Annals of Operations Research
https://doi.org/10.1007/s10479-022-04900-3

ORIGINAL RESEARCH

Identifying household finance heterogeneity via deep clustering

Yoontae Hwang[1] · Yongjae Lee[1] · Frank J. Fabozzi[2]

Accepted: 28 July 2022

Abstract

Households are becoming increasingly heterogeneous. While previous studies have revealed many important insights (e.g., wealth effect, income effect), they could only incorporate two or three variables at a time. However, in order to have a more detailed understanding of complex household heterogeneity, more variables should be considered simultaneously. In this study, we argue that advanced clustering techniques can be useful for investigating high-dimensional household heterogeneity. A deep learning-based clustering method is used to effectively handle the high-dimensional balance sheet data of approximately 50,000 households. The employment of appropriate dimension-reduction techniques is the key to incorporate the full joint distribution of high-dimensional data in the clustering step. Our study suggests that various variables should be used together to explain household heterogeneity. Asset variables are found to be crucial for understanding heterogeneity within wealthy households, while debt variables are more important for those households that are not wealthy. In addition, relationships with sociodemographic variables (e.g., age, education, and family size) were further analyzed. Although clusters are found only based on financial variables, they are shown to be closely related to most sociodemographic variables.

Keywords Household finance · Heterogeneous household · High-dimensional data · Clustering · Machine learning · Deep learning

1 Introduction

Households are becoming increasingly heterogeneous, due to increasing wealth inequalities (Atkinson et al., 2011; Piketty, 2013), financial crisis (Krueger & Perri, 2006), or the COVID-19 pandemic (Blundell et al., 2020; Dizioli & Pinheiro, 2021). Krueger et al. (2016) found

✉ Yongjae Lee
yongjaelee@unist.ac.kr

✉ Frank J. Fabozzi
fabozzi321@aol.com

[1] Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), 50 UNIST gil, Ulju gun, Ulsan 44919, Republic of Korea

[2] EDHEC Business School, 393 Promenade des Anglais, 06202 Nice Cedex 3, France

Published online: 21 September 2022

that households in different segments of the wealth distribution had different reactions to the 2007–2008 Global Financial Crisis, and Eichenbaum et al. (2021) reported that households have different COVID-19 pandemic mortality rates depending on their income levels. Consequently, many researchers have investigated the heterogeneity of household finances in various aspects. For example, heterogeneity in portfolio composition (Mankiw & Zeldes, 1991; Heaton & Lucas, 1997; Krusell & Smith, 1997; Case et al., 2005, 2011), income level (Constantinides & Duffie, 1996; Krueger et al., 2016; Lucas, 1994; Ahn et al., 2018), wealth level (Bricker et al., 2021; Case et al., 2005, 2011; Krueger et al., 2016), and demographics (Campbell, 2006; Berton et al., 2018; Calvet et al., 2021; Das et al., 2020) have been identified and analyzed.

However, Jappelli and Pistaferri (2014) and Krueger et al. (2016) pointed out the limitations of existing studies that separately investigate household heterogeneity in each dimension (e.g., income and wealth). That is, considering a few variables would not be enough to have a more detailed understanding of complex household heterogeneity. Krueger et al., (2016, p. 67) further noted that additional dimensions of household heterogeneity should be introduced to "better capture the joint distribution of wealth, income, and expenditure we observe in the data."

Figure 1 illustrates the average asset allocation of Korean households with respect to their wealth percentile from 2017 to 2020. Panel a of Fig. 1shows the results for the entire dataset. The proportions of deposit savings and long-term rental deposits almost monotonically decrease as households become wealthier. The proportion of residential housing increases up to middle class households, but it suddenly decreases. Instead, the proportion of nonresidential real estate increases. It is clear that the relationship between households' asset allocation and wealth level is nonlinear. Panels b and c of Fig. 1 represent the results from the bottom 20% and the top 20% income households, respectively. The relationship is clearly not simplified even if we look at subgroups partitioned by income level. This shows why conventional approaches would have difficulties in investigating the heterogeneity in household finance, which involves nonlinear relationships that are entangled in a multi-dimensional space.

Consequently, in this study, we perform a comprehensive analysis of household finance heterogeneity in various dimensions using an advanced clustering method. Since household wealth, income, and consumption are known to have skewed marginal distributions (Campbell, 2006), it would be difficult to fit such data using standard probability distributions. We believe that clustering methods can be helpful because these methods are specifically

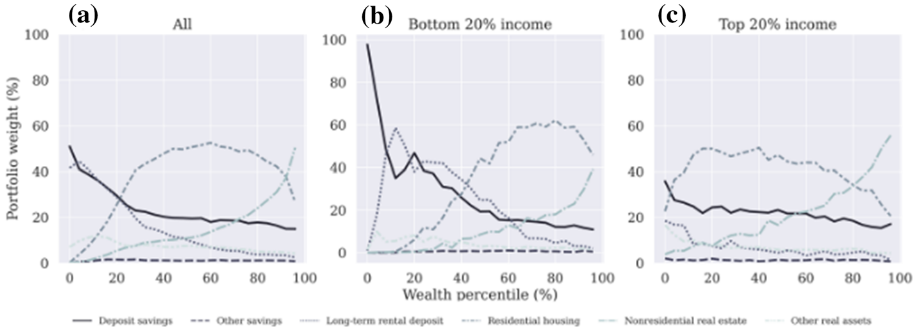


Fig. 1 Average portfolio weights of Korean households in 2017–2020

designed to find representative clusters based on the multidimensional joint distribution of data points. Because household financial data would have a complex dependence structure between a large number of items, deep learning-based and manifold learning-based dimension reduction techniques are employed along with conventional clustering methods. Many studies have shown that deep learning and manifold learning methods are helpful for handling complex nonlinear dependent structures (Bengio et al., 2012).

While we use only the financial aspects (as reported in the balance sheets) of households to identify the representative clusters, the clusters are analyzed in terms of multiple criteria. That is, the clusters are analyzed in terms of household demographics (age, gender, education, family size, and employment) as well as households' balance sheets (income, expenditure, assets, and debt). Our analysis shows that financial heterogeneity is closely related to demographic heterogeneity.

Korean household finance and living condition survey data were used in this study. Annual data from 2017 to 2019 consist of balance sheets (including income, expenditure, assets, and debt) and demographics (including age, gender, education, and employment status of householder, family size) of around 20,000 households each year. The Republic of Korea has shown remarkable growth since the Korean War in the 1950s to become the world's 10th largest economy in 2020 according to the World Bank (2021). However, such rapid growth has been accompanied by various social issues. Currently, Korea has the world's lowest fertility rate (OECD, 2021) and severe inter- and intra-generational wealth inequality compared to other developed countries (OECD, 2018). Hence, Korea offers a good example of a clearer heterogeneity in household finance.

The remainder of this paper is organized as follows. Section 2 introduces the clustering method employed in this study, Sect. 3 discusses the data and experimental setting, and Sect. 4 presents findings from the numerical experiments. Finally, Sect. 5 concludes the study.

2 Deep clustering

Consider a household i's balance sheet data $\boldsymbol{x^i} \in \mathbb{R}^d$, which consists of asset variables $\boldsymbol{x^i_A} \in \mathbb{R}^{d_A}$, debt variables $\boldsymbol{x^i_D} \in \mathbb{R}^{d_D}$, and expenditure variables $\boldsymbol{x^i_E} \in \mathbb{R}^{d_E}$. Hence, $\boldsymbol{x^i} = [\boldsymbol{x^i_A}; \boldsymbol{x^i_D}; \boldsymbol{x^i_E}] \in \mathbb{R}^d$. Our purpose is to find k clusters that divide N households based on their balance sheet data $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ so that each cluster would contain households that are similar in terms of their financial status. Hence, we apply clustering algorithms to households' balance sheet data $\mathbf{X} \in \mathbb{R}^{N\times d}$.

Clustering is one of the most popular unsupervised machine learning tasks that clusters through the similarity of data points without any label information (i.e., uses an unlabeled data). The objective of clustering is to maximize intra-group similarities and minimize inter-group similarities. Clustering methods have been shown to be useful in various tasks, such as images, medical, and finance (Ahmad & Khan, 2019).

The well-known clustering methods such as k-means, DBSCAN, hierarchical clustering, and Gaussian mixture model (GMM) have been successfully employed in various fields.[1] However, such conventional methods are not suitable for handling high-dimensional data.

Recently, many studies have shown that deep learning methods can be useful for enhancing clustering methods to effectively handle high-dimensional datasets. The so-called "deep clustering" methods have been proposed. Ghasedi Dizaji et al. (2017) and Caron et al.

[1] Saxena et al. (2017) and Ahmed and Khan (2019) provide a comprehensive review of conventional clustering algorithms.

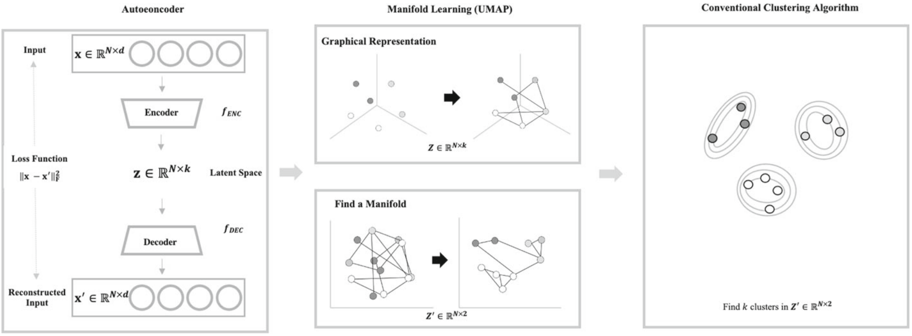

Autoeoncoder
Input
$\mathbf{x} \in \mathbb{R}^{N\times d}$
Encoder
f_{ENC}
Loss Function
$\|\mathbf{x} - \mathbf{x}'\|_F^2$
$\mathbf{z} \in \mathbb{R}^{N\times k}$
Latent Space
Decoder
f_{DEC}
Reconstructed Input
$\mathbf{x}' \in \mathbb{R}^{N\times d}$
Manifold Learning (UMAP)
Graphical Representation
$Z \in \mathbb{R}^{N\times k}$
Find a Manifold
$Z' \in \mathbb{R}^{N\times 2}$
Conventional Clustering Algorithm
Find k clusters in $Z' \in \mathbb{R}^{N\times 2}$

Fig. 2 N2D framework for deep clustering by McConville et al. (2021) (Created by the authors)

(2018) proposed clustering neural network models that utilize extracted important features from high-dimensional image data using a convolution neural network and an autoencoder,[2] respectively, which are jointly learned by interacting with conventional clustering methods (e.g., k-means). Guo et al. (2017) and Mukherjee et al. (2019) proposed clustering methods based on latent modeling using an autoencoder and generative adversarial networks,[3] respectively, and tested them on tabular data and image data. However, there was no significant performance improvement compared to conventional clustering methods.

McConville et al. (2021) proposed a simple deep clustering framework called N2D that directly uses conventional clustering algorithms (e.g., GMM) in a latent space found by deep learning and manifold learning techniques (see Fig. 2). Unlike other deep clustering methods mentioned earlier, the clustering step is separated from the dimension-reduction step. The N2D approach has been shown to achieve similar or even better performance compared to other deep clustering methods as well as conventional approaches. The key trick was to combine deep learning and manifold learning techniques to reduce the dimensionality of data by capturing complex nonlinear dependency structures. Therefore, we follow the N2D framework proposed by McConville et al. (2021) to find representative clusters of household balance sheet data.

For a household i's balance sheet data x^i, we first find its k-dimensional embedding $z^i \in \mathbb{R}^k$ via an autoencoder, and we further reduce it into a two-dimensional embedding $z'^i \in \mathbb{R}^2$ via UMAP. Then, clustering is performed with the two-dimensional embeddings z'^i of all households (i.e., for all i). The following subsections will explain in detail the two steps: (1) dimension reduction (autoencoder and UMAP) and (2) clustering (GMM).

[2] Convolutional neural networks refer to neural networks with specific structures that are known to be effective for handling image data (see Alzubaidi et al. (2021) for more detailed information). Autoencoders refer to a wide range of neural network models for dimension reduction tasks, and we will discuss these models further in Sect. 2.1.1.

[3] Generative adversarial networks (Goodfellow et al. 2014) are generative models that try to achieve high performance via adversarial training of two different neural networks. Xia et al. (2021) provides a summary of their variants and application examples.

