Master Thesis in Statistics and Data Mining

Alzheimer's disease heterogeneity assessment using high dimensional clustering techniques

Konstantinos Poulakis

Division of Statistics and Machine Learning

Department of Computer and Information Science

Linköping University

June 2016

Supervisors

Linköping University
Jose M. Peña

Karolinska Institute
Joana Braga Pereira

Examiner
Mattias Villani

Table of Contents

Abstract

This thesis sets out to investigate the Alzheimer's disease (AD) heterogeneity in an unsupervised framework. Different subtypes of AD were identified in the past from a number of studies. The major objective of the thesis is to apply clustering methods that are specialized in coping with high dimensional data sets, in a sample of AD patients. The evaluation of these clustering methods and the interpretation of the clustered groups from a statistical and a medical point of view, are some of the additional objectives.

The data consist of 271 MRI images of AD patients from the AddNeuroMed and the ADNI cohorts. The raw MRI's have been preprocessed with the software Freesurfer and 82 cortical and subcortical volumes have been extracted for the needs of the analysis.

The effect of different strategies in the initialization of a modified Gaussian Mixed Model (GMM) (Bouveyron et al, 2007) has been studied. Additionally, the GMM and a Bayesian clustering method proposed by Nia (2009) have been compared with respect to their performances in various distance based evaluation criteria. The later method resulted in the most compact and isolated clusters. The optimal numbers of clusters was evaluated with the Hopkins statistic and 6 clusters were decided while 2 observations formed an outlier cluster.

Different patterns of atrophy were discovered in the 6 clusters. One cluster presented atrophy in the medial temporal area only (n=37,~13.65%). Another cluster resented atrophy in the lateral and medial temporal lobe and parts of the parietal lobe (n=39,~14.4%). A third cluster presented atrophy in temporoparietal areas but also in the frontal lobe (n=74,~27.3%). The remaining three clusters presented diffuse atrophy in nearly all the association cortices with some variation in the patterns $(\mathrm{n}_1 = 40, \sim 14.7\%, \mathrm{n}_2 = 58, \sim 21.4, \mathrm{n}_3 = 21, 7.7\%)$. The 6 subtypes also differed in their demographical, clinical and pathological features.

keywords: Alzheimer's, heterogeneity, atrophy, clustering, Bayesian, morphometry, dimension

Acknowledgements

I take this opportunity to express my gratitude to several people whose expertise, sincere and valuable comments guided me in the writing of this thesis.

First, I would like to thank Karolinska Institute and especially the Division of Clinical Geriatrics, for giving me the opportunity to work with them and also providing the data set for the thesis. I would like to express my great appreciation to Erik Westman for defining and interesting problem as well as motivating me with interesting conversations.

Special thanks to my supervisor from Karolinska Institute Joana Braga Pereira for her continuous support, patience and immerse knowledge. Her guidance enlightened my way during the research and writing of this thesis.

I would like to thank my supervisor from Linköping University, Jose M. Peña whose contribution is profound all over the thesis and especially in the methodology part.

I wish to express my sincere thanks to the Division of Statistics and Machine Learning for providing me with all the necessary facilities for the research.

Last but not least I would like to thank my class mates for our funny, strange but interesting lunch discussions during this trip.

1 Introduction

1.1 Background

As the world population ages, neurodegenerative diseases that affect the elderly are becoming the focus of attention. Alzheimer's disease (AD) is currently the most common cause of dementia, with an increasing prevalence that nearly doubles every twenty years (Reitz et al., 2011; Karantzoulis et al., 2011; Meek et al., 1998). Current estimations[1] on the economic impact of AD suggest that vast amounts of resources will be necessary to take care adequately for patients afflicted with it.

AD is a progressive and ultimately fatal neurodegenerative disease, defined by loss of memory and other cognitive functions. Pathologically, it is characterized by decreases of β-amyloid (Aβ) peptides, reflecting amyloid plaque pathology, and increases of tau and phosphorylated tau proteins, indicating neurofibrillary tangles (Shaw et al., 2009). These pathological changes may start long before the patient experiences any symptoms, and usually lead to structural changes in the brain such as grey matter atrophy (Zhou et al., 2011). Medical imaging techniques allow studying brain structural changes in AD, providing a valuable diagnostic tool to monitor disease progression and treatment. In particular, structural Magnetic Resonance Imaging (MRI) is a non-invasive technique that has become very useful in medical research by providing information on different aspects of brain morphometry such as gray matter volume or cortical thickness. Volumetric and cortical thickness measures of atrophy can be extracted with automated morphometric analysis methods. These measures can be analysed using a wide variety of methodologies and improve our understanding on the AD-related changes in brain anatomy (Dickerson et al., 2001).

The research on AD prevalence and incidence mainly focuses on individuals that develop the disease after 65 years of age (late onset AD), where episodic memory loss is considered the most important cognitive symptom (Van der Flier al., 2011; Lehmann et al., 2013). However, some patients can develop AD before 65 years (early onset AD) and tend to present focal, non-amnestic clinical syndromes. Interestingly, these patients show more heterogeneous neuropsychological deficits, including fluent and non-fluent aphasia, apraxia, dyscalculia and executive dysfunction (Koedam et al., 2010; Galton et al., 2000; Gorno-Tempini et al., 2004; Johnson et al., 1999). A better understanding of the brain changes underlying the clinical heterogeneity in AD could provide critical insights into the disease mechanisms and improve the diagnosis of atypical syndromes.

[1] "The global cost of dementia is currently estimated on US$ 818 billion, which represent around 1.1% of global GDP."http://www.worldalzreport2015.org/downloads/world-alzheimer-report-2015-summary-sheet.pdf

The statistical analysis of biomedical data has received increasing attention in recent years as a result of the advances in both statistics and computer science. The use of population inference methods in the study of various diseases has revealed a lot of similarities in the patterns of their numerical representations. Such models allow assessing the structure of a disease from a totally numerical perspective, which can provide useful insights that help prevention, prognosis, and even treatment. The main task of inferential statistical methods is the extraction of knowledge to derive conclusions from data that are subjected to random variation. Both supervised and unsupervised techniques are popular tools among data analysts specialized in biomedical research.

In the machine learning terminology (Alpaydin, 2014), a supervised classification assigns new data object to one or more (finite) set of discrete class labels, on the basis of a training set of data objects whose category membership is known. In contrast, unsupervised learning (e.g. clustering) aims to explore the unknown nature of data through the separation of a finite dataset into a finite and discrete set of "natural" hidden data structures, with little or no ground truth based on inherent similarities or distances of the data (Everitt et al., 2001). Both clustering and supervised classifications reflect the human act of learning from data retrieved from observations and measurements. Clustering methods have become very popular due to their ability in dealing with the large amount of data that can be extracted from brain anatomy.

1.2 Objectives

The aims of this thesis are:

i. To investigate whether the heterogeneity of AD can be assessed with unsupervised statistical methods (clustering).

ii. To quantitatively and objectively evaluate the obtained unsupervised structure.

iii. To compare the characteristics of different subgroups of patients and investigate possible different phenotypes.

The next chapter briefly reviews the literature on the exploration of the AD subtypes, with an emphasis to the clinical, pathological characteristics and the atrophy patterns of the different AD subtypes and the statistical methods used to address this clustering problem. Chapter 3 describes the data sources and preprocessing, while chapter 4 includes the methodological framework of the study. In chapter 5 and 6 the most important results are presented and discussed. Finally, chapter 7 contains the conclusions and contributions of the thesis.

2 Literature review

This chapter is devoted to a short review of the existing literature on the exploration of the AD subtypes. Patients with AD pathology, while typically conceived as having an illness mainly of episodic memory, may evidence more prominent dysfunction in other cognitive domains, sometimes as the most salient trait of the illness (Dickerson et al., 2011). Interest in the investigation of the clinical heterogeneity of AD has been expressed from the scientific community for over two decades (Becker et al., 1988; Hof et al., 1989; Johnson et al., 1999). Both individual patient and population-based approaches have been employed in the strain of discriminating the diverse clinical subtypes of the disease. Neuroimaging studies have demonstrated that disease-specific atrophy patterns intimately match functional connectivity maps in cognitive normal individuals, suggesting that AD and other neurodegenerative disorders target specific functional networks (Seeley et al., 2009; Zhou et al., 2012). The comparison of patients with an early-onset (EOAD) to patients with a late-onset (LOAD) of AD has shown the different patterns of brain atrophy (Van der flier et al., 2011). Compared to the classic amnestic LOAD, EOAD shows greater atrophy in the temporo-parietal, medial parietal and lateral prefrontal cortex, while the medial temporal lobes are not affected significantly (Frisoni et al., 2005; Shiino et al., 2008).

The study of different neuropathological subtypes in AD remains ambiguous. Some studies reported that the pathological burden of senile plaques and neurofibrillary tangles was greater in EOAD compares to LOAD patients (Rossor et al., 1984; Nochlin et al., 1993; Bigio et al., 2002). However the focus of these studies is on the severity rather than the distribution of the pathology, thus they might suffer when matching the disease severity with the differential longevity (Van der flier et al., 2011). One study attempting to cluster the pathological changes of 80 patients into different AD subtypes using principal component analysis (PCA) did not find distinct subtypes (Armstrong et al., 2000). In contrast, another study found that the neuropathological features could separate AD patients into three different variants: a) a limbic predominant AD group characterized by high hippocampal neurofibrillary tangle counts and low neurofibrillary tangle counts in three association cortices (superior temporal, inferior parietal, middle frontal); b) a hippocampal-sparing AD group characterized by high neurofibrillary tangle counts in the three association cortices (where the limbic predominant subtype had lower values) and low neurofibrillary tangle counts in the hippocampus; and c) a typical AD group characterized by high neurofibrillary tangle counts both in the hippocampus and the association cortices (Murray et al., 2011). This pathological classification approach found that 25% of the AD patients present a non-typical distribution of neurofibrillary tangles. Followed by this description of subtypes, Whitwell et al. (2012) investigated the neuroimaging differences between hippocampal-sparing AD, limbic predominant

AD and typical AD. In that study, an association was found between neurofibrillary tangle deposition and brain atrophy, with the AD hippocampal-sparing cases showing the most cortical atrophy and the AD limbic predominant cases showing the greatest hippocampal atrophy (Whitwell et al., 2012). The algorithm used in that study of different AD classified the patients using three scores: a) the ratio of median hippocampal to cortical neurofibrillary tangle counts; b) the hippocampal to neurofibrillary tangle counts and c) the cortical neurofibrillary tangle counts. By assuming a predefined threshold to these scores, each patients has been classified in one of the three AD subtypes.

Concentrations of amyloid β (Aβ_{1_42}), microtubule associated protein tau (tau) and tau phosphorylated at threonine-181 (P-tau$_{181}$) in cerebrospinal fluid (CSF), are some of the biomarkers under substantial consideration in the progression of AD and MCI and have been investigated by various studies (Mattsson et al., 2009; Mulder et al., 2010; Sluimer et al., 2010). Although Aβ_{1-42} is typically lowered in the CSF, while tau and P-tau$_{181}$ are increased in patients with AD, no significant differences have been found between EOAD and LOAD in the level of these proteins (Andreasen et al., 1999; Bouwman et al., 2009). Moreover the CSF profile is similar between patients with atypical non-memory and typical memory phenotypes (Dickerson et al., 2010).

Besides neuropathological changes, several studies have examined extensively the underlying genetic factors associated with different patterns of neurodegeneration in AD. The apolipoproteinε4 (ApoE ε4) allele is the most important genetic risk factor for sporadic AD (Corder et al., 1993; Davidson et al., 2007). Although in a previous study, Armstrong et al. (2000) did not find distinct subtypes of AD using PCA analysis, individual differences in pathology were related to apolipoprotein E (ApoE) genotype, with e4 carriers showing greater senile plaque severity in frontal and occipital regions. ApoE e4 status might have different effects on disease progression in patients with EOAD and LOAD (Van der Flier et al., 2007). In patients with EOAD, e4 carriers had worse results in memory tests than did non-carriers (Marra et al., 2004; Lehtovirta et al., 1996). ApoEε4 might have different effects on disease progression in patients with EOAD and LOAD (Van der Flier et al., 2007). In a study where the whole brain atrophy rate was used as a measure of progression of the disease, the lack of the ApoEε4 in addition to younger age at disease onset was also associated with increased loss of brain volume (Sluimer et al., 2008).

Various methods have been applied to assess the heterogeneity of AD on brain morphometry. One study classified AD patients in four subgroups by analyzing brain volume loss with Voxel-based Morphometry (VBM) (Shiino et al., 2006). The four subgroups that were found in that study were characterized by atrophy in: the amygdala/hippocampal formations (Hipp); in the Hipp and Posterior Cingulate Cortex (Hipp-PCC); in the Hipp and posterior cortices (Hipp-TOPa) and in the PCC and posterior cortices (PCC/-TOPa). Another study used Surface-based Morphometric analysis (SBA) combined with an unsupervised learning approach based on Ward's clustering linkage method in 152 patients (Noh et al., 2014). In this study, the optimal number of clusters was defined from previous neuropathological studies and since the clustering was hierarchical agglomerative, the cut off could be specified in the desirable height (similarity). The authors chose the level corresponding to three and six clusters. In the 3-cluster level the patients were divided in the following subtypes: medial temporal (MT subtype), parietal dominant (P subtype) and diffuse atrophy (D subtype). In the next cut off, four clusters were defined, with the MT and P subtypes being the same, but the D subtype was divided in two further subtypes: medial frontal/temporal (MF) and the D subtypes.

3 Data

3.1 *Data sources*

The dataset used for this thesis consists of two large multicenter cohorts: the AddNeuroMed and the Alzheimer's disease Neuroimaging Initiative (ADNI). AddNeuroMed, a part of InnoMed (the Innovative Medicines Initiative) is an Integrated project funded by the European Union Sixth Framework program (Lovestone et al., 2007, 2009). The main objective of AddNeuroMedis to identify biomarkers or experimental models that can improve diagnosis, prediction and monitoring of disease progression in AD. The neuroimaging section of AddNeuroMed uses magnetic resonance imaging (MRI) and magnetic resonance spectoscopy (MRS) to extract valuable information for the identification of biomarkers for AD (Westman et al., 2011). The MRI data was collected from different centres across Europe (Lovestone et al., 2009; Simmons et al., 2001, 2011): University of Perugia (Italy), King's College London (United Kingdom), Aristotle University of Thessaloniki (Greece), University of Kuopio (Finland), University of Lodz (Poland) and University of Toulouse (France).

The ADNI dataset is an ongoing, longitudinal, multicenter study designed to develop imaging, clinical, genetic and biochemical biomarkers for the early detection and tracking of AD. ADNI began in 2004 with a six years fund of $67 million provided both by the private and public sectors, and comprises brain-imaging techniques such as positron emission tomography (PET) and structural MRI. These imaging techniques and biological markers are useful in clinical trials of mild cognitive impairment (MCI) and early AD. Similarly to the AddNeuroMedstudy in Europe, ADNI strives to reveal sensitive and specific markers of AD progression in U.S. patients from various sites across the country, so as to support the development of new treatments and monitor their effectiveness, as well as to reduce the expenditures of clinical trials.[2]

3.2 *MRI data acquisition*

The MRI data from ADNI and AddNeuroMed were acquired using a similar protocol (Jack et al., 2008). The imaging protocol for both studies consisted of a high resolution sagittal 3D T1-weighted MPRAGE volume (voxel size 1.1×1.1×1.2 mm^3) (Westman et al., 2011). Full brain and scull coverage was required for both datasets, and image quality control took place immediately after the images had been acquired at each site according to the AddNeuroMed’s quality control procedure (Simmons et al., 2009, 2011).

[2] More information about ADNI is available at http://adni.loni.usc.edu and http://www.adni-info.org

3.3 MRI surface based morphometric analysis

After acquiring the T1-weighted images, they were preprocessed using the FreeSurfer software (version 5.3), which provides cortical and subcortical measures of gray matter volume that can be later used for statistical analyses.

The FreeSurfer preprocessing stream consists of several steps: correction of motion artefacts and spatial distortions; removal of non-brain tissue (Segonne et al., 2004); automated transformation into the Talairach standard space; intensity normalization (Sled et al., 1998); segmentation of subcortical white matter and deep gray matter structures; tessellation of the gray/white matter boundary; automated topology correction (Segonne et al., 2007); and surface deformation to place the gray/white and gray/CSF borders (Fischl and Dale, 2000) (figure 1). Once the cortical models were complete, registration to a spherical atlas took place, which utilizes individual cortical folding patterns to match cortical geometry across subjects (Fischl et al., 1999). This was followed by parcellation of the cerebral cortex into 68 cortical regions using the atlas by Desikan et al. (2006) (Figure 2B/2C).

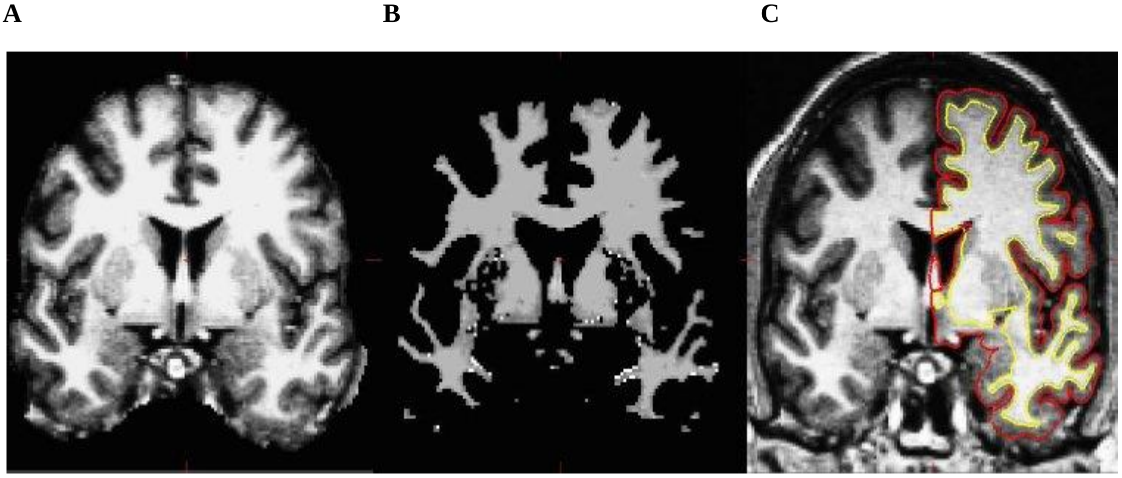


Figure 1. Three stages from the Freesurfer cortical analysis pipeline: A. skull stripped image. B. white matter segmentation. C. surface between the white and gray matter (yellow line) and between the gray and pia surface (red line) overlaid on the original volume.[3]

In addition to these 68 regions, seven subcortical structures from each hemisphere were also included: hippocampus, amygdala, thalamus, caudate, putamen, accumbens and pallidum (Figure 2)

[3] Source: http://www.freesurfer.net/fswiki/FreeSurferAnalysisPipelineOverview

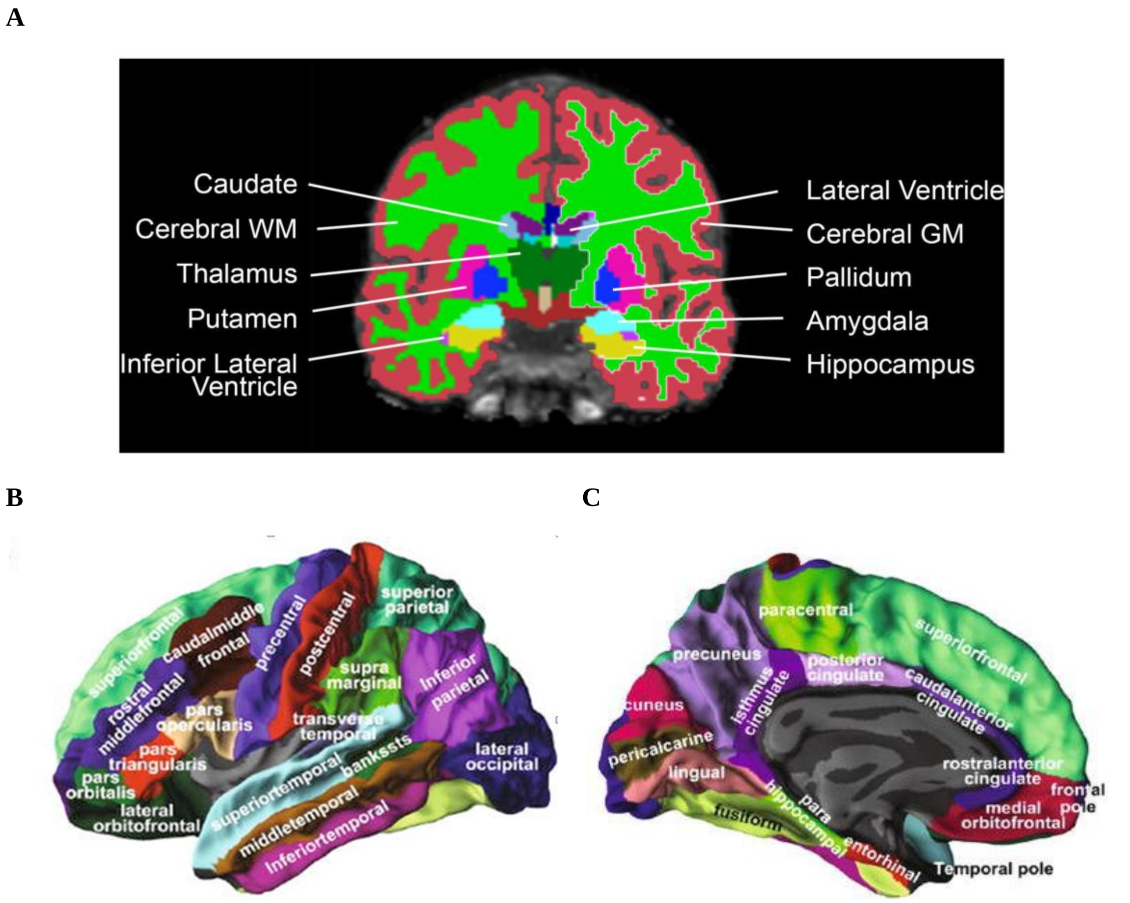


Figure 2. A. Volumetric stream generated tissue classes from a healthy subject illustrated by Freesurfer[4] (coronal view), the lateral (B) and medial (C) view of the grey matter surface illustrating the 34 regional cortical measures for one hemisphere.

Hence, the final dataset consists of 82 cortical and subcortical volume measures.[5] In a previous study, the cortical regions of interest (ROIs) provided by FreeSurfer have also been identified manually in order to calculate the similarity between the automated and manual process. The results were fairly promising, with a high accuracy between the two calculations. More specifically, the intraclass correlation coefficient (ICC) was 0.835 across all the ROIs and the mean distance error was less than 1mm (Desikan et al. 2006). This result suggests that this automated method is anatomically reliable and can be useful in studies assesing the cerebral cortex as well as in clinical studies aimed at tracking the evolution of disease-induced changes over time (Desikan et al. 2006).

[4] Source: Pinzka et al.(2015)

[5] More information on Freesurfer pipeline available at: http://www.freesurfer.net

3.4 Normalization of the data (ICV correction process)

In addition to the cortical and subcortical volume measures, FreeSurfer also provides the estimated total intracranial volume (eTIV) for each subject, which can be used to correct the volume measures by head size. The eTIV measure obtained by FreeSsurfer has been used in several studies for normalization (Westman et al., 2011, 2012, 2013) and is in good agreement with ICV inference segmentation acquired from proton density weighted images (Nordenskjöld et al., 2013). The importance of ICV correction is discussed by numerous authors and remains as a topic of investigation in the latest research (Zatz and Sernigan., 1983; Arndt et al., 1991; Sanfilipo et al., 2004; Pintzka et al., 2015).

