumption[9].

[5] Here the quality of the house includes the location etc.

[6] The evolution of (G_{it}, A_{it}, F_{it}) for stable homeowners where $G_{it} = p_t H_{it}$ can be seen as the evolution of (A_{it}, F_{it}) with $H_{it} = \overline{H_i}$ being time-invariant. However, we can also view H_{it} as endogenous and rational and the optimal choice is not to change H_{it} in the timespan (3 waves) we consider here.

[7] In a finite horizon setting, the policy functions change through time. But in our data which only covers 3 waves (6 years), this effect for a household should be less important.

[8] This might be due to the change of the general macro economic environment.

[9] The change of subsistence level of consumption may happen when there are new children born in the family or when children grow up and leave the family.

Households have different (G_{it}, A_{it}, F_{it}) because of household-specific (τ_{1i}, τ_{2i}) associated with the borrowing constraints, individual idiosyncratic income process, different initial wealth W_{it}, different ages and hence different planning horizon and different preference parameters.

Putting this theoretical setting into a statistical framework, given the timespan in our data is short (only 3 waves, i.e., 6 years), it means the cross sectional variation of (G_{it}, A_{it}, F_{it}) at each period t is likely to be more significant than the time series variation for each household. This is why we analyse the data by wave later in this paper[10].

2.1.2 Possible asset allocation regimes

From equations (1) and (2), we can see that there are four possible solution regimes in which different sets of constraints bind or are slack for stable homeowners (Table 1). For a particular household, the graph of regimes are shown in Figure 2.

It is tempting to identify the signs of λ_1 and λ_2 (Lagrangian multipliers associated with loan-to-value ratio constraint and loan-to-income ratio constraint) so that we can divide the sample into regimes and use proper econometric specification. However, if we just rely on the first order conditions from the theoretical model, then it is hard to distinguish the error terms of the moment conditions and λ_1 and/or λ_2 unless by making very strong parametric assumptions for λ_1 and λ_2 (Whited and Wu, 2006) and assuming that λ_1 and λ_2 are independent of the errors of the moment conditions. In this paper, instead of trying to identify the signs of λ_1 and λ_2,we use Taylor approximation to obtain non-parametric solution for each regime and try to identify the corresponding joint distribution of (G_i, A_i, F_i) for each regime, allowing for household heterogeneity which explain the cross-sectional variations of assets allocations across households. Without loss of generality, let the decision rules for F_i and A_i in Regime r for household i at each period be[11].

$$A_i = A_r(\boldsymbol{\varepsilon}_i)$$

$$F_i = F_r(\boldsymbol{\varepsilon}_i)$$

where $\boldsymbol{\varepsilon}_i = [\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i}]$, ε_{1i} is heterogeneity in initial wealth, ε_{2i} is heterogeneity in preferences, ε_{3i} is household idiosyncratic shocks.

One challenge to link the empirical model with the theoretical model is that τ_1, τ_2 may be partially individual specific[12] and are only partially observable to the econometrician. In other words, looking at the data, we have no exact idea whether the borrowing constraints (either one of the loan-to-value and loan-to-income constraints or both) are binding. We can see from Figure 4 (Section 3) that there is some boundary (upper limit) for the loan-to-value ratio. And in reality, choice of loan value affected by the loan-to-value constraint may affect the loan-to-income constraint and such restrictions are imposed differently by different lenders[13]. This complicated relationship between loan-to-value and loan-to-income constraints contributes to

[10] There is also the consideration of meeting the assumption that datapoints are i.i.d. for the Gaussian mixture model.

[11] The time subscript t is omitted here since we consider the cross-sectional variation in the same period.

[12] The values of τ_1 and τ_2 depend on occupation, income, age, credit record, etc., but the information in the dataset is not sufficient to reflect exact τ_1, τ_2.

[13] One example is in 2015, the loan-to-income restriction imposed by Barclays is initially 4.5 for all, later relaxed to 4.5 if loan value <£300000 and 5 if >£30000. (Bank of England, July 2015, Financial Stability Report)

the unknown nature of borrowing constraint faced by heterogenous households in the sense that τ_2 is not only unobservable but also "endogenous" depending on personal preference for lenders and choice of mortgage. For this reason we change the borrowing constraint in the optimisation problem as

$$A_i \geq B_r(\boldsymbol{\varepsilon}_i)$$

where B_r is a function of household heterogeneity and shocks whose functional form is regime specific[14].

Assume the expectation over observations who belong to regime r $E_r(\boldsymbol{\varepsilon}_i) = 0$. Taking a first order Taylor expansion around $E_r(\boldsymbol{\varepsilon}_i) = 0$ yields

$$\begin{aligned} A_i|\text{Regime r} &\sim A_r(E_r(\boldsymbol{\varepsilon}_i)) + [\boldsymbol{\varepsilon}_i - E_r(\boldsymbol{\varepsilon}_i)]\nabla A_r(E_r(\boldsymbol{\varepsilon}_i)) \\ &= A_r(W_i, E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla A_r(E_r(\boldsymbol{\varepsilon}_i)) \end{aligned}$$

$$\begin{aligned} F_i|\text{Regime r} &\sim F_r(E_r(\boldsymbol{\varepsilon}_i)) + [\boldsymbol{\varepsilon}_i - E_r(\boldsymbol{\varepsilon}_i)]\nabla F_r(E_r(\boldsymbol{\varepsilon}_i)) \\ &= F_r(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_r(E_r(\boldsymbol{\varepsilon}_i)) \end{aligned}$$

where $\nabla A_r(E_r(\boldsymbol{\varepsilon}_i)) = \begin{bmatrix} \frac{\partial A_r}{\partial \varepsilon_1}|_{\varepsilon_1 = E_r(\varepsilon_1)} \\ \frac{\partial A_r}{\partial \varepsilon_2}|_{\varepsilon_2 = E_r(\varepsilon_2)} \\ \frac{\partial A_r}{\partial \varepsilon_3}|_{\varepsilon_3 = E_r(\varepsilon_3)} \end{bmatrix}, \nabla F_r(E_r(\boldsymbol{\varepsilon}_i)) = \begin{bmatrix} \frac{\partial F_r}{\partial \varepsilon_1}|_{\varepsilon_1 = E_r(\varepsilon_1)} \\ \frac{\partial F_r}{\partial \varepsilon_2}|_{\varepsilon_2 = E_r(\varepsilon_2)} \\ \frac{\partial F_r}{\partial \varepsilon_3}|_{\varepsilon_3 = E_r(\varepsilon_3)} \end{bmatrix}$.

Note that in Regimes 1 and 2, the decision rules for F_i are reduced to

$$F_i|\text{Regime r} = 0$$

where r=1,2.

Notice here the observed F_i in regimes 1 and 2 is a mass point at 0, while the latent counterpart F_i^* is the solution obtained as if the no-sohrt-selling constraint $F_i \geq 0$ is not present, i.e.,

$$F_i^*|\text{Regime 1} \sim F_{3i}(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_{3i}(E_r(\boldsymbol{\varepsilon}_i)) \leq 0$$

$$F_i^*|\text{Regime 2} \sim F_{4i}(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_{4i}(E_r(\boldsymbol{\varepsilon}_i)) \leq 0$$

On the other hand, in Regimes 1 and 3, the decision rules for A_i are reduced to

$$A_i|\text{Regime r} \sim B_r(E_r(\boldsymbol{\varepsilon}_i)) \ + \boldsymbol{\varepsilon}_i \nabla B_r(E_r(\boldsymbol{\varepsilon}_i))$$

where r=1,3,$\nabla B_r(E_r(\boldsymbol{\varepsilon}_i)) = \begin{bmatrix} \frac{\partial B_r}{\partial \varepsilon_1}|_{\varepsilon_1 = E_r(\varepsilon_1)} \\ \frac{\partial B_r}{\partial \varepsilon_2}|_{\varepsilon_2 = E_r(\varepsilon_2)} \\ \frac{\partial B_r}{\partial \varepsilon_3}|_{\varepsilon_3 = E_r(\varepsilon_3)} \end{bmatrix}$.

In summary, the approximation of decision rules in each solution regime are shown in Table 2. Meanwhile, though housing quantity stays constant for each household in this model, housing wealth varies across

[14] The regime-specific functional form B_r is a result of normalisation of $E_r(\varepsilon_i) = 0$ for all regimes.

