Unsupervised Learning for Large Scale Data: The ATHLOS Project

Petros Barmpas[a], Sotiris Tasoulis[a], Aristidis G. Vrahatis[a], Panagiotis Anagnostou[a], Spiros Georgakopoulos[a], MatthewPrina[b,c], José Luis Ayuso-Mateos[d,e,f], Jerome Bickenbach[g,h], Ivet Bayes[m,d], Martin Bobak[i],Francisco Félix Caballero[j,k], Somnath Chatterji[l], Laia Egea-Cortés[m], Esther García-Esquinas[j,k],Matilde Leonardi[n], Seppo Koskinen[o], Ilona Koupil[p,q], Andrzej Pająk[r], Martin Prince[c,s], WarrenSanderson[t,u], Sergei Scherbov[t,v,w], Abdonas Tamosiunas[x], Aleksander Galas[y], Josep MariaHaro[m,d], Albert Sanchez-Niubo[m,d], Vassilis P. Plagianakos[a], Demosthenes Panagiotakos[z]

[a]Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece.

[b]Social Epidemiology Research Group. Health Service and Population Research Department, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.

[c]Global Health Institute, King's College London, London, UK.

[d]Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM, Madrid, Spain.

[e]Department of Psychiatry, Universidad Autónoma de Madrid, Madrid, Spain.

[f]Hospital Universitario de La Princesa, Instituto de Investigación Sanitaria Princesa (IIS Princesa), Madrid, Spain.

[g]Swiss Paraplegic Research, Guido A. Zäch Institute (GZI), Nottwil, Switzerland.

[h]Department of Health Sciences & Health Policy, University of Lucerne, Lucerne, Switzerland.

[i]Department of Epidemiology and Public Health, University College London, London, UK.

[j]Department Preventive Medicine and Public Health, Universidad Autónoma de Madrid/Idipaz, Madrid, Spain.

[k]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, CIBERESP, Madrid, Spain.

[l]Information, Evidence and Research, World Health Organization, Geneva, Switzerland.

[m]Research, Innovation and Teaching Unit. Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, Spain.

[n]Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy.

[o]National Institute for Health and Welfare (THL), Helsinki, Finland.

[p]Centre for Health Equity Studies, Department of Public Health Sciences, Stockholm University, Stockholm, Sweden.

[q]Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden.

[r]Department of Epidemiology and Population Studies, Jagienllonian University, Krakow, Poland.

[s]Centre for Global Mental Health. Health Service and Population Research Department, Institute of Psychiatry, Psychology& Neuroscience, King's College London, London, UK.

[t]International Institute for Applied Systems Analysis, World Population Program, Wittgenstein Centre for Demography and Global Human Capital, Luxemburg, Austria.

[u]Department of Economics, Stony Brook University, Stony Brook, NY, United States of America.

[v]Austrian Academy of Science, Vienna Institute of Demography, Vienna, Austria.

[w]Russian Presidential Academy of National Economy and Public Administration (RANEPA), Moscow, Russian Federation.

[x]Lithuanian University of Health Sciences, Kaunas, Lithuania.

[y]Department of Epidemiology and Preventive Medicine, Jagiellonian University, Krakow, Poland..

[z]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece.

1 Abstract

Recent technological advancements in various domains, such as the biomedical and health, offer a plethora of big data for analysis. Part of this data pool is the experimental studies that record various and several features for each instance. It creates datasets having very high dimensionality with mixed data types, with both numerical and categorical variables. On the other hand, unsupervised learning has shown to be able to assist in high-dimensional data, allowing the discovery of unknown patterns through clustering, visualization, dimensionality reduction, and in some cases, their combination. This work highlights unsupervised learning methodologies for large-scale, high-dimensional data, providing the potential of a unified framework that combines the knowledge retrieved from clustering and visualization. The main purpose is to uncover hidden patterns in a high-dimensional mixed dataset, which we achieve through our application in a complex, real-world dataset. The experimental analysis indicates the existence of notable information exposing the usefulness of the utilized methodological framework for similar high-dimensional and mixed, real-world applications.

2 Introduction

In recent years, the embracement of technology in various domains has experienced remarkable growth, and as a result, the data generation has increased significantly and is expected to multiply in the near future (Alharthi, Krotov and Bowman 2017) (Gupta and Rani 2019) (Gu, et al. 2017) . Simultaneously, utilizing the innovations in the collection, recording, and storage, huge databases are being constructed, including various data types from different sources (Roh, Heo and Whang 2019) (Plageras, et al. 2018) (Wang, Ng and Brook 2020). However, along with

the vast increase in size, it is not easy to simultaneously produce and categorize them, thus leading to heterogeneity phenomena. Therefore, advanced processing techniques that allow the management of such data become necessary, aiming to retrieve patterns, reduce dimensionality and at the same time automate the processing (Hsu and Glass 2018) (Usama, et al. 2019) (Wang, et al. 2019). This development constantly reinforces the “Big Data” term, enabling the research community to enhance insight, decision-making, and process automation while simultaneously necessitating cost-effective, novel means of information processing (Sagiroglu and Sinanc 2013). The arising Big Data characteristics though may weaken prediction accuracy and pattern discovery by imposing noise through measurement errors and false correlations (Fan, Han and Liu 2014).

Diverse dimensionalities and heterogeneous structure are two key features that intensify large-scaled data complexity. When we consider massive heterogeneous data (Zhu, et al. 2018), the features may represent different types of information of the same individual (Fan and Fan 2008). These issues become significant challenges when trying to enable data aggregation (Fan and Fan 2008) (Fan, Guo and Hao 2012) (Li, et al. 2016).

In an early stage of data centralized information systems, the focus is on finding the best feature values to represent each observation (Tufekci 2014) (Keyes and Westreich 2019). This type of sample feature representation inherently treats each individual as an independent entity without considering their social connections. In a dynamic world, the features used to represent the individuals and the social ties used to represent their connections may also evolve concerning temporal, spatial, and other factors. Such complexity is becoming part of Big Data applications' reality, introducing significant computational challenges for data analytics (Jin, et al. 2015) (Marx 2013).

Nevertheless, techniques and methods that fall into the category of unsupervised learning

have shown encouraging results and are able to overcome some of the difficulties of vast, heterogeneous data (Casolla, et al. 2019) (Hameed, et al. 2018) (Ma, et al. 2017) (Xiang, et al. 2018). In this study, we focus on the area of unsupervised learning, presenting a complete methodological procedure that utilizes recent advances in the field. We begin with a review of state of the art methods for clustering and dimensionality reduction and conclude with their utilization in a real world dataset, characterized by the aforementioned challenges. The organization of this paper is as follows. Chapter 3 presents the basic principles and characteristics of unsupervised learning and introduces some indicative techniques for two subcategories, dimensional reduction, and clustering respectivelly. Chapter 4 presents the methodology followed for the experimental procedure, in an attempt to extract knowledge from the data set at hand, that of the ATHLOS Cohort. Finally, in chapter 5, we conclude with a discussion upon the results and potential perspectives.

3 Unsupervised Learning Methods as an Approach for Knowledge Extraction

Unsupervised learning is one of the main categories of Machine Learning, along with supervised and reinforcement learning (Hinton, Sejnowski, et al. 1999), and hybrid methods like semi-supervised learning (Zhu and Goldberg 2009) (X. J. Zhu 2005). Unsupervised learning can be divided into three main application fields (Ghahramani 2003). The first one concerns data samples segmentation by some shared attributes, next is the outlier detection (Both can be attributed to Clustering methods (Jiang and An 2008) (Chawla and Gionis 2013)), while the last is dataset simplification by aggregating variables with similar attributes, a procedure known as Dimensionality Reduction accompanied with Feature Selection (Wei and Billings 2006) (Masaeli, Fung and Dy 2010) (Mladenić 2005). In summary, Unsupervised Learning aims to study the intrinsic structure of the data to find patterns that should not be considered plain, unstructured

noise (Ghahramani 2003).

Each of the subcategories has the potential to extract helpful information regarding a dataset. However, the combination of them has been previously shown to produce encouraging results (Diaz-Papkovich, Anderson-Trocmé and Gravel 2019) (Allaoui, Kherfi and Cheriet 2020) (Hozumi, et al. 2021). In what follows, we highlight the some of most representative techniques from each subcategory that we also utilize on our experimental analysis. Our aim is to incorporate both well-established and recent state-of-the-art and popular methods.