There are two main branches in the head size adjustment literature: the proportion approach and the residual approach (Pintzka et al., 2015). The former approach normalizes the volumes of interest (VOI's) by simply computing the ratio between each VOI and the ICV to predict the ICV-adjusted volumes. The suitability of this method has been questioned by Barnes et al. (2010), on their investigation of the association between the head size and the number of cerebral structures in a group of control subjects. The residual method uses a linear regression to estimate the volume of a neuroanatomical structure and ICV calculated either by the control group or the entire dataset (O'brien et al., 2006). The adjusted volumes are obtained as follows:

$$Volume_{adj} \ = Volume - b(ICV - \overline{ICV})$$

where $Volume_{adj}$ is the ICV-adjusted volume, Volume is the original uncorrected volume, b is the slope of the linear regression of Volume to the ICV, ICV is the subject’s ICV, and $\overline{ICV}$ is the mean ICV across all subjects.

There is one study that investigates the different correlations between ICV and VOI under the two normalization methods (Voevodskaya et al., 2014). In that study, the proportional method resulted in coherent negative correlations between ICV and the gray matter structures. The residual method eliminated the correlation between ICV and volumes completely. Another study attempted to investigate possible marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures, using a cohort of 966 healthy subjects (Pintzka et al., 2015). The results concluded that the proportional method suffers from systematic errors due to lack of proportionality between neuroanatomical volumes and ICV, resulting in methodological mis-assignment of volumes smaller or larger than their actual size. Since the residual method has been proved to outperform the other ICV-correction techniques, we will use it for correcting the VOI's.

3.5 Data description

The data set used for the analysis is the combined ADNI and AddNeuroMed data set. Some observations has been excluded due to the presence of missing values, since the purpose of the thesis is not related to imputation methods. Alzheimer's patients have been listed in one data set (Table 1). A representative sample of control patients from both data sets has been listed in another dataset to be used for comparative and visualization reasons in the meta analyses.

Table 1 ADNI and AddNeuroMed AD patients characteristics[6]

	ADNI	AddNeuroMed	ADNI/AddNeuroMed
Number	155	116	271
Age	75.19 ± 7.22	75.56 ± 6.04	75.35 ± 6.73
Female/Male	76/79	78/38	154/117
Education	14.7 ± 3.1	8.03 ± 4.01	11.8 ± 4.8
MMSE	23.3 ± 2	20.9 ± 4.8	22.28 ± 3.65
CDR	0.74 ± 0.25	1.15 ± 0.47	0.92 ± 0.42
ADAS 1	6.06 ± 1.42	6.63 ± 1.49	6.30 ± 1.475

Data are presented int eh form mean ± sd, MMSE: Mini Mental State Examination, CDR: Clinical Dementia Rating, ADAS 1: Word list non-learning (mean).

The ROI's obtained for each subjects and used in the cluster analysis are presented in Table 2. Thirty four cortical thickness and seven volumetric measures from each hemisphere composed a dataset of eighty two variables. The measurement unit for the ROIs is mm^3.

[6] Some observations has been excluded from the data set due to the presence of missing values

Table 2. Variables included in the cluster analysis.

Cortical thickness measures		Volumetric measures	
1.	Banks of superior temporal sulcus	1.	Accumbens
2.	Caudal anterior cingulate	2.	Amygdala
3.	Caudal middle frontal gyrus	3.	Caudate
4.	Cuneus cortex	4.	Pallidum
5.	Entorhinal cortex	5.	Putamen
6.	Frontal pole	6.	Thalamus
7.	Fusiform gyrus	7.	Hippocampus
8.	Inferior parietal cortex		
9.	Inferior temporal gyrus		
10.	Insula		
11.	Isthmus of cingulate cortex		
12.	Lateral occipital cortex		
13.	Lateral orbitofrontal cortex		
14.	Lingual gyrus		
15.	Medial orbitofrontal cortex		
16.	Middle temporal gyrus		
17.	Paracentral sulcus		
18.	Parahippocampal gyrus		
19.	Parsopecularis		
20.	Parsorbitalis		
21.	Parstriangularis		
22.	Pericalcarine cortex		
23.	Postcental gyrus		
24.	Posterior cingulate cortex		
25.	Precentral gyrus		
26.	Precuneus cortex		
27.	Rostral anterior cingulate cortex		
28.	Rostral middle frontal gyrus		
29.	Superior frontal gyrus		
30.	Superior parietal gyrus		
31.	Superior temporal gyrus		
32.	Supramarginal gyrus		
33.	Temporal pole		
34.	Transverse temporal cortex		

These variables refer to the 34 cortical regions and the 7 subcortical regions of one hemisphere. Since we consider both the left and right hemisphere in the analysis, the final data set consists of 82 variables.

4 Methodology

This chapter is organized in line with the progress of the thesis. Initially, we address the high dimensional clustering problem and motivate the choice of methods to cope with it. Later on four sections are devoted in a description of the clustering algorithms and the semi supervised approach. Following the natural flow of the thesis the evaluation methods for the clustering algorithms are described.

4.1 High dimensional clustering

Cluster analysis seeks to discover groups or clusters of similar objects. The similarity between different objects is usually determined in terms of a distance measure. Objects that have small distances between them should be grouped together. Technology advances have made the data collection an easier process, resulting in larger and more complex datasets with many objects and dimensions. Traditional clustering algorithms consider all the dimensions of a data set in order to learn as much information as possible about each object in a dataset. However in high-dimensional datasets some of the dimensions are usually non informative. These dimensions can misguide clustering algorithms by hiding clusters in noisy data (Parsons et al., 2004).

Another reason why high-dimensional data are in need of particular data analysis methods is the well known *curse of dimensionality* phenomenon. By adding dimensions without adding objects the empty hypercubes increase exponentially (Bellman et al., 1961). As a result the objects spread out until, in very high-dimensional spaces they are almost equidistant from each other. This makes the distance measures increasingly meaningless and the study of the clustering properties a very problematic process. Hopefully, it can be assumed that high dimensional data live in subspaces with a dimension lower than the original one. This assumption is called *the empty space* phenomenon (Scott and Thomson, 1984).

Previous research on the subtypes of AD supports that approximately three distinct groups exist, each of them developing degeneration in different regions of the brain in general (Whitwell et al., 2012). This encourages the assumption that each cluster lives in a partially or totally different subspace of the 82 variables. Moreover the fact that each cluster might be more or less heterogeneous than another reinforces the chance of different variation within each cluster.

The clustering algorithms proposed in this thesis aim to fill the gap of the previous studies to account these important particularities, in the strain of studying the AD heterogeneity.

4.2 A Gaussian model for high-dimensional data

A popular clustering technique uses Gaussian Mixture Models (GMM), assuming that each class is represented by a Gaussian probability density (McLachlan, 1992). Having a data set $\{x_1, x_2, \dots, x_n\}$ of n data points in p dimensions, clustered in k homogeneous classes, then the mixture model has a density,

$$f(x,\theta) = \sum_{i=1}^{k} \pi_i \phi(x, \theta_i), \tag{4.2.1}$$

where ϕ is a p-variate Gaussian density with parameters $\theta_i = \{\mu_i, \Sigma_i\}$ and π_i is the mixing proportion of each cluster. The estimation of the mean vector, the covariance matrix and the mixing proportion of each cluster defines a k-cluster model. This model requires the estimation of the full covariance matrix for each cluster and consequently the parameters increases with the square of the dimension.

4.2.1 Low-dimensional class-specific subspaces GMM (HDDC)

The number of parameters to estimate in the GMM can be reduced if we introduce low-dimensional class-specific subspaces, which is the main idea of the high-dimensional Gaussian model of Bouveyron et al. (2007). Let Q_i be the orthogonal matrix with the eigenvectors of Σ_i as columns, then the class conditional covariance matrix Δ_i is defined in the eigenspace of Σ_i by $\Delta_i = Q_i^t \, \Sigma_i Q_i$.

Therefore, the matrix Δ_i is a diagonal matrix holding the eigenvalues of Σ_i. In addition, it is assumed that Δ_i matrix is divided in two parts[7]

$$\Delta_i = \begin{pmatrix} \begin{pmatrix} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{pmatrix} & 0 \\ 0 & \begin{pmatrix} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{pmatrix} \end{pmatrix},$$

where the first part has dimension d_i and the second has $(p - d_i)$. Furthermore, it is assumed that $a_{ij} > b_i$, $j = 1, \dots, d_i$ and where $d_i \in \{1, \dots, p-1\}$ is unknown. The class specific subspace E_i is defined as the affine space spanned by the d_i eigenvectors associated to the eigenvalues a_{ij} and such that $\mu_i \in E_i$. In the same trend, the affine space subspace $E_i^\perp$ is such that $E_i \oplus E_i^\perp = \Re^p$ and $\mu_i \in E_i^\perp$. In this subspace $E_i^\perp$ the variance of each class i is modeled with the single parameter b_i. Also, let $P_i(x) = \widetilde{Q}_i \widetilde{Q}_i^{\,t}(x - \mu_i) + \mu_i$, and $P_i^\perp(x) = \overline{Q}_i \overline{Q}_i^{\,t}(x - \mu_i) + \mu_i$ be the projection of x in E_i and $E_i^\perp$, respectively, where $\widetilde{Q}_i$ is made of the d_i columns of Q_i

[7] Here the word assumption does not refer to a mathematical assumption, but the motivation of the authors to assume two parts in the matrix Δ_i in order to study their properties.

supplemented by $(p - d_i)$ zero columns and $\bar{Q}_i = (Q_i - \tilde{Q}_i)$. Therefore, E_i is called the specific subspace of the ith group since most of the data live around or on this subspace. In addition, the dimension d_i of the subspace E_i can be considered as the intrinsic dimension of the ith group, i.e. the number of dimensions needed to describe the main features of group i. Following this notation system the mixture model of Bouveyron et al (2007) is denoted by $[a_{ij} b_i Q_i d_i]$.

Different variations of the mixed model $[a_{ij} b_i Q_i d_i]$ are obtained if we fix some parameters to be common within or between classes, yielding in models with different regularizations. For instance if we fix $Q_i = Q$ for all the groups, then the orientation of the groups will be common. The family $[a_{ij} b_i Q_i d_i]$ is divided in three categories: models with free orientation, common orientation and common covariance matrices. When we refer to common covariance matrices two models are considered from the family above: the model $[a_j, b, q, d]$ and the model $[a, b, q, d]$. In the first model we have more than one eigenvalues j for the intrinsic dimension of the cluster i, but these are common between the clusters i. This means that all the clusters have the same covariance matrix. The second model, namely $[a, b, q, d]$, refers to a model where the covariance matrix is expressed only by one eigenvalue and also it is common for all the clusters.

The advantage of estimating the different variations of this family for our data set is the ability to examine various possible underlying structures in the resulting clusters. The model that corresponds to the best scoring for different clustering evaluation criteria will be one that has the most appropriate underlying structures for these data. For instance, if the natural clusters of the data have common covariance matrices, the best model will belong in the category of the models with common covariance matrices.

Since the objective of this variation of mixture models is to shrink the number of estimated parameters, a symbolic example would provide a numerical comparison between the number of parameters of the classic GMM and the proposed version. In the particular case where of 100-dimensional data, made of 4 classes and with common intrinsic dimension d_i equal to 10, the model $[a_{ij} b_i Q_i d_i]$ needs the estimation of 4231 parameters, while the full GMM (the complete GMM model where all the parameters for each variable are considered) requires the estimation of 20603 parameters. Considering that $[a_{ij} b_i Q_i d_i]$ model is the most complex of this family, its variations have less parameters to be estimated for the previous data set varying from 4228 (for the model $[a_{ij} b Q_i d_i]$) to 1351 for the least complex model, namely $[a, b, q, d]$. All the additional models of this family have a complexity between those of the two models above.

4.2.2 The EM algorithm for the proposed GMM and its sub-models

The Expectation Maximization (EM) algorithm is often used in model based clustering for the estimation of the parameters $\theta = \{\pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k\}$ with $\theta_i = \{\mu_i, \Sigma_i\}$. The algorithm repeats iteratively the E step, which creates a function for the expectation of the log-likelihood evaluated using the current estimates for the parameters and the M step, which computes parameters maximizing the expected log-likelihood found on the E step, until the maximum number of iterations defined by the user is reached or a termination criterion is achieved. More information about the theory of EM and its extensions can be found in McLachlan and Krishnan (1997). In the case of the GMM proposed by Bouveyron et al (2007), the parameters to be estimated by the EM algorithm are $\theta = \{\pi_i, \mu_i, \Sigma_i, a_{ij}, b_i, Q_i, d_i\}$. A short presentation of the EM algorithm for the $[a_{ij}\, b_i Q_i d_i]$ model is detailed here (Table 3).

Table 3 High dimensional data clustering. The EM algorithm

```
1:   Initialize $\theta = \{\pi_i, \mu_i, \Sigma_i, a_{ij}, b_i, Q_i, d_i\}$
2:   Repeat
3:       E step
4:           for each $i = 1, \dots, k$ and $j = 1, \dots, n$
5:           $t_{ij}^{(q)} = P(x_j \in C_i^{(q-1)} | x_j)$ for the fuzzy class $C_i$, which can be written from (4.2.1) as:
6:           $t_{ij}^{(q)} = 1/ \sum_{\ell=1}^{k} \exp\left(\frac{1}{2}\left(K_i(x_j) - K_\ell(x_j)\right)\right)$, where $K_i(x)$
7:           $K_i(x) = ||\mu_i - P_i(x)||_{A_i}^2 + \frac{1}{b_i}||x - P_i(x)||^2 + \sum_{j=1}^{d_i} \log(a_{ij}) + (p - d_i)\log(b_i) - 2\log(\pi_i)$
8:       M step
9:           $\pi_i^{(q)} = \frac{n_i^{(q)}}{n}$, $\hat{\mu}_i^{(q)} = \sum_{j=1}^{n} t_{ij}^{(q)} x_j$ , where $n_i^{(q)} = \sum_{j=1}^{n} t_{ij}^{(q)}$
10:          $W_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} \left(x_j - \hat{\mu}_i^{(q)}\right)\left(x_j - \hat{\mu}_i^{(q)}\right)^t$
11:          The $d_i$ first columns of $Q_i$ are estimated by the eigenvectors associated with the $d_i$ largest
             eigenvalues $\lambda_{ij}$ of $W_i$
12:          The estimator of $b_{ij}$ is $\hat{b}_i = \frac{1}{(p-d_i)}\left(Tr(W_i) - \sum_{j=1}^{d_i} \lambda_{ij}\right)$
13:          The estimator of $a_{ij}$ is $a_{ij} = \lambda_{ij}$
14:          The estimation of the number of intrinsic dimensions $d_i$ of each subclass is defined with the
             scree-test of Cattel (1966)
15:  until q times
```

In line number 7, $K_i(x)$ is called the cost function and it is defined mainly by two distances: the distance between the projection of x on the subspace E_i and the mean of the class and the distance between the observation and the subspace E_i. This cost function favors the assignment of a new observation to the class for which it is close to the subspace (first distance) and for which its projection is on the class subspace is close to the mean of the class (second distance). The variance terms a_{ij} , b_i balance the importance of both distances. For instance if the data are quite noisy, b_i which represents the noise variance will be large and therefore it is natural to balance the distance $||x - P(x)||^2$ by $1 \,/\, b_i$, so as to take into account the variance in $E_i^{\perp}$.

In the line 10, $W_i^{(q)}$ stands for the empirical covariance matrix of the fuzzy class C_i. The maximization step changes for different parameterizations of the $[a_{ij} b_i Q_i d_i]$ model. The estimators of a, b for the remaining models can be found in Bouveyron et al. (2007). The proofs of the results presented above are in Bouveyron et al. (2006).

In the line 14 the estimation of d_i is discussed. Their approach is based on the eigenvalues of the class conditional covariance matrix Σ_i of the class C_i. The j_{th} eigenvalue of Σ_i corresponds to the fraction of the full variance carried by the j_{th} eigenvector of Σ_i. The class specific dimension d_i, for $i = 1, \dots, k$ is estimated with the aid of the scree-test of Catell which looks for a break in the eigenvalues scree. The dimension for which the subsequent eigenvalues differences are smaller than a threshold is selected as d_i. The threshold used is the Bayesian information criterion (BIC) (Schwarz, 1978), which consists on minimizing $BIC(m) = -2\log(L) + v(m)\log(n)$. Here, L is the log likelihood of the model m, $v(m)$ is the number of variables of model m, n is the number of observations. Since the choice of number of clusters can be understood as a model choice, the BIC criterion may help to define the optimal number of clusters later in the evaluation.

4.3 Correlation clustering

At this place I will introduce the semi supervised approach. The correlation clustering followed by its incorporation in the GMM is explained in this section.

Correlation clustering introduced by Bansal et al. (2004), is an NP-hard task of separating a given data set of objects into groups based on a known pairwise similarity measure between the objects. More intuitively, we can consider the input of this problem as an undirected graph over a set of nodes (representing the data objects to be clustered), with positive or negative edges indicating that two objects are similar or dissimilar, respectively. The objective of correlation clustering, is to minimize the sum of the number of positive edges between different partitions and the number of negative edges within the partitions. In this direction, a partition will host objects that are similar. At the same time, different partitions will not host similar points. This is the main goal of this task and if we consider the two main properties of a good clustering which are compactness (to form compact clusters) and isolation (isolated clusters from each other) (Jain and Dubes. 1988), the aim of the algorithm is to fulfill these properties. In the literature of correlation clustering many different approximate and local search algorithms have been proposed in the effort of addressing this minimization problem (Ailon et al., 2008; Charikar et al., 2005; Giotis and Guruswami, 2006). Unfortunately, such implementations do not provide optimality guarantees on the produced clustering.

4.3.1 Correlation clustering formulation

Berg and Järvisalo (2013) introduced a framework for correlation clustering using the Maximum satisfiability (MaxSAT) Boolean optimization paradigm. Their approach is based on formulating the correlation clustering task in an exact manner as MaxSAT, and then using a MaxSAT solver for finding clusterings by solving the MaxSAT formulation. This approach discovers optimal clusterings.

Following the definition of Bansal et al. (2004), for the binary similarity case, a correlation clustering case consists of a set $V = \{u_1, u_2, \dots u_N\}$ of objects, and a binary similarity function $x: E \rightarrow \{0,1\}$ over a subset $E \subset V \times V$ of the ordered pairs of the objects. We assume that $s(u_i, u_j)$ is symmetric, i.e., that $s(u_i, u_j) = s(u_j, u_i)$ for any two objects u_i, u_j (undirected graph). Two objects are considered similar if $s(u_i, u_j) = 1$, and dissimilar if $s(u_i, u_j) = 0$. A correlation clustering instance (V, s) can be interpreted as an undirected graph with the set V of nodes and two types of labelled edge relations: $E^+ = \left\{\{u_i, u_j\} \middle| s(u_i, u_j) = 1\right\}$ (representing similar pair of nodes) and $E^- \left\{\{u_i, u_j\} \middle| s(u_i, u_j) = 0\right\}$ (representing dissimilar pair of nodes).

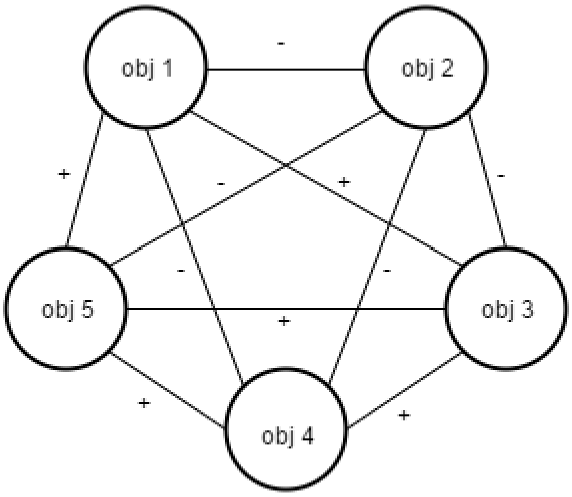


Figure 3: A graph representation of the correlation clustering instance with a binary similarity measure

Any function $cl: V \rightarrow \mathbb{N}$ is a solution to the correlation clustering instance, representing a clustering of objects into clusters indexed by numbers depending on the cluster that they belong. In correlation clustering the objective is to find a clustering for the objects in a way that correlates as well as possible with the similarity measure s, i.e, to find a function $cl: V \rightarrow \mathbb{N}$ minimizing the cost function

$$G(cl) = \sum_{\substack{(u_x, u_y) \in E \\ cl(u_x) = cl(u_y)}} \left(1 - s(u_x, u_y)\right) + \sum_{\substack{(u_x, u_y) \in E \\ cl(u_x) \neq cl(u_y)}} s(u_x, u_y).$$

A clustering cl of V is optimal iff $G(cl) \leq G(cl')$ for any clustering $cl'\ of\ V.$

4.3.2 Implementation of the optimal correlation clustering MaxSAT algorithm

Being inspired from Berg and Järvisalo algorithm I encoded a slightly different alternative in Answer Set Programming (ASP) language using the software CLINGO (Gebser et al., 2008). The algorithm above uses binary similarities, which means that two objects can be either similar or dissimilar. Although they employ a binary similarity measure, this cannot be applied to the data set of Alzheimer's patients because it oversimplifies the relationship of the patients characteristics. For the needs of this particular dataset an ordinal similarity measure has been employed under the assumption that one patient might present different levels of similarity with another patient in the patterns on atrophy. Different distance measures, including Euclidean, Minkowski and Mahalanobis, are proposed in the literature of clustering. The choice of Mahalanobis distance measure is motivated by the fact that the Euclidean and Minkowski distances are safely applied in data sets, where the variances are equal. This hypothesis is not considered accurate for the AD data set. In view of the fact that different regions of the brain may interact with each others with respect to the patterns of atrophy, the covariance of the variables may perhaps reveal useful intelligence. The squared Mahalanobis distance $\left(x_i - x_j\right)^T S^{-1}(x_i - x_j)$ includes this piece of information. This distance measure has been computed for the objects and recoded to values from one to eight. The similarity $s\left(u_i, u_j\right)$ takes finally the values $s\left(u_i, u_j\right) = \{1,2,3,4,5,6,7,8\}$.

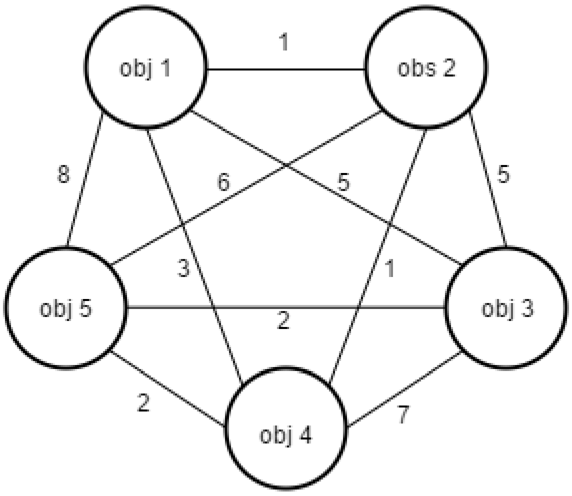


Figure 4: A graph representation of the correlation clustering instance with an ordinal similarity measure

The complexity of the algorithm for the ordinal similarity measure is quite high and thus prevents the computation of the optimal allocation of the whole data set in clusters. In order to provide a more intuitive perspective, an ASP implementation of the algorithm is presented and explained below (Table 4).

Table 4 ASP encoding for the correlation clustering with ordinal similarity measure.