2.1 Dimension reduction for clustering

Dimension reduction techniques are incorporated in most deep clustering methods to effectively handle high-dimensional data. The key to dimension reduction is to find low-dimensional representations (or features) lying in a high-dimensional space, which is often called latent modeling or feature extraction (Bengio et al., 2013). While other deep clustering algorithms jointly optimize latent modeling (or feature extraction) and clustering iteratively, McConville et al. (2021) separate the two tasks to simplify the overall process. However, to retain (or even improve) the performance of other deep clustering methods, they further divided the dimension reduction part into two. First, an autoencoder is used to find mid-dimensional embeddings to capture the global features. Second, manifold learning techniques, such as t-SNE and UMAP, are used to find low-dimensional manifolds to better capture local features. McConville et al. (2021) argue that such an approach can find more clusterable embeddings because both global and local features are crucial for clustering tasks.

2.1.1 Autoencoder

An autoencoder (AE) is a dimension reduction technique based on artificial neural networks and is often referred to as a deep learning version of principal component analysis (PCA), one of the most popular dimension reduction methods. While PCA is only able to capture linear dependence structures within data, AE is known to capture complex non-linear dependencies well (Bengio et al., 2013; Burges, 2010; Burges, 2010; Xie et al., 2016).

The AE is composed of an encoder function $f_{ENC} : \mathbb{R}^d \to \mathbb{R}^k$ and a decoder function $f_{DEC} : \mathbb{R}^d \to \mathbb{R}^k$. The encoder function f_{ENC} is a mapping from high-dimensional data $\mathbf{X} \in \mathbb{R}^{N \times d}$ with N samples and d features to corresponding embeddings $\mathbf{Z} \in \mathbb{R}^{N \times k}$ in a k-dimensional latent space with $k \ll d$. The decoder function f_{DEC} is a mapping from embeddings $\mathbf{Z} \in \mathbb{R}^{N \times k}$ to the original data $\mathbf{X} \in \mathbb{R}^{N \times d}$. AE is trained to minimize the following reconstruction loss:

$\ell_{\text{AE}} = \|\mathbf{X} - f_{DEC}(f_{ENC}(\mathbf{X}))\|_{\boldsymbol{F}}^2,$

where $\|\bullet\|_{\text{F}}^2$ is the Frobenius norm. While various neural network structures (e.g., convolutional neural networks and recurrent neural networks) can be used for both encoder and decoder functions, we use fully connected layers with a rectified linear unit (ReLU) for both functions. More details regarding the architectural choices are discussed in Appendix A.

Hence, the entire household balance sheet data $\mathbf{X} \in \mathbb{R}^{N \times d}$ is reduced to $\mathbf{Z} \in \mathbb{R}^{N \times k}$. Note that the embeddings are not separately found for asset, debt, and expenditure variables. Instead, each embedding incorporates all balance sheet variables so that the final clustering is done based on the entire balance sheet, not just subsets.

However, embeddings $\mathbf{Z} \in \mathbb{R}^{N \times k}$ found by AE do not necessarily preserve distances between data points $\mathbf{X} \in \mathbb{R}^{N \times d}$ in the original space, because AE is trained only in terms of minimizing the reconstruction loss. For any two data points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$ and their autoencoded embeddings $\mathbf{z}_i = f_{ENC}(\mathbf{x}_i), \mathbf{z}_j = f_{ENC}(\mathbf{x}_j) \in \mathbb{R}^k$, there is no relationship between $d(\mathbf{x}_i, \mathbf{x}_j)$ and $d(\mathbf{z}_i, \mathbf{z}_j)$, where d is an arbitrary distance measure. Then, autoencoded embeddings would not be appropriate for clustering because the objective of clustering is to find similar data points.

Therefore, in the N2D framework, clustering is not performed on the auto-encoded embeddings. Instead, AE is used to find intermediate embeddings with its dimension k not being too small, so that the distances in the original space are not fully lost. McConville et al. (2021) recommend using the dimension of autoencoded embeddings k as the desired number of clusters.

2.1.2 UMAP: uniform manifold approximation and projection

The manifold assumption in machine learning is that the observed data lie approximately on a low-dimensional manifold, and manifold learning refers to non-linear dimension reduction techniques based on such an assumption. Because a manifold is a topological concept in which every point is locally connected, manifold learning techniques are known to capture local features well. Many different models have been proposed, including isometric mapping (Tenenbaum et al., 2000), locally linear embedding (Tenenbaum et al., 2000), modified locally linear embedding (Zhang & Wang, 2007), Hessian eigenmapping (Donoho & Grimes, 2003), and t-distributed stochastic neighbor embedding (Van der Maaten & Hinton, 2008). While the last one (t-SNE) showed promising performance for complex datasets, it is often criticized for being too locally focused and lacks scalability (McConville et al., 2021).

In this regard, uniform manifold approximation and projection (UMAP) was recently proposed by McInnes et al. (2018), which is known to preserve the global structure as well as the local structure of data through a cross-entropy cost function. Let us consider a dimension-reduction task from $\mathbf{Z} \in \mathbb{R}^{N\times k}$ to $\mathbf{Z}' \in \mathbb{R}^{N\times 2}$. In other words, we wish to reduce k-dimensional dataset into two-dimensional embeddings. UMAP consists of three steps. First, graph construction. In this step, a graphical representation of $\mathbf{Z} \in \mathbb{R}^{N\times k}$ is presented. The relationship between two data points $z_i, z_j \in \mathbb{R}^k$ is represented as a probability

$$\mathrm{p}_{i|j} = \exp\left(-\frac{d\left(z_i, z_j\right) - \rho_i}{\sigma_i}\right),$$

where d is a distance measure, ρ_i is a local connectivity parameter, and σ_i is a normalization factor. Here, ρ_i is set as the average distance from z_i to its u nearest neighbors, where u controls the balance between local and global structure. If u is low, the UMAP model would focus on more detailed local structure, while a high u would ignore small details to represent global structure. Then, the global probability between the two data points is computed as.

$$\mathrm{p}_{ij} = \left(\mathrm{p}_{i|j} + \mathrm{p}_{j|i}\right) - \mathrm{p}_{i|j}\mathrm{p}_{j|i}$$

Second, graph embedding. For the corresponding embeddings $z'_i, z'_j \in \mathbb{R}^2$, the pairwise probability q_{ij} is computed as:

$$q_{ij} = \frac{1}{1 + a\|z'_i - z'_j\|^{2b}},$$

where a and b are hyper-parameters, and $\| \bullet \|$ is a norm function. Finally, cross-entropy is used as a loss function to find the optimal mapping $f_{UMAP} : \mathbb{R}^k \to \mathbb{R}^2$ from $\mathbf{Z} \in \mathbb{R}^{N\times k}$ to $\mathbf{Z}' \in \mathbb{R}^{N\times 2}$ from a fuzzy topological point of view. The cross-entropy loss function can be expressed as follows:

$$\ell_{\mathrm{UMAP}} = \sum_{i\neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right) + \left(1 - \mathrm{p}_{ij}\right)\log\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right)$$

McConville et al. (2021) tested various manifold learning techniques (isomapping, t-SNE, and UMAP) for their N2D framework, and N2D with UMAP demonstrated the best performance. Therefore, we use UMAP to find the final two-dimensional embeddings $\mathbf{Z}' \in \mathbb{R}^{N\times 2}$ from the intermediate embeddings $\mathbf{Z} \in \mathbb{R}^{N\times k}$ found by AE.

2.2 Clustering via Gaussian mixture model

Finally, the Gaussian mixture model (GMM) is employed to find clusters for the two-dimensional embeddings $\mathbf{Z}' \in \mathbb{R}^{N\times 2}$ found by AE and UMAP. Consider a k mixture of Gaussian distributions

$$\mathrm{p}(\mathrm{z}) = \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathrm{z} \mid \mu_i, \Sigma_i),$$

where $\mathcal{N}(\mathrm{z} \mid \mu_i, \Sigma_i)$ is a multi-dimensional Gaussian distribution with mean $\mathrm{\mu_i}$ and covariance matrix Σ_i, and $\mathrm{\pi_i}$ is a weight coefficient with $\mathrm{\pi_i} \geq 0$ and $\sum_{i=1}^{k} \pi_i = 1$. GMM finds the optimal parameters of the above Gaussian mixture that are most likely for the given data. That is, a log-likelihood given parameter θ_{GMM}

$$\ell_{\mathrm{GMM}} = \ln \mathrm{p}\big(\mathbf{Z}' \mid \theta_{\mathrm{GMM}}\big) = \sum_{j=1}^{N} \ln\left\{\sum_{i=1}^{k} \pi_i \mathcal{N}\Big(\mathrm{z}'_j | \mu_i, \Sigma_i\Big)\right\}$$

is maximized with respect to θ_{GMM}. Subsequently, the resulting k Gaussian distributions were considered as the optimal clusters.

Of course, conventional clustering methods would be subject to robustness issues with respect to initial points. Since k-means or GMM all start from random initial points and are not always guaranteed to converge to global optima, such clustering algorithms are often built to run multiple times with different random initial points and select the best one among them. We also use the same method to obtain more robust results.

3 Data and model

In this section, we describe our data and models (Sect. 3.1), and a simple analysis was performed to determine the appropriate number of clusters (Sect. 3.2). Also, we compare clustering performance of the deep clustering method with other popular clustering algorithms (Sect. 3.3).

3.1 Data and experimental settings

The Korean household finances and living conditions survey data were used in this study. This survey is conducted annually by the National Statistical Office of Korea, the Bank of Korea, and the Financial Supervisory Service of Korea to provide a solid ground for policymakers to account for households' financial soundness in terms of their level of income, assets, liabilities, and expenditures. Since the survey instrument was revised in 2017, we used data from 2017. The main analysis was done using survey data from 2017 to 2020. The total number of respondent households during that period was 54,920, and the number of unique households excluding multiple participation in different years was 26,907. In addition, the 2021 survey data of 18,187 households was used for out-of-sample analysis in Sect. 4.4. Note that the annual survey is conducted around every March. Hence, for example, the survey in 2020 is mostly based on households' financial activities in 2019. This means that our main analysis in Sects. 4.1–4.3 was done prior to COVID-19, and the out-of-sample analysis in Sect. 4.4 would show the changes after COVID-19.

For clustering purposes, we chose six asset-related variables, 12 debt-related variables, and seven expenditure-related variables for household balance sheets. The asset variables include deposit savings, other savings, long-term rental deposits, residential housing, non-residential real estate, other real assets. The debt variables include:

- *Mortgage loans*: Residential housing, nonresidential real estate, long-term rental deposit, living expenses, business, refinance
- *Credit loans*: Residential housing, nonresidential real estate, long-term rental deposit, living expenses, business, refinance

Expenditure variables include foodstuffs, housing, education, medical expenses, transportation, communication, other consumption expenditures. Other real assets include automobiles and valuables, and other consumption expenditures include spending on cultural life, clothing, alcohol, and tobacco. All variables are winsorized for the upper and lower 1% to handle extremely skewed distributions. In addition, they are divided by the total consumption expenditure to mitigate scale differences between households.

For demographic analysis, householder information (age, gender, education level, and employment status), number of household members, residential type, and location were used.

The specifications of the models are as follows: Both the encoder and decoder of the AE are fully connected multi-layer perceptrons (MLPs) with three hidden layers. All layers have rectified linear unit (ReLU) activation. The encoder MLP dimensions are d-100–100-200-k, where d is the dimensionality of the clustering variables and k is the number of clusters. That is, it receives a d-dimensional input, which goes through three hidden layers with 100, 100, and 200 neurons, respectively, and outputs a k-dimensional output. The decoder has an exactly opposite structure. Then, they are optimized using the Adam optimizer (Kingma & Ba, 2014). In Appendix A, we provide more detailed parameter settings and check the robustness of model outputs with respect to parameter choices. We confirm that our analysis would not be affected by small changes in parameters.

3.2 Number of clusters

We varied the number of clusters k from 4 to 12 to see how households are clustered as the number of clusters increases, and to determine the appropriate number of clusters for a more detailed analysis. Figure 3 shows the optimal clusters of household balance sheets obtained with different k, which is a hyperparameter that we should set before running the model. That is, circles with black color (label 4) represent optimal clusters when we set $k = 4$. Similarly, circles with light grey color (label 12) represent optimal clusters when we set $k = 12$. The location of a circle represents the median of total assets and total debt of households within each cluster, and the size of a circle indicates the average of the total expenditure of households within each cluster. Due to large scale differences in the total asset values of households, the asset axis is represented on a log-scale. The unit of all variables is KRW 10,000 ($\approx$ USD 10).

It can be seen from Fig. 3 that clusters are created along similar increasing curves of debt with respect to log(asset). In addition, there are a couple of clusters with very small total expenditures, while other clusters tend to have similar total spending. Hence, we would expect that there are more dimensions to household heterogeneity than total assets, total debt, and total expenditure. That is, we should investigate more detailed compositions of assets, debt, and expenditure to further understand household heterogeneity.