3.1 Dimensionality Reduction for Pattern Discovery through Visualization

Biomedical and health technologies are constantly evolving generating ultra-high dimensional data since we have several features for each record. Sampling techniques aim to reduce the dataset's size but still do not offer a solution for high-dimensional datasets. In such cases, Dimensionality Reduction precedes Clustering procedures as a preprocessing step (Kaski 1998) (Yan, et al. 2006). Dimensionality Reduction (DR) aims to solve the Curse of Dimensionality (Bellman n.d.) depicting that when the dimensionality increases, the volume of the space increases at such a rate that the dataset becomes sparse, opposing statistical methods. The goal is to find low-dimensional representations of the data that retain their fundamental properties, typically in two or three dimensions (Ghodsi 2006) (Sorzano, Vargas and Montano 2014). As such this process is also essential for data visualization in lower dimensions (Xia, et al. 2017). Visualization tools can assist in identifying the data structure while plotting the data in two dimensions allows researchers to pinpoint any remaining technical variability source between samples, which should be removed by normalization (Rostom, et al. 2017).

Meanwhile, well-established visualization techniques that have been proved effective for small or intermediate size data face a significant challenge when applied to big and high

dimensional data. Visualizing high-dimensional data could allow the discovery of hidden relationships between the hidden variables and numeric values (Xia, et al. 2017). Although there is remarkable progress in this field, identifying an extremely low-dimensional representation of large-scale and high-dimensional data remains a major challenge. Dimensionality reduction techniques able to handle large data are presented below, either traditional, established methods or state-of-the-art approaches specifically designed for Big Data scenarios.

Principal component analysis (PCA) is probably one of the most popular multivariate statistical technique used by almost all disciplines. It is also likely the oldest multi-variable technique. Its origins date back to Pearson (Pearson 1901). PCA is a statistical process that uses a rectangular transformation to convert a set of observations of possible associated variables (entities each receiving different numeric values) into a set of values of linearly unrelated variables called main variables. This transformation is defined so that the first principal component has the highest possible variance (i.e., gives the highest volatility to the data), and each subsequent one has the next highest variance, subject to the constraint that it is rectangular with the previous components. By visualizing the two main components, the user can apprehend some of the topologies that the data have while also being assured that most of the relevant information is still preserved. PCA, though, is mainly able to produce acceptable results in linear datasets (Shah, et al. 2013). For this reason, a large number of non-linear dimensionality reduction techniques have been created to preserve the topology of the dataset better. In what follows, we present some state-of-the-art tools that have seen adoption in recent years.

Van der Maaten and Hinton proposed t-distributed stochastic neighborhood embedding (t-SNE) in 2008. Until recently, it was considered to have vast applicability and great accuracy (Kobak and Berens 2019) (Rauber, et al. 2016). This technique is an extension of the SNE as

proposed by Hinton and Roweis in 2003 (Hinton and Roweis, Stochastic neighbor embedding 2003), which is a technique that minimizes the Kullback-Leibler (Kullback 1997) deviation of the scaled similarities among pairs of points both in high and low dimensional spaces. SNE uses a Gaussian kernel to compute similarities in a high and low dimensional space. The t-Distributed Stochastic Neighborhood Embedding improves SNE by using a t-Distribution as a kernel in low dimensional space. Because of the heavy-tailed t-distribution, t-SNE maintains local neighborhoods of the data better and penalizes wrong embeddings of dissimilar points (Maaten and Hinton 2008). This property makes it especially suitable to represent clustered data and complex structures in a few dimensions. The minimization of the Kullback-Leibler divergence with respect to the points is performed using gradient descent.

Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy and Melville 2018) is a new multi-dimensional learning technique for non-linear manifolds. The UMAP algorithm is competitive with t-SNE in terms of visualization quality and, according to the authors, maintains a more comprehensive structure with superior runtime performance. Furthermore, UMAP has no computational restrictions on the embedding dimension, making it viable as a general-purpose dimension reduction technique for machine learning. What characterizes UMAP is that it uses local approximations of the dataset and then links them with fuzzy unions to construct simplicial sets representing the high-dimensional data's topological geometry.

LargeVis (Tang, et al. 2016) is another novel visualization technique. Many recent approaches like the t-SNE, as mentioned above, construct a K-nearest neighbor graph and then project the graph into the 2-d space. LargeVis follows a similar procedure. First, LargeVis produces an accurately approximated K-nearest neighbor graph from the data and then layouts the graph in the low-dimensional space but, in contrast, uses an efficient algorithm for K-nearest

neighbor graph construction and a principled probabilistic model for graph visualization. The whole procedure thus could scale to millions of high-dimensional data points. According to the authors, LargeVis outperforms state-of-the-art methods in both efficiency and effectiveness.

3.2 Unsupervised Learning through Clustering

Broadly defined clustering aims to identify subgroups (clusters) in the data that are distinguished by an appropriate measure of similarity (or regularity), without any previous knowledge about the assignment of observations to clusters or even the presence of clusters (Everitt, et al. 2011). The main goal is to group sets of objects so that samples in the same group are more similar to each other than samples in different groups. Clustering is among the most used exploratory data analysis techniques (Berkhin 2006) (Gan, Ma and Wu 2020) (Jajuga, Sokolowski and Bock 2012). The recent explosion of data availability leads to an ever-growing tendency to "let the data speak" (Cios, Pedrycz and Swiniarski 1998). However, the properties of novel data sources and the increasing size, dimensionality, and speed at which data are captured pose challenges for established methods. Applications for cluster analysis include gene sequence analysis, market research, and object recognition. In general, clustering techniques aimed at big data can be categorized into single-machine and distributed clustering algorithms (Shirkhorshidi, et al. 2014).

Partitioning clustering algorithms aim to divide the space that the data points lie on into sub-spaces that each contains a set of data points, according to a pre-specified number. Their simplicity and performance have attracted the research community's interest even in recent years proposing sophisticated variations and big data capable versions (Sreedhar, Kasiviswanath and Reddy 2017). One significant advantage of such approaches is that every set of data points has a distinct center (representative) As a result, a new point can be efficiently assigned to the

appropriate set after the fact. Usually, Partitioning methods (K-means, PAM clustering) are most suitable for finding spherical or convex clusters, meaning they work well only for compact and well-separated clusters. Moreover, they can be severely affected by the presence of noise and outliers in the data. For such cases, density-based approaches are usually employed (Gao, et al. 2016) (Hahsler, et al. 2017); however, they typically fall short in the presence of vast amounts of high dimensional data. The computational and memory requirements in Big Data scenarios are prohibitive for density-based algorithms. Simultaneously, we lose the ability to extract representatives that allow a straightforward procedure of allocating new samples into clusters.

Kmeans (Forgy 1965) algorithm is an iterative algorithm that tries to partition a dataset in *K* pre-defined discrete non-overlapping clusters where each data point belongs to only one group. K-means is a simple algorithm that has been used in a variety of fields. The goal is to make the data points belonging to the same cluster as similar as possible while keeping the clusters as distant from each other as possible. For a set of samples $(x1, x2, \dots, xn)$ in the d-dimensional space, cluster the dataset into clusters $C = \{C1, C2, \dots, Ck\}$ by minimizing the within-cluster sum of squares or variance. The algorithm assigns data points to a cluster in such a way that the total sum of the squared distances among the data points and their cluster's centroid (mean value of the data points that belong to a cluster) is minimized. The less variation there is within groups, the more similar data points are within each cluster.

Hierarchical clustering constructs hierarchies of clusters in a top-down (agglomerative) or bottom-up (divisive) fashion. The former starts from n clusters, where n stands for the number of data points, each containing a single data point and iteratively merging the clusters satisfying certain closeness measures. Divisive algorithms follow a reverse approach, starting with a single cluster containing all the data points and iteratively split existing clusters into subsets. Hierarchical

clustering algorithms have been shown to result in high-quality partitions, especially for applications involving clustering text collections. Nonetheless, their high computational requirements usually prevent their usage in big data scenarios. However, more recent advancements in both agglomerative (Murtagh and Legendre 2014) (Zhang, Zhao and Wang 2013) and divisive strategies (Sharma, López and Tsunoda 2017) (Tasoulis, et al. 2014) have exposed their broad applicability and robustness. In particular, when divisive clustering is combined with dimensionality reduction (Hofmeyr 2016) (Pavlidis, Hofmeyr and Tasoulis 2016), we can still get methods capable of indexing large data collection that allow fast sample allocation due to their tree structure.