1: **Fact**	const n=4.
2: **Fact**	node(1..100).
3: **Rule**	{ color(X, 1..n) } = 1 :- node(X).
4: **Rule**	cluster(X, Y, C) :- color(X, C) , color(Y, C) , X<Y.
5: **Optimization statement**	: ∼ cluster(X, Y, C), dist(X, Y, W). [W, X, Y]

In line 1 we define the number of cluster which is 4 in this example. In line 2 we define the number of data points of the data set which is 100 in our case. In line 3 we assign exactly one color ($1 \leq color \leq n$) to each object. We can see it as assigning each object in exactly one cluster. The word color could be the word cluster instead, but I do not use it in order to avoid confusion with the cluster word later in the encoding. In line 4 we have a rule stating that if X has the color C, Y has the color C and $X<Y$, then X and Y are assigned to the same cluster C. Finally, line 5 denotes an optimization process (the sign $\sim$). This optimization process is a minimization one. We ask from the ASP solver to look into the possible solutions of the filtered variable *cluster(X, Y, C)* and also look into the input variable of the algorithm, namely *dist(X,*

Y, *W)*[8] (in this case we focus in minimizing the intracluster distances/ similarities). The solver will try to minimize the sum of the costs W, which is the ordinal similarity measure $s(u_i, u_j) = \{1,2,3,4,5,6,7,8\}$. This optimization will result to the optimal clustering allocation of the objects X and Y for a given cost W. Line 8 can also be written as: #minimize { W, X, Y, cluster (X,Y,C), dist(X, Y, W) }.

The optimal allocation was mined for the first 100 objects only, defining the best clustering for these patients given the similarity measure explained above.

4.3.3 A combination of two clustering methods: the semi supervised approach

The ASP encoded correlation clustering is a rather powerful algorithm to find an optimal allocation for objects in a data set. However, its computational inability to cluster the whole data set of patients makes it impossible to use it as a universal clustering method for our data set. Basu et al. (2002) on their study in semi-supervised clustering methods, discussed the capability of using an amount of labeled data in the process of the unsupervised learning in order to increase the clustering quality. In semi-supervised clustering, a number of labeled data is used along with the unlabeled data to obtain better clustering.[9] Proper seeding biases the clustering towards a superior region of the search space, thus decreasing the odds of it getting stuck in poor local optima, while at the same time producing a clustering result comparable to the user specified labels.

The high dimensional data clustering method explained in chapter 3 uses an EM algorithm to estimate the parameters of the Gaussian mixture model for clustering the data set. The choice of initial values is of great importance in the algorithm, since it can profoundly influence the ability to discover the global maximum and effect the speed of convergence of the algorithm (Karlis and Xekalaki, 2003).

The strategy under consideration in this thesis is as follows. Firstly, 100 patients will be assigned to clusters with the aid of the optimal correlation clustering via ASP explained in section 4.3.3. After allocating these patients into clusters, we apply the HDDC algorithm from section 4.2.2 using the optimal allocation of the correlation clustering as initial values for the first 100 patients.

4.4 Bayesian clustering with variable selection

When clustering high-dimensional data sets, the variable selection or projection into a subspace seems inevitable in an attempt to cope with the *curse of dimensionality* and the *empty space phenomenon*. The clustering method explained in chapter 3.2

[8] The $dist(X, Y, W)$ is the variable that holds the actual similarities of the observations

[9] The semi-supervised learning approach followed here refers to the semi-supervised by seeding and it is motivated by the study of Basu et al. (2002)

confronts the high-dimensional data clustering problem from the subspace projection perspective. In the process of comprising a comparative analysis, another clustering algorithm that tackles the same problem from a variable selection viewpoint is applied to the AD data set. The assessment of different statistical clustering perspectives, gave space to a method that challenges the high-dimensional problem in a Bayesian framework with prior distributions both for the allocation of objects into clusters and for the model parameters.

In the clustering literature the approaches of allocating data in different groups abide by two main strains, the distance and the model based clustering. The former reflects in the use of a clustering algorithm given a distance measure between the data objects, while the main goal is to group data objects that have small distance together. The model based approach also called parametric approach assumes that each cluster follows a probability distribution, thus the problem of clustering converts to a problem of estimating the parameters of each cluster distribution (Fraley and Raftery. 2002). A new method that combines the two main approaches in clustering, incorporating the variable selection feature in a Bayesian framework has been created by Nia (2009).

4.4.1 *A Bayesian model for clustering data*

The Bayesian regression created by Nia (2009) has its priors chosen in such way that the marginal posteriors are analytically tractable, yielding in a fast algorithm serving the needs of a fast clustering method. The marginal posterior is used here as a measure of suitability of a grouping. The allocation that maximizes the marginal posterior is the optimal under this notion, but it is not feasible to compute this value for all the possible divisions of data objects. The approximation suggested by Nia (2009) is the employment of an agglomerative algorithm in the search for the maximum aposteriori clustering. This means that the agglomerative hierarchical clustering is applied, but the measure of distance is not a physical distance but the marginal posterior. In such a way, we can overcome weaknesses arising in the agglomerative hierarchical model with respect to natural distances in high dimensional spaces.

In mathematical terms we consider c clusters, $c \in \{1, \dots, C\}$, that consist of T_c observations, and R_{ct} replicates of the t-th data object. Moreover, we assume that V variables are measured for each replicate, $t \in \{1, \dots, T_c\}$. The total number of data objects is $T = \sum_{c=1}^{C} T_c$, total number of replicates is $\sum_{c=1}^{C} \sum_{t=1}^{T} R_{ct}$, and the total number of measurements equals to $V = \sum_{c=1}^{C} \sum_{t=1}^{T} R_{ct}$, since for each variable we need to observe the values in replicate. The basic linear model proposed by Nia (2009) for the underlying model of the data has the following form

$$y_{uctr} = \mu + \gamma_{uc}\theta_{uc} + \eta_{uct} + \epsilon_{uctr}\,, \qquad (4.4.1)$$
$$u = 1, \dots, V\,, c = 1, \dots, C\,, t = 1, \dots, T_c\,, r = 1, \dots, R_{ct}$$

where$-\infty \leq \mu \leq \infty$ and $\theta_{uc} \overset{iid}{\sim} N(0,\ \sigma_\theta^2)$, $\eta_{uct} \overset{iid}{\sim} N(0,\ \sigma_\eta^2)$, $\epsilon_{uctr} \overset{iid}{\sim} N(0\,, \sigma^2)$, with $\sigma^2, \sigma_\theta^2 > 0\,, \sigma_\eta^2 \geq 0\,, \gamma_{uc}$is a Bernoulli distributed variable satisfying $P(\gamma_{uc} = 1) = p$. In equation (4.4.1), μ stands as a general mean for all the variables and data objects. Referring to the case where $\gamma_{uc} = 1$, then the analogous variable u, class c combination is said to be *active* and in the ideal configuration its mean would be $\mu + \theta_{uc}$. In the complementary case, where $\gamma_{uc} = 0$, the variable-class combination is *inactive* and in the ideal situation its mean would be μ. Nevertheless, no realizable setting is optimal, and additional variation between data objects, maybe as a result of the experimental conditions can be modeled by the normally distributed variable η_{uct}, pointing to a mean $\mu + \theta + \eta_{uct}$ for the t-th data object and the variable-class combination. Additional unpredictability between replicates is assumed to come from the measurement error ϵ_{uctr}. According to the description above we can conclude that the model (4.4.1) is a variant of the classical mixed effects model in which a random component disappears if the Bernoulli variable $\gamma_{uc} = 0$.

Although the model (4.4.1) corresponds to the problem in a variable level pretty well, it assumes that the variable cluster combinations are independent. A natural extension is to model whether each variable is *active* with a second Bernoulli level, yielding

$$y_{uctr} = \mu + \delta_u\gamma_{uc}\theta_{uc} + \eta_{uct} + \epsilon_{uctr}, \qquad (4.4.2)$$
$$u = 1, \dots, V\,, c = 1, \dots, C\,, t = 1, \dots, T_c\,, r = 1, \dots, R_{ct}$$

where δ_u are independent Bernoulli distributed variables with probability q and all the rest are the same as in formula (4.4.1). Therefore q is the proportion of *active* variables and p is the proportion of *active* classes, given that a variable is *active.* In this way a build in variable selection feature is added, solving the problem of the *curse of dimensionality.* Spike-and-slab models are often used as a variable selection tool in Bayesian regression models (George and McCulloch., 1997). The name of these models comes from the structure of the Spike-and-slab distribution which is a mixture of a point mass (at zero) and a continuous distribution away from zero (the slab) for the regression parameters. In the variable selection model designed by George and McCulloch a mixture of two Gaussian distributions is considered, assigning a mixture of distributions having the same support, so allowing a Gibbs sampler to sample from the posterior distribution. This formulation helps to separate negligible from true effects, because negligible effects are expected to appear from the spike prior which does will not affect the classification, while the true effects are expected to yield from the slab prior, which directs the classification. In this way the classification procedures which are robust to uninformative variables. In the model (4.4.2) of Nia (2009), a Gaussian prior is always assigned to the spike, but various distributions can be assigned in the slab, depending on the modeling of the effects assumptions. For the needs of the thesis a Gaussian slab is assumed for the effects since there is no prior knowledge indicating heavy tails for the variable-class combinations. By managing the appearance of a variable-class effect with the variable δ_u in the formula (4.4.2), the model can be described in a hierarchy as follows:

$$\begin{gathered} y_{uct}\,|\eta_{uct} \overset{iid}{\sim} N(\eta_{uct}, \sigma^2), \\ \eta_{uct}\,|\theta_{uc} \overset{iid}{\sim} N\big(\theta_{uc}, \sigma_\eta^2\big), \\ \theta_{uc}\,|\gamma_{uc} \overset{iid}{\sim} N\big(\mu, \gamma_{uc}\sigma_\theta^2\big), \\ \gamma_{uc} \overset{iid}{\sim} B(\delta_u p), \\ \delta_u \overset{iid}{\sim} B(q) \end{gathered} \tag{4.4.3}$$

In the data set used for the needs of the thesis, there are no replicates of data objects, so $R_{ct} = 1$. For simplicity, with fewer indices let y be a vector of measured quantities. For instance, y_u denotes the data available for variable u, y_c denotes the data in class c, y_{uc} denotes the data available for the uth variable and the cth class. Now let f refer to a generic probability density. The model (4.4.3), imposes independent variables, thus $f_y = \prod_{u=1}^{V} f(y_u)$. The joint density of the data y_u for a variable u is

$$f(y_u) = q(f(y_u|\delta_u = 1) + (1 - q)f(y_u|\delta_u = 0) \tag{4.4.4}$$

The mixture proportions of the distribution (4.4.4) are $(1 - q)$ and q for the spike and slab components respectively.

For the $f(y_u|\delta_u = 0)$ component of this mixture distribution we have $f(y_u|\delta_u = 0) = \prod_{c=1}^{C}\prod_{t=1}^{T_c} f_0(y_{uct})$, because when $\delta_u = 0$, no variable class combination is active. Also, $f_0(y_{uct}) = f(y_{uct}|\delta_u = 0) = f(y_{uct}|\delta_u = 1, \gamma_{uc} = 0)$. For the second component of the mixture (4.4.4) we have $f(y_u|\delta_u = 1) = \prod_{c=1}^{C} f(y_{uc})$, because in this case the variable u is active and only data in different classes are independent, but data within a class are not independent so as to simplify by adding one more product as in the case of $f(y_u|\delta_u = 0)$

For the active variables $\delta_u = 1$ we have

$$f(y_{uc}|\delta_u = 1) = pf_1(y_{uc}) + (1-p)\prod_{t=1}^{T_c} f_0(y_{uct}) \qquad (4.4.5)$$

where $f_1(y_{uc}) = f(y_{uc}|\delta_u = 1, \gamma_{uc} = 1)$, is a density with an *active* variable-class(cluster) combination, sharing the same θ_{uc}. The second density is the $f(y_{uc}|\delta_u = 1, \gamma_{uc} = 0)$, but it is expressed as $\prod_{t=1}^{T_c} f_0(y_{uct})$, because when the variable-class(cluster) combination is *inactive*, the data objects inside a cluster are independent. Consequently, the definition of $f_0(y_{uct})$ and $f_1(y_{uc})$ is needed so as to completely define formula (4.4.3). For $f_0(y_{uct})$ and $f_1(y_{uc})$ we have

$$f_0(y_{uct}) = \left[2\pi\left(\sigma_\eta^2 + \sigma^2\right)\right]^{-1/2} \times \exp\left(-\frac{(\bar{y}_{uct} - \mu)^2}{2(\sigma_\eta^2 + \sigma^2 R_{ct})}\right) \qquad (4.4.6)$$

$$f_1(y_{uc}) = \left[2\pi\left(\sigma_\theta^2 + \sigma_\eta^2 + \sigma^2\right)\right]^{-1/2} \times \exp\left(-\frac{(\bar{y}_{uct} - \mu)^2}{2(\sigma_\theta^2 + \sigma_\eta^2 + \sigma^2 R_{ct})}\right) \qquad (4.4.7)$$

The proofs of these formulas are listed in (Nia and Davison, 2014), while further information on the formulas and the model can be found in (Nia, 2009; Nia and Davison, 2012).

The hyperparameters $\phi = (\mu, \sigma^2, \sigma_\eta^2, \sigma_\theta^2, p)$ of the prior density can be estimated by maximizing the log likelihood

$$\ell(\phi) = \sum_{u=1}^{V}\sum_{c=1}^{C} logf(y_{uc};\phi) \qquad (4.4.8)$$

Here, it has to be noted that in this case where the data are unreplicated, only $\sigma^2 + \sigma_\eta^2$ are estimable, but by fixing the variance $\sigma_\eta^2 = 0$ we can estimate the other parameters. After maximizing the hyperparameters of the model, the marginal density is fixed and this yields to a fast algorithm used in the agglomerative clustering.

4.4.2 The clustering prior and a clustering paradigm

In the agglomerative hierarchical clustering we initially allocate all the data objects in different clusters and then we merge data objects having the minimum distance between them, until there is only one cluster. This approach has the advantage that through a dendrogram, the visual representation of the cluster merging in different cluster-levels is possible. However, a distance measure is required for computing the minimum distance between clusters. This measure is provided from a probability model through the chance in posterior when the clusters are merged.

Considering a partition C of T data objects partitioned into $|C| = C \in \{1, \dots, T\}$ blocks, with T_1, T_2 data objects belong in clusters 1, 2 respectively and so forth. Under the assumption of prior exchangeability in the grouping of data objects, we need to specify only a prior for the number of blocks in a partition and for their sizes. A uniform discrete prior $\Pr(C) = \frac{1}{T}(C = 1, \dots, T)$, for the distinct clusters of the partition, and a Dirichlet multinomial prior for the cluster sizes $T_1, \dots, T_C$ given C is proposed by (Heard et al., 2006). This prior parameterization yields to

$$\Pr(C) \propto \frac{(\mathrm{C}-1)!\,\mathrm{T}_1!\dots\mathrm{T}_\mathrm{C}!}{\mathrm{T}(\mathrm{T}+\mathrm{C}-1)!}, \qquad \sum_{c=1}^{C} T_C = T \tag{4.4.9}$$

In the algorithm of Nia and Davison, (2015) every data object is initially regarded having its own cluster, so the first partition has T blocks, reflecting to the T data objects. At each step the algorithm calculates every possible merger of pairs of blocks and the posterior probability is calculated. The merger with the maximum posterior probability of the resulting partition is applied. For a current partition C and data comprising its $T + 1 - C$ blocks denoted by $y_1, \dots, y_{T+1-C}$, each of the blocks contains $T_1, \dots, T_C$ data objects respective. Assume that a new partition C' merging blocks y_i, y_j of C is proposed to form a new block whose data are y_{ij}. Therefore, since the only change between the partition C and the partition C' regards the blocks y_i and y_j, the ratio of the posterior probabilities for C and C' is

$$\frac{\Pr(C') \prod_{c \in C'} f(y_C; \phi)}{\Pr(C) \prod_{c \in C} f(y_C; \phi} = \frac{(T + C - 1)(T_i + T_j)!}{(C-1) T_i! \, T_j!} * \frac{f(y_{ij}; \phi)}{f(y_i; \phi) f(y_j; \phi)}, \tag{4.4.10}$$

where $f(y_j; \phi)$ indicated the marginal density of the data for the data objects in block y_j. The new partition is the one that maximizes (4.4.10) over all the possible pairs of blocks of C. Comparisons between different partitions can be made with the aid of a natural scale as the log (4.4.10). This scale can be used as a distance measure for building the dendrogram of the clustering. A signed difference between marginal log posteriors will be used as a monotone height function that serves for the dendrogram drawing. That is, $\hat{\ell}'_C = \hat{\ell}_C - \max_C \hat{\ell}_C$, where $\hat{\ell}_C$is the log marginal posterior for partition C. Consequently, $\hat{\ell}'_{\hat{C}} = 0$ is the log marginal posterior of the optimal clustering $\hat{C}$ found by the algorithm. Finally, the length $|\hat{\ell}'_C - \hat{\ell}'_{C'}|$, between two successive partitions $C \subset C'$ is used to build the dendrogram of the agglomerative hierarchical clustering.

The number of computations for this algorithm is of order $O(VT^3)$, apart from the initial maximization of the (4.4.8) needed for the estimation of the parameters of the model. This approach in the line with hierarchical clustering, does not challenge the exhaustive search of all the possible dendrograms in the effort to find the one with the maximum marginal posterior. Different algorithms for searching on the global optimal clustering using the marginal posterior optimization criterion can be explored using Markov Chain Monte Carlo methods.

4.5 Evaluation of the clustering

A clustering algorithm is used to group together data objects which are "close" to one another in a multidimensional feature space, often aiming to uncover the inherent structure which the data possesses. A variety of clustering methods exists, many of which have shown good results in datasets where the data allocation is known. The measures of clustering quality can be divided in *internal* and *external* criteria. In an optimal scenario where the data allocation is already known, *external criteria* like the RAND index are used to evaluate the optimal clustering method and therefore the method that results in an allocation which lies closer to the real one is the preferable. Unfortunately in real world data sets, information about the object allocation is usually unknown. A more pragmatic approach, would be to assume that such knowledge is not available and try to look at the evaluation of a clustering from a different perspective. The criteria that use no information about the true allocation of the objects are called *internal* criteria.

A variety of criteria aiming at the validation of a clustering procedure, as well as the determination of the optimal allocation of objects for a particular experiment have been proposed in the literature. More information on different criteria can be found in Ben-David and Ackerman (2009). As previously explained in section , the clustering quality is defined in terms of compactness and isolation. A successful clustering allocation aims to produce groups of objects that are compact within themselves and also separated from each others. For assessing the compactness of a cluster, *intra cluster* measures can provide an adequate metric whereas for assessing the separation of a clustering outcome, intercluster measures can be used. Measures that incorporate both the intercluster and intracluster behavior can also be employed. The motivation for developing indices that combine compactness and separation yields from the fact that these two measures demonstrate opposing trends (separation decreases with the number of clusters while compactness increases), therefore their study separately will return opposing optimal clusterings (Datta et al., 2008).

4.5.1 Internal validity measures

Various indices have been suggested in the literature to cover the clustering validation gap. The measures used for the needs of the thesis are in two tracks: measures of connectedness and measures of non linear combinations of compactness and separation.

Connectedness

Connectedness measures to what extent objects are placed in the same cluster as their nearest neighbor in the data space. The connectivity measure used for the evaluation of the connectedness of the clusters is the connectivity measure introduced by Handl et al. (2005). It evaluates the degree to which neighboring objects have been placed in the same cluster. Define $nn_{i(j)}$ as the nearest neighbor of the object i and let $x_{i,nn_{i(j)}}$be zero if $nn_{i(j)}$ and i are in the same cluster and $1/j$ otherwise. More formally

$$x_{i,nn_{i(j)}} = \begin{cases} \frac{1}{j} \text{ if} \nexists C_k \colon i, nn_{i(j)} \in C_k \\ 0 \qquad\qquad otherwise, \end{cases}$$

where $C = \{C_1, \dots, C_K\}$ is a specific clustering partition, of the N observations in K disjoint clusters. Then the connectivity index is defined as

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} nn_{i(j)},$$

where L is the parameter determining the number of neighbors that contribute to the connectivity measure (Handl and Knowles., 2005). The connectedness should be minimized.

Combination of compactness and separation measures

Silhouette index

A quite popular measure for evaluating clustering quality diachronically is the silhouette width, that is the average of each object's silhouette value (Rousseeuw., 1987). The silhouette value accounts the degree of confidence in the clustering assignment of a specific object, with poorly clustered observations having values near -1 and well-clustered objects having value around 1. For object i two distances have to be counted firstly,

$$\alpha_i = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist(i,j), \quad b_i = \min_{(C_k \in C \backslash C(i))} \sum_{j \in C_k} \frac{dist\,(i,j)}{n(C_k)},$$

where α_i is the average distance between i and all other observations in the same cluster and b_i is the lowest average distance between i and the objects in the "nearest neighboring cluster". $C(i)$ is the cluster containing the observation i, $dist(i,j)$ is the distance between objects i and j and $n(C)$ is the cardinality of cluster C. Then the silhouette index is

$$S(i) = \frac{(b_i - a_i)}{\max\,(b_i, a_i)}, -1 \le S(i) \le 1$$

The silhouette index should be maximized and this intuitively can be understood by looking at the fraction of the silhouette measure, as maximizing the distance between an object from one cluster and the objects from another with the aid ob b_i (separation), while minimizing the distance between an object of a cluster and the remaining objects at this cluster with the aid of a_i (compactness).

Calinski-Harabasz index

Another index that provides valuable evidence in the process of comparing clustering allocations is the Calinski-Harbasz criterion (Calinski and Harabasz., 1974). The Calinski-Harabasz (CH), also called *error-variance criterion*, relies in the assumption that well defined clusters are meant to have a large between cluster variance and a small within-cluster variance.

Namely the overall between cluster variance $SS_B$is defined as

$$SS_B = \sum_{i=1}^{k} n_i \parallel m_i - m \parallel^2,$$

where k is the number of clusters, m_i is the centroid of cluster i, m is the overall mean of the sample data, and $\parallel m_i - m \parallel$ is the L^2 Euclidean distance between vectors in case that we use Euclidean distance as distance measure between objects. The SS_B can be referred to us, as the intercluster component of the CH formula.

The overall within cluster variance SS_W equals to

$$SS_W = \sum_{i=1}^{k} \sum_{x \in C_i} \parallel x - m_i \parallel^2,$$

where k is the number of clusters, C_i is the i_{th} cluster, m_i is the centroid of cluster i, x is an object, and $\parallel x - m_i \parallel$ is the L^2 is the Euclidean distance between the two vectors in the case where Euclidean distance is the distance measure under consideration. It is considered as the intracluster part of CH formula.

Finally, the Calinski-Harabasz index can be defined as

$$VRC_K = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1},$$

where N is the number of objects. We can understand from the fraction above that the the larger is the index the better is the clustering, since what serves a good clustering is a large sum of the squares of the variance between different clusters and the opposite within clusters.

Davies-Bouldin measure

Another index that is present in most works in clustering validation is the Davies-Bouldin (DB) index (Davies and Bouldin., 1979). The contribution of this index in the clustering evaluation literature can be addressed in its generality. Instead of simply proposing a cluster index, Davies and Bouldin formulated a general framework of outcomes of clustering algorithms. The DB index may be defined as

$$DB(K) = \frac{1}{K}\sum_{k=1}^{K} R_k \ \ for\ K \in \mathbb{N},$$

where

$$R_k = \max_{j=1,\dots,K, j\neq k} \left\{ \frac{\sigma_k + \sigma_j}{d(c_i, c_j)} \right\}, for\ k \in [1, \dots, K]$$

and

$$\sigma_k = \frac{1}{\sum_{i=1}^{N} w_{k,i}} \sum_{i=1}^{N} w_{k,i} \parallel x_i - \overline{x_k} \parallel, for\ k \in [1, \dots, K]$$

and also

$$d(c_i, c_j) = \parallel \overline{x_k} - \overline{x_j} \parallel$$

Here, σ_k denotes the dispersal of a cluster, calculated as the mean distance between a centrotype (centroid or medoid) and cluster points. For each cluster an almost similar cluster regarding their intracluster error sum of squares is searched, leading to the R_k. Then as we can see from the formula of $DB(K)$, that the metric equals to the average of the R_k values. It is desirable for the clusters to have the minimum possible similarity to each other. Consequently, the minimal observed index indicates the best cluster solution (Halkidi et al. 2001).

