Using Cluster Analysis, Cluster Validation, and Consensus Clustering to Identify Subtypes of Pervasive Developmental Disorders

By

Jess Jiangsheng Shen

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
November 2007

Abstract

Pervasive Developmental Disorders (PDDs) are neurodevelopmental disorders characterized by impairments in social interaction, communication and behaviour [Str04]. Given the diversity and varying severity of PDDs, diagnostic tools attempt to identify homogeneous subtypes within PDDs.

The diagnostic system *Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition* (DSM-IV) divides PDDs into five subtypes. Several limitations have been identified with the categorical diagnostic criteria of the DSM-IV. The goal of this study is to identify putative subtypes in the multidimensional data collected from a group of patients with PDDs, by using cluster analysis.

Cluster analysis is an unsupervised machine learning method. It offers a way to partition a dataset into subsets that share common patterns. We apply cluster analysis to data collected from 358 children with PDDs, and validate the resulting clusters. Notably, there are many cluster analysis algorithms to choose from, each making certain assumptions about the data and about how clusters should be formed. A way to arrive at a meaningful solution is to use consensus clustering to integrate results from several clustering attempts that form a *cluster ensemble* into a unified consensus answer, and can provide robust and accurate results [TJPA05].

In this study, using cluster analysis, cluster validation, and consensus clustering, we identify four clusters that are similar to – and further refine – three of the five subtypes defined in the DSM-IV. This study thus confirms the existence of these three subtypes among patients with PDDs.

Co-Authorship

Sections 3.2, 3.3 and 5.4 contain methods and results that were developed, obtained and published in collaboration with authors in the following paper:

J. Shen, P. Lee, J. A. Holden, H. Shatkay: Using Cluster Ensemble and Validation to Identify Subtypes of Pervasive Developmental Disorders. Proc. of the Annual Symposium of the American Medical Informatics Association (AMIA), 2007. (To appear)

Acknowledgements

I would like to express my sincerest thanks to my supervisor, Dr. Hagit Shatkay, for her great guidance, support, patience and understanding throughout the course of this thesis. I would like to thank her for all her help and advice over the years I have pursued my education and research at Queen's University.

I gratefully thank Drs Jeanette Holden, David Skillicorn, and Ira Cohen for kindly offering me their invaluable expertise and comments in the development of my thesis.

I would also like to extend my gratitude to the Ontario Graduate Scholarship Program and the School of Computing at Queen's University for funding my studies.

I would like to thank my thesis examination committee, Dr. Blostein, Dr. Graham, Dr. Holden, and Dr. Troje for their insightful suggestions. I also thank the chair of my defense, Dr. Olivo, for steering the ship.

Thanks to Fengxia and Phil for offering me their help, suggestions, and spiritual support. I will always remember the days we spent together.

Thanks to Larry for having faith in me, giving me unwavering support, and being there for me throughout all the difficulties.

Finally, I would like to express my sincerest gratitude to my parents for giving me their endless and unconditional love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Pervasive Developmental Disorders (*PDDs*) are a group of neurodevelopmental disorders of varying severity that affect three core areas: communication skills, social interaction, and behaviour patterns [Str94]. They have a large range of manifestations without a well-understood aetiology, although multi-genetic factors are believed to play a key role in causing PDDs [MTR04]. Dividing PDDs into homogeneous subgroups through the examination of symptomatic behaviours of subjects with PDDs can shed light on the discovery of genetic causes behind them, and guide clinicians to select appropriate treatments. Given the diversity of PDDs, current diagnostic standards, such as those included in the widely used *Diagnostic and Statistical Manual of Mental Disorders - 4th Edition* (DSM-IV) [APA94] attempt to provide diagnostic criteria to divide PDDs into relatively homogeneous subtypes by evaluating the three core areas affected by PDDs. The DSM-IV distinguishes among five categories: Autistic disorder (Autism), Childhood disintegrative disorder, Rett's disorder, Asperger's disorder, and Pervasive Developmental Disorder - Not Otherwise Specified (PDD-NOS).

The DSM-IV assigns one of the above five subtypes to patients, based on whether a cut-off threshold for certain criteria is met or not, and sets discrete boundaries among subtypes and between normal and abnormal conditions. However, empirical findings suggest that the subtypes thus defined may not represent distinct diagnostic groups [TD05, SOBD05, MCC01]. Moreover, using a cut-off threshold to categorically distinguish normal from abnormal values may cause the loss of potentially important information.

This thesis is about using *Cluster Analysis* to find subtypes for PDDs. Cluster analysis is a multivariate unsupervised machine learning method that partitions a dataset into subsets sharing common patterns. In contrast to the DSM-IV which divides PDDs into subgroups based on discrete and mutually exclusive impairments, cluster analysis evaluates the condition of patients based on a continuous view. That is, it partitions subjects according to the severity of their impairment, without using a cut-off value as a threshold to distinguish between the normal and the abnormal conditions [Kes02].

In cluster analysis, the underlying cluster structure of the data is *a-priori* unknown. Almost every clustering algorithm will find clusters in a dataset, even if there is no cluster structure in it. Moreover, the discovered subsets can be arbitrary when clustering methods are applied to the data. Therefore, we want to evaluate the clustering solutions quantitatively and objectively, a process called *Cluster Validation* [JD98]. Another issue with cluster analysis is that different algorithms may lead to different partitions of the data into clusters. Recent studies on cluster analysis attempt to circumvent this problem using consensus clustering, which puts multiple clustering results in a cluster ensemble to reach a single consensus solution [HK06].

Previous studies that employed cluster analysis for the subtyping of PDDs either used a very small dataset (30-50 cases) [SOHH99], used non-standard diagnostic tools [SACB86], or included in their studies patients with other conditions that are not PDDs [Res88], thus making the results less applicable to PDDs. Many of the previous studies did not use objective validation methods that are well-justified [SACB86, WMAD96, SFDW00], or did not use validation methods at all [Res88, PBM04]. In addition, previous studies on the subtyping of PDDs all employ a single clustering method.

1.1 Motivation

Our work is motivated by the need to build a well-validated diagnostic framework that employs a continuous view of the data, to identify the subtypes in PDDs. Such a framework may be an alternative to the diagnostic criteria for categorizing the subtypes of PDDs in the DSM-IV [APA94].

As part of our diagnostic framework, multiple clustering results produced with cluster analysis are validated and combined into one unique solution using consensus clustering. Formal methods of cluster validation examine how well a clustering fits a dataset (*fitness validation*) and how robust the clustering is in the face of perturbation in the data (*stability validation*) [TSK05]. Consensus clustering then puts multiple validated clustering results into an ensemble and combines them into a single consensus solution [HK06a]. Thus, consensus clustering can improve clustering performance by consolidating the output on which several clustering algorithms agree, typically leading to more robust results than those produced by any individual method [TLJF04].

The goal of this thesis is to identify the subtypes of PDDs using the combination of cluster analysis, cluster validation, and consensus clustering.

1.2 Contribution

In this thesis, we make several contributions. We first provide a broad survey of the general background of PDDs, including previous work on the subtyping of PDDs. We then review cluster analysis, cluster validation, and consensus clustering, which are applied to the data collected using a well-validated tool called the *Autism Diagnostic Interview – Revised* (*ADI-R*) to identify the subtypes in PDDs. To the best of our

knowledge, this is the largest study performed so far on subtyping PDDs based on the ADI-R data. It is also the first one to apply consensus clustering and a full-scale internal cluster validation, as discussed in Section 2.3. We identify four clusters that roughly correspond to – and further refine – three main subtypes of PDDs, namely Autism, Asperger's disorder, and PDD-NOS.

1.3 Thesis Organization

The rest of the thesis is organized as follows: An overview of PDDs and cluster analysis, cluster validation, consensus clustering, and related work on the subtyping of PDDs is presented in Chapter 2. Raw data collected from patients suffering from PDDs using the ADI-R, as well as the pre-processing procedure performed on the raw data, are described in Chapter 3. Chapter 4 discusses in detail the methodology of clustering the pre-processed data, focusing on cluster analysis, cluster validation, and consensus clustering. Experimental results and analysis are provided in Chapter 5, followed by conclusions and future study in Chapter 6.

Chapter 2

Pervasive Developmental Disorders and Cluster Analysis: Background

Our study on subtype identification of pervasive developmental disorders lies in the intersection of research on these disorders and on cluster analysis. In this chapter, we introduce pervasive developmental disorders, their diagnostic criteria, and previous research on the subtype identification of them. We then survey cluster analysis, cluster validation, and consensus clustering, which are the three major components of the methodology used in this thesis. We conclude this chapter with a survey of applications of cluster analysis in the subtyping of pervasive developmental disorders.

2.1 Pervasive Developmental Disorders

Pervasive developmental disorders (PDDs) are characterized by impairment in three core areas, namely communication skills, social interaction, and behaviour patterns, and may also be accompanied by deficiencies in cognitive ability, usually reflected by a lower IQ compared to the general population, epilepsy, and other co-morbidities [SLC03]. Consequently, individuals suffering from PDDs form a heterogeneous group which shows a wide range of social impairments, behavioural problems, communicational and cognitive difficulties [ASC05].

Current research suggests that multiple genes are associated with these disorders [FH05, SLMT07]. Different combinations of damaged genes interact with multiple environmental factors to cause PDDs [FRS01]. In fact, PDDs may be the final common

result for several different genetic abnormalities that share similar symptoms [Ash07]. Due to the wide-range of symptoms related to PDDs, a good diagnostic system is desired in order to divide PDDs into more homogenous subtypes.

2.1.1 Clinical Subtypes of PDDs

Since there is no confirmed biological/medical indicator for PDDs, diagnosis relies on experienced evaluators to identify predictive features based on medical history, observation, and the use of assessment tools [MCBM03].

Current psychiatric classification standards, such as the widely-used Diagnostic and Statistical Manual of Mental Disorders – 4th Edition (DSM-IV), provide diagnostic criteria by evaluating the three core areas affected by PDDs [APA94]. The five clinical subtypes of PDDs proposed by the DSM-IV are:

(1) *Childhood disintegrative disorder*;

(2) *Rett's disorder*;

(3) *Autistic disorder* , which is also referred to as *Autism*[1] in the DSM-IV;

(4) *Pervasive developmental disorder - not otherwise specified* (*PDD-NOS*);

(5) *Asperger's disorder*.

Subtypes (3) to (5) are the three subtypes of PDDs that are commonly observed, while Subtypes (1) and (2) are rare. Assessment tools are applied to individuals to collect information for making diagnoses on the subtypes of PDDs.

[1] Although *autism* is frequently used as a broad term to refer to PDDs in literatures, we adhere to its DSM-IV definition to reduce confusion.

2.1.2 Assessment Tools

Three main types of assessment tools are developed to gather clinical information and to enforce uniformity of assessment and diagnosis among clinicians: questionnaires, observational schedules, and interviews. Questionnaires are forms containing a fixed set of questions. They are submitted to patients (or caregivers who know the developmental history of the patients well) to collect responses. Interviews are performed by researchers who pose interviewees (patients or caregivers) with questions. In an interview, the researchers can clarify the interviewees' understanding of a given question. The interviewees are allowed to respond in their own words, and in greater detail. Interviews can result in much more information than questionnaires. Observational schedules are sessions designed for the researchers to observe and evaluate the behaviours of the subjects. Questionnaires are more suitable for screening for potential patients with PDDs before formal diagnosis can be made. Interviews and observational schedules are often used as diagnostic tools to decide whether an individual has PDDs or not, and what subtype of PDDs he has. Observational schedules are essential in a diagnostic process, but they are limited to evaluating behaviours at the point of observation without taking history into account; this limitation can be overcome using interviews [LCRL89].

Among many assessment tools, a few are diagnostic tools, and may be used as part of a formal diagnosis process. Here we introduce two representative diagnostic tools: an interview called *Autism Diagnostic Interview-Revised* (ADI-R) [LRC94] and an observational schedule called *Autism Diagnostic Observation Schedule* (ADOS) [LRGH89].

Both the ADI-R and the ADOS assessment tools use diagnostic algorithms based on separate cut-offs for scores received along the communicational, social, and behavioural areas, allowing for separate quantification of severity in each of these areas. The ADI-R also takes into consideration early history of development such as the time of the start of symptoms.

In the ADI-R, each of the three core areas is divided into four sub-areas that address different aspects of the same area. Each of the sub-areas is further divided into two to five specific impairments. According to the algorithm accompanying the ADI-R, scores are computed for the sub-areas, then the areas. That is, the scores of the specific two to five impairments in a sub-area are summed up to be the score of this sub-area; the scores of the four sub-areas in each area are then summed up to be the score of this area [LRL94]. By comparing the computed area scores and the starting time of the abnormalities[2] for an individual to the corresponding thresholds, this accompanying algorithm of the ADI-R effectively distinguishes between Autism and all other subtypes of PDDs, as well as between PDDs and other psychological disorders [Lor95, DBSK04]. However, it does not identify the subtypes of PDDs other than Autism. Notably, this algorithm is not the only approach to analyzing the ADI-R data. Researchers have taken advantage of the raw ADI-R scores to perform additional analysis [ZSNH03, VGRR06, TRD98].

Aside from its good performance in diagnosing Autism, the ADI-R also includes questions about the early development history of the subjects. Therefore, the data collected with the ADI-R have been widely used in assessing boundaries between the subtypes of PDDs, identifying new subgroups of PDDs, and quantifying symptoms

[2] For a diagnosis of Autism to be made, the starting time of abnormalities should be before an individual is 36 months old.

related to PDDs [TRD98, SGBZ06]. Data analyzed in our study are collected using the ADI-R, and are therefore considered standardized, comprehensive, and truly reflect the impairments of the patients with PDDs.

The ADOS is a test based on the direct observation of the behaviour of a child by her examiner. It consists of a semi-structured assessment of play, interaction, and social communication. Among the observation-based diagnostic tools for PDDs, the ADOS is the most widely supported by researchers [LRLC00]. Algorithms for the diagnosis of Autism for the ADI-R and the ADOS both employ a particular view on PDDs called the categorical view, which is discussed in detail in the next section.

2.1.3 Categorical View

The categorical view is held by researchers who consider mental illnesses as discrete conditions with clear boundaries, represented by cut-off thresholds, between different types and subtypes, and between normal and abnormal conditions [TD05]. The categorical view is employed by the diagnostic criteria in the DSM-IV, and determines how the diagnostic subtypes of PDDs are defined. For example, as shown in Appendix A, a subtype of PDDs is usually defined as having *abnormalities* in *n* items among the required diagnostic items. In such a definition, the value of *n* is an explicit threshold. An implicit threshold lies in the use of "abnormalities" because a cut-off value is used to separate the normal from the abnormal. The DSM-IV tries to clearly define each subtype of PDDs as a distinct entity that can be separated from other subtypes of PDDs.

The categorical view is reflected in the diagnostic tools which implement the DSM-IV. For example, the accompanying algorithm of the ADI-R sets a cut-off threshold

specific for each area of PDDs. If an individual's score of an area exceeds the corresponding threshold, he is considered to be abnormal with respect to this area; otherwise, the score is considered to be normal. For an individual to be diagnosed with a specific subtype, for example Autism, he should show abnormality in all of the three cores areas, and the abnormality should begin before the age of 36 months.

Empirical findings suggest that the disorders described according to the categorical view may not represent distinct diagnostic entities [TD05, SOBD05, MCC01]. Moreover, using an arbitrary threshold to categorically distinguish normal from abnormal values may cause the loss of potentially important information [SOBD05]. In fact, the American Psychiatric Association, which publishes the DSM-IV, acknowledges that an "alternative perspective to the categorical approach is the dimensional perspective that personality disorders represent maladaptive variants of personality traits that merge imperceptibly into normality and into one another" [TD05]. Some empirical data analysis methods, such as cluster analysis, are based on a "dimensional perspective". In this thesis, we refer to this perspective as a *continuous view* of the subgroups, and introduce it in the following section.

2.1.4 Continuous View

According to the continuous view, the severity of symptoms has a continuous distribution in a population, and there is no attempt to enforce a cut-off threshold to distinguish between the normal and abnormal individuals [SOBD05].

Consequently, an individual suffering from PDDs can be viewed as demonstrating impairments of varying degrees. For example, the impairment in social interaction can

include seeking affection inappropriately, having problems with imitation and joint referencing, or complete withdrawal. We do not identify them as normal or abnormal by using a threshold [Myh98].

Cluster analysis [HMS01], which will be reviewed in Section 2.2, and *factor analysis*[3] are two of the approaches that are developed based on a continuous view of the data being analyzed, and are commonly used by researchers. Some researchers collected data on symptoms related to PDDs, and performed factor analysis to identify major factors underlying these symptoms [TRD98, WE05]; others performed cluster analysis to identify sub-groups of patients [EHE94, PELW98, VGRR06].

Aside from using different subtyping techniques, researchers also obtained data from multiple areas where irregularities related to PDDs can be identified. There are three main source areas as introduced in the next section.

2.1.5 Data Sources for the Subtyping of PDDs

Subtyping studies aim to provide more homogeneous groups of PDDs or to verify the subtypes already identified. These studies usually focus on one of the three following areas or their combination [BS01]:

(1) Social interaction/communication/behaviours;

(2) Intellectual or adaptive functioning abilities;

(3) Medical/biological conditions.

[3] Factor analysis is a statistical approach that can be used to analyze inter-relationships among a large number of variables, based on their common underlying dimensions (factors). The approach involves mapping the original variables into a smaller set of representative dimensions (factors) with a minimum loss of information [Hai92].

Social interaction/communication/behaviour

This is the area that is currently used to characterize and diagnose PDDs, as discussed previously in Section 2.1. In 1979, Wing and Gould [WG79] identified three subtypes of PDDs in this area. These subtypes were defined by specific patterns of symptoms which the investigators called: *aloof, passive,* or *active-but-odd*. The aloof subtype characterized the highest proportion of children with Autism and a lower range of IQs. Children of the passive subtype did not initiate social interaction themselves, but could accept it when others initiate it. Children of the active-but-odd subtype sought interactions with others to serve their own narrow interests rather than regular social needs.

Wing and Gould broadened the concept of Autism to a spectrum of Autism-like disorders, which are characterized by impairments of different severity in the three core areas of Autism, and were later named *Autism Spectrum Disorders* (*ASDs*) or PDDs. They also suggested that there was no clear separation between Autism and other disorders on the autism spectrum, which was a fundamental change in the conceptualization of PDDs [WG79], and supported the continuous view in the subtyping of PDDs.

Wing and Gould's subtypes received extensive examination, and were generally supported. In most studies, two phenomena were observed. First, these subtypes did *not* map to the five diagnostic categories in the DSM-IV. Second, the existence of the aloof and active-but-odd groups was more evident, as found by later studies, than that of the passive group [BO94, BS01].

IQ and adaptive functioning

Pre-treatment IQ scores and language ability prior to the age of 5 or 6 were identified by some researchers to be effective predictive variables related to the outcome of PDDs [GS97, Rog98]. The strong relationship between subtype assignment and the level of intellectual functioning was also supported by other studies [WG79, VCBH89]. Therefore, although IQ data are not part of the data we analyze, we may include them in future work to verify the subtypes obtained in this study.

Medical/biological conditions

In one of their reviews, Rutter *et al.* [RBBL94] concluded that among the patients with PDDs, some medical conditions such as rubella encephalitis and fragile X syndrome which affected the central nervous system were associated with PDDs, and were suspected to have aetiological relationship with PDDs [RMFP90]. The rate of such conditions was about 10%, and might be even higher in the cases of PDDs associated with profound mental retardation, as well as in the cases of PDD-NOS.

Reviews of sibling, twin and family studies all concluded that a genetic aetiology existed for many cases of PDDs. However, since the identity and number of genes involved in PDDs was unknown, there were no conclusions as to whether the presence of specific genetic abnormalities was associated with particular subtypes of PDDs [BS01].

2.1.6 Current Status of the Subtyping of PDDs

A number of different subtyping approaches have led to several consistent findings:

(1) Developmental level (measured using the IQ score) and the three core areas of PDDs, including communication, social interaction, and behaviour patterns, are

the most important factors in explaining the variation in the manifestation of PDDs [GS97, VCBH89].

(2) The number of subtypes tends to be similar across studies, usually three subtypes when the three core areas of PDDs are measured and analyzed. When the IQ is also taken into consideration, the number of subtypes identified tends to increase to four [BS01].

(3) Most research supports the continuous view that different subtypes of PDDs fall along a continuum of severity ranging from almost normal to severely impaired rather than having distinct symptom profiles [PELW98].

As mentioned in Section 2.1.4, some researchers apply techniques such as cluster analysis to data collected from patients with PDDs in order to identify the subtypes of PDDs following the continuous view. Cluster analysis is also one of the main techniques – the other two being cluster validation and consensus clustering – we employ in this thesis. Cluster analysis and validation as well as consensus clustering are discussed in the next sections.

2.2 Cluster Analysis

Clustering methods partition objects into groups so that the objects in one group are similar to each other, and as dissimilar as possible from the objects in other groups [HMS01]. These objects are usually represented by multi-dimensional variables, also known as features or attributes. The similarity, or dissimilarity, between two data objects is typically measured as the distance between the multi-dimensional feature vectors that represent the objects. In this thesis, the feature vectors to be clustered are also referred to

as data points or data items. The larger the distance between two data points, the less similar they are to each other. We focus on three categories of clustering methods which are summarized by Han and Kamberand [HK06a], and are employed in our study:

(1) Partitioning methods [HK06].

These methods partition data into *k* subsets, where each subset is a cluster. Each object must belong to exactly one cluster, and no cluster can be empty. There are various objective functions that are optimized under this type of methods, with the most common being the sum of squared distances between every data point and its cluster centroid[4]. Clusters found with this type of methods are spherical. A typical example for such a method is the *k*-means clustering [KR90].

(2) Hierarchical methods [Joh67].