Table 1: Possible solution regimes for stable homeowners

	No-short-selling constrained ($\lambda_{3t} = 0$)	Borrowing constrained ($\lambda_{1t} > 0$ or $\lambda_{2t} > 0$)
Regime 1	Yes	Yes
Regime 2	Yes	No
Regime 3	No	Yes
Regime 4	No	No

Table 2: Approximation of decision rules in each solution regime

	A_i	F_i	F_i^*
Regime 1	$B_1(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla B_1(E_r(\boldsymbol{\varepsilon}_i))$	$\mathbf{0}$	$F_1(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_1(E_r(\boldsymbol{\varepsilon}_i)) \leq 0$
Regime 2	$A_2(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla A_2(E_r(\boldsymbol{\varepsilon}_i))$	$\mathbf{0}$	$F_2(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_2(E_r(\boldsymbol{\varepsilon}_i)) \leq 0$
Regime 3	$B_3(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla B_3(E_r(\boldsymbol{\varepsilon}_i))$	$F_3(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_3(E_r(\boldsymbol{\varepsilon}_i))$	$F_3(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_3(E_r(\boldsymbol{\varepsilon}_i)) > 0$
Regime 4	$A_4(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla A_4(E_r(\boldsymbol{\varepsilon}_i))$	$F_4(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_4(E_r(\boldsymbol{\varepsilon}_i))$	$F_4(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla F_4(E_r(\boldsymbol{\varepsilon}_i)) > 0$

Note: For regime r, $G_r \sim G_r(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla G_r(E_r(\boldsymbol{\varepsilon}_i))$

households. To capture sufficient information of a household in a period, we further assume the self-reported housing wealth is a function of $\boldsymbol{\varepsilon}_i$ with functional form G_r which varies with regime r[15].

$$G_r(\boldsymbol{\varepsilon}_i) \sim G_r(E_r(\boldsymbol{\varepsilon}_i)) + \boldsymbol{\varepsilon}_i \nabla G_r(E_r(\boldsymbol{\varepsilon}_i))$$

where $\nabla G_r(E_r(\boldsymbol{\varepsilon}_i)) = \begin{bmatrix} \frac{\partial G_r}{\partial \varepsilon_1}|_{\varepsilon_1 = E_r(\varepsilon_1)} \\ \frac{\partial G_r}{\partial \varepsilon_2}|_{\varepsilon_2 = E_r(\varepsilon_2)} \\ \frac{\partial G_r}{\partial \varepsilon_3}|_{\varepsilon_3 = E_r(\varepsilon_3)} \end{bmatrix}$.

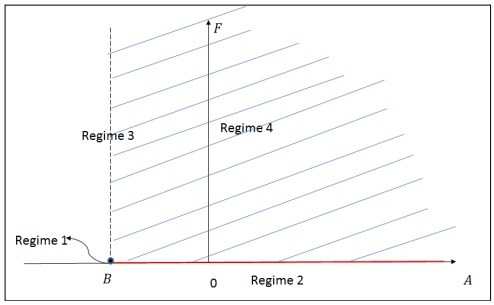

Figure 2: Possible solution regimes in the A-F space

Note: The point where A=B and F=0 is Regime 1. The bold line is Regime 2. The dotted line is Regime 3. The shaded area is Regime 4.

In order to derive the joint distribution of $(G_r, A_r, F_r^*)^T$ in each solution regime r, we first assume the joint distribution of household heterogeneity and shocks.

[15] The regime-specific functional form G_r is a result of normalisation of $E_r(\varepsilon_i) = 0$ for all regimes.

Assume the 3×1 vector $\boldsymbol{\varepsilon}_i^T$ of heterogeneity effects and shocks for each household i is drawn from a multivariate normal distribution

$$\boldsymbol{\varepsilon}_i^T \tilde{} N(\mathbf{0}, \Omega) \tag{3}$$

where $\Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$ is the covariance matrix for the heterogeneity and shocks.

Based on the normality assumption (3) and the approximation of decision rules in each solution regime as shown in Table 2, the joint distribution of $(G_r, A_r, F_r^*)^T$ for regime r is

$$\begin{bmatrix} G_r \\ A_r \\ F_r^* \end{bmatrix} \tilde{} N(\boldsymbol{\mu}_r, \Sigma)$$

where $\boldsymbol{\mu}_r = [G_r(E_r(\boldsymbol{\varepsilon}_i)), A_r(E_r(\boldsymbol{\varepsilon}_i)), F_r^*(E_r(\boldsymbol{\varepsilon}_i))]^T, \boldsymbol{\alpha}_r = \begin{bmatrix} \nabla G_r(E_r(\boldsymbol{\varepsilon}_i)) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \nabla A_r(E_r(\boldsymbol{\varepsilon}_i)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \nabla F_r(E_r(\boldsymbol{\varepsilon}_i)) \end{bmatrix}, \Sigma = \boldsymbol{\alpha}_r^T \Phi \boldsymbol{\alpha}_r, \Phi = \begin{bmatrix} \Omega & \Omega & \Omega \\ \Omega & \Omega & \Omega \\ \Omega & \Omega & \Omega \end{bmatrix}$,

Note that if $r = 1, 3$, then $A_r(E_r(\boldsymbol{\varepsilon}_i)) = B_r(E_r(\boldsymbol{\varepsilon}_i))$ and $\nabla A_r(E_r(\boldsymbol{\varepsilon}_i)) = \nabla B_r(E_r(\boldsymbol{\varepsilon}_i))$.

3 Data

The data applied in this paper is from Wealth and Asset Survey (WAS). WAS is a longitudinal survey, including information on level of various types of assets (savings, share investments, property wealth, mortgage, pension, etc.), and different sources of income flows (labour income, benefit income, pension income, etc.) of households in Great Britain. Currently there are three waves of data available (2006-2012) with each wave covering two years. In addition, demographics variables such as age, education qualification, household characteristics are available in the dataset. The advantage of this dataset is the comprehensive information on multiple asset stocks. This makes it feasible to study household finance decisions. As we don't study the tenure decisions in this paper, we select only homeowners for analysis.

3.1 Conceptual definition

For purpose of estimation, we group some important financial assets into the following 2 categories.

1. Net risk-free asset (A):

The risk-free asset includes household value of cash ISA, household value of national savings Product, household value of savings accounts, and household value of current accounts in credit. Net risk-free asset is defined as risk-free asset net of mortgage.

2. Risky asset (F):

Risky asset includes household value of Investment ISA, household value of UK Shares, household value of employee shares, household value of fixed term investment bonds, and household value of unit investment trusts.

All the monetary variables are converted to real values by Retail Price Index (RPI) setting the year 2006 as the base year.

3.2 Infrequency of home purchase decisions

Previous literature suggests an average time between house purchases is 20 to 30 years with a conservative estimate of transaction cost of 5% of value of the house sold (Grossman and Laroque, 1990). Statistical evidence shows that the annual average turnover of housing stock in the UK had fallen from over 12% in 1980s to 4.5% in 2010s. This means, on average, houses changed hands once every 8 years in 1980s and every 23 years now[16]. Flavin and Yamashita (2002) argues that though housing purchase decision is endogenous and rational, it is infrequent due to transaction cost. And in our data, the majority of the homeowners (over 97%) are non-movers.

3.3 Sample selection, household wealth and demographics

Table 3 reports the summary statistics for key variables in the sample of all the homeowners, which is a balanced panel tracing 8067 homeowners across three waves (24201 observations in total). Table 11 in the Appendix shows a detailed description of all these variables. The age range for all the homeowners is quite big, ranging from 21 to 101. To exclude the impact of pension wealth on asset allocations and focus on behaviours of non-movers, we further select households aged under 65, not retired without pension income and did not move homes during the three waves, which is a balanced panel tracing 2593 homeowners across three waves (7779 observations in total). Among the stable homeowners under 65, only 19% own their homes outright while the majority (81%) have mortgages, compared to 58% owning their homes outright and 42% having mortgages among all the homeowners. About 37% of the observations have risky assets. This participation rate in risky asset is highest in wave 2 (39%) and lowest in wave 3 (35%), while Table 3 shows the 50th percentile of risky asset holding in the whole sample of homeowners is positive. This means for the stable homeowners under 65 both the borrowing constraints and the no-short-selling constraint probably have a more important role to play compared with older households. Table 4 presents the summary statistics for these households. Figure 3 shows the histograms of housing wealth, net safe asset and risky asset for these households[17]. The distribution of housing value, net safe asset and risky asset are all skewed with long tails. Table 5 reports the correlations among age, education and asset holdings. Figure 4 is a scatter plot of mortgage against house value, where the red line represents the combinations of mortgage and house values with the loan-to-value ratio equal to 90%. We can see that while the loan-to-value ratio varies among households, most of the observations are below the red line. This is consistent with the financial market regulation. The correlation coefficients show that as households age, all the asset holdings rise. Housing

[16] Source: The new 'normal'- one year on (Is the march back to a sustainable market on track?), April 2015, Intermediary Mortgage Lenders Associations (imla) Report

[17] For the figure to be more presentable, we exclude the top 10% for each histogram.

wealth is positively correlated with both safe and risky assets, but negatively correlated with net safe asset, which is probably due to the bigger positive correlation between housing wealth and mortgage. Having a degree is positively correlated with housing wealth and risky asset.

