Department of Computer Science
Series of Publications A
Report A-2012-2

Mixture Model Clustering in the Analysis of Complex Diseases

Jaana Wessman

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XIV, University of Helsinki Main Building, on 13 April 2012 at noon.

University of Helsinki
Finland

Supervisor

Heikki Mannila, Department of Information and Computer Science and Helsinki Institute of Information Technology, Aalto University, Finland and Leena Peltonen (died 11th March 2010), University of Helsinki and National Public Health Institute, Finland

Pre-examiners

Sampsa Hautaniemi, Docent, Institute of Biomedicine, University of Helsinki, Finland
Martti Juhola, Professor, Department of Computer Science, University of Tampere, Finland

Opponent

Tapio Elomaa, Professor, Department of Software Systems, Tampere University of Technology, Finland

Custos

Hannu Toivonen, Professor, Department of Computer Science, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.helsinki.fi
URL: http://www.cs.Helsinki.fi/
Telephone: +358 9 1911, telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-7897-2 (paperback)
ISBN 978-952-10-7898-9 (PDF)
Computing Reviews (1998) Classification: I.5.3, J.2
Helsinki 2012
Unigrafia

Mixture Model Clustering in the Analysis of Complex Diseases

Jaana Wessman

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
Jaana.Wessman@iki.fi

PhD Thesis, Series of Publications A, Report A-2012-2
Helsinki, Mar 2012, 119+11 pages
ISSN 1238-8645
ISBN 978-952-10-7897-2 (paperback)
ISBN 978-952-10-7898-9 (PDF)

Abstract

The topic of this thesis is the analysis of complex diseases, and specifically the use of *k*-means and mixture modeling based clustering methods to do it.

We concern ourselves mostly with the modeling of complex phenotypes of diseases: the symptoms and signs of diseases, and the other multiple co-phenotypes that go with them. The two related questions we seek answers for are: 1) how can we use these clustering methods to summarize the complex, multivariate phenotype data, for example to be used as a simple phenotype in genetic analyses and 2) how can we use these clustering methods to find subgroups of sufferers of a particular disease, such that might share the same causal factors of the disease.

Current methods for studies on medical genetics ideally call for a single or at most handful of univariate phenotypes to be compared to genetic markers. Multidimensional phenotypes cannot be handled by the standard methods, and treating each variable as independent and testing one hundred phenotypes with unclear true dependency structure against thousands of markers results into problems with both running times and multiple testing correction. In this work, clustering is utilized to summarize a multi-dimensional phenotype into something that can then be used in association studies of both genetic and other type of potential causes.

I describe a clustering process and some clustering methods used in this work, with comments on practical issues and references to the relevant literature. After some experiments on artificial data to gain insight to the properties of these methods, I present four case-studies on real data, highlighting both ways to succesfully use these methods and problems that can arise in the process.

Computing Reviews (1998) Categories and Subject Descriptors:

I.5.3 [Pattern Recognition]: Clustering
J.2 [Applications]: Life and Medical Sciences—medical genetics, psychiatry

General Terms:

Clustering, Experimentation, Applications

Additional Key Words and Phrases:

Medical genetics, Psychiatry

Acknowledgements

I have been extremely lucky in being able to work with not just one but two world-class academy professors as my supervisors. I cannot properly express my gratitude to professor Heikki Mannila at the Department of Computer Science and the late professor Leena Peltonen-Palotie at the Department of Medical Genetics for all their encouragement and support. I also wish to thank Drs. Tiina Paunio and Mikko Koivisto, whose supervisory role in shaping this work has been crucial.

Drs. Pauli Miettinen and Stefan Schönauer provided valuable comments on the semi-final draft. My pre-examiners Sampsa Hautamäki and Martti Juhola did a thorough job and without them, this work would be of a much lower quality. Any errors that remain are definitely not their fault, but that of the author.

Everyone I have come to contact with during this multi-disciplinary work has been enthusiastic, supportive, and willing to share their expertise in their own fields. Especial mention needs to go to all my co-authors in the schizophrenia paper; to Dr. Verneri Anttila and professor Aarno Palotie for their dedication to finding out what was going on with the migraine data; and to Dr. Nelson Freimer and his group in the University of California for collaboration and hospitality during the temperament work.

Many fellow students shared this long road. To mention just a few, I want to thank MSc Jukka Kohonen and Drs. Matti Kääriäinen, Niina Haiminen, Pauli Miettinen, Jussi Kollin, Pekka Parviainen and Esa Junttila for both the friendship and the science. In addition to students in the same department, Drs. Marylka Uusisaari and Laura Uusitalo, MSc Suvi Ravela, and my sister Dr. Jenni Antikainen, among others, have been able to share my ups and downs without sharing my fields of sience.

My parents Airi and Pentti Antikainen have always encouraged me to study and learn and have been able to convince me it is fun, which is one of the best heritages one can give to one's children. My husband Petri Wessman made this work possible by providing not only encouragement and financial stability, but also all of that which really matters in life.

Notation and abbreviations

$\mathbf{1}_x$	function with value 1 if x is true and 0 otherwise
A, B	datasets
a, b	arbitrary values, functions, or random variables, as specified in the text
C, D	partitions of a set of observations, or equivalently, clusterings
d	number of variables in a dataset, or, equivalently, number of columns in a dataset, or, as a result, number of dimensions in a model
$E[a]$	expected value of a
$E[a\|b]$	expected value of a given b
F	a set of functions
f	any function, as specified in the text
g	counter used for clusters in a clustering model, or groups in a population, or components in a mixture model
H	entropy
i, j	indices
K	the larger one of the values k for two alternative clusterings
K'	the smaller one of the values k for two alternative clusterings
k	number of clusters in a clustering model, subgroups in a population, or components in a mixture model
L	likelihood
M	a clustering model
N	number of observations (individuals) in a data set, or, equivalently, number of rows in a data matrix
N_{ab}	number of pairs of observations satisfying certain conditions, as specified in the text
n	used for any integer, as specified in the text
n_a^C	number of observations having cluster label a in clustering C

$n_{a,b}$	number of observations having cluster label a in one clustering and cluster label b in another
o	number of parameters in a model
$p(a)$	probability of a
$p(a,b)$	joint probability of a and b
$p(a\|b)$	probability of a given b.
$S(C,D)$	a similarity function between clusterings C and D
T	arbitrary time period
t	counter for iterations in an algorithm
$a^{(t)}$	value of a on the t'th iteration of an algorithm
v	number of partitions in a cross-validation scheme
Y	a data matrix, or equivalently a data set of size N individuals × d variables
$\mathbf{y}_j$	the j'th row in a datamatrix Y, or, equivalently, the j'th individual in a data set
$\mathbf{y}_{ji}$	the ith element of y_j, or equivalently, the value of the i'th variable for the j'th individual
$\mathbf{y}_{j,obs}$	the observed (non-missing) values in $\mathbf{y}_j$
$\mathbf{y}_{\cdot l}$	the l'th column (variable) for any row (individual) in Y
z_{gj}	class label, taking value 1 if the j'th individual belongs to the g'th cluster / subgroup / mixture component
$\hat{z}_{gj}$	an expected value estimate of z_{gj}
$\mathbf{z}_j$	a class label vector for the j'th observation/individual
$\hat{\mathbf{z}}_j$	an expected value estimate of $\mathbf{z}_j$

θ	parameter vector for any model, as specified in the text
κ	used for Cohen's κ, a measure of agreement between two classifications
$\boldsymbol{\mu}$	the mean, either the vector of a multivariate Gaussian distribution, or equivalently, the mean of observations in a particular cluster
π_g	mixing proportion of the g'th component of a mixture model or g'th subgroup in a population, or, equivalently, the cluster propability of the g'th cluster
$\mathbf{\Sigma}$	the covariance matrix of a multivariate Gaussian distribution
ϕ	the likelihood function of a Gaussian distribution

ADI-R	Autism Diagnostic Interview - Revised
AMI	Adjusted Mutual Information
BIC	Bayesian information criterion

$C\kappa$	Cohen's κ
DISC1	a candidate gene for psychotic disorders (Disrupted in Schizophrenia 1)
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition
DTNBP1	a candidate gene for schizophrenia, also known as dysbindin 1 (Dystrobrevin Binding Protein 1)
EM	Expectation-Maximization (algorithm)
HA	Harm Avoidance, a scale in TCI
ID	identification number/code for an individual in a dataset
JI	Jaccard Index
MH	Meila H-index
MI	Mutual Information
NFBC1966	Northern Finland Birth Cohort 1966
NMI	Normalized Mutual Information
NS	Novelty Seeking, a scale in TCI
RD	Reward Dependency, a scale in TCI
P	Persistence, a scale in TCI
PC	Pairwise Concordance (Rand Index)
PCA	Principal Component Analysis
SD	Standard Deviation
TCI	Temperament and Character Inventory
YF	The Cardiovascular Risk in Young Finns study

Contents

Chapter 1

Introduction

> *"No catalog of techniques can convey the willingness to look for what can be seen, whether or not anticipated."*
> (John W. Tukey)

In this thesis, we describe the use of clustering methods in the analysis of complex diseases. Specifically, we concentrate on mixture model clustering, including the special case of k-means where suitable, and using them to summarize the complex phenotypes and co-phenotypes of these diseases or phenomena. Such summarizations can be of help when looking for causal factors (genetic or otherwise) for the phenomena.

This work is in the intersection of computational methods for data analysis and the medical science of etiology. In addition, the practical field of computer programming is necessary to implement the procedures described. All this combined makes for a large field, and thus it has been necessary to restict ourselves to a particular clustering method (mixture model clustering), as well as to not delve very deeply into any particular field of medicine. We hope, however, that even with these restrictions the description of the work performed here will also give general insights to a practical clustering process in the study of complex diseases.

This thesis has been written mainly with a computer science audience in mind; basic programming skills and reading skills of mathematical notation are assumed, and medical information is kept on a fairly basic level. The author has, however, also attempted to make the thesis readable for the medical reseacher audience.

1.1 Complex diseases

The definition of “a disease” is a matter of some debate in itself [Ems87]. For the purposes of this work, we define disease as a condition of an organism that

1. is considered abnormal,
2. causes impairments of bodily (including mental) functions,
3. follows from a specific set of causes, and
4. is associated (though not necessarily deterministically) with specific symptoms and signs.

Etiology, the study of origins of diseases, is concerned in defining diseases such that the causes[1] and the probabilities of symptoms and signs are known. Understanding the etiology of diseases is the key to alleviating suffering caused by them: we can prevent a disease by breaking the causal chain leading into it and we can cure a disease by removing a cause upholding it. When we cannot do either, easing symptoms can be more feasible when we understands their mechanisms, and often simply understanding what is happening and what to expect alleviates the mental suffering associated with diseases.

Many syndromes (collections of symptoms that seem to go together) that we think of as diseases are not diseases in the sense of the definition given above. For example “the common cold” is a collection of diseases, each caused by a separate microbiological entity [SM97]. On the other hand, sometimes different diseases are related to overlapping causes. Again as a simple example, over-consumption of alcohol is a causative agent of several possible complications, ranging from a common hangover to liver disease and even having consequences to the next generation in the form of fetal alcohol syndrome and the psychological consequences of being raised up by alcoholic parents [SM97]. In addition, various factors can alter the probabilities of particular symptoms in diseases ultimately caused by the same causes: sometimes we say that there exists two (or more) forms or *subtypes* of the same disease, when most causes and symptoms are the same, but a minor variation in the causative environment affects the exact expression of the symptoms [SM97].

[1] I use “cause” here in a broad way to mean both necessary and sufficient causes, and also a collection of factors that increase the probability of a disease. The philosophical concept of causation in disease is way beyond the scope of this work.

The fact that similar symptoms are often caused by different causes, and similar causes sometimes cause different symptoms, obviously greatly interferes with studies of etiology. In medical genetics, a disease is called *complex*, where it seems likely or clear that it is unlikely to follow a clear Mendelian inheritance pattern [Hun05]. Currently, it seems that such simple diseases are actually the exception rather than the rule: most diseases are caused by several mutations and even more environmental factors together, and some things that we think of as the same disease might actually be caused by two or more separate mutations that lead to a similar disturbance in the body, independently.

Genetics that relies on tracking established diagnoses or single symptoms will generally fail to establish the genetic etiology of such diseases, due to such studies requiring a much larger sample size than the more simple cases [Hun05]. Regardless, many current genetic analyses methods expect a single phenotype, the associations of which to the genetic markers under consideration are then studied, and for practical reasons, medical dataset sizes have an upper limit in the order of thousands, at most tens of thousands individuals. Hence, the initial purpose of this work: exploring one potential way to build from symptoms, signs, and other observations of individuals new subtypes for genetic analysis, in the hopes that these phenotypically homogeneous subtypes would correlate better with subgroups of syndromes with similar (genetic) etiology.

1.2 Nature of the data

The data we are working on typically concerns individuals. For each individual, ideally, the same variables have been measured, resulting in a data matrix. Sizes of datasets described in this study count individuals in thousands and variables in tens or low hundreds. Individual variables can be of any type: binary (for example presence or absence of a symptom), class-valued (types of symptoms of background information), ordered (answers to questionnaire items on a scale from strongly agree to strongly disagree), or continuous (age, various blood tests).

The information can come from several different sources. We can separate these sources, roughly, into three: self-reported, register-based, and measured data. Self-reported data is, obviously, information that the individual gives of him/herself. Register-based data is obtained (with the individual's permission) from various national registries such as the Hospital Discharge Registry utilized in this study. Measured data is data that has somehow been measured and confirmed for this study in particular. It can include physical examination data (blood tests, measurements performed by a medical professional), various structured ways of interviewing the patient (by trained and controlled interviewers), and variables constructed from case notes in a systematic way.

In addition to different sources, the data can concern different timelines. Some variables relate to the patient's status "now", others relate to his or her past history, or even his or her parents' or ancestors' history. In addition to this, data about history might have been obtained at different times: retrospectively, or in a follow-up during the time when it was current.

The way individuals are *recruited* to the study has effects on the data. All studies begin with identifying some sort of group of interest, be it individuals with a disease, individuals belonging to families with the disease, members of a population, inhabitants of a region, or something else. Then this group, or a random sample there of, is contacted and an attempt to recruit them for the study is made. Obviously, the way the group is identified in the first case and the response rate to recruitment affects whether the study actually contains a sample from the population originally under study, or some subpopulation there of. For example, it is very typical that the individuals worst affected by a given disease are not in the shape to respond, thus eliminating extreme cases of disease from the data.

All these make the technically simple data matrix actually quite a complex structure. Sources, timing, and recruitment all cause their own biases in the data and affect the reliability of the variables.

1.3 The role of clustering in genetics

Current methods for studies on medical genetics ideally call for a single or at most handful of univariate phenotypes to be compared to genetic markers. Multidimensional phenotypes cannot be handled by the standard methods, and treating each variable as independent and testing one hundred phenotypes with unclear true dependency structure against thousands of markers results into problems with both running times and multiple testing corrections.

When the obvious phenotypes (such as diagnoses) have been tested for associations with the markers and a suspicion remains that they do not capture all information about the causative links between disease and genes, researches typically want to look at phenotypes that are more directly or more strongly associated to certain genes. Such groups can be so-called endophenotypes: directly genetically associated phenotypes that predispose to disease [GG03]. Alternatively, they can be a redefinition of a diagnosis to weed out “noise” to find a core group of patients with a more similar disease [HKB+05].

In the typical case, we do not have before-hand information on what these endophenotypes or relevant subgroups might be. If we did, they could in many cases be measured or constructed directly (at least as well as the original diagnosis can be defined). The need to look for these phenotypes arises when it seems likely that the analysis of the etiology of a particular syndrome is confounded by the existence of multiple causal factors, both genetic and otherwise, but we do not exactly (or at all) understand how.

Constructing this kind of alternative phenotypes means summarizing often multidimensional data in novel ways, and has often been done manually with a domain specialist with a good “hunch.” Thus the question is or can be translated as “are there subgroups in this data that we are not yet aware of”. From the point of view of data analysis or machine learning this question naturally translates as the problem of clustering, which can be roughly defined as the unsupervised learning question of “division of a set of objects into subgroups such that objects in the same group are similar to each other, while being as different as possible from objects in other groups”.

In the studies presented in this thesis, clustering is utilized in summarizing a multi-dimensional phenotype into something that can then be used in association studies of both genetic and other type of potential causes. The work presented is by nature exploratory, in the sense that its purpose is to discover hypotheses that can then be tested by conventional statistical means, or finding questions that can be answered by further studies. Such exploration requires a different thinking than confirmatory data analysis – though not less care to be taken to be aware of biases possibly introduced. John W. Tukey, in 1980 [Tuk80], wrote:

> “If we need a short suggestion of what exploratory data analysis is, I would suggest that
>
> 1. It is an attitude, AND
> 2. A flexibility, AND
> 3. Some graph paper (or transparencies, or both).
>
> No catalog of techniques can convey the willingness to look for what can be seen, whether or not anticipated. The graph paper [—] not as a technique, but rather as a recognition that the picture-examining eye is the best finder we have of the wholly unanticipated.”

Computerized methods have allowed us to “see” some things that weer impossible to see with just graph paper and transparencies, but the principle still holds. While the approaches can (and should) borrow methods from each other, if one does not differentiate in one’s mind clearly between exploration and confirmation, the temptation arises to do both on one go: to first seek for a hypothesis, and then seeing it in the data at hand, to “confirm” it by hypothesis testing in the same. This leads to a sort of circular reasoning and indeed a rigorous confirmation of a hypothesis would require a separate dataset.

For example, the schizophrenia study presented in Section 4.1 was started as exploration of subgroups of the disease, and ended with the suggestion that individuals with the disease might have different genetic background depending on the presence of mood symptoms[WPTH$^+$09]. We could not have arrived to that suggestion by performing a study that would have required us to predefine exactly what we were looking for, but on the other hand this study alone cannot conclusively show that what is suggested is the case.

1.4 Overview of this work

In the work that lead to this book, I or co-workers have performed clustering studies of four medical datasets from the Finnish population: 1) schizophrenia patients and their relatives, 2) migraine sufferers from families with several migraine cases, 3) children with autism spectrum disorders and healthy controls, 4) a population sample of individuals assessing the associations of temperament and various lifestyle and health measurements.

In addition, during the course of this work, I have performed various experiments on artificial data as “sanity checks” for how well the selected cluster scoring, validation, and replication techniques perform.

Based on these, I describe a practical process for clustering, from preprocessing to postprocess visualization and statistical analysis, attempting to guarantee that the above three points have been taken into account. Matlab code implementing the parts of the process can be provided by the author.[2]

The process includes:

- preparatory analyses to familiarize the researcher with the data, identify features that suggest mistakes in the data (outliers, non-random

[2] One should not expect to take this code and simply run it on their data, however. For reasons that will become apparent, a lot of such code is data-dependent. Performing this kind of studies without at least one person who can program would be madness.

patterns of missing data, illogical distributions of variables), help to select variables, and to decide on missing-data handling procedures

- a mixture-model clustering process (though this can be easily, and has in one of the studies been, replaced by another method)
- scores for selecting the number of clusters
- randomization-based analyses to ensure the stability and validity of the clustering
- visualizations and statistical analyses to present the clustering to a domain specialist.

This thesis is organized as follows:

Chapter 2 first gives a comprehensive description of the clustering methods and techniques used in this work, together with references to relevant literature. We begin by describing the pre-processing stage of data cleaning. Especial attention is given to typical features of medical data, such as the role of demographic and diagnostic information, and the usually fairly large amount of missing data.

The basics of mixture model clustering and its special case the k-means clustering method are then explained. As missing data is a concern in most medical datasets, attention is paid for how to handle it. We describe general solutions to the problem, and give an overview for how to incorporate missing data handling into the mixture model clustering methods. Methods for selecting the number of clusters and to compare clusterings are described. Again, we first give a general overview and then describe in more detail the methods used in this work, namely v-fold cross-validation and the Bayesian information criterion score for cluster number selection, and pairwise concordance, adjusted mutual information, and Cohen's κ for clustering comparisons.

Finally, ways to analyze the quality of clustering are discussed. We separate this process into two questions: whether the clusters are real, and whether they are interesting. For the first question, we describe the concept of cluster stability on various levels of the process, and propose the procedure of randomly dropping individuals and variables for to assess it. We also recommend replication in a separate dataset whenever possible. For interestingness of the clustering, we describe some simple summarization and visualization procedures, as well as practical considerations of working with experts from fields other than our own.

Chapter 3 describes some original experiments on the behaviour of the described algorithms on artificial data. The artificial data was generated using a model similar to that used in clustering, but with added noise and missing data. We perform tens of test runs of the algorithm in various conditions. The tests reported include:

1. comparison of Bayesian Information Criterion and cross-validation as methods for cluster number selection, demonstrating that for realistic N both give acceptably good results;

2. experiments on the observation of ”natural hierarchies”: in the presence of a cluster structure in the data the clusters observed for different cluster number tend to form a hierarchical structure even for non-hierarchical methods;

3. experiments on replication of clusterings in a separate data sample, confirming that doing so can in many circumstances not only validate our prior clustering, but also indicate the lack of a clear cluster structure in the data;

4. a study on the effects of missing data, giving some insight into how much of the data can be unobserved for this kind of methods to still work; and

5. an experiment confirming that the method of randomly dropping data rows to explore cluster stability does produce reliable results, at least when model assumptions are somewhat met, even in the presence of originally missing data and noise.

Chapter 4 describes the original real-data studies, successes and failures, together with medical and methodological lessons learned. As our prime success story, we present a schizophrenia study where clustering was able to shed light on controversial results in medical genetics. In this study, Finnish individuals from families with individuals with schizophrenia were clustered, and the resulting clusters used as an alternative grouping in an association analysis of genetic markers in known candidate genes for schizophrenia. This study demonstrates that clustering can reveal groups with a more homogenous causal background and thus aid in detecting, for example, the genes involved.

As another succesful example, we present a clustering of sample of Finnish population into temperament groups. Here, the clustering is based on a questionnaire of adult temperament, and we show striking associations of these clusters to a wide variety of variables about health, lifestyle, and social status. This study demonstrates that sometimes clustering can simplify a multidimensional characteristic of individuals while keeping intact all or almost all of the associations of dimensions to relevant medical variables. Here we also demonstrate the use of a second sample to replicate the clustering results.

We also present two cautionary stories about where clustering works suboptimally: a migraine study where missing data proved to be a problem, and a study on autism where cluster structure was not discovered. While naturally of much less medical interest, from the computer science point of view these stories are at least of equal importance to the previous two, demonstrating the shortcomings and pitfalls of these methods.

The major contributions of this work are, besides the practical real-data studies (Chapter 4), the practical experience gained for clustering studies (Chapter 2) and the insights gained on simulated data into some hands-on details of clustering algorithm behavior (Chapter 3). All simulations presented in Chapter 3 were performed and reported by the author alone. In the studies in Chapter 4, the author has performed all clustering, validations, and genetic analyses for the schizophrenia study (Chapter 4.1), all clustering and validation involved in the migraine study (Chapter 4.3), and about half of the analyses in the temperament study (Chapter 4.2) together with co-author Stefan Schönauer, as well as taught the method and the validity analyses to and reviewed the results by co-author Ulrika Roine performing the clustering in the autism study (Chapter 4.4).