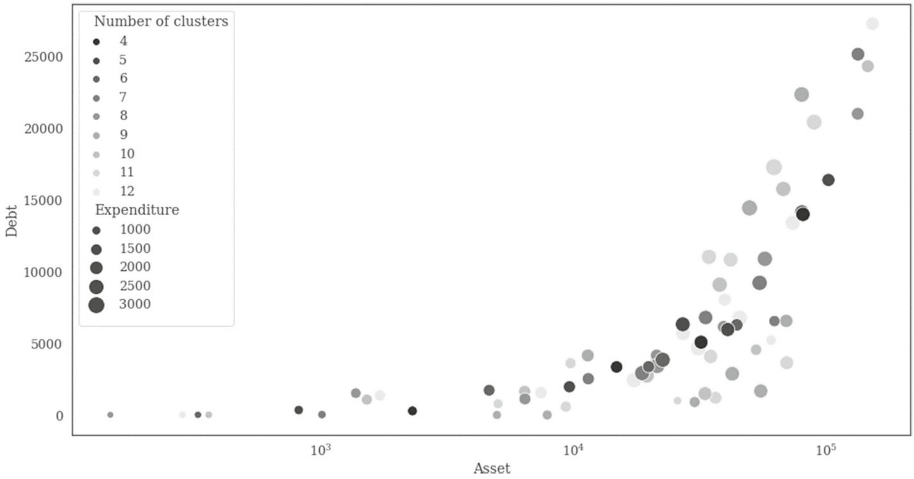


Fig. 3 Optimal household clusters with different number of clusters

Next, we determined the most appropriate *k* (number of clusters) for further analysis. There are households that appear in multiple years of the survey (17,887 out of 26,907). If they are assigned to different clusters in different years, it would result from either a significant change in the household balance sheet or unstable clustering. Thus, we keep track of these households and calculate the average of absolute changes in asset, debt, and expenditure variables. If the changes in the variables are small, it would mean that clustering is unstable. On the other hand, if the changes in the variables are large, it would imply that a household's cluster would change mostly when they had a significant change in their financial status, and thus, clustering would be stable. While the asset, debt, and expenditure variables are all used together for clustering, we calculated the changes in variables separately so that we may see more detailed aspects of the clustering results.

Table 1 shows the average absolute changes in assets, debt, expenditure variables in cluster

Table 1 Variable deviations and total count of cluster label changes

Experiments (*k*)	Average absolute difference of variables			Total count
	Asset	Debt	Expenditure	
4	0.261	0.391	0.071	7,473
5	0.235	0.308	0.069	8,587
6	0.233	0.335	0.069	9,554
7	0.239	0.333	0.070	12,202
8	**0.242**	**0.334**	**0.071**	**12,772**
9	0.226	0.347	0.067	14,458
10	0.230	0.313	0.070	15,767
11	0.222	0.316	0.067	16,212
12	0.231	0.308	0.070	16,548

changes, and total count of cluster changes. The average absolute differences indicate that cluster changes are caused by significant changes in debt and asset variables, while the effect of expenditure variables is relatively small. In terms of cluster numbers, note that the average absolute change of variables naturally decreases as the number of clusters increases because there are more clusters. For a similar reason, the total count of cluster changes tends to increase as the number of clusters increases. In this regard, the case of $k = 8$ (represented in bold) is particularly interesting because all variable changes are larger than in the case of $k = 7$ while the increment of total count is marginal compared to $k = 7$. That is, we would achieve relatively robust clusters when $k = 8$, thus, we fixed $k = 8$ for further analyses.

3.3 Model comparisons

Although we explained the reasons why we use a deep clustering method in Sect. 2, they should be backed up by performance comparisons. We compare our method (deep clustering via N2D) with four popular clustering methods, k-means, DBSCAN, hierarchical clustering (Ward's method), and hierarchical DBSCAN. k-means clustering would be the most well-known clustering algorithm that tries to separate data samples into k groups by choosing centroids that minimize the within-cluster variances. DBSCAN (Ester et al., 1996) is the acronym of density-based spatial clustering of applications with noise, which sums up its characteristics. It gathers points that are close to each other, while leaving out outliers. Hierarchical clustering methods aim to find clusters by building a hierarchy of clusters. There are various approaches depending on the linkage criterion that determines the dissimilarity between clusters. We use Ward's method, which can be seen as the hierarchical version of the k-means method. Lastly, the hierarchical DBSCAN is a hierarchical version of DBSCAN proposed by Schubert et al. (2017).

Clustering is a typical unsupervised learning task, and thus, the performance evaluation of clustering algorithms is not as trivial as regression models and classification models. The two most popular metrics are the Silhouette index and Davies-Bouldin index. The Silhouette index, proposed by Rousseeuw (1987), measures how each data point is similar to its own cluster compared to other clusters. The Davies-Bouldin index (Davies & Bouldin, 1979) represents the average similarity between each cluster and its closest cluster. Hence, good clusters would have a high Silhouette index but a low Davies-Bouldin index.

Table 2 summarizes the clustering performances of different methods. For each method, the number of clusters k is chosen to maximize the Silhouette index and minimize the Davies-Bouldin index. It is clear that the deep clustering method shows the best performance compared to other popular clustering methods in terms of two indexes in our dataset.

Table 2 Clustering performance comparison

	k-means	DBSCAN	Hierarchical clustering	Hierarchical DBSCAN	Deep clustering
	($k = 10$)	($k = 13$)	($k = 7$)	($k = 7$)	**($\boldsymbol{k = 8}$)**
Silhouette (↑)	0.317	0.065	0.292	0.154	**0.381**
Davies-Bouldin index (↓)	1.418	1.278	1.515	1.553	**0.816**

4 Analysis of household heterogeneity via deep clustering

In this section, we find representative clusters of household balance sheets via deep clustering and analyze them. The optimal clusters are analyzed in detail in terms of financial (Sect. 4.1) and demographic (Sect. 4.2) perspectives. The inter-cluster mobility is discussed in Sect. 4.3. Finally, we present an out-of-sample analysis in Sect. 4.4.

4.1 Household heterogeneity in balance sheets

As we have seen from Fig. 1, the relationship between asset allocation and wealth level is highly nonlinear, and dividing households in terms of wealth level was not helpful in simplifying the relationship. We present the same results for all five income quintiles, four age groups (under 40, 40 to 50, 50 to 60, above 60), and 20 income-age groups in Appendix B. While households are often classified in terms of their income or age, these results indicate that such groups do not do much to reduce within-group heterogeneity.

Figure 4 represents the average portfolio weights with respect to wealth level of different household clusters found by the deep clustering method. We can clearly see that the relationship has become much simpler. In particular, asset allocations seem almost constant within Clusters 1 to 4. This shows what deep learning can do in analyzing complex household finance data. Deep learning has been exceptional in capturing nonlinear dependencies within data. Hence, it was able to group households accounting for complex relationships, and thus, groups have much higher within-group homogeneity.

We now investigate the financial heterogeneity of households in more detail. Table 3 summarizes the financial variables of eight clusters with units of KRW 10,000 (≈ USD 10).[4] Clusters are sorted with respect to the average total asset value in descending order. Hence, Cluster 1 was the wealthiest group and Cluster 8 was the poorest group. The numbers in parentheses are proportions of each variable within the asset, debt, and expenditure categories. Values with relatively large proportions compared to other clusters are highlighted in bold.

For assets shown in Panel A of Table 3, there is a clear tendency that the wealthy-half (Clusters 1, 2, 3, 4) hold more than 50% of their assets in real estate (residential and non-residential), while non-wealthy-half (Clusters 5, 6, 7, 8) hold more than 50% of their assets in financial assets (deposit savings, other savings, long-term rental deposits). Among the wealthy-half, the wealthiest two (Clusters 1 and 2) have a significant amount of nonresidential real estate, but the other two (Clusters 3 and 4) do not. As for the non-wealthy-half, Cluster 5 has more than 60% of their assets in long-term rental deposits, whereas Clusters 6 and 7 are more concentrated in savings and other real assets. Cluster 8 seems to be the poorest group with a very small amount of assets. Overall, the major asset classes of different household groups are summarized in Fig. 5.

It is widely known that Korean household wealth is excessively concentrated in real estate compared to other developed countries (Fredriksen, 2012; Park, 2020). However, our analysis reveals that this statement is true only for the wealthy-half groups. This shows the importance of analyzing heterogeneous household groups, because aggregated values would be naturally biased towards wealthy groups that possess large amounts of assets.

A similar tendency can be found for the debt variables (Panel B of Table 3). More than 30% of loans in Clusters 1 and 2 are for nonresidential real estate, and more than 60% of loans in Clusters 3 and 4 are for residential housing. Approximately 70% of the loans for Cluster 5 are for long-term rental deposits, and more than 70% of loans in Clusters 7

[4] More detailed statistics of household balance sheets of different clusters are given in Appendix C.

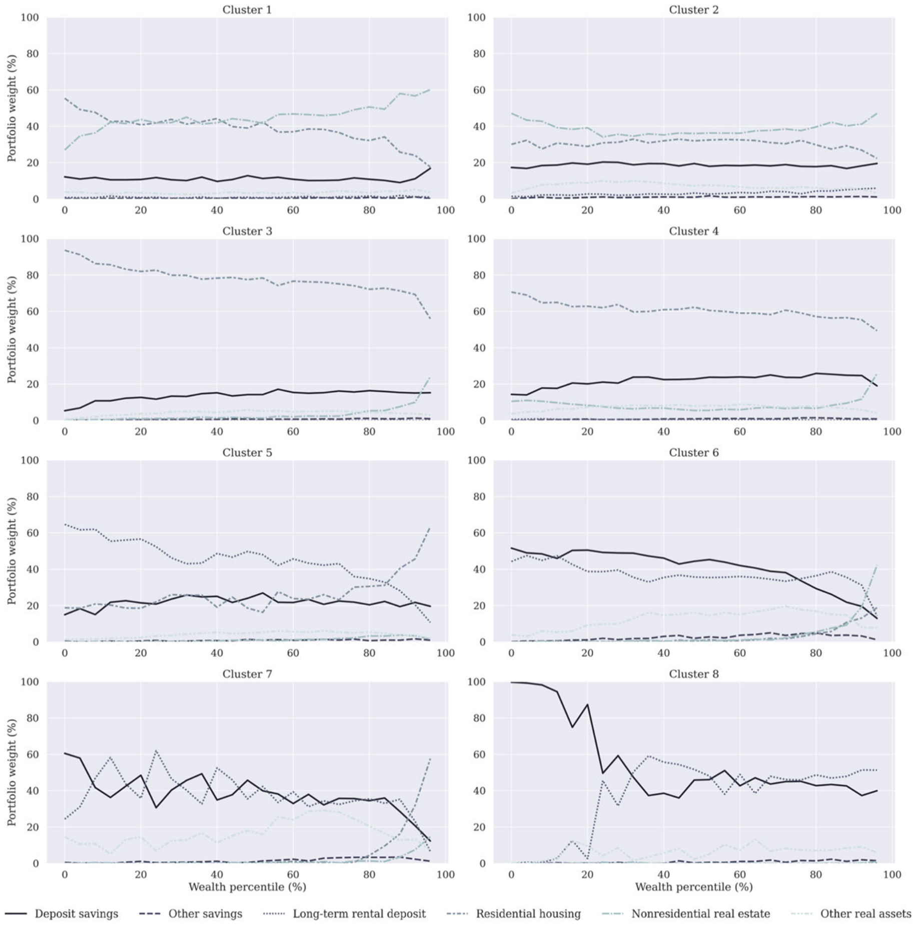


Fig. 4 Average portfolio weights of different household clusters

and 8 are for living expenses, business funds, and refinances. Hence, the purpose of loans changes from urgent financial liquidity to purchasing real estate as the household wealth level increases. In addition, more than 70% of the loans in Clusters 1 to 5 are mortgage loans, but the other clusters have more credit loans. Clusters 7 and 8 rarely have mortgage loans ($\leq$ 10%), probably due to a lack of underlying assets. Figure 6 summarizes the findings.

Panel C of Table 3 shows the expenditure variables for different household clusters. While the overall proportions are not as heterogeneous as in the asset and debt variables, a few interesting observations can be found. First, the poorest two clusters (7 and 8) spent a relatively large amount on housing ($\geq$ 20%) compared to the others. Second, Clusters 2 to 5 tended to invest more on education ($\geq$ 10%). Third, more than 10% of the expenditure of the poorest group (Cluster 8) is for medical purposes. Fourth, wealthy groups (Clusters 1 to 5) tend to spend slightly more (around 25%) for cultural life, clothing, alcohol, tobacco, etc. (categorized as ‘others’).