There are two main types of hierarchical methods: agglomerative or divisive. Agglomerative clustering starts with each data point forming its own singleton cluster, and iteratively merges pairs of clusters that are closest to one another, until one cluster is formed. The divisive approach goes in the opposite direction. That is, it starts with a single cluster with all the data points in it, and iteratively divides it into two clusters that are furthest apart from one another. There are several methods to calculate the distance between two clusters, such as *single, complete*, or *average linkage* [Joh67], *and Ward's method* [War63], which minimizes the increase in total within-cluster sum of squared distances when two

[4] Centroid is the centre, or representative, of a cluster. The definition of centroid may depend on the clustering algorithm being used. It is often the mean vector of the objects (feature vectors) in the cluster [HK06].

clusters are merged. Cluster structure found using hierarchical clustering can be represented graphically by a dendrogram.

(3) Model-based methods.

Model-based methods generate distribution models for data. There are *k* distributions in each model; each distribution represents one of the *k* clusters. The whole dataset is assumed to be a sample from the *k* distributions. Objective functions used for such methods typically favour models that are likely to generate the given dataset. A representative of these methods is *Expectation maximization for Gaussian mixture model* (*EM for GMM*, or *EM* for brevity) [DLR77].

K-means, agglomerative hierarchical (*hierarchical* for short in this study), and EM clustering will be introduced in detail in Chapter 4. Aside from these three categories of clustering methods, other methods have also been discussed in the literature [HK06a].

2.3 Cluster Validation

The process of evaluating the results of cluster analysis in a quantitative and objective way is called cluster validation [JD88]. It has four main components [TSK05]:

(1) Determine whether there is non-random structure in the data;

(2) Determine the number of clusters;

(3) Evaluate how well a clustering solution fits the given data when the data is the only information available;

(4) Evaluate how well a clustering solution agrees with partitions obtained based on other data sources.

Among these, Component (3) is known as *internal validation* while Component (4) is referred to as *external validation*.

Component (1) is fundamental for cluster analysis because almost every clustering algorithm will find clusters in a dataset, even if there is no cluster structure in it. However, this component is not the focus of this study because earlier research on the subtyping of PDDs has shown that there are subgroups of symptom profiles which correspond to the subtypes of PDDs [BS01].

We can use internal and/or external validation to determine the number of clusters in a dataset. In the context of this study, we do not have a ground-truth partition of the data to which we can compare our solution. Therefore, the number of clusters in our data is determined using internal validation only. We hence focus on introduction to internal validation methods first, and briefly discuss external validation in Section 2.3.2. We demonstrate the use of internal validation to determine the number of clusters in our data in Chapter 4.

2.3.1 Internal Validation

Two main measures are used to evaluate clustering solutions internally: *fitness* and s*tability*, both of which are employed in this study.

Evaluating the fitness of a clustering solution

Fitness refers to the quality of a clustering solution, usually evaluated by indices that are based on geometrical properties of clusters such as compactness, separation, and connectedness, because these criteria are the ones being optimized by most clustering methods [HK06b, EKX96, KR90, Joh67, DLR77].

Using hierarchical clustering, Milligan and Cooper offered the most comprehensive comparative study of 30 validation indices [MC85]. They found six indices that were shown to be better than the rest as the measure of fitness, but also pointed out that since their study was performed on a dataset generated in a specific way, one would *not* expect to find the same best indices if a different data generating strategy was adopted. Besides, this early study did not include indices devised and popularized in the last 20 years, which means the indices we use in our study were not explored in that survey.

Aside from the 30 validation indices mentioned above, other fitness indices exist for the purpose of estimating the number of clusters, including *Average Silhouette Width*, and *Bayesian information criterion* (BIC) [KR90, Raf00b], both of which are employed in this study, and are discussed in detail in Chapter 4. The average silhouette width evaluates the quality of a clustering solution by considering both *compactness* (distance between data points within the same cluster) and *separation* (distance between data points in two neighbouring clusters) [KR90]. The BIC is derived from Bayes' theorem [Sti82], and is used to determine which probability-based mixture model is the most appropriate [Raf00b].

Evaluating the stability of a clustering solution

The stability of a clustering solution, which usually refers to how robust a clustering solution is under perturbation or sub-sampling of the original data [BMC00, BEG02, LRB04, LD01, FD01], is another commonly-used validation criterion. A stable clustering solution is considered to have captured the underlying structure of a dataset, under the assumption that the clustering solution which captures the actual structure of the data source should be reproducible on other datasets drawn from the same source [LRB04].

In this study, one of the methods we use to estimate the number of clusters for our dataset is to assess the stability of clustering results. The assessment is based on the early work by Breckenridge [Bre89], and on the extensions by other researchers [TWBB01, LRB04, Wu04], all of which are introduced in this section. The procedure proposed by Breckenridge [Bre89] is called *Replication Analysis.* To perform replication analysis, a dataset is split into two equal subsets. As the core part of the analysis, the partition performed on one subset is used as the "ground truth" to group the items in the other subset via a machine learning technique called *Classification*. Therefore, we introduce classification first before we describe replication analysis in detail.

Classification is a supervised machine learning method. A classification algorithm is also called a classifier. In classification, each data item in the given dataset has a class label usually assigned to it by an expert or obtained from previous knowledge. The complete set is thus partitioned into several subgroups called classes. Typically the data are separated into two subsets called *training set* and *test set*. The training set provides the ground truth of the class membership of the data items in it. A model is built by learning a predictive pattern from the classified data in the training set, and is then used to predict class membership for the data items in the test set [HMS01]. The predicted membership of the test set is then compared to the membership that is assigned to it by an expert or obtained from previous knowledge to evaluate the performance of the classification. Classification is an important step in replication analysis, whose procedure is defined as follows:

1. Two disjoint subsets, A and B, are selected at random from a dataset D;
2. Subset A is grouped into *k* disjoint clusters, $<A_1, A_2 \ldots A_k>$, such that $A = \bigcup_{1 \le i \le k} A_i$; we denote this partition of subset A as Clu(A);
3. Subset B is also grouped into *k* disjoint clusters, $<B_1, B_2 \ldots B_k>$, such that $B = \bigcup_{1 \le i \le k} B_i$; We denote this partition of subset B as Clu(B);
4. A classification model is built to learn the class structure of subset A, assuming A is the training set, and Clu(A) is ground truth;
5. The data points in subset B are classified using the classification model learnt in Step 4. We denote the partition of subset B by Pred(B);
6. The degree of replication between A and B is measured by the agreement between the two partitions of subset B, Pred(B) and Clu(B).

In a simulation study, Breckenridge measured both the degree of replication and the actual recovery of the underlying cluster structure of the data he analyzed, and found that the two were positively correlated – the level of cluster replication from one part of the data to the other indicated the ability of a clustering method to recover the real cluster structure. Other researchers made use of Breckenridge's result to validate the stability of clustering solutions with varying number of clusters, and chose k_{opt} to be the *k* in the most stable solution [TWBB01, LRB04, Wu04]. In their studies of stability validation, the procedure described above for replication analysis was repeated *multiple* times for each number of clusters *k*. During each repetition, the original dataset was randomly split into two disjoint subsets A and B of equal size as mentioned in Step (1) of cluster replication. In Step (6), a pair of partitions Pred(B) and Clu(B) was produced before the agreement

between them was measured. That is, for each k, there were multiple pairs Pred(B) and Clu(B), and hence multiple evaluations of the agreement between them. The k that led to the best agreement between Pred(B) and Clu(B) is considered to be k_{opt} for the original dataset. Since there were multiple evaluations of the agreement for each k, a certain method, such as taking their average, was used in order to compare the agreement between Pred(B) and Clu(B) for every two k values. The k value that was associated with a higher agreement was considered to be more appropriate.

Previous researchers used different types of classification algorithms to produce Pred(B), and employed various measures to assess the agreement between Pred(B) and Clu(B) [Bre89, TWBB01, LRB04, Wu04]. These algorithms and measures are summarized in Table 2.1, and are introduced next.

In Step (4) of the procedure of replication analysis, a strong classification algorithm which has a small empirical classification error is needed, so that the agreement between Pred(B) and Clu(B) can be attributed to the intrinsic stability of the clustering solution, without considering the influence a poor classifier may have on the agreement. However, there is *no* known and agreed-upon optimal classification algorithm [Bre89, LR04].

Lange *et al.* suggested that an intuitive choice was a classifier that mimicked the clustering algorithm used to analyze subsets A and B. When no such choice was available, they suggested that the *k*-nearest neighbour classifier, which assigned a class label to a new data item by examining the k neighbours nearest to it (with known class labels), could be employed [LR04]. Following these guidelines, if the clustering method used was *k*-means clustering, Lange *et al.* chose to use the nearest centroid classification algorithm, which assigned each new data item to the class whose centroid was closest to this new

item. If other clustering methods were used, they chose the *k*-nearest neighbour algorithm. They pointed out that the nearest centroid classifier used by Tibshirani *et al.* [TWBB01] was only suitable for clustering methods which employed the concept of centroids, such as *k*-means clustering [LRB04].

Since the choice of the classifier in stability assessment is based on empirical considerations, we experiment with other types of strong classifiers such as *Random Forests* [Bre01], to classify subset B based on subset A and the cluster labels for subset A. The Random Forest classification method is discussed in detail in Chapter 4.

After using the classification algorithms mentioned above to produce Pred(B), previous researchers used a variety of measures to evaluate the agreement between Pred(B) and Clu(B) [TWBB01, LRB04, Wu04]. These measures are introduced next.

Aside from the classification algorithms, Table 2.1 also included two main categories of measures used in previous studies to evaluate the agreement between two partitions. The first category assumes that the clusters in Pred(B) can be mapped to those in Clu(B) in a one-to-one fashion. The agreement is then based on directly counting the total number of data items that are assigned to the corresponding clusters, one in Pred(B) and the other one in Clu(B). The other measure is based on viewing all the data points in pairs, and counting all the events when the two data points in each pair are put in the same cluster, in both Pred(B) and Clu(B).

Table 2.1: Four studies (column 1) using different classification algorithms (column 2) and measures of the agreement between Pred(B) and Clu(B) (column 3).

Name of study	**Classification algorithm**	**Measure of the agreement between Pred(B) and Clu(B)**
Replication analysis (Breckenridge)	*K*-nearest neighbour	Cluster mapping, Sum of overlapping objects
Prediction strength (Tibshirani *et al.*)	Nearest centroid	Co-membership
Stability-based validation (Lange & Roth)	Nearest centroid, *K*-nearest neighbour	Cluster mapping, Normalized sum of non-overlapping objects (disagreement)
Bootstrapping-based validation (Wu)	Study specific	Adjusted Rand Index

The measure used by Breckenridge and by Lange *et al.* belongs to the first category, which does not lie in the core of this study. The second category of measures was employed by Tibshirani *et al.* who performed cluster validation using *prediction strength*, and Wu who carried out bootstrapping-based validation (shown in Table 2.1). To be specific, Tibshirani *et al.* evaluated the agreement between Clu(B) and Pred(B) by examining *co-membership* of the data points [TWBB01], which was the relationship of objects falling into the same subset in a grouping process. The more data pairs assigned to the same cluster in both Pred(B) and Clu(B), the higher the agreement was between the two partitions.

Wu *et al.* employed the average of the Adjusted Rand Index (ARI) to measure the agreement between Clu(B) and Pred(B) [Wu04]. The ARI was calculated by counting the number of events in which pairs of objects belonged to the *same* or to *different* clusters in both Clu(B) and Pred(B), and correcting the count for chance agreement [AH85]. In Wu's study, the representative agreement between Pred(B) and Clu(B) for each *k* was the mean of the multiple evaluations.

The ARI was shown to be a robust index, and has been widely used [AH86]. While this is an interesting and succinct approach, we argue that taking the mean of multiple values of the ARI implies that the distribution of these values is unimodal, which is not necessarily true in all cases. We use the ARI as the measure of agreement between Clu(B) and Pred(B) in our study, but replace the calculation of the average ARI with a statistical significance test, as discussed in Chapter 4.

So far we have introduced the measures for fitness and stability validation, both of which are internal validation methods for cluster analysis, and are the focus of our study. We next briefly discuss external validation, as it is another important part of cluster validation.

2.3.2 External Validation

External validation is used to compare a clustering result to another set of membership labels (for the same objects) derived from a different data source. It can share measures used in stability validation as describe above, since both external validation and stability validation are concerned with measuring the agreement between two partitions. Other stability-validation measures that were not discussed above, such as contingency tables [EM97] and entropy-related indices [TSK05] can be employed for external validation as well.

Other than cluster analysis and cluster validation, consensus clustering is the third major component of the methodology we use in this study. The next section surveys the application of consensus clustering for combining multiple clustering results.

2.4 Consensus Clustering and Cluster Ensemble

Consensus clustering combines multiple individual clustering results into a single consensus solution to improve the accuracy and stability of clustering.

In practice, we may use a variety of clustering algorithms to partition a dataset into several clusters. Each of these clustering algorithms has its own clustering criteria, and imposes partitions on the data based on certain assumptions. Due to the lack of prior information about the underlying cluster structure, which is inherent to cluster analysis, we usually do not know which algorithm to choose in order to correctly identify this structure. Researchers have thus attempted to avoid selecting one particular criterion/algorithm by using instead a set of clustering solutions produced by different algorithms, called a *cluster ensemble*, and then incorporate them into a single partition referred to as the *consensus* solution [LOS04, SG02].

In general, the individual clustering solutions comprising the cluster ensemble can come from many different sources such as multiple clustering algorithms, multiple runs with random initializations of one clustering algorithm, subsets re-sampled from a dataset, various feature sets, etc. [LOS04]. A cluster ensemble improves clustering performance, as it can compensate for possible errors made by some clustering solutions by introducing the correct output of others; hence it can be more accurate and robust than each of the individual components [SSC98, SG02].

To arrive at a unique clustering solution based on a cluster ensemble with p best solutions obtained from cluster analysis and validation, each data point is represented as a p-dimensional vector, where the i^{th} position in the vector is the cluster label assigned to the data point by the i^{th} clustering solution (where $1 \leq i \leq p$). Consequently, the cluster

ensemble can be viewed as a *p*-dimensional categorical dataset, called the *ensemble dataset*, with the same number of objects as in each of the component clustering solutions. As an example, the ensemble dataset in Table 2.2 represents a cluster ensemble that contains six data points (rows) and four component clustering solutions (columns).

Minaei-Bidgoli *et al.* [MTP04] summarized five main categories of *consensus functions* that were used to map multiple individual clustering solutions to a consensus solution. We introduce three of them in this thesis, based on co-association, voting, and categorical clustering. The remaining two categories are based on information theory and on graph theory, and are not discussed here.

Table 2.2: An ensemble dataset for a cluster ensemble with four component clustering solutions and six data points.

	Solution 1	Solution 2	Solution 3	Solution 4
x_1	2	2	1	1
x_2	1	2	1	2
x_3	1	2	2	2
x_4	2	3	2	1
x_5	1	3	3	2
x_6	3	1	3	1

In *co-association-based* consensus clustering [FJ02], the co-association of each pair of data points is simply the count of events when the two points are assigned to the same cluster among all clustering solutions. An n x n co-association matrix can thus be constructed from the ensemble dataset with n data points and p clustering solutions to represent the similarity between each pair of data points in the ensemble dataset. The final clusters can be obtained by applying a similarity-based algorithm, such as the

commonly used hierarchical clustering (which is employed in this study), to the vectors in the co-association matrix.

A *voting* approach requires that the p clustering solutions in the ensemble have the same number of clusters, so that the clusters in each solution can be mapped to those in other solutions in a one-to-one fashion. To predict the cluster label of a data point using the voting mechanism, the cluster that is assigned to this point by the largest number of solutions is taken to be the cluster to which the point belongs [TLJF04].

In the *categorical* clustering approach, the combination of multiple partitions is viewed as a clustering task in and of itself [TJP03]. That is, each data point is represented by a categorical vector composed of cluster labels as shown in Table 2.2. These objects are clustered again using clustering methods such as k-modes [Hua98]; the clustering result then becomes the consensus solution.

Each consensus approach explores the structure of a cluster ensemble in its own way. Similar to choosing a clustering algorithm, choosing the best consensus approach greatly depends on the underlying structure of the ensemble dataset, which is not known. A reasonable approach is to use multiple consensus approaches to obtain several consensus solutions, and combine these solutions yet again [MTP04]. However, this is out of the scope of this study, and we do not pursue it here. In this study, we employ the co-association-based approach to build a similarity matrix from the ensemble dataset, and use hierarchical clustering to cluster the objects in the ensemble, as discussed in Chapter 4.

Among the three core parts of our methodology, that is, cluster analysis, cluster validation, and consensus clustering, cluster analysis has been applied by several groups

to identify the subtypes of PDDs [PBGP75, Res88, EHE94, PELW98, VGRR06], while cluster validation has been applied to a limited extent [SACB86, EHE94, WMAD96, SFDW00]. To the best of our knowledge, consensus clustering has not been previously performed in this area. In the following section, we review some of the previous studies that are most closely related to ours.

2.5 Previous Studies of the Subtyping of PDDs Using Cluster Analysis

In the studies of the subtyping of PDDs using cluster analysis, typically datasets contain n subjects (i.e. patients), each of which is represented by a p-dimensional feature vector. These feature values represent the characteristics of the patients in one or more of the three areas introduced in Section 2.1.5, namely social interaction/communication/behaviours, intellectual or adaptive functioning abilities, and medical/biological conditions.

We surveyed previous studies on the subtyping of PDDs using cluster analysis along some important factors such as the size of samples, the condition of subjects, the power of data-collection tools, and the cluster analysis and validation method used for subtyping. As no previous cluster analysis study for the subtyping of PDDs used consensus clustering, this topic is not surveyed here.

Regarding the sample size, most previous studies included between 100-200 subjects [BS01], but some included fewer than 50 patients [SACB86, SMCL95, SOHH99, PBM04]. Such small samples are unlikely to faithfully reflect the actual subgroup structure of the population that suffer from PDDs.

Regarding the condition of subjects for the subtyping of PDDs, it is clearly desirable to include patients with various forms of PDDs, and exclude subjects with conditions that are not PDDs, as the purpose of this study, and of several others, is to identify subtypes within PDDs. However, some of the previous studies included subjects with other conditions such as mental retardation, psychosis, schizophrenia, and language disorders [PBGP75, Res88, PBM04]. In such cases, refined subtypes within PDDs are difficult to identify due to the heterogeneity of the data.

Some data-collection instruments under different categories of assessment tools are reviewed in Section 2.1.2. Many assessment tools have been used, some of which were validated by more studies than others for the purpose of collecting data related to PDDs. The latter include the Wing Subgroups Questionnaire [CD93], the ADI-R [LRL94], and the ADOS [LRGH89]. However, many of the previous studies based their analysis on data collected using tools that were not specifically designed for identifying subtypes of PDDs [PBGP75, SACB86, PBM04, Res88], or using other non-standardized information sources such as hospital records [EHE94]. Since the subtypes identified by cluster analysis depend primarily on the initial raw data, non-standardized, non-specific, or inaccurate raw data cannot generate valid subtypes.

All previous studies of the subtyping of PDDs used a single clustering method, such as *k*-means [Res88, EHE94, PBM04], hierarchical clustering [SACB86, SFDW00, VGRR06], or model-based clustering [PBGP75, PELW98], without an explanation as to why the specific method was used. When choosing a clustering solution, many of them did not employ validation methods that were well-justified [SACB86, WMAD96, SFDW00], or did not use validation methods at all [Res88, PBM04].

Compared to previous studies (some of which we introduced above), our study is, to the best of our knowledge, the largest study done so far on subtyping PDDs based on data collected with the well-validated ADI-R. In the next chapter, we discuss the pre-processing of the raw data. Through the pre-processing, we obtain the dataset to which we apply the cluster analysis.

Chapter 3

Data Pre-processing[1]

The raw data for this study were obtained from 394 patients (referred to as *subjects*) with pervasive developmental disorders, collected by the *Autism Genetic Resource Exchange* project. To apply cluster analysis to the data, we pre-process and represent each subject as a feature vector. In this chapter we extend the discussion about the Autism Diagnostic Interview – Revised (ADI-R), introduced in Chapter 2, to provide information necessary for understanding the data pre-processing, and describe how we construct the dataset used for cluster analysis from the raw data collected using the ADI-R.

For 36 out of the 394 patients, there are certain questions that were not answered. The number of such questions ranges from 26 to 72. These patients are excluded from our dataset because of the high percentage of missing answers, leaving 358 subjects, each with a complete set of answers. The age of the 358 subjects ranges from 2 to 21 (mean = 6.9, std = 3.5, distribution shown in Figure 3.1); male to female ratio is 6:1[2]. Representing the subjects as feature vectors requires an in-depth understanding of the interview and its accompanying algorithm, which are discussed in the following section.

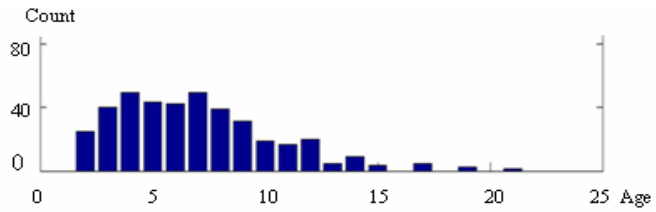

Figure 3.1: Age distribution of the 358 subjects.

[1] Special thanks to Drs. Jeanette Holden, Heidi Penning, and Ira Cohen, and Ms. Phil Hyoun Lee for helping me understand and pre-process the data.

[2] This male to female ratio is higher than seen in the PDD population. This higher ratio is an artifact of the preference for families with boy siblings in the original study that gave rise to this dataset.

3.1 Introduction to the Autism Diagnostic Interview – Revised

The Autism Diagnostic Interview – Revised (ADI-R) is an interview that consists of 3 parts. Part (1) includes 28 opening questions concerned with the early development of a child such as the starting time of the manifestation of PDDs. Part (2) forms the main component of the interview, containing 51 questions about the three core areas of PDDs, namely communication, social interaction, and behaviour patterns. Part (3) consists of 14 questions about general behaviours that do not belong to the core areas of PDDs, such as memory skills, motor skills, and fainting. The interview is administered by a person who is trained to administer the talks with a parent or another caregiver who is familiar with both the developmental history and the current behaviour of the subject.