Chapter 2

The clustering process

> *"All models are false, but some are useful."*
> (George E. P. Box)

The main theme of this work is applying clustering methods to various medical datasets. This sort of work sits firmly in the overlap of various fields: theoretical computer science that describes the methods, the scientific field the data is applied to (referred to by the computer scientists as the "domain"), and the program engineering and practical data analysis skills needed to make those two meet. We limit ourselves here to the application of a particular family of clustering models (namely, mixture model clustering and its special case, k-means) to a particular domain (namely, that of certain fields of medicine). This removes us both from certain other fields of clustering familiar to medical researches (especially hierarchical methods) and some typical major domains familiar to clustering experts (market research, text classifications, gene expression), and hopefully provides some new insights to applicative computer science and the medical fields both.

When we are clustering data on a real-life medical field, we are generally never looking for "the real subgroups in the data". This is due to the simple fact that most of the time, the concept of "the real subgroups" is not realistic. Many ways to group the data meaningfully usually exist, each useful for different purposes. In the studies described in this thesis, we are looking for *a* subgroup structure that can tell us *something new* about the data or the domain. A good example of this are our results in the schizophrenia family data (described in detail in Chapter 4.1). A main result of that study is that when we draw the line between "psychosis in general" and "core schizophrenia in particular" differently, we find not that one categorization being better at detecting all associations, but that different categorizations reveal different etiological factors. Some candidate genes are associated to psychosis in general, and some to a very specific subset of schizophrenia. Neither of these categorizations is more “true” than the other, but which one should be used depends on the research question.

Classically, a clustering process is separated into three stages. The names used for these stages vary; in different textbooks they have been called, for example, “pre-processing, analysis, and post-processing” [HK06] or “exploration, model-building, and validation” [HL01]. In the first stage, the researcher first looks at the data, familiarizes herself with it, selects the variables to be used, performs necessary transformations on them, and runs initial tests to select the clustering methods and parameters to be used. Then, in the next stage, the method itself is applied on the selected data, a model is selected from among those produced by various parametrizations, and the validity of the model is studied. Finally, the models are analyzed, and scientific conclusions about the domain drawn. The process presented in this chapter follows these phases, too.

2.1 Preprocessing stage

2.1.1 Understanding the data

In this section we describe some of the steps in the preprocessing stage. This is a highly data-specific phase, and due to this reason we will rather informally present some observations from the studies described in Chapter 4, rather than attempt to give a full procedure and formal descriptions.

Before any data analysis, it is necessary to get familiar with the data enough to understand its features and peculiarities, to find possible biases, hidden dependencies, and other sources of error (see e.g. [HMS01, HK06]). To this end, the researcher should be aware of the basics of how the data at hand has been collected. This includes at least: how were the individuals sampled, when and where were various measurements obtained, how are the measurements coded (in what units or using which classes), what recodings have been performed on the variables (discretizations, normalizations, combining classes), what types of missing data are there, and how are these types designated in the data.

The most important distinction about the possible ways of recruiting individuals to the study, from a clustering point of view, is whether the data under study comes from a random sample from a population, from a case-control sample, or from some more complex design (for example from recruiting members of families or individuals from a particular region with a particular disease). Many statistical methods and clustering procedures have underlying assumptions of independence, which are violated in all but random sampling collection. This does not necessarily pose a problem for the clustering itself, but it can be crucial when interpreting the results.

Before more complex analysis, we then take a look at the variable descriptions of the data, including the ranges and possible values for each variable, as well as annotations to describe what the values mean. From this, the type—categorical, ordered, continuous—and range of each variable can be figured out. For the methods used in this thesis, it is simplest if all variables can be treated similarly, and either treated as continuous dimensions of a real space, or as unordered classes, even if this requires transformations. In many cases, however, this is not possible without making the transformations so artificial as not to be interpretable. At this point, we must also check that all variables actually match their description—meaning mostly, that there are no values other than the valid ones.

In preprocessing for clustering it is important to identify key demographic and data-specific variables that should not end up being the major determinants of the clustering. What these are depends on the exact application, but in most practical examples at least a running participant identification number in the study and row in the data matrix belong to this category. If data has been collected in several centers or phases, any variable identifying these will also be included in this set. Of variables related to the individual, age and sex typically belong to this category, as we are usually not interested in a clustering solution that reveals only such basic truths that old people are different from adolescents, or males from females. Depending on the application, also geographic location, ethnic group, level of education, or other such demographics might belong here.

In addition, we might want to identify a small set of (5, at most 10) variables of especial interest to be used in first-pass post-processing analyses for evaluating the interestingness of the clustering. These could be, for example, diagnoses or the most interesting symptoms. In some studies we have also opted to making the data analysts blind to diagnostic groups to begin with, to assure that we do not unconsciously steer the clustering process towards something that appeals to our prior understanding of the phenomenon.

On datasets where data missingness is expected, we can then proceed by looking at the patterns of missingness. Data missing at random is the exception, not the rule, in medicine. It is possible for data missingness to carry information [LR02], for example sometimes a “missing” value signifies that the variable cannot be recorded because it does not exist (if you do not have headaches, the severity of those headaches is not a meaningful concept). At this point, such missingness with information needs to be separated from really unknown data, for example by recoding it as a separate value.

Questions answered at this point include the following [LR02]. How many percent of data is missing per variable? Are there variables with more missing than recorded data? Is there a pattern to the percentages of missing data over the variables? For example, in questionnaire data, the later a question is on the questionnaire form, the more data is usually missing. Is there a pattern of these percentages over the individuals, and if so, is it related to some specific demographic variable? Any such discrepancies will then be gone over with a domain expert, preferably with the same people who provided the data.

After familiarizing ourselves with the missing data patterns, we then take a look at the distributions of each individual variable. Medical variables often have a natural minimum and/or maximum, and if the first sanity check over annotations did not already do this, any outliers beyond these are be recognized and either corrected by the domain experts or treated as missing data. These natural ranges are in the optimal case provided by the medical experts in charge of the data collection. Also values clearly beyond the typical range of values for that variable should be recognized at this point. How far away is “clearly” is not an easy question to answer, but a possible rule of thumb is that if the presence or absence of one value alone significantly changes the mean or variance of the variable, then that value needs to be removed.

Associations to the demographic and data-specific are then looked at. Any standard statistical test will do the job. If associations to arbitrary features of the data, such as running IDs or centers of data collection are found, they are reported to the domain experts before proceeding. If there is a small number of such variables, they can simply be dropped from the analysis, although we should obtain a good understanding of why such associations occur. Otherwise, there is the chance that other, more complex, associations with arbitrary features might go undetected. If there are many, however, sometimes the domain specialist can advice us that the association is natural and to be expected. For example, if cases tend to have a lower ID and controls a higher one, we can proceed to look for demographic effects for cases and controls separately, but ignore the general associations. Where such an explanation cannot be found, it can necessary to restrict the clustering into a subgroup of individuals, e.g. only to those from a particular data collection center, or to cluster groups separately.

If associations to demographic variables are found, the options to correct that include 1) adjusting the values of the associated variables by some standard way, 2) analyzing groups (e.g. males and females) separately, 3) dropping the variable completely, or 4) accepting the effect as inherent to the phenomenon and including the variable as is. All decisions to drop data, stratify analysis, or adjust variables, are made in communication with domain experts.

Once these basic considerations have been gone through, two major decisions are then made: first, which variables will be used for clustering, and second, which individuals will be included. In the studies presented here, we excluded from clustering any variables that directly code for diagnoses of interest. For example, in addition to the diagnosis itself, we would exclude the variable for case-control status in a case-control dataset. After all, in a clustering study we are usually not interested in replicating an existing classification scheme (which a diagnosis essentially is); if we were, we would be using classification methods instead.

Next, we want to check for redundant variables by looking at all pairwise correlations of variables. If two variables are identical or nearly identical (possibly apart from labeling or scale used), one of them can be dropped. If the missing data pattern for the variables is not identical, combining the two into a variable with less missing data than either of the original ones is also possible—which one to use for those individuals who have both recorded being, again, a domain expert's choice. Such highly correlating variables often result from features of the data collection process, for example the same thing having been measured twice on different visits to the clinic performing the studies or asked in separate parts of questionnaires, or sometimes from having been both measured and self-reported. (In the latter case, the difference between measurement and self-report can be an interesting variable in itself.)

Once all this is done, we divide the remaining variables into two parts: those to perform the clustering on, and those to use as comparison data for the clustering obtained. Sometimes, a clear division of phenotype variables to a clustering subset and comparison subset suggests itself. This, for example, is the case in our temperament clustering study described in Section 4.2, where the researchers were specifically interested in temperament groups and their associations to a large set of background variables, rather than clusters of that background. Sometimes we only leave out diagnoses and data specific variables (e.g., running IDs, collection center information). Sometimes to limit the amount of missing data we are forced to include variables with at least some cut-off percentage of non-missing data.

As to individuals, sometimes the medical interest lies in finding subgroups inside a particular diagnostic or demographic group, and the rest of the individuals can be excluded. For example if we are interested in subgroups of individuals with the disease, healthy controls can be ignored (though they can be included too to see if they form a separate cluster). It might also be necessary to exclude individuals who have been for some reason unable to participate fully (for example, individuals with mental retardation were excluded in the schizophrenia study described below). Other than that, the only exclusion criteria for individuals that we have considered is missing data. Medical datasets often include people who originally enrolled to the study, but did not show up for medical examinations or fill up the questionnaires sent to them. These individuals have most of their data beyond demographics and diagnoses missing, and thus do not provide useful information.

In the final dataset, as a rule of thumb, the number of variables should be a fraction of the number of individuals for most clustering algorithms to provide stable and meaningful results. If after removing redundant variables there still are more variables than what feels reasonable, or if a high number of variables used fails to provide a stable clustering, we can further prune the variables, starting from excluding variables with most missing data and/or those with a high correlation to another variable. Various dimension reduction techniques also could be used, but beyond simple combining of binary variables have not been utilized in the studies reported in this work. Such techniques have the downside of making the included variables harder to interpret. (One should note that in gene expression studies considerable progress has been made towards methods that are applicable even in the case when there are many more variables than observations, for example [Kii08]. However, in this study we do not tackle this issue.)

2.1.2 Selection of clustering method

For an overview of different clustering methods see, for example, [JMF99, HMS01] or [HK06]. It is not easy to suggest criteria for what clustering method should be used, beyond general guidelines of some methods being more suitable for continuous and some for class-labeled data. All methods come with some strengths combined with some assumptions, the violation of which can cause unexpected and, in the worst case, undetectable errors. As a principle, since clustering is by nature exploratory, it is crucial that the assumptions of the model are as explicit as possible and the results it produces are interpretable and understandable by the researches involved.

The data itself, obviously, poses some restrictions on the selection. For example, the k-means procedure [Mac67, Llo82] is widely spread and easily available, due to its being included in many (if not most) available software packages for this kind of analysis. Strictly speaking, the k-means procedure is applicable only when the data can be interpreted as points in some continuous (typically Euclidean) space, and when the amount of missing data is relatively small. Hierarchical methods [JMF99] also require a way to formulate a distance measure, and are best suitable for domains where the data points can be assumed to form a hierarchy (as, for example, genetic sequences can be assumed to do, based on evolution).

Mixture model methods [MP00, HJ03] require that a joint distribution given the group the individual belongs to can be formulated. This usually means that some explicit assumptions of distributions and independence between the variables have to be made. Mixture model methods combine nicely explicit assumptions with interpretability of the results, which is the reason for why they have been used in this thesis whenever possible, with the only “exception” of reverting to simple k-means (a special case of mixture models) when the data consists of continuous variables and is complete.

The selection of a clustering method might incur further needs of preprocessing. For example, for many distance measures, including the Euclidean distance, it is necessary to normalize or scale the variables so that one dimension will not span a much higher range than some other, inadvertently gaining more weight. As another example, the strong independence assumptions many models, for example the naïve-Bayes model (e.g. [DP97]) used in this work, might call for some way to combine highly correlated variables, to preserve some of the dependency structure of the original data.

Once the clustering method is selected, before proceeding any further, we need to specify the process in detail. This must be done in order to avoid problems with multiple testing. The process description should include at least

- which clustering method is to be used,
- which variables the clustering is based on, what preprocessing will be done on them, and how will missing values be treated,
- what score will be used for model selection among different parameters the process requires (most notably, number of clusters),
- how will cluster validity/stability be assessed, and
- under which conditions will groups (for example, males and females) be reclustered separately, or variables excluded from clustering.

Paradoxically, for the selection of the variables, the clustering method, and the method for missing value handling, it might be necessary to perform a couple of initial runs of the algorithms on the data under consideration and to assess the stability of the results, in order to avoid using huge amounts of energy into providing an unstable clustering. When this is done, we should avoid looking at the results other than stability before the final selection of methodology. Optimally the person performing this stage should be as blind as possible to data semantics; in the very least we should blind them to diagnosis or case/control status.

2.2 Mixture model clustering

2.2.1 Mixtures of distributions as clustering

In statistics, one very basic method of describing data is to make the assumption that the data comes from a certain model (say, is normally distributed), and then to look for the parameter estimation of that distribution that make the data best fit the model (or vice versa). Two groups of subjects can then be compared by comparing these parameters and calculating whether the differences are statistically significant or likely to have arisen by chance alone.

In fitting of mixtures of distributions, the underlying assumption is that the subjects come from a population of k groups with proportions $\pi_1, ..., \pi_k$ (summing to one). Each group has a similar distribution (say, the variables for each subject come from a multivariate normal distribution) but with different, unknown, parameters for each group. The probability of the observed values for a particular subject are defined on the unobserved group (termed the “latent class” in the classification context) of the subject. The task is then to simultaneously find the distribution parameters for each group, the mixing proportions, and the group of each subject such that the data fit to the model is maximized. [MP00]

In the case of distributions whose parameters can be found in closed form, and using the maximum likelihood setting as the definition for best fit, the task can be achieved for a given k with the Expectation-Maximization algorithm [DLR77, MP00]. The output for this algorithm is 1) the parameters of the distribution for each group, and 2) for each individual the probability of belonging to each group (summing up to one, naturally). These probabilities can then be used as a probabilistic (soft) clustering of the subjects, or, when a deterministic (hard) clustering is required (as often is the case for interpretability), the cluster of each subject can be taken to be the one with the highest probability. [MP00]

2.2.2 The Gaussian and Naïve Bayes models

Given an $N \times d$ data matrix Y in which all rows $\mathbf{y}_j$, $j = 1, ..., N$ correspond to a d-dimensional data vector describing one individual, the task is now to specify the mixture model in detail and to find the maximum likelihood estimate for it. Given a context where we assume each subject to be "really" produced from one of the components of the mixture, we can approach this as thinking the problem as a problem of estimation with missing data [DLR77, MP00, HJ03].

The probability distributions used in this work are 1) multivariate Gaussian distributions, and 2) the Naïve Bayes model with point distributions. The latter assumes every variable to be a class-valued one, and independent from all other variables given the class assignment (this independence assumption is why it is called "naïve", or sometimes "simple") [DP97].

In a finite mixture of k d-dimensional Gaussian distributions, denote for each $g = 1, ..., k$ the mixing proportions by π_g and the parameters by $\boldsymbol{\mu}_g$ (the mean vector) and $\boldsymbol{\Sigma}_g$ (the covariance matrix). The value of the probability distribution function for an observation $\mathbf{y}_j$ is [MP00]

$$f(\mathbf{y}_j|\theta) = \sum_{g=1}^{k} \pi_g \phi(\mathbf{y}_j|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2.1}$$

where $\theta = (\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the parameter vector of the model, containing the mixing proportions $\pi = \pi_1, ...\pi_k$ and the parameters $\boldsymbol{\mu} = \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_k$of the normal distributions, and ϕ is the probability distribution function of the multivariate Gaussian distribution:

$$\phi(y|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{(\mathbf{y}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mu)}{2}} \tag{2.2}$$

That is, as the groups in the data are mutually exclusive and exhaustive, the joint density for observing the values of $\mathbf{y}_j$ is the sum of the densities for observing the same in each group, weighted by the proportions of the groups. One can think of $\mathbf{y}_j$ having been sampled by first sampling one of the groups, and then sampling the values of $\mathbf{y}_j$ from the distribution with that group's parameters.

For each subject $j = 1, ..., N$, consider the class label $\mathbf{z}_j$: a k-dimensional binary vector, where $z_{gj} = 1$ or 0 according to whether the j'th subject came from the g'th group or not. These data are unknown, and to obtain a clustering, we want to estimate them together with the distribution parameters. For estimation purposes, we allow the estimates $\hat{z}_{gj}$ to have values *between* and including 0 and 1. In general, a case could be allowed to belong to several classes, but for the purposes of this work, we require $\mathbf{z}_j$ and the estimate $\hat{\mathbf{z}}_j$ to sum up to exactly one.

In a discrete Naïve Bayes model of k components and d variables, with the mixing proportions of π_g, denote by θ the collection of all the parameters of the model (mixing proportions and point probabilities $p(\mathbf{y}|z_j = 1)$) the probability distribution function at an observation $\mathbf{y}_j$ is

$$\begin{aligned} f(\mathbf{y}_j|\theta) &= \sum_{g=1}^{k} \pi_g P(\mathbf{y}|z_{gj} = 1) \qquad (2.3) \\ &= \sum_{g=1}^{k} (\pi_g \prod_{i=1}^{d} P(y_{ij}|z_{gj} = 1)) \end{aligned}$$

where $P(y_{ij}|\mathbf{z}_{gj} = 1)$ is the point probability of the i'th element in vector $\mathbf{y}$ given that the group assignment for the individual is g.

When we assume that all individuals are independent from each other, the value of the probability function for the whole data is simply the product of the probabilities for the individuals:

$$f(Y|\theta) = \prod_{j=1}^{N} f(\mathbf{y}_j|\theta) = \prod_{j=1}^{N} (\sum_{g=1}^{k} \pi_g \phi(\mathbf{y}_j|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)) \qquad (2.4)$$

for the Gaussian and

$$f(Y|\theta) = \prod_{j=1}^{N} f(\mathbf{y}_j|\theta) = \prod_{j=1}^{N} (\sum_{g=1}^{k} (\pi_g \prod_{i=1}^{d} P(y_{ji}|z_{gj} = 1)))$$

for the discrete Naïve Bayes case.

2.2.3 Expectation-Maximization algorithm for fitting Mixture Models

The Expectation-Maximization (`EM`) algorithm was first proposed by Dempster et al. in 1977 [DLR77]. The below follows the presentation for mixture model clustering by Hunt and Jorgensen [HJ03].

Suppose first that the data matrix Y is complete, that is, no data is missing. In this case, for a fixed k, we can obtain a maximum likelihood estimate for the missing class labels and the parameters of the distribution with a variation of the general Expectation Maximization-algorithm.

Intuitively explained, the `EM`-algorithm is an iterative process which alternatively improves our current estimates of the parameters, until no additional improvement can be made. We start by picking some arbitrary values for the model parameters[1]. On each iteration, the algorithm first replaces the class labels by their expected values based on the current parameters (Expectation-step). Then it updates the parameters using this filled-in data (Maximization-step, meaning the maximization of the complete-data log-likelihood given the estimates for the missing data). Due to the properties of the model setting, this iteration cannot make the likelihood of the observed data given the current parameters worse, and it can improve it [DLR77]. The procedure is repeated several times, until no considerable improvement is achieved, or the pre-set maximum number of iterations is reached.

Denote by $E[a|b]^{(t)}$ the expected value of a given b at iteration t. Now, more formally, for a mixture of Gaussian distributions, the algorithm works as follows:

> **Initialization**: set $\pi_i^{(0)}$, $\boldsymbol{\mu}_i^{(0)}$, $\boldsymbol{\Sigma}_i^{(0)}$ to some arbitrary values. Set $t = 1$.
>
> **E-step**: set the class labels to their expected values : $\hat{\mathbf{z}}_j^{(t)} = E[\mathbf{z}_j|\pi_i^{(t-1)}$, $\boldsymbol{\mu}_i^{(t-1)}$, $\boldsymbol{\Sigma}_i^{(t-1)}]$ for each $j = 1, ..., N$.
>
> **M-step**: calculate $\pi_i^{(t)}$, $\boldsymbol{\mu}_i^{(t)}$, $\boldsymbol{\Sigma}_i^{(t)}$ for each $i = 1, ..., k$ as maximum likelihood estimates based on Y and $\hat{\mathbf{z}}_j^{(t)}$.
>
> **Convergence**: check if the algorithm has converged or a user-specified maximum t has been reached. If not, increase t by one and repeat the E- and M-steps.

The calculation of the necessary values for Gaussian distributions is straightforward, as follows [MP00, HJ03].

The expectation for individual j belonging to class g is the likelihood of y_j given that class and the class parameters, divided by the sum of the likelihoods of y_j in each class:

$$\hat{z}_{gj} = E[z_{gj}] = \frac{\pi_g \phi(\mathbf{y}_j|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{i=1}^{k} \pi_i \phi(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}. \tag{2.5}$$

The mean vector of each class is the mean of the data for all individuals weighted by the class probabilities:

$$\boldsymbol{\mu}_g = \frac{\sum_{j=1}^{N} \hat{z}_{gj} \mathbf{y}_j}{\sum_{j=1}^{N} \hat{z}_{gj}}. \tag{2.6}$$

[1] Alternatively, we could start from an arbitrary class assignment, followed by first an M-step and then an E-step.

The covariance of two variables l and m in class g is the covariance of the two variables over all individuals, weighted by the class probabilities:

$$\sigma_{g(lm)} = \frac{\sum_{j=1}^{N} \hat{z}_{gj}(y_{jl} - \boldsymbol{\mu}_g)(y_{jm} - \boldsymbol{\mu}_g)}{\sum_{j=1}^{N} \hat{z}_{gj}}. \tag{2.7}$$

The mixing proportions are the sums of group probabilities in each group, divided by N:

$$\pi_i = \frac{\sum_{j=1}^{N} \hat{z}_{ij}}{N}. \tag{2.8}$$

Since the estimates $\hat{z}_{ij}$ sum up to one for each subject j, the final values of the class label vectors $\hat{\mathbf{z}}_j$ can be used as the probabilities that a certain subject belongs to a certain class, and hence, they give the desired clustering.

It can be shown that this process will always converge to a *local* maximum of the log-likelihood [MP00]. No guarantee about finding the global maximum exists; in fact, in many cases a global maximum itself does not exists, as certain pathological cases putting an increasingly narrower distribution over one data point can achieve infinite likelihoods. To counter these problems, the algorithm is restarted and ran with different beginning values several times, and there is a maximum number of iterations. In the end either the most frequent (the one found the most times) or the best (in the sense of the observed-data likelihood) of the solutions that actually converged is picked as the ”correct” one. (Often, but not always, the most frequent and the best solution are equal.)