For the needs of the comparisons between different models a distance matrix is needed. The Mahalanobis distance is an acceptable choice, since it incorporates the correlation information of the data set apart from the distance of the sublects (Mahalanobis. (1936).

4.5.2 In the measurement of clustering tendency

Although the literature is abound with a wide diversity of clustering algorithms, less attention has been given to related issues such as clustering validity and clustering tendency. Some indices and measures of clustering validity has been discussed in previous paragraphs. Before partitioning a data set, one needs to examine whether the data set exhibits a predisposition to cluster into natural partitions without identifying the group themselves. This is what formally establishes the clustering tendency domain (Jain and Dubes. 1988). A data set with no natural clusters could be thought of, as a random collection of feature vectors. Such data sets should not be a subject of collection. A clustering tendency study might give useful insights about the general nature of the data and therefore reinforce a clustering algorithm choice. Another contribution of this domain is the validation of clustering results in a qualitative and observable manner. Clustering tendency studies do not provide information about optimal number of clusters, but only evidence of absence or existence of natural clusters. Moreover, it is well known that any partitioning algorithm irrespective to the natural groups of objects, produces a-priori specified number of cluster ranging from 2 to n-1, where n is the number of objects. The best partitioning is therefore the most natural partitioning, in the sense that the inherent natural grouping of the data is captured.

In the effort of assessing the data set nature from different perspectives, Barnerjee and Dave (2004) borrowed a statistical concept from the field of clustering tendency and examined its applicability to validate results from a clustering scheme. The problem of testing for clustering tendency can also be described as a problem of spatial randomness. Unlike statistic based cluster validity measures, clustering tendency tests are stated in terms of an internal criterion and no apriori information is brought into the analysis (Jain and Dubes. 1988). The null hypothesis of such tests is most often a random position hypothesis, that is

> H_0: The patterns are generated by a Poisson process with an intensity of L patterns per unit volume.

Under the null hypothesis, the number of patterns falling in a region of volume V has a Poisson distribution with mean LV. *S*ince the number of patterns falling in disjoint regions of V are independent random variables and L is a constant, the Poisson process is a reasonable model to address the absence of structure behavior (Barnerjee and Dave. 2004). These models are called sparse sampling tests and they are quite promising for their high power against clustered alternative hypotheses. In contrast with tests based on small inter-pattern distances like nearest neighbor distance tests that depend heavily on the intensity L of the Poisson process assumed under the null hypothesis. Spatial sampling tests are based on sampling origins randomly identified in a sampling window.

Hopkins statistic

From several tests that have been proposed in the literature, Hopkins statistic (Hopkins and Skellam. 1954) seems to be an easy to use and comprehend one, while it has been compared to Holgate statistic with good results (Panayirci and Dubes. 1983).

Let $\{y_i\}$ be m sampling origins placed at random in the d dimensional sampling window and $\{x_i\}$ be a collection of n patterns, $m \ll n$. Two distances are defined here. Let u_j be the minimum distance from y_j to points in x_i , $j = 1,2, \dots , m$. Also w_j is the minimum distance from a randomly selected pattern in $\{x_i\}$ to each nearest neighbor (m out of the n patterns are marked at random for this purpose). The Hopkins statistics is defined then as

$$H = \frac{\sum_{j=1}^{m} u_j^d}{\sum_{j=1}^{m} u_j^d + \sum_{j=1}^{m} w_j^d},$$

The statistic compares the nearest neighbor distribution of randomly selected locations to that for the randomly selected patterns. When the patterns are clustered, on the average distances from patterns to nearest patterns are smaller than distances from sampling origins to nearest patterns because the sampling origins are selected uniformly. Therefore H values close to 1 suggest aggregation. Similarly values close to 0 suggest repulsion, or regular spacing. Under the null hypothesis, H_0, the distances from the sampling origins to their nearest patters should, on average be the same as the interpattern nearest neighbor, implying a Poisson process behavior and hence H is around 0.5.

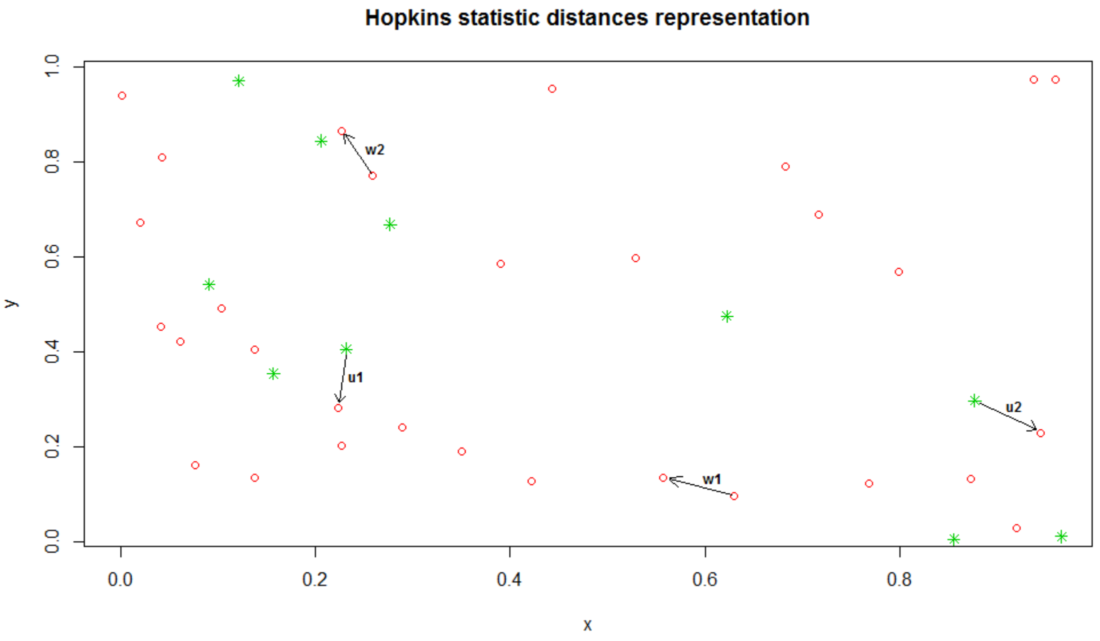


Figure 5. Hopkins statistic w and u distances representation. Here, w_i represent's the distance between a random origin in the d dimensional window (green stars) and the closest neighbor observations (red points), while u_i represent's the distances between a randomly chosen observation (red points) and its nearest neighbor (red points).

The contribution of Barnerjee and Dave (2004), can be addressed in the applicability of the random position test for cluster validity using the Hopkins statistic. A natural cluster is *unusually* compact and *unusually* isolated. A clustered data set is ordered due to the existence of natural groups; in the absence of natural groups it is a random set of data objects approximating a Poisson process. The additional assumption that the authors do is that apart from isolated and compact, a natural cluster is random within itself. Let H_i be the Hopkins statistic for the i^{th} cluster at a paricular level of clustering, the average value of the statistic is then

$$H_{av} = \frac{1}{c}\sum_{i=1}^{c} H_i \text{, for fixed } c$$

A non rejection of the null hypothesis as described above would mean that the value of H_{av} is around 0.5. Proceeding from $c = 2$ until $c = n - 1$, where $n - s$ the number of objects, the lowest of c that generates $H_{av} \approx 0.5$ indicates the generation of a partition that identifies the natural clusters in the data set. Under the assumption that the number of natural clusters is the optimal number of clusters, this derivative of the Hopkins statistic can provide valuable support in the clustering validation.

5 Results

The results chapter is organized as follows. As a first step, the results of the evaluation procedure in different levels of the clustering process are discussed. Secondly, the results of the clustering process are presented together with interpretations of the different groupings. Finally, clinical features of each group that did not take part in the clustering process are available for further comparisons and interpretations of the resulting groupings.

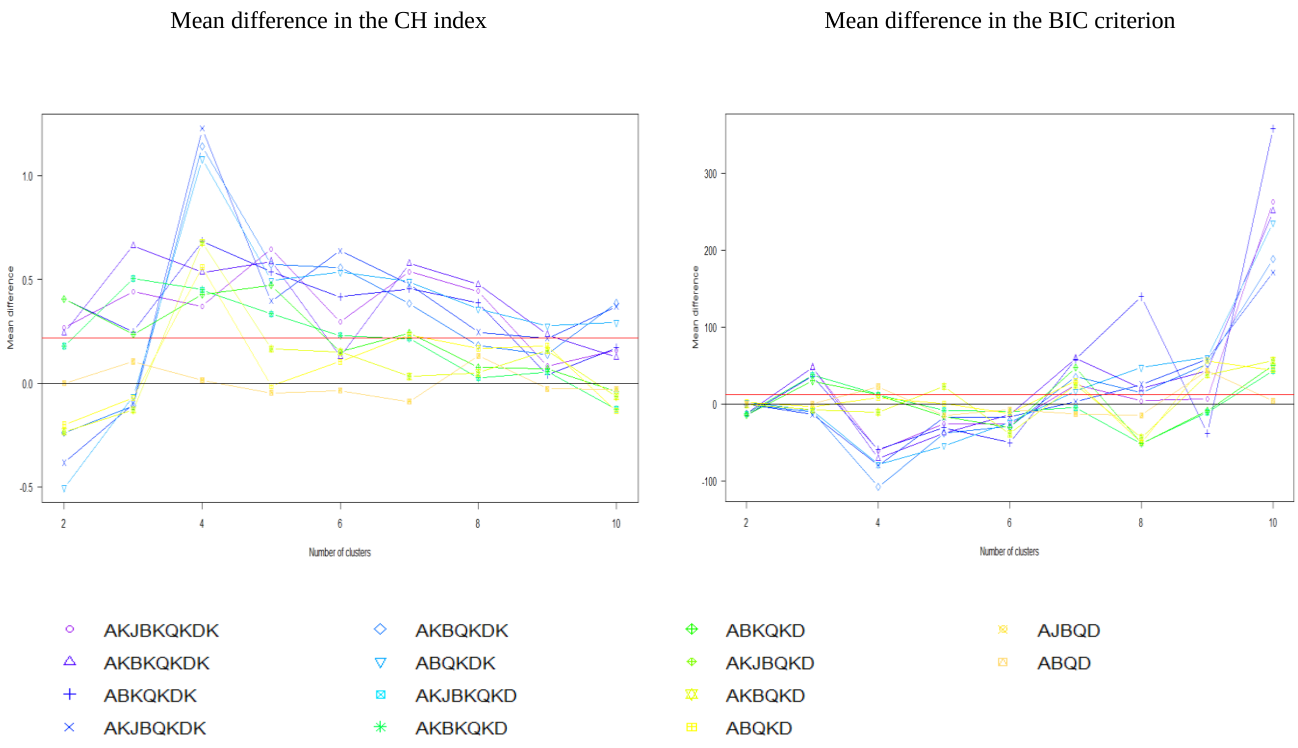
5.1 Evaluation procedure

At this section of the results the evaluation of the model is discussed. The evaluation process is organized in three parts. To begin with, following the methodology we explore the impact of different initialization strategies of the HDDC EM algorithm. The initialization of the EM algorithm with the results of the correlation clustering for the first 100 subjects will be compared to the initialization with random allocation of the subjects into groups. Moreover, the initialization of the EM algorithm with the results of the correlation clustering for the whole dataset will be compared to the initialization with random allocation of the subjects into groups. As a second task we proceed in the evaluation of the choice between the high dimensional data clustering (chapter 4.2) and the Bayesian hierarchical clustering (chapter 4.5). After presenting the results of the winning model the validation of the optimal number of clusters is commented.

*5.1.1 **The initialization of the EM algorithm***

The initialization of the EM algorithm as it is mentioned in the methodology chapter, is an important step for a successful transition to a well formed clustering outcome. The approach followed for the thesis consists of the initialization of the EM algorithm of the high dimensional data clustering in two different ways. Under the first approach the algorithm is fed with a random initial allocation of the subjects in clusters. Under the second approach the algorithm is fed with the following allocation of subjects: the correlation clustering presented in the methodology chapter is applied to the first 100 subjects and these are located to the clusters signified as the optimal ones. As for the remaining subjects they are randomly assigned in clusters, completing an initial placement that allows the EM algorithm to iterate until the expected result is achieved. However, in such experiments in often happens that one initialization in random locations might be extreme and result in very unusual clusters yielding a bad clustering. An admissible way of overcoming this difficulty, is to repeat the experiment of the initialization plenty of times in order to obtain a more complete understanding of the outcome. From now on, in order to make the interpretation easier we will refer to the initialization with correlation clustering as *Init 1* and the random initialization as *Init 2*.

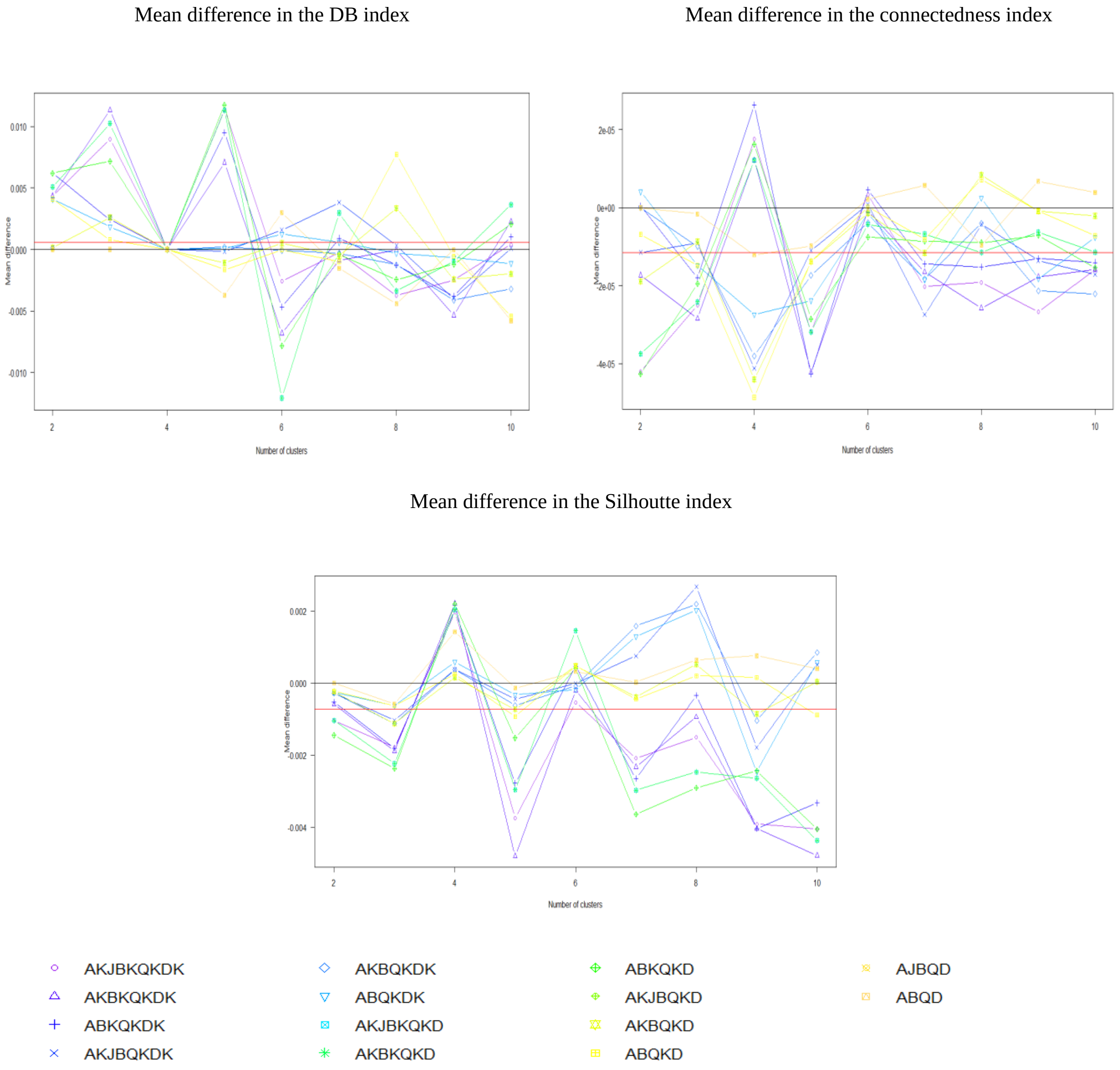
The measures used for the evaluation are the ones described in the respective section of the methodology chapter. The initialization has been repeated 100 times for each model of the GMM family $[a_{ij}\, b_i Q_i d_i]$ and the different indices have been computed. Three of the evaluation criteria (Silhouette score, Bayesian information criterion and Calinski-Harabasz index) demonstrate the optimal cluster when they take the highest value, while the remaining two criteria (Davies-Bouldin and Connectedness) point to the optimal clustering when they take the lowest value. Intuitively, we can think of this problem as an optimization one. The allocation that maximizes the first three indices and minimizes the two last is the optimal under these criteria. An easy way to turn this problem to a general maximization problem is to invert the values of the last two indices. Such a transformation will be useful in the interpretation of the results.


Mean difference in the CH index

Mean difference in the BIC criterion

Figure 6. Mean difference on CH index and BIC criterion for the two initializations over the 14 models of the $[a_{ij}\, b_i Q_i d_i]$ GMM family.

In Figure 6 the difference of the indices is formulated as $\text{diff} = \frac{1}{100}\sum_{i=1}^{100}(I_{Init\ 1} - I_{Init\ 2})$, where $I = \{CH, BIC\}$. The red line in the plots indicates the mean difference for all the methods and clusters from which we can see how the measures of clustering validity perform for the 14 models of the family $[a_{ij}\, b_i Q_i d_i]$. The actual values of these Figures correspond to the mean difference of each measure for each one of the two initializations (random, correlation clustering). For instance, on the left graph in Figure 6: the CH index has been computed for 100 initializations of the EM algorithm for each of the 14 models and for the initial values obtained by the correlation clustering (the first 100 subjects are allocated with the correlation clustering and the rest of the data set is randomly allocated into clusters). This procedure has been repeated again, but now the initial allocations were randomly selected for the whole data set. Finally the difference on the mean indices has been computed. All the mean values for each model and each number of clusters between two and ten has been computed and plotted. In the same trend as the Figure 6 for the CH and index and BIC criterion, Figure 7 presents the results for the remaining three indices.



Mean difference in the DB index

Mean difference in the connectedness index

Mean difference in the Silhoutte index

Figure 7. Mean difference on DB, the Silhouette and the connectivity indexes for the two initializations over the 14 models of the $[a_{ij}\, b_i Q_i d_i]$ GMM family.

This is an efficient way of presenting the most important attributes of these simulations because the resulting tables are too wide to be plotted individually. The mean value is chosen to represent the general behavior of the simulations after inspecting visually that the distribution of the results is approximately normal. The reason for computing the difference between the two initializations is that the sign of the difference provides a measure of evidence against or in favor of each initialization. When the difference is positive, this supports the correlation clustering initialization, while when the difference is negative, then the result is in favor of the random initialization. Also, the mean of all the resulting differences has been computed (red line in each graph), so as to provide a clear decision about the initialization yielding in the optimal clustering with respect to each validation index.

The left graph of the Figure 6 shows the CH mean difference of the two initializations for the 14 models of the HDDC algorithm. In general the values exceed 0 for most of the methods and clustering level. This means that the average distance between clusters for the Init 1 is larger than for the Init 2, while the average distance within clusters is smaller for the Init 1 than for the Init 2. This can be interpreted as follows: the clusters produced by Init 1 have more distance between them, while they are more compact within them than the clusters produces from the Init 2.

The right graph of the Figure 6 shows the mean BIC difference of the two initializations for the 14 models of the HDDC algorithm. The lowest value of this criterion shows the most preferable model. In our case the models have negative BIC values. Still we seek to maximize the difference in the means. The models seem to follow different trends for different levels of clustering (number of clusters). For two and three number of clusters the difference is almost zero while for 10 number of clusters the *best* models yield from the Init 1. As a mean difference (red line) the Init 1 returns better models with respect to the BIC criterion than the Init 2.

The top left graph (Figure 7) shows the mean DB difference of the Init 1 and Init 2 for the 14 different models of the HDDC algorithm. As described in section 4.6.1 the DB is a general measure that computes the similarity between clusters, with an optimal clustering when the clusters are as dissimilar as possible. However it is not just an intracluster measure since the intercluster structure is also incorporated in the formula. The clustering with the lowest similarity is the optimal. in our results the values have been inverted, that is the highest value shows the optimal clustering. Between levels 2 to 5 clusters the mean differences are positive and then negative for all the models. The mean is slightly positive indicating that Init 1 returns the best allocation for all the clustering levels.

The mean difference between the connectedness index for the Init 1 and Init 2 is plotted in the top right side of the Figure 7. This index measures in what extend data in a cluster are close to their nearest neighbor in the data space. The values in most cases are negative indicating Init 2 as the *winner* allocation. Although this result is in expense of the Init 1, we need to take into account in the interpretation that the nearest neighbor indexes put much weight in the data space neighborhoods around the observations into consideration. The weakness of these indexes can be addressed in the *curse of dimensionality* phenomenon (section 4.1), since the distance measure that they use to calculate the possible neighbors is not based in any subspace projection but in the whole dimensionality. Thus, their interpretability becomes increasingly meaningless when the dimensions where the distance measure is calculated increase too much.

Finally, the silhouette index graph (bottom, Figure 7) is a measure computed separately for each observation. It measures the intercluster and intracluster behavior of an observation. In our case the average silhouette for all the observations has been computed. The mean difference is slightly negative, but taking into account the absolute difference (0.0007), the two initialization do not differ significantly.

One more alternative in the initialization has been attempted in order to study deeper the behavior of the correlation clustering algorithm. Under the Init 1 initialization the algorithm found the optimal allocation for a subset of the total dataset with respect to the correlation clustering cost function, while the remaining subjects have been arbitrarily assigned into clusters. This is a good start for a semi supervised analysis, but the fact that the undirected graph of the correlation clustering instance only includes a subset of the total data makes the rest of the subjects *unknown* to the correlation clustering. The assignment of the remaining subjects to the *system* might decrease the clustering quality later on in the EM step.

In an attempt to address this problem, another strategy has been followed and compared again to the Init 2 (random allocation of all the subjects). This time the whole data set has been included to the correlation clustering graph. Unfortunately complexity issues described in the correlation clustering chapter, do not allow the algorithm to optimize the cost function exhaustively. However the algorithm tries to optimize the cost function in steps and this allows it to work until the system runs out of memory. The semi supervised clustering has been repeated one more time, but now the algorithm has been allowed to work until it run out of memory and the last allocation of subjects before the procedure stopped has been saved. In this way the EM algorithm is getting fed with a complete initial allocation (not only the first 100 subjects) for the data set, namely *Init 3* (incomplete correlation clustering). The results are presented in Table 5.

Table 5. Difference in the mean for the two alternative strategies in the initialization with correlation clustering.

	diff_1	diff_2
Silhouette index	-7.165e-05	0.947e-05
Davies-Bouldin index	5.71e+05	24.21e+05
Connectedness index	-1.15e+06	-0.91e+06
Calinski Harabasz index	0.22	0.50
BIC criterion	12.172	-5.400

In Table 5 the $\text{diff}_1 = \frac{1}{100}\sum_{i=1}^{100}(I_{Init\ 1} - I_{Init\ 2})$, refers to the mean values presented in Figures 6 and 7 with a red horizontal line. These mean values are representative of the mean difference between Init 1 and Init 2. On the other hand, $\text{diff}_2 = \frac{1}{100}\sum_{i=1}^{100}(I_{Init\ 3} - I_{Init\ 2})$ refers to the mean difference between Init 3, which is the correlation clustering process until the system runs out of memory and Init 2. It is clear that the results for diff_2 are more promising since they present better values for 4 out of the 5 indices. The only measure that performs worse under this initialization is the BIC criterion, but the difference is irrelevant if we take into account also the standard deviation of the differences apart from the mean. The general outcome of the comparison between diff_1 and diff_2 is that diff_2 performs better in general. The proportional Figures 6 and 7 of diff_1, for the case of diff_2 can be found in the Appendix (Appendix fig. 1, Appendix fig. 2).