The Normalized Cut Divisive Clustering (Ncutdc) (Hofmeyr 2016) algorithm is a computationally efficient divisive clustering algorithm relying on hyperplane separators. It generates a binary partitioning tree by recursively partitioning a dataset using a hierarchical collection of hyperplanes with low normalized cut measured across them. As in the minimum density hyperplane case, the projection pursuit problem is formulated as a minimization problem. The normalized cut (NCut) (Shi and Malik 2000) associated with a partition of X into clusters $C1, \ldots, Ck$ is expressed as,

$$\mathrm{NCut}(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \sum_{m=1}^{k} \frac{\mathrm{Cut}(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m)}{\mathrm{volume}(\mathcal{C}_m)}$$

By minimizing the normalized cut, this leads to solutions for which $Cut(Cm, X \setminus Cm)$ is small and volume (Cm) is large, for all m. Since $\mathrm{volume}(\mathcal{C}_m) = \mathrm{Cut}(\mathcal{C}_m, \mathcal{X} \setminus \mathcal{C}_m) + \sum_{i,j:\mathbf{x}_i,\mathbf{x}_j \in \mathcal{C}_m} \mathrm{similarity}(x_i, x_j)$, where the last term is the total internal similarity of points in $\mathcal{C}_m$, this implies that the similarity within clusters is high whereas the similarity between clusters is

low. However, the NCut problem is NP-hard, and, instead, a continuous relaxation of the problem known as spectral clustering (Shi and Malik 2000) (Von Luxburg 2007) is considered. This leads to a reduction in complexity, but this method remains applicable only to moderate size situations.

The Genie (Gagolewski, Bartoszuk and Cena 2016) algorithm is an alternative for the more classical, single-linkage criteria hierarchical clustering. The algorithm aims to offset the disadvantages of the single linkage scheme, that is, sensitivity to outliers, the creation of very skewed dendrograms, and consequently not reflecting the actual underlying data structure unless there are well-separated clusters. Simultaneously, to retain the single-linkage's simplicity and efficiency, the following linkage criterion, referred to as the Genie algorithm, was created. Let F be a fixed inequity measure (e.g., the Gini-index) and $g \in (0, 1]$ be some threshold. At step j: If $\mathrm{F}(c_{(n-j)}, \ldots, c_{(1)}) \leq g, c_i = |C_i^{(j)}|$, apply the original single linkage criterion, otherwise, if $\mathrm{F}(c_{(n-j)}, \ldots, c_{(1)}) > g$, restrict the search domain only to pairs of clusters such that one of them is the smallest. This modification prevents extreme increases of the chosen inequity measure and forces early merges of small clusters with others.

Finally, in density-based clustering, a cluster is a set of data objects spread in the data space over a contiguous region of high density of objects. Density-based clusters are separated from each other by contiguous regions of low density of objects. Data objects located in low-density regions are typically considered noise or outliers. Density-based clustering algorithms are able to discover arbitrary-shaped clusters, but usually suffer from increased computational costs, preventing them to be scalable. DensityPeaks (Rodriguez and Laio 2014) algorithm is a novel density-based approach. Similar to the K-medoids method, it has its basis only in the distance between data points. Like DBSCAN (Hahsler, et al. 2017) and the mean-shift process, it can detect non-spherical clusters and automatically find the correct number of clusters. As in the mean-shift method, the

cluster centers are defined as local maxima in the density of data points. However, unlike the mean-shift method, this procedure does not require embedding the data in a vector space and maximizing explicitly the density field for each data point. The algorithm assumes that cluster centers are surrounded by regions with lower local density and are relatively far away from higher local density points. For each data point, the algorithm computes two measures: its local density and its distance from samples of higher density. Both these measures depend exclusively on the intervals between data points, which are considered to satisfy the triangular inequality.

4 Experimental Analysis

To the best of the writer's knowledge, there is no singular unsupervised methodology that is able to extract all the available information from any dataset (Adam, et al. 2019). In many cases, methods from different fields are combined, resulting in more sophisticated techniques with benefits from all the components. This section presents some combinatorial methodologies and an example of such schemes in a Big Data scenario. More precisely, we will implement a complete unsupervised learning methodology for the ATHLOS cohort dataset while also utilizing very recent techniques for variable exploration.

4.1 Data Specification and Pre-processing

ATHLOS (Ageing Trajectories of Health: Longitudinal Opportunities and Synergies) is a project funded by the European Union's Horizon 2020 Research and Innovation Program, which aims to interpret aging's impact on health better. The ATHLOS project provides a harmonized dataset (Sanchez-Niubo, et al. 2019), built upon several longitudinal studies and originated from five continents. More specifically, it contains samples coming from more than 355,000 individuals who participated in 17 general population longitudinal studies in 38 countries. Based on the WHO healthy aging framework, researchers from the ATHLOS consortium reviewed measures of

functional ability in the aging cohorts and identified 47 items related to health, physical, and cognitive functioning. The consortium harmonized these 47 items into binary variables and used item-response theory modeling to generate a common measure for healthy aging across cohorts.

In this paper, we used 15 of these studies, namely the 10/66 Dementia Research Group Population-Based Cohort Study (Prina, et al. 2017), the Australian Longitudinal Study of Aging (ALSA) (Luszcz, et al. 2016), the Collaborative Research on Ageing in Europe (COURAGE) (Leonardi, et al. 2014),the ELSA (Steptoe, et al. 2013), the study on Cardiovascular Health, Nutrition and Frailty in Older Adults in Spain (ENRICA) (Rodríguez-Artalejo, et al. 2011), the Health, Alcohol and Psychosocial factors in Eastern Europe Study (HAPIEE) (Peasey, et al. 2006), the Health 2000/2011 Survey (Koskinen 2018), the HRS (Sonnega, et al. 2014), the JSTAR (Ichimura, Shimizutani and Hashimoto 2009), the KLOSA (Park, et al. 2007), the MHAS (Wong, Michaels-Obregon and Palloni 2017), the SAGE (Kowal, et al. 2012), SHARE (Börsch-Supan, et al. 2013), the Irish Longitudinal Study of Ageing (TILDA) (Whelan and Savva 2013) and the Longitudinal Aging Study in India (LASI) (Arokiasamy, et al. 2012). The 15 general population longitudinal studies utilized in this work consist of 990,000 samples in total, characterized by 184 variables. The version used here is a preprocessed dataset where a selection of variables has been removed along with several samples as described in (Anagnostou, et al. 2021). The resulting data matrix constituted by 770,764 samples and 107 variables has been imputed using the Vtreat (Zumel and Mount 2016) imputation method in order to populate the missing values in a meaningful manner. As a result 458 dummy variables has been created constituting the final data dimensionality.

4.2 Data Visualization for Pattern Recognition

To this end, we employ a series of dimensionality reduction algorithms for embedding the ATHLOS dataset in two dimensions. All methodologies are implemented using the R-project open-source environment for statistical computing, and experiments were conducted for each algorithm to tune its hyper-parameters. In what follows, we provide the details of each implementation tested, referring to the corresponding R packages when possible. For PCA, we employed the implementation found in the "dimRed" package (Kraemer, Reichstein and Mahecha 2018). The fast "Rtsne" implementation for tSNE (Krijthe 2015) was used, and the two main hyper-parameters were set as follows, "perplexity" at a range of 30 to 800 and "theta" at a range of 0.1 to 1. The "uwot" implementation of the UMAP algorithm (McInnes, Healy and Melville 2018) is used, and the two main hyper-parameters were examined, "cluster neighbors" at a range from 15 to 100 and "minimum distance" at a range from 0.01 to 0.15. Finally, the "largeVis" implementation of the LargeVis algorithm (Elberg 2020) was used, and the hyper-parameters, "Kapa" and "max iterations," were set at a range from 10 to 200 and from 10 to 50 accordingly.

In Figure 4.1, we observe the resulting visualization for all methods across the samples. It is evident that PCA tends to produce a more coherent representation without however distinguishing any particular clusters. The advantage of this technique, however, is that the coordinates have a strict and valuable definition. UMAP and LargeVis seem to produce very similar results creating several very distinct clusters. tSNE also created groups that were, however, larger, taking into account their in-between cluster distances. As depicted in Figure 4.1, the non-linear dimensionality reduction techniques, in this case, tend to separate more clearly groups of individuals that lie more closely in the intrinsic dimensionality manifold.