The algorithm for the Naïve Bayes model works in the same way. The expectation for the individual j belonging to class g is the likelihood of y_j given that class and the class parameters, divided by the sum of the likelihoods of y_j in each class:

$$\hat{z}_{gj} = E[z_{gj}] = \frac{\pi_g P(\mathbf{y}_j | z_{gj} = 1)}{\sum_{i=1}^{k} \pi_i P(\mathbf{y}_j | z_{ij} = 1)}. \tag{2.9}$$

The point probability that the l’th variable $y_{\cdot l}$ takes value A, given a group assignment g, is the number of individuals with that value weighted by their current class probabilities. $\mathbf{1}_x$ stands for a function that takes value 1 if x is true, and 0 otherwise:

$$P(y_{\cdot l} = A | z_g = 1) = \frac{\sum_{j=1}^{N} z_{gj} \mathbf{1}_{y_{jl}=A}}{\sum_{j=1}^{N} z_{gj}}. \tag{2.10}$$

The mixing proportions are the sums of group probabilities in each group, divided by N:

$$\pi_g = \frac{\sum_{j=1}^{N} \hat{z}_{gj}}{N}. \tag{2.11}$$

2.2.4 Implementational details

Beginning values. In practice, it is not possible to pick just *any* beginning values, as this would drive many probabilities too close to zero. Values too close to zero cause both practical problems (inaccuracies in floating point arithmetic) and theoretical ones (one or more clusters becoming empty). Instead, we pick values that we have a reason to believe to be reasonably in the same range as the correct ones. In the case of Gaussian mixtures, we have used a procedure recommended by McLahlan and Peel [MP00], where the means of each component are randomly sampled from a multivariate distribution with the sample mean and covariance as parameters, and all component covariances initially set to the sample covariance. (With missing data, calculated from the available data.)

Definition of convergence. For checking for convergence, we initially considered setting the required difference in log-likelihoods as small as possible (to smaller than `realmin` in Matlab; that is, practically zero). However, as the procedure involves some calculations with very small floating point numbers, the log-likelihoods estimated are not accurate enough to produce the small differences reliably. Also, after a certain change in difference the changes in parameters are also very minimal. In practice, such an extreme setting only serves to prolong the running time without changing accuracy. In the end, after some experimenting, we went back to the `netlab` [Nab01] default of considering the procedure converged if the difference between log-likelihoods is equal or less than 10^{-4}.

Number of restarts As the procedure requires an initial "guess" of either model parameters or cluster memberships, there is inherent randomness in it. The `EM` algorithm guarantees convergence into a local optimum, but not into a global one. To increase the probability that a global one has been find, we need to restart the algorithm from different guesses and keep the best model found in those restarts. (Naturally, it is not possible to guarantee that a global optimum has been found.) Instead of using a fixed number of restarts, I find it a good idea to use an adaptive one: repeat the process until for n consecutive repeats a model better than the best so far has not been found. I typically use n of 10.

2.2.5 The k-means algorithm

In k-means clustering, [Mac67, Llo82, HMS01] k centers are selected and each observation belongs to the cluster the center of which is closest. This can be seen as a special case of a mixture model where the cluster probability of each observation have a value of one for exactly one cluster (and zero for all others), and every cluster in the distribution is spherical in form.

The basic problem is to find a k-means clustering that minimizes the sum of distances of each point to the closest center [Mac67]. If $Y = (\mathbf{y}_1, ..., \mathbf{y}_n)$ is a set of observations, as above, and $C = C_1, ...C_k$ is a partition of the observations into k subsets, f some distance function, and μ_g the mean of the observations in C_g, the aim is to minimize

$$\arg\min_{C} \sum_{g=1}^{k} \sum_{\mathbf{y}_j \in C_g} f(\mathbf{y}_j, \boldsymbol{\mu}_g). \tag{2.12}$$

This minimization in the general case can be shown to be NP-hard in Euclidean space [ADHP09]. Various algorithms converging to local optima instead of the global one are in practice used. The standard method [Llo82] uses an iterative process very like the one described for mixture models in general earlier in this chapter: starting from an arbitrary collection of centers, assign each point to the closest center, and then recalculate the centers. The assignment-recalculation cycle is repeated until no cluster assignments change. More formally:

Initialization: Set $\boldsymbol{\mu}_g^{(0)}$ to some arbitrary values for all clusters $g = 1, \ldots, k$. Set $t = 1$.

Update labels[2]: Set $z_{gj}^{(t)} = 1$ if $f(y_j, \boldsymbol{\mu}_g^{(t)}) = \min_{h \in 1,\ldots,k} f(\mathbf{y}_j, \boldsymbol{\mu}_h^{(t)})$ and 0 otherwise for all $j = 1, \ldots, N$ observations and clusters $g = 1, \ldots, k$ clusters

Update centers[3]: Set $\boldsymbol{\mu}_g^{(t)} = \sum_{j=1}^{N} (z_{jg}^{(t)} \mathbf{y}_j) / \sum_{j=1}^{N} z_{jg}^{(t)}$.

Convergence: If $\boldsymbol{\mu}_g^{(t-1)}$ is $\boldsymbol{\mu}_g^{(t)}$ for all $g = 1, \ldots, k$, or if a user-specified maximum t has been reached, stop the process. Otherwise, increase t by one and repeat the update steps.

The selection of starting points will affect the outcome. Various heuristics have been suggested for to aid in selection. In any case, like above, the

[2] We assume here that the distance of an observation to two different cluster centers cannot be exactly the same. If this does not hold, the class label vectors can be checked and ties can be broken arbitrarily, ensuring that the class label vector for individual $\mathbf{z}_j^{(t)}$ has exactly one value of one, and the rest zeros.

[3] We assume here that no cluster is empty. An empty cluster would lead to division by zero and an undefined center. If an empty cluster happens, it can simply be removed from the process.

process is usually repeated several times from different starting points, and the best clustering (as per the above minimization criterion) chosen.

2.2.6 Handling missing values

For a good general handbook on handling missing data in statistical analyses, see the book by Little and Rubin [LR02].

Missing data in medical datasets

Data can be missing in a medical dataset for various reasons, leading to different patterns of missingness. It is absolutely vital to look at the patterns of missingness for each variable and each individual before starting any sort of analysis, and to identify plausible processes behind missing values. In addition to looking at simply percentage of missingness, we need to also look at these percentages against simple demographics – region, gender, age etc. A simple plot of the data matrix, individuals in rows ordered by running ID, age, etc. with a black square for a missing value and a white one for non-missing one (such as in Figure 2.1), reveals obvious blocks of missing values.

There obviously almost always is some randomly missing data. Complete datasets in medicine are the exception; some data missing is the normal state of things. Blood tests sometimes do not produce a result, or produce an obviously erroneous result, due to technical faults. An individual might skip a question in a questionnaire by accident. Some of this type of chance missingness is not *completely* random: variables have a different probability of failing (some tests fail more often than others, questions at the bottom of a page are skipped more often), and individuals have a different probability of failing too (some people are more careful than others in filling out forms).

Some variables for some individuals are unmeasurable; for example, frequency of a symptom is undefined if the person never experiences said symptom, and questions about past pregnancies are meaningless for males. In this case, missingness depends deterministically on the value of some (possibly in itself unmeasured) variable. In this case, a missing value for that variable identifies a group of individuals, far from random, and actually carries clear medical information.

Some variables might not have been measured for everyone in the dataset for reasons of study design. This might be, for example, because only persons in that subset agreed to have their blood taken or responded to a particular questionnaire, or because the test is expensive and was only applied to a (we hope) representative subsample, or because some new procedure was only

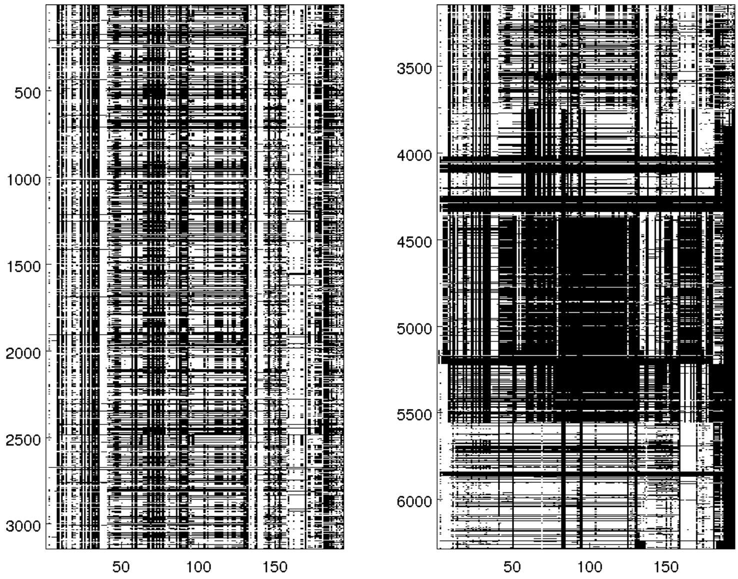

Figure 2.1: Matrix of the data used in the migraine study described in Chapter 4.3, observed values in white and missing values in black. X-axis, the 194 original variables. Y-axis, 6283 individuals in the dataset, ordered by ID. Note the obvious non-randomness of missing data. This picture was originally *not* drawn; only one with rows of the data file was used. Strong correlations of clusters and other variables to running ID very soon called for this one, too.

invented when a long-term data collection process was already underway and thus recorded only those enrolled at a later date. This leads to blocks of missing values: the same m variables are missing for the same n individuals.

We usually have some individuals with data *completely* or almost completely missing. Participants in a medical study are recruited with the aim of covering a particular group of people—for example, in a case-control setting, we formulate a way of sampling cases (e.g. “all patients with this disease who visited clinic a during time T”, or “all patients with the diagnosis of b in the National Hospital Discharge Registry during time T”), and then a way of sampling controls for the cases—but not everyone researchers contact for interest to participate will consent to do so. It would be a mistake to simply disregard the existence of these individuals. While methodologically, we need to proceed with the assumption that our sample with actual data in it is representative of the original population of intent, when interpreting the results we also need to keep in mind that the people who declined to participate probably did not necessarily or even typically do so by random. For example, those worst affected by disease are also more likely to decline because the disease itself prevents or discourages them from participation.

In longitudinal studies, variables collected early in the study are usually more complete than those collected late, due to individuals dropping out of the study, losing contact to the researchers, or dying. In this case, the missingness proportion of a variable correlates to the time of data collection. Similar pattern can often be observed in questionnaire studies: the further towards the end of a questionnaire the variable was asked about, the more people have gotten bored with the form and stopped filling it in. Again, the assumption that the people who drop out or do not finish a questionnaire do so by random is usually necessary to simplify analyses, but nevertheless not valid.

The cases above mostly describe a situation where the missingness of values is a “feature, not a bug”, but data can also be missing by error. Random failures of laboratory tests can be considered to be in this category. In addition, though this is changing, medical data is practically always obtained not by instruments feeding directly to a computer, but by paper questionnaires and manual laboratory tests, which are then coded to an electronic format by an error-prone human being. This can produce not only errors, but also missingness, if the coder accidentally puts in meaningless values or skips some individuals.

A specific case of user error causing missing data is a “file-matching error”, where coding has been done in separate stages (either by different individuals or at different times or both, sometimes even using different software), and when matching those stages an error occurs such that variables are not matched correctly. This leads to a data matrix where a particular variable occurs twice (or more times), with each individual having a missing value in one occurrence and their correct value (possibly also missing) in the other, depending on which patch they were coded in.

Many coding errors can be fixed fairly easily if noticed early enough. Recognizing them only after analyses have been started or done typically means the necessity to redo the whole thing.

Basic approaches to handling missing data

Basic approaches to handling missing data can be roughly categorized into three: ignore missing data, fill it in with some "guesstimate", and include it in the model as is [LR02].

Ignoring values that have missing data work for some applications, where required model parameters or scores can be calculated based on existing data only. Effectively, this equals to only considering the marginal distributions of observed data [LR02]. This method is generally not practical if the proportion of missing data is high, or if the missing data carries a lot of information which is thus destroyed. Hence, it is usually not practical in complex diseases applications.

Possibly the most typical approach to missing data in medical studies is to impute it. Imputation refers to the practice of replacing missing values by some estimated values [LR02]. As simplistic schemes, for example mean value of the variable or the mean of a group of similar individuals (same age group and sex, for example) could be used. In the case of a missing questionnaire item the mean of the answers of the same individual to similar questions might be appropriate. Various more complex schemes based on for example regression on other, observed items, can also be used (see for example [TCS$^+$01, LR02])

Shrive and colleagues [SSQG06] analyze various methods of imputation in a 20-question depression scale, for fairly simple scenarios of missingness. While basic statistics such as means and SDs could be regenerated from imputed data, the best method analyzed (multiple imputation) had mis-classification rates (compared to full data) for depression between 5 and 10 percent, depending on the scenario. Note that this mis-classification is incurred by the missing data alone, and for fairly simple missing data assumptions.

The third alternative is to include missing data in the clustering model. This is can be done by at least two different ways: treating missing data as an additional value for the variable (applicable when the variable is categorical), or trying to learn the missing values at the same time as the clustering assignment (works well in the EM context, see below). Wagstaff [Wag01] has also suggested a modified KSC algorithm that treats variables with missing values as constraints on a clustering performed with complete variables; however, as it is usual that all or almost all variables have missing data, this approach is not practical for complex diseases analysis.

Whatever approach we end up choosing, when drawing conclusions from the results, it is necessary to keep in mind that some data was originally missing. It is important to, as far as possible, flag any originally missing data [TCS$^+$01, Wag01], and take care to analyze clustering results with regards to their dependency on data missingness.

Methods used in this work

The following is based on [GJ94] and [HJ03].

In the discussion in Chapter 2.2 of the EM algorithm for the mixture of Gaussian distributions or discrete the Naïve Bayes model, we considered a situation where we have no missing data in the initial data set, that is, the only hidden data are the class labels of the model setting. The EM-algorithm, however, can also be used in situations where data is missing from some of the subject vectors y_j.

In this case, the expected values of group probabilities can be calculated as

$$E[z_{ij}] = \frac{\pi_i \phi(\mathbf{y}_{j,obs} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{g=1}^{k} \pi_g \phi(\mathbf{y}_{j,obs} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)} \tag{2.13}$$

where $\mathbf{y}_{j,obs}$ are the observed values of the j'th subject. These calculations are done using the appropriate marginal distributions and pose no great problems.

Concerning the other necessary calculations, the idea immediately suggests itself that the missing values in Y should also be replaced by their expected values given the group parameters for the calculations in each group. The conditional expectation for a variable m, missing for subject y_j, supposing that it originates from the g'th mixture component ($E[y_{jm} | y_{j,obs}, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g]$), can be calculated for arbitrary normal distributions by the sweep operator (see [HJ03, LR02]). Indeed doing so will lead to unbiased estimates for $\boldsymbol{\mu}_g$ [HJ03].

Doing so in this straight-forward way will force many values artificially close to the means, and thus the calculations would underestimate the covariances [HJ03]. The bias for variables m and l is known to be $(N - 1)^{-1} \sum_{j=1}^{N} \sigma_{lm|obs,j}$, where $\sigma_{lm|obs,j}$ is the covariance of the observed values of the variables m and l when y_{jl} and y_{jm} are both missing, and 0 otherwise

[HJ03]. Intuitively, the covariances must be increased by the amount of covariance that is not explained by the observed variables to avoid underestimating them.

In practice, it is not reasonable to calculate all the values exactly, but on each E-step calculate the sufficient statistics for each component distribution. The calculation can be sped up by using a so-called matrix sweep operator for deriving the conditional expected values and residual covariances. Even then, the procedure is time consuming for large amounts of data.

A `MATLAB` implementation of the procedure was programmed by the author for the analyses presented in this book, based on the publicly available `netlab` package [Nab01]. In a class data case when using Naïve Bayes models, I have handled missing data by an additional label rather than any other imputation scheme, prior to or during clustering. The additional missing value label allows the model to adjust to data that is not missing by random.

2.3 Selecting the number of clusters

2.3.1 Overview

Most clustering methods in general use, including all those used in this work, require the number of clusters to be passed to the basic algorithm as a parameter. Selection of the number of clusters based on minimizing the same criteria that one uses to select between models of the same number of clusters is generally not possible. Namely, these criteria rely directly or indirectly on the average distance between the closest cluster center and the data points, and hence tend to go to zero as cluster number approaches the number of data points.

Hence, the question whether a real cluster structure exists in a data being clustered, and if so, how many clusters are there, is of fundamental importance in real-data clustering studies. Various validity indices have been developed to answer this question; for a good overview see, for example, [BSH$^+$07].

Generally, these measures can be divided into three types: internal, relative, and external [BSH$^+$07]. Internal measures are such that measure some quality of the obtained clustering itself, for example the Silhouette index [Rou87, GB03], which measures how compact and separate the obtained clusters are, or the Bayesian Information Criterion [Sch78], a score for selecting between models of different parameters based on data likelihood given the model and a penalty score for model complexity.

Relative measures compare alternative clusterings of the same data with different subsets of the data. For example, stability schemes where clusterings obtained on partial data are compared [LRBB04] and other cross-validation procedures [Smy96] fall into this category (see also Chapter 2.5.2). External measures rely on the comparison of the obtained clustering to some "golden standard" and as such are usually only relevant in method development [BSH$^+$07], as when the correct classification solution is already known, clustering is rarely needed.

In addition to exact scores, various visual aids have been suggested. Such methods are based on looking at scores or other measures of the alternative models and identifying a point where the rate of improvement between subsequent cluster numbers changes, as first suggested by Thorndike in 1953 [Tho53].

In this work, we have used two methods in particular, namely the Bayesian Information Criterion (`BIC`) [Sch78] and the 10-fold cross-validation scheme [Smy96]. These scores are described here in detail; for experiments comparing the two, see Chapter 3.2.

2.3.2 The Bayesian information criterion

A number of scores have been developed for selecting between clustering models based on the idea that a model should be scored by minimizing the error between the data and the model plus a penalty on the complexity of the model. A complex model that fits the data perfectly will have zero error, but a big penalty on the complexity, while a simple model might have a large error, but still score better because of the complexity penalty being low.

The Bayesian information criterion [Sch78] is suitable for selecting between models that can express data likelihood given the model, or where the model errors can be assumed to be normally distributed. The basic form of the `BIC` score is

$$-2 \ln L + o \ln(N) \tag{2.14}$$

where L is the likelihood for the model, o is the number of free parameters in the model, and N is the number of data points. The first term is often called error, and the second one penalty.

For Gaussian mixtures, the number of free parameters is

$$o = kd + kd(d-1)/2 + (k-1) \tag{2.15}$$

where d is the number of variables in the data and k is the number of components. The first term in the number of parameters comes from the k

cluster centers of dimensionality d, the second term from the k covariance matrices, and the third from the cluster probabilities (the -1 is there because once all but one of them are known, the final one is fixed).

For a Naïve Bayes mixture model we have

$$o = k \sum_{i=1}^{d} (a_i - 1) + (k - 1) \tag{2.16}$$

where k is the number of clusters, d is the number of variables, and a_i is the number of possible values for the i'th variable. The first term in the number of parameters are for the k probability tables for d variables (-1, because once $a_i - 1$ probabilities are fixed, the last one is known), and the second term is for the cluster probabilities.

In the case of k-means, which does not directly produce a likelihood estimate for the data, the `BIC` score can still be used if we treat the clusters as spherical Gaussians [CG92].

Like for many other scores for cluster validity and/or similarity, it is not always easy to clearly describe what value of `BIC` is "a good one", as the range of the values is greatly data-dependent. However, this is not a limitation in the case of this work, since we only use `BIC` to compare models of the same class built with the same data, and hence it is enough to see if one score is bigger or smaller than another.

2.3.3 Cross-validation

Another commonly used way to address the problem of over-fitting is that of cross-validation [Smy96, MP00]. In v-fold cross-validation, the dataset is partitioned to v non-overlapping groups of equal sizes at random [Smy96]. One partition at a time is designated as the test set and the rest of the data used as the training set. Models parameters are estimated on the training set, and then evaluated on the test set. For example, the average distance of the points in the training set to the k centers in the learning set is calculated for k-means and the data likelihood of the test set given the model learned on the learning set is calculated for mixture models. Alternative models of different number of clusters are learned with each training dataset. The sums (or averages) scores for each number of clusters are then compared, and the best one chosen as the number of clusters. In Monte-Carlo cross-validation, repeated partitions of the same size are constructed at random, with the possibility for overlap allowed [Smy96].

The idea is that as long as the model reasonably can represent the *underlying distribution* or complete object space from which the samples are drawn, the likelihood of the test set should also remain high, while if the model overfits to data, it might be very good on the training set but will fail on the test set, because it has adjusted to the outliers. In practice, we have used $v = 10$ and number of clusters from 1 to 12 with success.

2.3.4 Visual aids

In addition to scores, various visual aids have been used to figure out the correct number of clusters. The most common of these is the so-called "elbow criterion".

As stated above, for most of the clustering methods in common use it holds that by increasing k one can always improve on the goal function (until $k = N$). For example, in k-means clustering, by increasing the number of cluster centers, the optimal sum of the distances of the points to those centers will decrease, and in mixture model clustering the overall data log-likelihood decreases if we introduce more components.

However, this change tends to be rapid in the beginning of this process of increasing k, while after a certain point adding more clusters will result in only a slight change. Looking at the plot of the goal function, one can often identify a point where an initially rapid decrease in the function to be minimized flattens off. Looking for such a point and taking it to represent the correct number of clusters is titled an "elbow criterion". (The idea was introduced by Robert L. Thorndike in a humorous speech "Who belongs in the family?" in 1953 [Tho53], though he introduces it as something to base a score on, freely admitting he could not actually achieve that.)

Obviously, this method is far from foolproof: the elbow might not exist, but instead we observe a smoothly flattening curve, or there might be more than one elbow. However, it does always make sense to look at not just the minimum value of whatever score is being used, but also the raw values of the goal function and the score for each k tested. It can easily happen that a score suggests a particular number of clusters, but looking at the actual goal function one sees a plateau rather than a dip around this point, in which case also the adjacent numbers should be considered as a possibility. In the scope of this work, we have always looked at the goal function plots, but not really utilized the elbow criterion as a method to select the number of clusters.

In addition to looking at the goal function, we can look at other measures describing the clustering solutions. In the studies of this thesis it turned out to be useful to examine the similarity between the clustering solutions for subsequent values of k. See Chapter 3.3 for more discussion and Figure 4.10 for an example.

2.4 Comparing clusterings

In the course of clustering, we often have to compare two alternative clustering models for the same data to find out if they are similar or not. For example, we might want to ask if two models obtained from different starting points for an algorithm are the same or similar, or if the clusterings obtained for different values of k resemble each other or not. In stability analysis (see next section) we want to know how much a clustering obtained with partial data differs from one obtained with full data, and in replication studies we want to compare cluster labels obtained based on the model learned on one dataset to those obtained by a model learned on an independent dataset. Sometimes we also want to compare a clustering result to some known classification, in medicine typically diagnoses, and in experiments with artificial data to the known “correct” reference clustering.

Various scores have been developed for comparing two clusterings based on the labels they give for individuals; for an overview see, for example, [Mei05]. These have been roughly categorized into three varieties [VEB09]: pair-counting, set-matching, and information-based.

2.4.1 Pair-counting measures

Pair-counting measures look at all pairs of items (individuals) in the dataset, and count the times when the two items are, or are not, in the same cluster in each of the clusterings [VEB09]. Four counts are obtained when comparing clusterings C and D:

- N_{11}: the number of pairs that are in the same cluster in both C and D
- N_{00}: the number of pairs that are in a different cluster in both C and D
- N_{10}: the number of pairs that are in the same cluster in C, but in a different cluster in D
- N_{01}: the number of pairs that are in a different cluster in C but in the same cluster in D.