Table 3: Summary statistics for all the homeowners from 2006-2012

			Coefficient of	Percentiles		
Variable	Observations	Mean	variation	25th	50th	75th
employ	24201	3.61	0.80	1.00	2.00	7.00
nkids	24201	0.43	1.96	0.00	0.00	0.00
degree	24201	0.30	1.53	0.00	0.00	1.00
quali	24201	0.52	0.96	0.00	1.00	1.00
Age	24201	58.41	0.25	47.00	59.00	70.00
marital	24201	2.08	0.75	1.00	1.00	3.00
totHval	24201	259205.80	0.97	134240.10	191771.50	294791.70
A	24201	-333.58	-380.92	-41354.22	5072.39	32924.44
cash	24201	37809.20	2.45	3211.29	13144.87	40000.00
mortgage	24201	38142.78	2.10	0.00	0.00	52568.43
risky	24201	38672.60	3.97	0.00	88.60	22248.01
hhNetFin	24201	83106.23	3.07	4313.18	24850.68	82907.62
GrossEmploy	24201	20778.50	1.53	0.00	9794.81	33808.89
GrossSE	24201	1940.26	5.70	0.00	0.00	0.00
Invest	24201	1822.34	5.91	0.00	48.73	664.47
income	24201	25024.18	1.47	0.00	14400.00	39918.80
lvratio	24201	0.17	1.56	0.00	0.00	0.29
hhsize	24201	2.30	0.51	2.00	2.00	3.00
bedrooms	24201	3.14	0.31	3.00	3.00	4.00
hsetype	24201	1.37	1.55	1.00	2.00	2.00

Table 4: Summary statistics for stable homeowners under 65 from 2006-2012

			Coefficient of	Percentiles		
Variable	Observations	Mean	variation	25th	50th	75th
employ	7779	1.32	0.76	1.00	1.00	1.00
nkids	7779	0.97	1.10	0.00	1.00	2.00
degree	7779	0.33	1.42	0.00	0.00	1.00
quali	7779	0.58	0.86	0.00	1.00	1.00
Age	7779	44.65	0.20	38.00	45.00	51.00
marital	7779	1.95	0.77	1.00	1.00	3.00
totHval	7779	200661.00	0.66	120000.00	166860.10	239714.40
A	7779	-44660.75	-1.82	-84316.98	-41041.96	12.24
cash	7779	17803.36	2.44	1052.83	4897.40	16651.17
mortgage	7779	62464.11	1.03	14318.45	49613.59	92225.33
risky	7779	12483.43	5.53	0.00	0.00	2876.57
hhNetFin	7779	35762.21	2.65	221.14	8840.79	37642.45
GrossEmploy	7779	35502.52	0.87	17664.38	31833.13	48008.09
GrossSE	7779	2276.80	4.13	0.00	0.00	0.00
Invest	7779	540.72	5.43	0.00	9.59	185.30
income	7779	41678.97	0.82	22200.00	37104.00	54700.13
lvratio	7779	0.35	0.86	0.08	0.31	0.56
hhsize	7779	2.93	0.44	2.00	3.00	4.00
bedrooms	7779	3.13	0.29	3.00	3.00	4.00
hsetype	7779	1.58	1.34	1.00	2.00	3.00
north_esat	7779	0.03	5.30	0.00	0.00	0.00
north_west	7779	0.14	2.49	0.00	0.00	0.00
yorkshire_humb	7779	0.12	2.69	0.00	0.00	0.00
east_mid	7779	0.10	2.96	0.00	0.00	0.00
west_mid	7779	0.10	2.99	0.00	0.00	0.00
east_england	7779	0.12	2.66	0.00	0.00	0.00
london	7779	0.09	3.26	0.00	0.00	0.00
south_east	7779	0.16	2.25	0.00	0.00	0.00
south_west	7779	0.06	3.93	0.00	0.00	0.00
wales	7779	0.07	3.75	0.00	0.00	0.00

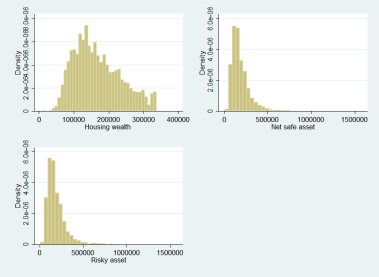

Figure 3: Histograms of housing wealth, net safe asset and risky asset for stable homeowners under 65

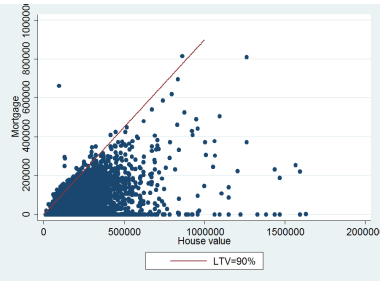

Figure 4: Scatter plot of mortgage vs. house value

Table 5: Correlation matrix for stable homeowners under 65

Variables	Age	Age2	degree	risky	totHval	A	cash	mortgage	income	lvratio
Age	1.00									
Age2	0.99	1.00								
degree	-0.08	-0.09	1.00							
risky	0.11	0.11	0.09	1.00						
totHval	0.14	0.13	0.25	0.25	1.00					
A	0.38	0.38	-0.05	0.18	-0.08	1.00				
cash	0.17	0.18	0.15	0.26	0.36	0.62	1.00			
mortgage	-0.36	-0.37	0.16	-0.06	0.35	-0.85	-0.11	1.00		
income	0.01	0.00	0.25	0.13	0.40	-0.08	0.23	0.25	1.00	
lvratio	-0.53	-0.52	0.03	-0.13	-0.18	-0.68	-0.24	0.70	0.03	1.00

4 The econometric model

4.1 A multivariate Gaussian mixture model for asset allocation patterns

We aim to estimate the assets allocations patterns. Specifically, we try to fit the data on assets holdings with a multivariate Gaussian mixture model via a censored data EM algorithm. We proceed with the assumption

for a multivariate Gaussian mixture in a clustering context that any nonnormal features in the data result from some underlying group structure (McLachlan and Peel, 2000). We will illustrate the necessity of using a censored data EM algorithm rather than the widely applied standard EM algorithm.

4.1.1 A standard EM algorithm for a multivariate Gaussian mixture model

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$ be a set of independently and identically distributed (i.i.d.) observations on a d-dimensional space R^d. In our case, $d = 3$ and

$$\boldsymbol{y}_n = (G_n, A_n, F_n^*)^T$$

where G_n, A_n are observed housing wealth and net safe asset for household n, and F_n^* is the latent counterpart of observed risky asset F_n with the observation rule as follows[18].

$$\begin{aligned} F_n &= F_n^* \text{ if } F_n^* > 0 \\ &= 0 \text{ otherwise} \end{aligned} \tag{4}$$

In this paper, we use capital letters $\boldsymbol{Y}_n$ and $\boldsymbol{Z}_n$ to represent random variables and the corresponding lower letters $\boldsymbol{y}_n$ and $\boldsymbol{z}_n$ to denote the realisations of them, respectively. The subscript n here denotes the n-th data point. When there is no subscript n, both the capital letters and lower letters represent the entire sample. The probability density function of an observation under a K-component Gaussian mixture model is written in parametric form as

$$f(\boldsymbol{y}_n; \Psi) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_n; \theta_k) \tag{5}$$

where π_k are scalars of positive mixing proportions summing to unity[19], f_k are multivariate normal density functions for component k parametrised by θ_k, and $\Psi = (\pi_1, ..., \pi_K, \theta_1, ..., \theta_K)$ is the vector containing all unknown parameters in the mixture model, $\theta_k = (\boldsymbol{\mu}_k, \Sigma_k)$ with $\boldsymbol{\mu}_k$ being the vector of means and Σ_k being the covariance matrix of component k.We use maximum likelihood (ML) to fit this mixture model via a widely applied approach, EM algorithm.

The EM algorithm is first introduced by the seminal paper by Dempster et al. (1977), which aims to find the maximum likelihood estimate from incomplete data. It is useful in incomplete data problems where algorithms such as the Newton-Raphson method may be more complicated. In our case, the Gaussian mixture model can be viewed as a model for the joint distribution of elements of $\boldsymbol{Y}_n$ depending on some unobservable (latent) vector $\boldsymbol{Z}_n$ which indicates the membership of observation n belonging to one of the K components for each observation, i.e., the complete data is

$$\boldsymbol{y_c} = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T$$

[18] This is due to the no-short-selling constraint in risky asset investment. Here F_n^* represent the optimal amount of investment in risky asset for household n. In the discussion of the standard EM algorithm, we assume that the latent F_n^* is observable.

[19] Since $\sum_{k=1}^{K} \pi_k = 1$, one of the mixing proportions π_k is redundant.

where $\boldsymbol{z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ and $\boldsymbol{z}_n$ is a K-dimensional component-label vector with its k-th element $z_n^k = 1$ if $\boldsymbol{y}_n$ is generated from component k and 0 otherwise. In our case, the missing data $\boldsymbol{z}$ is the membership indicator to the regimes that we conjecture in Section 2.

The missing Z_n can be thought of as one draw from K categories with probabilities $\pi_1, ..., \pi_K$.

That is, we assume Z_n is distributed according to a multinomial distribution:

$$\mathbf{Z}_n \tilde{} Mult_K(1, \boldsymbol{\pi})$$

where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)^T$.

The complete-data log likelihood function is

$$L_c(\Psi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_n^k [\ln \pi_k + \ln f_k(\boldsymbol{y}_n; \theta_k)] \tag{6}$$

On the other hand, the incomplete-data log likelihood function is[20]

$$\begin{aligned} L(\Psi) &= \sum_{n=1}^{N} \ln f(\boldsymbol{y}_n; \Psi) \\ &= \sum_{n=1}^{N} \ln [\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_n; \theta_k)] \end{aligned} \tag{7}$$

The standard EM algorithm proceeds iteratively in two steps, E (for expectation) and M (for maximisation). Let $\Psi^{(0)}$ be the initial value for Ψ and $\Psi^{(p)}$ be the value of Ψ after the p-th EM iteration.