5.1.2 In the search of the optimal model

In this step of the evaluation, since the initialization of the EM algorithm for the $[a_{ij} b_i Q_i d_i]$ GMM family is now decided, the two main clustering models will be compared under the AD patients data set.

For notational simplicity we will refer to the GMM clustering as HDDC and in its variations as submodels of HDDC from now on. Also the high dimensional Bayesian clustering with variable selection explained in chapter 4.5 will be annotated as Bayesian clustering from now on. The measures under consideration for this comparative analysis are: the Silhouette index, the DB index, the CH index and the connectedness index introduced by Handl et al. (2005). For the needs of this step of the evaluation the DB and the Connectedness index have not been inverted as in the previous evaluation step, that is the lowest values of these two indices optimize the clustering result. The remaining two indices indicate the optimal clustering instance when they recieve their maximal values.

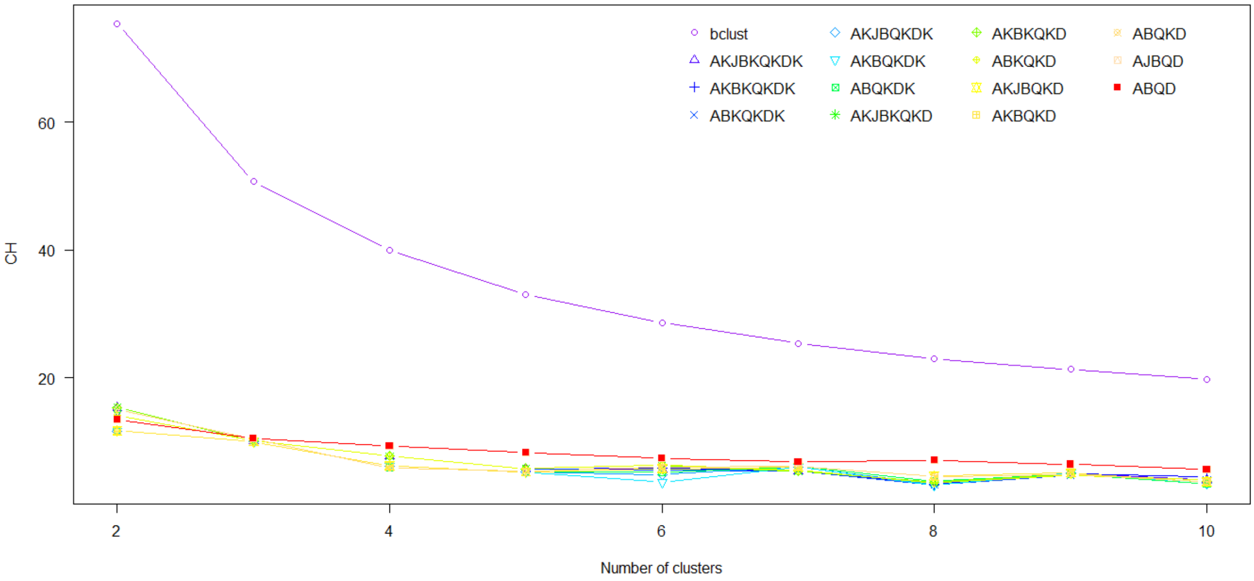


Figure 8. Calinzki-Harabasz (CH) index for the Bayesian clustering and the HDDC sub-models for different number of clusters. Higher scores indicate better clustering.

In Figure 8 the calculation of the CH index for the 14 sub-models of the HDDC and the Bayesian clustering is plotted for different parameterizations of number of clusters. The first 13 sub-models of HDDC seem to follow the same trend while the $[a, b, Q, d]$ sub-model outperforms them (red line). The superiority of this sub-model against the remaining 13 can be interpreted with the aid of its four parameters. It seems that most successful interpretation for the different clusters underlying structure is that all the clusters have common intrinsic dimension numbers, common noise components, common intrinsic components, and common orientations. More intuitively, there is enough evidence that due to the CH index the clusters under the HDDC have the shame shape. However the Bayesian clustering result (purple line), outperforms all the 14 HDDC sub-models with much greater CH scores for all the different clustering parameterizations.

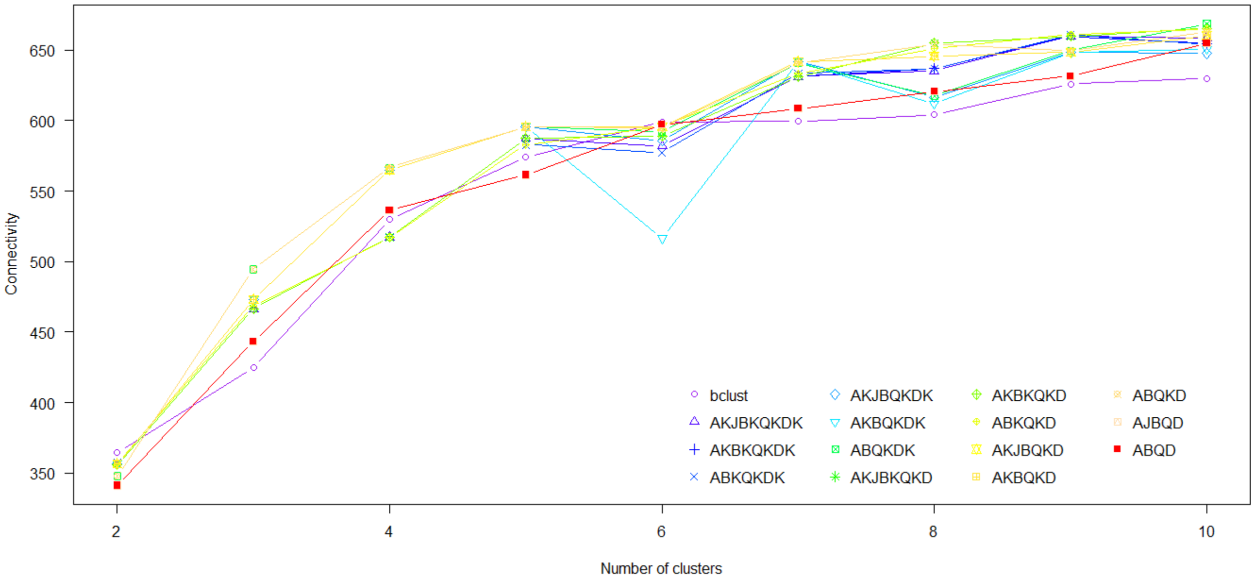


Figure 9. Connectedness index for the Bayesian clustering and the HDDC sub-models for different number of clusters. Lower scores indicate better clustering.

The connectedness index is plotted in Figure 9 for the 14 HDDC sub-models and the Bayesian clustering. In general, it returns more diverging values for different parameterizations of the clustering number than the CH index (Figure 8). In clustering level 2, the $[a, b, Q, d]$ model has the lowest connectedness and all the remaining models are slightly higher. In clustering level 3 the Bayesian clustering returns the lowest connectedness and the other models vary more than in the level 2. The Bayesian clustering has the lowest values for 5 out of 9 parameterizations of the cluster number and therefore it outperforms the HDDC family in most of the clustering number cases.

The values for the DB (Figure 10) index in different cluster numbers parameterization seem to be clearly discriminative between HDDC and Bayesian clustering. The former model family returns scores between 24 and 33 with no clear patterns between the 14 sub-models. On the other hand, Bayesian clustering is totally separated with values deviating between 15 and 21. Lower scores attribute better performance with respect to the compactness and isolation properties of the clusters and therefore Bayesian clustering outperforms the HDDC.

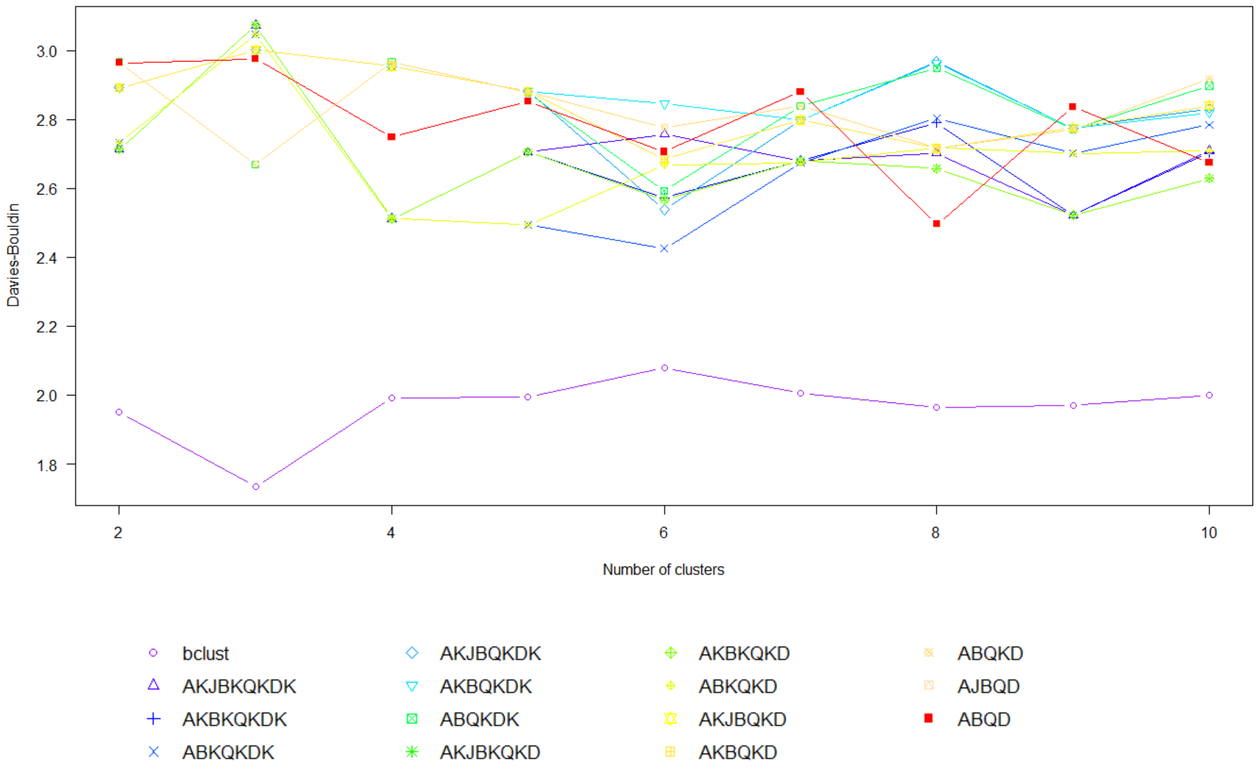


Figure 10. Davies Bouldin (DB) index for the Bayesian clustering and the HDDC sub-models for different number of clusters. Lower scores indicate better clustering

The only competitive model from the HDDC family to the Bayesian clustering, is proved to be the simplest and least complex one, namely $[a, b, Q, d]$. In contrast with the CH, DB and connectedness indices, Silhouette index seems to indicate the model $[a, b, Q, d]$ as the optimal one (Figure 11). Although, this model appear to be better than all the competitive ones with respect to the average Silhouette index, the absolute difference of this model with all the rest is smaller than 0.015, thus it is insignificant[10]. In addition, some weaknesses of this particular index to cope with high dimensional data sets due to the nearest neighbor calculations, probably make it much different with respect to the resulting values than the other three indices. However the descending trend of its values while the number of cluster increases, can be translated as a similarity of Silhouette with the other three indices.

[10] Silhouette index receives values between -1 and 1. Taking this into account a difference of 0.015 is accounted as insignificant.

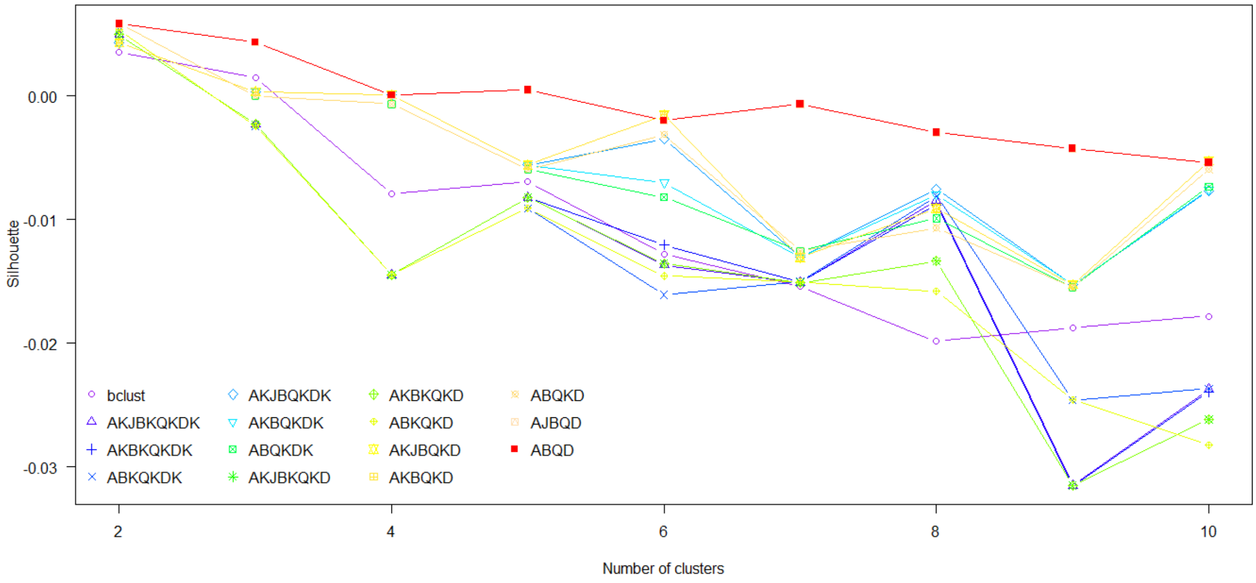


Figure 11. Silhouette index for the Bayesian clustering and the HDDC sub-models for different number of clusters. Higher scores indicate better clustering.

The comparison of the 14 models of the HDDC family between themselves and the Bayesian clustering with respect to four criteria, provides enough evidence for the superiority of the Bayesian clustering against the other models for this dataset. However, the model $[a, b, Q, d]$ is the only competitor of the Bayesian model since it performs better than the remaining 13 sub-models of the HDDC family, almost for all the *index-number of clusters* combinations.

5.2 Clustering results

This section facilitates the results of the AD data set agglomerative Bayesian hierarchical clustering analysis together with the exploration of the optimal number of clusters for this method.

From results of the comparative analysis between the HDDC sub-models and the Bayesian clustering model, the latter has the best performance and therefore it is chosen as the optimal clustering result between the 15 considered. For the further analysis of the allocation of the AD patients data set into clusters, the method under consideration is the Bayesian clustering since it holds the most optimized properties with respect to the compactness and isolation of its clusters.

5.2.1 *Agglomerative hierarchical clustering*

The ability to represent the construction of a clustering result in its step can be accounted as a major advantage, because it reinforces the interpretation of its merge or split of the subjects in groups. Agglomerative hierarchical clustering has this valuable property since the tree structure of the results is a visual guide of the clustering of the subjects into groups.

The tree structure of the results of the Bayesian clustering (Figure 12), uncovers two main clusters. This can be seen by the high distance between the merge of the two clusters in one (more than 10000 thousand). In level 3 we can observe that one of the clusters of level 2 keep its formation but the second splits in two other clusters (Figure 12).

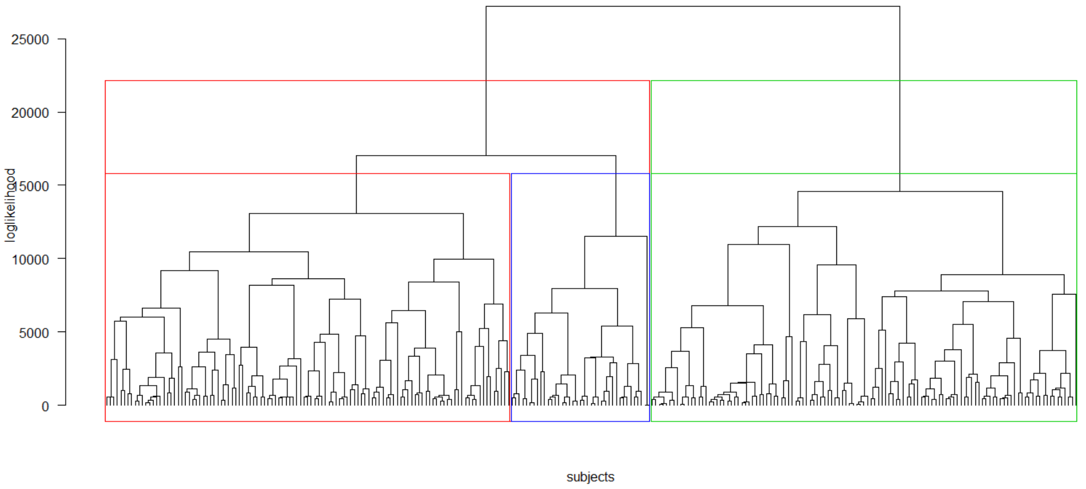


Figure 12. Dendrogram of the Bayesian hierarchical clustering for the AD dataset with rectangles around the cluster level 2 and the cluster level 3.

The resulting dendrogram discloses some important information about the groups of patients clustered together. Firstly, the distances of the subjects within each groups are not large (very tall branches), that is the clusters are compact in almost all the levels. However, some branches are tall and this reveals that subjects clustered together within such clusters are distant from each others. For instance inside the blue rectangle (cluster 2) in clustering level 3, we can observe two subjects in the *far right* borders of the cluster that have a very high merge with the remaining subjects of the cluster. These two subjects are quite close to each other and at the same time far away from the rest of the subjects of this cluster with respect to the distance measure used for the clustering (log posterior). From the resulting dendrogram we can summarize that with respect to the distance measure used there is both compactness and isolation until the clustering level 4.

5.2.2 *In the definition of the optimal clustering level*

From the Figures 8, 9, 10, and 11 apart from the differences in the model performances, one can collect useful information about the scores that each model gets for different parameterizations in the number of clusters. For instance, with respect to the connectedness index all the models with 2 clusters have the least connectivity between them. Also with respect to the CH index, the Bayesian clustering with 2 clusters is the one that has the highest score. Thus, one can say that 2 clusters is the optimal separation of data into clusters. For the DB criterion, the Bayesian clustering again receives the minimal value but now with 3 clusters. Finally, through the Figure 11, we can observe that all the methods with 2 and 3 clusters have their highest Silhouette scores, showing that probably the optimal separation is in that level. Moreover, from Figure 12 which represents the agglomerative tree of the Bayesian clustering the isolation and compactness of the clustering level 2 and 3 agrees with the evaluation results of the section 2.5.1. That is, the most isolated and compact clusters are in clustering level 2 and 3.

The reason that different clustering evaluation criteria return different optimal parameterizations yields from the fact that they examine the clustering result quality from different perspectives. For instance the connectedness index and the CH index examine similar properties of the clustering and thus their results are more or less similar. The DB index is a general measure of clustering quality, since for each cluster it searches for an almost similar cluster regarding their intracluster error sum of squares and then calculates the dispersal of each cluster. It can be understood that this is a dispersal measure and thus it might return different results from the previous two indices. The Silhouette index on the other hand searches for nearest neighbors to calculate the clustering quality.

Although, the four clustering evaluation criteria under consideration did their duty for the clustering method comparison task well, one property of clustering quality is not examined by them. As discussed in the evaluation section, natural clusters are compact, isolated but also uniform within themselves. When we look for the optimal number of clusters considering groups that comply with the first two properties and not the third one we might result in groupings that facilitate outliers or heterogeneous formations. For example if we observe more carefully the different branches of the agglomerative tree (Figure 12), we can easily identify that if we cut it in four clusters instead of three, the isolation of the clusters is not as great before. However, the two clusters that emerge from the last cluster to the right (green rectangle) are two totally different branches with much distance between them. Therefore the green hatched cluster can be ascribed as a cluster which is much different from the red and blue hatched clusters and facilitates within itself two different clusters. Although the four evaluation indices (section 5.1.2) managed to evaluate the clustering quality of the 15 different models, they fail in the recognition of the 4th cluster. Motivated by these observations in the clustered subjects it has been decided that in the search of the optimal number of clusters, these measures will not be taken into account.

Encouraged by the assumption that a cluster is random within itself, the Hopkins statistic will be used to explore after which number of clusters there is absence of clustered behavior. The Hopkins variation H_{av} is applied in different levels of clustering until the desirable threshold is reached indicating the number of natural clusters in the data set under the Bayesian clustering outcome.

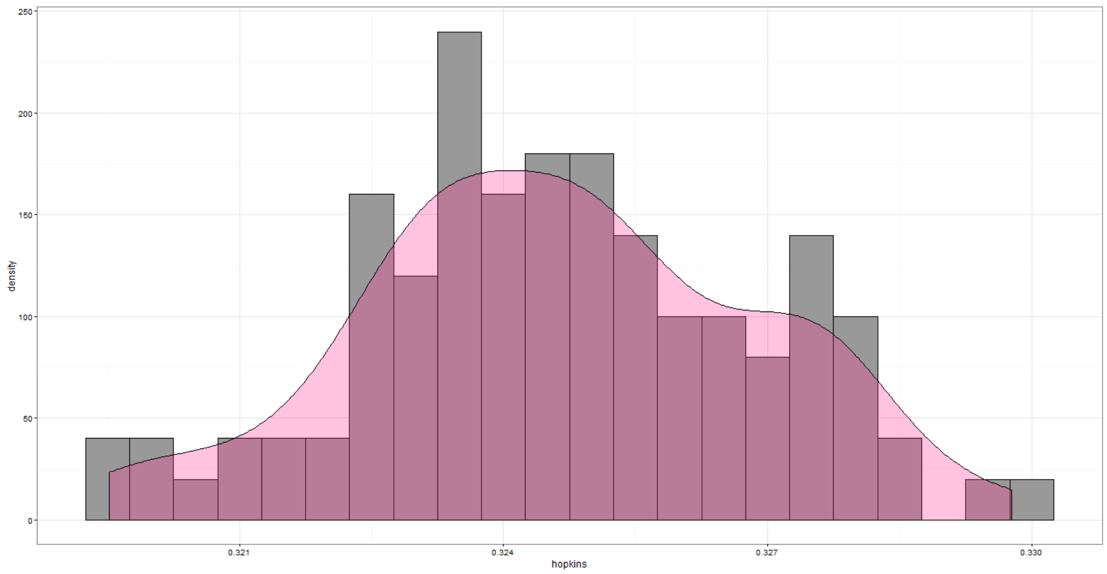


Figure 13. Hopkins statistic value for the clustering tendency of the AD dataset.[11]

The Hopkins statistic for the examination of the clustering tendency of the AD dataset has been sampled 100 times to increase our certainty about its value (Figure 13). The reason is that the results of sparse sampling tests might deviate and the distribution of a sufficient sampling window provides a comprehensive insight in the behavior of the statistic. Most of the mass lies in the space between 0.32 and 0.33; it exhibits quite stable behavior, allowing to conclude that the AD dataset is clusterable with regularly spaced data, data that are neither clustered nor random.

The Hopkins statistic has been simulated for different number of clusters 100 times and the results are presented as a mean together with 95% confidence intervals (Figure 14).