Based on this counts, various scores can be formulated to measure the degree of agreement between C and D. The possibly most intuitive of these, the “pairwise concordance” or the Rand Index [Ran71] similarity measure counts the times the clusterings agree on a pair out of all possible pairs:

$$S_{PC}(C,D) = \frac{N_{00} + N_{11}}{N_{11} + N_{00} + N_{10} + N_{01}} \tag{2.17}$$

The problem with this measure is that while the score is one if and only if the two clusters exactly agree, it does not in practice ever reach zero; in fact it tends to concentrate on a narrow interval close to one [HA85, VEB09]. The expected value for a random clustering varies with the clustering parameters (N, k, mixing proportions). This expected value can be calculated, and adjustments have been suggested based on it [HA85], scaling the value between 0 and 1, with a baseline of 0 for similarity of random clusterings, independent of k [VEB09].

Another pair-counting score is the Jaccard Index [BHEG02]

$$S_{JI}(C, D) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{2.18}$$

This index leaves out of consideration the case where both clusterings assign the pair to different clusters. This can be argued to be appropriate in the case where there are numerous clusters and the case of two items belonging to a different one is not very informative. As an example, in gene expression studies the result that two genes do *not* share an expression pattern could be considered to be of little interest.

2.4.2 Set-matching methods

Set-matching based scores compare the clustering labels directly by counting the contingency table $n_{c,d}$ of clustering labels where $c = 1, ..., k_C$ and $d = 1, ..., k_D$ are the clustering labels of the two clusterings C and D[VEB09]:

		Label in C				
		1	2	...	k_C	total
Label in D	1	$n_{1,1}$	$n_{2,1}$	...	$n_{k_C,1}$	n_1^D
	2	$n_{1,2}$	$n_{2,2}$	...	$n_{k_C,2}$	n_2^D
	...	...	...	...		
	k_D	n_{1,k_D}	n_{2,k_D}	...	n_{k_C,k_D}	$n_{k_D}^D$
	total	n_1^C	n_2^C		$n_{k_C}^C$	N

Many standard statistical tests (for example, the χ^2 test) for testing significance of the deviation of a cross-tabulation from the null hypothesis that the table is produced by two independent random variables (see e.g. [HL01]) could also be applied to this table. It is, however, obvious that two alternative clusterings produced with the same method of the same data cannot be independent, so the meaning of this kind of statistical testing is somewhat unclear when comparing clusterings with alternative parameters.

In comparing clusterings of replication samples (see Chapters 2.5.3 and 3.4) the statistical question of independence becomes more valid.

Many scores for comparing cross-tabulations (for example mis-classification rates or rater agreement measures) assume that a clear relationship exists between labels in C and those in D [HL01, MH01]. In clustering, cluster labels are typically not semantically meaningful or fixed, and so for example "cluster 1" in C cannot be assumed to be the equivalent of "cluster 1" in D. For many scores this difficulty can be overcome by looking at all possible ways of matching up the labels in C and D and taking the maximum/minimum (whichever is appropriate) over those [MH01]. In clustering, since k is typically fairly small for both C and D, an exhaustive search over all possible mappings is usually possible.

For example: let K be the maximum and K' the minimum of k_C and k_D, and F the set of all injective mappings f of $\{1, \ldots, K\}$ into $\{1, \ldots, K'\}$. The Meila H-index [MH01] is then defined as

$$S_{MH}(C, D) = 1 - \frac{1}{n} \max_{f \in F} \sum_{g=1}^{K} n_{g,f(g)} \tag{2.19}$$

or, in other words, the mis-classification rate minimized over all possible ways to associate the clusters in one of the clusterings to the clusters in the other.

As another example, pertaining to the domain of the real-data studies in this work, Cohen's κ [Coh60] is a score often used in psychology to compare two "raters" or classifiers. It can be defined for clusterings in the same way as Meila's H above. Note though that Cohen's κ is only valid when $k_C = k_D = k$. Let $f : \{1, ..., k\} \to \{1, ..., k\}$ now be a bijection telling how clusters in C are matched with clusters in D. Define the expected probability of agreement between two classifications given f as

$$p(a|f) = \sum_{g=1}^{k} (\frac{n_g^C}{N} \frac{n_{f(g)}^D}{N}) \tag{2.20}$$

that is, the probability of agreement given the marginal distributions of the classifications. Cohen's κ in the clustering case is defined by:

$$S_{C\kappa}(C, D) = \max_{f \in F} \frac{\sum_{g=1}^{k} (n_{g,f(g)}/N) - p(a|f)}{1 - p(a|f)}. \tag{2.21}$$

The first term $\sum_{g=1}^{k} (n_{g,f(g)}/N)$ of the numerator tells the proportion of observations clustered according to the pairing f, and F is the set of all possible bijections f.

One problem with this kind of measures is defining what constitutes a good similarity result, and what does a certain difference between observed values mean. For example for Cohen's κ, based on simulations, it seems not possible to give one good cut-off value for what is an "acceptable" result [BQMR97]. In literature, various rules of thumb such as 0.6 for "substantial" and 0.8 for "almost perfect" agreement [LK77] have been suggested. Like Bakeman and colleagues [BQMR97], we suggest using such high values of Cohen's κ as indicative of the presence of similarity, but avoiding comparisons of models based on values that both make the cut-off (considering for example the value of 0.85 as "significantly higher" similarity than 0.75).

2.4.3 Information-based measures

Finally, two clusterings can be compared based on their mutual information. Let

$$p(g,h) = \frac{n_{g,h}}{N} \tag{2.22}$$

be the joint probability of cluster label g in C and cluster label h in D, and similarly let the probability of cluster label g in C be

$$p(g|C) = \frac{n_g^C}{N}. \tag{2.23}$$

Here $n_{g,h}$ is the number of observations having cluster label g in C and cluster label h in D, and n_g^C is the number of observations with label g in C, as before.

The mutual information of two clusterings is defined as the mutual information between the associated random variables [CT91]:

$$S_{MI}(C,D) = \sum_{g=1}^{k_C} \sum_{h=1}^{k_D} p(g,h) \log \frac{p(g,h)}{p(g|C)p(h|D))} \tag{2.24}$$

`MI` is non-negative and takes value 0 if the two clusterings are independent ($p(g,h) = p(g|C)p(h|D)$ always). Otherwise, if logarithm of base 2 is used, it can be thought of as a measure for how many bits of information knowing the clustering label of an item in C reveals of its label in D (or vice versa; `MI` is symmetric). [CT91]

The upper limit of `MI` depends on the entropies of the two clusterings—in practice, this means, on N and k. Hence, it is not immediately intuitively usable for figuring out how close to equal two clusterings are. This can be avoided by normalizing the score, for example according to [SG03]:

$$S_{NMI}(C,D) = \frac{S_{MI}(C,D)}{\sqrt{H(C)H(D)}} \tag{2.25}$$

where $H(C)$ is the entropy of C:

$$H(C) = -\sum_{g=1}^{k_C} p(g|C) \log p(g|C) \tag{2.26}$$

(using the same logarithm base as for the mutual information, naturally).

Vinh and Bailey [VEB09] observe that while the theoretical baseline of the (normalized) mutual information is zero, this is rarely achieved for random clusterings when k exceeds $N/100$, because by chance the clusterings will resemble each other. To counter this, they calculate the expected value of the Mutual Information between two clusterings with the observed marginal counts

$$E[I(M_{C,D})] = E[I(M)|n_1^c \dots n_{k_c}^c, n_1^d, \dots n_{k_d}^d] \tag{2.27}$$

Here M simply denotes any contingency table consistent with the marginal counts. Based on it, the normalized Mutual Information can then be adjusted for the expected value as follows.

$$S_{AMI}(C,D) = \frac{S_{MI}(C,D) - E[I(M_{C,D})]}{\sqrt{H(C)H(D)} - E[I(M_{C,D}]} \tag{2.28}$$

This for the case where the normalization is based on the square root of the entropies; other normalization possibilities exist [VEB09]. This measure is called the Adjusted Mutual Information or `AMI` in the following.

`AMI` has the desirable property that it it takes value of 1 if and only if the clusterings are exactly the identical, and the value of 0 when the mutual information between the models is what you would expect by chance [VEB09]. As the normalization and the adjustment are both monotonic, two values of the `AMI` can be directly compared just like two values of mutual information; higher value implies higher similarity.

2.5 Cluster validation

2.5.1 Validity

By validating a clustering we mean, in a loose sense, the procedure of evaluating the results of a clustering algorithm [HBV01], especially when the aim is to do so in a quantitative and objective manner [RLBB02]. The main questions are:

- Are the clusters *real*, in the sense that they describe an actual structure in the data or in the population from which the data was sampled, whichever is appropriate?

- Are the clusters *interesting*, in the sense that describing the data as clusters benefits the goals of the domain researchers?

An example of a clustering that is real, but not interesting, is one that differentiates between males and females in a sample, or other division along known groups. Examples of clusterings which are interesting, but not real, can be obtained from a random clustering; the human ability to see patterns everywhere is amazing.

The question of whether a clustering is real can be approached by various measures of clustering stability [HMS01, HK06]. As already mentioned in Chapter 2.3, such validation can be based on internal, relative, or external features of the data.

If we assume that we have sampled every individual in the population of interest, the question whether a structure really exists can be answered by scores used for cluster number selection (see Chapter 2.3) by comparing the trivial solution of all individuals in one cluster to other solutions. However, in medicine (and many other fields of science) this is usually not the case, and we have to answer the harder question of whether the obtained clustering corresponds to something real in the underlying population; this is where the concept of stability becomes relevant.

The interestingness of a clustering will usually be established by a domain specialist, with the help of various visualizations and statistics (see Chapter 2.6.2). To avoid over-estimating the value of the clustering because it seems interesting, the question of validity in the “reality” sense must be established before looking at interestingness.

2.5.2 Stability

Cluster stability refers to the variability (or lack of it) in clustering solutions, when the process is repeated. In practice, the term can refer to four different concepts.

One, there is the stability related to the clustering algorithm: whether consequent runs of the same algorithm will result in similar results. In many clustering algorithms, there is some randomness involved; for example in the case of k-means and mixture models via the `EM` algorithm, one has to pick the starting point at random. In some other algorithms randomness can be introduced by the process. For example, in hierarchical clustering the methods require the selection of a pair of clusters based on a distance measure, and ties are usually broken at random.

Several restarts should be performed with different random initializations and random choices; standard practice is to report the best of these as the result [HMS01, HK06]. However, in the presence of a clear, real cluster structure the vast majority of these subsequent runs tend to result in the same or similar solution. This phenomenon is the first meaning of “cluster stability”. While such stability cannot be considered proof of a real, interesting cluster structure, the absence of such basic stability should lead to doubts of the presence of such a clustering, assuming one has no reason to doubt the basic applicability of the clustering algorithm used to the data at hand.

We have found that in practice (see studies reported in Chapter 4), for both k-means and mixture model clustering, 10–20 random restarts for each number of clusters is sufficient to establish this sort of stability in the presence of a strong cluster. These restarts are usually performed in any case to search for the best solution for the particular number of clusters and so analyzing this kind of stability causes no additional time requirements.

Second, stability can refer to the similarity of clustering results obtained when individuals are randomly removed from the dataset. If randomly removing even a small number of individuals tends to change the clustering results a lot, this suggests that the original results are highly dependent on particular outlier individuals, which is usually not the desirable case.

This sort of analysis requires a high number of re-clusterings with subsets of the data. We have found it a good and informative practice to randomly drop subsequent tens of percent of individuals (keeping 90, 80, 70, ... percent of the original rows, down to 10 percent), and to perform 10 separate random drops of each size, using as k the number of clusters the full data suggests. The clustering labels obtained on these alternative models are then compared to ones obtained with full data. This process is similar to Monte-Carlo cross-validation for cluster-number selection [Smy96], but instead of using it to select the number of clusters, we use it to assess the stability of a clustering solution.

A simple plot of the average similarity for each sample size (for any measure that seems suitable, see Chapter 2.4) will establish the cluster solution dependency on particular individuals. A good result is one where the similarity starts to drop clearly only when the proportion of removed individuals is of such a size that it is likely that most members of a particular cluster have been removed (hence, dependent on the minimum cluster size). Similar analysis can be performed for dropping variables.

Third, “real” clustering structures are usually stable relative to changes in the model parameters, most notably in the case of this work, changes in cluster number k. A true clustering structure in the population often (though not necessarily quite always) leads to clustering solutions of subsequent k being hierarchical refinements of each other, and hence clusterings of different k resembling each other (see Chapter 3.3 for observations of this in simulations).

Since clusterings for subsequent k are in any case obtained in the process of deciding k, it is easy to look at similarities between these different models. In the absence of similarity, more care needs to be taken in selecting the number of clusters, and the question whether the clustering actually is “real” must be carefully considered, while in the presence of similarity k can be more freely selected from the models that score close to best, and some confidence can be gained by the clustering being something other than arbitrary.

Fourth, “stability” might refer to stability in an actual replication study; this is discussed in the next subsection.

2.5.3 Replication

Clustering is sometimes thought of as partitioning a set of objects that, implicitly or explicitly, are thought to cover the whole object space. The goal is defined simply as obtaining a good partition of the objects at hand. However, in reality, at least in medical sciences, most clustering studies are performed not to obtain a good partition of the individuals in the sample, but with the hope that such a partition could reveal something about an underlying, unmeasured, structure in a population from which that sample was drawn. In natural sciences in general, the replication of results in a new sample from the same or similar population is considered the ultimate test for whether the results represent facts.

The idea of replicating clustering studies is as such hardly new: Ketchen and Sook mention this idea in their paper on application of cluster analysis in strategic management research [KS96], referring as far back as to a study in 1982 where it has been done. They also comment, however, that obtaining second samples can be difficult or even impossible.

Despite the practical difficulties, whenever a second sample *can* be obtained, replication remains the method of choice also for validating clustering results of population samples. Given datasets A and B, a simple way to approach the replication is as follows:

- Make sure all normalizations, data transformations, and missing data imputations are done in the exact same manner in both datasets.

- Learn a clustering model M_A and a clustering C_A^A using dataset A, exactly as if you would if B did not exist.
- Similarly, obtain a clustering model M_B and a clustering C_B^B using dataset B.
- Assign a cluster to every individual in A using model M_B (resulting in clustering C_B^A),
- Assign a cluster to every individual in B using model M_A (resulting in clustering C_A^B).
- Now treat $C_A^A \cup C_A^B$ and $C_B^A \cup C_B^B$ as two alternative clusterings over the combined sample. If the clusterings reflect a true cluster structure in the population, they should be similar (though rarely the exact same).

When the number of clusters in M_A and M_B is different, we can additionally compare not only the best models for both, but the models with the corresponding number of clusters for both cases, resulting in a total of three comparisons of similarity.

Since it is usually the case that the clustering study on the first dataset is already complete when the second sample is obtained, we must take care that the preprocessing phase is replicated properly. For example, if variables in sample A was discretized into five classes using percentiles from sample A, it must be considered whether the correct solution to use these same boundaries in discretizing B, instead of using a separate set of boundaries obtained from the second sample. The argument for this is that the boundaries are somewhat arbitrary and might be affected by outliers, and we are trying to replicate the clustering, not the percentiles. The best way, naturally, would be to use boundaries from the combined sample, but this would require repeating the clustering on A.

Please see Chapter 3.4 for an example showing that replication can not only separate between clusterings of different populations, but also detect the case where clusterings are erroneously detected in a population where no subgroup structure exists, and Chapter 4.2 for a study where replication was used.

2.6 The final model

In the end of a clustering process, we arrive to one (or a few) models that we then consider the “true” one. We then analyze what that model tells us about the data or about the population where the data was sampled from.

2.6.1 The selection of the “final model”

In a practical data analysis situation, sooner or later we have to fix one model, or at most a handful of models, as the final clustering, based on which conclusions of pertaining to the domain can be drawn. Clustering is, by nature, exploratory data analysis. It does not render itself easily into formal hypothesis testing setting, and the concept of correction for multiple testing does not always easily apply.

The strict approach is to require that the selection process is fixed exactly beforehand, in order to avoid biases caused by the researcher’s preferences for results. We should indeed strive to write down the process by which this selection will be done in the beginning of the process and during it document all steps and any changes. In the very least, we must fix which score will be used for model selection, e.g. to select between different k.

I have, however, yet to see any analysis process on non-trivial data where at least one detail would not have been changed during the analysis. Typical reasons for such changes include realizing too late that known subgroups (e.g. males and females) have different data ranges and so should probably be clustered separately, or finding out in the middle of the process that clustering solutions for $k = 2, ..., 4$ have practically equal scores. In such cases, it makes no sense to stick to a predefined process that does not fit the requirements of the data.

Based on these experiences, to blindly choose the best model (k) based on a score does not seem advisable. Instead, we should look at the scores and consider all models that have a score close to the best. As shown in Chapter 3.3 it is typical that clusterings into subsequent k are refinements of each other, and when this holds, it can be argued that one of them is not any truer than the others, but that one may freely choose the detail level that provides best or most thought-provoking insight into the medical field under study.

2.6.2 Visualization and statistical analysis

Once a model has been selected, we face the tasks of explaining and interpreting it. First step of this process is to gain a reality check for ourselves; the last step is to explain the final model in a scientific article or presentation. These steps and those in between require all sorts of visualizations and statistical analyses to describe the groups identified.

It is obviously impossible in the scope of this work to list all possible ways, or even all ways ever so far used, to present a clustering solution. In addition, a single correct way for visualization does not exists, but the correct way depends on what is being presented to who. Next we describe some simple approaches. See, for example, [TSK06] for more material on visualization.

In the first stage, we want to gain insight to the clustering solution at a high level. We can start with this with simple tables: number of individuals in each cluster, and the breakdown of demographic and "important" variables such as gender, diagnoses, etc. in each group, giving proportions for categorical and mean and SD for continuous variables. Bar charts and boxplots will help in presenting this work to those who prefer visual representations at this stage. At this point, we also perform certain reality checks: we look at things like running ID, order in data matrix, family numbers, and percentages of missing data per individual to see if the clusters seem to associate to any of these. Plotting the data into the first principal components can help to see how clear the separation of the clusters is, though when the dimensionality reduction is an order of magnitude or more, a lack of separation in the first few dimensions does not necessarily mean anything.

The aim of this stage is to be able to describe the clustering in basic, concrete terms, such as: we have k clusters, with $\pi_1, ..., \pi_k$ percent of the individuals in each cluster, males and females do / do not separate into different clusters, individuals in cluster g tend to be older/younger than others, individuals with this-or-that diagnosis cluster separately from / together with those of another one. We will, hopefully, be able to say that the clustering is not directly dependent on out-of-domain variables such as running IDs or non-informative missingness status.

The second stage of looking at the solution is to look at all variables in the dataset (both those on which the clustering is based on, and any other information), and see how they associate to the clustering. Tables with means and SDs or proportions (whichever is appropriate) as well as a bar graph or a box plot are produced for each variable. In addition to the values for each cluster, we should include similar figures / graphs for the whole data, for comparison. We then need to go through all these and noting down which characterize each cluster. In addition to observed values of the data, we also look at percentages of missing data broken down per cluster.

This is a tedious task, especially when the number of variables is large. Due to missing data, special value concerns, the need to specify what type each variable is, and so forth, a special script or program to allow for such special cases and to print out all the information in a format suitable for this particular dataset is usually necessary, instead of using a general data analysis program.

In addition to providing statistics and visualizations, we would strongly prefer such a program to automatically call our attention to the 'most significant' or 'best characterizing' variables. Unfortunately, the question of how to assess the significance of the association of a variable to a clustering is not easy to answer by standard statistical methods, because the concept of independence is not clear where we have specifically tried to construct a clustering that separates the individuals. Only when the variables analyzed at this stage are clearly separate from those that the clustering is based on (for example, analyzing the association of childhood social class with adult temperament, see Chapter 4.2), the use of simple p-values with conservative multiple correction can be completely justified as a means to confirm that an association of a variable to the clustering is statistically significant.

Due to the fact that it is rare that researchers collect data that they do not primarily believe to have something to do with the phenomenon of interest, we face a lot of gray area when trying to answer the question whether two clusters "significantly differ" from each other regarding a particular variable. One solution is to pick a suitable statistics (for example, a χ^2 test for contingency tables, a t-test for independence of means [HL01]), and simply use the order of variables to tell which of them are the most important for that cluster, without drawing further statistical conclusions.

Sometimes it is not individual variables we are interested in, but patterns of variables. Especially when describing cluster centers, we want to get some over-all idea of how the variables go together. A table can help in this for those who are good at reading them. As visualization, we have found the rarely used starplot (see Figure 4.8) of good use, because of the (admittedly fairly subjective) ease of spotting patterns without the (also subjective) distraction of variables catching attention differently depending on their relation to the left / right / center of the graph (or table).

After this stage of analysis, we at least to some degree can describe the clusters in terms of individual (clustering and background) variables, such as whether clusters differ on amount of missing data, what variables have different statistics between clusters, and which variables for each cluster are the ones that separate them.

At this point, the research group has very probably developed nicknames for the clusters, such as calling some "the elders" and another "the psychotic". As it seems to be in practice impossible to force everyone to talk about e.g. cluster A, B, and C all the time (if we try, we will find ourselves going "Who are they again?" "The urban businessmen." very soon), it is a good idea to check the nicknames against data, and if necessary change them to something that fits better.

2.6.3 Things to take into account when working with domain experts

A real-data clustering study should in my opinion never be performed without a domain expert, that is, a scientist trained in the field the data describes. The closer that expert is familiar with the particular dataset, the better. The closer they can work with the group performing the actual analysis, the better.

First of all, expertise is needed for to simply understand the data in the pre-processing stage. For a trivial example, most medical variables have natural boundaries of possible values: just like a person's age cannot be less than zero, a person's systolic blood pressure cannot be above 300 (without death occurring instantly). Where discretization is needed, instead of automatic schemes a natural boundary can sometimes be defined, and these natural boundaries should be used whenever possible. Categorical variables often include particular values that are of the most critical interest, so sometimes it makes sense to reduce the number of values to include only these and "the rest". Secondly, expertise is critical in looking at the results. Obviously, it gives us a way to establish the interestingness of the results in practical studies.

There are two caveats related to working with experts of a medical field – these probably hold for experts in other fields too, but the author of this has no personal experience there. First of all, human beings are extremely good at detecting patterns, regardless of whether any exist in the data they are seeing, and experts are even better at seeing patterns in their own field (this is part of what makes them experts). The whole field of statistics can be viewed partly as an answer to this problem: to discern a spurious pattern that a human being sees but which actually rises out of random chance from patterns produced by actual, interesting features of the data.

We must, therefore, resist the temptation to establish the validity of a clustering solution by showing it to a domain expert and asking if it makes sense, as this question will also in some cases where the solution is arbitrary result in an affirmative answer. Instead, we must first convince ourselves that the clusters reflect a structure in the data and/or population (see Chapters 2.3 and 2.5), and only then present the domain scientists with the question of what interesting patterns they see in this existing clustering.

Another standard problem of exploratory data analysis is the temptation to do the analysis every which way and then pick the most appealing solution as the correct one. Naturally, some 'fiddling with it' is impossible to avoid.

The question of whether to analyze, e.g., men and women separately or to adjust variables for age depends a lot on whether the results when doing so differ from the 'raw clustering'. Which variables to include is subject to some debate and it is a rare real-life study where this decision is fully made before first passes at analysis have already been performed. Sometimes one notices that a clustering solution is dependent on a particular uninteresting variable or on missingness status, and a new clustering excluding these must then of course be performed.