A rough decision tree is shown in Fig. 7 to summarize the multidimensional heterogeneity of household finance. We can see that asset and debt variables are more crucial for representing

Table 3 Average values (proportions) of asset, debt, expenditure variables of different household clusters

No	Deposit savings	Other savings	Long-term rental deposit	Residential housing	Nonresidential real estate	Other real assets
Panel A. Assets						
1	16,089.8 (12.1%)	662.4 (0.5%)	1343.1 (1.0%)	**42,702.6 (32.2%)**	**67,093.6 (50.6%)**	4673.1 (3.5%)
2	10,386.6 (18.4%)	801.0 (1.4%)	2470.8 (4.4%)	**16,426.5 (29.1%)**	**22,199.5 (39.3%)**	4167.7 (7.4%)
3	6507.1 (16.6%)	353.3 (0.9%)	1.1 (0.0%)	**30,408.0 (77.8%)**	227.3 (0.6%)	1589.6 (4.1%)
4	5189.1 (24.4%)	219.9 (1.0%)	82.1 (0.4%)	**12,662.5 (59.5%)**	1500.5 (7.1%)	1612.5 (7.6%)
5	**5822.6 (28.5%)**	327.1 (1.6%)	**12,501.0 (61.3%)**	1.1 (0.0%)	674.7 (3.3%)	1076.2 (5.3%)
6	**2735.1 (43.2%)**	**336.3 (5.3%)**	**2016.5 (31.8%)**	18.5 (0.3%)	95.4 (1.5%)	**1132.5 (17.9%)**
7	**469.7 (34.3%)**	**33.0 (2.4%)**	**474.8 (34.7%)**	31.3 (2.3.%)	18.5 (1.4%)	**339.4 (24.8%)**
8	**62.7 (42.5%)**	1.3(0.6%)	**78.2 (52.9%)**	0.0 (0.0%)	0.2 (0.1%)	5.2 (3.6%)
No	Residential housing	Nonresidential real estate	Long-term rental deposit	Living expense	Business funds	Refinance
Panel B-1. Debts (Mortgage loans)						
1	2185.7 (19.1%)	**4906.1** (42.9%)	87.4 (0.8%)	130.9 (1.1%)	**3208.7 (28.1%)**	123.9 (1.1%)
2	1699.0 (26.3%)	**1959.3** (30.3%)	228.2 (3.5%)	118.1 (1.8%)	**1382.6 (21.4%)**	108.4 (1.7%)
3	**4189.9 (81.8%)**	24.2 (0.5%)	27.5 (0.5%)	142.9 (2.8%)	259.5 (5.1%)	82.4 (1.6%)
4	**1712.2 (62.2%)**	90.3 (3.3%)	26.2 (1.0%)	136.3 (5.0%)	252.6 (9.2%)	75.3 (2.7%)
5	9.7 (0.3%)	75.9 (2.2%)	**2456.7** (70.7%)	56.8 (1.6%)	62.5 (1.8%)	16.9 (0.5%)
6	0.4 (0.1%)	3.7 (0.5%)	**146.4 (20.9%)**	60.6 (8.7%)	74.1 (10.6%)	14.9 (2.1%)
7	7.7 (1.3%)	0.5 (0.1%)	8.2 (1.3%)	15.5 (2.5%)	18.3 (3.0%)	5.1 (0.8%)
8	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
Panel B-2. Debts (Credit loans)						

Table 3 (continued)

No	Residential housing	Nonresidential real estate	Long-term rental deposit	Living expense	Business funds	Refinance
1	57.1 (0.5%)	232.3 (2.0%)	22.7 (0.2%)	100.5 (0.9%)	342.8 (3.0%)	36.1 (0.3%)
2	102.9 (1.6%)	213.4 (3.3%)	55.0 (0.9%)	161.6 (2.5%)	380.0 (5.9%)	50.9 (0.8%)
3	95.9 (1.9%)	15.5 (0.3%)	5.8 (0.1%)	135.4 (2.6%)	119.2 (2.3%)	25.4 (0.5%)
4	71.6 (2.5%)	32.9 (1.2%)	9.9 (0.4%)	187.6 (6.8%)	124.8 (4.5%)	31.4 (1.1%)
5	64.4 (1.9%)	125.4 (3.6%)	**336.1** (9.7%)	119.5 (3.4%)	136.5 (3.9%)	16.0 (0.5%)
6	6.0 (0.9%)	20.1 (2.9%)	**76.7 (11.0%)**	**141.7 (20.2%)**	**126.9 (18.1%)**	28.5 (4.1%)
7	1.6 (0.3%)	9.0 (1.5%)	**44.8 (7.3%)**	**188.5 (30.8%)**	**201.2 (32.9%)**	**111.8 (18.3%)**
8	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	**0.6 (100.0%)**	0.0 (0.0%)	0.0 (0.0%)

No	Foodstuffs	Housing	Education	Medical	Transportation	Communication	Others
Panel C. Expenditures							
1	712.3 (31.4%)	292.0 (12.9%)	155.7 (6.9%)	189.3 (8.4%)	218.1 (9.6%)	125.0 (5.5%)	**572.7 (25.3%)**
2	856.1 (27.8%)	325.7 (10.5%)	**388.3 (12.5%)**	219.1 (7.0%)	327.3 (10.5%)	181.7 (5.8%)	**801.4 (25.8%)**
3	708.7 (31.4%)	257.9 (11.4%)	**230.9 (10.2%)**	164.2 (7.3%)	217.3 (9.6%)	142.5 (6.3%)	**537.5 (23.8%)**
4	885.2 (28.5%)	296.5 (9.5%)	**418.5 (13.5%)**	227.9 (7.3%)	323.7 (10.4%)	196.6 (6.3%)	**759.9 (24.4%)**
5	677.6 (31.2%)	260.5 (12.0%)	**234.3 (10.8%)**	117.5 (5.3%)	209.0 (9.6%)	145.5 (6.7%)	**528.3 (24.3%)**
6	604.3 (29.4%)	368.2 (17.9%)	165.8 (8.1%)	116.8 (5.7%)	207.5 (10.1%)	147.5 (7.2%)	447.2 (21.7%)
7	503.2 (29.5%)	**368.4 (21.6%)**	106.1 (6.2%)	107.6 (6.3%)	160.9 (9.4%)	125.7 (7.4%)	332.2 (19.5%)
8	339.7 (36.3%)	**210.5 (22.5%)**	18.4 (2.0%)	**110.0 (11.8%)**	54.1 (5.8%)	54.1 (5.8%)	148.5 (15.9%)

(Unit: KRW 10,000)

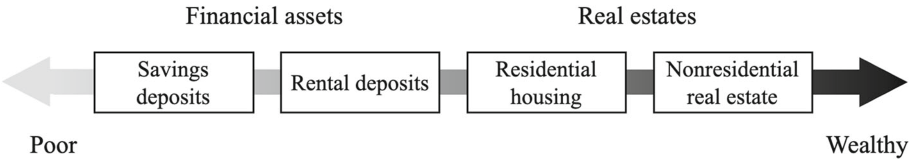


Fig. 5 Major asset class of households with different level of wealth

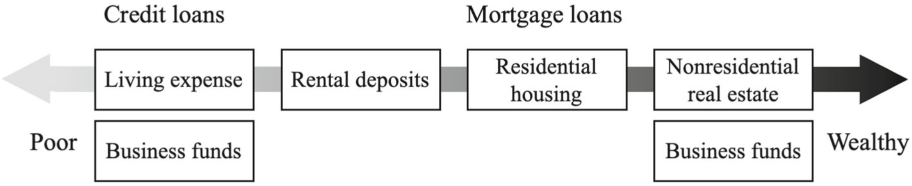


Fig. 6 Major loan types of households with different level of wealth‘

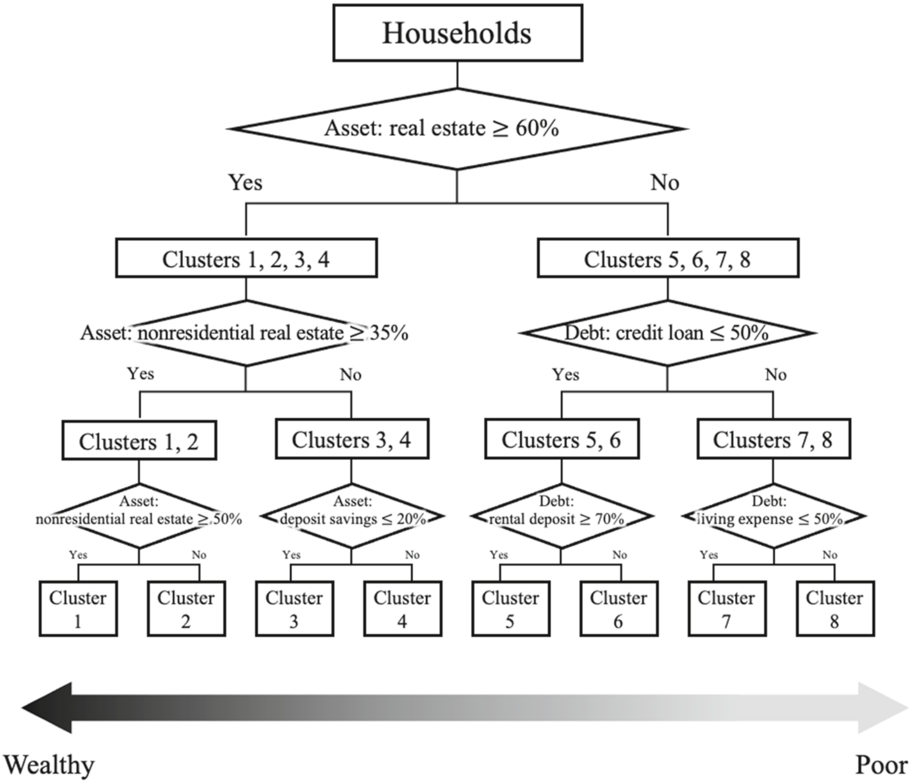


Fig. 7 Decision tree for household clusters

the heterogeneity of households than expenditure variables. For more detailed classifications, asset compositions (especially real estate) are important for wealthy groups, whereas the purpose and type of debt are important for non-wealthy groups.

4.1.1 Clustering quality and variable importance

Here we further check the quality of the clustering results and the importance of each variable by investigating how variables are distributed within and between groups. Recall that the objective of clustering is to find clusters with high within-cluster similarities and low between-cluster similarities. We believe that the Gini coefficient and its decomposition can be useful in this regard. The Gini coefficient is a popular measure of inequality in the distribution of income or wealth, and some researchers have decomposed the Gini coefficient to investigate the causes of disparity in income distributions with different populations and educational backgrounds (Deaton & Paxson, 1994, 1997). There are two popular approaches to decomposition: Pyatt (1976) and Shorrocks (1982). While the former directly compares the Gini coefficients of different groups, the latter linearly decomposes the Gini coefficient into within-group, between-group, and overlapping inequalities. We use the latter approach because it quantifies within-group and between-group inequalities that are exactly in line with the clustering objective.

Let us consider *k* groups (or clusters) and a variable Y. $\mathrm{Y_I}$ represents the variable within group i with mean $\mathrm{\mu_I}$ and cumulative distribution $\mathrm{F_i(Y_i)}$. Then, the overall population $\mathrm{Y_u} = \mathrm{Y_1 \cup Y_2 \ldots \cup Y_k}$ is the union of all groups with $\mathrm{F_u(Y_u)} = \sum_{\mathrm{i}} \mathrm{p_i F_i(Y_i)}$, where $\mathrm{p_I}$ is the population share of group i, with mean $\mathrm{\mu_u}$. The Gini coefficient of the overall population is defined as.

$$\mathrm{G} = \frac{2\mathrm{cov}(\mathrm{Y_u}, \mathrm{F_u}(\mathrm{Y_u}))}{\mathrm{\mu_u}},$$

and Mookherjee and Shorrocks (1982) decomposed it into

$$\mathrm{G} = \mathrm{G_W} + \mathrm{G_B} + \mathrm{G_O}.$$

Here, within-group inequality $\mathrm{G_W}$ is defined as $\mathrm{G_W} = \sum_{\mathrm{i}} \mathrm{p_i q_i G_i}$, where $\mathrm{q_i}$ is the variable share of group i, $\mathrm{G_i} = \frac{2\mathrm{cov}(\mathrm{Y_i}, \mathrm{F_i}(\mathrm{Y_i}))}{\mathrm{\mu_i}}$ is the Gini coefficient within group i. Between-group inequality $\mathrm{G_B}$ is defined as $\mathrm{G_B} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \frac{\mathrm{p_i p_j} |\mathrm{\mu_i} - \mathrm{\mu_j}|}{2\mathrm{\mu_u}}$, and overlapping inequality $\mathrm{G_O}$ is the remainder.