The ADI-R is accompanied by three versions of a validated algorithm that were developed solely for the diagnosis of Autism, as follows:

(1) A lifetime version for assessing a patient's entire developmental history;

(2) A current version for evaluating a patient's current behaviour;

(3) A version specifically designed for patients younger than 4 years old.

The presence (or absence) of Autism is diagnosed using the version appropriate for the patient. All three versions of algorithm are structured the same: each of the three core areas of PDDs is further divided into four sub-areas, each of which is represented by 2 to 5 questions in the ADI-R [LLR03]. The structure is demonstrated in Figure 3.2 using the area of social interaction in the algorithm version for children younger than 4 years old. The answers provided by the caregiver are scored by the interviewer according to criteria shown in Table 3.1 [LRL94]. A summary score is calculated for each sub-area by summing up all corresponding answers to the questions belonging to this sub-area.

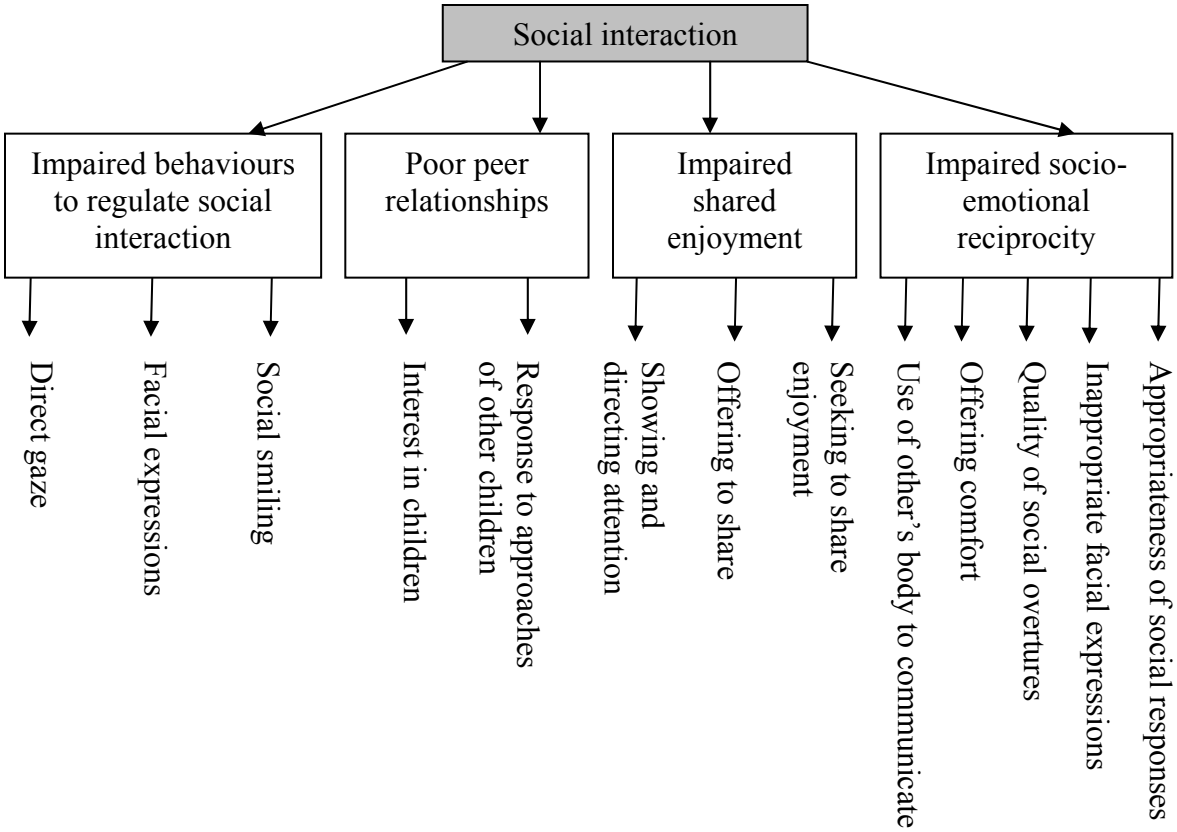

Figure 3.2 Structure of the area of social interaction in the accompanying ADI-R algorithm for children younger than 4 years old. Grey rectangle: core area of social interaction. Clear rectangle: sub-area. Specific questions are listed under each sub-area.

Table 3.1: Scoring criteria for answers provided by the caregiver to questions in the ADI-R. Left column: answer provided by the caregiver. Right column: score given by the interviewer.

Answer	Score
No definite behaviour of the type specified.	0
Behaviour of the type specified probably present but defining criteria not fully met.	1
Definite or extreme abnormal behaviour of the type described.	2

Similarly, a summary score is computed for each area based on the scores of the respective sub-areas. A child is diagnosed as having Autism if each of her scores for the three core areas reaches some pre-defined threshold, and her early manifestation of PDDs starts before the age of 36 months [LRL94]. Notably, only questions in Parts (1) and (2) of the ADI-R, but not those in Part (3) which ask about a patient's general behaviour, are included in the accompanying algorithm for diagnosing Autism.

It is important to note that some questions in the ADI-R consist of two alternative sub-questions. Specifically, if a question is about the *presence* of an *abnormal* behaviour, it contains two sub-questions that ask about the *current* condition and about the condition when the behaviour persisted in the history of a subject for at least 3 months. If a question is about the *absence* of a *normal* behaviour, it contains two sub-questions that ask about the *current* condition and about the condition that existed when the subject was between the ages of 4 and 5 years, which is the period when the behaviour of an individual with PDDs is likely to be *the most abnormal* [RLL03]. Depending on which version of the accompanying ADI-R algorithms is used by the interviewer, the answer to one of the two alternative sub-questions is taken to be the answer to the question. Figure 3.3 demonstrates the two types of questions (ellipses) in the ADI-R: some of the questions have two sub-questions (circles); the rest of the questions do not contain sub-questions. The answers to the questions or sub-questions represented by darkened ellipses and circles are used to construct the dataset for the subtyping of PDDs. Note that only one of the sub-questions for each question is selected.

In this study, we construct the dataset for cluster analysis using the raw ADI-R data gathered from 358 subjects following a scheme that is extended from the accompanying

ADI-R algorithm lifetime version, with reference to the version that is designed for children younger than 4 years old, both proposed by Lord *et al.* [LRL94]. In the next section we discuss the choice of sub-questions used in this study in order to obtain the data for our cluster analysis.

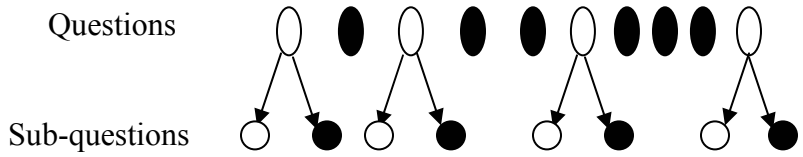

Figure 3.3 Structure of the questions in the ADI-R. Ellipse: question. Circle: sub-question. The filled items represent the questions or sub-questions whose answers are used by a specific version of the ADI-R algorithm.

3.2 Selection of Sub-questions

Based on the lifetime version of the accompanying ADI-R algorithm, if a question is about the *absence* of a certain normal behaviour, the score that denotes the condition of the child when she is 4 to 5 years old is used. For children that are younger than 4 years old, according to the version of the algorithm for this age group, the score that denotes the *current* condition is used. If a question is about the *presence* of a certain abnormal behaviour, the score that denotes the condition of the child when he initially exhibited this behaviour is used. Two exceptions are the questions about "reciprocal conversation" and "social chat", whose scores for the *current* condition are always used regardless of the version of the algorithm, because they have been proven to have high distinguishing power [LLR94]. For questions that do not contain two sub-questions, the score for the answer is directly used.

It has been proposed [YSF94] that relying on the caregiver's retrospection for answering questions about a child's earlier conduct may lead to errors. For example, parents may tend to recall fewer symptoms if their child is currently functioning well, and more symptoms otherwise, thus biasing the answer. A study included an external validation to justify the use of retrospective data, by assessing the agreement between the parental retrospective reports and independent ratings based on previous medical charts [YSF94]. The two correlated at $r = 0.72$ ($p < 0.05$), which suggested that the ratings based on the parents' recollection of their children's symptoms were positively correlated with those objectively measured, and could thus be used to assess past behaviours [YSF94]. Indeed, many other studies included parental recollection as a data source and showed meaningful and justifiable results [BP95, GOF03, GMPC98, CTDC05].

After the selection of sub-questions, our dataset contains 93 variables, each of which represents the severity of an impairment obtained as an answer to a question or to a sub-question in the ADI-R. We then perform *feature selection* and *feature extraction* to reduce the dimensionality of the dataset, as discussed next.

3.3 Dimensionality Reduction

There are many benefits for dimensionality reduction including easier data visualization, reduced storage requirements, increased computing speed and improved classification/clustering performance [GE03].

To perform *feature selection*, domain knowledge can be used to determine the subset of attributes that are relevant to the data analysis task. Such knowledge can also be employed to combine multiple attributes into new features to accomplish *feature*

extraction [WH03]. Both feature selection and feature extraction are practiced in the diagnosis of Autism using the accompanying ADI-R algorithm [LRL94, RLL03]. For example, the algorithm only uses the scores of a portion of the 93 questions in the ADI-R. This is a form of feature selection. Specifically, depending on its version, the algorithm chooses 39 to 42 questions from Parts (1) and (2) of the ADI-R. As shown in Figure 3.2, the scores of the answers to the chosen questions are combined first to obtain scores for the sub-areas, and then for the areas to which the sub-areas belong. That is, feature extraction is performed on these questions to generate features to represent the sub-areas and the areas.

With the guidance of an expert on the diagnostic procedures for PDDs [HP05], we use much more of the data than is typically used by any version of the accompanying ADI-R algorithm. Particularly, we use 64 of the 93 questions in the ADI-R, and combine them into 22 features for cluster analysis (Appendix B). These features are conceptually similar to the sub-areas in the accompanying ADI-R algorithm, and are therefore named according to the sub-areas to which they correspond. Since we use more questions than the accompanying ADI-R algorithm, some of the 22 features do not have corresponding sub-areas. We refer to these features as *add-on features*, and name them according to the impairments described by the questions pertaining to them. The 22 features are:

(1) Early onset of symptoms;

(2) Impaired early development;

(3) Delayed acquisition age of language;

(4) Abnormal conversational interchange;

(5) Stereotyped speech;

(6) Impaired receptive communication;

(7) Impaired gesture expressive communication;

(8) Impaired behaviours to regulate social interaction;

(9) Poor peer relationships;

(10) Impaired shared enjoyment;

(11) Impaired socio-emotional reciprocity;

(12) Impaired social development;

(13) Lack of initiation of activities;

(14) Encompassing preoccupation;

(15) Stereotyped motor mannerisms;

(16) Ritualistic behaviour;

(17) Sensory issues;

(18) Adherence to routine;

(19) Symptoms of Rett's disorder;

(20) Aggression;

(21) Epilepsy;

(22) Demonstrated savant skills.

As shown above, we include the add-on features such as those pertaining to *aggression* (*Feature 20*) and *demonstrated savant skills* (*Feature 22*). We also include a few non-redundant questions that are not used by the accompanying ADI-R algorithm in several sub-areas, such as questions pertaining to *attention to voice* (Question 46) and *unusual attachment to objects* (Question 76). There are four main reasons for including more questions and features in this study than those included in the accompanying ADI-R

algorithm. First, it has been pointed out that certain behaviours, such as hyperactivity, aggressiveness, and destructiveness, are currently not included in the diagnosis of Autism, while they are likely to be relevant [FMBB03]. Second, some researchers observed that symptoms such as isolated (savant) skills and epilepsy are related to certain subtypes or severity levels of PDDs [OWL05, GPA05, NDTH03]. Third, the questions about mid-line hand movement[3] and hyperventilation are included because they are symptomatic of Rett's disorder. Fourth, some questions about the impairment of early development, such as delayed walking, are also included because early development is related to certain subtypes of PDDs such as Autism and Childhood disintegrative disorder [LRL94].

Questions about age-specific manifestation are not used to form the feature set in our study because these questions only apply to patients of certain ages. We also do not add on the features that are about the regression of symptoms in patients. In the ADI-R, the questions about relapse aim to provide more accurate information for differential diagnosis between Autism, Rett's disorder, and Childhood disintegrative disorder [LRL94]. Meanwhile, Feature 1, which asks about the "early onset of symptoms" in a patient, and Feature 19, which asks about some unique symptoms associated with Rett's disorder, are already included to differentiate these three subtypes of PDDs. Besides, the regression is not unique to patients with Rett's disorder, nor is it unique to those with Childhood disintegrative disorder. It is also common among patients with Autism with an occurrence rate ranging from 20% to 50% [FC01].

Similar to the way the score of a sub-area is calculated in the accompanying ADI-R algorithm, the value of a feature in our study is obtained by summing the scores for the

[3] A particular way of moving one's hands in front of one's body. For example, hand wringing or turning hands from side to side together as if washing them.

questions pertaining to the feature. The scores for all features are then normalized to the range [0, 1], by mapping the maximum score to 1, the minimum to 0, and the intermediate values linearly to the (0, 1) interval.

After applying the data pre-processing steps described above to the raw ADI-R data, we obtain a dataset corresponding to the 358 subjects, each of which is represented by a 22-dimensional feature vector, where each feature is a real number in the range [0, 1]. Cluster analysis, cluster validation and consensus clustering can then be performed, as shown in the next chapter, to identify the subtypes of PDDs based on the dataset.

Chapter 4

Methods

The pre-processing stage, described in Chapter 3, results in a set of feature vectors to which we apply the actual analysis. Three clustering methods, representative of three categories of clustering algorithms, are applied to partition this dataset into subsets as shown in Section 4.2. Two validation methods introduced in Sections 4.3 and 4.4 are used to evaluate the clustering results. The results obtained by using different combinations of clustering and validation methods are then integrated into a unified consensus solution, as shown in Section 4.5, in order to reach an agreement on the assignments of subjects into clusters. In Section 4.6, the clusters in the consensus solution are identified, analyzed and compared to clinical diagnoses previously obtained for the subjects. The complete pipeline diagram is included in Appendix C.

4.1 Overview of Methods

As described in Chapter 3, we pre-process the original data to obtain a set representing the 358 subjects using 22 numerical features (ranging in value from 0 to 1). Each feature value represents the severity of impairment in one of the 22 areas listed in Section 3.3 for the patients suffering from PDDs. The three clustering methods we apply to the set are *k*-means, agglomerative hierarchical (hierarchical for short), and EM for Gaussian mixture model (EM for short), which represent three categories of algorithms: partitioning, hierarchical and model-based, respectively. One of the core issues in our study is to find the appropriate number of clusters, k_{opt}, through cluster validation. This is achieved by

checking the fitness and the stability of the partitions produced by each of the three clustering methods. The combination of the three clustering methods and two validation criteria produces six *best* partitions of the dataset, each of which has the number of clusters that is considered to be the most appropriate for the specific combination of clustering method and validation criterion. The other core issue in our study is to *consolidate* these best results. To achieve this, the six best partitions are treated as the components of a cluster ensemble, and are integrated to form one consensus solution.

In the following sections, we introduce the criteria and procedures that we use for clustering, validation and result integration.

4.2 Clustering Methods

In this section, we provide the details of the *k*-means, the hierarchical, and the EM clustering algorithms which have been briefly presented in Section 2.2. We use the implementation of the *k*-means [Tek06] and of the hierarchical clustering methods provided in *Matlab* [MW05], and the implementation of the EM method provided in the *Weka* machine learning software package [WF05]. Throughout the description, we frequently use the term *data point* to refer to a feature vector in an *n*-dimensional space.

4.2.1 *K*-means Clustering

In the term *k*-means clustering, *k* represents the number of clusters. The value of *k* is usually unknown *a priori* and has to be chosen by the user. Each cluster has a centroid, which is usually computed as the mean of the feature vectors in the cluster. The cluster

membership of each data point under the *k*-means clustering algorithm is decided based on the cluster-centroid nearest to the point. As centroids cannot be directly calculated until clusters are formed, the user specifies *k* initial values for the centroids at the beginning of the clustering process. The actual centroid values are calculated once clusters have been formed.

The *k*-means algorithm partitions a dataset into *k* clusters using the following steps [Mac67]:

(1) Initialize the cluster centroids with *k* initial values.

(2) For each data point in the dataset, find the closest centroid, and assign the point to the cluster associated with this centroid.

(3) Calculate the centroid for each of the *k* clusters based on the new cluster memberships.

(4) Iterate through steps (2) and (3) until a termination condition is met.

Several termination conditions can be used in step (4). Each condition compares the value(s) of a certain measure computed in the current iteration to the value(s) of the same measure computed in the previous iteration. Three commonly-used conditions are:

(i) The centroids do not change;

(ii) The sum of squared distances from each of the data points to their respective centroids does not change;

(iii) The cluster membership of the data points does not change.

We apply Condition (iii) as the termination condition for *k*-means clustering. When the memberships of the data points do not change, the centroids computed based on these

memberships do not change. Consequently, the sum of squared distances from the data points to their centroids does not change either.

In Step (2) of the clustering procedure, the *k*-means algorithm finds the *nearest* centroid for each point. The word "nearest" is meaningful only when there is a pre-defined distance metric. The metric we use is the *Euclidean distance*. Although many other metrics can be used, we adopt the Euclidean distance for both *k*-means and hierarchical clustering because it was used in several previous studies in which cluster analysis was employed to find subtypes in PDDs, and led to meaningful results [Les88, CT85].

The Euclidean distance between two data points X_1 and X_2 , each represented by a *p*-dimensional vector, $X_1 = (X_{1_1}, X_{1_2}, ..., X_{1_p})$ and $X_2 = (X_{2_1}, X_{2_2}, ..., X_{2_p})$, is denoted as $d\_Euc\ (X_1, X_2)$, and defined as follows:

$$d\_Euc\ (X_1, X_2) = \sqrt{\sum_{i=1}^{p} (X_{1_i} - X_{2_i})^2} \ . \qquad (4.1)$$

The *k*-means clustering procedure iteratively moves data points between clusters, minimizing the sum of squared distances, denoted by *J*, from each data point to its cluster centroid. Denote the i^{th} cluster by C_i, then the sum of squared distances for C_i, denoted by J_i, is defined as follows:

$$J_i = \sum_{X \in C_i} d\_Euc(X, Y_i)^2 \ ,$$

where $d\_Euc(X, Y_i)$ is the Euclidean distance from a data point X in C_i to C_i's centroid Y_i.

We can then calculate the sum of squared distances for all the k clusters, denoted by J, as:

$$J = \sum_{i=1}^{k} J_i \tag{4.2}$$

Starting the k-means algorithm from various sets of initial values may lead to varying local minima of J. We wish to find the one that is the global minimum. However, it is unrealistic to exhaust all the sets of initial values. Therefore, we run k-means clustering multiple times, starting from different initial values at each run and choose the solution that minimizes the sum of squared distances, J. By using multiple runs, the algorithm is more likely to converge to the global minimum of J, or at least to a local minimum that is the closest to the global minimum among the multiple local minima.

4.2.2 Hierarchical Clustering

To perform hierarchical clustering on a set of n data points, a symmetric n x n distance matrix consisting of pair-wise distances between the data points is generated. The main steps of hierarchical clustering are [Joh67]:

(1) Assign each data point to a separate cluster to obtain n clusters, each of which contains one data point (a singleton).

(2) Find the closest pair of clusters; merge the two clusters to form a new cluster.

(3) Compute distances between the new cluster formed in step (2) and each of the remaining clusters;

(4) Iterate through steps (2) and (3) until all of the n points in the dataset are merged into a single cluster.

In addition to a metric for measuring distance between individual data points, which is the Euclidean distance in this study, we also need a method to compute the distance between clusters in Step (2), so that the two most similar clusters can be merged. Such a method is referred to as *linkage* [Joh67].

The linkage method we employ here is called *average link* [SS73]. Let C_1 and C_2 be two clusters; denote the distance between any two data points X_1 and X_2 by $d(X_1, X_2)$; the distance between clusters C_1 and C_2, measured by the average link method, is defined as the average distance between all pairs of points from C_1 and C_2, where one point in the pair is in C_1 and the other is in C_2. The distance is denoted by $d_{avg}(C_1, C_2)$ and is calculated as:

$$d_{avg}(C_1, C_2) = \frac{\sum_{X_2 \in C_1, X_2 \in C_2} d(X_1, X_2)}{n_{C_1} * n_{C_2}}, \tag{4.3}$$

where n_{C_1} and n_{C_2} denote the number of points in clusters C_1 and C_2, respectively.

Agglomerative hierarchical clustering, as introduced in Section 2.2, merges all the data points in a dataset into sub-trees, and ultimately into one single tree. Figure 4.1 shows this process. There are 20 sub-trees (clusters, labelled from 1 to 20) on the *x*-axis. The *y*-axis represents the distance between sub-trees measured with the average linkage method. Two sub-trees that are closest to each other are then connected to form one larger sub-tree; this process is iterated until a single tree that includes all the data points in the dataset is formed. In general, when employing hierarchical clustering to build a single tree, the user does not need to specify the desired number of clusters, *k*. However, to obtain a clustering result with a specific number of clusters, the user does need to provide the value of *k*, so that the algorithm can cut the tree at a certain level where there

are exactly k sub-trees under this level. Figure 4.1 shows how to obtain three clusters, marked as clusters 1, 2, and 3, from the complete tree.