In the absence of a formal procedure of penalizing such fiddling (such as e.g. correction for multiple testing provides for hypothesis testing), it is difficult to give any hard and fast rule of when to stop this iterative fiddling-with-it process. It is absolutely essential though that the people involved are aware of the temptation and do not imagine themselves immune to it. Some people see the watching out for this as the domain expert's job: they are seen as 'the customer' of the data analysts, and the analyst's job is to simply perform whatever analyses are requested. In the opinion of the author of this thesis, however, it is primarily the responsibility of the data analyst to draw the line; they know the method, and knowing when to stop is part of knowing your trade.

Chapter 3

Simulations

> *"A fool is a man who never tried an experiment in his life."*
> (Erasmus Darwin)

As seen in Chapter 2, there are many possible choices in the selection of a clustering method and its setting. Hence it is important to understand how the methods chosen perform on different types of data. It is useful to test the behaviour of the algorithms on datasets that have a known structure and at the same time resemble real data in many respects. This chapter describes such experiments on simulated data.

To begin with, the artificial data used will be described. Then we explore the behaviour of two scores for selecting the number of clusters, namely the `BIC` score and 10-fold crossvalidation (see Chapter 2.3). The second set of simulations concerns the observation that in the presence of a true clustering structure, even a non-hierarchical clustering method tends to provide approximately hierarchical results. Next, we explore the use of a separate sample from the same population in confirming the clustering results. Then, effects of missing data on these clustering methods is tested, and finally, we shortly explore the practice of removing data rows to confirm cluster stability.

3.1 The artificial data

In generating the datasets we experiment on, we have tried to imitate real data, but do so staying close to model assumptions, except when deliberately testing the effect of breaking them. Thus the artificial data is generated by using parameters that stem from the real datasets. Unless mentioned otherwise in what follows, the data for these simulations was generated from mixture models of multivariate Gaussian distributions of 12 dimensions (same number as in our real-life temperament dataset in Chapter 4.2).

The cluster centers were generated from a spherical Gaussian distribution of zero mean and variances of one. Covariances were sampled from the experimental distribution of covariances of normalized variables in two real-life datasets (those described in chapters 4.2 and 4.3). If this sampling gave a matrix that was not positive definite (as covariance matrices must be) the closest positive definite matrix was used (computed by the `netlab` [Nab01] implementation of the algorithm described in [Luc01]).

For class-valued data, the obtained values where then discretized into five according to percentiles (20, 40, 60, 80). This was done rather than directly generating class-valued data because our impression is that a lot of actual medical class-valued data actually comes from a continuous but unmeasurable distribution (for example, severity of pain on a scale of 1-5).

In 'clean' datasets, models of three centers were used without added noise. In 'noisy' datasets, one fifth of the data points were replaced by points sampled from a uniform distribution over a ball with a zero mean and radius of the maximum distance of any point from zero in the original sample. Datasets with 'no structure' have only one component in the mixture model, with zero mean and covariances sampled from the real-life experimental distribution. In 'gradient' datasets, all points where sampled from a distribution elongated along one dimension.

Note that it would of course be possible to generate more complex artificial datasets, either by a more complex data generation method or by re-sampling heavily from real-life datasets. However, in most cases we did not want to do this, because of the uncertainty it causes for interpretation of whether results are due to method properties or some unknown feature of the data. Instead where feasible, the same simulation has been performed with an available, suitably modified real life data set, to give an idea on how the results obtained compare to those seen in realistic settings.

3.2 `BIC` score versus 10-fold cross-validation

In this section, we explore empirically the result that the Bayesian Information Criterion and the k-fold cross-validation criterion for model selection should asymptotically be the same. The key questions are on one hand how large N should be for the asymptotic result to hold, and on the other hand which criterion seems better for realistic N. (In the studies on real life data presented in this thesis, both criteria have been successfully used.)

For each scenario (clean, noisy, and no-structure, as explained in the previous Section), five samples of each size were created for $N = \{100, 200, 300, 400, 500, 600, 700, 800, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$ to cover realistic dataset sizes for the types of data analyzed in this thesis. In addition, random samples of $N = \{100, 200, 300, ..., 2000\}$ were drawn from a real life dataset (the replication data used and described in Chapter 4.2).

A mixture model of Gaussians was then fit to the data using the Expectation-Maximization (EM) algorithm (see Section 2.2). Random starts were repeated until no improvement was found in five consequent restarts and the best model was selected. This process was repeated for k from 1 to 5 and the scores calculated for each.

Bayesian Information Criterion (Section 2.3.2) was calculated assuming model parameters for the centers of the model and full covariance matrices. Ten-fold cross-validation (Section 2.3.3) was performed and the sum of the test set data likelihoods calculated on the best model found for each training set.

Figure 3.1 shows the average BIC and 10-fold cross-validation scores per individual, over different N, for different dataset types for the three-cluster models. The pattern is similar for any other number of clusters (not shown). We can see that when N reaches 1000, regardless of the dataset type, BIC and 10-fold cross-validation give roughly equal results. Some difference remains in the case of the generated datasets with noise even for $N = 10,000$. One can speculate that this might be due to the noise being modeled slightly differently for each cross-validation round, and suggest that if both BIC and 10-fold cross-validation scores are calculated for some dataset and a difference found, there might be heavy noise present.

Figure 3.2 shows the average best number of clusters predicted by each method for scenario, per N. For the real life data, the number of clusters predicted by both methods for $N = 2000$ is three, but BIC stabilizes to this value sooner than 10-fold cross-validation. Of course, we do not know the real number of clusters for this data, though the real-life analysis suggests that numbers 2-4 can be considered useful. We can see that also for the artificial data with cluster structure, for low N, 10-fold cross-validation predicts cluster numbers lower than BIC.

We cannot consider the occasional prediction of four clusters for the noisy data true "over-estimation", as typically the fourth cluster is an attempt to predict the noise. The more there are data points, the likelier the "background noise" is to get its own cluster. For the data without real cluster structure, BIC consistently predicts 2 or more clusters for even for high N, while 10-fold cross-validation does not err from the (correct)

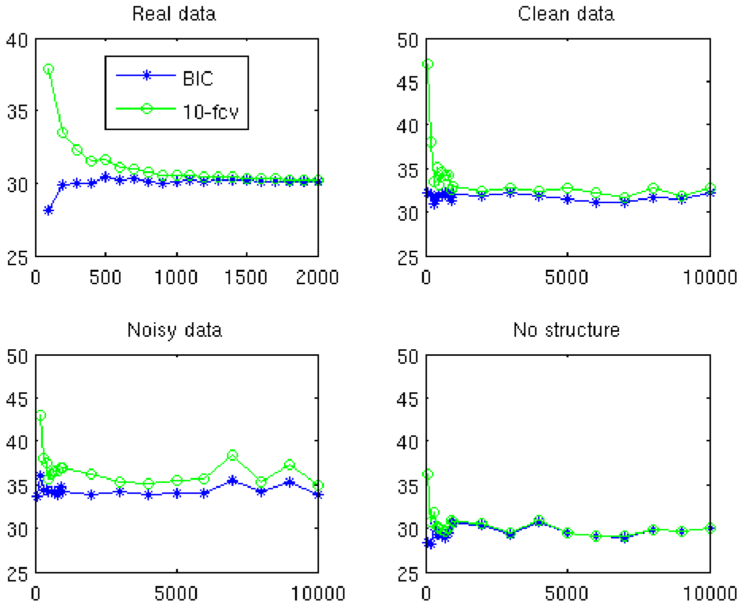

Figure 3.1: Average BIC and 10-fold cross-validation scores per individual over different N, for different scenarios. Note that the X-axis is different for the real data case than the three artificial-data cases. Convergence is clearly seen around $N = 1,000$, except in the presence of heavy noise.

estimate of one even once.

As a conclusion we can say that in the presence of a cluster structure, for low N, BIC is more likely to show the true number of clusters while k-fold cross-validation is likely to underestimate it. However, in the absence of structure, BIC is likely to overestimate the number. As it is not possible to know, for real-life settings, whether a clustering structure exists or not, and what a sufficient N for the convergence of the scores in the presence of one is, we cannot really recommend one over the other in the general case. In any specific case, we need to consider which type of error (predicting a cluster structure in the absence of one, or underestimating the number of true clusters) is more preferable to avoid.

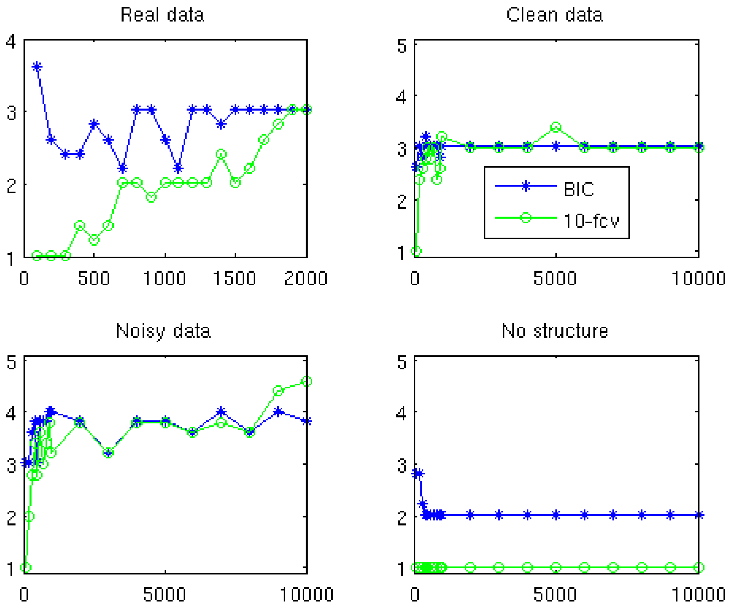

Figure 3.2: Average number of clusters predicted by the BIC and 10-fold cross-validation scores for different N. True number of clusters is unknown for real data, three for clean data, three or four for noisy data (depending on if one is used to model noise), and one for data without structure. For low N, 10-fold cross-validation tends to underestimate the number of clusters in the presence of a clear cluster structure, while the BIC score seems to overestimate the number in the absence of structure.

3.3 Natural hierarchies

In this section, we explore the idea that if a real subgroup structure (as opposed to say a single cloud or a gradient without clear demarcations) exists in the data, some non-hierarchical clustering algorithms can be expected to give hierarchical structured clusterings when the number of clusters changes. That is, from the best clustering into k we can arrive to the best clustering into $k+1$ by splitting one of the clusters. We initially observed this to be the case on many real life datasets, and it intuitively makes sense.

Assume that a dataset consists of three distinct, spherical clusters. These clusters are trivially found by the standard algorithms for $k = 3$. If one attempts to fit a k-means cluster model for $k = 4$, it seems unlikely that a model using a cluster to mix two of the real clusters would obtain a better score than one which simply assigned one center to two of the real clusters each and two centers for one of them. On the other hand, if the data, instead of clear demarcated subgroups, contains a gradient, it might make sense to use whatever k you have to evenly cover the whole range of data. Thus one would expect that the clusterings for $k = 3$ and $k = 4$ do not overlap hierarchically. Similarly, in the case of a mixture model of Gaussians, for $k = 3$ each real cluster should gets its own distribution and for $k = 4$ the maximum likelihood is achieved by simply allocating two distributions for one of the clusters. In the case of no real clusters the distributions end up overlapping and changing places arbitrarily when cluster number increases. (See Figure 3.3 for an example.)

Since the standard procedure for non-hierarchical clustering already includes building a model for various values of k and using some scoring system to pick among those, calculating such a similarity comparison for clusterings with k and $k+1$ clusters does not add to the complexity of the clustering process.

To test this idea, we created 100 artificial datasets of 2000 points. Of those, there were 25 of each of the clean, noisy, and gradient cases as described in Chapter 3.1. The last 25 datasets were purposefully designed to confound the k-means clustering process, consisting of points randomly spread on the surfaces of four 12-dimensional spheres of diameters 1 to 3 ("Concentric" below). All datasets where clustered with 1) fitting a mixture model of Gaussian distributions with the EM algorithm, and 2) the k-means algorithm, for cluster numbers 2 to 8. The adjusted mutual information (Section 2.4.3) of clusterings into k and $k+1$ was then calculated for each dataset to compare the clusterings.

Figures 3.4 (Mixtures of Gaussians) and 3.5 (k-means) show the adjusted

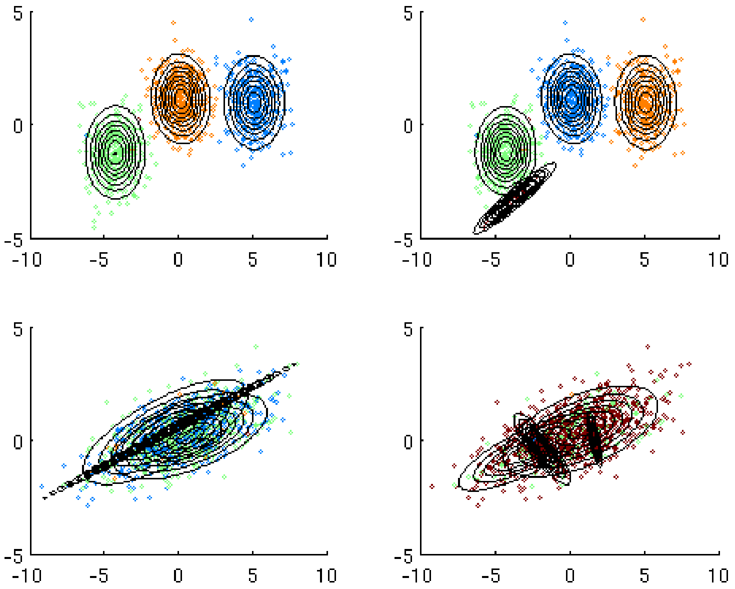

Figure 3.3: A clustering of toy datasets with a mixture of Gaussians into 3 (left) and 4 (right) of two 2-dimensional artificial datasets, one with a clear cluster structure (top) and one without (bottom).

mutual information of k clusters compared to $k+1$ clusters for each of the 25 runs in the four scenarios. As can be seen, the mutual information between clusterings into subsequent k is clearly higher in datasets with real structure as compared to those without clear group structure (the ”Gradient” case).

In the Mixture Model case this difference is extreme, but it can be seen also in the k-means case where more model assumptions are violated. Since the k-means algorithm is able to produce only spherical clusters, in the gradient case it tends to fit models where these are evenly spread on the gradient, resulting with more overlap than in the fitting of mixtures of Gaussians. It is also noteworthy that even though the k-means algorithm is unable to reliably reproduce the concentric spheres as clusters, the similarity graphs are nevertheless able to somewhat differentiate between the datasets between this sort of structure and the datasets without any sort of cluster structure.

As for the Gaussian mixtures the similarity tends to peak around or right after the true number of clusters, this observation could also be used as a further model selection aid. Formulating an exact score with a cut-off similarity value for real-life datasets is non-trivial. It is not possible based on artificial experiments to decidedly say how much depends on the true number of clusters, their overlaps, various data parameters (such as numbers of observations or variables, and missing data proportions), model assumption violations, etc. Nevertheless, extremes of the scale of whichever comparison method used can clearly be considered indicative of the presence or absence of cluster structure. If desired, artificial datasets with parameters appropriate to data at hand and varying structure assumptions could be generated and experimental values obtained this way for each study separately.

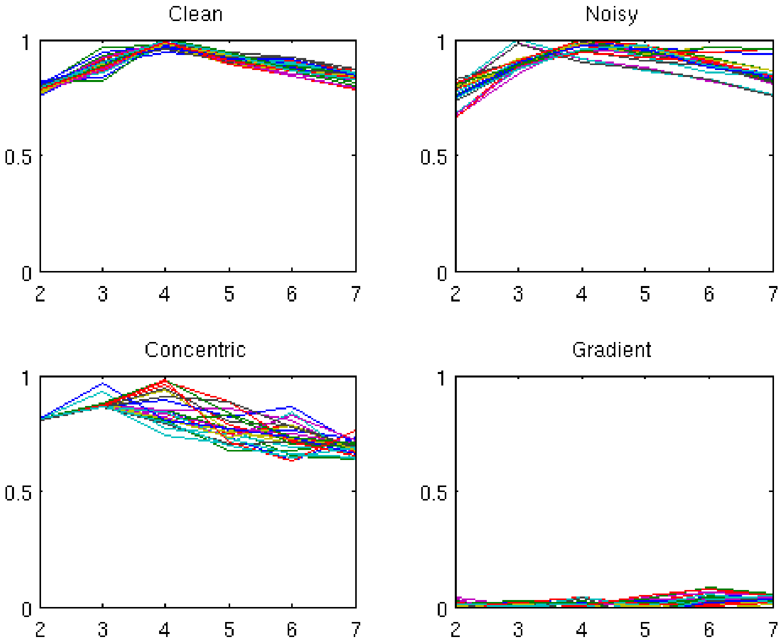

Figure 3.4: Adjusted mutual information (y-axis) between a clustering into k and clustering into $k + 1$ (x-axis) in each of the 25 datasets (one line per dataset) of four different scenarios (clean data conforming to model assumptions, same with added noise, concentric spheres, and gradient without cluster structure). Models fit with the EM algorithm for Gaussian Mixtures. Real number of clusters is three for the first three scenarios and one for the last one.

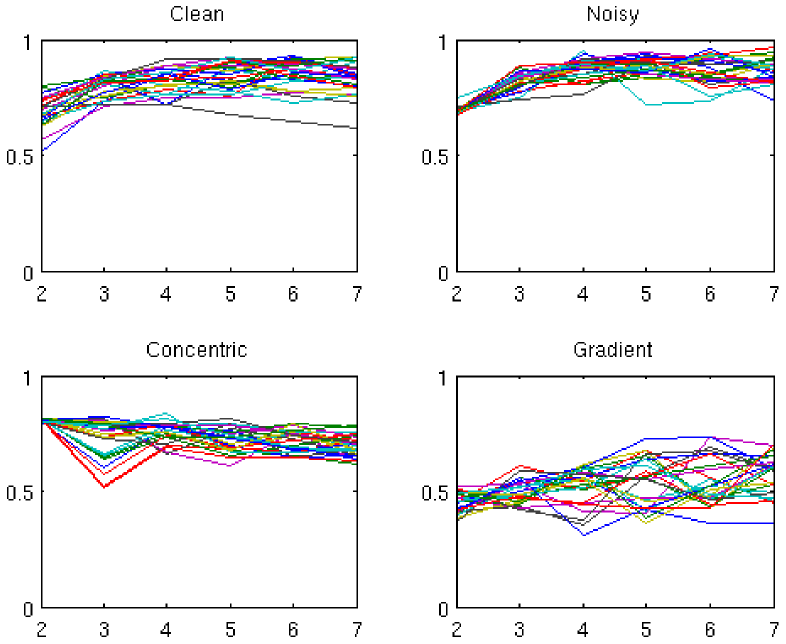

Figure 3.5: Adjusted mutual information (y-axis) between a clustering into k and clustering into $k+1$ (x-axis) in each of the 25 datasets (one line per dataset) of four different scenarios (clean data conforming to model assumptions, same with added noise, concentric spheres, and gradient without cluster structure). Models fit with the k-means algorithm. Real number of clusters is three for the first three scenarios and one for the last one.

3.4 Replicability on a separate sample

If a clustering is "real" in the sense that it reflects an underlying subgrouping of individuals in the original population, then two samples from that population should reveal a similar (though probably not, due to random effects, completely identical) clustering model. This idea was used in our analysis of temperament data, where a clustering model based on a sample from the Finnish population was validated by comparing to a clustering in a second sample. In this section we describe the simulations that are used to study whether this intuitive property actually holds. Such a check is crucial for the applicability of clustering methods.

Next we describe the process at a high level. We create three datasets, two by sampling from the same distribution and one from sampling from a similar class of distributions but with different parameters. Call these datasets the "original", "replication", and "independent" samples. We learn a clustering model based on each of the datasets. We then use the model built on the original dataset to give clustering labels to individuals in the replication and independent samples, and the models built on those two to give two clustering labels to individuals in the original dataset.

We combine these labels to obtain four clusterings: A1) labels for individuals in the original dataset and the replication dataset, based on the original model, A2) labels for individuals in the original dataset and the replication dataset, based on the replication model, B1) labels for individuals in the original dataset and the independent dataset, based on the original model and B2) labels for individuals in the original dataset and the independent dataset, based on the independent model. Comparing a similarity score between A1 and A2 to that of between B1 and B2 we get insight to whether two clusterings from the same population can be expected to be more similar to each other than two clusterings from different populations, given that the populations have an underlying clustering structure.

First, data was generated from a mixture of multivariate normal distributions as described in 3.1, for a fixed number $k = \{3, 4, 5\}$ of true centers. This distribution was sampled for two datasets of size 1000 each ("original" and "replication" samples). In addition, another ("independent") sample was created from another distribution with the number of centers randomly selected to be between $k - 1$ and $k + 1$ (inclusive).

These datasets were clustered independently using the `EM` algorithm to fit mixtures of Gaussian distributions (see Section 2.2). The best k for the original sample was determined by the Bayesian information criterion (Section 2.3.2), and the replication and independent sample where forced to the same k.

For to compare this to a situation where there is no cluster structure in the original population, data was generated from a single multivariate normal distribution for the original and replication samples, and the independent dataset was generated from a distribution of 1–5 centers, at random. These data were similarly clustered, but k forced in turns to 3, 4, or 5, in order to simulate a situation where we mistakenly identify a cluster structure that is not there.

Alternative clusterings for original/replication and original/independent datasets were compared by cross-tabulating the cluster labels and calculating the χ^2 statistic. One hundred experiments were performed for existing cluster structure and for the case of no cluster structure in the original, and for each possible value of k involved.

Figures 3.6 to 3.8 show observed χ^2 values for the four cases (original population does or does not have a cluster structure, second sample is a replication or an independent sample) for $k = 3$ to 5. Note that this k refers to the true number of clusters in the case where cluster structure exists and to the arbitrarily picked k in the case of no cluster structure.

We see the basically same phenomenon in all of the cases. Where there is no cluster structure and the replication dataset comes from an independent sample, the statistics stays below $N = 1000$. Where there is a cluster structure but the second sample is an independent one, or where there is no cluster structure but the replication is from the same distribution, we see a distribution of observed χ^2 that resembles the theoretical χ^2 distribution, maximum observed values falling at about $2 * N = 2000$, which is what you would expect for two random clusterings.

In the case of a replication from the same distribution with a cluster structure, we see a spread-out of values, going up to extremes close to what we would expect with a perfect replication. The higher the number of clusters, the more there are poorer-quality replications. Since our N does not increase with k, this is to be expected: the smaller N/k, the more likely it is that the sample contains only a small amount of representatives from a particular cluster, making clustering more random. In any case, the distribution does not significantly overlap with those obtained in other scenarios.

This means that should we observe a high similarity between the replication clusterings, we can fairly safely assume that 1) there is a clustering structure in the population, and 2) we have managed to replicate the sampling procedure accurately. However, should we observe a low similarity, we cannot based on it alone deduce whether this is because of lack of a real cluster structure or failure of replication.