In the (p+1)-th iteration, the E-step estimates the complete-data sufficient statistics, which is the conditional expectation of $L_c(\Psi)$ given $\boldsymbol{y}$ using $\Psi^{(p)}$ for Ψ.

$$Q(\Psi; \Psi^{(p)}) = E[\ln L_c(\Psi) | \boldsymbol{y}, \Psi^{(p)}] \tag{8}$$

In order to get $Q(\Psi; \Psi^{(p)})$, we need to compute $E(Z_n^k | \mathbf{y}, \Psi^{(p)})$ as follows.

$$\begin{aligned} E(Z_n^k | y, \Psi^{(p)}) &= \Pr(Z_n^k = 1 | \boldsymbol{y}, \Psi^{(p)}) \\ &= \frac{\pi_k^{(p)} f_k(\boldsymbol{y}_n; \theta_k^{(p)})}{\sum_{j=1}^{K} \pi_j^{(p)} f_j(\boldsymbol{y}_n; \theta_j^{(p)})} \\ &= w_n^k(\Psi^{(p)}) \end{aligned} \tag{9}$$

where we denote $E(Z_n^k | \mathbf{y}, \Psi^{(p)})$ by $w_n^k(\Psi^{(p)})$, the posterior probability that the n-th observation of the sample belongs to the k-th component of the mixture.

[20] $L(\Psi)$ can be seen as the log of joint density when marginalising out the unknown $\mathbf{Z}$.

Hence the conditional expectation of the complete data likelihood can be written as[21]

$$Q(\Psi;\Psi^{(p)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} w_n^k(\Psi^{(p)})[\ln \pi_k + \ln f_k(\boldsymbol{y}_n;\theta_k)] \tag{10}$$

The M-step of the (p+1)-th iteration involves maximising equation (10) with respect to Ψ. Here the update rule for $\pi_k^{(p+1)}$ is computed independently of the updated estimates $\theta^{(p+1)}$.The update rules in the M-step are in closed form:

$$\pi_k^{(p+1)} = \frac{1}{N}\sum_{n=1}^{N} w_n^k(\Psi^{(p)})$$

$$\boldsymbol{\mu}_k^{(p+1)} = \frac{\sum_{n=1}^{N} w_n^k(\Psi^{(p)})\boldsymbol{y}_n}{\sum_{n=1}^{N} w_n^k(\Psi^{(p)})}$$

$$\Sigma_k^{(p+1)} = \frac{\sum_{n=1}^{N} w_n^k(\Psi^{(p)})(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(p+1)})(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(p+1)})^T}{\sum_{n=1}^{N} w_n^k(\Psi^{(p)})}$$

The E-step and the M-step are alternated until convergence. Dempster et al. (1977) show the monotonicity of the EM algorithm; that is,

$$L(\Psi^{(p+1)}) \geq L(\Psi^{(p)})$$

4.1.2 The censored data EM algorithm for a multivariate Gaussian mixture model

In the application of standard EM algorithm, the data points $\boldsymbol{y}_n$ are all fully observed and the only missing data is the component memberships $\boldsymbol{z}_n$. However, as we mentioned above, the risky asset F_n is censored according to the observation rule (equation (4)). That is, one coordinate (F_n^*) of our data $\boldsymbol{y}_n = (G_n, A_n, F_n^*)^T$ is not fully observable because of the no-short-selling constraint. If we pretended $F_n = F_n^*$ all the time, the model would be misspecified. For this reason we apply the censored data EM algorithm introduced by Lee and Scott (2012) where they apply this algorithm to synthetic and flow cytometry data and use simulations to show that their algorithm outperforms the standard EM algorithm when there is truncation and censoring. The censored data EM algorithm deals with both the missing component memberships and the loss of exact values of the censored data. The missing component memberships are formulated as missing data as in the standard EM algorithm, while the censoring problem is addressed by integrating out the density of unknown latent values of F_n^* in the likelihood function.

We can express our observed data $\boldsymbol{x}_n$ in the following form:

$$\begin{aligned} \boldsymbol{x}_n &= \boldsymbol{y}_n \text{ if } F_n^* > 0 \\ &= \boldsymbol{x}_{mn} \text{ otherwise} \end{aligned} \tag{11}$$

where $\boldsymbol{y}_n = (G_n, A_n, F_n^*)^T$ denote the fully observed observations that preserve their latent values of positive F_n^* while $\boldsymbol{x}_{mn} = (G_n, A_n, 0)^T$ denote the observations with corner solutions of $F_n = 0$.

[21] $Q(\Psi;\Psi^{(p)})$ is obtained by replacig the unknown z_n^k in $L_c(\Psi)$ by its expected value $w_n^k(\Psi^{(p)})$.

The additional complication of censoring added to the standard EM algorithm is dealt with by identifying whether the pattern of each observation is $\boldsymbol{y}_n$ or $\boldsymbol{x}_{mn}$ according to equation (11) and then modify the likelihood contribution of an observation as opposed to equation (5). That is, now the likelihood contribution of the observed $\boldsymbol{x}_n$ is:

$$\begin{aligned} f(\boldsymbol{x}_n;\Psi) &= \sum_{k=1}^{K}\pi_k f_k(\boldsymbol{x}_n;\theta_k) \text{ if } F_n^* > 0 \\ &= \int_{-\infty}^{0}[\sum_{k=1}^{K}\pi_k f_k(\boldsymbol{x}_n;\theta_k)]dF_n \text{ otherwise} \end{aligned}$$

The posterior probability is

$$w_n^k(\Psi^{(p)}) = \frac{\pi_k^{(p)} f_k(\boldsymbol{x}_n;\theta_k^{(p)})}{\sum_{j=1}^{K}\pi_j^{(p)} f_j(\boldsymbol{x}_n;\theta_j^{(p)})}$$

Applying the EM machinery, the update rule in the M-step change from the standard EM algorithm accordingly. These are analysed in details in the work by Lee and Scott (2012).

$$\pi_k^{(p+1)} = \frac{1}{N}\sum_{n=1}^{N} w_n^k(\Psi^{(p)})$$

$$\boldsymbol{\mu}_k^{(p+1)} = \frac{\sum_{n=1}^{N} w_n^k(\Psi^{(p)})[1(F_n>0)\boldsymbol{x}_n + 1(F_n=0)\begin{bmatrix} G_n \\ A_n \\ E(F_n^*|\boldsymbol{x}_n, z_n^k=1)\end{bmatrix}]}{\sum_{n=1}^{N} w_n^k(\Psi^{(p)})}$$

$$\Sigma_k^{(p+1)} = \frac{\sum_{n=1}^{N} w_n^k(\Psi^{(p)}) S_n^k}{\sum_{n=1}^{N} w_n^k(\Psi^{(p)})}$$

where

$$\begin{aligned} S_n^k &= \{[1(F_n>0)\boldsymbol{x}_n + 1(F_n=0)\begin{bmatrix} G_n \\ A_n \\ E(F_n^*|\boldsymbol{x}_n, z_n^k=1)\end{bmatrix}] - \boldsymbol{\mu}_k^{(p+1)}\} \\ &\quad \{[1(F_n>0)\boldsymbol{x}_n + 1(F_n=0)\begin{bmatrix} G_n \\ A_n \\ E(F_n^*|\boldsymbol{x}_n, z_n^k=1)\end{bmatrix}] - \boldsymbol{\mu}_k^{(p+1)}\}^T \\ &\quad + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ 0 & R_n^k \end{bmatrix} \end{aligned}$$

$$R_n^k = 1(F_n=0)\{E(F_n^{*2}|\boldsymbol{x}_n, z_n^k=1) - E(F_n^*|\boldsymbol{x}_n, z_n^k=1)[E(F_n^*|\boldsymbol{x}_n, z_n^k=1)]^T\}$$

In our application, to choose the initial parameters $\Psi^{(0)}$, we implement k-means clustering algorithm 5

times with different starting points and choose the set of mixture model parameters from k-means that gives the maximum complete-data log likelihood. We terminate the algorithm when the increase of the complete-data log likelihood between two successive iterations is smaller than the tolerance parameter we set, or when the number of iterations reaches a fixed number.

5 Main empirical results

Tables 6 and 7 show the estimation results for the multivariate Gaussian mixture model for asset allocation patterns on each wave of data and on pooled data, respectively. Here the labels of the components are assigned by ascending order of housing wealth. The means of the four components show that on average, approximate net worth $(A_n + G_n + F_n)$ rises when housing wealth rises. Thus the first component is the poorest while the fourth is the richest. The estimated mixing proportions suggest that the unconditional probability of belonging to component 1 is the highest, while the unconditional probability of belonging to component 4 is the lowest.

Since we use censored data EM algorithm for estimation, the estimated parameters are associated with the latent data $\boldsymbol{y}_n = (G_n, A_n, F_n^*)^T$.For example, the estimated mean of risky assets for each component is the mean of F_n^* as if the observation rule (11) is not present and we could always observe the optimal risky asset holdings F_n^*. From the estimation results, we can see in all the three waves and the pooled data, the first two components both have negative means of F_n^*, which suggests that on average the first two components are no-short-selling constrained in risky asset investment. The mixing proportions π_1 and π_2 add up to about 80%, which suggests about 80% of the households are no-short-selling constrained on average.

Comparing the estimated means from wave 1 and wave 2 data, the mean of latent risky asset increases for the first two components while the housing wealth decreases for all the components. The drop of housing wealth could reflect the drop of house prices due to the financial crisis in wave 2 (2008-2009), while the increase of risky asset investment may show substitution effect of risky asset for housing for the first two components. For the third and fourth components, on the other hand, the mean of risky asset investment drops by about 91% and 32% in wave 2, which may show lack of confidence in risky asset in the face of financial crisis. In wave 3, however, all components except component 1 increase holdings of risky asset. This may reflect the recovery of confidence on risky asset for the second, third and fourth components, while the first component (the poorest) is less resilient in the post-crisis period[22].