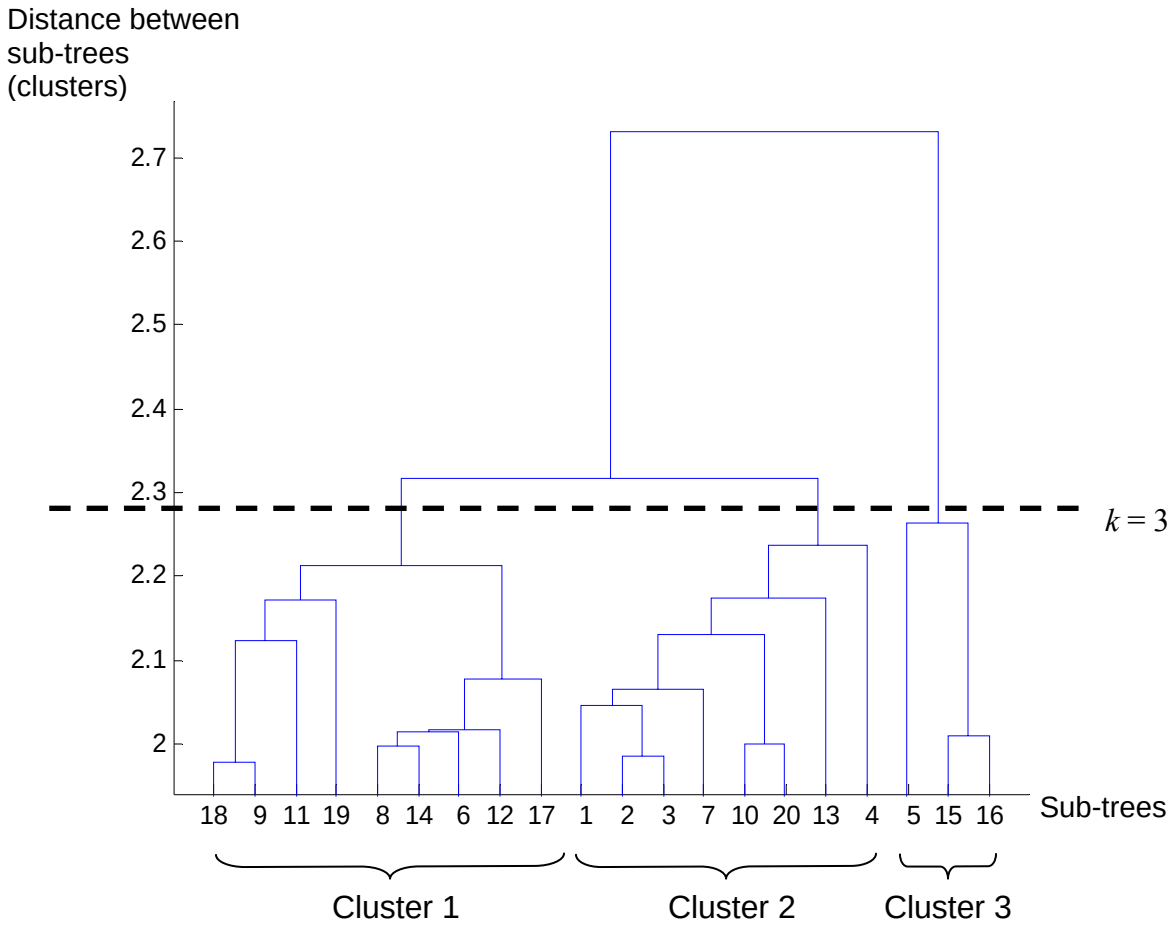

Figure 4.1: A hierarchical tree built from 20 sub-trees, then cut at a certain level (shown by a horizontal dashed line), resulting in three clusters marked as Clusters 1, 2 and 3 under the x-axis.

4.2.3 Expectation Maximization for Gaussian Mixture Models

Expectation maximization for Gaussian mixture models (referred to here as EM for short), is a probabilistic clustering algorithm that makes use of the finite Gaussian mixture model. There are k multivariate probability distributions in the model, where each distribution represents a cluster. As in k-means and in hierarchical clustering, k is specified by the user. In the implementation of the EM algorithm in the Weka package, it

is assumed that all features are independent random variables given the cluster. The main steps of the EM method are [DLR77]:

(1) Initialize the parameters (mean and variance) for the k Gaussian distributions.

(2) Compute the probability density for each feature vector under each of the k clusters, using the probability density function of the Gaussian distribution for each cluster.

(3) Re-estimate the parameters for each of the k Gaussian distributions using the probability densities computed in Step (2).

(4) Iterate through Steps (2) and (3) until convergence.

By updating the mean and variance of the Gaussian distribution for each cluster in step (3), the probability of the data given the parameters, also known as the *likelihood* of the parameters to generate the data, is maximized. The higher the likelihood is, the better the clustering model fits the data.

We now show how the likelihood (L) is calculated with respect to the parameters (θ) for two clusters C_1 and C_2, modeled by two Gaussian distributions. Let $\{X_1, X_2, ..., X_n\}$ be the n feature vectors (data points) in a dataset D; the probability density value of a feature vector X_i under the distribution model associated with cluster C_1 is denoted as $p(X_i \mid C_1)$, while the prior probability of C_1 is $p(C_1)$. The likelihood, denoted by L, is the probability density of the data, D, given the parameters, denoted by $p(D \mid \theta)$. The latter is computed as the joint probability density of the data points in D given the two Gaussian distributions that model the two clusters C_1 and C_2. Therefore, L is expressed as:

$$\mathrm{L} = p(D \mid \theta) = \prod_{i=1}^{n} p(X_i \mid \theta)$$

$$= \prod_{i=1}^{n} \left(\sum_{j \in \{1,2\}} p(X_i \mid C_j) * p(C_j) \right) \qquad (4.4)$$

Similar to *k*-means clustering, the EM algorithm may converge to the local maxima of the likelihood function L [DHS01]. The maximum to which the EM algorithm converges depends on the initial values assigned to the parameters in step (1) above. By using multiple sets of starting parameters and taking the solution that yields the highest L, EM clustering is more likely to converge to a global optimum or an optimum that is the closest to the global optimum among the multiple values of L we obtain.

As discussed previously in Sections 4.2.1 to 4.2.3, the three clustering algorithms we use all require the user to specify the number of clusters, *k*. For the *k*-means and the EM algorithms, *k* needs to be specified *before* the clustering process starts; while for hierarchical clustering, it can be specified after the process is finished. Although, as is the case for most clustering problems, we do not know the exact value for *k*, we do know the range of the values for *k* based on previous research on the subtypes of PDDs. In this study, we set the range of *k* to be from 3 to 7, based on the *a priori* knowledge obtained about the number of clusters from more than 15 studies on the subtyping of PDDs [BS01]. We then use the two validation methods, fitness validation and stability validation, which are briefly introduced in Section 2.3.1, to locate the most appropriate number of clusters, k_{opt}, for each of the three clustering algorithms. The two validation methods are discussed in further detail next.

4.3 Fitness Validation

In this section, we talk about the measures and procedures that we use in the fitness validation of the 3 clustering algorithms.

4.3.1 Fitness Measures

For clustering methods such as *k*-means and hierarchical clustering, in which each element is assigned to exactly one cluster, fitness is usually evaluated based on geometrical properties such as the compactness and separation of clusters. For probabilistic methods such as EM clustering, fitness can be evaluated using measures derived from probability [LD01]. When a partition produced by a clustering method results in an optimal value of the fitness measure for this method among multiple partitions examined, this partition is considered to be the best; k_{opt} is then set to be the number of clusters in this partition. We summarize the measures used for the fitness validation methods of the three clustering methods in Table 4.1. The range of values for these two measures is also listed in the table. For the *k*-means and hierarchical clustering methods, a fitness measure called the *average silhouette width* is used. For EM clustering, a measure called *Bayesian Information Criterion* (BIC), which is derived from the likelihood L, is used. Both of the measures are introduced in detail next.

Table 4.1: Measures for the fitness validation of the three clustering methods

Method	Measure	Value range
K-means	Average silhouette width	[-1, 1]
Hierarchical	Average silhouette width	[-1, 1]
EM	BIC	NA

Average Silhouette Width

Silhouette width measures how much more similar a data point is to the points in its own cluster than to points in a neighbour cluster. The *average silhouette width* is calculated as the average of the silhouette width for all the points in the data [KR90]. To formally

define the average silhouette width, we first formally define the terms *neighbour cluster* and *silhouette width.*

In a clustering solution, a neighbour cluster for a data point X in a cluster C, is the cluster whose data points have the shortest average distance to X among all the clusters other than C. Let $\{X_1, X_2, ..., X_n\}$ be the n data points in a dataset D, and $d(X_i, X_j)$ be the distance between two data points X_i and X_j. If $C(X_i)$ is the cluster to which X_i belongs, then the neighbour cluster for X_i, denoted by $C_{neig}(X_i)$, is defined as:

$$C_{neig}(X_i) = \widehat{C}_m = \underset{C_m \neq C(X_i)}{argmin} \left[\frac{\sum_{X_k \in C_m} d(X_i, X_k)}{|C_m|} \right], \tag{4.5}$$

where $\widehat{C}_m$ is the cluster whose points have the minimum average distance to X_i.

The silhouette width for X_i, denoted by S_i, is defined as:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{4.6}$$

where $a_i = \frac{\sum_{X_j \in C(X_i)} d(X_i, X_j)}{|C(X_j)|}$ and $b_i = \frac{\sum_{X_k \in C_{neig}(X_i)} d(X_i, X_k)}{|C_{neig}(X_i)|}$.

That is, a_i is the average distance between X_i and the points in the same cluster as X_i, and b_i is the average distance between X_i and the points in X_i's neighbour cluster. The value of S_i ranges from -1 to +1. When it is close to -1, X_i is likely to be assigned to the wrong cluster; when it is close to 0, X_i is equally likely to be assigned to any of the two clusters; when it is close to +1, X_i is considered to be clustered correctly.

The average silhouette width of a dataset is calculated as the mean of the silhouette widths over all data points in a set. Formally, it is defined as:

$$Average\ silhouette\ width = \frac{\sum_{i=1}^{n} S_i}{n}.$$

This measure evaluates the quality of the clustering by taking into consideration both the compactness and the separation of the clusters.

Bayesian Information Criterion

Given the observed data, the EM algorithm maximizes the likelihood of the parameters to generate the data for the k Gaussian distributions, where k is the number of clusters assumed for the data. When we use a large number of Gaussian distributions, they can then model the data very well and yield high likelihood. However, in most cases, the objective of cluster analysis is to describe the general population from which the data are sampled, rather than the specific dataset. The phenomenon of an over-complex model fitting the observed data too closely and not being able to generalize to unseen data is called *overfitting* [HMS01]. Therefore, we usually want to avoid over-complex clustering models..

The Bayesian information criterion (BIC) is devised to avoid overfitting, and is defined as [Raf86]:

$$\mathrm{BIC} = -2\ln(\mathsf{L}) + v\ln(n), \qquad (4.7)$$

where n is the number of data points, L is the likelihood of the parameters to generate the data in the model, as illustrated in Section 4.2.3, and v is the number of free parameters (means and standard deviations) in the Gaussian mixture model. The BIC takes into account both the fit of the model to the data and the complexity of the model.

A model that has a *smaller* BIC is preferred. In this study, we take the inverse of the BIC so that a model that has a *larger* 1/BIC is preferred, which is consistent with the criterion that a *larger* average silhouette width is preferred for the *k*-means and the hierarchical methods. Based on Eq. 4.7, the 1/BIC is expressed as:

$$1/\mathrm{BIC} = (-2\ln(\mathsf{L}) + v\ln(n))^{-1} \quad . \tag{4.8}$$

In the computation of the 1/BIC, $v\ln(n)$ is a penalty term because it penalizes a model for its complexity in order to avoid overfitting. While the likelihood term, ($-2\ln(\mathsf{L})$ in Formula 4.8), causes the 1/BIC to favour a more complex model because a more complex model can increase the likelihood L, the penalty term causes the 1/BIC to favour a simpler model because a smaller v decreases the penalty.

We use the above measures to introduce the procedures used for fitness validation.

4.3.2 Procedure for Fitness Validation

Two major differences exist between the procedures used for the fitness validation of *k*-means/EM clustering, and the validation procedure used for hierarchical clustering. First, for methods such as the *k*-means and the EM clustering methods, we need to specify the number of clusters, *k,* before starting the clustering process. In contrast, a hierarchical method can build a dendrogram without a pre-specified *k*. Second, the *k*-means and the EM algorithms should be started with multiple initial values in order to obtain a close-to-optimal solution.

Hierarchical clustering may also produce different solutions under certain conditions. However, the cause for the variability is not the same as the cause in the cases of the *k*-means and EM algorithms. In hierarchical clustering, multiple solutions can be generated

when the distance between two sub-trees[1], measured by the linkage method introduced in Section 4.2.2, is equal to the linkage distance between another pair of sub-trees, and thus creates a tie of distances. The existence of such ties for multiple pairs of sub-trees means that a decision of which pair of sub-trees to merge among equally good candidates should be made. If the decision is made arbitrarily, the clustering results may change accordingly. Different implementations of hierarchical clustering use different decision rules, which can lead to different solutions. For the implementation we use in Matlab, when ties exist, changing the order of the data points may change the memberships of the points involved in the ties due to the method of tie-handling in this implementation. This is not a desired phenomenon as the order of the subjects in our study, represented by the data points, should not affect their membership. Therefore, we use a number of random permutations of the input data to find whether they result in different clustering solutions [VSH05]. For each k (number of clusters), we performed 10 random permutations on the data points to verify that the permutations make no difference to results. Therefore, we do not use multiple runs of hierarchical clustering in this study.

For the k-means and the EM methods, pseudocode for the fitness validation procedure is shown in Table 4.2. The following notation is used:

(1) D: the dataset;

(2) k: the number of clusters;

(3) C: the clustering solution that has the smallest sum of squared distances or the largest likelihood for a fixed k;

[1] In this context, sub-trees also include those that have a single data point in them when agglomerative hierarchical clustering first starts.

(4) *optimum*: the smallest sum of squared distances for *k*-means clustering, or the largest likelihood of the parameters for EM clustering. It is initialized as *infinite* for *k*-means clustering, and *negative infinite* for EM clustering.

(5) *fitness*: the value of the fitness measure for the solution, *C*.

Table 4.2: Pseudocode for the fitness validation of the k-means and the EM clustering

```
for k = 3 to 7                                  % We assume 3 to 7 clusters in the dataset.
    Initialize optimum;
    % Use 20 sets of different initial values to cluster the dataset, D, using k-means/EM clustering.
    % For k-means, the k initial centroids are selected at random from D.
    % For EM, the initial means and variances are computed from clusters obtained by running k-
    % means on D.
    for i = 1 to 20
        temp1 = cluster(D, the i-th set of initial values); % temp1 stores the cluster membership of D
        % For k-means, temp2 is the sum of squared distances. For EM, it is the likelihood of the
        % data given the parameters
        temp2 = cost_fuction(D, temp1)
        % For k-means,  find_optimum is a min() function. For EM, it is a max() function.
        optimum = find_optimum(temp2, optimum);
        if (optimum is updated)
                C = temp1;
        endif;
    endfor i;
    % calculate_fitness calculates the average silhouette width for k-means and the 1/BIC for EM.
    fitness =  calculate_fitness(C);
    output fitness and C;
endfor k.
```

As shown in Table 4.2, for the *k*-means and the EM methods, *k* is specified before the clustering procedure starts. For each *k*, 20 different settings of the initial values are used. The clustering solution that minimizes the sum of squared distances (for *k*-means clustering) or maximizes the likelihood (for EM clustering) is taken to be the *k*-cluster solution, and a fitness measure value is calculated for it.

For hierarchical clustering, pseudocode for fitness validation is shown in Table 4.3. There is no need for the variables *i* and *optimum* since no multiple starting points are used in this procedure. A variable, *tree*, is used to represent the single hierarchical tree with all the data points in D merged into it. As shown in the pseudocode, a dendrogram is built based on the data, and the value of the average silhouette width is computed for the clustering solutions with 3 to 7 clusters.

Table 4.3: Pseudocode for the fitness validation of hierarchical clustering

```
% Build a hierarchical tree from dataset D using hierarchical clustering
tree = build_tree(D);
for k = 3 to 7
    % Cut the tree at a certain level to obtain k clusters.
    C = cut_tree(tree, k)
    % calculate_fitness calculates the average silhouette width for hierarchical clustering
    fitness = calculate_fitness(C);
    Output fitness and C;
endfor k.
```

Experimental conditions for the three clustering methods are summarized in Table 4.4. Both the *k*-means and the hierarchical methods use the Euclidean distance as the distance

metric, and both use 20 different sets of initial values. The linkage method used by hierarchical clustering is the average link defined in Section 4.2.2.

Table 4.4: Experimental conditions for the three clustering algorithms

	Distance metric	Multiple Starting points	Linkage method
K-means	Euclidean	20	NA
Hierarchical	Euclidean	NA	Average link
EM	NA	20	NA

As we have shown, to find the appropriate number of clusters, k_{opt}, we perform fitness validation on clustering solutions with different number of clusters. The number of clusters in the solution that has the best fitness score is taken to be the k_{opt}. The procedure for stability validation is presented in the next section.

4.4 Stability Validation

The procedure we use for stability validation extends the method called *replication analysis*, which was carried out by Breckenridge [Bre89] and has been introduced in detail in Section 2.3.1. Our procedure is also closely related to several other stability validation methods previously proposed by Lange *et al.* [LRB04], Tibshirani *et al.* [TW05], and Wu [Wu04], all of which are also based on cluster replication analysis and have been introduced in Section 2.3.1. We provide a formal description of our procedure next.

As before, $\{X_1, X_2, ..., X_n\}$ are the n feature vectors in the dataset D. Let k be the number of clusters. Denote the cluster labels of the n vectors by B, where $B = \{b_1,..b_n\}$. To be more specific, if the feature vector X_i is assigned to cluster C_j ($1 \leq j \leq k$) then b_i =

C_j. When a clustering method M assigns the set of labels B to the data set D, we denote it as $B = M(D)$. Pseudocode for the validation procedure is first listed in Table 4.5 and explained in more detail in the following two paragraphs.

Table 4.5: The pseudocode of the procedure for stability validation

```
for i = 1 to N
    for k = 3 to 7
        % Split D into two disjoint subsets D1 and D2, where |D1| = |D2|.
        {D1, D2} = split_in_half(D);
        B1 = M(D1); B2 = M(D2); % k clusters in both D1 and in D2
        Use (D1, B1) to train a classifier φ;
        B2' = φ(D2); % k classes in D2
        % agreement() evaluates the agreement on the cluster memberships of the
        % data points between B2 and B2'.
        Output agreement(B2, B2');
    endfor k;
endfor i.
```

The core of the pseudocode, namely, the inner loop, includes the main steps of replication analysis discussed in Section 2.3.1. To be specific, a dataset D is split into two equal subsets denoted by D_1 and D_2. D_1 and D_2 are partitioned into k clusters independently using a clustering method to obtain cluster labels B_1 and B_2, respectively. The cluster labels obtained for the D_1 set, B_1, are viewed as "ground truth", and supervised learning is used to train a classifier based on this clustering solution, where the cluster labels are viewed as the classes. In this work we use *Random Forests* [Bre01] as the classifier. The classifier is then used to classify the D_2 set and obtain class labels

B_2'. Agreement between the cluster labels assigned to the D_2 dataset by the unsupervised clustering algorithm, B_2, and by the Random Forest classifier, B_2', is calculated.

As shown in the pseudocode, the split of the dataset is repeated N times. That is, for each of the N splits of the dataset, we compute the agreement between the cluster labels (B_2) and class labels (B_2') under each of the five values of k (from 3 to 7). As a result, we have an N x 5 matrix in which the rows are the N data splits, the columns are the different k values from 3 to 7, and the elements are the agreement values calculated between B_2 and B_2'.

In this study, we measure the agreement value using the *adjusted Rand index* (ARI). The higher the agreement, the higher the ARI score. The number of clusters for which the ARI values are statistically significantly larger than those produced for any other number of clusters, checked by the Wilcoxon signed rank test [RDCT05], is taken to be the optimal number of clusters, k_{opt}. The best clustering solution with k_{opt} clusters is used as a component in the cluster ensemble in this study.

To explain the procedure for stability validation in detail, we introduce Random Forest and the ARI next.

4.4.1 Random Forest

Random forest is a classifier formed by an ensemble of *Decision Trees* [Kir02]. Each tree is constructed based on sub-samples randomly drawn from the training set with replacement. That is, a data item can be sampled multiple times. This process is known as *Bootstrap Aggregating*, which ensures that the sample is drawn independently from the

same distribution for building each of the trees in the forest. Bootstrap aggregating improves the accuracy and the stability of classification [Bre96].

To predict the class label of a data item, each tree in the random forest assigns a class label to the item, and the class assigned by the largest number of trees is taken to be the class to which the item belongs.

The performance of random forest classifiers was demonstrated to be at least as good as that of other classifiers such as decision tree, *Support Vector Machine* [SLTW04], and the highly accurate classification ensemble *Adaboost* [FS96]. The random forest classifiers are also more robust with respect to noise, and have only a small number of parameters to tune [Bre01]. Therefore, we use a random forest classifier as part of the stability validation procedure.

4.4.2 Adjusted Rand Index

As discussed in Section 2.3.1, there are multiple ways to calculate agreement between partitions. The ARI calculates it by computing the percentage of pairs of objects belonging to the *same* or to *different* subsets in each of the two partitions among all pairs of data points, and correcting for chance agreement. The ARI is defined as follows.