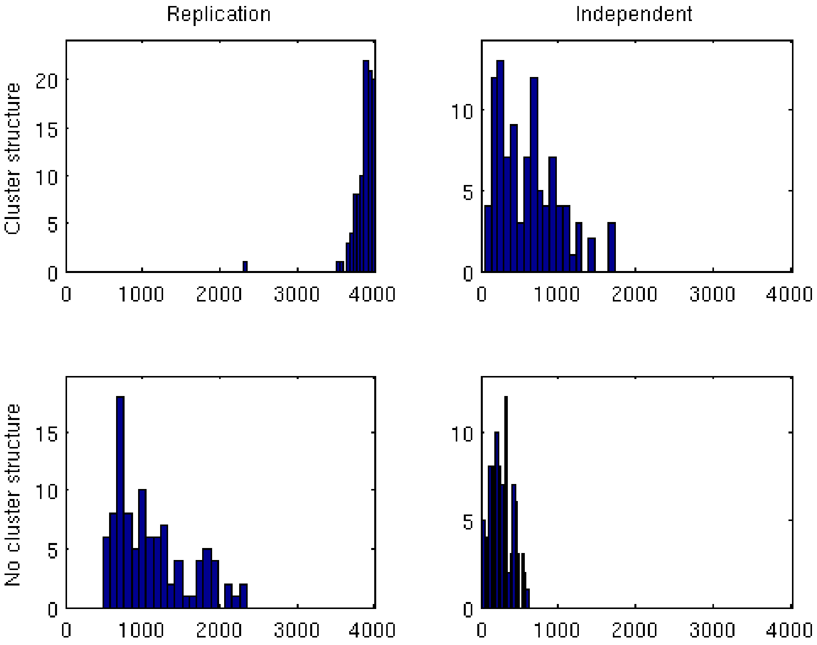

Figure 3.6: Histograms of chi-square values comparing clustering based on a model obtained on a sample from a distribution with or without cluster structure, to either a model obtained on another sample from the same distribution (replication) or to a model obtained with a sample from a different distribution with cluster structure (independent). Three clusters in the original distribution.

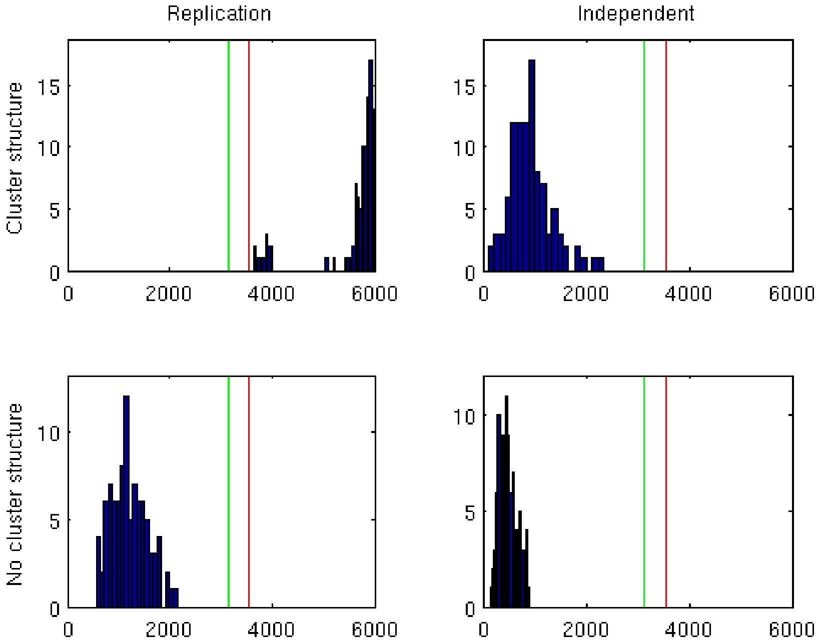

Figure 3.7: Histograms of chi-square values comparing clustering based on a model obtained on a sample from a distribution with or without cluster structure, to either a model obtained on another sample from the same distribution (replication) or to a model obtained with a sample from a different distribution with cluster structure (independent). Four clusters in the original distribution. Red and green lines show observations from a real life replication experiment. [WSM$^+$12]

For the sake of interest, in the case of four clusters, I have shown two values observed in a real life dataset (described in [WSM$^+$12] and Chapter 4.2). In the case of real data, what structures exist, they probably do not conform very closely to the model assumptions. This might explain why we observe values that are between the observed distributions for the different simulated scenarios.

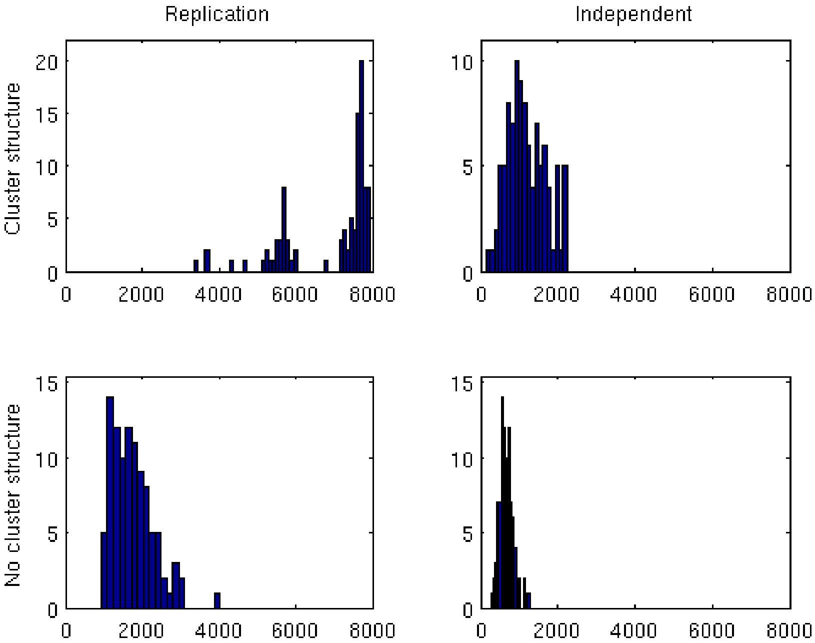

Figure 3.8: Histograms of chi-square values comparing clustering based on a model obtained on a sample from a distribution with or without cluster structure, to either a model obtained on another sample from the same distribution (replication) or to a model obtained with a sample from a different distribution with cluster structure (independent). Five clusters in the original distribution.

3.5 Effects of missing data

In medical datasets missing data is very common. This is usually handled by more or less refined imputation methods (see Section 2.2.6), but the effect on results is not very well understood. In the experiments of this section, we progressively remove more and more points at random from the data, cluster the remaining points, and compare the results to the clustering obtained on full data. The generated datasets are either mixture models of Gaussians, or mixture models of multinomials. In the first case, missing data is treated as missing in the `EM` sense and imputed on each cycle; in the second case missing data is treated as an additional value for each variable.

The aim is to find out how sensitive these methods are for missing data, and whether there is a difference in the effect of missing data in the case of data with a real clustering structure, and data without structure.

For the real life data, I used the temperament dataset from Section 4.2 for the Gaussian case, and a migraine dataset described in Section 4.3 for the multinomial one. The migraine dataset has over 6000 individuals and 194 variables, but a lot of missing data. A greedy selection process yielded 1000 individuals and 83 variables such that the initial dataset is complete. The temperament dataset has complete data for the 12 variables for 2000+ individuals, of which a random 1000 was selected as a test case.

Six different kinds of artificial datasets were generated: the clean, noisy, and no-structure scenarios described in Chapter 3.1, for both Gaussian and class-valued data. There were ten datasets of each kind.

A reference clustering for each dataset was first found. Number of clusters from 2 to 5 was tried (even though 1 is the correct answer for some datasets, it was excluded, as obviously the similarity between clusterings into 1 is always perfect and so it is not of interest). Restarts from random starting points were performed for each number of clusters until data likelihood did not improve on six consequent restarts. The best overall model was chosen from among the best clusterings into 2 to 5 clusters by using the Bayesian information criterion (Section 2.3.2). Re-clusterings of each dataset with increasing number of missing data were then performed with the same procedure.

First re-clustering of each dataset was performed with the full data, to give a baseline of what to expect from replicability of results in general. Each dataset was then re-clustered with progressive ten of percent of observed values in the dataset removed at each stage. In these re-clusterings, the cluster number was assumed to be known and equivalent to the one in the reference clustering. Two kinds of removal simulations were performed. In the first case, observations were removed by random. In the second case, to achieve a more realistic pattern of missingness, observations in a block in the data consisting of half of the columns and quarter of the rows was given a weighted probability of missingness (five times that of other values).

Figure 3.9 shows the average adjusted mutual information (Section 2.4.3) results for the discretized and Gaussian data, for both random and non-random missingness.

The good news is that even when data is missing not at random, the situation is not much worse than it missing at random. For the discrete data case, the situation seems very good: modelling the missing data as another value for the clusters does not disturb the models much when cluster structure is actually present in the data. For the Gaussian case, where missing data is imputed on each round of the `EM`-algorithm, similarity is relative to missing data proportion.

We can also see that the reference clustering for the artificial datasets with no real clustering fails to be replicated at all, even in the case of no missing data. (Note that in a real life situation, the `BIC` might have suggested selecting just one cluster as the correct model for this data; however, as stability results for just one cluster do not make sense, that possibility is excluded here, leading to totally arbitrary two-cluster solutions and hence no similarity.)

One should note that it is difficult to draw conclusions of the quality of the data or interestingness of the clusters based on these curves. For the class-valued data, the adjusted mutual information `AMI` for the dataset with most intentional noise falls faster initially than the clean dataset, but ends up higher, and for the Gaussian case, while it starts slightly lower, it remains higher when data points are removed. This is most likely due to the fact that the algorithm ends up modeling noise with a fairly stable cluster, and reminds us that a stable cluster is not necessarily an interesting one.

It would be tempting to draw from this the conclusion that you actually gain by discretizing, since the similarities for the discretized case are so much higher than those of the Gaussian case. This would, however, be incorrect, as each case is compared to the reference clustering obtained by said method on full data, instead of the true clustering in the data-generating model.

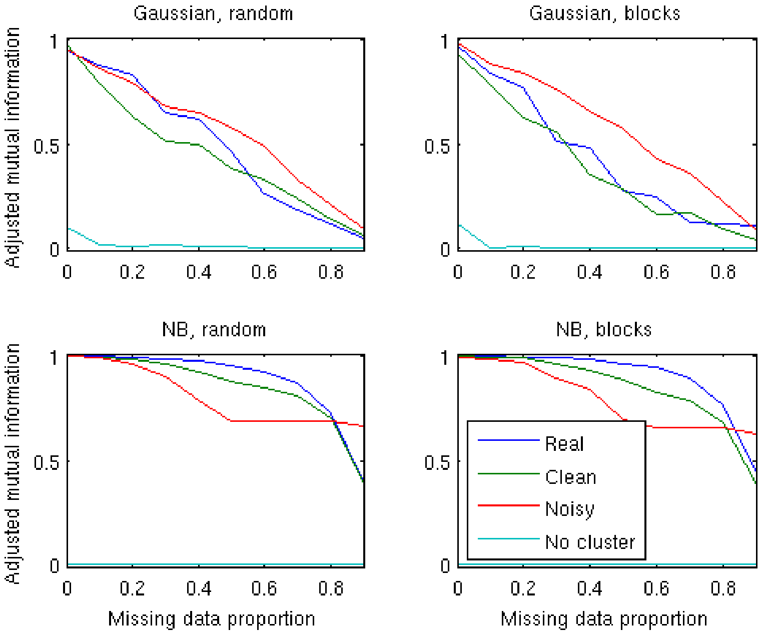

Figure 3.9: Adjusted mutual information for re-clusterings with various percentages of missing data. Top figures: mixture models of Gaussians, with missing data treated as missing in the `EM`-sense. Bottom figures: the discrete naive Bayes model, with missing data treated as another value for the variable. On the left, data missing at random. On the right, blocks of data having higher probability of missingness. X-axis, missing data proportion. Y-axis, adjusted mutual information. Four different simulation scenarios were considered: based on real data, a 'clean' model conforming to model assumptions, one with added noise, and one without cluster structure. Note that the Gaussian and Naive Bayes cases are not comparable.

3.6 Stability analysis by random drops

In the real data analyses described in this thesis, we have often utilized the idea, explained in detail in Chapter 2.5.2, that a clustering that reflects a true phenomenon in the underlying population rather than over-fitting into features of the data should not be sensitive to removal of individuals at random. In this section, we study this phenomenon on artificial data.

Artificial datasets of $N = 2000$ were generated as described in Chapter 3.1. Each dataset was clustered with a Gaussian mixture model. Each dataset was first clustered as is to get a reference clustering, and then re-clustered first intact and then with a 10 percent of rows removed at a time, until only 200 rows remained. There were 25 datasets of each kind, 100 sets in total. Note that the difference of these experiments to the missing data ones reported in the previous section is that here we remove complete rows as opposed to individual points of data. For real-life data, we used a sample of $2,000$ data points from the temperament dataset described in Chapter 4.2.

As Figure 3.10 shows, when there is a clear cluster structure in the data, with these parameters, the clusterings can be expected to resemble the one obtained on full data for until about $N = 400$, with Adjusted Mutual Information of above 0.9 for until about half of the data points are removed. For the artificial datasets we see a pattern where the similarity falls relatively steadily when the missing data proportion increases. For the real-life simulation case, we see a pattern where the similarity first stays high and then falls abruptly. This or the complete absence of stability is in our experience the typical pattern for real-life data in general. Gradient-type data and and data without any structure behave similarly with each other, `AMI` staying close to zero (implying similarity no higher than that which would be expected by chance) all the time.

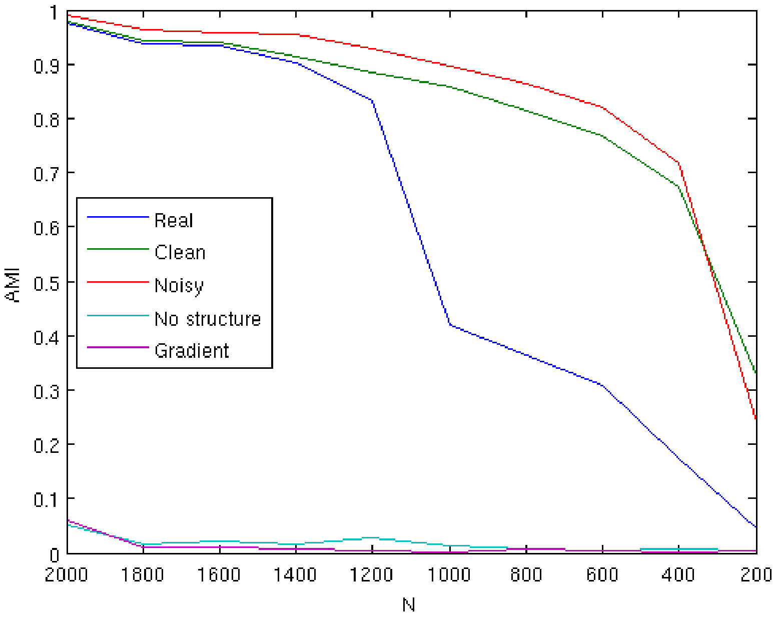

Figure 3.10: Adjusted mutual information to the reference clustering on full data, when an increasing amount of data rows are removed. Average over 25 simulated datasets is shown. X-axis shows remaining number of rows in the dataset, Y-axis shows the adjusted mutual information between the clusterings on the reduced dataset and the original.

Chapter 4

Studies on real data

> *"As a physician and as a pilot, I think it lets me be a pretty good translator having one foot in the medical world and one foot in the flying world. Sometimes when the medical guys come in and speak medical stuff to the pilots, the pilots really don't know what they're saying."*
> (David M. Brown)

In this chapter we consider some studies where clustering has been used in the analysis of medical data. We describe the underlying medical phenomena and the datasets, present some of the findings, and describe the methodological implications of the studies.

4.1 Case 1: Schizophrenia subtypes

In this study, we performed a clustering of individuals with schizophrenia, and their relatives (some healthy, some with schizophrenia or another mental health problem), based on various measures of symptoms and signs of disease. The aim, which was realized, was to identify subgroups of sufferers with specific genetic associations. Three candidate genes were analyzed and we showed an association of one of them (*DTNBP1*) to a specific subtype of schizophrenia, characterized by lack of emotional symptoms and a more severe course of illness. The results of this study have been published in [WPTH$^+$09].

4.1.1 Background

Schizophrenia (from Greek *skhizein*, "to split", and *phren*, "mind") is a chronic psychotic disorder with varying course and symptoms. Its most common symptoms include hallucinations (typically auditory, but also others), paranoid or otherwise bizarre delusions, and disorganized thought and speech. It is often associated with significant social and cognitive disability and blunted emotional states [KNW95].

Diagnosis of schizophrenia is based on anamnesis and status, that is, descriptions of experiences by the patient and possibly his family, and observations by professionals. Laboratory, brain imaging, or other measurement tests for schizophrenia do not exist, but several structured, scored interview procedures and structured ways to describe past symptoms have been devised to aid with diagnosis and give measurable definitions for research. The diagnostic criteria for schizophrenia have been set for clinical practice, but genetic liability seems to extend to a broader phenotype with abnormalities in multiple dimensions, such as cognitive capabilities [KNW95, TSF00].

The disease has been known for long to have a familial, most likely genetic, association [KMG$^+$93, CKL$^+$98, TWM$^+$94]. However, genetic studies have failed so far to bring forth a clear picture on which genes are involved [FK08]. Some promising genes, like *DISC1*, have been found to not be associated only to schizophrenia but also to a wide variety of other mental disorders [CBS$^+$08]. Others, like *DTNBP1*, seem to show association to specific symptoms in schizophrenia [FK08] and the association to the disease itself has been replicated in some populations and studies [FFP$^+$04, VCS$^+$08], but not in others [FFP$^+$04, SDL$^+$08]. To understand this heterogeneity, both on the phenotype and the genetic levels, efforts have been targeted to identify endophenotypes [GG03] or other more homogeneous subgroups [COO06, HKB$^+$05, PTHH$^+$04, CDN$^+$08], with the hope that such subgroups would be more directly associated with underlying biology.

In this study, we had the luxury of having a dataset where both information about a wide variety of possibly related phenotypes and genetic data about promising candidate genes was available. This allowed us to first seek for clusters of the phenotypes, and then analyze how those phenotype groups are associated to the genetic data. Comparing this to associations to diagnoses, we could shed some light on reasons for the unclear findings of previous genetic studies. Latent class analysis has been previously used to investigate symptoms of psychotic illnesses [CSWM94, MNL$^+$04, MMM$^+$05], but rarely to investigate the broader phenotypes related to schizophrenia [LMR07].

4.1.2 Data

Data in this study [HVS$^+$99, VLH$^+$00, WPTH$^+$09] consisted of 904 individuals and 203 variables, most of them binary. Most individuals had at least one other family member in the dataset. There were 288 families in total. For details of the data, see [HVS$^+$99, VLH$^+$00, WPTH$^+$09].

The clustering dataset had 203 variables. The first 73 of them were binary variables describing the presence or absence of symptoms during the patient's lifetime. A structured interview had provided 111 variables of mostly of class nature, and finally, 17 continuous variables described the results of neuropsychological testing. See [WPTH$^+$09] and the supplements of the article for details. Genetic information was not used for clustering, but analyzed for its association to the clusters. The genetic dataset consisted of 53 single nucleotide polymorphisms (essentially, binary variables) over three different candidate genes.

4.1.3 Methods

Variables with continuous scales, or ordered scales with more than 10 different values, were discretized into variables with five different values, using the 20th, 40th, 60th and 80th percentile as the cut-off points. Missing values were assumed to form an additional class for each variable. Pairs of binary variables were combined to form new variables with 8 values (missing values were considered a third value). This was done in a "greedy" manner, always combining first the pair with highest correlation. This reduces the number of variables and aims to reduce the problems resulting from the violation of the independence assumptions in the model. No other pre-processing of the data was performed.

The data were clustered using a Naïve Bayes mixture model (see Chapter 2.2). Family information, direct information on existing diagnoses, or or any genetic variant information, were not used for clustering. The number of clusters was chosen by the 10-fold cross-validation procedure. Cluster robustness was tested by random removals of individuals and variables (separately), as described in Chapter 2.5.2.

We also performed the clustering separately for males and females, and separately for individuals in the two cohorts of the data. These results are practically equivalent to the one obtained by full data, within at most 10 individuals clustered differently, and will not be presented.

The association of the clusters to the candidate gene single-nucleotide polymorphisms was analyzed after clustering with the program Mendel [LSS05], using its allele-sharing non-parametric linkage option. In these analyses, we considered one of the psychosis groups of the three-cluster solution (see below) as affected, the non-psychosis group as affected, and the third cluster unknown. For the clinical diagnoses, a DSM-IV [Ame00] based definition for schizophrenia spectrum disorder was used as definition

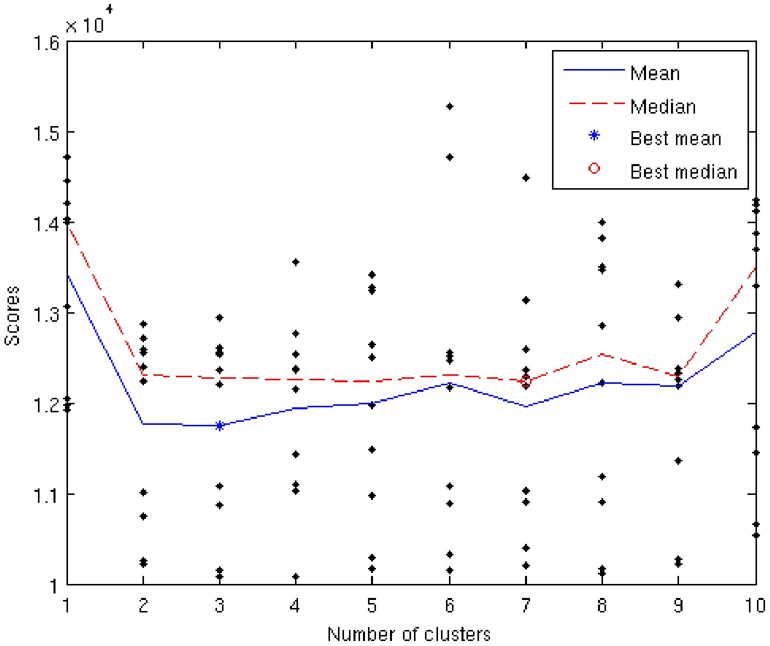

Figure 4.1: 10-fold cross-validation scores for the schizophrenia clusters. X-axis, number of clusters. Y-axis, data log-likelihood score in 10-fold cross-validation (smaller is better). Dots show individual experiments, and in addition mean and median are shown.

of affected status. The p-values presented in the article were obtained via permutation and corrected for multiple testing by the Bonferroni correction.

4.1.4 Clustering results

Figure 4.1 shows the 10-fold cross-validation scores over different number of clusters (the smaller the score, the better the test-set data likelihood given the training-set model). We can see that the scores are very similar for some variety in the number of clusters. This is related to the fact that the alternate models at the flat area of the mean and median curves are hierarchical refinements of each other. We can see this from Figure 4.2: the adjusted mutual information between subsequent number of clusters is always above 0.5 (on this scale, one means complete similarity and zero means similarity between random clusterings of the size).