4.3 Clustering for Verification

The next step of our analysis is to determine the existence of clusters that could be verified through visualization. As a first step, we looked at the clustering tendency of the dataset. That is whether the data contain any inherent grouping structure. For this purpose, we calculated the well-known "Hopkins" statistic (Lawson and Jurs 1990), for which values close to 1 indicate a clusterable dataset. In our case, using the “clusterland” R package (YiLan and RuTong 2015)(, the corresponding calculated value is *83%,* suggesting a high degree of clusterability. Subsequently, for brevity and the sake of the reader's convenience, we choose to determine a representative number of clusters to accompany visualizations instead of providing extensive parameter analysis, which would hinder visual interpretation. Apparently, this task is not trivial, and there are over thirty different approaches in the literature regarding this exact case. After both heuristic experiments and counseling the results provided by the "factoextra" R package (Kassambara, Mundt and others 2017), we propose that any number between *4* to *12* clusters is an appropriate choice. Consequently, we proceed to the clustering step. For the Kmeans algorithm, we chose the memory efficient implementation found in (Emerson and Kane 2020), optimized for large scale applications. For the Genie algorithm, the corresponding R package “genie” (Gagolewski, Bartoszuk and Cena 2016) was used. For the Ncutdc algorithm, the implementation found in the authors ‘s “PPCI” (Hofmeyr and Pavlidis, PPCI: an R Package for Cluster Identification using Projection Pursuit 2019) package was used. Lastly, any attempt to use popular density-based approaches in the original dimension space, unfortunately, failed due to the dataset's scale. In particular, we employed the recent implementation of the DensityPeaks algorithm found in (Pedersen, Hughes and Qiu 2017) but hardware requirements exceeded 1TB of RAM usage. Thus, we chose to firstly reduce the number of dimensions down to 50 with the PCA method in order to

be able to run all the algorithms in a reasonable amount of time. Afterward, we extracted a uniformly random subsample of the dataset of 20000 samples. The DensityPeaks algorithm lacks the ability to tune the parameters “rho” and “delta” that affect the number of retrieved clusters. Instead, provided is a graphical tool with which the user can manually set the respective values through a scatter plot's visual investigation. In our tests, we set these as the mean values across the parameters calculated for all samples.

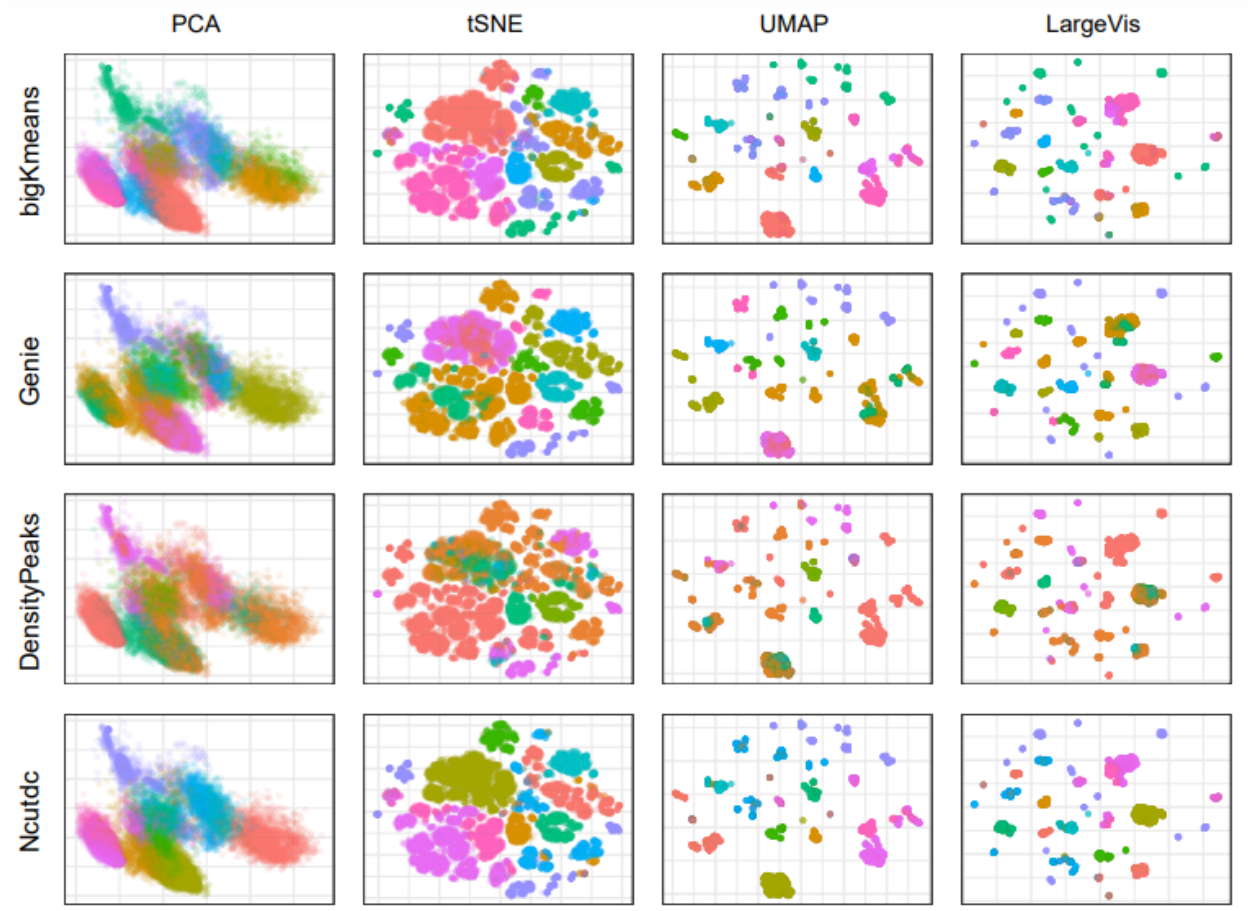

Figure 4.1: 2D visualization of PCA, tSNE, UMAP and LargeVis embeddings (column-wise) on the ATHLOS dataset clustered with “bigKmeans”, “Genie”, “DensityPeaks”, and “Ncutdc” clustering algorithms (row-wise) accordingly. Circular points represent each sample of data colored by the corresponding cluster ID resulting from the respective clustering algorithms.

Cluster memberships depicted in Figure 4.1 arise from the application of bigKmeans, Genie, DensityPeaks and Ncutdc clustering algorithms on the ATHLOS dataset in the 50-d dimensional space, with k=10. It is observed that there is a clear correlation between the embeddings in the two-dimensional space and the clusters assigned by bigKmeans. Also, the Ncutdc algorithm produces very distinct clusters, similar to bigKmeans, with minor tangling, as it is the case of Genie. Lastly, for DensityPeaks, a large number of data points appear to be mixed regarding the allocated cluster and the corresponding location within the embedding. Thus, this result further verifies that the visualizations were representative and that there are groups of individuals with some distinct characteristics. However, which variables are the dominant separation attributes is not easy to interpret only with statistical measures or exhaustive search for each feature in every cluster.

To numerically validate the clustering result of the aforementioned algorithms, we employ an implementation of the Silhouette index (Rousseeuw 1987) Dunn index (Bezdek and Pal 1995) and Separation index specifically designed for Large scale data provided by the “fpc” R package (Hennig 2020). Silhouette is a method of interpretation and validation of consistency within a cluster. The silhouette index compares how similar an object is to its group compared to other clusters. The silhouette has a range of *-1* to *+1*, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. The Dunn index depicts the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance and has a value between zero and infinity. The Separation Index is defined based on the distances for every data point to the closest neighbor not in the same group. The separation index then depicts the mean of the smallest proportion of these distances. This allows formalizing separation less sensitive to a single or a few ambiguous points. Considering the results of the

aforementioned metrics presented in Table 1, it is verified that “DensityPeaks” do not separate clusters as effectively as the other three methods with “Ncustdc” achieving the highest scores.

	Silhouette	**Dunn**	**Separation**
BigKmeans	0.1265699	0.738058	8.343318
Genie	0.04942602	0.7035942	6.956037
DensityPeaks	-0.1252029	0.4331565	8.033905
Ncutdc	**0.1301279**	**0.7423549**	**8.436816**

Table 1: Internal Validation metrics for the Clustering algorithms. Larger values depict grater similarities within the same cluster, while the clusters been more separate.

In an attempt to extract knowledge based on the identified patterns, the 2d embedding of the tSNE method was colored according to some categorical variables of interest (see Figure 4.2). The variables were selected according to the following scheme. For every categorical variable, we calculated the Silhouette (Rousseeuw 1987), Dunn (Bezdek and Pal 1995) and Separation (Hennig 2020) indexes for the 2d embeddings , considering the variables as a typical clustering results. Then, we chose six variables that presented the highest separability according to these metrics. There are apparent correlations with some of the variables and the resulting clumps in the visualization, implying that individuals have some distinct characteristics. We interestingly observe that we are able to identify separable regions characterized by particular levels of the categorical variables, potentially leading to straight forward cluster characterization.