For a dataset D, let C and Q be two partitions: $C=\{C_1,..C_N\}$ and $Q=\{Q_1,..Q_M\}$, where

$$D = \bigcup_{i=1}^{N} C_i = \bigcup_{j=1}^{M} P_j \ .$$

Let $X_i \in D$ be a data point in D. We denote by $C(X_i)$ the subset to which X_i belongs under partition C, and $Q(X_i)$ the subset to which X_i belongs under partition Q. Let A be

the set of pairs of points $X_i, X_j \in D$ that are placed in the *same* subset according to both partitions, formally:

$$A = \{\langle X_i, X_j \rangle \mid C(X_i) = C(X_j) \text{ AND } Q(X_i) = Q(X_j)\},$$

and B be the set of pairs of points $X_i, X_j \in D$, that are placed in *different* subsets according to both partitions, formally:

$$B = \{\langle X_i, X_j \rangle \mid C(X_i) \neq C(X_j) \text{ AND } Q(X_i) \neq Q(X_j)\}.$$

Denote by $|A|$ and $|B|$ the number of pairs in the sets A and B, respectively. The ARI is then defined as:

$$ARI = \frac{R - E(R)}{1 - E(R)},$$

where $R = \dfrac{|A| + |B|}{\text{Total number of pairs} < X_i, X_j > \text{ in } D}$, $E(R)$ is the expected value of R under chance agreement between C and Q, and 1 is the maximum value that R can obtain (when C and Q are the same). The value of the ARI ranges between -1 and 1. The larger the ARI is, the better the agreement between the partitions C and Q.

By applying three clustering methods to our ADI-R dataset, and evaluating the clustering results of each method using fitness and stability validation, we obtain six best solutions, each of which is the result of applying the combination of a clustering method and a validation procedure to the dataset. We then use the six solutions as components of the cluster ensemble we construct.

4.5 Consensus Clustering and Cluster Ensemble

As described in Section 2.4, we represent each of the 358 subjects as a 6-dimensional vector, where the i^{th} position in the vector is the cluster label assigned to the subject by the i^{th} clustering solution (where $1 \leq i \leq 6$). We call the dataset formed by these 358 six-dimensional vectors the *ensemble dataset*. A consensus function is then applied to this dataset to reach a unified solution.

In our ensemble dataset, the cluster labels that make up the feature vectors are nominal rather than numerical. Nominal data consist of a finite number of distinct, non-ordinal values that do not have a numerical meaning, and can only be tested for equality [HMS01]. Therefore, the similarity of two objects represented by nominal vectors cannot be simply measured by standard metrics such as the Euclidean distance. For the consensus approach we employ in this study, which is co-association-based and is introduced in Section 2.3.1 and Section 2.4, the similarity of the feature vectors in the cluster ensemble can be represented by their co-membership [TJP03]. A similarity-based clustering algorithm, typically hierarchical clustering, can then be used to find the consensus solution for the ensemble dataset. However, we modify the co-association-based approach slightly to suit the specific implementation of hierarchical clustering.

The merging of the clusters in hierarchical clustering, as introduced in Section 4.2.2, is based on the n x n distance matrix which stores the distances between every two data points, where n is the number of data points in a dataset. The distances in fact measure the dissimilarities rather than the similarities between the data points. Therefore, rather than evaluating the similarity between every two nominal objects in the ensemble dataset with their co-membership, we measure the dissimilarity between them. That is, we

examine whether the two objects share a cluster in each component solution in an ensemble, and count the times that they *do not belong* to the same cluster. The ratio between this sum and the total number of the component solutions in an ensemble is the measure of dissimilarity between the two objects. In other words, the distance between every two feature vectors in the ensemble dataset is represented by the *lack of co-association* between them, and is formally defined next.

For the n nominal vectors in a dataset, the distances between every two of them form an n x n distance matrix, denoted by S. As before, X_i and X_j represent two p-dimensional vectors, where $X_i = (X_{i\_1}, X_{i\_2}, ..., X_{i\_p})$, $X_j = (X_{j\_1}, X_{j\_2,...,} X_{j\_p})$, and p is the dimensionality of the ensemble dataset. The distance between X_i and X_j is then denoted by S_{ij}, and is defined as:

$$S_{ij} = \frac{1}{p}\sum_{d=1}^{p}\delta(X_{i\_d}, X_{j\_d}) \quad , \tag{4.9}$$

where

$$\delta(a,b) \equiv \begin{cases} 0, \text{ if } a = b; \\ 1, \text{ otherwise.} \end{cases}$$

We can then use hierarchical clustering to partition the ensemble dataset into several clusters based on the n x n distance matrix defined by Eq. 4.9. This is another clustering problem where the number of clusters in the consensus solution is unknown. Another round of fitness and stability validation, similar to what we introduced in Sections 4.3 and 4.4, is carried out to determine the most appropriate number of clusters, k_{opt}, for the ensemble dataset.

Each of the fitness and stability validation methods is able to find a consensus solution for the ensemble dataset. If the two methods agree with each other, the common solution can be taken to be the final partition of the dataset; otherwise, further analysis is performed to find a solution that can balance the fitness and the stability, and shows good results in regard to both criteria. Statistical tests and domain knowledge are then applied to this unique final consensus to analyze the characteristics of the clusters.

4.6 Analyzing the Clusters in the Consensus Solution

We characterize the impairments shared by subjects within each cluster based on the magnitude of the feature values for each cluster, statistical significance tests on the difference in feature values between every two clusters, and domain knowledge. The magnitude of a feature value for a cluster is represented by the mean of the feature values for the individual subjects in the cluster. The statistical significance test we use in this context is the Wilcoxon rank-sum test [MW05]. For each cluster, we compare its distribution of every feature value to that of another cluster, and find the features that have statistically significantly different distributions between the two clusters.

Among the 358 subjects in this study, the Autism Genetic Resource Exchange project provides clinical diagnoses assigned to 168 of them according to the DSM-IV diagnostic criteria. To demonstrate the agreement and discrepancy between the clinical diagnoses and cluster assignments, we use a contingency table[2] [EM97], and analyze the table with the *Chi*-square test [Che54] and with Fisher's exact test [Fis22].

[2] A contingency table is a table of counts or frequencies. It lists the number of times that an entity falls into a variety of different categories.

Chapter 5

Results and Analysis

In this chapter, we present the results obtained by applying cluster analysis, cluster validation, and consensus clustering to our pre-processed multi-dimensional dataset of 358 patients with PDDs. We first present in Section 5.1 the six best clustering solutions obtained using the *k*-means, hierarchical, and EM clustering methods and the fitness and the stability validation methods. In Sections 5.2 and 5.3, we discuss the integration of these solutions into a single consistent clustering consensus. In Section 5.4, we characterize each cluster in the consensus, and associate it with a subtype of PDDs according to its characteristics. Since some of the subjects in this study have been previously assigned a clinical diagnosis, we compare the cluster assignments to the clinically diagnoses of these subjects.

5.1 Results from Cluster Analysis and Cluster Validation

The results obtained using *k*-means clustering are presented first, followed by those obtained using hierarchical and EM clustering. All these results are validated using fitness and stability validation. For each method, we first present the results obtained using fitness validation, followed by those obtained using stability validation.

5.1.1 *K*-means Clustering Results Obtained Using Fitness Validation

In this study, the *k*-means clustering process stops iterating once the cluster assignment of the data points stops changing. As shown in Table 4.3.4, 20 sets of random initial values

are used for each k, each of which leads to a clustering solution. Among the 20 solutions, the one that produces the lowest sum of squared distances between the data points and their respective cluster centroids, denoted by $SqDist_{min}$, is taken to be the solution with respect to the specified number of clusters, k.

Figure 5.1 shows a plot of the $SqDist_{min}$ as a function of the number of clusters, k. Clearly, the value of $SqDist_{min}$ monotonically decreases as k increases. This is expected, because as the number of clusters increases, the algorithm can find more compact clusters for the data. In the extreme case, every point forms its own cluster and the sum of squared distances becomes zero. Therefore, the sum of squared distances is not a suitable criterion in-and-of-itself for determining the correct number of clusters, $k_{opt,}$ when we use k-means clustering. As discussed in Section 4.3.1, we employ the average silhouette width measure.

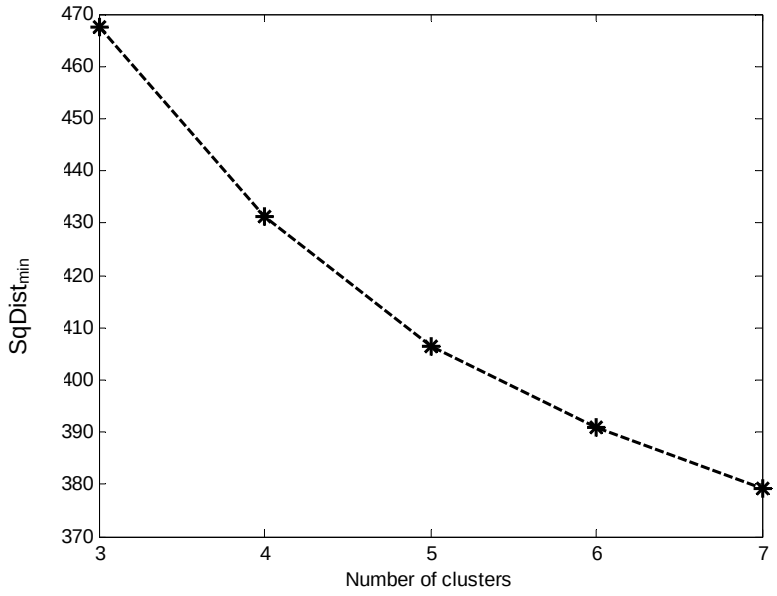

Figure 5.1: The value of $SqDist_{min}$ for each k, obtained using k-means clustering, as a function of the number of clusters.

Figure 5.2 shows a plot of the value of the average silhouette width for *k*-means clustering, as a function of the number of clusters, *k*. The average silhouette width for a given *k* is calculated based on the clustering solution that has the lowest sum of squared distances, $SqDist_{min}$, among the 20 clustering solutions produced for this *k* using different sets of initial values. The value of the average silhouette width is highest when *k* is 3; therefore, the value k_{opt} for our dataset, based on *k*-means clustering with fitness validation, is 3. The corresponding clustering solution is saved as one of the components in the cluster ensemble described in Section 5.2.

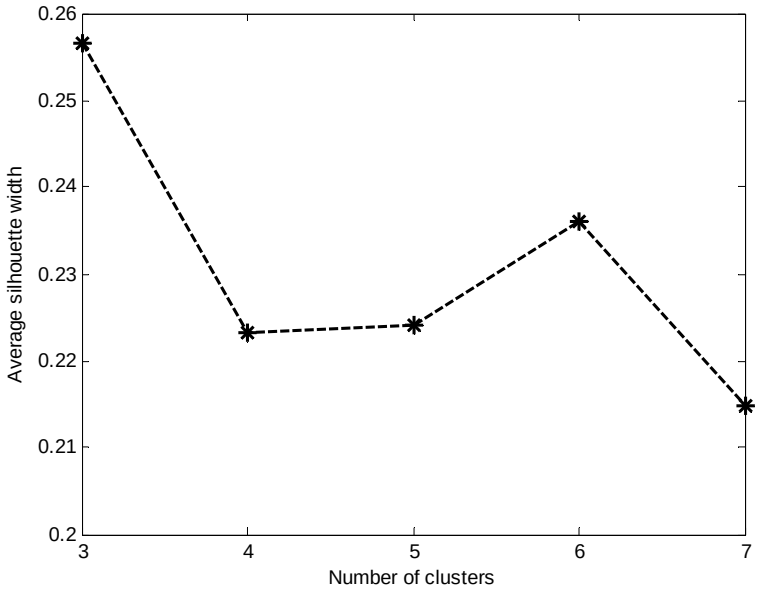

Figure 5.2: The average silhouette width (for k-means clustering) as a function of the number of clusters.

5.1.2 *K*-means Clustering Results Obtained Using Stability Validation

We use 200 random splits of the data to identify the number of clusters *k* for which the Adjusted Rand index (ARI) values are statistically significantly larger than those

produced for any other value of k ($3 \leq k \leq 7$) according to the Wilcoxon signed-rank test, which is a non-parametric form of the paired t-test. The identified k is then taken as the most appropriate number of clusters, k_{opt}. Each split is used to compare the stability of clustering solutions obtained under every two options of k ($3 \leq k \leq 7$). We decide on the number of random splits of the data in an *ad hoc* manner by gradually increasing it until at least the ARI values for a certain k ($3 \leq k \leq 7$) are statistically significant larger than the ARI values for each of the other options of k ($3 \leq k \leq 7$).

As shown in Table 4.5, for each value of k, the N splits of the data lead to N values of the measure that evaluates the agreement between the two halves, namely B_2 and B_2', obtained from each split. In the current context, $N = 200$ and the measure that evaluates the agreement is the ARI. Figure 5.3 shows the plot of the mean and the median taken over the 200 values of the ARI as a function of k. When there are 3 clusters, the corresponding mean and median are the largest.

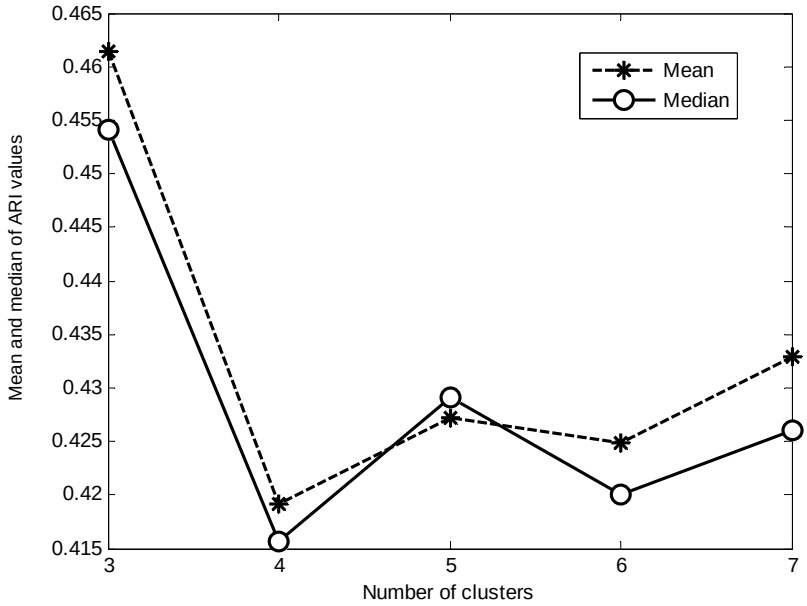

Figure 5.3: The mean and median of the 200 values of the ARI (obtained using k-means clustering) as a function of the number of clusters.

Since 200 ARI values are produced for each k, there are 200 pairs of the ARI values for every two values of k being compared using the Wilcoxon signed-rank test, each pair coming from the same split of the dataset. The Wilcoxon test is used to compare the ARI values in pairs. Table 5.1 contains the p values calculated using the test. If the p value in the row where $k = k_1$ ($3 \leq k_1 \leq 7$) and in the column where $k = k_2$ ($3 \leq k_2 \leq 7$) is greater than 0.05, there is insufficient evidence to reject the hypothesis that there is no difference between the 200 ARI values for the clustering solution with k_1 clusters and those for the solution with k_2 clusters. The alternative hypothesis is that the 200 ARI values obtained for the clustering solution with k_1 clusters are greater than those obtained for the solution with k_2 clusters. The p values on the diagonal of Table 5.1 are obtained when the ARI values for a solution are compared to themselves using the Wilcoxon signed-rank test.

These p values do not contribute to choosing the solution with the most appropriate number of clusters for a dataset, and are therefore not provided. As observed in the table, all the off-diagonal values in the first row show $p \leq 0.05$. This means that the ARI values obtained for $k = 3$ are statistically significantly higher than those obtained for $4 \leq k \leq 7$. Therefore, the best clustering solution with 3 clusters is used as a component in the cluster ensemble in this study.

Table 5.1: The p values obtained from the Wilcoxon signed-rank test on every 200 pairs of the ARI values obtained using k-means clustering. The k values are the numbers of clusters in the clustering solutions.

k	3	4	5	6	7
3	NA	0.0003	0.0047	0.0034	0.0093
4	0.9997	NA	0.8184	0.7853	0.9185
5	0.9953	0.1818	NA	0.3930	0.6145
6	0.9966	0.2151	0.6075	NA	0.7870
7	0.9907	0.0817	0.3860	0.2133	NA

In the previous section, we showed that k_{opt} was identified to be 3 as well using the fitness measure. Hence, when we use k-means clustering as the clustering method, the two criteria, fitness and stability, arrive at the same number of clusters, namely 3. This means that the two solutions obtained using k-means clustering for the cluster ensemble are identical, because there is only one candidate clustering solution for each k ($3 \leq k \leq 7$), which yields the smallest *SqDist* (denoted by $SqDist_{min}$) when 20 sets of initial values are tested, as shown in Section 5.1. We consider them both as distinct solutions in the integration step, as two separate criteria identified this solution as optimal, and as such it should carry twice the weight in the integration step.

5.1.3 Hierarchical Clustering Results Obtained Using Fitness Validation

Figure 5.4 shows a plot of the value of the average silhouette width as a function of the number of clusters k. When k is 3, the clustering yields the highest value of the average silhouette width. Therefore, k_{opt} based on hierarchical clustering using fitness validation is 3. The corresponding clustering is saved as one of the components for the cluster ensemble.

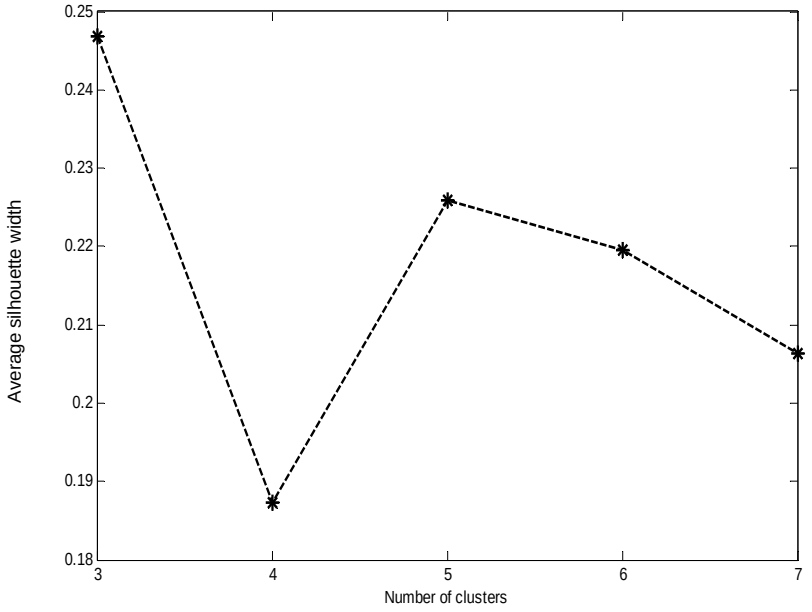

Figure 5.4: The average silhouette width (obtained using hierarchical clustering) as a function of the number of clusters.

5.1.4 Hierarchical Clustering Results Obtained Using Stability Validation

Similar to the stability validation of k-means clustering, we use 500 random splits of the data to identify the most appropriate number of clusters k_{opt}. Since the number of splits is decided in an *ad hoc* manner as mentioned in Section 5.1.2, it can have a varying value in different contexts. This explains why 200 and 500 splits are used for k-means and

hierarchical clustering, respectively. Figure 5.5 shows a plot of the mean and median taken over the 500 ARI values as a function of the number of clusters. The median for k = 6 and the mean for k = 7 are the highest among those obtained for all values of k ($3 \leq k \leq 7$).

Similar to Table 5.1, Table 5.2 shows the p values obtained from the comparison of the ARI values, produced by every two clustering solutions with different number of clusters k ($3 \leq k \leq 7$), using the Wilcoxon signed-rank test. When k = 6 (4th row), the p values off the diagonal are all smaller than 0.05. This means that the ARI values obtained for k = 6 are statistically significantly larger than those obtained for *k=3, 4, 5, or* 7. Therefore, the best clustering solution with 6 clusters is used as a component of the cluster ensemble.

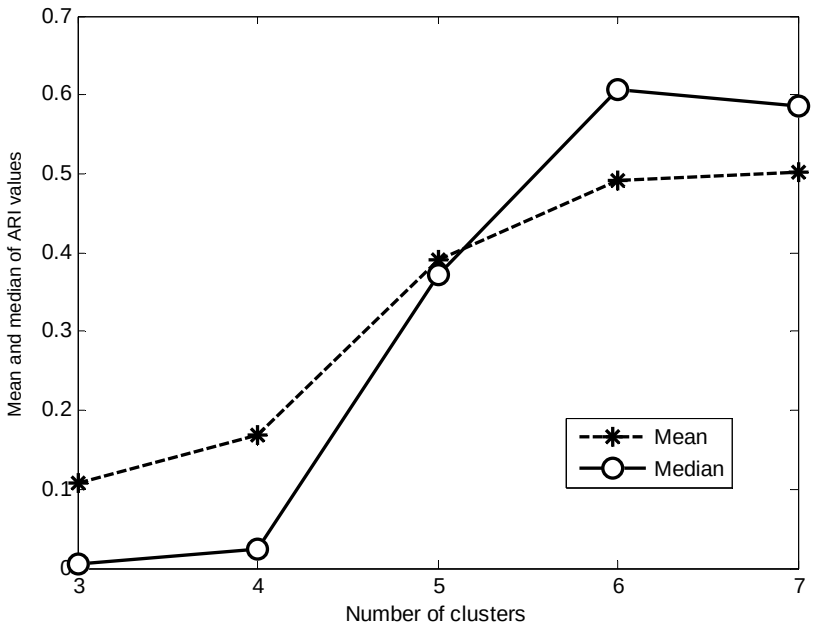

Figure 5.5: The mean and median taken over 500 ARI values (obtained using hierarchical clustering) as a function of the number of clusters.

Table 5.2: The p values from the Wilcoxon signed-rank test on 500 pairs of ARI values obtained using hierarchical clustering. The k values are the numbers of clusters in the clustering solutions.

k	3	4	5	6	7
3	NA	0.9380	0.9999	0.9999	0.9999
4	0.0621	NA	0.9999	0.9999	0.9999
5	0.0000	0.0000	NA	0.9627	0.9894
6	0.0000	0.0000	0.0373	NA	0.0000
7	0.0000	0.0000	0.0106	0.9999	NA

5.1.5 EM Clustering Results Obtained Using Fitness Validation

Figure 5.6 shows a plot of the 1/BIC value as a function of the number of clusters *k*. When *k* is 3, the clustering solution attains the highest 1/BIC value; therefore, k_{opt} based on EM clustering and fitness validation is 3. The corresponding solution is then used as a component in the cluster ensemble.