We calculated within-group inequality ($\mathrm{G_W}$), between-group inequality ($\mathrm{G_B}$), and overlapping inequality ($\mathrm{G_O}$) for all cluster variables, and the proportions of the three inequalities are shown in Fig. 8. Three important observations were made. First, we can see that all within-group inequalities are less than 20% and are mostly much less than between group inequalities. This indicates that the quality of clustering is good because all variables tend to have high within-group similarities and low between-group similarities. Second, there are some variables in which between-group inequality accounts for more than 60% of the Gini index. For example, long-term rental deposits, residential housing, nonresidential real estate, mortgage loans for nonresidential housing, long-term rental deposits, business funds, and credit loans for long-term rental deposits. All these variables were shown to be very important in interpreting the clustering results. Third, all expenditure variables exhibited more than 60% of the overlapping inequalities. That is, these variables do not contribute much to clustering, which is consistent with our previous discussion.

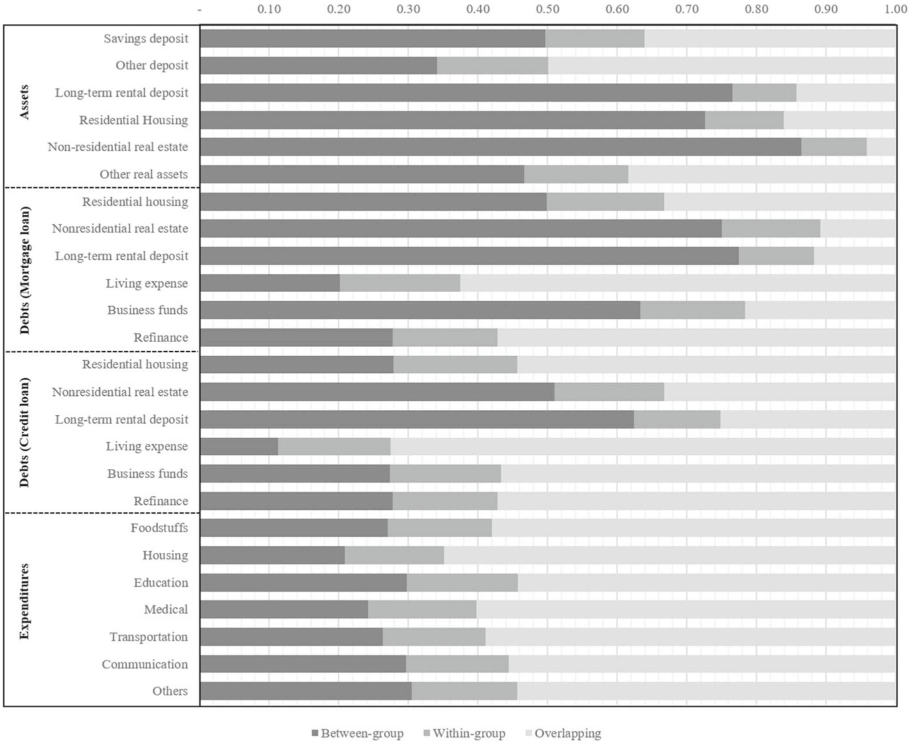


Fig. 8 Decomposition of Gini coefficients into between-group, within-group, and overlapping inequalities

Table 4 List of independent variables for logistic regression

Independent variable	Description
Area of residence	Living in Seoul metropolitan area or not
Gender of householder	Male or not
Number of family members	(Numbers are directly used for regression)
Education level of householder	Under middle school, high school, or higher education
Home ownership	None (includes monthly rental or free company housing), long-term rental, or homeowner
Age of householder	Under 39, 40 ~ 49, 50 ~ 59, or upper 60
Income level	Low-income (1st and 2nd income quintiles), mid-income (3rd income quintile), or high-income (4th and 5th income quintiles)
Employment status	Employed or not (includes freelancers or helping family business)

4.2 Sociodemographic characteristics of clusters

Although optimal clusters are found only with respect to a financial perspective, there is no doubt that household finance is closely related to sociodemographics, such as householder's age, education level, and so on. Therefore, we conducted logistic regressions for all clusters

Table 5 Logistic regression results of clusters with respect to socio-demographic variables

Variables	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio
Constant	− 5.224***	0.005	− 3.972***	0.019	− 2.833***	0.059	− 3.025***	0.049
Metropolitan area	*0.642****	*1.900*	− **0.646*****	**0.524**	*0.622****	*1.863*	− **0.977*****	**0.376**
Gender (male)	*0.224****	*1.251*	*0.474****	*1.607*	− .109***	0.897	− 0.014	0.986
Number of members	− **0.349*****	**0.705**	0.062***	1.064	− 0.047***	0.954	*0.338****	*1.402*
Education (under middle school)								
High school	0.246***	1.279	− 0.064*	0.938	− 0.053	0.949	− **0.353*****	**0.702**
Higher education	*0.773****	*2.166*	0.129***	1.138	0.134***	1.143	− **0.751*****	**0.472**
Home ownership (none)								
Long-term rental	0.489***	1.630	0.597***	1.816	0.063	1.065	− 0.463***	0.629
Homeowner	1.486***	4.421	0.656***	1.927	*2.806****	*16.547*	*1.648****	*5.197*
Age (under 39)								
40 ~ 49	0.665***	1.945	0.596***	1.815	− **0.304*****	**0.738**	*0.385****	*1.470*
50 ~ 59	1.330***	3.780	1.000***	2.718	− **0.584*****	**0.558**	0.164***	1.178
Upper 60	*2.539****	*12.66*	*1.235****	*3.440*	− **0.472*****	**0.624**	− 0.485***	0.616
Income level (low-income)								
Mid-income	0.338***	1.403	0.356***	1.428	− 0.119***	0.888	− 0.031***	0.734
High-income	*0.812****	*2.252*	*0.799****	*2.224*	− 0.358***	0.699	− 0.257	0.773
Employment	0.012	1.012	*0.628****	*1.875*	− **0.491*****	**0.612**	0.193***	1.213
Number of households	5937		10,644		10,699		10,001	

Table 5 (continued)

Variables	Cluster 5		Cluster 6		Cluster 7		Cluster 8	
	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio	Coeff	Odds ratio
Constant	− 1.402***	0.246	− 0.367***	0.693	− 0.088	0.916	− 0.107***	0.899
Metropolitan area	*0.794****	*2.212*	0.013	1.013	0.021	1.022	− 0.205**	0.815
Gender (male)	**− 0.411*****	**0.663**	**− 0.225*****	**0.799**	0.005	1.005	0.137**	1.147
Number of members	− 0.098***	0.906	0.004	1.004	− 0.068***	0.934	**− 0.525*****	**0.592**
Education (under middle school)								
High school	− 0.050	0.951	0.204***	1.227	0.017	1.017	**− 0.599*****	**0.549**
Higher education	*0.310****	*1.363*	− 0.188***	0.829	− 0.435***	0.647	**− 1.330*****	**0.264**
Home ownership (None)								
Long-term rental	*2.197****	*8.995*	**− 0.979*****	**0.376**	**− 2.421*****	**0.089**	**− 2.414*****	**0.089**
Homeowner	− 1.674***	0.187	**− 2.840*****	**0.058**	**− 2.911*****	**0.054**	**− 2.758*****	**0.063**
Age (under 39)								
40 ~ 49	**− 0.578*****	**0.561**	**− 0.170*****	**0.843**	**− 0.177*****	**0.838**	− 0.139	0.870
50 ~ 59	**− 0.850*****	**0.427**	**− 0.368*****	**0.692**	**− 0.251*****	**0.778**	− 0.019	0.981
Upper 60	**− 0.957*****	**0.384**	**− 0.823*****	**0.439**	**− 0.810*****	**0.445**	− 0.168	0.845
Income level (low-income)								
Mid-income	− 0.043	0.958	− 0.326***	0.722	**− 0.855*****	**0.425**	**− 1.753*****	**0.173**
High-income	− 0.331***	0.718	− 1.018***	0.361	**− 2.040*****	**0.130**	**− 2.299*****	**0.100**
Employment	− 0.031	0.970	*0.328****	*1.388*	**− 0.262*****	**0.769**	**− 1.091*****	**0.336**
Number of households	5614		6204		4223		1598	

*$p < .05$, **$p < .01$, ***$p < .001$

to investigate their sociodemographic characteristics. Consider a logistic regression for a cluster. The dependent variable y_i is defined to represent whether a household is in a cluster. The independent variables are presented in Table 4. (Detailed statistics with the percentage see Appendix A.1.)

Table 5 summarizes the results of logistic regressions. Regression coefficients with statistical significance and corresponding odd ratios are shown. Notable variables are highlighted with shadows: positive (italic) and negative (bold) relationships. We can see that most variables are statistically significant, while having both positive and negative values. It shows a strong relationship between the multidimensional heterogeneity of household finance and sociodemographics.

Clusters 1 and 2, the wealthiest two groups, were shown to consist of older households compared to others. They both tend to have a highly educated male householder, living in their own houses, and have a high income. While Cluster 2 households live outside the Seoul metropolitan area and are employed, Cluster 1 households live in or near Seoul and have a small number of family members with mixed employment status. Cluster 2 was also more likely to have more family members.

Cluster 3 is quite unique in that it is one of the wealthiest groups with their own houses in metropolitan areas, but its households are likely to be unemployed (including freelance or helping family business) and have low income. They can also be characterized as highly educated young households. Perhaps this peculiar cluster represents young households who inherited houses early.

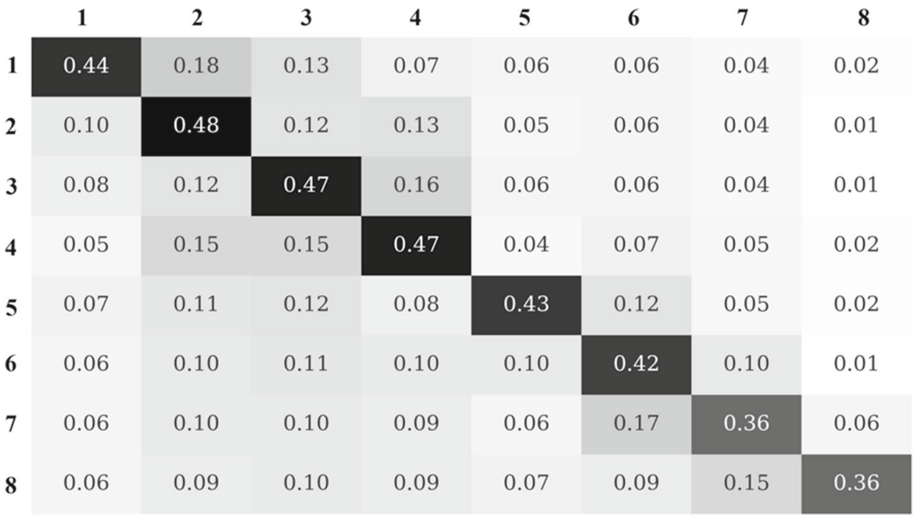
Clusters 4 and 5 can be regarded as two middle-class groups. Cluster 4 can be characterized as living outside metropolitan areas, large families, homeowners, and low education, while Cluster 5 can be characterized as living in metropolitan areas, small families, long-term rental housing, high education, and high income. These reflect typical rural–urban differences in family size (Key, 1961), income (Lipton, 1977), education (van Maarseveen, 2020), and housing affordability (Lee & Jun, 2018).

Clusters 6 and 7 both consist of poor households who are relatively young, under temporary housing (mostly monthly rent), with no higher education. However, the former is likely to be employed, whereas the latter is not.

Cluster 8 clearly represents the most vulnerable households with very small families (high probability of being alone), low education, low income, low education, unemployed, and under temporary housing, regardless of their age. This cluster had the smallest number of constituents.

Let us summarize the findings with respect to variables.