It also turned out that beyond three clusters, the additional clusters are almost all small subgroups that would not be of interest or practical in

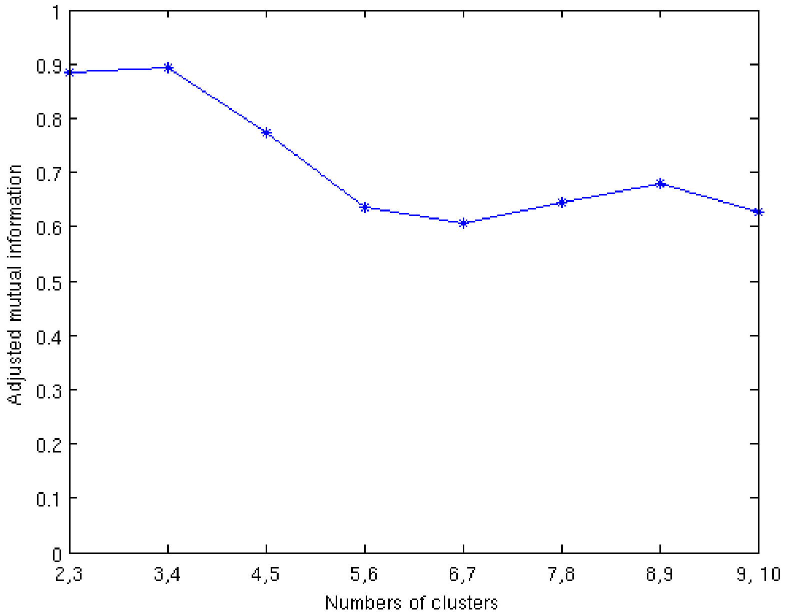

Figure 4.2: Adjusted mutual information between models with subsequent number of clusters k for the schizophrenia data. X-axis, numbers of clusters for the models compared. Y-axis, adjusted mutual information: value of 1 signifies identical models, and value of 0 signifies expected value for random clusterings of similar size.

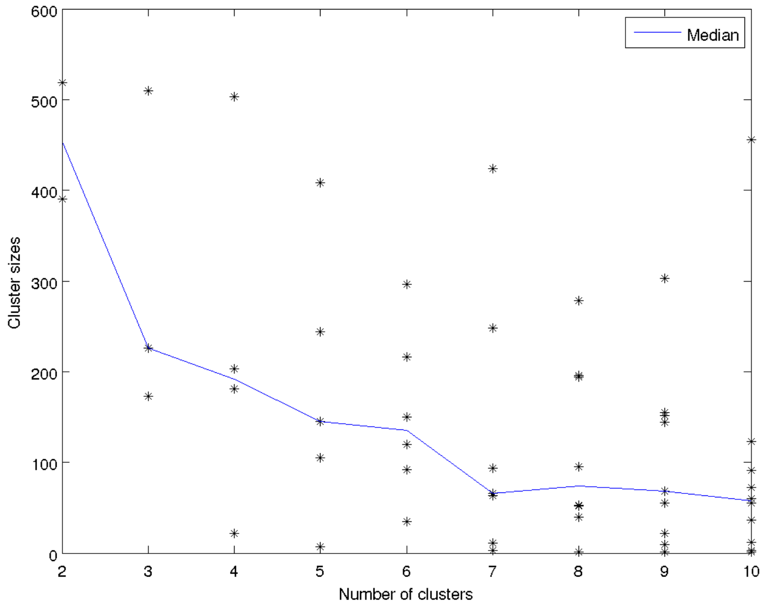

Figure 4.3: Cluster sizes for different number of clusters in the schizophrenia dataset. X-axis, number of clusters. Y-axis, number of individuals in a cluster. Shown are values for each cluster in black and the median cluster size for each model.

further analysis. As Figure 4.3 shows, when the cluster size goes up, many very small clusters are added. In fact (compare to the previous figure), these models tend to be approximate hierarchical refinements of each other. For example, the model with four clusters can be formed from the model with three clusters simply by separating a cluster of 22 individuals from one of the clusters and moving nine other individuals from one cluster to another (see Figure 4.4. On the other hand, the clustering into two is a fairly trivial division of cases into psychotic and non-psychotic and hence, while obviously valid, hardly interesting in the medical sense. We ended up looking further into the model with three clusters, which turned out to provide an interesting division of the psychotic cases into two groups (as explained later).

To analyze the stability of this model, we formed a clustering into three with random subsets of the cases, and compared this to the original clustering. Figure 4.5 shows the results of this experiment. We also performed a similar experiment removing variables instead of cases, the results of which are

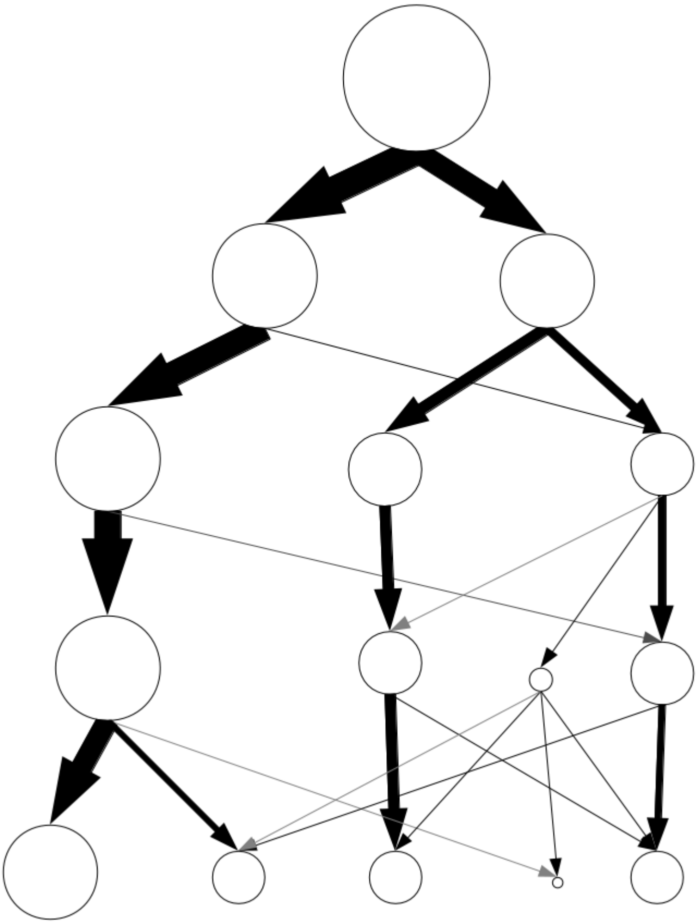

Figure 4.4: The hierarchy of clusterings into subsequent number of clusters in the schizophrenia study. Each row of circles corresponds to a clustering into as many clusters. Circle areas are relative to number of individuals in the cluster. Arrow widths are relative to number of individuals that are in both of the two clusters connected. Grey arrows correspond to less than five individuals.

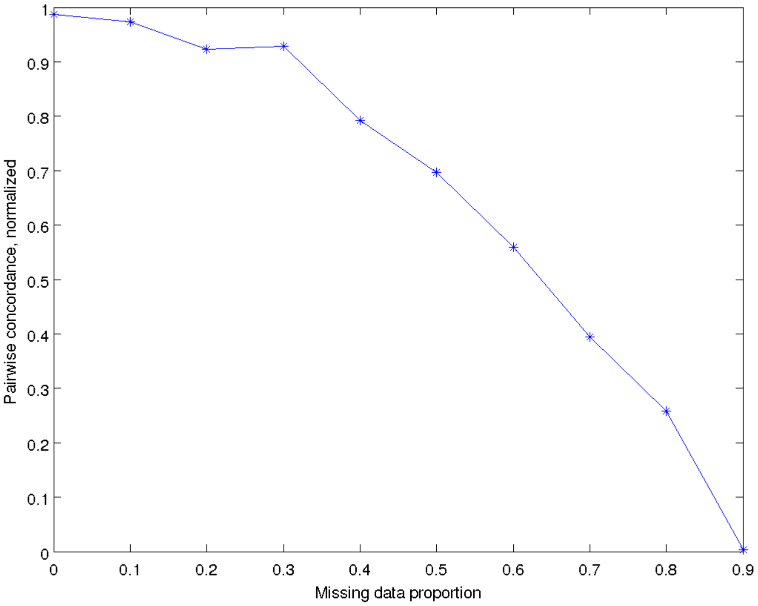

Figure 4.5: Results of the experiment where cases were randomly removed from the schizophrenia set and a new clustering obtained. X-axis, percentage of data removed. Y-axis, normalized pairwise concordance between the original clustering and the reduced dataset; value of 1 signifies full similarity and a value of 0 the similarity of random assignment to clusters of similar size. The error bars show two standard deviations over 10 repeats of the experiment for each percentage.

shown in 4.6.

The first phenomenon we can see that there are a certain number of cases that are "between clusters" so that the exact assignment of them varies even with the full data. This is an expected feature given the algorithm used: the restarting point is random and we stop the iteration process when the likelihood changes very little, meaning that individuals whose likelihood given two different clusters is similar might end up in a different cluster. As can be seen from the similarity being almost exactly one, though, the number of such individuals in this dataset is very low (below 10).

Second, we can see that as long as we have about 70 % of the individuals we receive reasonably similar clusterings, and that the similarity starts to steeply decline after that. This suggests that the clustering is not dependent on some outlier cases but actually represents a real phenomenon in the

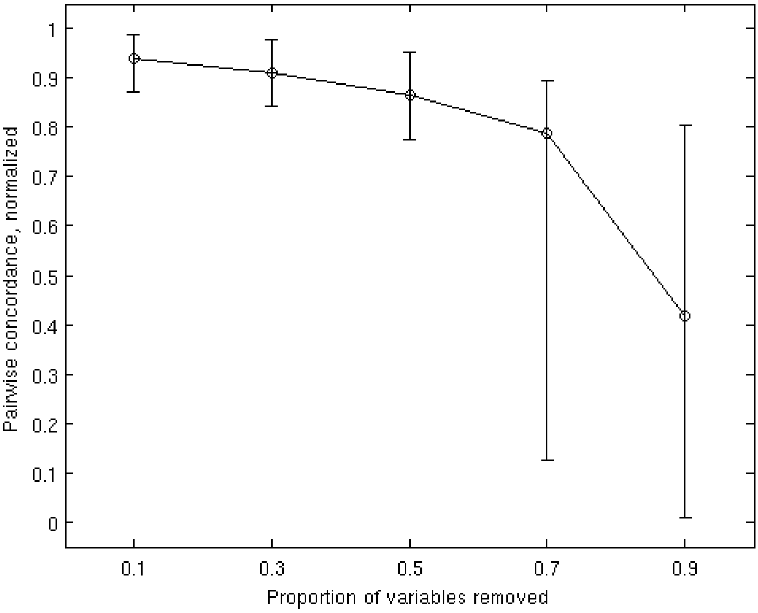

Figure 4.6: Results of the experiment where variables were randomly removed from the schizophrenia set and a new clustering obtained. X-axis, percentage of data rows removed. Y-axis, normalized pairwise concordance between the original clustering and the reduced dataset; value of 1 signifies full similarity and a value of 0 the similarity of random assignment to clusters of similar size. The error bars show minimum and maximum observed over 10 repeats.

data. Also, the amount of data is enough for this kind of analysis, with a certain safety buffer, but that the amount could not have been significantly smaller without affecting the results. We can also see that about half of the variables can be removed without the results changing much, but when 70 percent of them are removed, the quality of the clustering suffers a lot.

4.1.5 Medical implications

For detailed discussion of the results, see [WPTH+09].

We identified three clusters of individuals; 1) one that we ended up calling "core schizophrenia", 2) one for psychotic illness characterized by mood symptoms, and 3) one for individuals without psychotic symptoms. Looking at the clusters versus established diagnoses (Figure 4.7) we can see that the cluster division does not follow the "established" division of psychoses into schizophrenia spectrum and psychotic mood disorders. Many schizophrenia and schizophrenia spectrum cases cluster with individuals with mood disorders instead.

After noticing that the clusters do not conform to the diagnostic divisions, we analyzed the difference between individuals with schizophrenia in cluster 1 and those in cluster 2. It turns out that individuals with schizophrenia in cluster 1 had less mood symptoms, more cognitive impairments, and a more severe course of illness than individuals with schizophrenia in cluster 1. They seem to represent a "core schizophrenia" subgroup, while the schizophrenic individuals in cluster 2 represent "less severe schizophrenia with mood symptoms".

Of the candidate genes, we found in [WPTH+09] that *DISC1*, a well-established candidate gene for several mental disorders, is associated to the broader schizophrenia spectrum disorder, confirming previous results obtained also on this same dataset. However, *DTNBP1*, another candidate gene with more inconsistent results from previous studies, showed association to the "core schizophrenia" group but not to the broader diagnosis. This gene has also in other studies been shown to have an association to symptoms that characterize our "core schizophrenia", for example earlier age of onset, absence of manic symptoms, and more severe psychopathology [FK08].

This is consistent with the interpretation that the partially contradictory results of earlier studies (for example, [FFP+04, CDN+08, FK08, VCS+08, SDL+08]) are most likely partially caused by the association of candidate genes to different aspects or subtypes of schizophrenia. The conclusion can be drawn that attention to details of the disease phenotype, such as prevalence of mood symptoms, age of onset, severity of the disease, and cognitive functions, and analysis per subgroups based on them, is necessary

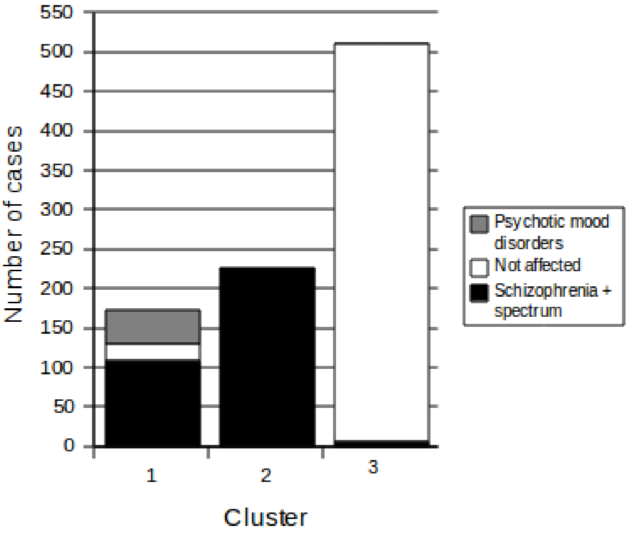

Figure 4.7: Clusters versus diagnostic groups in the schizophrenia study. The bars show the number of individuals in each cluster, colored by diagnostic group. We see that Cluster 3 consists mostly of individuals without a psychotic disorder and Cluster 2 of individuals with a schizophrenia spectrum disorder, while Cluster 1 has individuals from all groups. Figure reproduced from the article [WPTH$^+$09].

to understand the full picture of the genetics of schizophrenia.

However, the association of *DISC1* to a much broader genotype cautions against seeking for only one “true subtyping” or definition of the disease, narrowing down the search, since doing so would miss the role of that gene in a wide variety of psychiatric disorders. This implies that understanding the genetics of even the most hereditary psychiatric disorders will require big samples of detailed phenotypes from relatively homogeneous populations, which will provide challenges for data collection and management as well as statistical analysis.

4.1.6 Methodological implications

This study clearly demonstrated that clustering methods have their place in genetic studies of complex diseases. We were able to demonstrate that the dataset clusters to subgroups that are clinically meaningful. Moreover, one of these has a clear association to a known candidate gene that is not observed in the same dataset for the clinical diagnosis. Trying to reach the same conclusion by analyzing each of the multiple variables separately would have lead to dilution of the results after correction for multiple testing, and the results would have been inconclusive. The Naïve Bayes mixture model worked well in this case, despite the naive assumptions from which it gets its name.

We have also here confirmed in practice that cluster number selection based on the cross-validation techniques works. When such techniques do not give one clear minimum score, comparing solutions with subsequent number of clusters is a good idea. Here it turned out, as it did in our simulations (see Chapter 3.3), that these solutions are refinements of each other. Such an observation both gives us confidence that the observed clustering structure is real, and makes the choice between these solutions less critical.

Naturally, we nevertheless need to estimate the robustness of the model. Here we did it by dropping cases or variables in increasing amounts, and comparing the solutions obtained with partial data to the original one. When the mutual information between the original clustering and those obtained with minor parts of the data removed is high, we can conclude that the clustering solution is not likely dependent on some outlier observations. Here, we found that we can remove up to 30 percent of the cases and up to 50 percent of the variables without a large decline in the similarity of the clustering to the original.

The latter observation might pose the question whether there were ‘too many’ variables. Each variable increases the complexity of the model and the runtime of the analysis, and so if up to 50 percent of them are not needed to reconstruct the clustering, could we have done better with further dimension reduction before analysis? It is of course difficult to answer this question post-hoc, but I would here say that since the models obtained with partial variable data were similar, but not the same, the additional variables still carried information.

One possible downside of this kind of analysis is the fairly long time it takes to perform all the calculations. The clustering process is iterative in itself, and even to obtain one clustering we need to restart the process several times because of the EMalgorithm requiring a random start that might potentially affect results (see Chapter 2.2.3). To select between cluster numbers, we need to rerun this process for each possible k, and in the random drop experiments then repeat all of that for all of the random partial datasets. This is by no means prohibitive though: on a Linux cluster, using Matlab's Parallel Computing Toolbox to run four labs in parallel, the computation time is in order of days rather than weeks.

The comparison of the clustering scheme to the established diagnostic scheme became very important here. Visualization and detailed analysis of the multiple variables required work, a lot of it manual. It also required some work and careful explanation of the genetic results to convince people that the point of the clustering was not as such to try and present a “better” or “more correct” way to group individuals than the diagnostic grouping, but rather that we demonstrated that probably no general-purpose grouping even exists.

4.2 Case 2: Temperament groups

In this study, we clustered individuals based on their answers to a questionnaire assessing temperament. The clustering structure was replicated in a separate sample. The individuals in each cluster were then compared regarding variables from various life domains, such as mental and physical health, marriage status, education and employment, as well as scores on other psychiatric scales. We were able to show that 1) the 12 temperament “scales” traditionally calculated from the questionnaire are not independent in the population, and 2) that people in different clusters have very different outcomes in life. This section is based on [WSM$^+$12].

4.2.1 Background

For a more detailed description of the area, see [WSM$^+$12]. Temperament refers to aspects of personality that are considered inherited or innate, rather than learned (e.g. [ZB08], [Clo87]). It consists of early-appearing variation in general mood or emotional reactivity; the innate tendency of a person to react in a particular way in particular situations. In contrast, personality in addition to temperament includes learned responses, preferences, and opinion, while temperament is considered to be the biologically-based, inherited core of personality. Temperament affects how good a fit a person is to a particular society and social environment, and both temperament alone and this interaction are assumed to function as a predisposing or protecting factors to psychiatric disorders (see below).

Several approaches have been developed for classifying temperament and personality [ZB08]. A number of scaling systems are available for measuring temperament in adults. Such systems typically consist of a questionnaire, usually filled out by the person him/herself, containing questions assumed to be related to temperament. The questions are typically either binary ("Does the following statement describe you?") or answered with a small set of ordered options ("On a scale of 1 to 5, how well does the following statement describe you?"). The answers to these questions are then combined, by summing or averaging the individual answers to certain predefined (usually non-overlapping) sets of questions assumed to relate to the same facet of a temperament, into "scales". When (again, usually non-overlapping) subsets of questions belonging to one such scale are analyzed separately, they are termed "subscales".

One such scaling system is the Temperament and Character Inventory (TCI) [CvP93, CPvW94]. The version used in this study tests four temperament dimensions defined by basic stimulus-response characteristics: novelty seeking (NS), harm avoidance (HA), reward dependence (RD) and persistence (P). These four are further divided into 12 subscales. Individual scales measured by the TCI are normally distributed in the population, with sex-dependent differences [MKE+04, MVL+07].

For the role of temperament in mental and somatic health, see [SM06, SKM+07, HbJ+09], though the causal factors and directions are complex and unclear. In particular, a high level of HA has been related to a number of psychiatric disorders [RS04, EBS+04, BPBC96, FADA+02, KSB+04]. Studies have also consistently shown a genetic component for all domains of temperament, with heritability of 50 to 65% in Western populations [HCM94, SHC+96, AOY+02, GCHM03]. Research concerning the relationships between various domains of temperament has been less consistent, however [MLK+08].

In [WSM+12], we had the opportunity to cluster individuals from two Finnish populations sample based on TCI domains. This allowed for a comparison of cluster structures learned in the separate samples, an excellent way to validate the clustering. The goal was to explore whether specific subgroups of individuals with certain patterns of scores could be found, and if so, whether these groups differ with respect to other measurements. Such measurements were available for the first sample at the time of the study.

4.2.2 Data

The data in [WSM$^+$12] consisted of two datasets, the Northern Finland Birth Cohort 1966 [Ran69] (NFBC1966) and The Cardiovascular Risk in Young Finns study [RJR$^+$08] (YF).

The NFBC1966 is a prospective cohort study, originally including all live-born individuals from Oulu and Lapland whose expected year of birth was 1966 (N = 12,058) [Ran69]. The individuals have been followed up at several points in their lives, ranging from prenatal surveys of parents to a 31-year follow-up. The data used in this study comes from the 31-year surveys, in which a questionnaire assessing temperament was sent to all participants [SKM$^+$07], limited to those individuals with complete personality questionnaire, and individuals with mental retardation excluded (final N = 3,711).

The YF project started in 1980. In it, 3,596 individuals from six different age cohorts (3 to 18 years) from the Social Insurance Institution population registry were chosen randomly [RJR$^+$08]. The TCI scores for this study were measured in 2001 (N = 2,097 for those with complete scale data) [HbJ$^+$09].

The samples differ from each other in three major ways. 1) The age in NFBC1966 is 31 years for all participants, while participants in the YF study are between 24 and 40 years old, at the time of measurements. 2) NFBC1966 population covers the two Northernmost provinces of Finland, while YF individuals are sampled from all over Finland. 3) The version of TCI used in the two studies is not the same.

Age and geographic distribution turned out not to be significantly correlated to the TCI scores in YF, so we chose to simply the ignore age- and location-related differences.

The major difference between the TCI versions is that the one used in NFBC1966 questionnaire presents the questions in a binary format ("Does this statement describe you?") while the one used in YF presents them in a scaled format ("On a scale of 1-5, how well does this statement describe you?"). Hence, the summed scales are not directly comparable. Nevertheless, as for the k-means clustering algorithm the data were normalized to mean zero and variance 1, the models learned on one dataset could be directly applied to the normalized data of the other, and this turned out not to pose any problems.

The data on NFBC1966 concerning current life domains can be divided into five parts: 1) the so-called 15D questionnaire, assessing self-reported well-being in 15 domains (such as mobility, vision, hearing, sleeping, eating, depression, distress, sexual activity), 2) education and social status, 3) physical well-being and health, including life habits such as smoking and alcohol use, 4) confirmed diagnoses (both somatic and psychiatric), and 5) other psychological scales.

4.2.3 Methods

The data in [WSM+12] were clustered by the k-means procedure (see Chapter 2.2.5), and the number of clusters chosen by the Bayesian Information Criterion (Chapter 2.3.2) from among 2–12 clusters. (The k-means methods was able to, in this fairly normally distributed data without missing values sample, produce robust results in a short time.) The Euclidean distance between the 12-dimensional, normalized temperament subscale vectors was used as the similarity measure. Males and females were analyzed separately, as previous studies have established there are significant differences in scores between genders [MVL+07].