In the results on wave 1, the estimated covariances of housing wealth and risky asset are negative for the first component, but positive for other components. This suggests for the poorest component, housing wealth and risky asset investment move in the same direction. However, for the other components, the more housing wealth owned the less investment in risky asset, which is consistent with the argument that house price risk crowds out stockholdings (Cocco, 2005).

This analysis is soft clustering where for each observation n we obtain posterior probability of belonging to the k-th component conditional on the data, w_n^k. For the purpose of visualising the results of soft clustering,

[22]One should bear in mind that the above comparisons between waves are based on the average behaviours of each cluster rather than the average behaviours of a particular group of people.

Table 6: Estimated parameters from censored data EM algorithm by wave

	Wave 1	(N=2593)			Wave 2	(N=2593)			Wave 3	(N=2593)		
	Mixing proportions											
	π_1	π_2	π_3	π_4	π_1	π_2	π_3	π_4	π_1	π_2	π_3	π_4
	0.41	0.42	0.14	0.02	0.45	0.30	0.20	0.05	0.42	0.39	0.15	0.04
	Means											
	$E(G)$	$E(A)$	$E(F^*)$		$E(G)$	$E(A)$	$E(F^*)$		$E(G)$	$E(A)$	$E(F^*)$	
μ_1	133317.5	-47985.7	-3055.63		126399.4	-45112.4	-2448.93		111619	-40104	-3342.14	
μ_2	219935.2	-58148.1	-3065.45		194802	-35403.8	-16783.7		182992.1	-34850	-14071.9	
μ_3	358255.9	-52165.2	21455.74		292327.3	-73757.4	1980.043		326184.5	-40461	5918.564	
μ_4	653977.6	-39814.3	154306.6		508513.3	13728	104694.7		526712.5	7118.211	170426.6	
	Covariance											
Σ_1	G	A	F^*		G	A	F^*		G	A	F^*	
G	1.39E+09	-4.1E+08	21133465		1.36E+09	-3.6E+08	33646917		1.02E+09	-3E+08	31272845	
A	-4.1E+08	1.44E+09	18321741		-3.6E+08	1.5E+09	13601821		-3E+08	1.31E+09	6837128	
F^*	21133465	18321741	10635291		33646917	13601821	12580937		31272845	6837128	13321620	
	G	A	F^*		G	A	F^*		G	A	F^*	
Σ_2	totHval	A	risky		totHval	A	risky		totHval	A	risky	
G	4.32E+09	-6.4E+08	-6.9E+07		3.42E+09	1.18E+08	2.19E+08		2.8E+09	-6.8E+08	-3.4E+08	
A	-6.4E+08	4.38E+09	2.03E+08		1.18E+08	4.23E+09	1.72E+09		-6.8E+08	3.51E+09	1.03E+09	
F^*	-6.9E+07	2.03E+08	2.02E+08		2.19E+08	1.72E+09	2.4E+09		-3.4E+08	1.03E+09	1.34E+09	
Σ_3	G	A	F^*		G	A	F^*		G	A	F^*	
G	1.74E+10	-4.8E+09	-2.2E+09		1.08E+10	-2E+09	2.73E+08		9.57E+09	-2.1E+09	5.89E+08	
A	-4.8E+09	1.72E+10	2.54E+09		-2E+09	9.64E+09	4.32E+08		-2.1E+09	1.39E+10	1.54E+09	
F^*	-2.2E+09	2.54E+09	2.86E+09		2.73E+08	4.32E+08	1.85E+08		5.89E+08	1.54E+09	1.18E+09	
Σ_4	G	A	F^*		G	A	F^*		G	A	F^*	
G	1.28E+11	1.06E+10	-2.5E+10		8.65E+10	-1.9E+10	2.8E+09		1.17E+11	-8.2E+09	-2.8E+10	
A	1.06E+10	7.64E+10	2.5E+10		-1.9E+10	4.84E+10	1.51E+10		-8.2E+09	5.53E+10	1.77E+10	
F^*	-2.5E+10	2.5E+10	1.17E+11		2.8E+09	1.51E+10	7.21E+10		-2.8E+10	1.77E+10	1.77E+11	

Table 7: Estimated parameters by censored data EM algorithm on pooled data

Pooled data (N=7779)				
Mixing proportions				
	π_1	π_2	π_3	π_4
	0.48	0.31	0.17	0.042
Means				
	$E(G)$	$E(A)$	$E(F^*)$	
μ_1	128888.8	-44189.4	-3515.64	
μ_2	206786.6	-39257.7	-19295	
μ_3	311812.8	-67574.2	3244.545	
μ_4	534198.9	1075.424	125359.8	
Covariance				
Σ_1	G	A	F^*	
G	1.56E+09	-3.9E+08	51069621	
A	-3.9E+08	1.53E+09	16220325	
F^*	51069621	16220325	23127503	
Σ_2	G	A	F^*	
G	4.14E+09	-2.2E+08	1.91E+08	
A	-2.2E+08	4.43E+09	1.8E+09	
F^*	1.91E+08	1.8E+09	2.84E+09	
Σ_3	G	A	F^*	
G	1.31E+10	-2.4E+09	1.84E+08	
A	-2.4E+09	1.2E+10	4.56E+08	
F^*	1.84E+08	4.56E+08	2.25E+08	
Σ_4	G	A	F^*	
G	9.96E+10	-1E+10	-1E+10	
A	-1E+10	5.47E+10	1.7E+10	
F^*	-1E+10	1.7E+10	1.13E+11	

we label each observation with the label of component with the highest posterior probability. For example, if the n-th observation has the posterior probabilities $w_n^1 = 0.8, w_n^2 = 0.03, w_n^3 = 0.07, w_n^4 = 0.1$, then this observation is labelled as belonging to group 1. That is, the group labels we assign here is in the context of hard clustering. Figures 5, 6, 7 and 8 show the histograms of posterior probabilities of belonging to the corresponding component that each observation is labelled. The peaks around 1 in these histograms means the maximum of w_n^k is near to one for most of the observations $\boldsymbol{y}_n$, which shows evidence that the components are well separated. Figures 9 and 10 show the distributions of housing wealth, net safe asset for each group (hard cluster) for each wave of the data and the pooled data, where the solid curves are fitted kernel density. In Figures 9 and 10, the distributions of housing wealth and net safe asset in each hard cluster are quite symmetric and are most concentrated in the first group and most dispersed in the fourth group. This is consistent with the estimated variances of housing wealth and net safe asset in each component. With the labels of component membership, we find most of the observations do not change the label across the three waves using the results from both the pooled data and each wave of the data. This means most of the poor households stay poor while most of the rich households stay rich in the timespan we cover (3 waves), reflecting the inertia.