Age	Old clusters are likely to be wealthy, which is natural in a sense that households would accumulate wealth during working ages. However, there were also two strong exceptions (Clusters 3 and 8)
Education	The three most wealthy clusters are highly educated while the three most poor clusters are poorly educated. For the two middle class groups, one in metropolitan area (Cluster 5) is highly educated and the other outside metropolitan area (Cluster 4) is poorly educated. Also, Cluster 3 is highly educated but has low income
Income	The two most wealthy clusters have high income, and the three most poor clusters have low income. However, three clusters in the middle exhibit mixed results (especially Cluster 3)



	1	2	3	4	5	6	7	8
1	0.44	0.18	0.13	0.07	0.06	0.06	0.04	0.02
2	0.10	0.48	0.12	0.13	0.05	0.06	0.04	0.01
3	0.08	0.12	0.47	0.16	0.06	0.06	0.04	0.01
4	0.05	0.15	0.15	0.47	0.04	0.07	0.05	0.02
5	0.07	0.11	0.12	0.08	0.43	0.12	0.05	0.02
6	0.06	0.10	0.11	0.10	0.10	0.42	0.10	0.01
7	0.06	0.10	0.10	0.09	0.06	0.17	0.36	0.06
8	0.06	0.09	0.10	0.09	0.07	0.09	0.15	0.36

Fig. 9 Transition matrix between household clusters

Number of family members	While there is no clear linear relationship between family size and wealth, it is interesting to note that the wealthiest and the poorest clusters are highly likely to consist of small families
Area of residence	No overall trend is found, but typical rural–urban differences can be seen between the two middle class groups (Clusters 4 and 5)

Previous studies have focused on finding a linear relationship between two variables. For example, researchers have reported the existence of a linear relationship between income and wealth (Lee et al., 2020), between education and wealth (Brückner & Gradstein, 2013; Boshara et al., 2015), and the absence of a linear relationship between income and wealth (Mueller, Buchholz, & Blossfeld, 2011). However, our results show that even if there is an overall trend between two variables, there is always a strong exception, making the relationship non-linear. Hence, considering multiple variables is crucial for understanding the complex relationship between financial and sociodemographic variables.

4.3 Mobility between clusters

We analyze the mobility between clusters by tracking the cluster movements of households who participated in the survey multiple times. From 2017 to 2020, clusters of 12,272 households out of 52,920 total respondent households changed. Figure 9 shows the transition matrix of the clusters. The number in cell (i, j) represents the probability of a household moving from cluster i to cluster j in the next survey.

Some block-diagonal shapes can be observed. Two large blocks can be seen within Clusters 1, 2, 3, 4 and within Clusters 5, 6, 7, 8. That is, not many households move from the wealthy groups to the non-wealthy groups and vice versa, which indicates that there are two separate classes that are not reachable to each other in a few years of term. It is interesting to note that Clusters 1, 2, 3, 4 mostly own their houses and Clusters 5, 6, 7, 8 do not.

In addition, there are small blocks between most adjacent clusters (e.g., between clusters 1–2, 3–4, 5–6, 6–7, 7–8). However, we can find another weak link between Clusters 2 and 3. Recall that the key difference between the two clusters was that Cluster 2 had a substantial amount of nonresidential real estate, but Cluster 3 had almost none. Therefore, real estate is not only a crucial factor for classifying households, but also a huge hurdle for households who wish to climb up the class ladder.

4.4 Out-of-sample analysis after COVID-19

Lastly, we show the out-of-sample results using the survey data in 2021, which is mostly based on financial activities of households in 2020. Hence, it will allow us to see the changes after COVID-19 pandemic.

Figure 10 represents the variable importance weights of between-group inequalities of Gini coefficients of asset and debt (mortgage and credit loans) variables. We can see from the figure that after COVID-19, between-group inequalities are decreased in asset variables, but they are increased in debt variables. This means that the changes in debt after COVID-19 are quite different for different clusters, while changes in assets would not. Hence, we can see that the impact of COVID-19 was quite asymmetric for household debt, but it was relatively even for household assets. This makes sense because COVID-19 caused immediate damage to the income of households who have their own business (e.g., restaurants or coffee shops), and many of them had to obtain additional loans.

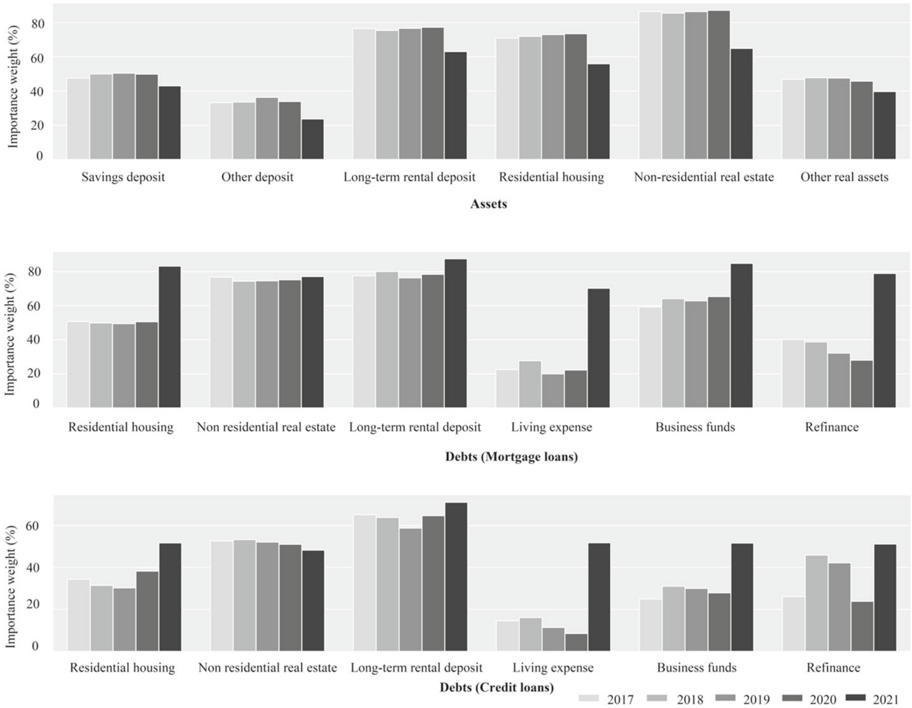
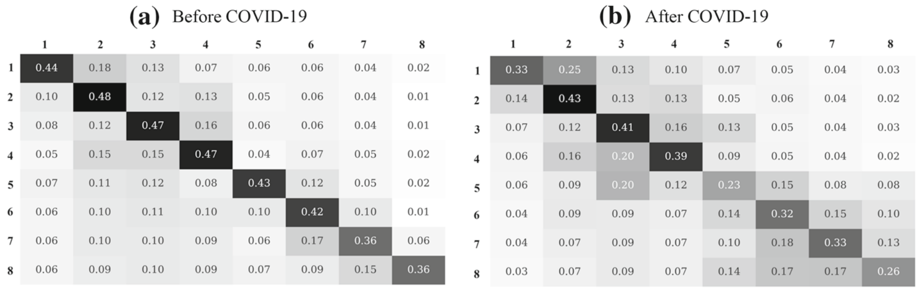


Fig. 10 Variable importance weights of between-group inequalities of Gini coefficients in different years



(a) Before COVID-19

	1	2	3	4	5	6	7	8
1	0.44	0.18	0.13	0.07	0.06	0.06	0.04	0.02
2	0.10	0.48	0.12	0.13	0.05	0.06	0.04	0.01
3	0.08	0.12	0.47	0.16	0.06	0.06	0.04	0.01
4	0.05	0.15	0.15	0.47	0.04	0.07	0.05	0.02
5	0.07	0.11	0.12	0.08	0.43	0.12	0.05	0.02
6	0.06	0.10	0.11	0.10	0.10	0.42	0.10	0.01
7	0.06	0.10	0.10	0.09	0.06	0.17	0.36	0.06
8	0.06	0.09	0.10	0.09	0.07	0.09	0.15	0.36

(b) After COVID-19

	1	2	3	4	5	6	7	8
1	0.33	0.25	0.13	0.10	0.07	0.05	0.04	0.03
2	0.14	0.43	0.13	0.13	0.05	0.06	0.04	0.02
3	0.07	0.12	0.41	0.16	0.13	0.05	0.04	0.03
4	0.06	0.16	0.20	0.39	0.09	0.05	0.04	0.02
5	0.06	0.09	0.20	0.12	0.23	0.15	0.08	0.08
6	0.04	0.09	0.09	0.07	0.14	0.32	0.15	0.10
7	0.04	0.07	0.09	0.07	0.10	0.18	0.33	0.13
8	0.03	0.07	0.09	0.07	0.14	0.17	0.17	0.26

Fig. 11 Transition matrix between household clusters before (left) and after (right) COVID-19

Next, we investigate the change in mobility between clusters. Figure 11 compares the transition matrix before and after COVID-19. It is clear that the mobility is increased after COVID-19, because every diagonal term became smaller (i.e., probabilities of staying in the same cluster are reduced).

$$\frac{\text{Average probability of moving into a poorer cluster}}{\text{Average probability of moving into awealthier cluster}}$$

However, if we look into the above ratio[5] for the transition matrix, there is a significant difference between the wealthy half (Clusters 1,2,3,4) and the non-wealthy half (Clusters 5,6,7,8). Before COVID-19, the average of the above ratio for the transition matrix for the wealthy half and the non-wealthy half was 0.621 and 0.653, respectively. After COVID-19, however, they become 0.639 and 1.151. While the direction of cluster mobility for the wealthy half was not affected by COVID-19, it is clear that the probability of the non-wealthy half going into poorer clusters became much higher. Hence, we can see that COVID-19 had a much greater adverse impact for the non-wealthy half than the wealthy half.

5 Conclusion

This study has shown how advanced clustering techniques, especially that involve deep learning models, can be useful for understanding the complex heterogeneity of household finance. By utilizing a deep learning-based clustering N2D framework proposed by McConville et al. (2021), we were able to efficiently handle high-dimensional data to find representative clusters. More specifically, we could capture and decompose the nonlinear relationships in data through deep clustering, whereas conventional age or income groups could not.

The key implication of this study is that various variables should be considered together to analyze household heterogeneity. For example, real estate ownership was shown to be critical for the broad classification of wealthy and non-wealthy Korean households. Within the wealthy group, nonresidential real estate was shown to be the next key factor, while credit loans were found to be important explanatory variables for further classifications within the non-wealthy group. We used the Gini coefficients and their decompositions to further verify

[5] For the wealthy half (Clusters 1,2,3,4) before COVID-19, the numerator would be the average of the values on the right side of the first four diagonal values (0.18 + 0.13 + 0.07 + … + 0.04 + 0.07 + 0.05 + 0.02) / 22 = 0.0673. On the other hand, the denominator would be the average of the values on the left side of the first four diagonal values (0.10 + 0.08 + 0.12 + 0.05 + 0.15 + 0.15) / 6 = 0.1083. Hence, the ratio becomes 0.0673 / 0.108 = 0.621. The other ratios can be calculated similarly.

the quality of clustering and the relative importance of the variables. In addition, the multidimensional heterogeneity of households was shown to be closely related to sociodemographic variables, and the relationships were non-linear.

Since this study was conducted based on Korean household data, detailed findings should be interpreted carefully and might not be directly applicable to households in different countries. Hopefully, however, our study will encourage other researchers to search for more multidimensional aspects of household heterogeneity. Such findings are crucial for developing more accurate macroeconomic models with heterogeneous agents and deriving appropriate economic policies.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1C1C1010456). Fabozzi was funded by EDHEC Business School for the 2020-2021 academic year.

Appendix A: Hyperparameters for deep clustering

For autoencoder, we used a three-layer fully connected networks with rectified linear units (ReLU). While there are many different network architectures and activation functions, we have chosen one of the simplest forms to mitigate the architecture specific results. Note that ReLU is known to be more appropriate for sparse data compared to the sigmoid function (Glorot et al., 2011), and households' debt data are very sparse since there are many households who do not have any or some type of loan. In addition, we set the dropout rate as 0.05 for all layers for the robustness of the model.

Table 6 shows the range of hyperparameters we used to train deep clustering algorithm (autoencoder and UMAP). To find the optimal parameters, we used a random search approach (Bergstra & Bengio, 2012) by randomly sampling 200 models within the range. Note that the number of all combination is $4 \times 4 \times 4 \times 3 \times 2 = 384$, and thus, the random search covers more than 50%.