	Silhouette	**Dunn**	**Separation**
Cardiovascular_History	0.2239903	1.2507654	1.1195133
Angina_History	0.2198167	1.2698951	1.2975713
Respiratory_History	0.2059388	1.3212471	1.0268838
Hypertension_History	0.1861763	1.2968042	0.9150790
Frequent_Contacts	0.1800876	1.2259719	0.1931220
Group_Sports	0.1533813	1.2334624	4.9907552

Table 2: Internal Validation metrics for the most separated features. Larger values depict grater similarities within the samples of with the same value for that particular feature, while the samples not in the same category been more separate.

In more detail, the features depicted in Table 2 are the following: Cardiovascular_History: Is a binary variable and refers to the history of stroke or myocardial infarction (heart attack) of an individual. Angina_History: Is a binary variable and refers to the "h_angina" variable of the original dataset that depicts whether or not an individual had a history of angina. Respiratory_History: Refers to the categorical "h_respiratory" variable of the original dataset that depicts one's history of chronic respiratory diseases such asthma, CPD, COPD, bronchitis, etc. This variable is then transformed with the "Vtreat" package to an encoding that expresses the within-group deviation of the outcome conditioned on each categorical level in the original data and thus, amend the high cardinality of the variable. Hypertension_History: Refers to the categorical "h_hypertension" variable of the original dataset that depicts one's history of hypertension for an individual. This variable is then transformed with the "Vtreat" package to an encoding that expresses the within-group deviation of the outcome conditioned on each categorical level in the original data and thus, amend the high cardinality of the variable. Frequent_Contacts: Is a binary variable and refers to the "cont_fr" variable of the original dataset that depicts if an

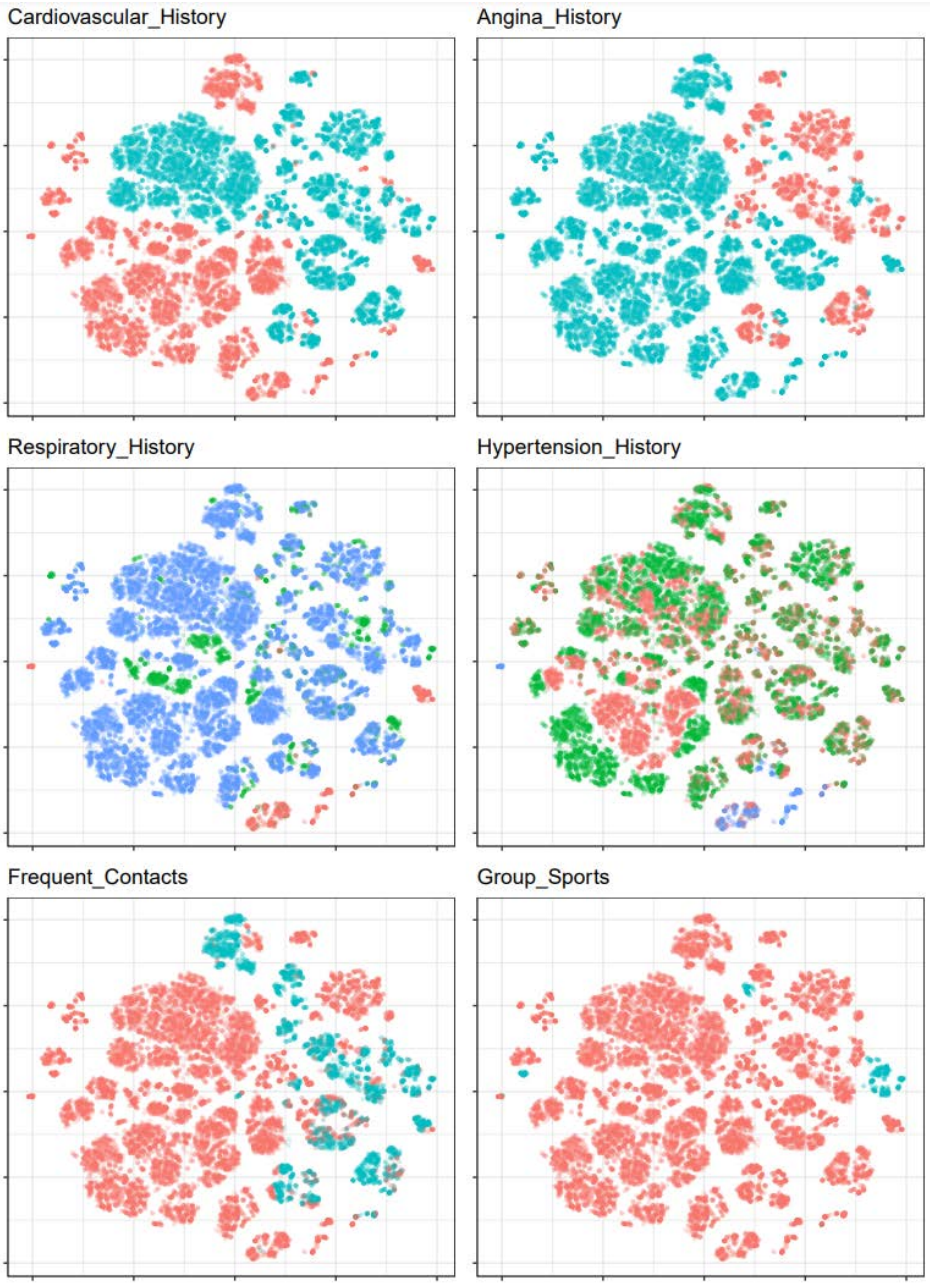

Figure 4.2: Visualizations of the 2d embedded dataset using t-SNE with respect to a selection of variables found in the ATHLOS dataset. Each subplot corresponds to different variables, while different colors correspond to different values for every variable.

individual has frequent contacts with friends/neighbours. Group_Sports: Is a binary variable and refers to the "sport" variable of the original dataset that depicts if an individual participates currently in group sport activities.

4.4 Variable Importance through Heatmaps

Motivated by the previous sections' findings, we introduce an additional step to visualize the regions' variable differences. For the final step, we chose to include a novel method so far utilized only for Genomics in Gene Expression (Linderman, Rachh, et al. 2019). The process incorporates both tSNE and clustering in order to produce a Heatmap that visualizes many variables (instead of genes) of interest at the same time. To build the t-SNE Heatmap introduced in (Linderman, Rachh, et al. 2019), we initially compute t-SNE embeddings of the Variables of interest into one dimension. This implementation incorporates the FIt-SNE (Linderman, Rachh, et al. 2017), which is scalable to millions of Variables in terms of computational time. Then, the 1D t-SNE embeddings are discretized in 100 bins, and the representation of each variable is produced by the sum of its expression in the samples contained in each bin, while each variable corresponds to a vector in R^100. Hierarchical clustering upon the aforementioned vectors produces even more meaningful results. Subsequently, for some set of variables of interest, the algorithm "enriches" the variables that have a similar expression pattern in the t-SNE (see Figure 4.3). Afterward, these vectors are transformed into Heatmap format, with each row being a variable and each column a bin using the heatmaply R package (Galili, et al. 2018).