[11] The kernel of the density is Gaussian and the bandwidth is the default one (Silverman's ‘rule of thumb).

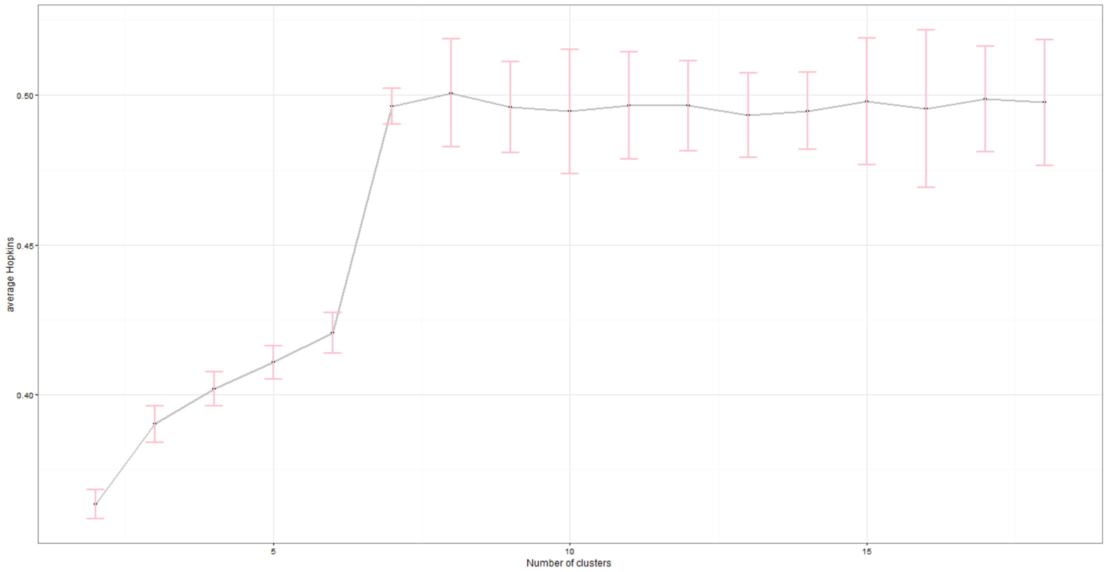


Figure 14. Average Hopkins statistic with 95% confidence intervals for different number of clusters.

Starting from the left of the graph we can observe that for 2 clusters the average value of the Hopkins statistic is around 0.36. This means that each one of the clusters still has patterns within itself, which means that it does not form one natural cluster but more. The nonexistent natural clusters demonstrates the necessity to look deeper in the hierarchy, until the random formations come up in the clustering result. The H_{av} constantly increases for the clustering levels three to six and the confidence intervals do not overlap significantly; that is the values are quite stable and provide a good estimate of the clustering tendency in every level with respect to the previous and next parameterization (number of clusters). Something unusual is detected between the six and 7 number of clusters. At first place it has to be remarked that as it is also explained in the section 4.5.2, describing the Hopkins statistic when it receives scores close to 0.5 then we are in the level of natural clusters. As it can be noticed, in the level of 7 clusters the values of the statistic did reach the threshold and in some case they even got over it (intervals). After this level and for all the remaining levels until the 18 clusters we can see that the values are always around 0.5 with some fluctuations occasionally. The question to be answered at this step of the analysis is, what makes the Hopkins statistic increases very rapidly between the level 6 and 7 of the clustering.

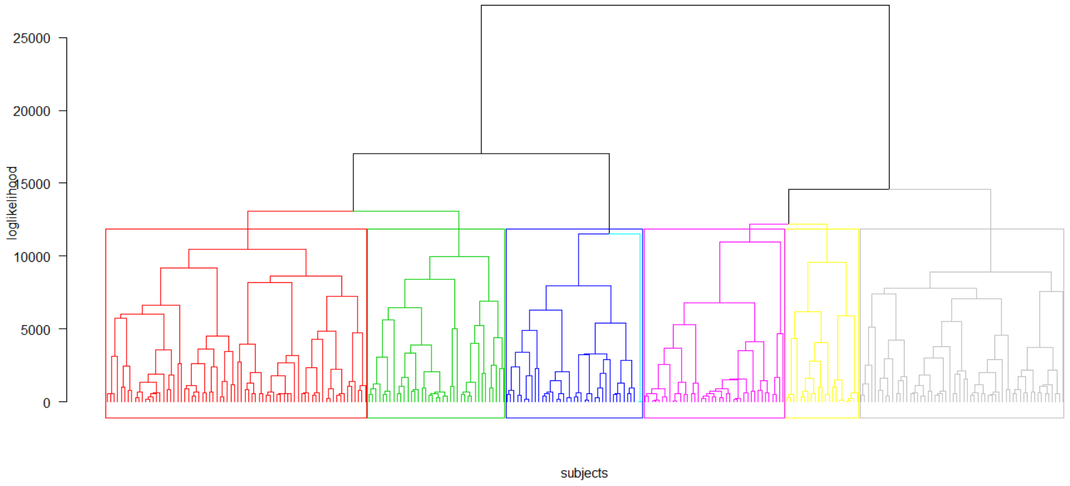


Figure 15. Dendrogram of the Bayesian hierarchical clustering for the AD dataset with rectangles around the level of 6 clusters and colored branches for the level of 7 clusters.

The number of naturals groupings has been identified by the Hopkins statistic in the 7 clusters level. However if we inspect the Figure 15 carefully, we can see that in the blue rectangle (cluster 3 in 6 cluster level), there are two main branches of subjects. The left is in blue color while the right is in sky blue color. The latter subjects are only two and from their distance to the remaining subjects of cluster 3 in 6 cluster level we can see assume that they are extreme outliers. If we cut at the tree in the 7 cluster level (colored branches) the only cluster that splits is the blue hatched cluster in a blue and a sky blue colored branch. To sum up, this means that in the level of 6 clusters, the two outliers described above are clustered together with the rest of the subjects of cluster three; the Hopkins statistic recognizes this anomaly and keeps the average Hopkins low in order to inform that the number of natural clusters has not yet reached and more splits are inevitable. At the level of seven clusters these two outliers are clusters together and all the remaining clusters stay untouched. Therefore, it can be concluded that in the level of seven clusters we have six natural groups of subjects and two outliers clustered in a separate group.

5.2.3 Results for different clusters

A sufficient clustering in high dimensional data is meant to find groups of subjects that are highly discrete in each dimension. A summary of the most important descriptive statistics for the resulting groups, is an efficient way to present some of the main differences between them. Demographical, clinical and pathological differences exist between the 6 clusters and are summarized in table 6.

Table 6. Demographical, clinical and pathological characteristics of the 6 clusters.

	$\mathrm{AD}_{\mathrm{cl}1}$	$\mathrm{AD}_{\mathrm{cl}2}$	$\mathrm{AD}_{\mathrm{cl}3}$	$\mathrm{AD}_{\mathrm{cl}4}$	$\mathrm{AD}_{\mathrm{cl}5}$	$\mathrm{AD}_{\mathrm{cl}6}$
N (%)	39 (14.4)	74 (27.3)	37 (13.65)	40 (14.7)	58 (21.4)	21 (7.7)
Women^{*a}	22 (56)	43 (58)	21 (57)	30 (75)	28 (48)	9 (42)
Disease duration	4 (2-6)	3 (2-5)	2.95 (1.9-4.4)	4 (2-5.7)	2.5 (1.8-4)	3 (2-5)
Age	76.0 (68.9-79.4)	75.9 (71.4-80.2)	72.0 (70.6-76.0)	77.0 (72.4-80.2)	76.5 (71.0-80.5)	77.5 (72.6-81.9)
Onset	71.7 (65.7-75.6)	72.5 (67.7-77.3)	70.0 (66.3-73.7)	73.1 (68.3-75.7)	72.8 (68.0-78.0)	73.0 (68.0-76.6)
Education	12 (8-16)	12.5 (10-15)	12 (8-16)	9.5 (6-16)	12.5 (8-16)	12 (6-15)
ApoE carriers^{*b}						
ApoE ϵ3 carriers	28 (72%)	60 (81%)	28 (77%)	31 (77.5%)	44 (75.8%)	19 (90.4%)
ApoE e4 carriers	22 (56.4%)	44 (59.4%)	25 (67.6%)	27 (67.5%)	33 (56.9%)	12 (57.1%)
2 ApoE e4 carriers	10 (25.6%)	12 (16.2%)	6 (16.2%)	7 (17.5%)	10 (17.2%)	2 (9%)
MMSE	23 (21-25)	23 (21-25)	24 (23-26)	21 (17-23)	23 (21-25)	20 (17-23)
Q1	6.0 (5.3-7.3)	6.0 (5.3-7.0)	5.3 (4.3-6.3)	7.0 (6.3-7.7)	7.0 (5.0-8.0)	7.0 (6.0-7.3)
CSF^{*c} values						
TTAU	118 (100-188)	131 (98-174)	114 (85-127)	80 (74-189)	115 (89-146)	109 (71-134)
PTAU_{181}	41 (30.5-69.5)	41 (33-58.5)	32 (28-38)	34 (31.5-52.5)	37.5 (30.5-47)	36 (35-45.5)
TAU	133 (101-206)	136 (99-160)	106 (64-121)	87 (84-179)	122 (85-160)	93 (64-143)
ABETA_{142}	127 (117-137)	142 (125-158)	138 (122-163)	145 (127-150)	143 (114-158)	124 (120-128)
CDR^{*d} domain						
CDR general	0.86 (0.41)	0.86(0.36)	0.73 (0.25)	1.11 (0.49)	0.91 (0.38)	1.19 (0.56)
Memory	0.95 (0.43)	1.07 (0.62)	0.77 (0.42)	1.38 (0.66)	1.13 (0.57)	1.21 (0.77)
Orientation	0.82 (0.56)	0.71 (0.45)	0.5 (0.39)	0.91 (0.61)	0.9 (0.64)	1.1 (0.66)
Judgment	0.70 (0.48)	0.8 (0.47)	0.54 (0.36)	1.04 (0.49)	0.84 (0.49)	1 (0.65)
Community Affairs	0.81 (0.57)	0.7 (0.46)	0.5 (0.41)	0.98 (0.61)	0.72 (0.5)	1 (0.65)
Home and hobbies	0.97 (0.73)	0.76 (0.49)	0.5 (0.5)	1.1 (0.75)	0.81 (0.62)	1.07 (0.75)
Personal care	0.36 (0.63)	0.27 (0.48)	0.08 (0.28)	0.65 (0.83)	0.31 (0.54)	0.71 (0.85)
$ADAS^{*d}$ scale						
Naming objects and fingers	0.85 (1.14)	0.54 (0.71)	0.43 (0.7)	0.78 (0.97)	0.6 (0.59)	0.67 (0.8)
Following commands	0.82 (1.05)	0.72 (0.77)	0.51 (0.56)	1.52 (1.18)	0.84 (0.97)	1 (0.9)
Constructional praxis	0.92 (0.74)	0.89 (0.87)	0.65 (0.59)	1.3 (1.18)	1 (0.77)	1.38 (1.07)
Ideational praxis	0.46 (1)	0.55 (0.9)	0.24 (0.49)	1.33 (1.22)	0.55 (86)	0.9 (0.83)
Orientation, total incorrect	2.41 (1.93)	2.68 (2)	1.84 (1.71)	3.38 (1.98)	2.48 (2)	3.76 (1.76)
Word recognition, mean incorrect	6.72 (3.32)	6.22 (3.18)	4.9 (2.63)	7.2 (3.31)	6.64 (3.42)	6.8 (2.66)
Recall the test instructions	0.51 (1.1)	0.53 (1.26)	0.14 (0.42)	1.12 (1.2)	0.62 (0.93)	1.1 (1.26)
Spoke language ability	0.56 (0.88)	0.42 (0.72)	0.22 (0.53)	0.55 (0.88)	0.52 (0.68)	0.67 (0.73)
Word finding difficulty in spontaneous speech	0.87 (1.1)	0.7 (0.87)	0.68 (0.75)	0.88 (1)	1.16 (0.9)	1.1 (1)
Comprehension in spoken language	0.54 (0.85)	0.38 (0.7)	0.32 (0.53)	0.65 (0.95)	0.6 (0.82)	0.86 (0.96)

Note: Data is presented on the form $\tilde{X}$ (Q_1-Q_3), where Q_1 (1_{st} quartile), Q_3 (3_{rd} quartile) and $\tilde{x}$ (median), unless otherwise stated.

$*_a$ For women the data are presented in the form n(%), where % refers to the percentage of women in the cluster.

$*_b$ For the ApoE carriers the data are presented in the form n(%), where % refers to the percentage of carriers in the sample.

$*_c$ For the CSF values, data available for a subset of the ADNI patients: $\{n_1 = 11, n_2 = 27, n_3 = 17, n_4 = 7, n_5 = 16, n_6 = 7\}$.

$*_d$ For the CDR (Clinical Dementia Rating) domain and the ADAS (Alzheimer's Disease Assessment Scale) cognitive the values are presented in the form, mean(sd).

The data used for the calculations of Table 6 differ from the data used for the clustering. More specifically, the data of Table 6 is a summary of an external data set that has been used to examine possible differences between the clustered patients more than the ROI atrophy patterns.

In order to avoid any confusion between the Table 5, Figure 15 the matches of cluster numbers from table 5 to the branch color of the clusters in Figure 15 follows: $AD_{cl_1} \rightarrow green\ color,$ $AD_{cl_2} \rightarrow red\ color,$ $AD_{cl_3} \rightarrow blue\ color,$ $AD_{cl_4} \rightarrow pink\ color,$ $AD_{cl_5} \rightarrow grey\ color$ and $AD_{cl_6} \rightarrow yellow\ color.$ The blue sky color (Figure 15) inside the cluster 3 refers to the outlier cluster with 2 observation that has been discussed in previous section and is not included in analysis of the results.

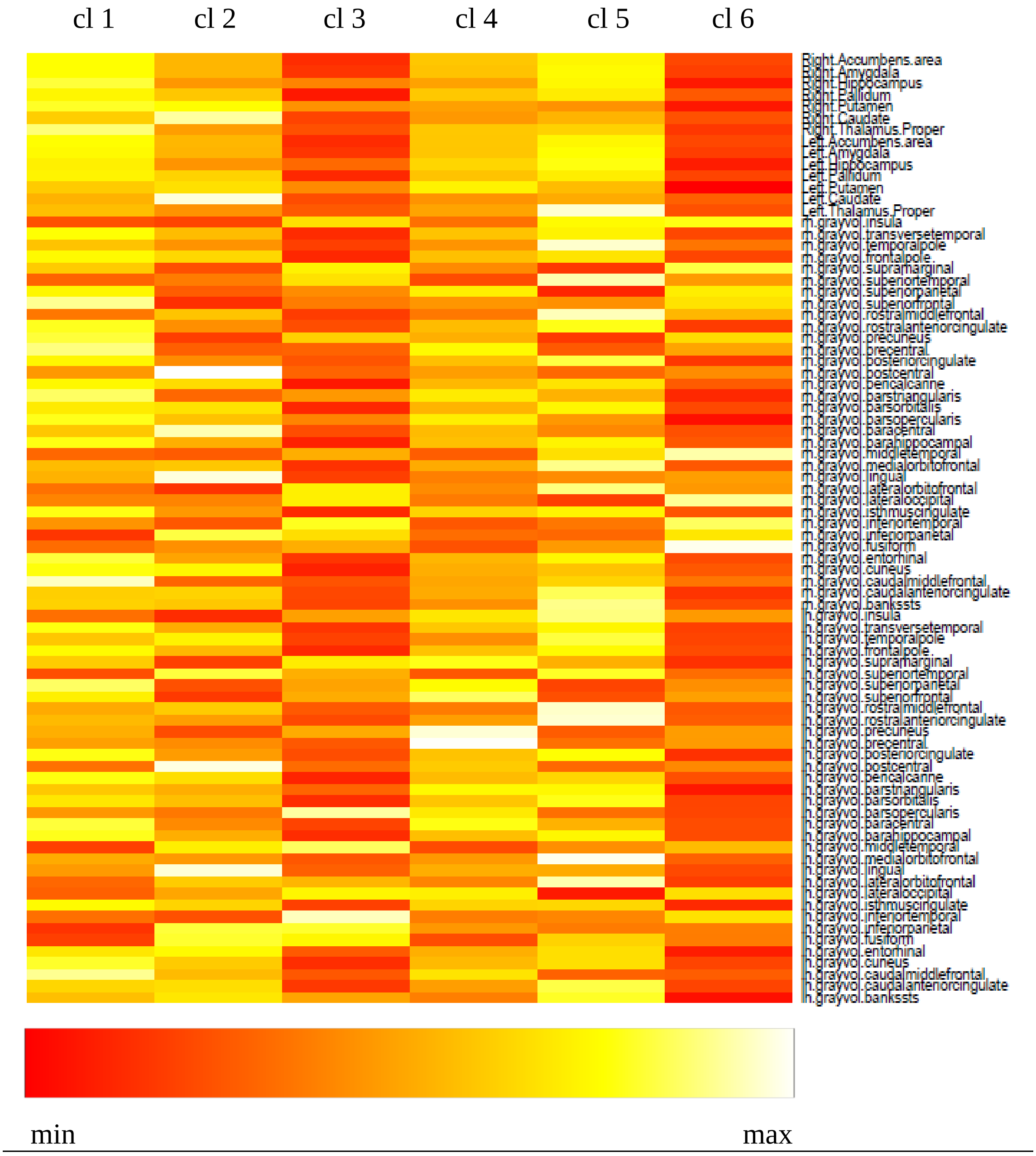


Figure 16 Heat map of the variable-cluster Bayes factors for the 6 clusters and the 82 variables.

As explained in section 4.5 the Bayesian clustering model that is used for the needs of the thesis has two different effects with respect to its variable, one for the important variables in the formation of the clustering result (variable level) and one for the important variables in a cluster level (variable-cluster level). For the variable-cluster level the Bayes factors are plotted in the heat man of Figure 16. From the different colors we can see that there are clusters where many variables are important while there are other clusters where few variables are important. For example cluster three has very few variables of great importance while cluster 5 have many more. Also cluster 6 has few important variables, but these have very high Bayes factor (right middle frontal, right fusiform and right lateral occipital) and also cluster 1 that has many important variables that but few with very big scores.

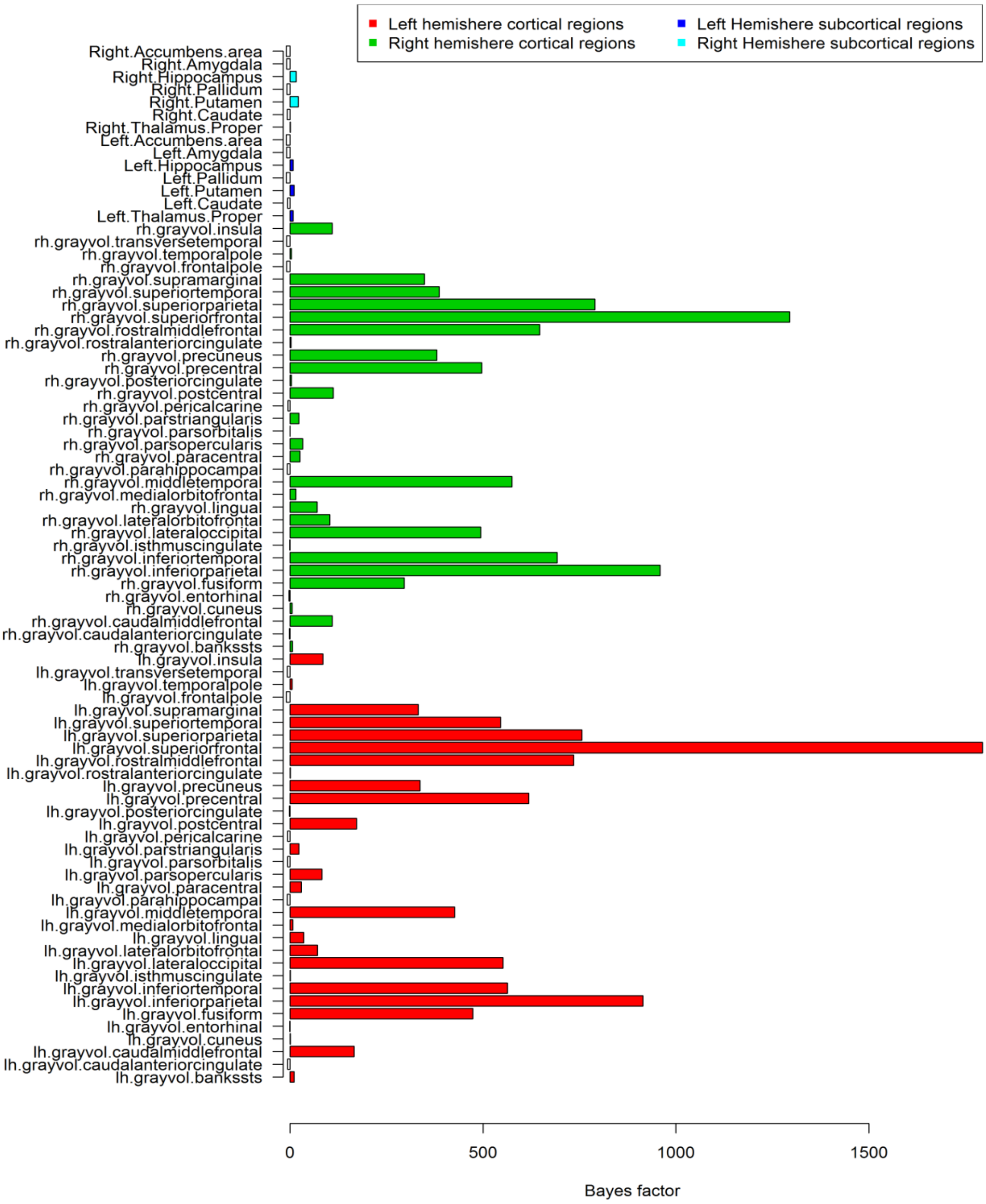


Figure 17. Variable level importance using the Bayes factor for the variables of the AD data set (the vertical axis includes the full names of the variables).

For the variable level effect it is possible to extract a measure of importance of each variable in the clustering. This is done through the computation of the Bayes factors as explained in section 4.4. The variables that hold the most importance can be interpreted as the ones that contribute the most in the discrimination of the patients in the 6 groups. Intuitively, these features can be understood as the dimensions where the clusters have the most distance between them.

The variables of importance hold a degree of symmetry with respect to the left and right hemispheres (Figure 17). Only variables that have positive Bayes factors indicate that the evidence in favor of the variable is higher than the evidence against it and also only these variables are colored in the plot. One can observe that variations exist in the importance of the same variables in different hemispheres. For instance the variable superior frontal referring to the left hemisphere has the highest importance, while its equivalent in the right hemisphere (variable rh.grayvol.superiorfrontal) is the second most important but with lesser Bayes factor than the one in the left hemisphere. The subcortical regions, both from the left and right hemisphere do not seem to contribute a lot in the discrimination of the clusters since they have small values (Figure 17).

Any attempt to identify subtypes of AD using particular patterns of regional brain atrophy should employ gender and age normative values of regional brain volumes obtained from a generous cognitively normal (CN) population (Byun et all, 2015). Motivated by this, the resulting clusters have been visualized, with the help of a CN population (Figure 18). The resulting figures represent the comparison between the atrophies in the cortical regions of a cognitive healthy population and each cluster in a vertex level. Although the clustering algorithm has been to the VOI's, the comparison of the different groups is in a vertex level for two reasons. Firstly, the MRI images of the patients are available for visualizing them. In addition in a vertex level the accuracy of the visualizations increases significantly and this allows a better interpretation of the results.