We have chosen the best model configuration among 200 randomly sampled model settings with respect to the Silhouette index and Davis-Bouldin index. Here, we show that our results are not restricted to this particular choice. In Fig. 12, we compare the box plot of Euclidean distance from the cluster centers of the best model for top 20 model configurations and the whole 200 random samples. To be more specific, we ordered clusters in terms of their average wealth for each model setting. Then, for a model configuration, we would have clusters 1 to 8. Next, we calculated the distance between the cluster $\boldsymbol{i}$ in each model configuration and the cluster $\boldsymbol{i}$ in the best model configuration, and we took the summation for all $\boldsymbol{i}$. In the figure, it is clear that the top 20 models have cluster centers that are very much close to the cluster centers from the best model. Hence, it means that the results shown in Sect. 4 would

Table 6 Hyperparameter search range

Hyperparameter	Range
Batch size	[16, 24, 32, 64]
Learning rate	[0.0001, 0.001, 0.001, 0.01]
Epochs	[20, 50, 75, 100]
# of nodes in each layer	[10–1000, 10–1000, 10–1000]
α, β(UMAP)	[0.9 ~ 1, 0.5–1.0]

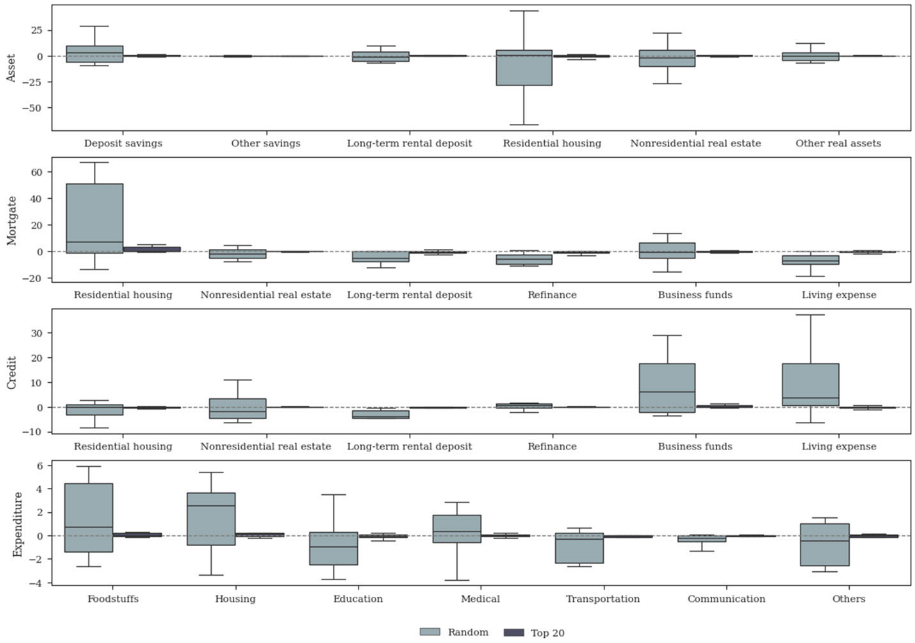


Fig. 12 Euclidian distance between the models

not change much even if we choose another model configuration within the top 20 model configurations with respect to the Silhouette index and Davies-Bouldin index.

Appendix B. Average portfolio weights of different income and age groups

See Figs. 13, 14, and 15.

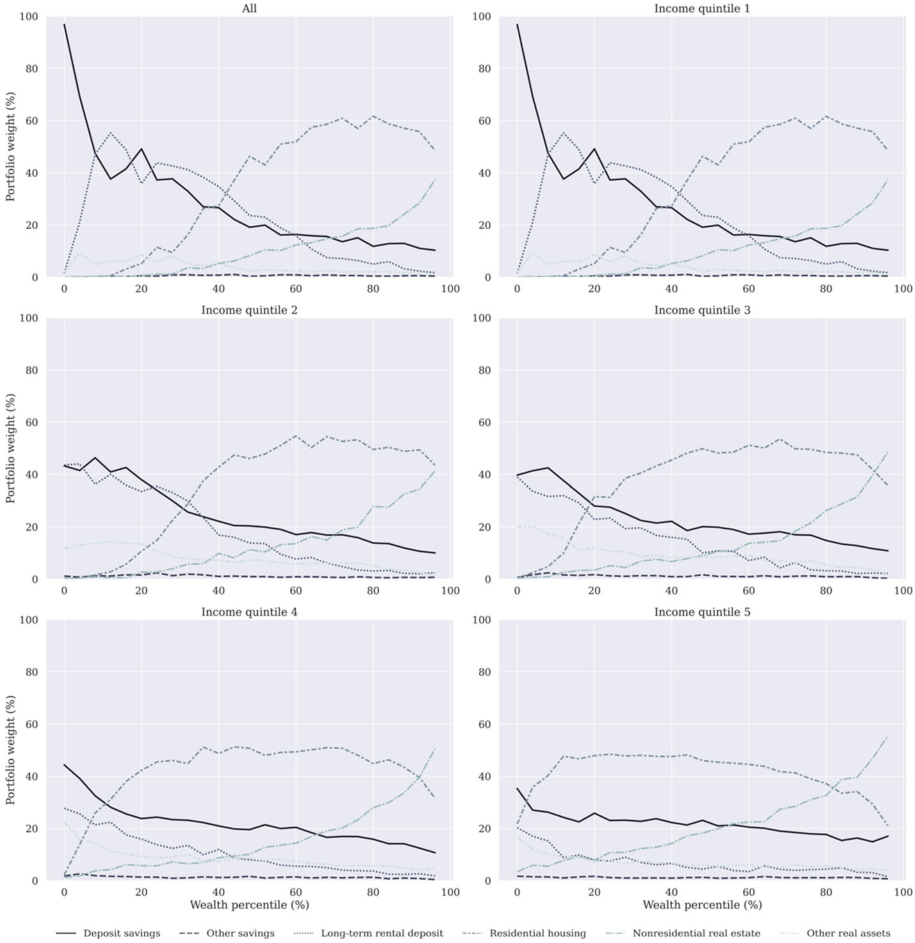


Fig. 13 Average portfolio weights of different income quintiles

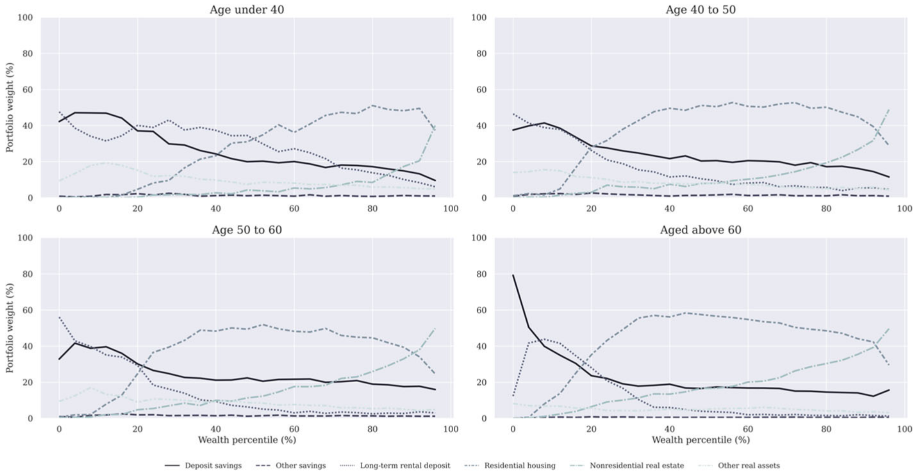


Fig. 14 Average portfolio weights of different age groups

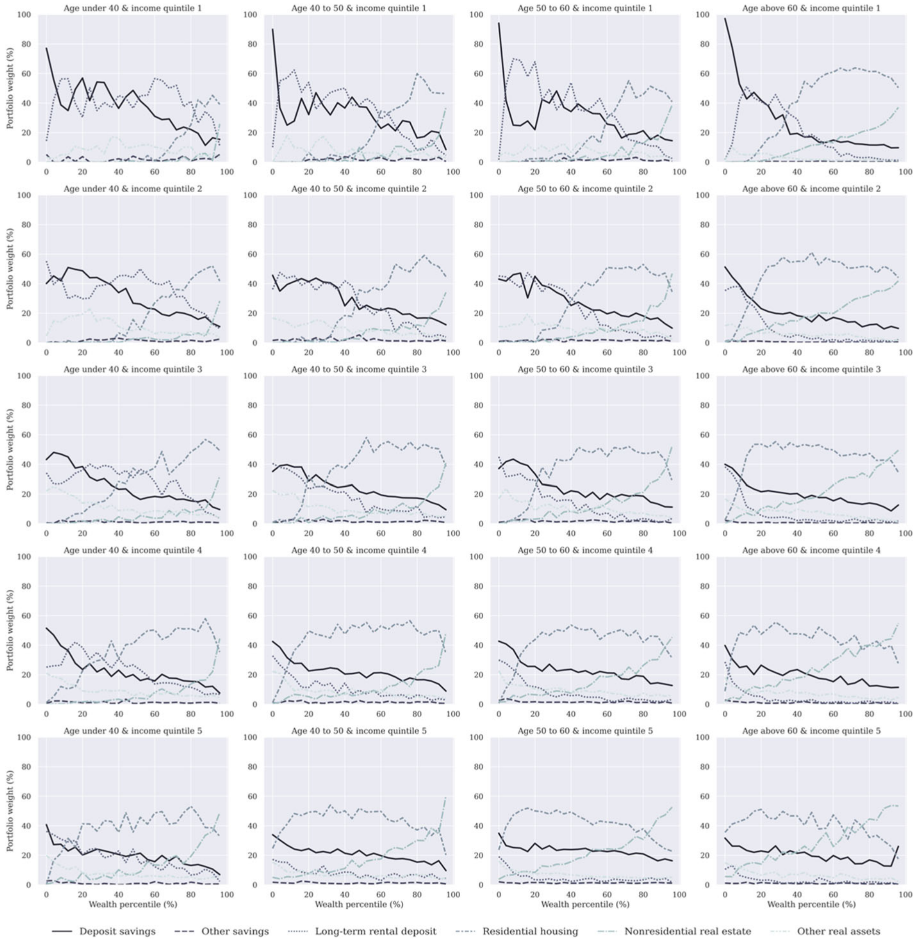


Fig. 15 Average portfolio weights of different income and age groups

Appendix C. Summary statistics of household balance sheet of clusters

See Table 7.

Table 7 Summary statistics of household balance sheet of clusters

Panel A. Assets												
No	Deposit savings		Other savings		Long-term rental deposit		Residential housing		Nonresidential real estate		Other real assets	
	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median
1	0.105 (0.128)	0.061	0.003 (0.019)	0.000	0.007 (0.039)	0.000	0.390 (0.303)	0.330	0.461 (0.294)	0.491	0.028 (0.068)	0.007
2	0.178 (0.159)	0.132	0.012 (0.047)	0.000	0.032 (0.091)	0.000	0.304 (0.226)	0.313	0.382 (0.242)	0.373	0.080 (0.137)	0.030
3	0.143 (0.139)	0.106	0.006 (0.029)	0.000	0.000 (0.010)	0.000	0.813 (0.157)	0.845	0.004 (0.020)	0.000	0.036 (0.056)	0.014
4	0.229 (0.157)	0.213	0.008 (0.032)	0.000	0.004 (0.029)	0.000	0.613 (0.246)	0.650	0.074 (0.175)	0.000	0.072 (0.087)	0.0447
5	0.254 (0.249)	0.173	0.010 (0.050)	0.000	0.654 (0.271)	0.712	0.000 (0.002)	0.000	0.015 (0.069)	0.000	0.043 (0.063)	0.017
6	0.467 (0.312)	0.441	0.002 (0.103)	0.000	0.359 (0.316)	0.297	0.001 (0.019)	0.000	0.005 (0.039)	0.000	0.130 (0.182)	0.474
7	0.418 (0.348)	0.327	0.011 (0.067)	0.000	0.389 (0.355)	0.322	0.004 (0.054)	0.000	0.003 (0.045)	0.000	0.174 (0.262)	0.000

Table 7 (continued)

Panel A. Assets

No	Deposit savings		Other savings		Long-term rental deposit		Residential housing		Nonresidential real estate		Other real assets	
	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median
8	0.683 (0.408)	1.000	0.003 (0.042)	0.000	0.273 (0.401)	0.000	0.000 (0.021)	0.000	0.001 (0.029)	0.000	0.037 (0.147)	0.000

No	Residential housing		Nonresidential real estate		Long-term rental deposit		Living expense		Business funds		Refinance	
	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median
Panel B-1. Debts (Mortgage loans)												
1	0.295 (0.432)	0.000	0.328 (0.446)	0.000	0.017 (0.125)	0.000	0.042 (0.192)	0.000	0.229 (0.406)	0.000	0.021 (0.134)	0.000
2	0.341 (0.450)	0.000	0.270 (0.421)	0.000	0.036 (0.178)	0.000	0.052 (0.211)	0.000	0.228 (0.407)	0.000	0.021 (0.139)	0.000
3	0.814 (0.375)	1.000	0.004 (0.060)	0.000	0.006 (0.078)	0.000	0.060 (0.230)	0.000	0.043 (0.195)	0.000	0.021 (0.136)	0.000
4	0.669 (0.456)	1.000	0.032 (0.168)	0.000	0.011 (0.099)	0.000	0.098 (0.287)	0.000	0.100 (0.293)	0.000	0.030 (0.168)	0.000
5	0.003 (0.059)	0.000	0.012 (0.103)	0.000	0.868 (0.328)	0.000	0.054 (0.220)	0.000	0.027 (0.157)	0.000	0.008 (0.088)	0.000