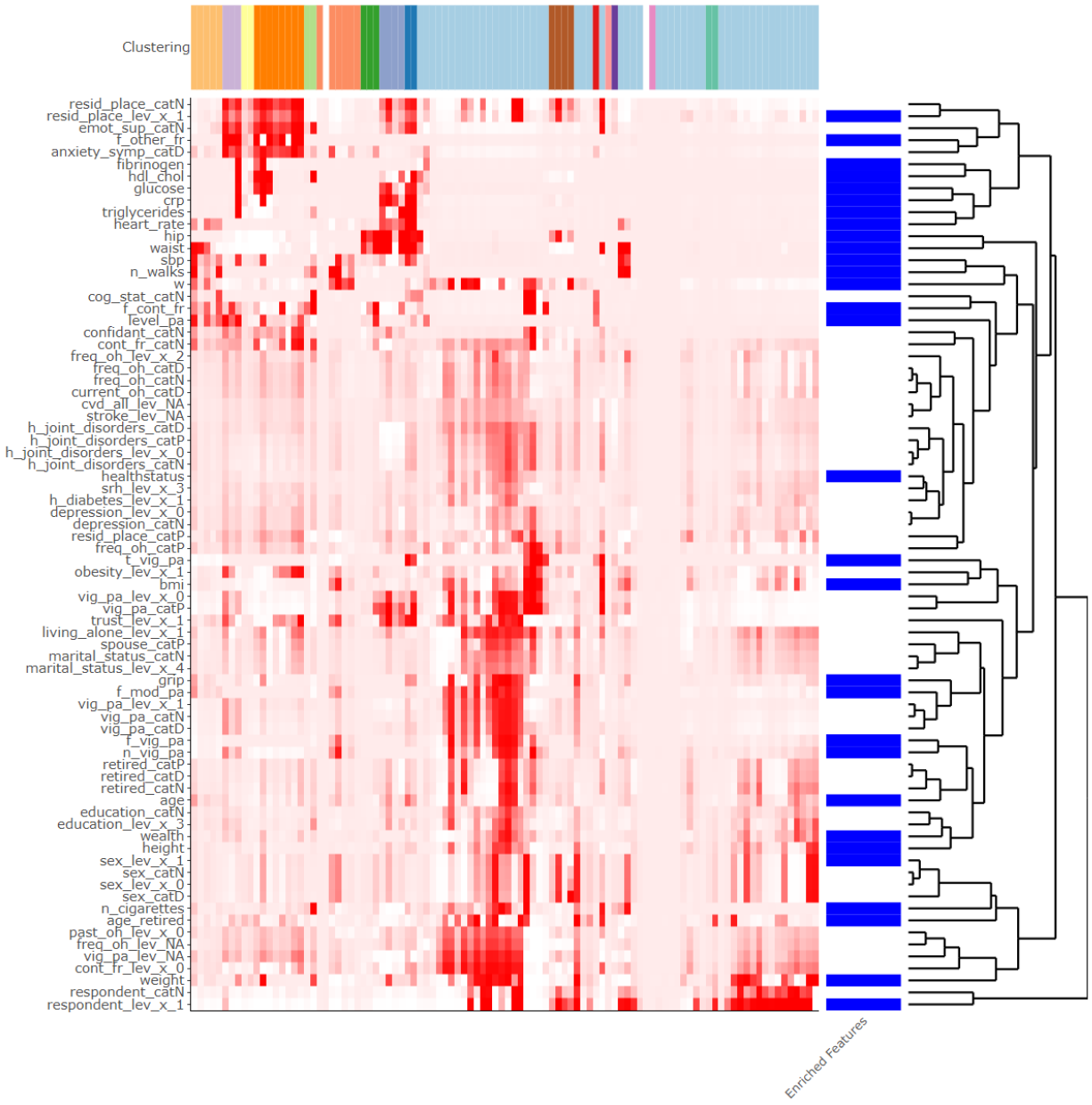

Figure 4.3: Heatmap Visualization of the Athlos dataset. Depicted are the Variables in every row, colored according to their respective value, with red depicting a larger value. The columns represent 100 bins of samples and are arranged according to their 1d-tSNE embedding. Also, the "Clustering" bar depict the cluster assignment of the DBSCAN algorithm run in the 2d-tsne embedding of the dataset. The "Enriched_Features" bar depicts the additional variables returned by the 1d-tsne-heatmap function where in blue color are the variables of interest as an input for the function and with white color are the additional variables returned for the "enrich" parameter equal to 5.

This is possible because it has been previously shown that t-SNE preserves the cluster structure of well-clustered data regardless of the embedding dimension (Linderman, Rachh, et al. 2019), and thus 1D t-SNEs contains the same information as 2D t-SNEs. The resulting vectors visualized in the produced Heatmap presented in Figure 4.3 provides a clear depiction of the variables' behavior among the clusters, with hundreds of variables visualized at the same time.

5 Conclusions

This study provided insight into recent Unsupervised Learning methods and their popular implementations through their application on a real word complex dataset. A shown, based on these methods we were able to provide a comprehensive example of knowledge extraction and pattern recognition analysis. In addition, we utilized a promising novel unsupervised learning approach from the Gene Expression field to provide useful information of Variable expression for the ATHLOS cohort dataset at hand, exposing the method’s usefulness in similar Big Data tasks.

6 Acknowledgements

This work is supported by the ATHLOS (Aging Trajectories of Health: Longitudinal Opportunities and Synergies) project, funded by the European Union's Horizon 2020 Research and Innovation Program under grant agreement number 635316.

References

Adam, Stavros P., Stamatios-Aggelos N. Alexandropoulos, Panos M. Pardalos, and Michael N. Vrahatis. 2019. "No free lunch theorem: A review." *Approximation and optimization* (Springer) 57–82.

Alharthi, Abdulkhaliq, Vlad Krotov, and Michael Bowman. 2017. "Addressing barriers to big data." *Business Horizons* (Elsevier) 60: 285–292.

Allaoui, Mebarka, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study." *International Conference on Image and Signal Processing.* 317–325.

Anagnostou, Panagiotis, Sotiris Tasoulis, Aristidis G. Vrahatis, Spiros Georgakopoulos, Matthew Prina, Jose Luis Ayuso-Mateos, Jerome Bickenbach, et al. 2021. "Enhancing the Human Health Status Prediction: the ATHLOS Project." *medRxiv* (Cold Spring Harbor Laboratory Press).

Arokiasamy, P., David Bloom, Jinkook Lee, Kevin Feeney, and Marija Ozolins. 2012. "Longitudinal aging study in India: Vision, design, implementation, and preliminary findings." In *Aging in Asia: findings from new and emerging data initiatives*. National Academies Press (US).

Bellman, R. n.d. "Corporation, R.(1957) Dynamic Programming." *Corporation, R.(1957) Dynamic Programming.* Princeton University Press, NJ, USA.

Berkhin, Pavel. 2006. "A survey of clustering data mining techniques." In *Grouping multidimensional data*, 25–71. Springer.

Bezdek, James C., and Nikhil R. Pal. 1995. "Cluster validation with generalized Dunn's indices." *Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems.* 190–193.

Börsch-Supan, Axel, Martina Brandt, Christian Hunkler, Thorsten Kneip, Julie Korbmacher, Frederic Malter, Barbara Schaan, Stephanie Stuck, and Sabrina Zuber. 2013. "Data

resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE)." *International journal of epidemiology* (Oxford University Press) 42: 992–1001.

Casolla, Giampaolo, Salvatore Cuomo, Vincenzo Schiano Di Cola, and Francesco Piccialli. 2019. "Exploring unsupervised learning techniques for the Internet of Things." *IEEE Transactions on Industrial Informatics* (IEEE) 16: 2621–2628.

Chawla, Sanjay, and Aristides Gionis. 2013. "k-means–: A unified approach to clustering and outlier detection." *Proceedings of the 2013 SIAM International Conference on Data Mining.* 189–197.

Cios, Krzysztof J., Witold Pedrycz, and Roman W. Swiniarski. 1998. "Data mining and knowledge discovery." In *Data mining methods for knowledge discovery*, 1–26. Springer.

Diaz-Papkovich, Alex, Luke Anderson-Trocmé, and Simon Gravel. 2019. "UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts." *PLoS genetics* (Public Library of Science) 15: e1008432.

Elberg, Amos B. 2020. "largeVis: High-Quality Visualizations of Large, High-Dimensional Datasets." https://github.com/elbamos/largeVis.

Emerson, John W., and Michael J. Kane. 2020. "biganalytics: Utilities for 'big.matrix' Objects from Package 'bigmemory'." https://CRAN.R-project.org/package=biganalytics.

Everitt, Brian S., Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster analysis.* John Wiley & Sons.

Fan, Jianqing, and Yingying Fan. 2008. "High dimensional classification using features annealed independence rules." *Annals of statistics* (NIH Public Access) 36: 2605.

Fan, Jianqing, Fang Han, and Han Liu. 2014. "Challenges of big data analysis." *National science review* (Oxford University Press) 1: 293–314.

Fan, Jianqing, Shaojun Guo, and Ning Hao. 2012. "Variance estimation using refitted cross-validation in ultrahigh dimensional regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (Wiley Online Library) 74: 37–65.

Forgy, Edward W. 1965. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications." *biometrics* 21: 768–769.

Gagolewski, Marek, Maciej Bartoszuk, and Anna Cena. 2016. "Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm." *Information Sciences* 363: 8–23. doi:10.1016/j.ins.2016.05.003.