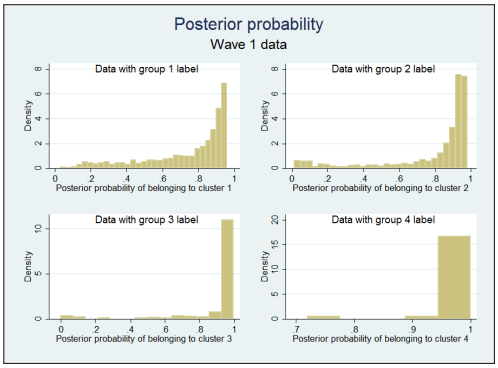

Figure 5: Histogram of posterior probabilities in each group for wave 1 data

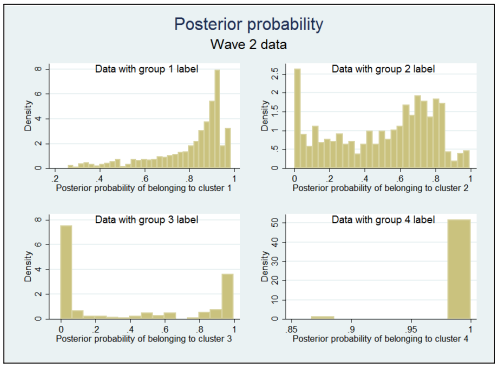

Figure 6: Histogram of posterior probabilities in each group for wave 2 data

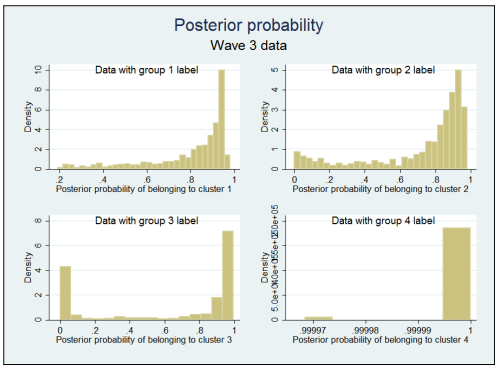

Figure 7: Histogram of posterior probabilities in each group for wave 3 data

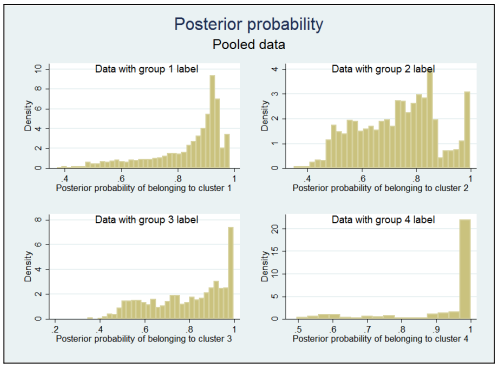

Figure 8: Histogram of posterior probabilities in each group for pooled data

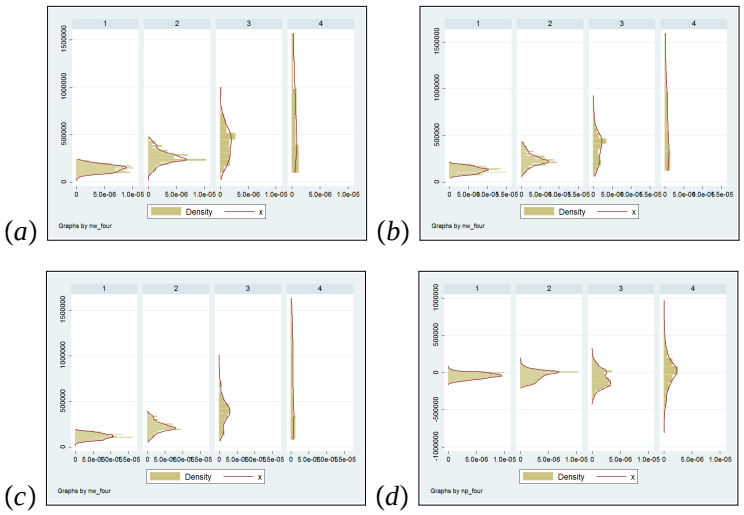

Figure 9: Histograms of housing wealth for each group

Note: (a) Wave 1 data; (b) Wave 2 data; (3) Wave 3 data; (d) Pooled data.

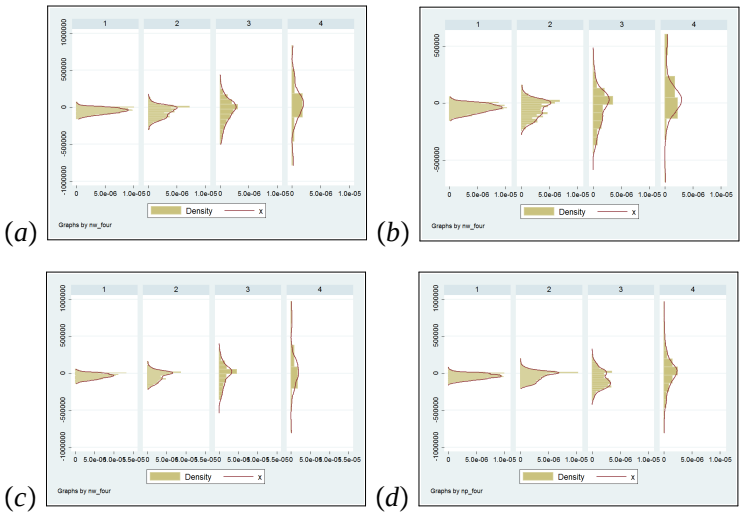

Figure 10: Histograms of net safe asset for each group

Note: (a) Wave 1 data; (b) Wave 2 data; (3) Wave 3 data; (d) Pooled data.

5.1 Determinant of component membership

We study the determinant of component membership by regressing the posterior probability of component membership on demographics and regional dummies. Since there are four components, there are four posterior probabilities correspondingly. Therefore we use a multivariate regression model, i.e., a system of linear probability models.

$$w_n^k = \boldsymbol{\beta}_k \mathbf{X}_n + \varepsilon_{nk}$$

where w_n^k is the estimated posterior probability that observation n belongs to component k obtained from the estimation of the multivariate Gaussian mixture model; $\mathbf{X}$ is a column vector of demographic variables including age, age squared, income, degree dummy, number of children under 18, and regional dummies; $\boldsymbol{\beta}_k$ is the row vector of coefficients on $\mathbf{X}$ for the k-th equation. Here the base category of regional dummies is Wales. This model is equivalent to doing k OLS regressions for each equation separately (in our model 4 regressions), except that between-equation covariances of the residuals is also estimated using this system estimator.

Tables 8 and 9 report the signs and statistical significance of the coefficients on each wave of data and the pooled data. The detailed results of regressions are shown in Tables 12, 13, 14 and 15 in Appendix. The definitions of the variables are shown in Table 11 in the Appendix. With this joint estimator instead of regressing each equation separately, we can estimate the between-equation covariance. The Breusch–Pagan test rejects the null hypothesis that the residuals of the four equations are independent of each other on all the datasets we estimate. The F test shows that all the regressors as a whole is strongly significant on all the data sets we use. The results show that households who are younger, less educated with lower income are more likely to belong to component 1, which is on average no-short-selling constrained and with lower net worth. There is also a regional effect. In general, households living in regions outside Wales are less likely to belong to the poorest component (component 1). Households living in London are more likely to belong to the richest component (component 4), compared to households living in Wales.

Table 8: The signs and statistical significance of coeficients in multivariate regressions by wave

	Wave 1 (N=2593)				Wave 2 (N=2593)				Wave 3 (N=2593)			
	w^1	w^2	w^3	w^4	w^1	w^2	w^3	w^4	w^1	w^2	w^3	w^4
Age	-	+	+	-	-	+	+	-	-	+	+	-
	**	*	n.s.	n.s.	*	n.s.	**	n.s.	***	***	**	n.s.
Age2	+	-	+	+	+	-	-	+	+	-	-	+
	n.s.	*	n.s.	*	n.s.	n.s.	*	**	**	**	n.s.	*
income	-	+	+	+	-	+	+	+	-	-	+	+
	***	n.s.	***	***	***	n.s.	***	***	***	n.s.	***	***
degree	-	+	+	+	-	+	+	+	-	+	+	+
	***	n.s.	***	n.s.	***	n.s.	***	***	***	n.s.	***	***
nkids	-	+	+	+	-	-	+	+	-	-	+	+
	***	*	**	***	***	n.s.	***	**	***	n.s.	***	n.s.
northeast	-	+	+	-	+	+	-	-	+	+	-	-
	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
northwest	-	+	+	-	-	+	-	-	+	+	-	-
	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
yorkshirehumb	-	+	+	-	-	+	-	+	-	+	+	-
	**	*	n.s.	n.s.	n.s.	*	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
eastmid	-	+	+	-	-	+	-	-	-	+	+	-
	**	**	n.s.	*	*	**	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
westmid	-	+	+	-	-	+	+	-	-	+	+	-
	***	**	n.s.	n.s.	**	**	n.s.	n.s.	**	*	n.s.	n.s.
eastengland	-	+	+	-	-	+	+	+	-	+	+	+
	***	***	**	n.s.	***	***	***	n.s.	***	***	***	n.s.
london	-	+	+	+	-	+	+	+	-	+	+	+
	***	***	***	**	***	**	***	***	***	**	***	***
southeast	-	+	+	+	-	+	+	+	-	+	+	+
	***	***	***	n.s.	***	***	***	n.s.	***	***	***	*
southwest	-	+	+	-	-	+	+	-	-	+	+	+
	***	***	***	n.s.	***	**	***	n.s.	***	***	***	n.s.
cons	+	-	-	+	+	-	-	+	+	-	-	+
	***	n.s.	**	n.s.	***	n.s.	***	n.s.	***	*	***	n.s.

Note: n.s. not significant, *p<0.05, **p<0.01, *** p<0.001.

Table 9: The signs and statistical significance of coeficients in multivariate regressions

Pooled data (N=7779)				
	w^1	w^2	w^3	w^4
Age	-	+	+	-
	***	***	***	**
Age2	+	-	-	+
	n.s.	*	**	***
income	-	+	+	+
	***	n.s.	***	***
degree	-	+	+	+
	***	**	***	***
nkids	-	+	+	+
	***	n.s.	***	***
northeast	-	+	+	-
	n.s.	n.s.	n.s.	n.s.
northwest	-	+	+	-
	n.s.	n.s.	n.s.	n.s.
yorkshirehumb	-	+	+	-
	**	**	n.s.	n.s.
eastmid	-	+	+	-
	**	***	n.s.	n.s.
westmid	-	+	+	-
	***	***	n.s.	n.s.
eastengland	-	+	+	+
	***	***	***	n.s.
london	-	+	+	+
	***	***	***	***
southeast	-	+	+	+
	***	***	***	**
southwest	-	+	+	-
	***	***	***	n.s.
cons	+	-	-	+
	***	n.s.	***	n.s.

Note: n.s. not significant, *p<0.05, **p<0.01, *** p<0.001.

5.2 Alignment in theoretical regimes and empirical components

Figure 11 shows the non-participation rate in risky asset and the percentage of mortgage holders for each group. We can see that on all the datasets, the non-participation rate decreases from Group 1 to Group 4 monotonically. The majority of households with the Group 1 label (about 80%) do not invest in any risky

asset, while about half (waves 1, 2, and 3 data) or more than a half (pooled data) of households with Group 2 label do not participate in the risky asset investment. On the other hand, only a small proportion of the households with Group 3 and Group 4 labels do not hold risky asset. This suggests that Groups 1 and 2 are likely to be no-short-selling constrained, i.e., their optimal holding of risky asset would be non-positive if the no-short-selling constraint is not present. This is consistent with signs of estimated means of F^* in Tables 6 and 7.