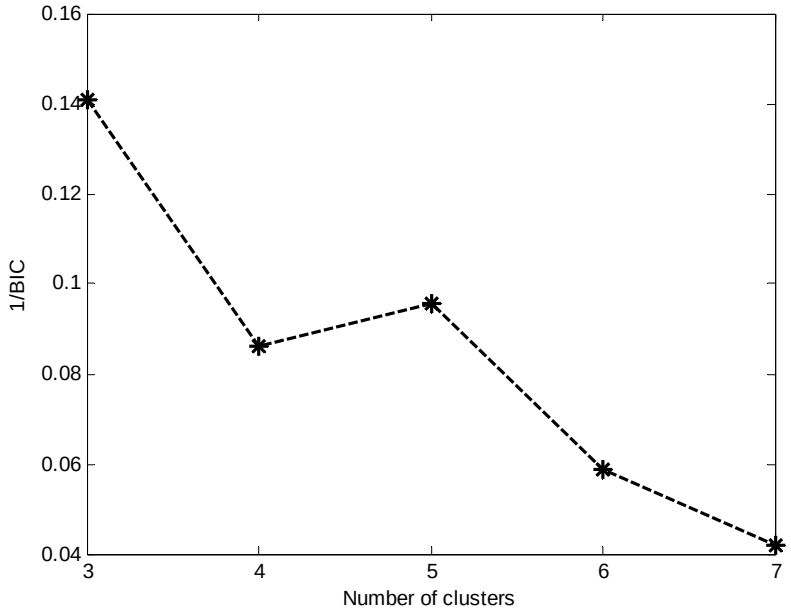

Figure 5.6: The value of the 1/BIC (obtained using EM clustering) as a function of the number of clusters.

5.1.6 EM Clustering Results Obtained Using Stability Validation

Similar to what we have demonstrated with the *k*-mean and hierarchical methods, 200 random data splits are employed to find k_{opt} using the stability validation of EM clustering. Figure 5.7 shows a plot of the mean and median taken over the 200 ARI values as a function of the number of clusters. The mean and median are the highest for the 3-clusters solution. Similar to Tables 5.1 and 5.2, Table 5.3 shows the *p* values when the ARI values, produced by every two clustering solutions with different number of clusters *k* ($3 \leq k \leq 7$), are compared in pairs using the Wilcoxon signed-rank test. We observe that when $k = 3$ (1st row), the off-diagonal *p* values are all smaller than 0.05. This means that the ARI values obtained when *k* is *3* are statistically significantly larger than those obtained for *k* from 4 to 7. Therefore, k_{opt} is taken to be 3 when EM clustering and stability validation are used. The clustering result with three clusters is then included as one of the six components for the cluster ensemble. Similar to what has been explained for *k*-means clustering, the two solutions obtained using EM clustering for the cluster ensemble are identical, with the number of clusters being 3 in both cases.

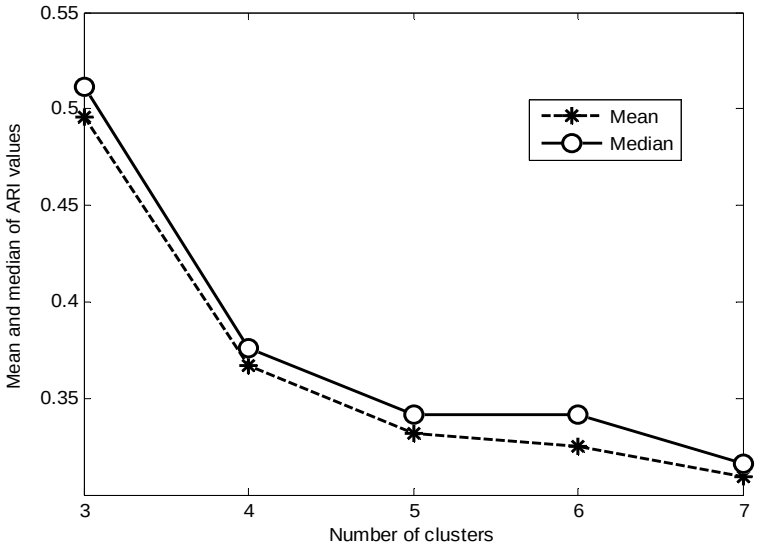

Figure 5.7: The mean and median taken over the 200 ARI values (obtained using EM clustering) as a function of the number of clusters.

Table 5.3: The p values from the Wilcoxon signed-rank test on 200 pairs of the ARI values obtained using EM clustering. The k values are the numbers of clusters in the clustering solutions.

k	3	4	5	6	7
3	NA	0.0000	0.0000	0.0000	0.0000
4	0.9999	NA	0.0067	0.0010	0.0000
5	0.9999	0.9933	NA	0.3206	0.0185
6	0.9999	0.9990	0.6794	NA	0.0247
7	0.9999	0.9999	0.9815	0.9753	NA

5.2 Consensus Clustering Solution

The six best clustering solutions described above form a cluster ensemble. We follow the process described in Section 4.5 to construct an ensemble dataset. There are 358 subjects in this dataset, each of which is represented as a 6-dimensional categorical vector, where the i^{th} position in the vector is the cluster label assigned to the subject by the i^{th} clustering

solution (where $1 \leq i \leq 6$). We obtain two consensus solutions using hierarchical clustering (with the average linkage method) and the fitness and stability validation methods as discussed in Section 4.5. The fitness and stability validation methods, similar to the validation performed on the ADI-R dataset, are used to determine the optimal number of clusters, k_{opt}, for the ensemble dataset. Ideally, the values of k_{opt} obtained using the two validation methods are the same, so that we can simply use the clustering solution with the number of clusters set to k_{opt}. In our study, however, the two validation methods lead to two different values of k_{opt} . Therefore, we choose a solution that can balance the fitness and the stability, and performs well when both the fitness and the stability measures are taken into consideration. In the following sections, we first present the solution obtained using hierarchical clustering and fitness validation for the ensemble dataset, followed by the solution obtained using the same clustering method and stability validation.

5.2.1 Hierarchical Clustering and Fitness Validation for the Ensemble Dataset

Figure 5.8 shows a plot of the value of the average silhouette width as a function of the number of clusters. When k is 7, the clustering yields the highest average silhouette width. Therefore, k_{opt} is taken to be 7 when hierarchical clustering and fitness validation are applied to the ensemble dataset.

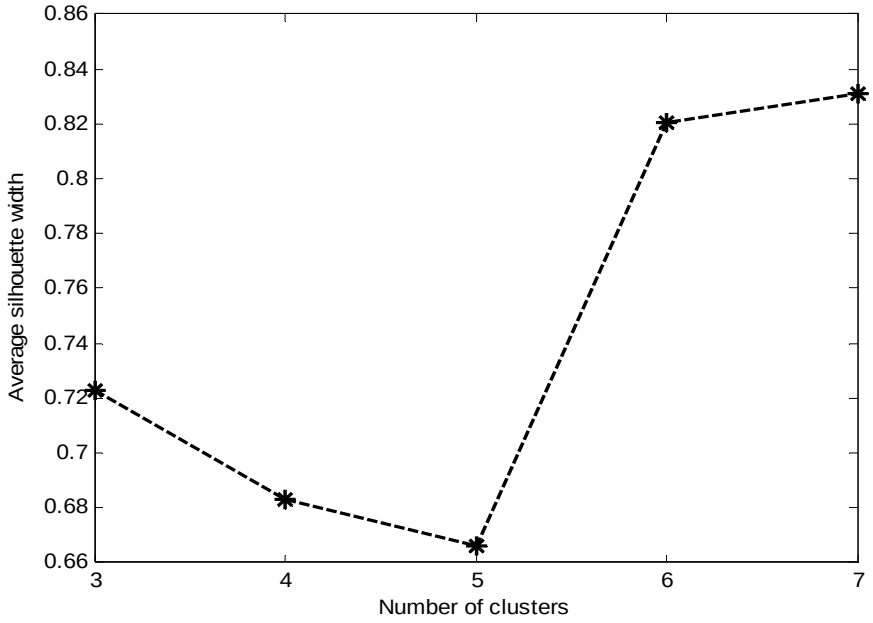

Figure 5.8: The average silhouette width obtained for the ensemble dataset using hierarchical clustering, as a function of the number of clusters

5.2.2 Hierarchical Clustering and Stability Validation for the Ensemble Dataset

We use 200 random data splits to find k_{opt} using stability validation for the ensemble dataset. Figure 5.9 shows a plot of the mean and median taken over the 200 ARI values as a function of the number of clusters. The mean is highest when there are three clusters, while the median is highest when there are four. Similar to what has been described for Tables 5.1, 5.2 and 5.3, in Table 5.4, each row corresponds to a clustering solution with k_1 clusters ($3 \leq k_1 \leq 7$). In each cell of the row in Table 5.4, the ARI values of the solution with k_1 clusters are compared to the ARI values of the solutions with k_2 ($3 \leq k_2 \leq 7$) clusters as specified in the column using the Wilcoxon signed rank test, which yields a p value. The more cells with $p \leq 0.05$ in a row, the higher the ARI values of the solution this row corresponds to compared to the solutions other rows correspond to. We observe

that when k = 3 (1st row), the off-diagonal p values are all smaller than 0.05, which suggests that the ARI values of the solution with three clusters are larger than those of the solutions with 4 to 7 clusters. Consequently, k_{opt} based on hierarchical clustering and stability validation is 3.

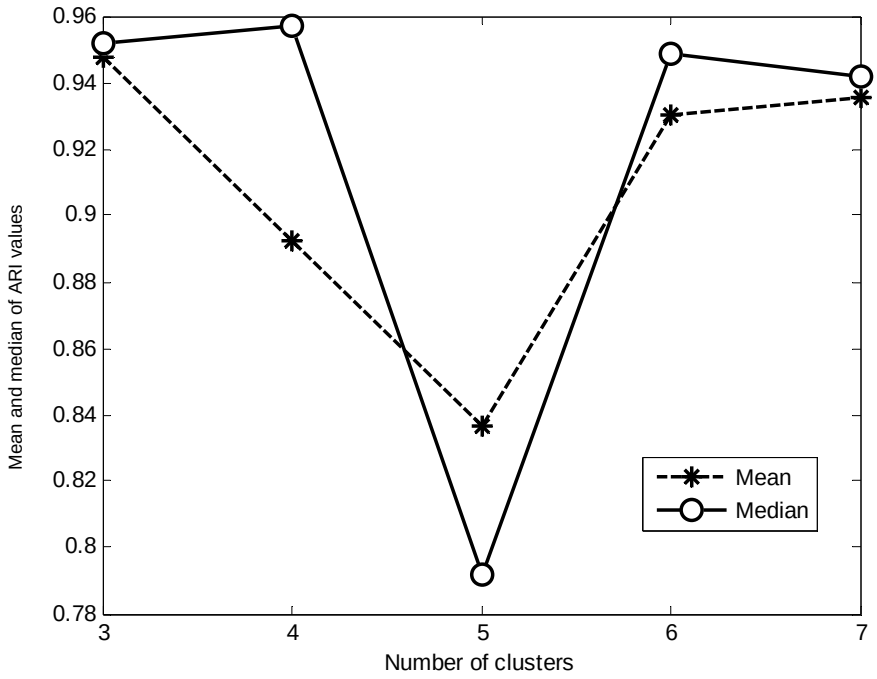

Figure 5.9: The mean and median taken over the 200 ARI values (obtained for the ensemble dataset using hierarchical clustering) as a function of the number of clusters.

Table 5.4: The p values from the Wilcoxon signed-rank test on 200 pairs of the ARI values obtained for the ensemble dataset using hierarchical clustering.

k	3	4	5	6	7
3	NA	0.0002	0.0000	0.0045	0.0000
4	0.9998	NA	0.0000	0.9790	0.9741
5	0.9999	0.9999	NA	0.9999	0.9999
6	0.9955	0.0210	0.0000	NA	0.0021
7	0.9999	0.0259	0.0000	0.9979	NA

As the two consensus partitions identified using the two validation methods are different from each other, one with 7 and the other with 3 clusters, it is practically reasonable to choose a solution that performs well when both the fitness and the stability measures are taken into consideration.

5.3 Choosing a Clustering Solution

To decide on a clustering solution that is considered to be the best when both validation methods are taken into account, we rank the different clustering solutions obtained according to their fitness and the stability[1], and assign them score values equal to their ranks. Thus, for each value of k, the clustering solution with k clusters obtains two rank scores since there are two validation measures used to evaluate the solutions. We then take the sum of the two rank scores, and take the solution that has the largest sum as the consensus solution.

5.3.1 Scoring Solutions based on Ranks

The rank scores of the solutions in the first row in Table 5.5 are based on the fitness of the solutions as measured by the average silhouette width (Figure 5.8). Each cell in the first row has a value equal to the fitness-rank of the solution, ranging from 1 to 5. For the five clustering solutions, each having a different number of clusters ranging from 3 to 7, the higher the average silhouette width for a solution, the higher this solution is ranked with respect to its fitness, therefore the higher its rank score in the first row of Table 5.5.

[1] For the five clustering solutions with $3 \leq k \leq 7$, the ranks are from 1 to 5. A solution that is considered to be the best receives a rank of 5; a solution that is considered to be the poorest receives a rank of 1.

The rank scores in the second row in Table 5.5 are obtained from Table 5.4 based on the number of the cells in each row of Table 5.4 that have $p \leq 0.05$. These scores are again integers ranging from 1 to 5. As shown in Table 5.4 and discussed in Section 5.2.2, each row corresponds to a clustering solution with k_1 clusters ($3 \leq k_1 \leq 7$). The larger the number of cells that have $p \leq 0.05$ in a row, the higher the solution corresponding to this row is ranked among the five solutions with three to seven clusters, and the higher the rank score given to the solution.

Table 5.5: The rank scores obtained for the clustering solutions with different number of clusters. The k values are the numbers of clusters in the clustering solutions.

	k=3	k=4	k=5	k=6	k=7
Rank score (fitness)	3	2	1	4	5
Rank score (stability)	5	2	1	4	3
Sum of rank scores	8	4	2	8	8

We sum up the rank scores in the first and the second row to obtain the scores in the 3rd row in Table 5.5. As three different solutions, namely those with three, six, or seven clusters, all have the same (highest) sum of rank scores (8), we perform further selection to obtain a single solution.

We choose the solution with 6 clusters, based on the following three reasons: First, it balances the fitness and the stability, as reflected by the second highest scores along both these criteria, as shown in the first and second rows of Table 5.5. In contrast, the 3-cluster or the 7-cluster solution scores the highest according one of the criteria. Second, when we compare the 6-cluster solution and the 3-cluster solution, the 6-cluster solution scores much higher in terms of fitness than the 3-cluster solution, while it scores only slightly lower in terms of stability than the 3-cluster solution. To be specific, while the 3-cluster

solution is ranked one place lower than the 6-cluster solution along the stability dimension, as shown in Table 5.4 and in the 2nd row of Table 5.5 and, the mean and the median of the 200 ARI values of the 6-cluster solution are not much lower than those of the 3-cluster solution, as shown in Figure 5.9. In contrast, while the 6-cluster solution is ranked only one place higher than the 3-cluster solution according to the fitness score, as shown in the 1st row of Table 5.5, the value of the average silhouette width of the 6-cluster solution is much higher than that of the 3-cluster solution as shown in Figure 5.8. Therefore, we deem the six-cluster solution to be better than the 3-cluster solution when both their fitness and stability are considered. Third, Figures 5.8 and 5.9 both show that the solutions with six and seven clusters indeed have similar average silhouette width values, and obtain similar mean and median values for the ARI. Since the two solutions have the same sum of scores and similar overall performance with respect to both the fitness and the stability, for parsimony reasons we prefer the solution with fewer clusters. The six-cluster solution is analyzed in detail in the next section.

5.3.2 Analysis of the Six-cluster Solution

Table 5.6 shows the distribution of the 358 subjects across the 6 clusters. Aside from four large clusters with at least 48 subjects in each, there are two very small clusters, one containing 2 and the other 6 subjects. Given the small cluster size, these subjects are as likely to be outliers as to represent distinct subgroups. Indeed, cluster analysis is commonly known as a method for outlier detection [LTS04]. Therefore, our study focuses on the four larger clusters, which contain the majority, (350 out of 358), of the subjects. We refer to a vector that acts as a representative of all the data points assigned

to a cluster as a *prototype*. The distinction among these four major clusters is then confirmed by the distribution of the major prototypes in Table 5.7, and is discussed in the following paragraph.

Table 5.6: The number of the subjects (and respective percentage) in each cluster within the six-cluster consensus solution.

Cluster	Number of subjects (percentage)
1	139 (38.83%)
2	48 (13.41%)
3	6 (1.68%)
4	2 (0.56%)
5	108 (30.17%)
6	55 (15.36%)

Each prototype in Table 5.7 is a unique 6-dimensional nominal vector. The table shows the distribution of prototypes across the data where the leftmost column shows the frequency of each prototype. There are 22 unique prototypes in the table. The majority, 287 out of 358, of the subjects are represented by the first four *distinct* prototype vectors (shown in the first 4 rows of Table 5.7), each of which has at least *two* feature values that are different from the feature values of the other three prototypes. The remaining 18 prototypes either have low frequencies (fewer than 10), or are only different from one of the four major prototypes by *a single* feature value. Therefore, the cluster structure of the ensemble dataset, judged by the prototype vectors, is 4-modal. Each mode corresponds to one of the four major prototypes, and is the centre of a cluster. The subjects characterized

by each of the remaining prototypes, denoted by P, join the subjects characterized by the prototype that is the most similar to P among the four major prototypes.

Table 5.7: The frequencies of the prototypes in the ensemble dataset (the 4 modes are shown in boldface)

Prototype No.	Prototype Vector						Subject Frequency
1	1	1	1	3	3	3	**123**
2	2	2	1	2	2	2	**90**
3	3	3	1	2	2	2	**51**
4	3	3	1	2	1	1	**23**
5	3	3	1	3	1	1	12
6	3	3	1	1	1	1	9
7	1	1	1	3	2	2	9
8	2	2	3	6	2	2	8
9	2	2	1	3	2	2	6
10	1	1	1	3	1	1	5
11	2	2	2	5	2	2	4
12	2	2	1	2	1	1	4
13	3	3	2	4	1	1	3
14	3	3	1	3	2	2	3
15	3	3	2	5	1	1	1
16	3	3	1	3	3	3	1
17	3	3	1	1	2	2	1
18	2	2	1	3	3	3	1
19	2	2	1	3	1	1	1
20	2	2	1	2	3	3	1
21	2	2	1	1	1	1	1
22	1	1	1	2	1	1	1

We observe that the four modes identified from analyzing the distribution of the prototypes in the dataset correspond to the four large clusters in the six-cluster solution found using hierarchical clustering. Specifically, mode 1 corresponds to cluster 1, and modes 2, 3, and 4 correspond to clusters 5, 6, and 2, respectively.

Next, to better understand the results, we summarize the characteristics of the patients in the four large clusters, namely clusters 1, 2, 5 and 6 in the six-cluster solution.

5.4 The Characteristics of the Four Large Clusters

Figure 5.10 shows the feature values associated with each of the four large clusters in the six-cluster solution. The rows correspond to the 22 features of PDDs; the columns correspond to the four clusters. The names of the 22 features are listed beside the rows, and the cluster labels are listed above the columns in the figure. The number of subjects in each cluster is shown underneath each column. As shown by the horizontal bar at the bottom of the figure, the values of the features range from 0 to 1 with darker shades corresponding to higher feature values. The Wilcoxon rank-sum test, a non-parametric form of the *t*-test, in the *R package for statistical computing* [RDCT05], is used to evaluate the statistical significance of the difference between the distributions of feature values for every two clusters. In this section, "significant" and "statistically significant" are thus used interchangeably.

Each feature is referred to as *f* followed by its ordinal number in the 22-feature list in Figure 5.10. For example, "*early onset of symptoms*" is the first feature on the list, and is thus referred to as *f1*. Each of the three core areas of PDDs, namely communicational, social, and behavioural, as well as the area of early development, is represented by several features. The communicational area includes features *f4* - *f7*; the social area includes features *f8*- *f13*; the behavioural area includes features *f14* - *f18*; the area of early development is represented by features *f1* - *f3*. Features *f*19 - *f*22 are mostly add-on features about general behaviours that are related to PDDs, as discussed in Section 3.3. Each cluster is denoted by *Ci* where *i* is the ordinal number of the cluster in Table 5.6 and in Figure 5.10. The four clusters we discuss are thus denoted by *C1, C2, C5* and *C6*. We

first give an overview of the characteristics of the four clusters; and then analyze each in detail.