Galili, Tal, Alan O'Callaghan, Jonathan Sidi, and Carson Sievert. 2018. "heatmaply: an R package for creating interactive cluster heatmaps for online publishing." *Bioinformatics* (Oxford University Press) 34: 1600–1602.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. 2020. *Data clustering: theory, algorithms, and applications.* SIAM.

Gao, Jing, Liang Zhao, Zhikui Chen, Peng Li, Han Xu, and Yueming Hu. 2016. "ICFS: An Improved Fast Search and Find of Density Peaks Clustering Algorithm." *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)* 537-543.

Ghahramani, Zoubin. 2003. "Unsupervised learning." *Summer School on Machine Learning.* 72–112.

Ghodsi, Ali. 2006. "Dimensionality reduction a short tutorial." *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada* 37: 2006.

Gu, Dongxiao, Jingjing Li, Xingguo Li, and Changyong Liang. 2017. "Visualizing the knowledge structure and evolution of big data research in healthcare informatics." *International journal of medical informatics* (Elsevier) 98: 22–32.

Gupta, Deepak, and Rinkle Rani. 2019. "A study of big data evolution and research challenges." *Journal of Information Science* (SAGE Publications Sage UK: London, England) 45: 322–340.

Hahsler, Michael, Matthew Piekenbrock, S. Arya, and D. Mount. 2017. "dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and related algorithms." *R package version* 1–0.

Hameed, Pathima Nusrath, Karin Verspoor, Snezana Kusljic, and Saman Halgamuge. 2018. "A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration." *BMC bioinformatics* (Springer) 19: 129.

Hennig, Christian. 2020. "fpc: Flexible Procedures for Clustering." https://CRAN.R-project.org/package=fpc.

Hinton, Geoffrey E., and Sam T. Roweis. 2003. "Stochastic neighbor embedding." *Advances in neural information processing systems.* 857–864.

Hinton, Geoffrey E., Terrence Joseph Sejnowski, Tomaso A. Poggio, and others. 1999. *Unsupervised learning: foundations of neural computation.* MIT press.

Hofmeyr, David P. 2016. "Clustering by minimum cut hyperplanes." *IEEE transactions on pattern analysis and machine intelligence* (IEEE) 39: 1547–1560.

Hofmeyr, David P., and Nicos G. Pavlidis. 2019. "PPCI: an R Package for Cluster Identification using Projection Pursuit." *The R Journal.* doi:10.32614/RJ-2019-046.

Hozumi, Yuta, Rui Wang, Changchuan Yin, and Guo-Wei Wei. 2021. "UMAP-assisted K-means

clustering of large-scale SARS-CoV-2 mutation datasets." *Computers in biology and medicine* (Elsevier) 104264.

Hsu, Wei-Ning, and James Glass. 2018. "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 5614–5618.

Ichimura, Hidehiko, Satoshi Shimizutani, and Hideki Hashimoto. 2009. "JSTAR first results 2009 report." Tech. rep., Research Institute of Economy, Trade and Industry (RIETI).

Jajuga, Krzystof, Andrzej Sokolowski, and Hans-Hermann Bock. 2012. "Classification, clustering, and data analysis: recent advances and applications." (Springer Science & Business Media).

Jiang, Sheng-yi, and Qing-bo An. 2008. "Clustering-based outlier detection method." *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery.* 429–433.

Jin, Xiaolong, Benjamin W. Wah, Xueqi Cheng, and Yuanzhuo Wang. 2015. "Significance and challenges of big data research." *Big Data Research* (Elsevier) 2: 59–64.

Kaski, Samuel. 1998. "Dimensionality reduction by random mapping: Fast similarity computation for clustering." *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227).* 413–418.

Kassambara, Alboukadel, Fabian Mundt, and others. 2017. "Factoextra: extract and visualize the results of multivariate data analyses." *R package version* 1: 337–354.

Keyes, Katherine M., and Daniel Westreich. 2019. "UK Biobank, big data, and the consequences of non-representativeness." *Lancet (London, England)* (NIH Public Access) 393: 1297.

Kobak, Dmitry, and Philipp Berens. 2019. "The art of using t-SNE for single-cell

transcriptomics." *Nature communications* (Nature Publishing Group) 10: 1–14.

Koskinen, S. 2018. "Health 2000 and 2011 Surveys—THL Biobank. National Institute for Health and Welfare." *Health 2000 and 2011 Surveys—THL Biobank. National Institute for Health and Welfare.*

Kowal, Paul, Somnath Chatterji, Nirmala Naidoo, Richard Biritwum, Wu Fan, Ruy Lopez Ridaura, Tamara Maximova, et al. 2012. "Data resource profile: the World Health Organization Study on global AGEing and adult health (SAGE)." *International journal of epidemiology* (Oxford University Press) 41: 1639–1649.

Kraemer, Guido, Markus Reichstein, and Miguel D. Mahecha. 2018. "dimRed and coRanking—Unifying dimensionality reduction in R." *R Journal* (R Foundation) 10: 342–358.

Krijthe, Jesse H. 2015. "Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation." *R package version 0.13, URL https://github. com/jkrijthe/Rtsne.*

Kullback, Solomon. 1997. *Information theory and statistics.* Courier Corporation.

Lawson, Richard G., and Peter C. Jurs. 1990. "New index for clustering tendency and its application to chemical problems." *Journal of chemical information and computer sciences* (ACS Publications) 30: 36–41.

Leonardi, Matilde, Somnath Chatterji, Seppo Koskinen, Jose Luis Ayuso-Mateos, Josep Maria Haro, Giovanni Frisoni, Lucilla Frattura, et al. 2014. "Determinants of health and disability in ageing population: the COURAGE in Europe Project (collaborative research on ageing in Europe)." *Clinical psychology & psychotherapy* (Wiley Online Library) 21: 193–198.

Li, Miaomiao, Xinwang Liu, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. 2016. "Multiple kernel clustering with local kernel alignment maximization."

Linderman, George C., Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. 2017. "Efficient algorithms for t-distributed stochastic neighborhood embedding." *arXiv preprint arXiv:1712.09005.*

Linderman, George C., Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. 2019. "Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data." *Nature methods* (Nature Publishing Group) 16: 243–245.

Luszcz, Mary A., Lynne C. Giles, Kaarin J. Anstey, Kathryn C. Browne-Yung, Ruth A. Walker, and Tim D. Windsor. 2016. "Cohort Profile: The Australian Longitudinal Study of Ageing (ALSA)." *International journal of epidemiology* (Oxford University Press) 45: 1054–1063.

Ma, Fenglong, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang. 2017. "Unsupervised discovery of drug side-effects from heterogeneous data sources." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 967–976.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing data using t-SNE." *Journal of machine learning research* 9: 2579–2605.

Marx, Vivien. 2013. "The big challenges of big data." *Nature* (Nature Publishing Group) 498: 255–260.

Masaeli, Mahdokht, Glenn Fung, and Jennifer G. Dy. 2010. "From transformation-based dimensionality reduction to feature selection." *ICML*.

McInnes, Leland, John Healy, and James Melville. 2018. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426.*

Mladenić, Dunja. 2005. "Feature selection for dimensionality reduction." *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. 84–102.

Murtagh, Fionn, and Pierre Legendre. 2014. "Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?" *Journal of classification* (Springer) 31: 274–295.

Park, Joon Hyuk, Soo Lim, J. Lim, K. Kim, M. Han, In Young Yoon, J. Kim, et al. 2007. "An overview of the Korean longitudinal study on health and aging." *Psychiatry investigation* (YOUNG CHO CHUNG) 4: 84.

Pavlidis, Nicos G., David P. Hofmeyr, and Sotiris K. Tasoulis. 2016. "Minimum density hyperplanes." *The Journal of Machine Learning Research* (JMLR. org) 17: 5414–5446.

Pearson, Karl. 1901. "LIII. On lines and planes of closest fit to systems of points in space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (Taylor & Francis) 2: 559–572.

Peasey, Anne, Martin Bobak, Ruzena Kubinova, Sofia Malyutina, Andrzej Pajak, Abdonas Tamosiunas, Hynek Pikhart, Amanda Nicholson, and Michael Marmot. 2006. "Determinants of cardiovascular disease and other non-communicable diseases in Central and Eastern Europe: rationale and design of the HAPIEE study." *BMC public health* (Springer) 6: 255.

Pedersen, Thomas Lin, Sean Hughes, and Xiaojie Qiu. 2017. "densityClust: Clustering by Fast Search and Find of Density Peaks." https://CRAN.R-project.org/package=densityClust.