To start with one can consider the 3 clusters level of the Figure 18. Diverging patterns can be observes between the three clusters. In the top left of the figure we can see a cluster of patients that in comparison with the CN group present atrophy only in the parahippocamplal gyrus and entorhinal cortex. Cluster 2 present atrophy in many more areas excluding the primary motor cortex and mainly the occipital lobe. Moving on to cluster 3, there the atrophy is in an advance level and affects almost all the areas besides a small area in the paracentral sulcus and the pericalcarine cortex. In the level of 6 clusters two of the clusters in cluster level 3 split in more clusters while the cluster with the least atrophy remains unseparated in sub-clusters. In level 6, the cluster 1 and 2 clearly have different patterns of atrophy with cluster 2 having comparatively more atrophy both in the frontal lobe and the parietal lobe. Finally, cluster 4, 5 and 6 are the ones that have atrophy almost everywhere. However cluster 4 is the one that has the most expanded atrophy among all the clusters. Cluster 5 and 6 that have the closest patterns of atrophy but still differences can be found in the caudal middle frontal gyrus of the right hemisphere and supramarginal gyrus of the left hemisphere.

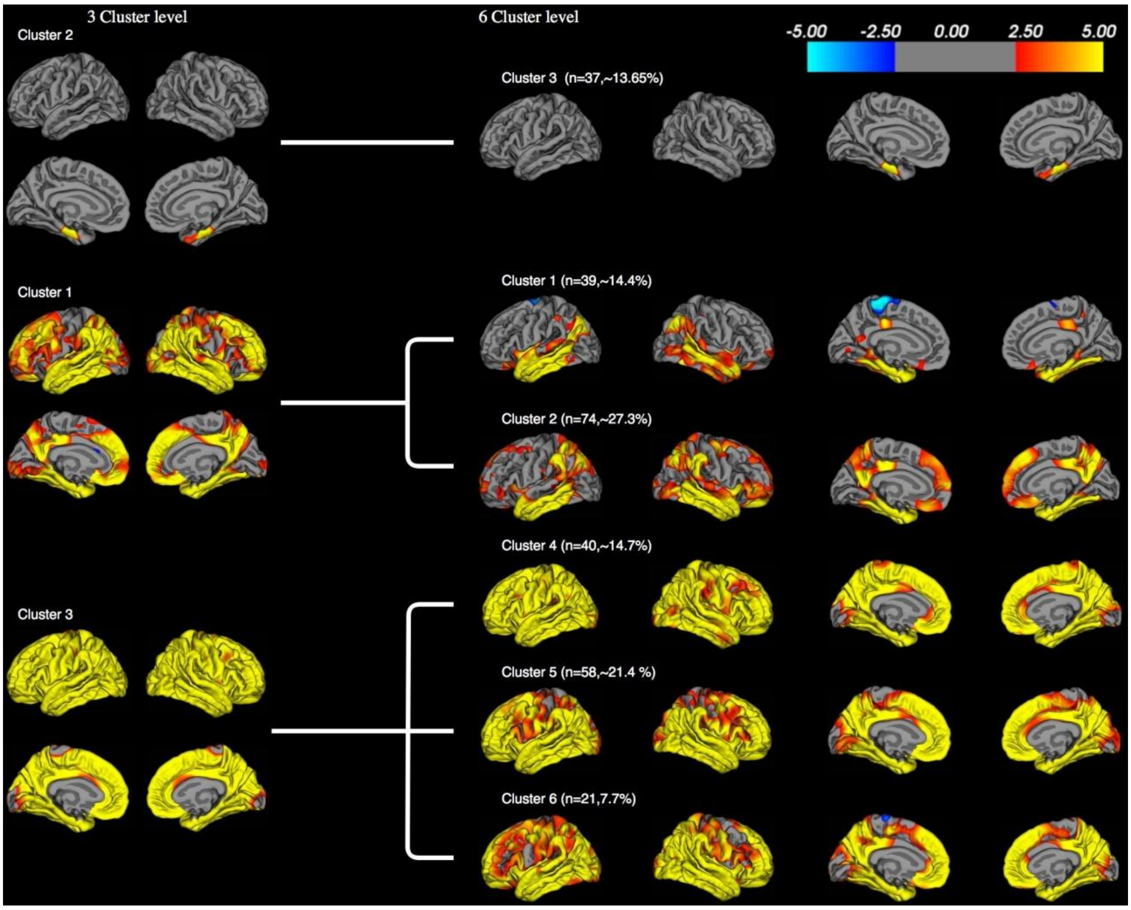


Figure 18 Comparison between the CN population and each cluster's atrophy for the cortical regions in clustering level 3 and 6.[12] Grey color represents no difference in the atrophy patents with the CN population, blue color points higher volume in a cluster in comparison to the CN population, red color represents lower volume of a cluster against the CN population and yellow color indicates much difference between the CN population and a cluster.

The differences between the volumetric measures for the 6 clusters are presented in the Appendix (Appendix fig. 3).

[12] The colour represents the $-\log(p)$, where p is the p_{value} of a t-test in a vertex level. This is a convenient way for a vertex level comparison in the means of two populations (AD cluster vs CN cluster).

6 Discussion

6.1 Clustering evaluation

The evaluation of clustering performance in contrast to supervised methods is always a rather complicated process with no standard routines. Although many different indexes have been discussed in the literature, the quality of a clustering result remains partially subjective when the golden standard is not available. One can address the relativity of the optimal clustering using different evaluation criteria for the same clustering result. In line with many applied clustering studies, the results of the comparisons between initializations for the EM algorithm using more than one evaluation criteria differ. For some criteria the $init$ 1 has enough evidence to be the optimal while for some other the differences between $init$ 1 and $init$ 2 are almost invisible (see Figures 6 and 7). Considering the different characteristics of a clustering result that each of the five criteria takes into account, an overview of the advantages and disadvantages of the two initialization can be concluded. When looking for optimal groupings from the perspective of intercluster and intracluster variance the CH criterion shows that $init$ 1 is optimal. When the perspective changes and the best clustering is defined as the one where neighbor data are clustered together, then the connectedness index results are in favor of $init$ 2. When the clustering results are getting penalized through the log likelihood and the number of parameters of their models, then $init$ 1 is once again picked out as the best alternative in the initialization question. As for the DB index, the results of the comparison between $init$ 1 and $init$ 2 do not differ in such a level that we could safely pick an alternative. The silhouette index assesses the confidence of the choice to cluster an object in a specific cluster and is computed individually for each object. The average for all the objects is calculated. The results suggest that $init$ 2 is better than $init$ 1. However we have to be mindful when applying this measure in high dimensional data sets, because in its intercluster part it calculates distances from the nearest neighbor cluster and might fall in the same trap of the *curse of dimensionality*.

The performance of the correlation clustering initialization is more clear when $init$ 3 is considered. The analysis of the results for the initialization of the EM clustering algorithm under the $[\alpha_{ij} b_i Q_i d_i]$ GMM family (HDDC) provides enough evidence to summarize in two main points:

i. The initialization using the correlation clustering has a better impact in the final allocation of the subjects than the random initialization, with respect to the five indices under consideration.

ii. In the study of the initialization with correlation clustering: when the whole data set is clustered until the computer runs out of memory, the resulting allocation is more successful than in the situation where only 100 subjects are clustered optimally and the remaining are randomly allocated.

The comparison between the HDDC and the Bayesian clustering with respect to different distance based evaluation criteria finalized in favor of the Bayesian clustering. In that comparison the BIC criterion is not used as in the previous step (initialization) due to the fact that even if both HDDC and Bayesian clustering are model based methods and the log posterior is computable, the number of parameters cannot be defined straightforward in the case of the latter model.

Both HDDC and the Bayesian clustering are constructed to accommodate high dimensional data sets. However the perspective from which they look into the high dimensional question is different. Some differences are presented below:

i. HDDC looks into subspaces of the whole dimensionality and assumes that for each cluster a subspace of the dimensions is intrinsic while the rest of the dimensions correspond to noise. The decision about the intrinsic and noisy part under this model is closely related to the Δ_i matrix (section 4.2.1), that contains the eigenvalues of the covariance matrix of cluster i, Σ_i. By looking at the eigenvalues one can think that the model measures the variability of linear combinations of the variables. Therefore, two of the main features that HDDC used to address the high dimensional problem are the dimensionality reduction and the linear combination of variables, through the eigenvalues. Moreover the HDDC assumes that each cluster can be expressed by a normal distribution which is the core of GMM models.

ii. Nia's model for Bayesian clustering on the other hand is a two level effect model. It uses an effect for each variable and an effect for each variable cluster combination. These effects correspond to the Bayesian variable selection scheme since they are modeled as spike and slab distributions. The main assumption of this model is that variables are independent from each other which is a rather strong assumption.

Although HDDC corresponds theoretically to a more pragmatic view of the underlying structure of the data set since the brain ROI's are not independent from each other, the variable selection method resulted in better clustering. One interpretation on why the Bayesian clustering performed better than the HDDC model for the AD data set can be addressed in the assumption of the Gaussian clusters for the later model. One may speculate that the clusters do not follow a normal distribution, and therefore assuming this decreases the performance of the clustering. In that sense, we can interpret that the Gaussianity assumption of the HDDC algorithm influences more harmfully the clustering than the independent variables assumption of Nia's Bayesian model.

The evaluation procedure for the optimal number of clusters indicated that 7 clusters is the optimal number of natural clusters in the data set under the Bayesian clustering. The reasons that at this level of the evaluation the Hopkins statistic has been used in contrast with the previous level are explained extensively in the section 4.5.2. It has been observed that the choice of 7 clusters manages to reveal 6 natural clusters and an outlier cluster of 2 observations. From that point of view the state of the art approach of Barnerjee and Dave (2004) used for the definition of the optimal number of clusters, proved to be superior against the classical distance based evaluation criteria. Despite the fact that the average Hopkins statistic worked well, one disadvantage with respect to all the evaluation criteria used in the thesis has been identified. It is addressed in one of the main objectives of the thesis, that is to address the clustering challenge of the AD data set in a high dimensional framework. For that reason non classical clustering algorithms are employed and the study focused in subspace clustering and variable selection. However the clustered groups have been mainly evaluated with respect to their distance characteristics. For the distance based evaluation criteria and the Hopkins statistic the Mahalanobis distance and the Euclidean distance have been used respectively. Therefore the high dimensional focus in the analysis, is limited to the modeling and not the validation. Oszust and Kostka (2015), in an effort to fill this gap in the literature, proposed a distance measure that computes the distance of two points taking into account only the dimensions that they share. However the two clustering algorithms used in the thesis do not result in the dimension that each of the data points lives, but only in their allocations into groups. To the best of our knowledge there are no clustering evaluation criteria specialized in high dimensional data sets. Despite that, after studying the Bayesian clustering in deep one could conclude that a more sophisticated way to compute the distance matrix after the clustering is by scaling the dimensions with respect to their variable importance. Since the Bayes factors are available for a variable and a variable-cluster level, we can compute the Euclidean distance in such a way that every dimension will be weighted with respect to its particular importance in the clustering. Future research in unsupervised learning should focus in the high dimensional clustering evaluation in addition to the high dimensional clustering methods.

6.2 Clustered groups

To the extent of our knowledge, this thesis is the first work that attempted to address the heterogeneity of Alzheimer's disease atrophy patterns from a high dimensional point of view, in the sense that the existing studies used clustering methods that treat all the dimensions as relevant. The resulting allocation defines groups of patients with diverse patterns of atrophy. In order to visualize the resulting atrophy patterns, as set of controls described in the data section is used. A compact visualization in section 5.2.3 allows to interpret the clusters further. At this point, is has to be reminded that the algorithm has been fed with atrophy measurements and its task is to define group with different patterns of atrophy. As is can be observed in the Figure 18 that presents the left and right lateral hemispheres and the Appendix fig. 3 that presents the subcortical regions for each group of patients, the clusters are quite different both in the level of 3 and 6 clusters. In level 6, patients with very few atrophy are allocated in cluster 3, while patients with the highest amounts of atrophy are grouped in cluster 4. A main characteristic of all the clusters is the atrophy in the lower temporal lobe which is also responsible for the memory functions. The clusters also have differences pathological characteristics. Patients in clusters with low atrophy have better scores in the cognitive tests (Table 6). In general the algorithm can be accounted as successful from a clustering perspective, since it is straightforward that the 6 clusters have different atrophies and thus the observations are well discriminated in many of the data set dimensions. Some of the groups define clusters with different median age. Patients do not differ in terms of disease duration so theoretically they should be at a similar stage of AD. Although they show a similar disease duration, implying that they might be in a similar disease stage, we found a large variability in the patterns of brain atrophy among AD patients and that these patterns were associated with the patient’s age and cognitive functions. This supports the notion that AD is heterogeneous.

In the 6 clusters level, cluster 3 is the one that concentrates the youngest patients if we take into account the Age (Table 6), and therefore the lowest atrophy can be observed in all the ROI's of the brain. Also the scores of this group in the CDR and ADAS are the lowest compared to the other groups revealing a group of patients with very good clinical characteristics and thus more preserved than the rest of the groups. Moreover, the MMSE of these patients has a median value of 24 which is only one degree lower than the threshold that discriminates the control with the AD cases. The next clusters with respect to increasing atrophy are clusters 1 and 2. These two clusters share very similar demographical, clinical and pathological characteristics. However they are located in different groups because their pattern of atrophy are slightly different both in the cortical and subcortical VOI's. Both their atrophy and additional characteristics place the patients of these two groups in the typical AD subtype. Moving on to the clusters 4, 5 and 6, the extended atrophy in most of the ROI's is easily observable. Again, some differences exist in the atrophy levels of these groups even if they present a diffuse atrophy subtype. One critical observation concerning all the clusters is the absence of atrophy in the lateral occipital cortex.[13] Moreover, the only directly comparable study to ours is the Hierarchical clustering of AD patients of Noh et al. (2014). In their case the clusters are not as compact as ours if we look separation from three to six clusters. If an agglomerative clustering is good, then when we split a cluster in two more clusters by cutting its branch in two parts, then these two new clusters should look alike in the sense that they belong to the same cluster higher in the hierarchy. This effect reveals the compactness of a cluster. In the case of our clustered groups visualization (Figure 18), one can easily observe that when a cluster in level 3 is split then the new clusters are quite close to each other. However this is

[13] The cortical and subcortical regions are presented in a brain map at the data section.

not the case in the study of Noh et al. (2014). In this sense the clustering of this thesis produced clusters that are more compact in level 3 and therefore more natural.

To summarize, the clustering of the AD patients succeeded in grouping the patients with respect to the extent of their brain atrophy. The resulting clusters do not present different stages of the disease because the duration of the disease in each cluster is more or less the same. However, we were not able to find all the AD subtypes described in the literature review. One reason can be addressed in the homogeneity of the data set. The data set used is a combination of the ADNI and AddNeuroMed data sets. The patients from the ADNI data set (155 out of 271) are very homogeneous in their pathology since they were chosen to be representative of the typical AD. Therefore the ability to find the atypical subtypes from this sample is reduced. Another reason is that in the studies described in the literature review that managed to find the three subtypes the authors always looked in specific regions of the brain and the methodologies were deterministic and not unsupervised apart from Noh et al. (2014). In this study, we put an effort in finding distinct clusters of atrophy by taking into account almost the whole brain in a ROI level. Therefore our approach is global in the sense that we studied the brain as an entirety. However, apart from the clusters 4, 5 and 6 that present the diffuse atrophy subtype, the remaining three clusters are quite promising to reveal the subtypes that we are looking for if we look lower in the dendrogram hierarchy.

7 Conclusions

This work assesses the ability of two clustering methods to reveal heterogeneous groups of patient in a high dimensional neuroimaging dataset. For the first method different initializations are examined and the best one is applied. The two methods are compared and the best one is evaluated for the optimal number of clusters. The results of the clustering are presented with interpretations and discussion.

The first model (GMM) uses an EM algorithm to find the clusters. The study of different initializations for the EM demonstrated that the incomplete correlation clustering returns better results, compared to the correlation clustering for the first 100 patients and the random allocation initialization.

The Bayesian clustering returns better scores than the GMM for the calculated distance based evaluation criteria. The optimal number of clusters was decided with the aid of the average Hopkins statistic and 6 is decided to be the optimal number of natural clusters, while one cluster of 2 outlier observations was excluded from the results.

The clusters correspond to different patterns and levels of atrophy. One cluster includes patients with atrophy only in the entorhinal cortex, the parahippocampal gyrus and the fusiform gyrus. This cluster has considerably good clinical and pathological characteristics. Two clusters have atrophy in the VOI's that the previous cluster has, but also in the temporal lobe and parts of the parietal lobe. These clusters have worse scores in their clinical tests than the medial temporal cluster and could be accounted as the typical AD subtype. Finally, three clusters have extent atrophy in most of the brain regions, excluding parts of the occipital lobe. These three clusters have the worst scores in the clinical tests and are considered as the diffuse atrophy Alzheimer's subtype.

One possible continuation in the exploration of the AD data set would be to repeat the analysis in a vertex level. Whilst VOI's are adequate to provide a lot of information about the atrophy levels in the brain regions, a vertex level analysis might provide more accurate information about the atrophy patterns. A vertex level analysis is feasible since the algorithms used for the thesis are able to accommodate high dimensional data sets. Another future extension is to add interactions of variables as covariates in the clustering in order to explore possible dependencies that are not explored by the clustering algorithms in the default settings.

8 Literature

Achard, S., & Bullmore, E. (2007). Efficiency and cost of economical brain functional networks. *PLoS Comput Biol, 3*(2), e17.

Ailon, N., Charikar, M., & Newman, A. (2008). Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM), 55*(5), 23.

Alpaydin, E. (2014). Introduction to machine learning. *MIT press*, Cambridge, MA.

Andreasen, N., Minthon, L., Clarberg, A., Davidsson, P., Gottfries, J., Vanmechelen, E., ... & Blennow, K. (1999). Sensitivity, specificity, and stability of CSF-tau in AD in a community-based patient sample. *Neurology, 53*(7), 1488-1488.

Armstrong, R. A., Nochlin, D., & Bird, T. D. (2000). Neuropathological heterogeneity in Alzheimer's disease: a study of 80 cases using principal components analysis. *Neuropathology, 20*(1), 31-37.

Arndt, S., Cohen, G., Alliger, R. J., Swayze, V. W., & Andreasen, N. C. (1991). Problems with ratio and proportion measures of imaged cerebral structures. *Psychiatry Research: Neuroimaging*,*40*(1), 79-89.

Banerjee, A., & Davé, R. N. (2004, July). Validating clusters using the Hopkins statistic. In *Fuzzy systems, 2004. Proceedings. 2004 IEEE international conference on* (Vol. 1, pp. 149-153). IEEE.

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89-113.

Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M., Lehmann, M., Hobbs, N., ... & Fox, N. C. (2010). Head size, age and gender adjustment in MRI studies: a necessary nuisance?. *Neuroimage*,*53*(4), 1244-1255.

Basu, S., Banerjee, A., & Mooney, R. (2002). Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*.

Becker, J. T., Huff, F. J., Nebes, R. D., Holland, A., & Boller, F. (1988). Neuropsychological function in Alzheimer's disease: pattern of impairment and rates of progression. *Archives of Neurology*, *45*(3), 263.

Bellman, R. (1961). Adaptive control: a guided tour. *Princeton University Press, Princeton, NJ.*

Ben-David, S., & Ackerman, M. (2009). Measures of clustering quality: A working set of axioms for clustering. In *Advances in neural information processing systems* (pp. 121-128).

Berg, J., & Jarvisalo, M. (2013). Optimal correlation clustering via MaxSAT. *IEEE 13th International Conference on Data Mining Workshops (ICDMW),* (pp. 750-757).

Bigio, E. H., Hynan, L. S., Sontag, E., Satumtira, S., & White, C. L. (2002). Synapse loss is greater in presenile than senile onset Alzheimer disease: implications for the cognitive reserve hypothesis.*Neuropathology and applied neurobiology*, *28*(3), 218-227.

Bouveyron, C., Girard, S., & Schmid, C. (2006). Class-specific subspace discriminant analysis for high-dimensional data. In *Subspace, Latent Structure and Feature Selection* (pp. 139-150). Springer Berlin Heidelberg

Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis*, *52*(1), 502-519.

Bouwman, F. H., Schoonenboom, N. S., Verwey, N. A., van Elk, E. J., Kok, A., Blankenstein, M. A., ... & van der Flier, W. M. (2009). CSF biomarker levels in early and late onset Alzheimer's disease. *Neurobiology of aging*,*30*(12), 1895-1901.

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science, AAAS*, *325*(5939), 414.

Byun, M. S., Kim, S. E., Park, J., Yi, D., Choe, Y. M., Sohn, B. K., ... & Lee, D. Y. (2015). Heterogeneity of Regional Brain Atrophy Patterns Associated with Distinct Progression Rates in Alzheimer's Disease. *PloS one*, *10*(11), e0142756.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276.

Charikar, M., Guruswami, V., & Wirth, A. (2003, October). Clustering with qualitative information. In*Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on* (pp. 524-533). IEEE.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G., ... & Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families.*Science*, *261*(5123), 921-923.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179-194.

Datta, S., Datta, S., Pihur, V., & Brock, G. (2008). clValid: an R package for cluster validation. *Journal of Statistical Software*, *25*(04).

Davidson, Y., Gibbons, L., Pritchard, A., Hardicre, J., Wren, J., Stopford, C., ... & Pendleton, N. (2006). Apolipoprotein E ε4 allele frequency and age at onset of Alzheimer's disease. *Dementia and geriatric cognitive disorders*, *23*(1), 60-66.

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

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... & Albert, M. S. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.*Neuroimage*, *31*(3), 968-980.

Dickerson, B. C., & Wolk, D. A. (2012). MRI cortical thickness biomarker predicts AD-like CSF and cognitive decline in normal adults. *Neurology,78*(2), 84-90.

Dickerson, B. C., Goncharova, I., Sullivan, M. P., Forchetti, C., Wilson, R. S., Bennett, D. A., & Beckett, L. A. (2001). MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. *Neurobiology of aging*, *22*(5), 747-754.

Dickerson, B. C., Stoub, T. R., Shah, R. C., Sperling, R. A., Killiany, R. J., Albert, M. S., ... & Blacker, D. (2011). Alzheimer-signature MRI biomarker predicts AD dementia in cognitively normal adults. *Neurology*, *76*(16), 1395-1402.

Everitt, B. S., Landau, S., & Leese, M. (2001). Cluster analysis. 2001. *Arnold, London*.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., ... Montillo, A. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341-355.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Caviness, V. (2004). Automatically parcellating the human cerebral cortex. *Cerebral cortex*, *14*(1), 11-22.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), 611-631.

Frisoni, G. B., Testa, C., Sabattoli, F., Beltramello, A., Soininen, H., & Laakso, M. P. (2005). Structural correlates of early and late onset Alzheimer's disease: voxel based morphometric study. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(1), 112-114.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, *2*(1-2), 56-78.

Galton, C. J., Patterson, K., Xuereb, J. H., & Hodges, J. R. (2000). Atypical and typical presentations of Alzheimer's disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases. *Brain*, *123*(3), 484-498.

Gebser, M., Kaminski, R., Kaufmann, B., Ostrowski, M., Schaub, T., & Thiele, S. (2008). A user's guide to gringo, clasp, clingo, and iclingo.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. Statistica sinica, 339-373.

Giotis, I., & Guruswami, V. (2006, January). Correlation clustering with a fixed number of clusters. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm* (pp. 1167-1176). Society for Industrial and Applied Mathematics.

Gorno- Tempini, M. L., Dronkers, N. F., Rankin, K. P., Ogar, J. M., Phengrasamy, L., Rosen, H. J., ... & Miller, B. L. (2004). Cognition and anatomy in three variants of primary progressive aphasia.*Annals of neurology*, *55*(3), 335-346.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, *17*(2), 107-145.