The percentage of mortgage holders decreases from Group 1 to Group 4 on all the datasets except for the pooled data (Figure 11). For the pooled data the percentage of mortgage holders in Group 3 (79%) is slightly higher than in Group 2 (72%). Similarly, in the estimation in each wave, the percentage of mortgage holders is the highest in Group 1 and lowest in Group 4. To investigate whether households in each group are borrowing constrained or not, we show the means and coefficients of variation of loan-to-value ratio for each group in Table 10. Note that only the sample with a loan-to-value ratio within the reasonable range (0,1) are included in Table 10, since a loan-to-value ratio outside this range could be a result of measurement error in the data and does not serve as sensible signal of borrowing constraint. Table 10 shows that on average Group 1 has the highest loan-to-value ratio with smallest variation while Group 4 has the lowest loan-to-value ratio with big variation. This indicates that households with the Group 1 label are most likely to be mortgage borrowing constrained since they borrow the most proportion of the house value on average and this proportion tends to concentrate around some maximum limit. As discussed in Section 3 the maximum borrowing limit is unobserved and can be individual-specific. However, the relative clustering of loan-to-value ratio in Group 1 compared to other groups may result from the fact that the majority of Group 1 mortgage holders borrow as much as they can to finance their houses. By contrast, households with Group 4 label are least likely to be borrowing constrained since they borrow the least proportion of the house value with the biggest variation of loan-to-value ratio. According to Figure 11 and Table 10, on the pooled data, Group 3 has higher percentage of mortgage holders and higher mean of loan-to-value ratio with bigger variation than Group 2, which may serve as evidence of Group 3 being more borrowing constrained than Group 2. However, this evidence is not as strong as the ones supporting Group 1 and Group 4, because opposite argument may be raised based on the descriptives on data in waves 1, 2 and 3.

In summary, there is strong evidence to support that components 1 and 4 are in line with the theoretical regimes 1 and 4. However, the evidence is weaker to align components 2 and 3 with the corresponding theoretical regimes.

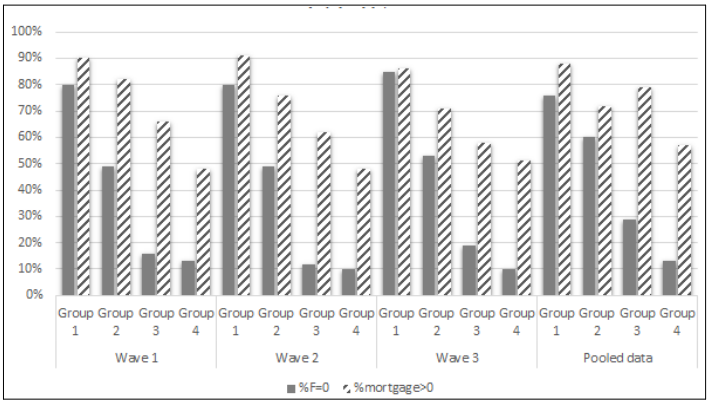

Figure 11: Non-participation rate in risky asset and percentage of mortgage holders for each group

Table 10: The means and coefficients of variation of loan-to-value ratio for mortgage holders in each group

Wave 1		
Group	Mean	Coefficient of variation
1	0.44	0.54
2	0.37	0.64
3	0.33	0.71
4	0.27	0.87
Wave 2		
Group	Mean	Coefficient of variation
1	0.46	0.53
2	0.39	0.62
3	0.36	0.69
4	0.24	0.78
Wave 3		
Group	Mean	Coefficient of variation
1	0.45	0.53
2	0.37	0.64
3	0.32	0.74
4	0.32	0.81
Pooled data		
Group	Mean	Coefficient of variation
1	0.45	0.53
2	0.36	0.67
3	0.40	0.60
4	0.30	0.79

Note: The sample selected for this table is with loan-to-value ratio within the (0,1) interval.

6 Conclusion

This paper starts with a theoretical model and derives four theoretical regimes of asset allocations depending on whether the borrowing constraint and no-short-selling constraint are binding or not. The theoretical model gives a steer to setting the number of components in the empirical work. Considering the complication of identifying the theoretical regime membership in the data, a censored data EM algorithm is used to estimate the multivariate Gaussian mixture model. Estimation results show distinct patterns of asset allocations across homeowners using the WAS data from the UK. The estimated parameters reveal that on average about 80% of the households are no-short-selling constrained in risky asset investment and with low net worth. A system of linear probability models are estimated to find determinants of component membership. Among other things, we find that households who are younger, less educated with lower income are more likely to be no-short-selling constrained in risky asset investment and with lower net worth. These findings

reflect a life-cycle effect as well as an education effect on asset allocation. The education effect could work through changing life cycle human capital and/or improving financial literacy. The estimation results and indicative evidence from loan-to-value ratio variation between components strongly suggest that the first empirical component is aligned with the first theoretical regime, while the fourth empirical component is in line with the fourth theoretical regime. There is weaker evidence to indicate that the other two components match the corresponding theoretical regimes. Apart from unobservable borrowing constraints and no-short-selling constraint, potentially, some random factors that are modeled in the theoretical model could account for the split of empirical components: heterogeneity in initial wealth and preferences (e.g. different marginal utility, expectation, risk aversion, etc.), and household idiosyncratic shocks.

The analysis in this paper is semi-parametric in the sense that the mixing proportions and component means are not parametrised. This enables the data to talk in a more flexible way than the fully parametric model. Nevertheless, it would be interesting to extend our work by parametrising mixing proportions and component means and comparing the estimation results with this paper.

A Appendix

Table 11: Variable definitions

Variable	Definition
employ	Employment Status of household representative person or partner. (1 if Employee, 2 if self-employed, 3 if unemployed, 4 if student, 5 if looking after family, 6 if sick or disabled, 7 if retired, 8 if other.)
nkids	Number of children under 18.
degree	1 if have a degree or above and 0 otherwise.
quali	1 if have qualification lower than the degree level and 0 otherwise.
Age	Age of the household repersentative person or partner.
Age2	Age squared.
marital	Marital status of household representative person or partner. (1 if married, 2 if cohabiting, 3 if single, 4 if widowed, 5 if divorced, 6 if separated, 7 if same sex couple, 8 if civil partner, 9 if former separated civil partner.)
totHval	Real value of the house owned.
A	Real net safe asset.
cash	Real safe asset.
mortgage	Real total mortgage on main residence.
risky	Real risky asset.
hhNetFin	Real household net financial wealth.
GrossEmploy	Real gross annual employee payment.
GrossSE	Real gross annual income from self employment.
Invest	Real total investment income.
income	The sum of GrossEmploy and GrossSE.
lvratio	Loan to value ratio calculated by mortgage devided by totHval.
hhsize	Number of people in household.
bedrooms	Number of bedrooms.
hsetype	Type of house (1 if detached, 2 if semi-detached, 3 if terraced).
northeast	1 if live in North East and 0 otherwise.
northwest	1 if live in North West and 0 otherwise.
yorkshirehumb	1 if live in Yorkshire and the Humber and 0 otherwise.
eastmid	1 if live in East Midlands and 0 otherwise.
westmid	1 if live in West Midlands and 0 otherwise.
eastengland	1 if live in East of England and 0 otherwise.
london	1 if live in London and 0 otherwise.
southeast	1 if live in South East and 0 otherwise.
southwest	1 if live in South West and 0 otherwise.
wales	1 if live in Wales and 0 otherwise.

Table 12: Multivariate regression results on wave 1 data

	w^1	w^2	w^3	w^4
Age	-0.021**	0.019*	0.0055	-0.0039
	(-2.95)	(2.55)	(1.02)	(-1.52)
Age2	0.00013	-0.00020*	0.0000042	0.000064*
	(1.63)	(-2.31)	(0.07)	(2.12)
income	-0.0000034***	0.00000049	0.0000022***	0.00000071***
	(-14.41)	(1.94)	(12.00)	(8.19)
degree	-0.14***	0.0046	0.13***	0.0094
	(-9.58)	(0.29)	(11.09)	(1.74)
nkids	-0.044***	0.017*	0.017**	0.0094***
	(-6.51)	(2.41)	(3.27)	(3.82)
northeast	-0.068	0.080	0.010	-0.022
	(-1.55)	(1.69)	(0.30)	(-1.33)
northwest	-0.040	0.036	0.025	-0.021
	(-1.27)	(1.07)	(1.01)	(-1.82)
yorkshirehumb	-0.083**	0.086*	0.0093	-0.012
	(-2.59)	(2.51)	(0.37)	(-1.03)
eastmid	-0.094**	0.096**	0.023	-0.026*
	(-2.84)	(2.74)	(0.91)	(-2.15)
westmid	-0.13***	0.11**	0.039	-0.015
	(-3.94)	(3.03)	(1.50)	(-1.25)
eastengland	-0.24***	0.18***	0.073**	-0.0093
	(-7.65)	(5.31)	(2.92)	(-0.79)
london	-0.38***	0.18***	0.16***	0.034**
	(-10.97)	(4.99)	(6.02)	(2.64)
southeast	-0.29***	0.17***	0.12***	0.00074
	(-9.46)	(5.05)	(5.21)	(0.07)
southwest	-0.29***	0.19***	0.12***	-0.019
	(-7.74)	(4.78)	(4.06)	(-1.42)
$_cons$	1.43***	-0.16	-0.31**	0.034
	(9.94)	(-1.01)	(-2.76)	(0.65)

Note: t statistics in parentheses. *p<0.05, **p<0.01, *** p<0.001.