Table 7 (continued)

No	Residential housing		Nonresidential real estate		Long-term rental deposit		Living expense		Business funds		Refinance	
	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median
6	0.002 (0.044)	0.000	0.007 (0.083)	0.000	0.439 (0.478)	0.000	0.288 (0.442)	0.000	0.133 (0.332)	0.000	0.041 (0.195)	0.000
7	0.024 (0.153)	0.000	0.004 (0.069)	0.000	0.186 (0.388)	0.000	0.428 (0.488)	0.000	0.175 (0.379)	0.000	0.038 (0.187)	0.000
8	0.000 (0.000)	0.000	0.000 (0.000)	0.000	0.000 (0.000)	0.000	0.000 (0.000)	0.000	0.000 (0.000)	0.000	0.000 (0.000)	0.000
Panel B-2. Debts (credit loans)												
1	0.054 (0.225)	0.000	0.124 (0.324)	0.000	0.016 (0.122)	0.000	0.323 (0.462)	0.000	0.375 (0.479)	0.000	0.022 (0.146)	0.000
2	0.061 (0.236)	0.000	0.110 (0.305)	0.000	0.037 (0.185)	0.000	0.316 (0.454)	0.000	0.327 (0.462)	0.000	0.030 (0.165)	0.000
3	0.130 (0.329)	0.000	0.013 (0.113)	0.000	0.015 (0.121)	0.000	0.497 (0.491)	0.333	0.156 (0.356)	0.000	0.040 (0.191)	0.000
4	0.081 (0.267)	0.000	0.021 (0.141)	0.000	0.010 (0.099)	0.000	0.506 (0.489)	0.500	0.178 (0.377)	0.000	0.046 (0.203)	0.000
5	0.018 (0.133)	0.000	0.049 (0.211)	0.000	0.294 (0.446)	0.000	0.343 (0.465)	0.000	0.104 (0.301)	0.000	0.025 (0.148)	0.000
6	0.006 (0.080)	0.000	0.012 (0.108)	0.000	0.132 (0.330)	0.000	0.419 (0.478)	0.000	0.232 (0.413)	0.000	0.054 (0.218)	0.000
7	0.001 (0.003)	0.000	0.001 (0.036)	0.000	0.092 (0.281)	0.000	0.436 (0.480)	0.000	0.168 (0.366)	0.000	0.109 (0.301)	0.000

Table 7 (continued)

No	Residential housing		Nonresidential real estate		Long-term rental deposit		Living expense		Business funds		Refinance	
	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median
8	0.000 (0.000)	0.000	0.000 (0.000)	0.000	0.000	0.000	0.430 (0.498)	0.152	0.000 (0.000)	0.000	0.000 (0.000)	0.000

Panel C. Expenditures

No	Foodstuffs		Housing		Education		Medical		Transportation		Communication		Others	
	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median	Mean (Std dev)	Median
1	0.335 (0.131)	0.323	0.157 (0.090)	0.136	0.033 (0.090)	0.000	0.098 (0.103)	0.064	0.094 (0.068)	0.080	0.060 (0.036)	0.054	0.219 (0.124)	0.200
2	0.296 (0.115)	0.283	0.121 (0.082)	0.098	0.081 (0.126)	0.000	0.084 (0.105)	0.046	0.110 (0.070)	0.097	0.063 (0.033)	0.057	0.241 (0.121)	0.226
3	0.338 (0.125)	0.326	0.139 (0.079)	0.118	0.060 (0.111)	0.000	0.089 (0.104)	0.057	0.093 (0.065)	0.059	0.066 (0.036)	0.059	0.213 (0.118)	0.196
4	0.302 (0.116)	0.290	0.110 (0.079)	0.088	0.096 (0.127)	0.000	0.087 (0.120)	0.041	0.106 (0.066)	0.095	0.066 (0.033)	0.061	0.230 (0.119)	0.215
5	0.337 (0.127)	0.322	0.146 (0.101)	0.115	0.062 (0.111)	0.000	0.064 (0.083)	0.033	0.095 (0.066)	0.065	0.073 (0.040)	0.065	0.220 (0.125)	0.205
6	0.310 (0.119)	0.297	0.203 (0.131)	0.176	0.053 (0.099)	0.000	0.063 (0.091)	0.029	0.095 (0.069)	0.081	0.073 (0.040)	0.066	0.200 (0.118)	0.181
7	0.341 (0.126)	0.297	0.235 (0.131)	0.217	0.037 (0.086)	0.000	0.070 (0.104)	0.031	0.087 (0.068)	0.069	0.072 (0.042)	0.063	0.181 (0.112)	0.163
8	0.381 (0.141)	0.376	0.234 (0.128)	0.215	0.009 (0.047)	0.000	0.121 (0.141)	0.070	0.054 (0.048)	0.041	0.055 (0.040)	0.044	0.142 (0.103)	0.118

References

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access, 7*, 31883–31902.

Ahn, S., Kaplan, G., Moll, B., Winberry, T., & Wolf, C. (2018). When inequality matters for macro and macro matters for inequality. *NBER Macroeconomics Annual, 32*(1), 1–75.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data, 8*(1), 1.

Atkinson, A. B., Piketty, T., & Saez, E. (2011). Top incomes in the long run of history. *Journal of Economic Literature, 49*(1), 3–71.

Bengio, Y., Courville, A. C., & Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. Technical Report arXiv:1206.5538, U. Montreal (2012). Available at http://arxiv.org/abs/1206.5538

Bengio, Y., Courville, A. C., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(2), 281–305.

Berton, F., Mocetti, S., Presbitero, A. F., & Richiardi, M. (2018). Banks, firms, and jobs. *The Review of Financial Studies, 31*(6), 2113–2156.

Boshara R, Emmons, W. R., & Noeth, B. J. (2015). The demographics of wealth. Available at http://www.stlouisfed.org/household-financial-stability/the-demographics-of-wealth. *Federal Reserve Bank of St. Louis*

Bricker, J., Krimmel, J., & Ramcharan, R. (2021). Signaling status: The impact of relative income on household consumption and financial decisions. *Management Science, 67*(4), 1993–2009.

Brückner, M., & Gradstein, M. (2013). *Income and schooling*. (No. DP9365) CEPR discussion papers. Available at SSRN. Available at https://ssrn.com/abstract=2224290

Burges, C. J. (2010). *Dimension reduction: A guided tour*. Now Publishers Inc.

Calvet, L. E., Campbell, J. Y., Gomes, F., & Sodini, P. (2021). *The cross-section of household preferences. (No, w. 28788).* National Bureau of Economic Research.

Campbell, J. Y. (2006). Household finance. *The Journal of Finance, 61*(4), 1553–1604.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In. *Lecture notes in computer science.* Proceedings of the European conference on computer vision (ECCV) (pp. 139–156)

Case, K. E., Quigley, J. M., & Shiller, R. J. (2005). Comparing wealth effects: The stock market versus the housing market. *The B.E. Journal of Macroeconomics*, *5*(1), 1–34.

Case, K. E., Quigley, J. M., & Shiller, R. J. (2011). *Wealth effects revisited 1978–2009*. *National Bureau of Economic Research* (No, w. 16848)

Constantinides, G. M., & Duffie, D. (1996). Asset pricing with heterogeneous consumers. *Journal of Political Economy, 104*(2), 219–240.

Das, S., Kuhnen, C. M., & Nagel, S. (2020). Socioeconomic status and macroeconomic expectations. *The Review of Financial Studies, 33*(1), 395–432.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*, 224–227.

Deaton, A., & Paxson, C. (1994). Intertemporal choice and inequality. *Journal of Political Economy, 102*(3), 437–467.

Deaton, A. S., & Paxson, C. H. (1997). The effects of economic and population growth on national saving and inequality. *Demography, 34*(1), 97–114.

Dizioli, A., & Pinheiro, R. (2021). Information and inequality in the time of a pandemic. *Journal of Economic Dynamics and Control, 130*, 104202.

Donoho, D. L., & Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America, 100*(10), 5591–5596.

Eichenbaum, M. S., Rebelo, S., & Trabandt, M. (2021). *Inequality in life and death. (No, w. 29063)*. National Bureau of Economic Research.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96).* AAAI Press. Pp. 226–231.

Fredriksen, K. B. (2012). *Less income inequality and more growth-are they compatible? Part 6. The distribution of wealth*

Ghasedi Dizaji, K., Herandi, A., Deng, C., Cai, W., & Huang, H. (2017). Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 5736–5745)

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315–323). JMLR workshop and conference proceedings.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*

Guo, X., Gao, L., Liu, X., & Yin, J. (2017). Improved deep embedded clustering with local structure preservation. In *International joint conference on artificial intelligence (IJCAI)*, pp. 1753–1759

Heaton, J., & Lucas, D. (1997). Market frictions, savings behavior, and portfolio choice. *Macroeconomic Dynamics, 1*(1), 76–101.

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

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization. arXiv preprint* arXiv:1412.6980

Krueger, D., Mitman, K., & Perri, F. (2016). Macroeconomics and household heterogeneity. In Taylor, J.B. & Uhlig, H. (eds) *Handbook of macroeconomics* (Vol. 2, pp. 843–921)1 Edn., Elsevier

Krueger, D., & Perri, F. (2006). Does income inequality lead to consumption inequality? Evidence and theory. *Review of Economic Studies, 73*(1), 163–193.

Krusell, P., & Smith, A. A. (1997). Income and wealth heterogeneity, portfolio choice, and equilibrium asset returns. *Macroeconomic Dynamics, 1*(2), 387–422.

Lee, K.-Y., & Jun, H.-J. (2018). Determinants of housing affordability among renters and homeowners: Comparison between the capital and non-capital regions. *Journal of Korea Planning Association, 53*(4), 143–161.

Lee, S. K., Shin, H. J., & Kim, C. H. (2020). Inequality of the household income and wealth in Korea: Research outcome and agenda. *Economy and Society, 127*, 60–94.

Lucas, D. J. (1994). Asset pricing with undiversifiable income risk and short sales constraints. *Journal of Monetary Economics, 34*(3), 325–341.

Mankiw, N. G., & Zeldes, S. P. (1991). The consumption of stockholders and nonstockholders. *Journal of Financial Economics, 29*(1), 97–112.

McConville, R., Santos-Rodriguez, R., Piechocki, R. J., & Craddock, I. (2021). (Not too) deep clustering via clustering the local manifold of an autoencoded embedding. In 2020. N2d 25th *international conference on* pattern recognition *(ICPR)*, pp. 5145–5152. IEEE

McInnes, L., Healy, J., & Melville, J. (2018). *Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint* arXiv:1802.03426

Mookherjee, D., & Shorrocks, A. (1982). A decomposition analysis of the trend in UK income inequality. *The Economic Journal, 92*(368), 886–902.

Mueller, N., Buchholz, S., & Blossfeld, H. P. (2011). *Wealth inequality in Europe and the delusive egalitarianism of Scandinavian countries*. University of Bamberg.

Mukherjee, S., Asnani, H., Lin, E., & Kannan, S. (2019). Clustergan: Latent space clustering in generative adversarial networks. In. *Proceedings of the AAAI conference on artificial intelligence.*, 33, 4610–4617 (Vol. 33, No. 01, pp. 4610–4617)

OECD, A. (2018). *A broken social elevator? How to promote social mobility. COPE Policy Brief*

OECD (2021). *Fertility rates (indicator)*. https://doi.org/10.1787/8272fb01-en

Park, C. G. (2020). Long-term trends in the Korean financial sector and Covid-19, *Korea Capital Market Institute (KCMI) Issue Report*, 20–22

Piketty, T. (2013). *Capital in the 21st Century*. President and Fellows, Harvard College.

Pyatt, G. (1976). On the interpretation and disaggregation of Gini coefficients. *The Economic Journal, 86*(342), 243–255.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., et al. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*, 664–681.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS), 42*(3), 1–21.

Shorrocks, A. F. (1982). Inequality decomposition by factor components. *Econometrica, 50*(1), 193–211.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319–2323.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(11), 2579–2605.

World Bank, *World Development Indicators* (2021). GDP (current US$) [Data file]. Available at https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Xia, W., Zhang, Y., Yang, Y., Xue, J. H., Zhou, B., & Yang, M. H. (2021). *Gan inversion: A survey. arXiv preprint* arXiv:2101.05278

Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning,* 478–487.

Zhang, Z., & Wang, J. (2007). MLLE: Modified locally linear embedding using multiple weights. In *Advances in Neural Information Processing Systems*, 1593–1600

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.