Handl, J., & Knowles, J. (2005, March). Exploiting the trade-off—the benefits of multiple objectives in data clustering. In *Evolutionary Multi-Criterion Optimization* (pp. 547-560). Springer Berlin Heidelberg.

Heard, N. A., Holmes, C. C., & Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, *101*(473), 18-29.

Hof, P. R., Bouras, C., Constandinidis, J., & Morrison, J. H. (1989). Balit's syndrome in Alzheimer's disease: specific disruption of the occipito-parietal visual pathway. *Brain Research*,*493*(2), 368-375.

Hopkins, B., & Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany*, *18*(2), 213-227.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., ... & Dale, A. M. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods.*Journal of Magnetic Resonance Imaging*, *27*(4), 685-691.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall Inc., Michigan state University, New Jersey, 201-202.

Johnson, J. K., Head, E., Kim, R., Starr, A., & Cotman, C. W. (1999). Clinical and pathological evidence for a frontal variant of Alzheimer disease. *Archives of Neurology*, *56*(10), 1233-1239.

Karantzoulis, S., & Galvin, J. E. (2011). Distinguishing Alzheimer’s disease from other major forms of dementia. *Expert review of neurotherapeutics*, Vol. 11, Iss. 11.

Karlis, D., & Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures.*Computational Statistics & Data Analysis*, *41*(3), 577-590.

Kerr, M. K., & Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical research*, *77*(02), 123-128.

Koedam, E. L., Lauffer, V., van der Vlies, A. E., van der Flier, W. M., Scheltens, P., & Pijnenburg, Y. A. (2010). Early-versus late-onset Alzheimer's disease: more than age alone. *Journal of Alzheimer's Disease*, *19*(4), 1401-1408.

Koss, E., Edland, S., Fillenbaum, G., Mohs, R., Clark, C., Galasko, D., & Morris, J. C. (1996). Clinical and neuropsychological differences between patients with earlier and later onset of Alzheimer's disease A CERAD analysis, part XII.*Neurology*, *46*(1), 136-141.

Lehmann, M., Ghosh, P. M., Madison, C., Laforce, R., Corbetta-Rastelli, C., Weiner, M. W., ... & Miller, B. L. (2013). Diverging patterns of amyloid deposition and hypometabolism in clinical variants of probable Alzheimer's disease. *Brain*, *136*(3), 844-858.

Lehtovirta, M., Soininen, H., Helisalmi, S., Mannermaa, A., Helkala, E. L., Hartikainen, P., ... & Riekkinen, P. J. (1996). Clinical and neuropsychological characteristics in familial and sporadic Alzheimer's disease Relation to apolipoprotein E polymorphism. *Neurology*, *46*(2), 413-419.

Li, C. M., & Manya, F. (2009). MaxSAT, Hard and Soft Constraints. *Handbook of satisfiability*, *185*, 613-631.

Lovestone, S., Francis, P., & Strandgaard, K. (2007). Biomarkers for disease modification trials-The innovative medicines initiative and AddNeuroMed. *The journal of nutrition, health & aging*, *11*(4), 359.

Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., ... & Ward, M. (2009). AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences*, *1180*(1), 36-46.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta), 2, 49-55.

Marra, C., Bizzarro, A., Daniele, A., De Luca, L., Ferraccioli, M., Valenza, A., ... & Masullo, C. (2004). Apolipoprotein E ε4 allele differently affects the patterns of neuropsychological presentation in early-and late-onset Alzheimer's disease patients. *Dementia and geriatric cognitive disorders*, *18*(2), 125-131.

Mattsson, N., Zetterberg, H., Hansson, O., Andreasen, N., Parnetti, L., Jonsson, M., ... & Rich, K. (2009). CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *Jama*, *302*(4), 385-393.

McLachlan, G. J. (1992). Discriminant analysis and statistical pattern recognition. Wiley publications.

McLachlan, G. J., & Krishnan, T. (1997). The EM Algorithm and Extensions. Wiley publications.

Meek, P. D., McKeithan, E. K., & Schumock, G. T. (1998). Economic considerations in Alzheimer's disease. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *18*(2P2), 68-73.

Mulder, C., Verwey, N. A., van der Flier, W. M., Bouwman, F. H., Kok, A., van Elk, E. J., ... & Blankenstein, M. A. (2010). Amyloid-β (1–42), total tau, and phosphorylated tau as cerebrospinal fluid biomarkers for the diagnosis of Alzheimer disease.*Clinical chemistry*, *56*(2), 248-253.

Murray, M. E., Graff-Radford, N. R., Ross, O. A., Petersen, R. C., Duara, R., & Dickson, D. W. (2011). Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *The Lancet Neurology*, *10*(9), 785-796.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167-256.

Nia, V. P. (2009). *Fast High-Dimensional Bayesian Classification and Clustering* (Doctoral dissertation, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE).

Nochlin, D., Van Belle, G., Bird, T. D., & Sumi, S. M. (1992). Comparison of the severity of neuropathologic changes in familial and sporadic Alzheimer's disease. *Alzheimer disease and associated disorders*, *7*(4), 212-222.

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., ... & Park, K. H. (2014). Anatomical heterogeneity of Alzheimer disease Based on cortical thickness on MRIs. *Neurology*, *83*(21), 1936-1944.

Nordenskjöld, R., Malmberg, F., Larsson, E. M., Simmons, A., Brooks, S. J., Lind, L., ... & Kullberg, J. (2013). Intracranial volume estimated with commonly used methods could introduce bias in studies including brain volume measurements.*Neuroimage*, *83*, 355-360.

O'Brien, L. M., Ziegler, D. A., Deutsch, C. K., Kennedy, D. N., Goldstein, J. M., Seidman, L. J., ... & Herbert, M. R. (2006). Adjustment for whole brain and cranial size in volumetric brain studies: a review of common adjustment factors and statistical methods. *Harvard review of psychiatry*,*14*(3), 141-151.

Oszust, M., & Kostka, M. (2015). Evaluation of Subspace Clustering Using Internal Validity Measures. *ADVANCES IN ELECTRICAL AND COMPUTER ENGINEERING*, *15*(3), 141-146.

Panayirci, E., & Dubes, R. C. (1983). A test for multidimensional clustering tendency. *Pattern Recognition*, *16*(4), 433-444.

Parsons, L., Haque, E., & Liu, H. (2004) . Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, *6*(1), 90-105.

Partovi Nia, V., & Davison, A. C. (2015). A simple model- based approach to variable selection in classification and clustering. *Canadian Journal of Statistics*, *43*(2), 157-175.

Pintzka, C. W., Hansen, T. I., Evensmoen, H. R., & Håberg, A. K. (2015). Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: a HUNT MRI study. *Frontiers in neuroscience*, *9*.

Pintzka, C. W., Hansen, T. I., Evensmoen, H. R., & Håberg, A. K. (2015). Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: a HUNT MRI study. *Frontiers in neuroscience*, *9*.

Reitz, C., Brayne, C., & Mayeux, R. (2011). Epidemiology of Alzheimer disease. *Nature Reviews Neurology*, *7*(3), 137-152.

Rossor, M. N., Iversen, L. L., Reynolds, G. P., Mountjoy, C. Q., & Roth, M. (1984). Neurochemical characteristics of early and late onset types of Alzheimer's disease. *Br Med J (Clin Res Ed)*, *288*(6422), 961-964.

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.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, *52*(3), 1059-1069.

Sanfilipo, M. P., Benedict, R. H., Zivadinov, R., & Bakshi, R. (2004). Correction for intracranial volume in analysis of whole brain atrophy in multiple sclerosis: the proportion vs. residual method. *Neuroimage*, *22*(4), 1732-1743.

Scott, D. W., & Thompson, J. R. (1983). Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proceedings of the fifteenth symposium on the interface* (Vol. 528, pp. 173-179). North-Holland, Amsterdam.

Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L., & Greicius, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks.*Neuron*, *62*(1), 42-52.

Shaw, L. M., Vanderstichele, H., Knapik- Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., ... & Dean, R. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of neurology*,*65*(4), 403-413.

Shiino, A., Watanabe, T., Kitagawa, T., Kotani, E., Takahashi, J., Morikawa, S., & Akiguchi, I. (2008). Different atrophic patterns in early-and late-onset Alzheimer’s disease and evaluation of clinical utility of a method of regional z-score analysis using voxel-based morphometry. *Dementia and geriatric cognitive disorders*, *26*(2), 175-186.

Shiino, A., Watanabe, T., Maeda, K., Kotani, E., Akiguchi, I., & Matsuda, M. (2006). Four subgroups of Alzheimer's disease based on patterns of atrophy using VBM and a unique pattern for early onset disease. *Neuroimage*, *33*(1), 17-26.

Simmons, A., Westman, E., Muehlboeck, S., Mecocci, P., Vellas, B., Tsolaki, M., ... & Evans, A. (2009). MRI measures of Alzheimer's disease and the AddNeuroMed study. *Annals of the New York Academy of Sciences*, *1180*(1), 47-55.

Simmons, A., Westman, E., Muehlboeck, S., Mecocci, P., Vellas, B., Tsolaki, M., ... & Evans, A. (2011). The AddNeuroMed framework for multi- centre MRI assessment of Alzheimer's disease: experience from the first 24 months. *International journal of geriatric psychiatry*, *26*(1), 75-82.

Sluimer, J. D., Bouwman, F. H., Vrenken, H., Blankenstein, M. A., Barkhof, F., van der Flier, W. M., & Scheltens, P. (2010). Whole-brain atrophy rate and CSF biomarker levels in MCI and AD: a longitudinal study. *Neurobiology of Aging*, *31*(5), 758-764.

Sluimer, J. D., Vrenken, H., Blankenstein, M. A., Fox, N. C., Scheltens, P., Barkhof, F., & van der Flier, W. M. (2008). Whole-brain atrophy rate in Alzheimer disease Identifying fast progressors.*Neurology*, *70*(19 Part 2), 1836-1841.

Van der Flier, W. M., Pijnenburg, Y. A., Fox, N. C., & Scheltens, P. (2011). Early-onset versus late-onset Alzheimer's disease: the case of the missing APOE ε4 allele. *The Lancet Neurology*,*10*(3), 280-288.

Van der Flier, W. M., Staekenborg, S., Pijnenburg, Y. A., Gillissen, F., Romkes, R., Kok, A., ... & Scheltens, P. (2007). Apolipoprotein E genotype influences presence and severity of delusions and aggressive behavior in Alzheimer disease.*Dementia and geriatric cognitive disorders*, *23*(1), 42-46.

Voevodskaya, O., Simmons, A., Nordenskjöld, R., Kullberg, J., Ahlström, H., Lind, L., ... & Westman, E. (2014). The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease.*Frontiers in aging neuroscience*, *6*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*,*393*(6684), 440-442.

Westman, E., Aguilar, C., Muehlboeck, J. S., & Simmons, A. (2013). Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. *Brain topography*, *26*(1), 9-23.

Westman, E., Muehlboeck, J. S., & Simmons, A. (2012). Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion.*Neuroimage*, *62*(1), 229-238.

Westman, E., Simmons, A., Muehlboeck, J. S., Mecocci, P., Vellas, B., Tsolaki, M., ... & Spenger, C. (2011). AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. *Neuroimage*, *58*(3), 818-828.

Whitwell, J. L., Dickson, D. W., Murray, M. E., Weigand, S. D., Tosakulwong, N., Senjem, M. L., ... & Jack, C. R. (2012). Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. *The Lancet Neurology*, *11*(10), 868-877.

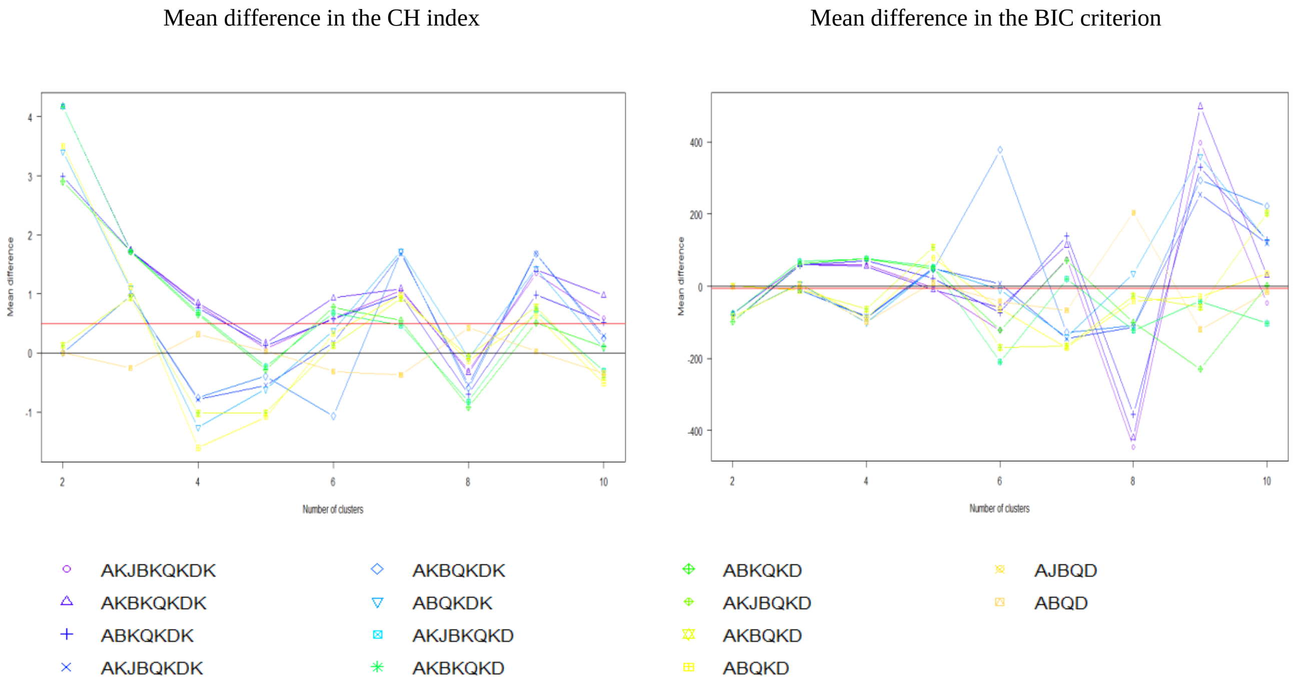
Zatz, L. M., & Jernigan, T. L. (1983). The ventricular-brain ratio on computed tomography scans: validity and proper use. *Psychiatry Research*, *8*(3), 207-214.

Zhou, J., Gennatas, E. D., Kramer, J. H., Miller, B. L., & Seeley, W. W. (2012). Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron*, *73*(6), 1216-1227.

Zhou, L., Wang, Y., Li, Y., Yap, P. T., Shen, D., & Alzheimer's Disease Neuroimaging Initiative (ADNI. (2011). Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. *PloS one*, *6*(7), e21935.

9 Appendix

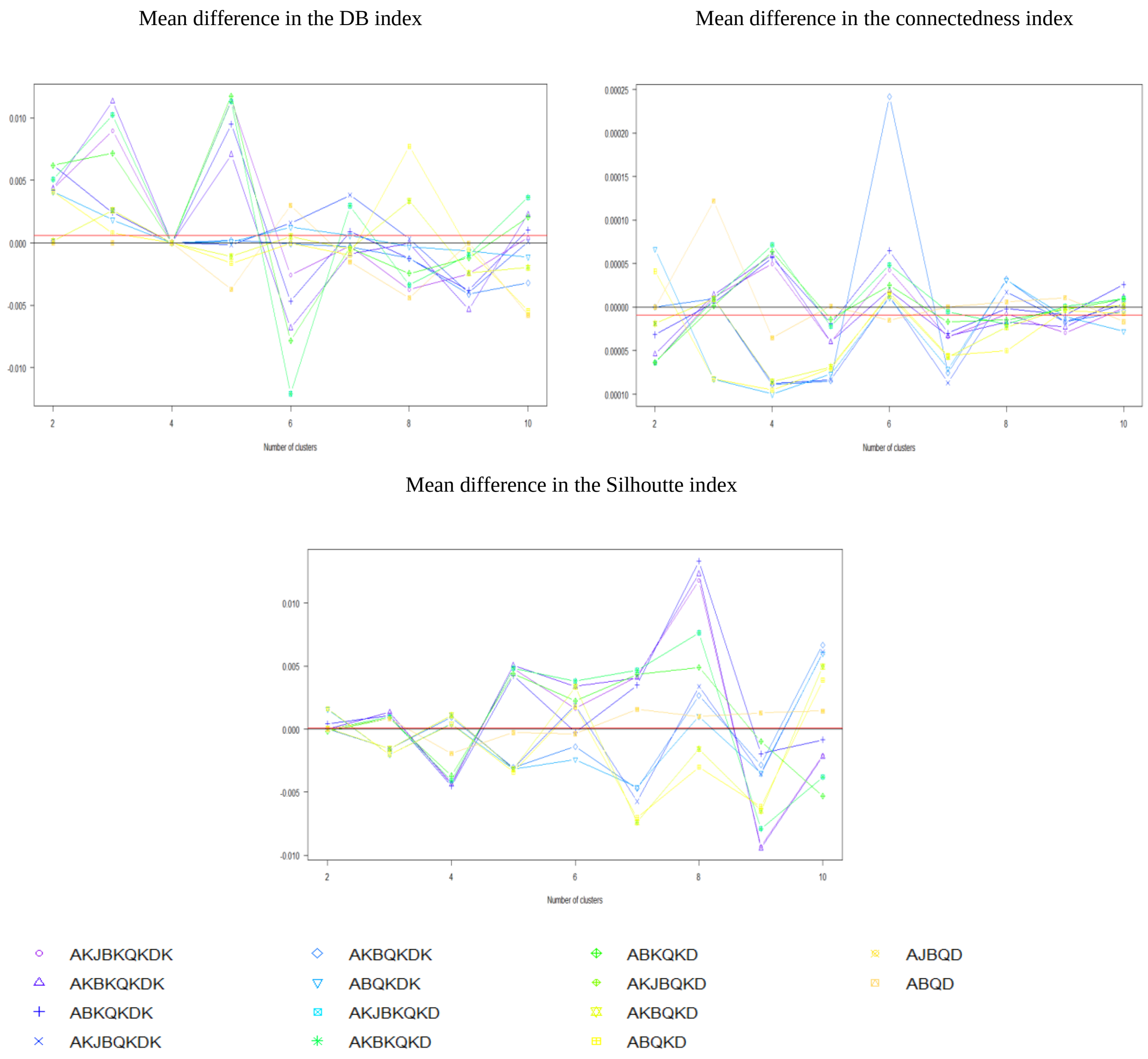
Here you can put detailed tables, program code etc.



Mean difference in the CH index

Mean difference in the BIC criterion

Appendix fig. 1 Mean difference on CH index and BIC criterion for the initialization (Init 3) over the 14 models of the $[a_{ij} b_i Q_i d_i]$ GMM family. Here the difference of the indices is formulated as diff $= \frac{1}{100}\sum_{i=1}^{100}(I_{Init\ 3} - I_{Init\ 2})$, where $I = \{CH, BIC\}$. *Init* 2 refers to random initialization of the EM, while *Init* 3 refers to the initialization with the correlation clustering allocation until the operating system runs out of memory. The red line in the plots indicates the mean difference for all the methods and clusters.

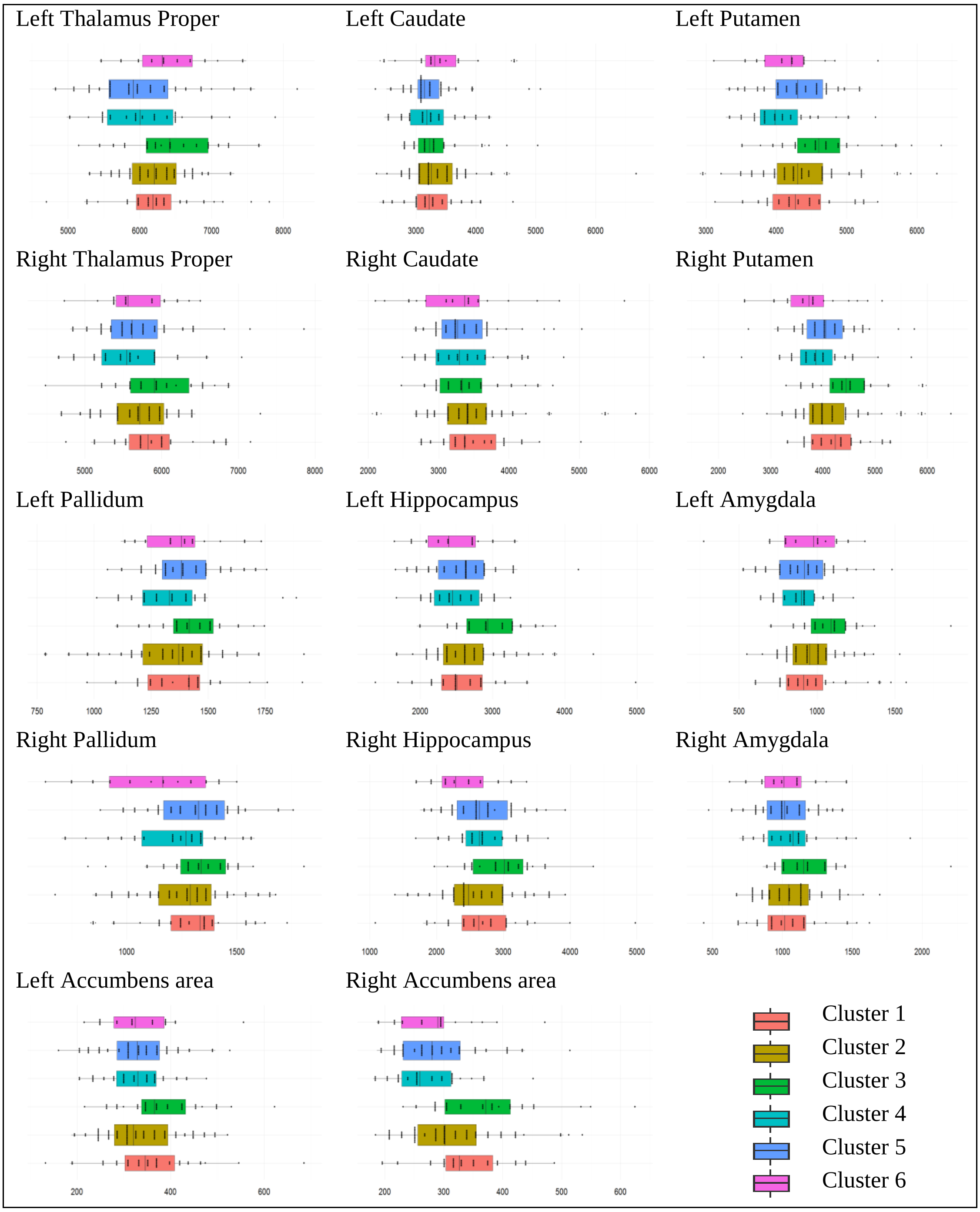


Mean difference in the DB index

Mean difference in the connectedness index

Mean difference in the Silhoutte index

Appendix fig. 2 Mean difference on DB, the Silhouette and the connectivity indexes for the two initializations over the 14 models of the $\left[a_{ij}\, b_i Q_i d_i\right]$ GMM family. Here the difference of the indices is formulated as $\mathrm{diff} = \frac{1}{100}\sum_{i=1}^{100}(I_{Init\ 3} - I_{Init\ 2})$, where $I = \{Conndectedness, DB, Silhouette\}$. *Init* 2 refers to random initialization of the EM, while *Init* 3 refers to the initialization with the correlation clustering allocation until the operating system runs out of memory. The red line in the plots indicates the mean difference for all the methods and clusters.



Appendix fig. 3 Box-plots of gray matter findings in the 14 subcortical region for the 6 clusters. Box height is related to the sample size of each cluster.

LIU-IDA/STAT-A--16/001—SE