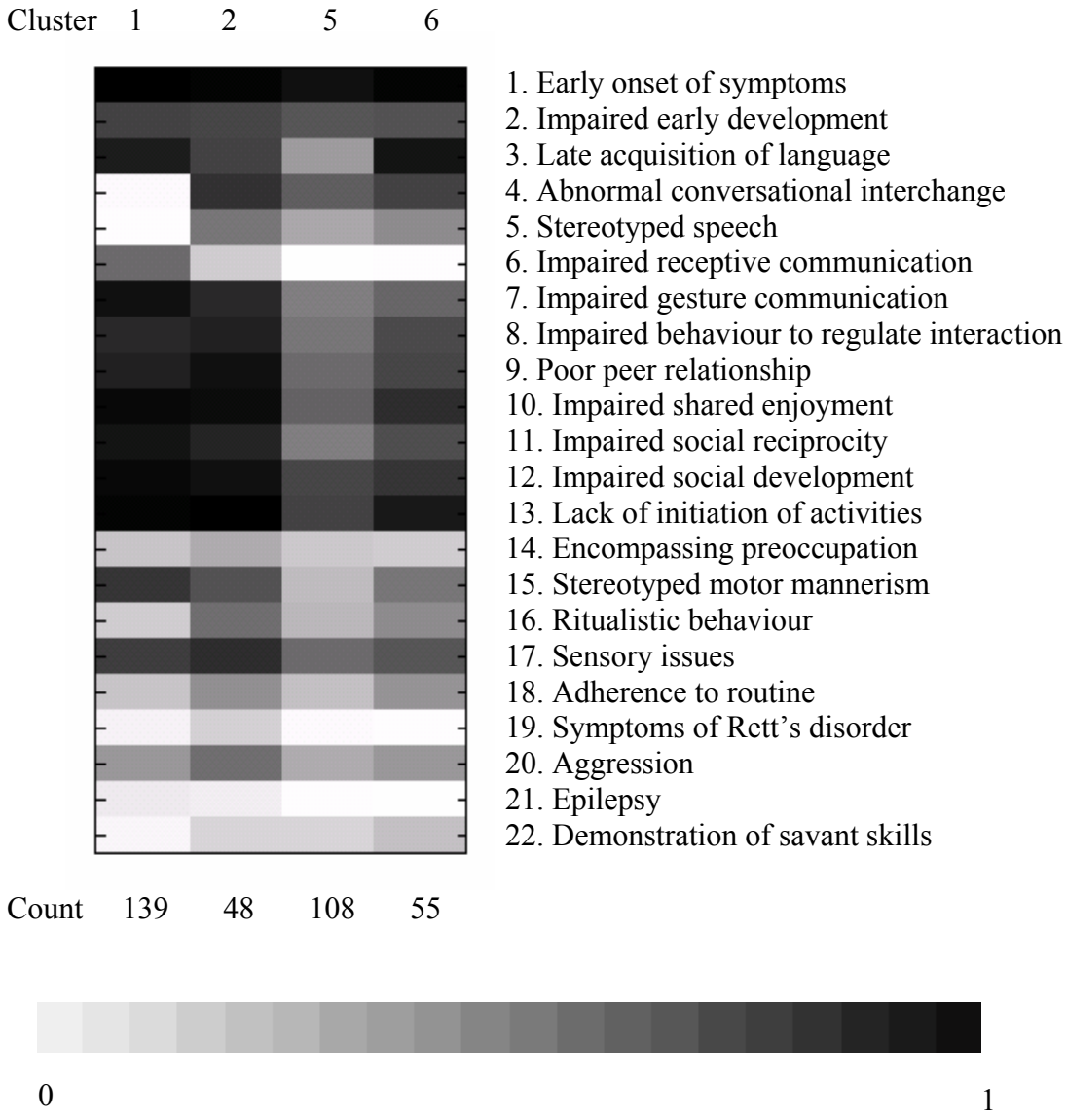

Figure 5.10: Feature values for the four large clusters in the six-cluster solution

5.4.1 Overview of Cluster Characteristics

Each of the clusters shown in Figure 5.10 is characterized by a distinct distribution of feature values. Certain patterns are clearly visible in the figure. First, Cluster *C5* is distinctly different from the other three by having low feature values on the whole. Second, Cluster *C1* has extremely low values for Features *f4* and *f5* that represent verbal abnormality. Third, some of the features have high values across all the clusters, such as "*early onset of symptoms*", while others have low values across all the clusters, such as "*symptoms of Rett's disorder*". Fourth, the general trend of the severity of impairment among the four clusters, from highest to lowest, is: *C1, C2, C6* and *C5*.

According to the results obtained using the Wilcoxon rank-sum test implemented in Matlab [MW05], with the exception of feature *f11* ("*impaired social reciprocity*"), no single feature is statistically significantly different for every pair of compared clusters. Feature *f11* is the only feature whose values show significantly different distribution between every two clusters. As shown in Figure 5.10, subjects in *C1* show the most severely impaired social reciprocity, followed by those in clusters *C2, C6* and *C5*, ordered by the severity of the impairment from high to low. This order is consistent with the overall order of the severity of impairments across the four clusters mentioned in the previous paragraph. Social reciprocity, such as offering comfort and showing appropriate social responses, is one of the cornerstones of healthy development in children [Gro02, CT00]. Aside for identifying the general patterns for the four clusters, we also characterize each of them in the following section.

5.4.2 Analysis of the Individual Clusters

Clusters *C1* and *C2* group together the most severely impaired patients, corresponding to the typical Autism subtype in the DSM-IV. Patients in both clusters show highly impaired social function (*f8-13*), *"stereotyped motor mannerism"* (*f15*) and are hypersensitive to stimuli (*f17*). However, the two clusters are still distinct.

Subjects in Cluster *C1* demonstrate late language acquisition and language impairment so severe that they are considered non-verbal. They have no functional use of three-word phrases, and at times are completely mute. Their scores on features *f4* and *f5* are close to zero, as almost all of them cannot be evaluated for their impairment in spoken language due to the lack of speech. Almost two thirds of them have problems understanding other people's language (*f6*), and their social reciprocity is significantly lower than that of the subjects in *C2*.

In contrast, the subjects in Cluster *C2* exhibit overly persistent (*f16*) and aggressive behaviours (*f20*), which are significantly different from those present among subjects in *C1*. While they do demonstrate some language skills, their verbal development is still delayed (*f3*), and severely impaired as demonstrated by stereotyped speech and poor conversational ability (*f4*, *f5*).

Cluster *C5* contains the least impaired subjects, corresponding to the typical subtype of the Asperger's disorder in the DSM-IV. Although these subjects also show early onset of symptoms (*f1*) and impaired early development (*f2*), the feature values for these subjects are lower compared to those for the subjects in all other three clusters. This difference is tested to be significant except when *C5* is compared to *C6* on early development (*f2*). Notably, subjects in *C5* show a close-to-normal age of language

acquisition (*f3*). However, they cannot be called normal because their social and communicational abilities (*f4*, *f7-13*) are still impaired, and they are hypersensitive to stimuli (*f17*).

Cluster *C6* is characterized by an intermediate level of severity. It is similar in characteristics to *C2,* but shows lower scores for almost all features except for delayed language acquisition (*f3*). For 10 of the features, the lower scores compared to *C2* are highly statistically significant ($p \leq 0.05$). We note that the characteristics of the subjects in this cluster are generally those of PDD-NOS, which stands for PDD-Not Otherwise Specified.

5.4.3 Comparison between Clinical Diagnosis and Cluster Assignment

For 168 of the 358 subjects studied here, the Autism Genetic Resource Exchange [AGRE] provides clinical diagnoses made by physicians based on the DSM-IV criteria. Some of the 168 subjects are assigned multiple diagnoses. In such cases, we use as the clinical diagnosis for each subject the one that is a subtype of PDDs and is listed first among the diagnoses for the subject. For example, when a subject is diagnosed to have "Behaviour disorder, Depression, Bi-polar, Asperger's disorder, Autism", we consider Asperger's disorder to be her clinical diagnosis because it is a subtype of PDDs, and is listed before the other subtype diagnoses of PDDs (Autism in this case).

Each of the two smaller clusters among the six-cluster consensus solution we identified, namely Clusters *C3* and *C4,* has only one subject that previously received a clinical diagnosis. Therefore, there are 166 subjects with clinical diagnoses among the

350 subjects in the four large clusters analyzed here. These 166 subjects are the focus of the comparison between clinical diagnoses and cluster assignments in this section.

Table 5.8 shows the correspondence between the clinical diagnoses (rows) and the cluster memberships (columns) for the four large clusters. Clusters *C1*, *C2* and *C6* are dominated by Autism diagnoses, while Cluster *C5* is dominated by Asperger's disorder diagnoses. The *Chi*-square test (or Fisher's exact test[2]) shows that the distributions of the clinical diagnoses are statistically significantly different ($p \leq 0.05$) between Clusters *C1*, *C5* and either *C2* or *C6*, which means that we can reject the hypothesis that the cluster assignments are independent of the clinical diagnoses for these comparisons, and accept the hypothesis that the assignments and the diagnoses are related by more than chance. Such a significant difference is expected because the feature patterns of *C1*, *C5* and either *C2* or *C6* are notably distinct as shown in Figure 5.10. However, there is no statistically significant difference between the distributions of the clinical diagnoses for clusters *C2* and *C6*.

Table 5.8: Correspondence between the clinical diagnoses (rows) and the cluster memberships (columns)

	C1	C2	C5	C6	Total
Autism	75	16	18	14	123
Asperger's	0	4	21	2	27
PDD-NOS	2	3	9	2	16
Total	77	23	48	18	166

[2] Fisher's exact test is used in place of the *Chi*-square test when there is at least one cell in a contingency table that has count zero, or more than 80% of the cells have counts fewer than 5 because the *Chi*-squared probabilities do not apply to these situations.

To summarize, using an ensemble of validated clustering solutions, we identified four clusters that roughly correspond to – and further refine – three main subtypes of PDDs, namely Autism, PDD-NOS and Asperger's disorder. In the next chapter, we conclude the thesis with a brief summary of our work, and then extend the discussion to future studies we plan to perform.

Chapter 6

Conclusion and Future Work

Cluster analysis for identifying subtypes of PDDs can lead to targeted aetiology studies and to effective type-specific intervention. The study presented throughout this work proposes a framework for subtyping PDDs based on the integration of multiple validated clustering results from three clustering methods. We identified six clusters, four of which are analyzed in detail based on the distribution of feature scores and certain domain knowledge of autism..

The dataset used here is the largest ADI-R dataset analyzed so far. Our clusters are characterized by a distribution of scores along many features, and thus distinguish among subgroups based on finer criteria than those defined by the DSM-IV. The clusters form a continuum of severity along the different impairments and thus agree with the opinion held by many researchers that the subtypes of PDDs should not be distinguished based on discrete, mutually exclusive, impairments but rather form a spectrum of disorders varying in severity from almost normal to highly impaired [BS01]. The difference we find between the cluster memberships and clinical diagnoses highlights the value of cluster analysis as a method for identifying subtypes in the data that may not be identified using a rule-based algorithm such as that defined for the ADI-R.

In future studies we plan to include data from other sources so that we can validate our clustering results externally. Data sources such as genomic data obtained in molecular biology labs, family information collected from the siblings of the patients in this study, and IQ data collected via IQ tests can be used for the purpose of external

validation[1] because genetic aetiology exists for many cases of PDDs, as concluded by major reviews of sibling, twin and family studies [MTR04]. The IQ data may also be used as a source of validation data because IQ is one of the most important factors (the other being language ability) in explaining the different symptoms of PDDs [GS97, VCBH89].

Alternatively, we can add clustering solutions obtained from other data sources to the cluster ensemble we built in this study. To achieve this, the data obtained from other sources should also be clustered, so that the resulting clustering solutions can then be treated as additional components in the cluster ensemble used in this study to produce a consensus clustering.

We may also assign different weights to different features according to their importance in the diagnosis of the subtypes of PDDs. This idea is reasonable because it is based on the finding that some of the features, such as those concerning language development and ability, and social reciprocity are more important than other features, such as aggression and epilepsy, in the diagnosis of the subtypes of PDDs [MSNA01, LRL94]. By giving the more important features higher weights, we may be able to detect clustering structure that is more suitable for the particular problem of finding the subtypes of PDDs.

By including data from the sources mentioned above, we will be able to base our clustering results on more comprehensive information than used in our present study to reach a potentially more objective and fully validated (both internally and externally) solution. This is important because PDDs are disorders that affect many aspects other

[1] Thanks to Dr. Ira Cohen for this suggestion.

than social interaction, communication, and behaviour. By weighing the features, we can exploit more of the in-depth knowledge previously gained about PDDs, by the numerous studies carried out by other groups. This process will require a close collaboration between researchers working in the areas of data analysis and autism.

Another direction of future work is to employ soft clustering. Soft clustering typically assigns a data point to belong to multiple clusters with different probabilities. This approach is likely to be a better fit for the continuous view of PDDs than the deterministic clustering method presented in this thesis, because it not only circumvents the need to fix a cut-off threshold for feature values, but also avoids setting strict boundaries between the resulting clusters.

Bibliography

[AGRE] Autism Genetic Resource Exchange. http://www.agre.org

[AH85] L. J. Arabie, P. Hubert. Comparing partitions. Journal of classification. 2: 193-218. 1985.

[APA94] The American Psychiatric Association. The American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition (DSM-IV). The American Psychiatric Association. 1994.

[ASC05] http://www.autismsocietycanada.ca/understanding_autism/ what_ are_asds/index_e.html. Autism Society Canada. 2005.

[Ash07] J. Asher. NIMH news press. March 15. 2007. http://www.nih.gov/news/pr /mar2007/nimh-15.htm

[BEG02] A. Ben-Hur, A. Elisseeff, I. Guyon. A stability based method for discovering structure in clustered data. In Aetman, R.B. (Ed.), *et al*. Pacific Symposium on Biocomputing. New Jersey World Scientific Publishing Co. 2002.

[BMC00] M. Bittner, P. Meltzer, Y. Chen. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. 406: 536–540. 2000.

[BO94] M. C. Borden, T. H. Ollendick. An examination of the validity of social subtypes in autism. Journal of Autism and Developmental Disorders. 24: 23–38. 1994.

[BP95] J. Brown, P. Prelock. The Impact of Regression on Language Development in Autism. Journal of Autism and Developmental Disorders. 25: 305–309. 1995.

[Bre01] L. Breiman. Random Forests. Machine Learning. 45 (1): 5-32. 2001.

[Bre89] J. Breckenridge. Replicating cluster analysis: Method, consistency, and validity. Multivariate Behavioural Research, 24: 147- 161. 1989.

[Bre96] L. Breiman. Bagging predictors. Machine Learning. 24(2):123-140. 1996.

[BS01] L. J. Beglinger, T. H. Smith. A Review of Subtyping in Autism and Proposed Dimensional Classification Model. Journal of Autism and Developmental Disorders. 31(4): 411-22. 2001.

[CD93] P. Castelloe, G. Dawson. Subclassification of children with autism and pervasive developmental disorder: A questionnaire based on Wing's subgrouping scheme. Journal of Autism and Developmental Disorders. 23: 229-241. 1993.

[Che54] Chernoff H, Lehmann E.L. The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit. The Annals of Mathematical Statistics. 25:579-586. 1954.

[CT00] J. N. Constantino, R. D. Todd. Genetic Structure of Reciprocal Social Behaviour. American Journal of Psychiatry. 157:2043-2045. 2000.

[CT85] J. F. Curry, R. J. Thompson, Jr. Patterns of behavioural disturbance in developmentally disabled and psychiatrically referred children: A cluster analytic approach. Journal of Pediatric Psychology. 10: 151-167. 1985.

[CTDC05] T. Charman, E. Taylor, A. Drew, H. Cockerill, J. Brown, G. Baird. Outcome at 7 Years of Children Diagnosed with Autism at Age 2: Predictive Validity of Assessments Conducted at 2 and 3 Years of Age and Pattern of Symptom Change over Time. Journal of Child Psychology and Psychiatry. 46(5): 500-513. 2005.

[DBSK04] A. De Bildt, S. Sytema, C. Ketelaars, D. Kraijer, E. Mulder, F. Volkmar, R. Minderaa: Interrelationship between Autism Diagnostic Observation Schedule-Generic (ADOS-G), Autism Diagnostic Interview- Revised (ADI-R), and the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) Classification in Children and Adolescents with Mental Retardation. Journal of Autism and Developmental Disorders. 34(2): 129-137. 2004.

[DF02] J. Dudoit, S. Fridlyand, A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology. 1-21. 2002.

[DHS01] R. O. Duda, P. E. Hart, D. G. Stork. Pattern Classification. John Wiley & Sons Inc. 2001.

[DLR77] A. Dempster, N. Laird, D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 39(1):1-38. 1977.

[DWM01] M. Daszykowski, B. Walczak, D. L. Massart. Looking for Natural Patterns in Data, Part 1: Density Based Approach. Chemometrics and Intelligent Laboratory Systems. 56: 83-92. 2001.

[EHE94] L.C. Eaves, H.H. Ho, D. M. Eaves. Subtypes of autism by cluster analysis. Journal of Autism and Developmental Disorders. 24: 3–22. 1994.

[EKSX96] M. Ester, H. P. Kriegel, J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining. 1996.

[Elk97] C. Elkan. Boosting and Naive Bayesian learning. Technical report. Department of Computer Science and Engineering. University of California, San Diego. 1997.

[EM97] V.J. Easton, J.H. McColl. Statistics Glossary, STatistical Education through Problem Solving (STEPS). 1997. http://www.cas.lancs.ac.uk/glossary_v1.1/catdat.html

[FD01] J. Fridlyand, S. Dudoit. Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Berkeley Department of Statistics. 2001.

[FH05] U. Frith, F. Happe. Autism spectrum disorder. Current biology. 15(19): 786-790. 2005.

[Fis22] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society. 85(1):87- 94. 1922.

[FJ02] A. Fred, A. K. Jain. Data Clustering Using Evidence Accumulation. Proceedings of the 16th International Conference on Pattern Recognition. 276-280. 2002.

[FM83] E. B. Fowlkes, C. L. Mallows. A method for comparing two hierarchical clusterings. Journal of American Statistics Association. 78:553-584. 1983.

[FMBB03] S. Fecteau, L. Mottron, C. Berthiaume, J. A. Burack. Developmental Changes of Autistic Symptoms Autism. 7(3): 255 - 268. 2003.

[Fom02] E. Fombone. Prevalence of childhood disintegrative disorder. Autism. 6(2): 149-57. 2002.

[Fom03] E. Fombonne. Epidemiological surveys of autism and other pervasive developmental disorders: an update. Journal of autism and developmental disorders. 33(4): 365-82. 2003.

[FRS01] S. Folstein, B. Rosen-Sheidley. Genetics of Autism: Complex Aetiology for a Heterogeneous Disorder. Nature Reviews Genetics. 2: 943-955. 2001.

[FS96] Y. Freund, R. E. Schapire. Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference. 148–156. 1996.

[GE03] I. Guyon, A. Elissee. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research. 3: 1157-1182. 2003.

[GOF03] W. Goldberg, K. Osann, P. Filipek. Language and other regression. Journal of Autism and Developmental Disorders. 33(6): 607-616. 2003.

[GPA05] L. Gabis, J. Pomeroy, M. Andriola. Autism and Epilepsy: Cause, Consequence, Comorbidity, Or Coincidence? Epilepsy Behaviour. 7(4): 652-656. 2005.

[GS87] C. Gillbert, S. Steffenburg. Outcome and prognostic factors in infantile autism and similar conditions: A population-based study of 46 cases followed through puberty. Journal of Autism and Developmental Disorders. 17(2): 273-287. 1987.

[Hai92] J. F. Hair. Multivariate data analysis (3rd edition). New York: Macmillan. 1992.

[Ham50] R. Hamming. Error detecting and error correcting codes. The Bell System Technical Journal 29: 147-160. 1950.

[HK06a] J. Han, M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann. 2006.

[HK06b] J. Handl, J. Knowles. Multiobjective clustering and cluster validation. Springer Series on Computational Intelligence 16:21-47, 2006.

[HP05] Personal Communication with Heidi Penning. 2005.

[HMS01] D. Hand, H. Mannila, P. Smyth. Principles of Data Mining. The MIT Press. 2001.

[Hua98] Huang Z. Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery. 2(2): 283–304. 1998.

[JD88] A. K. Jain, R. Dubes. Algorithms for clustering data. Prentice-Hall. 1988.

[Gro02] J. Groden. Book Reviews. Autism. 6:136 - 140. 2002.

[JMPC98] I. Jambaqué, A. L. Mottron, B. G. Ponsot, A. C. Chironc. Autism and Visual Agnosia in A Child with Right Occipital Lobectomy. Journal of Neurology, Neurosurgery, and Psychiatry. 65:555-560. 1998.

[Joh67] S. C. Johnson. Hierarchical clustering schemes. Psychometrika. 32: 241–254. 1967.

[Kan43] L. Kanner. Autistic disturbances of affective contact. The Nervous Child. 2: 217-250. 1943.

[Kes02] Kessler RC. The Categorical versus Dimensional Assessment Controversy in the Sociology of Mental Illness. Journal of Health and Social Behaviour. 43:171-188. 2002.

[Kir02] C. W. Kirkwood. Decision Tree Primer. 2002 http://www.public.asu.edu/~kirkwood/DAStuff/decisiontrees/.

[KR90] L. Kaufman, P. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley. 1990.

[Kuh55] H. Kuhn. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly. 2 : 83-97. 1955.

[LLC01] C. Lord, B. L. Leventhal, E. H. Cook Jr., Quantifying the Phenotype in Autism Spectrum Disorders. American Journal of Medical Genetics (Neuropsychiatric Genetics) 105: 36–38. 2001.

[LLR03] A. Le Couteur, C. Lord, M. Rutter. ADI-R Autism Diagnostic Interview – Revised. WPS Edition, Interview Protocol. Western Psychological Services. 2003.

[Lor95] C. Lord. Follow-up of two-year-olds referred for possible autism. Journal of Child Psychology and Psychiatry. 36(8): 1365-1382. 1995.

[LRB04] T. Lange, V. Roth, M. Braun. Stability-based validation of clustering solutions. Neural Computation. 16**:**1299–1323. 2004.

[LOS04] T. Li, M. Ogihara, S. Ma. On Combining Multiple Clusterings. In Proceedings of the Thirteenth Conference on Information and Knowledge. 47: 294-303. 2004.

[LRL94] C. Lord, M. Rutter, A. Le Couteur. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. Journal of Autism and Developmental Disorders. 24(5): 659-85. 1994.

[LRGH89] C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, E. Schopler. Autism diagnostic observation schedule: A standardized observation of communicative and social behaviour. Journal of Autism and Developmental Disorders. 19(2): 185-212. 1989.

[LRLC00] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr., B. L. Leventhal, P. C. DiLavore, A. Pickles, M. Rutter. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders. 30(3): 205-23. 2000.