Plageras, Andreas P., Kostas E. Psannis, Christos Stergiou, Haoxiang Wang, and Brij B. Gupta. 2018. "Efficient IoT-based sensor BIG Data collection–processing and analysis in smart

buildings." *Future Generation Computer Systems* (Elsevier) 82: 349–357.

Prina, A. Matthew, Daisy Acosta, Isaac Acosta, Mariella Guerra, Yueqin Huang, A. T. Jotheeswaran, Ivonne Z. Jimenez-Velazquez, et al. 2017. "Cohort profile: the 10/66 study." *International journal of epidemiology* (Oxford University Press) 46: 406–406i.

Rauber, Paulo E., Alexandre X. Falcão, Alexandru C. Telea, and others. 2016. "Visualizing Time-Dependent Data Using Dynamic t-SNE."

Rodriguez, Alex, and Alessandro Laio. 2014. "Clustering by fast search and find of density peaks." *science* (American Association for the Advancement of Science) 344: 1492–1496.

Rodríguez-Artalejo, Fernando, Auxiliadora Graciani, Pilar Guallar-Castillón, Luz M. León-Muñoz, M. Clemencia Zuluaga, Esther López-García, Juan Luis Gutiérrez-Fisac, et al. 2011. "Rationale and methods of the study on nutrition and cardiovascular risk in Spain (ENRICA)." *Revista Española de Cardiología (English Edition)* (Elsevier) 64: 876–882.

Roh, Yuji, Geon Heo, and Steven Euijong Whang. 2019. "A survey on data collection for machine learning: a big data-ai integration perspective." *IEEE Transactions on Knowledge and Data Engineering* (IEEE).

Rostom, Raghd, Valentine Svensson, Sarah A. Teichmann, and Gozde Kar. 2017. "Computational approaches for interpreting scRNA-seq data." *FEBS letters* (Wiley Online Library) 591: 2213–2225.

Rousseeuw, Peter J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* (Elsevier) 20: 53–65.

Sagiroglu, Seref, and Duygu Sinanc. 2013. "Big data: A review." *2013 international conference*

on collaboration technologies and systems (CTS). 42–47.

Sanchez-Niubo, Albert, Laia Egea-Cortés, Beatriz Olaya, Francisco Félix Caballero, Jose L. Ayuso-Mateos, Matthew Prina, Martin Bobak, et al. 2019. "Cohort profile: The Ageing trajectories of health–longitudinal opportunities and synergies (ATHLOS) project." *International journal of epidemiology* (Oxford University Press) 48: 1052–1053i.

Shah, Jamal Hussain, Muhammad Sharif, Mudassar Raza, and Aisha Azeem. 2013. "A Survey: Linear and Nonlinear PCA Based Face Recognition Techniques." *Int. Arab J. Inf. Technol.* 10: 536–545.

Sharma, Alok, Yosvany López, and Tatsuhiko Tsunoda. 2017. "Divisive hierarchical maximum likelihood clustering." *BMC bioinformatics* (BioMed Central) 18: 546.

Shi, Jianbo, and Jitendra Malik. 2000. "Normalized cuts and image segmentation." *IEEE Transactions on pattern analysis and machine intelligence* (Ieee) 22: 888–905.

Shirkhorshidi, Ali Seyed, Saeed Aghabozorgi, Teh Ying Wah, and Tutut Herawan. 2014. "Big data clustering: a review." *International conference on computational science and its applications.* 707–720.

Sonnega, Amanda, Jessica D. Faul, Mary Beth Ofstedal, Kenneth M. Langa, John W. R. Phillips, and David R. Weir. 2014. "Cohort profile: the health and retirement study (HRS)." *International journal of epidemiology* (Oxford University Press) 43: 576–585.

Sorzano, Carlos Oscar Sánchez, Javier Vargas, and A. Pascual Montano. 2014. "A survey of dimensionality reduction techniques." *arXiv preprint arXiv:1403.2877.*

Sreedhar, Chowdam, Nagulapally Kasiviswanath, and Pakanti Chenna Reddy. 2017. "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop." *Journal of Big Data* (Springer) 4: 27.

Steptoe, Andrew, Elizabeth Breeze, James Banks, and James Nazroo. 2013. "Cohort profile: the English longitudinal study of ageing." *International journal of epidemiology* (Oxford University Press) 42: 1640–1648.

Tang, Jian, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. 2016. "Visualizing large-scale and high-dimensional data." *Proceedings of the 25th international conference on world wide web.* 287–297.

Tasoulis, S., L. Cheng, N. Välimäki, N. J. Croucher, S. R. Harris, W. P. Hanage, T. Roos, and J. Corander. 2014. "Random projection based clustering for population genomics." *2014 IEEE International Conference on Big Data (Big Data).* 675-682. doi:10.1109/BigData.2014.7004291.

Tufekci, Zeynep. 2014. "Big questions for social media big data: Representativeness, validity and other methodological pitfalls." *Proceedings of the International AAAI Conference on Web and Social Media.*

Usama, Muhammad, Junaid Qadir, Aunn Raza, Hunain Arif, Kok-Lim Alvin Yau, Yehia Elkhatib, Amir Hussain, and Ala Al-Fuqaha. 2019. "Unsupervised machine learning for networking: Techniques, applications and research challenges." *IEEE Access* (IEEE) 7: 65579–65615.

Von Luxburg, Ulrike. 2007. "A tutorial on spectral clustering." *Statistics and computing* (Springer) 17: 395–416.

Wang, C. Jason, Chun Y. Ng, and Robert H. Brook. 2020. "Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing." *Jama* (American Medical Association) 323: 1341–1342.

Wang, Shiping, Jinyu Cai, Qihao Lin, and Wenzhong Guo. 2019. "An overview of unsupervised

deep feature representation for text categorization." *IEEE Transactions on Computational Social Systems* (IEEE) 6: 504–517.

Wei, Hua-Liang, and Stephen A. Billings. 2006. "Feature subset selection and ranking for data dimensionality reduction." *IEEE transactions on pattern analysis and machine intelligence* (IEEE) 29: 162–166.

Whelan, Brendan J., and George M. Savva. 2013. "Design and methodology of the Irish Longitudinal Study on Ageing." *Journal of the American Geriatrics Society* (Wiley Online Library) 61: S265–S268.

Wong, Rebeca, Alejandra Michaels-Obregon, and Alberto Palloni. 2017. "Cohort profile: the Mexican health and aging study (MHAS)." *International journal of epidemiology* (Oxford University Press) 46: e2–e2.

Xia, Jiazhi, Fenjin Ye, Wei Chen, Yusi Wang, Weifeng Chen, Yuxin Ma, and Anthony K. H. Tung. 2017. "LDSScanner: Exploratory analysis of low-dimensional structures in high-dimensional datasets." *IEEE transactions on visualization and computer graphics* (IEEE) 24: 236–245.

Xiang, Lingyun, Guohan Zhao, Qian Li, Wei Hao, and Feng Li. 2018. "TUMK-ELM: a fast unsupervised heterogeneous data learning approach." *IEEE Access* (IEEE) 6: 35305–35315.

Yan, Jun, Benyu Zhang, Ning Liu, Shuicheng Yan, Qiansheng Cheng, Weiguo Fan, Qiang Yang, Wensi Xi, and Zheng Chen. 2006. "Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing." *IEEE transactions on Knowledge and Data Engineering* (IEEE) 18: 320–333.

YiLan, Luo, and Zeng RuTong. 2015. "clustertend: Check the Clustering Tendency."

https://CRAN.R-project.org/package=clustertend.

Zhang, Wei, Deli Zhao, and Xiaogang Wang. 2013. "Agglomerative clustering via maximum incremental path integral." *Pattern Recognition* (Elsevier) 46: 3056–3065.

Zhu, Chengzhang, Longbing Cao, Qiang Liu, Jianping Yin, and Vipin Kumar. 2018. "Heterogeneous metric learning of categorical data with hierarchical couplings." *IEEE Transactions on Knowledge and Data Engineering* (IEEE) 30: 1254–1267.

Zhu, Xiaojin Jerry. 2005. "Semi-supervised learning literature survey." (University of Wisconsin-Madison Department of Computer Sciences).

Zhu, Xiaojin, and Andrew B. Goldberg. 2009. "Introduction to semi-supervised learning." *Synthesis lectures on artificial intelligence and machine learning* (Morgan & Claypool Publishers) 3: 1–130.

Zumel, Nina, and John Mount. 2016. "vtreat: a data. frame Processor for Predictive Modeling." *arXiv preprint arXiv:1611.09477*.