Table 13: Multivariate regression results on wave 2 data

	w^1	w^2	w^3	w^4
Age	-0.016*	0.0079	0.016**	-0.0072
	(-2.26)	(1.30)	(2.81)	(-1.87)
Age2	0.000069	-0.000045	-0.00014*	0.00012**
	(0.84)	(-0.65)	(-2.26)	(2.72)
income	-0.0000037***	0.00000014	0.0000024***	0.0000012***
	(-14.87)	(0.66)	(12.31)	(9.06)
degree	-0.12***	0.0054	0.081***	0.036***
	(-8.34)	(0.43)	(7.14)	(4.63)
nkids	-0.047***	-0.00046	0.037***	0.010**
	(-6.89)	(-0.08)	(7.11)	(2.77)
northeast	0.0054	0.021	-0.0057	-0.021
	(0.12)	(0.57)	(-0.17)	(-0.89)
northwest	-0.0061	0.025	-0.0059	-0.013
	(-0.19)	(0.95)	(-0.24)	(-0.79)
yorkshirehumb	-0.061	0.066*	-0.0087	0.0041
	(-1.91)	(2.43)	(-0.35)	(0.24)
eastmid	-0.065*	0.079**	-0.0052	-0.0088
	(-1.96)	(2.82)	(-0.20)	(-0.50)
westmid	-0.092**	0.073**	0.021	-0.0019
	(-2.78)	(2.60)	(0.82)	(-0.11)
eastengland	-0.21***	0.10***	0.094***	0.014
	(-6.48)	(3.71)	(3.79)	(0.79)
london	-0.38***	0.083**	0.19***	0.098***
	(-10.86)	(2.83)	(7.30)	(5.33)
southeast	-0.26***	0.087***	0.15***	0.029
	(-8.62)	(3.33)	(6.28)	(1.79)
southwest	-0.23***	0.10**	0.13***	-0.0033
	(-6.04)	(3.18)	(4.47)	(-0.16)
$_cons$	1.42***	-0.038	-0.42***	0.043
	(9.10)	(-0.29)	(-3.53)	(0.51)

Note: t statistics in parentheses. *p<0.05, **p<0.01, *** p<0.001.

Table 14: Multivariate regression results on wave 3 data

	w^1	w^2	w^3	w^4
Age	-0.036***	0.027***	0.015**	-0.0064
	(-4.78)	(3.58)	(2.63)	(-1.67)
Age2	0.00026**	-0.00025**	-0.00010	0.00010*
	(3.15)	(-3.08)	(-1.67)	(2.45)
income	-0.0000017***	-0.00000018	0.0000012***	0.00000071***
	(-10.55)	(-1.06)	(9.66)	(8.41)
degree	-0.14***	0.0034	0.11***	0.028***
	(-9.31)	(0.22)	(9.57)	(3.57)
nkids	-0.039***	-0.0012	0.034***	0.0061
	(-5.43)	(-0.17)	(6.24)	(1.68)
northeast	0.031	0.0062	-0.018	-0.019
	(0.70)	(0.14)	(-0.53)	(-0.84)
northwest	0.0031	0.0090	-0.0018	-0.010
	(0.10)	(0.28)	(-0.07)	(-0.63)
yorkshirehumb	-0.053	0.048	0.014	-0.0087
	(-1.62)	(1.44)	(0.56)	(-0.52)
eastmid	-0.048	0.045	0.022	-0.018
	(-1.43)	(1.31)	(0.84)	(-1.07)
westmid	-0.092**	0.075*	0.023	-0.0069
	(-2.69)	(2.17)	(0.90)	(-0.40)
eastengland	-0.25***	0.13***	0.093***	0.027
	(-7.75)	(4.03)	(3.72)	(1.63)
london	-0.40***	0.10**	0.23***	0.069***
	(-11.26)	(2.83)	(8.47)	(3.83)
southeast	-0.31***	0.11***	0.16***	0.039*
	(-9.90)	(3.60)	(6.59)	(2.42)
southwest	-0.26***	0.14***	0.10***	0.012
	(-6.69)	(3.61)	(3.57)	(0.61)
$_cons$	1.83***	-0.38*	-0.50***	0.054
	(10.76)	(-2.19)	(-3.90)	(0.63)

Note: t statistics in parentheses. *p<0.05, **p<0.01, *** p<0.001.

Table 15: Multivariate regression results on pooled data

	w^1	w^2	w^3	w^4
Age	-0.017***	0.012***	0.012***	-0.0062**
	(-4.28)	(-3.37)	(-4.11)	(-3.13)
Age2	0.000088	-0.000088*	-0.00010**	0.00010***
	(-1.93)	(-2.24)	(-3.15)	(-4.55)
income	-0.0000025***	1.5E-07	0.0000015***	0.00000090***
	(-21.44)	(-1.52)	(-17.65)	(-15.52)
degree	-0.15***	0.023**	0.094***	0.030***
	(-17.04)	(-3.09)	(-15.4)	(-7.06)
nkids	-0.043***	0.0043	0.029***	0.0099***
	(-10.88)	(-1.27)	(-10.31)	(-5.04)
northeast	-0.0088	0.03	0.00093	-0.022
	(-0.34)	(-1.34)	(-0.05)	(-1.72)
northwest	-0.014	0.023	0.005	-0.015
	(-0.75)	-1.47	-0.39	(-1.61)
yorkshirehumb	-0.052**	0.052**	0.0056	-0.0057
	(-2.77)	(-3.23)	(-0.42)	(-0.62)
eastmid	-0.061**	0.071***	0.0044	-0.014
	(-3.16)	(-4.23)	(-0.32)	(-1.44)
westmid	-0.11***	0.089***	0.025	-0.0072
	(-5.48)	(-5.29)	(-1.84)	(-0.76)
eastengland	-0.23***	0.12***	0.096***	0.0096
	(-12.05)	(-7.47)	(-7.23)	(-1.04)
london	-0.39***	0.13***	0.19***	0.073***
	(-19.35)	(-7.59)	(-13.02)	(-7.32)
southeast	-0.29***	0.12***	0.14***	0.027**
	(-16.02)	(-7.63)	(-11.19)	(-3.1)
southwest	-0.24***	0.14***	0.098***	-0.0019
	(-10.85)	(-7.53)	(-6.3)	(-0.17)
cons	1.41***	-0.13	-0.33***	0.043
	(-16.26)	(-1.69)	(-5.35)	(-1.02)

Note: t statistics in parentheses. *p<0.05, **p<0.01, *** p<0.001.

References

[1] Attanasio, O.P., Bottazzi, R., Low, H.W., Nesheim, L. and Wakefield, M., 2012. Modelling the demand for housing over the life cycle. Review of Economic Dynamics, 15(1), pp.1-18.

[2] Bonaparte, Y., Korniotis, G.M. and Kumar, A., 2014. Income hedging and portfolio decisions. Journal

of Financial Economics, 113(2), pp.300-324.

[3] Campbell, J.Y., 2006. Household finance. The Journal of Finance, 61(4), pp.1553-1604.

[4] Campbell, J.Y., Chan, Y.L. and Viceira, L.M., 2003. A multivariate model of strategic asset allocation. Journal of financial economics, 67(1), pp.41-80.

[5] Carroll, C.D., 2012. Solution Methods for Microeconomic Dynamic Stochastic Optimization Problems. Johns Hopkins University. Available at http://econ.jhu.edu/people/ccarroll/solvingmicrodsops.pdf.

[6] Cocco, J.F., 2005. Portfolio choice in the presence of housing. Review of Financial studies, 18(2), pp.535-567.

[7] Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), pp.1-38.

[8] Flavin, M. and Yamashita, T., 2002. Owner-occupied housing and the composition of the household portfolio. The American Economic Review, 92(1), pp.345-362.

[9] Grossman, S.J., and Laroque, G., 1990. Asset Pricing and Optimal Portfolio Choice in the Presence of Illiquid Durable Consumption Goods. Econometrica, 58(1), pp.25-51.

[10] Heaton, J. and Lucas, D., 2000. Portfolio choice and asset prices: The importance of entrepreneurial risk. The journal of finance, 55(3), pp.1163-1198.

[11] Kim, T.S. and Omberg, E., 1996. Dynamic nonmyopic portfolio behavior. Review of financial studies, 9(1), pp.141-161.

[12] Lee, G. and Scott, C., 2012. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. Computational Statistics & Data Analysis, 56(9), pp.2816-2829.

[13] Van Rooij, M., Lusardi, A. and Alessie, R., 2011. Financial literacy and stock market participation. Journal of Financial Economics, 101(2), pp.449-472.

[14] McLachlan, G. and Peel, D. (2000) Finite Mixture Models, John Wiley & Sons, Inc., Hoboken, NJ, USA.

[15] Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica, pp.867-887.

[16] Whited, T.M. and Wu, G., 2006. Financial constraints risk. Review of Financial Studies, 19(2), pp.531-559.

[17] Yao, R. and Zhang, H.H., 2005. Optimal consumption and portfolio choices with risky housing and borrowing constraints. Review of Financial studies, 18(1), pp.197-239.

[18] Zeldes, S.P., 1989. Consumption and liquidity constraints: an empirical investigation. The Journal of Political Economy, pp.305-346.