[LRLR89] A. Le Couteur, M. Rutter, C. Lord, P. Rios, S. Robertson, M Holdgrafer, J. McLennan. Autism Diagnostic Interview: A Standardized Investigator-Based Instrument. Journal of Autism and Developmental Disorders. 19(3): 363-387. 1989.

[LTS04] A. Loureiro, L. Torgo, C. Soares. Outlier Detection using Clustering Methods: a data cleaning application. Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. 2004.

[LW02] A. Liaw, M. Wiener. Classification and regression by random forest. R News. 2: 18–22. 2002.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceeding of the 5th Symposium on Mathematical Statistics and Probability. 1: 281–297. 1967.

[MO00] J. N. Miller, S. Ozonoff. The external validity of Asperger Disorder: Lack of evidence from the domain of Neuropsychology. Journal of Abnormal Psychology 109, 227-238. 2000.

[MTP04] B. Minaei-Bidgoli, A. Topchy, W. Punch. A Comparison of Resampling Methods for Clustering Ensemble. International conference on Machine Learning; Models, Technologies and Applications. 939-945. 2004.

[MC85] G. W. Milligan, M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 50:159-179. 1985.

[MC86] G. W. Milligan, M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behaviour Research. 21:846—850. 1986.

[MBM03] P. Manning-Courtney, J. Brown, C. A. Molloy. Diagnosis and Treatment of Autism Spectrum Disorders. Current Problems in Pediatric and Adolescent Health Care. 33(9): 283-304. 2003.

[MSA01] G. B. Mesibov**, V.** Shea**,** L. W. Adams**.** Understanding Asperger Syndrome and High Functioning Autism. The Autism Spectrum Disorders Library. 1: 136. 2001.

[MSMB98] W. Mahoney, P. Szatmari, J. Maclean, S. Bryson, G. Bartolucci, S. Walter, M. Jones, L. Zwaigenbaum. Reliability and accuracy of differentiating pervasive developmental disorder subtypes. Journal of the American Academy of Child and Adolescent Psychiatry. 37: 278–285. 1998

[MSNA01] K. Mildenberger, S. Sitter, M. Noterdaeme, H.Amorosa. The use of the ADI-R as a diagnostic tool in the differential diagnosis of children with infantile autism and children with a receptive language disorder. European Child & Adolescent Psychiatry. 10(4): 248-255. 2001.

[MTR04] R. Muhle, S. V. Trentacoste**,** I. Rapin. The Genetics of Autism. Pediatrics 11(3): 472-486. 2004.

[Myh98] G. Myhr. Autism and other pervasive developmental disorders: exploring the dimensional view. Canadian Journal of Psychiatry. 43: 589–95. 1998.

[MW05] The MathWorks Inc., MATLAB version 7. R14. 2005.

[NDTH03] E. Nurmi, M. Dowd, O. Tadevosyan-Leyfer, J. Haines, S. Folstein, J. Sutcliffe. Exploratory Subsetting of Autism Families based on Savant Skills Improves Evidence of Genetic Linkage to 15q11–Q13. Journal of the American Academy of Adolescence Psychiatry. 42:856–863. 2003.

[NIH06] http://www.nichd.nih.gov/publications/pubs/sos_autism/sub4.cfm. National Institute of Child Health and Human Development. National Institute of Health. USA. 2006.

[OSM00] S. Ozonoff, M. South, J.N. Miller. DSM-IV defined Asperger syndrome: Cognitive, behavioural and early history differentiation from High Functioning Autism. Autism 4, 29-46. 2000.

[OWL05] S. Ozonoff, B. Williams, R. Landa. Parental Report of the Early Development of Children with Regressive Autism: The Delays-Plus-Regression Phenotype. 9: 461-486.Autism. 2005.

[PBGP75] M. Prior, D. Boulton, C. Gajzago, D. Perry. The classification of childhood psychoses by numerical taxonomy. Journal of Child Psychology and Psychiatry. 16: 321–330. 1975.

[PBM04] M. Prior, S. Barrett, J. Manjiviona. Children on the borderlands of autism. Autism. 8(1):61–87, 2004.

[PELW98] M. Prior, R. Eisenmajer, S. Leekam, L. Wing, J. Gould, B. Ong, D. Dowe. Are There Subgroups within the Autistic Spectrum? A Cluster Analysis of a Group of Children with Autistic Spectrum Disorders. The Journal of Child Psychology and Psychiatry and Allied Disciplines. 39: 893-902. 1998.

[RDCT05] R Development Core Team. R: A language and environment for statistical computing, reference index version 2.4.0. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. 2005.

[Raf86] A. Raftery. A note on Bayes factors for log-linear contingency table models with vague prior information. Journal of the Royal Statistical Society. 48(2): 249-250. 1986.

[RBBL94] M. Rutter, P. Bailey, Bolton, A. Le Couteur. Autism and known medical conditions: myth and substance. *Journal of Child Psychology and Psychiatry.* 35**:** 311-322. 1994.

[Res88] L. Rescorla. Cluster analytic identification of autistic preschoolers. Journal of Autism and Developmental Disorders. 18: 475–492. 1988.

[RF68] F. J. Rohlf, D.R. Fisher. Testes for hierarchical structure in random data sets. Systematic Zoology. 17: 407-412. 1968.

[RKO01] M. Rutter, J. Kreppner, T. G. O'Connor, the ERA Team. Specificity and heterogeneity in children's responses to profound deprivation. British Journal of Psychiatry. 179: 97-103. 2001.

[RLL03] M. Rutter, A. Le Couteur, C. Lord. ADI-R: The Autism Diagnostic Interview-Revised. Western Psychological Services. 2003.

[RMFP90] E. R. Ritvo, A. Mason-Brothers, B. J. Freeman, C. Pingree, W. R. Jenson, W. M. McMahon, P. B. Petersen, L. B. Jorde, A. Mo and A. Ritvo. The UCLA-University of Utah epidemiologic survey of autism: the etiologic role of rare diseases. American Journal of Psychiatry. 147:1614-1621. 1990.

[Rog98] S. J. Rogers. Empirically supported comprehensive treatments for young children with autism. Journal of Clinical Child Psychology. 27:168-179. 1998.

[RTBI06] N. J. Rinehart, B. J. Tonge, J. L. Bradshaw, R. Iansek, P. G. Enticott, J. McGinley. Gait function in high-functioning autism and Asperger's disorder Evidence for basal-ganglia and cerebellar involvement? European Child & Adolescent Psychiatry. 15(5): 256-264. 2006.

[SACB86] B. Siegel, T. F. Anders, R. D. Ciaranello, B. Bienenstock, H. C. Kraemer. Empirically Derived Subclassification of the Autistic Syndrome. Journal of Autism and Developmental Disorders. 16(3): 275-293 .1986.

[SFDW00] M. C. Stevens, D. A. Fein, M. Dunn, D. Allen, L. H. Waterhouse, C. Feinstein, I. Rapin. Subgroups of children with autism by cluster analysis:

a longitudinal examination. Journal of American Academy of Child Adolescence Psychiatry. 39: 346-52. 2000.

[SGBZ06] P. Szatmari, S. Georgiades, S. Bryson, L. Zwaigenbaum, W. Roberts, W. Mahoney, J. Goldberg, L. Tuff. Investigating the structure of the restricted, repetitive behaviours and interests domain of autism._Journal of Child Psychology and Psychiatry. 47(6): 582-590. 2006.

[SLMT07] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge *et al.* Strong Association of De Novo Copy Number Mutations with Autism. 316: 445-449. Science. 2007.

[SLTW04] V. Svetnik, A. Liaw, C. Tong, T. Wang. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In F. Roli, J. Kittler, T. Windeatt (eds.). Multiple Classifier Systems, Fifth International Workshop. 334-343. MCS Proceedings. 2004.

[SMCL95] J. A. Sevin, J. L. Matson, D. Coe, S. R. Love, M. J. Matese, D. A. Benavidez. Empirically derived subtypes of pervasive developmental disorders: A cluster analytic study. Journal of Autism and Developmental Disorders. 25: 561-578. 1995.

[SOBD05] A. E. Skodol, J. M. Oldham, D. S. Bender, I. R. Dyck. Dimensional Representations of DSM-IV Personality Disorders: Relationships to Functional. American Journal of Psychiatry 162: 1919-1925. 2005.

[SOHH99] W. Stone, O. Ousley, S. Hepburn, K. Hogan, C. Brown. Patterns of adaptive behaviour in very young children with autism. American Journal on Mental Retardation. 104: 187–199. 1999.

[SS73] P. Sneath, R. Sokal. Numerical taxonomy. W. H. Freeman, San Francisco. 1973.

[SSC98] A. Sharkey, N. Sharkey, S. Cross. Adapting an Ensemble Approach for the Diagnosis of Breast Cancer. In Proceedings of ICANN 98. Springer-Verlag. 281-286. 1988.

[ST05] W. L. Stone, L. Turner. The Impact of Autism on Child Development. Encyclopedia on Early Childhood Development. Centre of Excellence for Early Childhood Development. Published Online August 24. 2005.

[Str04] M. Strock. Autism Spectrum Disorders (Pervasive Developmental Disorders). NIH Publication No. NIH-04-5511, National Institute of Mental Health, National Institutes of Health. 2004.

[Sti82] S. M. Stigler. Thomas Bayes' Bayesian Inference. Journal of the Royal Statistical Society, Series A. 145: 250–258. 1982.

[Sza00] P. Szatmari. The classification of autism, Asperger's syndrome, and pervasive developmental disorder. Canadian Journal of Psychiatry 45: 731-738. 2000.

[Tek06] K. Teknomo. K-Mean Clustering Code in Matlab.

http://people.revoledu.com/kardi/tutorial/kMean/matlab_kMeans.htm. 2006.

[TD05] T. J. Trull, C. A. Durrett. Categorical and dimensional models of personality disorder. Annual Review of Clinical Psychology. 1: 355-380. 2005.

[TJP03] A. Topchy, A.K. Jain, W. Punch. A Mixture Model of Clustering Ensembles. SIAM International Conference on Data Mining. 2003.

[TLJF04] A. Topchy, M. H. Law, A. K. Jain, A. Fred. Analysis of Consensus Partition in Cluster Ensemble. In Proceedings of the Fourth IEEE International Conference on Data Mining. 225-232. 2004.

[TRD98] P. Tanguay, J. Robertson, A. Derrick. A dimensional classification of autism spectrum disorder by social communication domains. Journal of the American Academy of Child and Adolescent Psychiatry. 37: 271–277. 1998.

[TSK05] P. N. Tan, M. Steinbach, V. Kumar. Introduction to Data Mining. Pearson Addison Wesley. 2005.

[TWBB01] R. Tibshirani, G. Walther, D. Botstein, P. Brown. Cluster validation by prediction strength. Technical Report 2001-21. Statistics Department. Stanford University. 2001.

[TWH01] R. Tibshirani, G. Walther, T. Hastie. Estimating the number of clusters in a dataset via the Gap statistic. Journal of the Royal Statistics Society, Series B. 63: 411–423. 2001.

[VCBH89] F. Volkmar, D. Cohen, J. Bregman, M. Hooks, J. Stevenson. An examination of social typologies in autism. Journal of the American Academy of Child and Adolescent Psychiatry. 28: 82–86. 1989.

[VGRR06] S. Verte, H. M. Geurts, H. Roeyers, Y. Rosseel, J. Oosterlaan, J. A. Sergeant**.** Can the Children's Communication Checklist differentiate autism spectrum subtypes? Autism. 10(3): 266-287. 2006.

[VLKK04] F. Volkmar, C. Lord, A. Klin, A. Klin.Autismand pervasive developmentaldisorders. Journal of Child Psychology and Psychiatry 45(1): 135-170. 2004.

[VSH05] W.A. Van Der Kloot, A. M. J. Spaans, W. J. Heiser. Instability of hierarchical cluster analysis due to input order of the data: The PermuCLUSTER solution. Psychological Methods, 10(4): 468-476. 2005.

[War63] J. H. Ward. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association. 58: 236-244. 1963.

[WE05] T. Williams Jr., D. Eaves. Factor analysis of the pervasive developmental Disorders rating scale with teacher ratings of students with autistic disorder. Psychology in the Schools. 42(2): 207-216. 2005.

[WF05] I. H. Witten, E. Frank. Data Mining: Practical machine learning tools and techniques. 2nd Edition. Morgan Kaufmann. 2005.

[WG79] L. Wing, J. Gould. Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. Journal of Autism and Developmental Disorders. 9: 11-29. 1979.

[WH03] A. B. Wilcox, G. Hripcsak. The Role of Domain Knowledge in Automating Medical Text Report Classification. Journal of The American Medical Informatics Association, 10(4): 330-338. 2003.

[WHO92] World Health Organization. International classification of diseases, tenth revision (ICD-10). World Health Organization. 1992.

[Wil45] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics. 1: 80-83. 1945.

[Win93] L. Wing. The Definition and Prevalence of Autism: A Review. European Child and Adolescent Psychiatry. 2(2): 61-74. 1993.

[WMAD96] L. Waterhouse, R. Morris, D. Allen, M. Dunn, D. Fein, C. Feinstein, I. Rapin, L.Wing. Diagnosis and classification in autism. Journal of Autism and Developmental Disorders. 26: 59-85. 1996.

[Wu04] F. Wu. Computational Methods for Analysis and Modeling of Time-Course Gene Expression Data. Ph. D. Thesis. University of Saskatchewan. 2004.

[YSF94] N. Yrmiya, M. Sigman, B. J. Freeman. Comparison between Diagnostic Instruments for Identifying High-Functioning Children with Autism. Journal of Autism and Developmental Disorders. 24: 281-291. 1994.

Appendix A

DSM-IV Definition of the Autistic Spectrum Disorders

The full diagnostic criteria defined by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) for Pervasive Development Disorders are outlined below. They are available on many websites about PDDs. The following content is copied from http://www.autismsocietyofwa.org/DSMIV.html in September, 2007.

DSM-IV Criteria, Pervasive Developmental Disorders

299.00 Autistic disorder

A. A total of six (or more) items from (1), (2), and (3), with at least two from (1), and one each from (2) and (3):
  (1) qualitative impairment in social interaction, as manifested by at least two of the following:
    (a) marked impairment in the use of multiple nonverbal behaviours, such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction.
    (b) failure to develop peer relationships appropriate to developmental level.
    (c) a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest).
    (d) lack of social or emotional reciprocity.
  (2) qualitative impairments in communication, as manifested by at least one of the following:
    (a) delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gesture or mime).
    (b) in individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others.
    (c) stereotyped and repetitive use of language or idiosyncratic language.
    (d) lack of varied, spontaneous make-believe play or social imitative play appropriate to developmental level.
  (3) restricted, repetitive, and stereotyped patterns of behaviour, interests, and activities as manifested by at least one of the following:
    (a) encompassing preoccupation with one or more stereotyped and restricted patterns of interest that is abnormal either in intensity or focus.
    (b) apparently inflexible adherence to specific, nonfunctional routines or rituals
    (c) stereotyped and repetitive motor mannerisms (e.g., hand or finger flapping or twisting or complex whole-body movements).
    (d) persistent preoccupation with parts of objects.

B. Delays or abnormal functioning in at least one of the following areas, with onset prior to age 3 years:

(1) social interaction.
(2) language as used in social communication.
(3) symbolic or imaginative play.

C. The disturbance is not better accounted for by Rett's disorder or childhood disintegrative disorder.

299.80 Pervasive developmental disorder, not otherwise specified

This category should be used when there is a severe and pervasive impairment in the development of reciprocal social interaction or verbal and nonverbal communication skills, or when stereotyped behaviour, interests, and activities are present, but the criteria are not met for a specific pervasive developmental disorder, schizophrenia, schizotypal personality disorder, or avoidant personality disorder. For example, this category includes "atypical autism" --presentations that do not meet the criteria for autistic disorder because of late age of onset, atypical symptomatology, or subthreshold symptomatology, or all of these.

299.80 Asperger's disorder

A. Qualitative impairment in social interaction, as manifested by at least two of the following:
(1) marked impairment in the use of multiple nonverbal behaviours, such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction.
(2) failure to develop peer relationships appropriate to developmental level.
(3) a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest to other people).
(4) lack of social or emotional reciprocity.

B. Restricted, repetitive, and stereotyped patterns of behaviour, interests, and activities, as manifested by at least one of the following:
(1) encompassing preoccupation with one or more stereotyped and restricted patterns of interest that is abnormal either in intensity or focus.
(2) apparently inflexible adherence to specific, nonfunctional routines or rituals.
(3) stereotyped and repetitive motor mannerisms (e.g., hand or finger flapping or twisting, or complex whole-body movements).
(4) persistent preoccupation with parts of objects.

C. The disturbance causes clinically significant impairment in social, occupational, or other important areas of functioning.

D. There is no clinically significant general delay in language (e.g., single words used by age 2 years, communicative phrases used by age 3 years).

E. There is no clinically significant delay in cognitive development or in the development of age-appropriate self-help skills, adaptive behaviour (other than in social interaction), and curiosity about the environment in childhood.

F. Criteria are not met for another specific pervasive developmental disorder or schizophrenia.

299.80 Rett's disorder

A. All of the following:
   (1) apparently normal prenatal and perinatal development.
   (2) apparently normal psychomotor development through the first 5 months after birth.
   (3) normal head circumference at birth.
B. Onset of all of the following after the period of normal development:
   (1) deceleration of head growth between ages 5 and 48 months.
   (2) loss of previously acquired purposeful hand skills between ages 5 and 30 months with the subsequent development of stereotyped hand movements (i.e., hand-wringing or hand washing).
   (3) loss of social engagement early in the course (although often social interaction develops later).
   (4) appearance of poorly coordinated gait or trunk movements.
   (5) severely impaired expressive and receptive language development with severe psychomotor retardation.

299.10 Childhood disintegrative disorder

A. Apparently normal development for at least the first 2 years after birth as manifested by the presence of age-appropriate verbal and nonverbal communication, social relationships, play, and adaptive behaviour.
B. Clinically significant loss of previously acquired skills (before age 10 years) in at least two of the following areas:
   (1) expressive or receptive language.
   (2) social skills or adaptive behaviour.
   (3) bowel or bladder control.
   (4) play.
   (5) motor skills.
C. Abnormalities of functioning in at least two of the following areas:
   (1) qualitative impairment in social interaction (e.g., impairment in nonverbal behaviours, failure to develop peer relationships, lack of social or emotional reciprocity).
   (2) qualitative impairments in communication (e.g., delay or lack of spoken language, inability to initiate or sustain a conversation, stereotyped and repetitive use of language, lack of varied make-believe play).
   (3) restricted, repetitive, and stereotyped patterns of behaviour, interests, and activities, including motor stereotypies and mannerisms.
D. The disturbance is not better accounted for by another specific pervasive developmental disorder or by schizophrenia.

Appendix B

Features extracted from the ADI-R questionnaire

Below are the 22 features we use in this study. The 22 items in bold face are the features. The items listed under each feature are the questions in the ADI-R that pertain to the feature.

1. Early onset of symptoms

2. Age when parents first noticed that something is not quite right in language, relationships or behaviour
4. Onset as perceived with hindsight
86. Age when abnormality first evident
87. Interviewer's judgment on age when developmental abnormalities probably first manifest (code in months)

2. Impaired early development

5. Walked unaided
6. Acquisition of bladder control: daytime
7. Acquisition of bladder control: night
8. Acquisition of bowel control

3. Late acquisition age of language

9. Age of first single words
10. Age of first phrases

4. Abnormal conversational interchange

34. Social vocalization/ "chat"
35. Reciprocal conversation (at whatever verbal level of complexity possible)

5. Stereotyped speech

33. Stereotyped utterances and delayed echolalia
36. Inappropriate questions or statements
37. Pronominal reversal
38. Neologisms/idiosyncratic language

6. Impaired receptive communication

29. Comprehension of simple language

7. Impaired gesture expressive communication

42. Pointing to express interest
43. Nodding
44. Head shaking
45. Conventional/ instrumental gestures

8. Impaired behaviours to regulate social interaction

50. Direct gaze
51. Social smiling
57. Appropriateness of social responses

9. Poor peer relationships

49. Imaginative play with peers
62. Interest in children
63. Response to approaches of other children

10. Impaired shared enjoyment

52. Range of facial expressions used to communicate
53. Offering to share
54. Seeking to share his/her enjoyment with others

11. Impaired socioemotional reciprocity

31. Use of other's body to communicate
55. Offers comfort
56. Quality of social overtures
58. Inappropriate facial expressions
59. Appropriateness of social responses

12. Impaired social development

46. Attention to voice
47. Spontaneous imitation of actions
48. Imaginative play
61. Imitative social play

13. Lack of initiation of activities

60. Initiation of appropriate activities

14. Encompassing preoccupation

67. Unusual preoccupations

76. Unusual attachment to objects

15. Stereotyped motor mannerisms

77. Hand and finger mannerisms
78. Other complex mannerisms or stereotyped body movements (do not include isolated rocking)

16. Ritualistic behaviour

39. Verbal rituals
70. Compulsions/rituals

17. Sensory issues

69. Repetitive use of objects or interest in parts of objects
71. Unusual sensory interests
72. Undue general sensitivity to noise
73. Abnormal idiosyncratic negative response to specific sensory stimuli

18. Adherence to routine

74. Difficulties with minor changes in subject's own routines or personal environment
75. Resistance to trivial changes in the environment (not directly affecting the subject)

19. Symptoms of Rett's disorder

79. Mid-line hand movements
84. Hyperventilation

20. Aggression

81. Aggression to caregivers or family members
82. Aggression to non-caregivers or non-family members
83. Self injury

21. Epilepsy

85. Faints/fits/blackouts

22. Demonstration of savant skills

88. Visual-spatial ability (i.e. in puzzles, jigsaws, shapes, patterns, etc.)
89. Memory skill (accurate memory for detail, as of dates or timetables)
90. Musical ability (recognition, composition, absolute pitch or performance)
91. Drawing skill (unusually skilled use of perspective or creative approach)

92. Reading ability (e.g. early sight reading)
93. Computational ability (e.g. mental arithmetic)

Appendix C

Pipeline Diagram