Clusterings between the NFBC1966 and YF data were compared as follows. First, clustering models were learned for both datasets independently and a cluster was assigned to each individual in that dataset. Second, a clustering was assigned to each individual based on the best model over-all obtained on the other dataset. Third, since the k obtained for each model by BIC did not agree, all individuals were assigned a cluster according to the model built on the dataset the individual belonged to with the best k on the *other* dataset. Thus, each individual had four cluster assignments: two based on the overall best models for each dataset, and two based on the best model on one data for the k decided based on the other data.

We then cross-tabulated these cluster assignments and compared the χ^2 statistics to those of random clusterings of individuals. Associations to life domain variables were assessed by χ^2 or one-way ANOVA statistic, whichever is appropriate, and corrected for multiple testing by the Bonferroni correction.

4.2.4 Results

A four-cluster solution had the smallest error for the NFBC1966 data, and a two-cluster solution for the (smaller) YF data. Tables 4.1 and 4.2 show cross-tabulations of the alternative models (for females and males). Cohen's

κ values are given, as they are the measurement for similarity of this kind typically used in the field. All these values pass the rules of thumb of 0.6 for "substantial agreement" and most of them are above the 0.8 limit for "almost perfect" agreement; as noted before, care must be taken in comparing two already high values with each other and we will refrain from doing so here.

We also calculated the χ^2 values for the four versus four cluster cross-tabulations for the combined datasets and compared them to values obtained for random clusterings. These experiments are described in Chapter 3.4. Figure 3.7 shows the value observed for males (red) and females (green), compared to values observed in various scenarios of replication. We can see that while the values do not reach those obtained on a full replication on artificial data that confirms to the model assumptions, they are also far from random.

Moreover, the patterns of the average TCI scales in each cluster were strikingly similar for males and females, so that it was possible to by simple visual inspection of the centers easily say which female cluster corresponds to which male cluster; Figure 4.8 shows the matched up clusters for the NFBC1966 data to illustrate this. (Cluster labels have been assigned to reflect this matching.)

For psychological interpretation of the clusters, see [WSM$^+$12].

4.2.5 Medical implications

It is usually assumed that the TCI subscales form independent, separate dimensions of temperament. That we can capture almost all associations to life domains with just four groups formed based on the 12 dimensions would suggest that instead of being independent, the subscales form typical patterns distributed fairly evenly in the population (the clusters were of almost equal size).

It is remarkable that such a simple temperament cluster is so strongly associated to a person's socioeconomic status and education, and to experiences of depression and anxiety, though the interpretation of this association into causal hypothesis is far from trivial. See [WSM$^+$12] for these results and further discussion.

4.2.6 Methodological implications

In this study, we had the opportunity to replicate the clustering in a separate sample. While not complicated, such a procedure is novel in practical applications. A potential replication sample of the same population with a

a)

	Cluster in YF Both datasets		Cluster in YF YF only		Cluster in YF NFBC1966 only	
Cluster in NFBC1966	F1	F2	F1	F2	F1	F2
I	514	336	189	124	325	212
II	739	17	245	11	494	6
III	208	447	23	67	185	380
IV	0	657	0	224	0	433

b)

	Cluster in YF Both datasets				Cluster in YF YF only				Cluster in YF NFBC1966 only			
Cluster in NFBC1966	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
I	111	17	625	97	48	13	391	70	63	13	391	70
II	150	565	41	0	86	392	22	0	86	392	22	0
III	543	10	27	75	457	10	24	74	457	10	24	74
IV	130	0	0	527	77	0	0	356	77	0	0	356

c)

	Cluster in YF Both datasets		Cluster in YF YF only		Cluster in YF NFBC1966 only	
Cluster in NFBC1966	F1	F2	F1	F2	F1	F2
F1	0	1262	0	361	0	901
F2	1461	195	457	65	1004	130

Table 4.1: [WSM+12] Clusterings based on NFBC1966 models vs YF models, females. a) NFBC1966 best model (four clusters) vs YF best model (two clusters). b) NFBC1966 model into four (best model) vs YF model into four (same number of clusters) (Cohen's κ 0.7, 0.67, and 0.71). c) NFBC1966 model into two (equal number of clusters) vs YF model into two (best model) (Cohen's κ 0.87, 0.85, and 0.87)

a)

	Cluster in YF Both datasets		Cluster in YF YF only		Cluster in YF NFBC1966 only	
Cluster in NFBC1966	M1	M2	M1	M2	M1	M2
I	168	495	64	150	104	345
II	0	631	0	247	0	384
III	421	566	208	266	213	300
IV	659	0	279	0	380	0

b)

	Cluster in YF Both datasets				Cluster in YF YF only				Cluster in YF NFBC1966 only			
Cluster in NFBC1966	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
I	0	5	1	657	0	3	0	211	0	2	1	446
II	0	556	29	46	0	212	14	21	0	344	15	25
III	52	66	607	262	33	36	279	126	19	30	328	136
IV	527	0	49	83	238	0	12	29	289	0	37	54

c)

	Cluster in YF Both datasets		Cluster in YF YF only		Cluster in YF NFBC1966 only	
Cluster in NFBC1966	M1	M2	M1	M2	M1	M2
M1	12	1591	6	620	6	971
M2	1236	101	545	43	691	58

Table 4.2: [WSM^{+}12] Clusterings based on NFBC1966 models vs YF models, males. a) NFBC1966 best model (four clusters) vs YF best model (two clusters). b) NFBC1966 model into four (best model) vs YF model into four (same number of clusters) (Cohen's κ 0.73, 0.70, 0.75). c) NFBC1966 model into two (equal number of clusters) vs YF model into two (best model) (Cohen's κ 0.92, 0.92, 0.92)

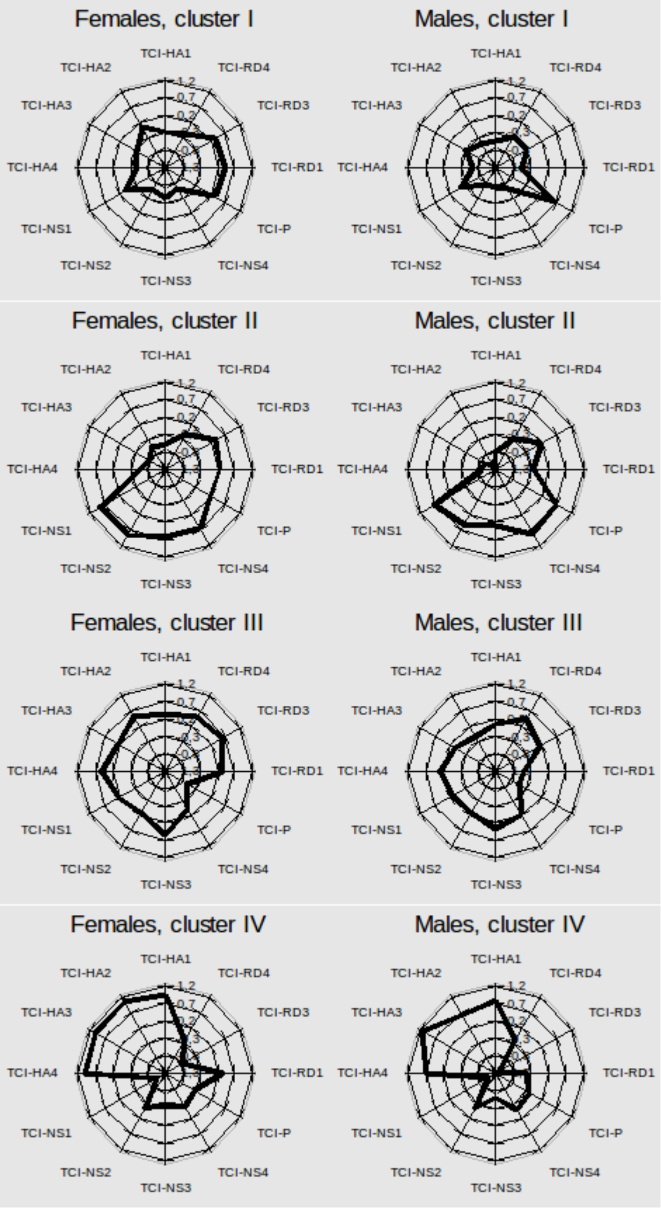

Figure 4.8: [WSM+12] Temperament cluster centers in the NFBC66 best model as starplots, normalized data. The male and female seem similar to each other, despite having been learned independently. The subscales are as follows: HA-1: anticipatory worry, HA-2: fear of uncertainty, HA-3: shyness, HA-4: fatigability; NS-1: exploratory excitability, NS-2: impulsiveness, NS-3: extravagance, NS-4: disorderliness; RD-1: sentimentality, RD-3: attachment, RD-4: dependence; P: persistence.

similar enough recruitment procedures is seldom available, but even when it is, replication of an exploratory data analysis procedure is rarely done.

As we have demonstrated here and in the experiments with artificial data (see Chapter 3.4) that it is a feasible and good method of establishing the validity of a clustering model, we hope that it will become part of standard procedures in clustering studies whenever replication data is available.

When a replication sample is not available, one might try simulating it by dividing the original data randomly into two parts. While this serves as a good check for the stability of the clustering solution, it fails in many respects to achieve the same goals as a replication sample. Namely, the important questions the replication answers is whether there is something in the recruitment or data collection procedure that causes the clustering structure. Rather than limit the sample size into half, the question whether the clusters are arbitrary is better answered by random dropping of variables / individuals, the existence of a natural hierarchy, and other methods of analysis of clustering stability.

4.3 Case 3: Migraine and the problems with missing and recoded data

4.3.1 Background

Migraine (from Old French *megrim*, Greek *hemikrania*, *hemi* “half” and *krania*, “skull”) is a syndrome of the central nervous system, the typical manifestation of which is recurrent attacks of pulsing, one-sided headache [GLF02]. The headache is often associated with other symptoms, such as nausea, sensitivity to light and sounds, or dizziness; sometimes more severe neurological symptoms also occur, such as numbness or even partial paralysis of the extremities. In aural migraine, the attacks are preceded by so-called aura symptoms, typically visual disturbances such as seeing bursts of color or zigzagging lines. In non-aural migraine, the aura symptoms are absent. In the USA and Western Europe, up to 11 % of people suffer from at least yearly migraine attacks [GLF02].

Migraine symptoms vary significantly between sufferers [Ant10]. Some individuals have one or two relatively mild attacks over years; some have severe, debilitating, hours-long attacks several times a week. As some sufferers have migraine auras without actual headache and the collection of related symptoms varies, it is possible that two individuals diagnosed with migraine do not share a single symptom.

Migraine is often triggered by specific experiences or situations, such as particular foods (red wine, chocolate, cheese), sleep deprivation, physical exercise, mental stress, normal hormonal changes (especially in women), etc. Very little is known about the exact causal mechanisms of triggers to the attack. The disorder is known to have a heritable component, but the exact genetics are unclear.

In this study, a large dataset has been analyzed as an attempt to form clusters of sufferers based on a diverse collection of aura and attack symptoms, triggers, course of disease, and other information of the individuals, such as co-morbidity. The dataset, however, is large and it has been collected over several years, and it has turned out that the clustering is very sensitive to differences in missing data probabilities and coding of variables that have been introduced during this time. For an overview of both the data used in this study and the disease, see the PhD thesis of Verneri Anttila [Ant10].

4.3.2 Data

The data in this study consists of a total of about 6500 individuals coming from families where at least three family members (among grandparents, parents, parent's siblings, siblings and children of an index case) suffer from migraine.

A group of neurologists under Mikko Kallela and Markus Färkkilä have been collecting a database of such families from headache clinics since 1992. All individuals, whether they themselves suffered of migraine or not, were asked to fulfill the validated Finnish Migraine Specific Questionnaire for Family Studies [KWF01] and to provide a blood sample. A neurologist also performed a physical examination of the index patient and sometimes other family members as well. Based on these, over 200 variables were recorded. The semantics of the variables varies widely; among other things migraine features, age of onset, and other diagnoses ("co-morbidity") have been recorded.

This dataset consists of 6,283 individuals, both healthy and with migraine. In our study, we limited ourselves to individuals with migraine and to variables with missing data proportion of under 50% and a clear annotation. This resulted in 135 variables and 2,500 to 3,500 individuals (see below). Unfortunately, the data collection process and the coding of various variables has not remained completely stable through-out the years, mainly due to the increasing understanding of migraine and thus shifting of interests.

4.3.3 Methods

We attempted to cluster the data using a mixture model of independent multinomial distributions (the Naïve Bayes model). The Bayesian information criterion was used to select the number of clusters from $k = 1, \ldots, 12$. Missing data was handled as a separate value for each variable. Random dropping of individuals and variables were performed in order to establish cluster stability.

Due to the missing data and recoding problems we performed the clustering in several subsets of individuals. Originally, we included all individuals with a migraine diagnosis and at least 50 percent of data recorded ($N = 2{,}661$) (dataset ALL in the following). In the course of preliminary clustering, the results showed that some of the clusters were related to data collection and recording procedures (rather than underlying disease-related phenomena). Thus we decided to exclude everyone with running ID over 4017, the point identified by a data expert as a significant change in those procedures. This resulted in a dataset with $N = 2500$, denoted below by OLD.

4.3.4 Results

The BIC scores over different number of clusters can be seen in Figure 4.9. The natural hierarchies test (see Chapter 3.3) also looks convincing for both models, as can be seen from Figure 4.10.

Based on the BIC score we selected six as the number of clusters. It is noteworthy that the natural hierarchies score peaks at four and five, and that the BIC score also flattens out before the rise from six. The natural hierarchies scores guarantee that this solution is highly similar to the solutions into four and five clusters, so it even if either of those is the “real” number of groups in the population, choosing six does not take us too far from the correct solution.

Cluster stability (Figure 4.11) was not particularly impressive, if not particularly terrible either. Most worrying was that the average Adjusted Mutual Information in 10 experiments does not reach over 0.9 even for re-clusterings with the full data. However, the number of clusters the BIC score suggested (Figure 4.12) remained at six in re-clusterings, only falling slowly when at least 20 percent of the data is removed.

Since the stability of the models was at question, and the amount of missing data was so great, we went on by looking carefully at the clusterings obtained. The aim was to understand how much missing data determines the cluster of each individual.

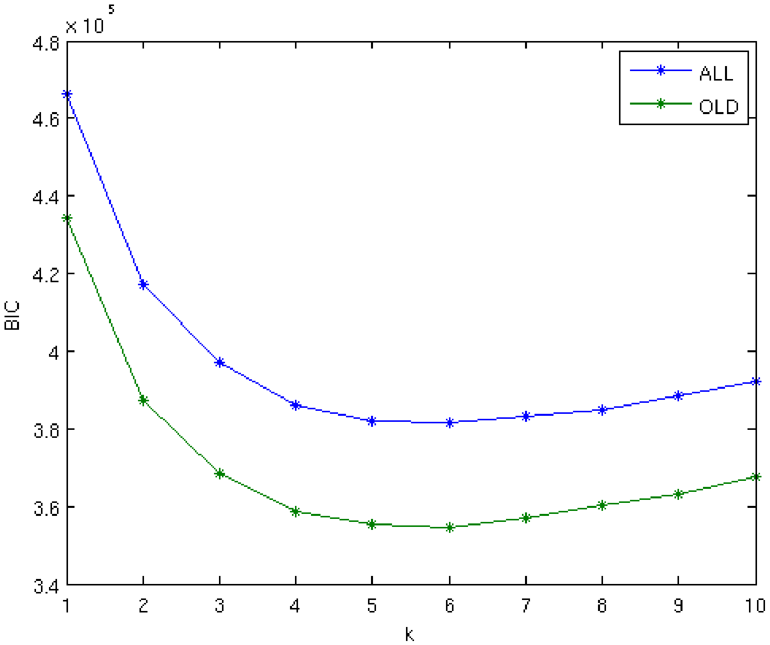

Figure 4.9: BIC scores for the two alternative datasets in the migraine data. X-axis, number of clusters. Y-axis, score. Due to difference in the number of individuals, these scores cannot be directly compared with each other. What is notable is the similar form of the curves: removing some individuals does not seem to have changed the optimal number of clusters.

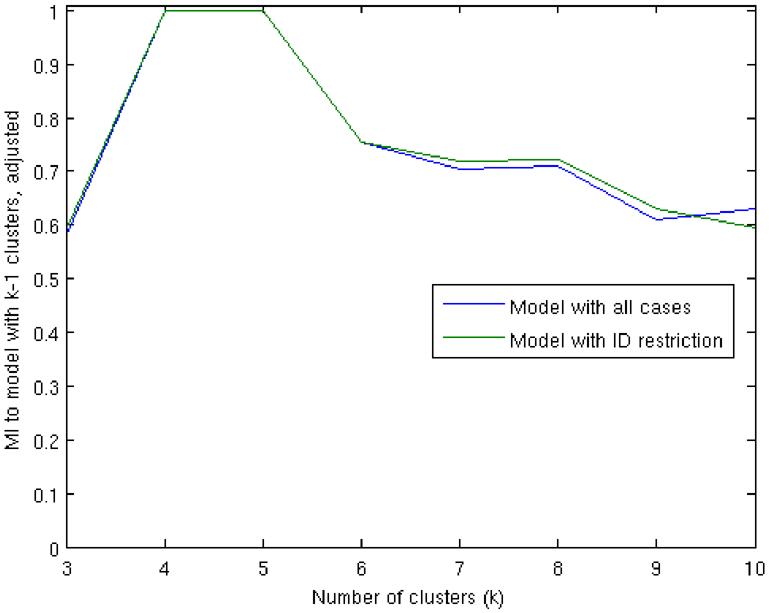

Figure 4.10: Adjusted mutual information between models with k and $k-1$ clusters for the two alternative migraine datasets. X-axis, number of clusters k. Y-axis, adjusted mutual information between the models of k and $k-1$ clusters. We can see that the models with 3-5 clusters are perfect refinements of each other, and the rest are quite similar too, suggesting a real cluster structure in the data. We can also see that there is not much difference between the behavior of the two datasets.

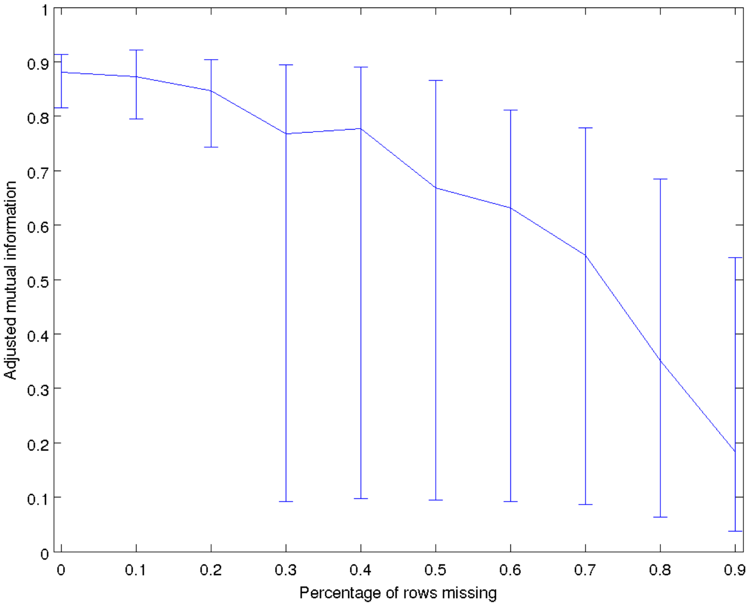

Figure 4.11: Average AMI to the original best model in 10 experiment series of re-clustering with increasingly missing data in the migraine data. X-axis, percentage of data removed at random. Y-axis, adjusted mutual information (1 = perfect similarity, 0 = similarity of random cluster assignments of the same size. Bars show minimum and maximum. The ALL dataset was used for this experiment. Note that the average AMI is below 0.9 even for re-clusterings of the original dataset.

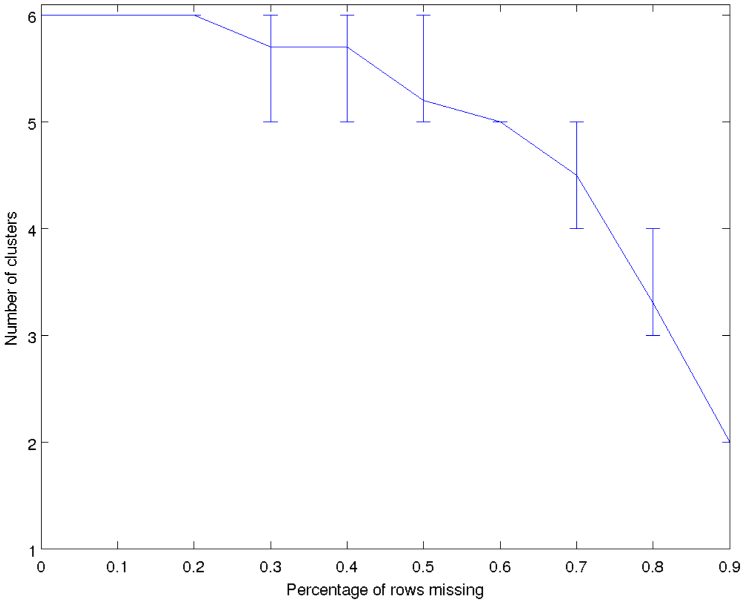

Figure 4.12: Average number of clusters in 10 experiment series of re-clustering with increasingly missing data in the migraine data. X-axis, percentage of data removed at random. Y-axis, average number of clusters for the 10 repetitions. Bars show minimum and maximum. Note that six clusters remains the suggested number even when a significant portion of data is removed.

a)

		Cluster, missing data as value					
		1	2	3	4	5	6
Cluster, missing data ignored	1	443	0	0	0	0	0
	2	12	401	0	0	0	0
	3	408	9	195	0	0	0
	4	40	0	0	323	0	28
	5	10	33	0	2	328	14
	6	174	0	0	0	0	241

b)

		Cluster, missing data as value					
		1	2	3	4	5	6
Cluster, missing data ignored	1	221	0	0	0	241	0
	2	2	302	0	34	1	0
	3	0	0	235	1	202	0
	4	0	1	0	383	14	0
	5	1	0	0	0	448	0
	6	0	1	32	1	37	343

Table 4.3: Comparison of clustering models when class is assigned based on only available data, or treating missing data as an additional value (similarly to when building the model). a) All cases with at least 50 percent of data (ALL), b) all cases with at least 50 percent of data and ID < 4018 (OLD).

Table 4.3 shows a comparison of two different ways to assign a cluster status to individuals in the two training sets. First, we can assign a cluster for each individual by treating missing data as an additional value for the cluster (as was done when building the models). Second, we can assign a cluster for each by simply ignoring the variables with missing data, by considering the probability of obtaining the values for the rest of the variables given the model.

We can see from this that there seems to be a cluster in both datasets (namely, cluster 1 for ALL and cluster 5 for OLD) that does not remain intact when the use of missing data information is forbidden, suggesting that this cluster label is based on missing data information only. In addition, there are two clusters (columns 2 and 6 in a), 3 and 4 in b)) that lose some members when missingness information is discarded. Three clusters remain intact in both models.

In Table 4.4 we see a comparison of the clusterings obtained by the two datasets, here using all the $N = 3{,}410$ cases of migraine in the data, when a cluster label is assigned to each case by ignoring missing data. This

		Cluster with the ALL model					
		1	2	3	4	5	6
Cluster with the OLD model	1	79	0	8	1	1027	3
	2	0	0	0	452	2	0
	3	255	0	2	0	3	0
	4	29	0	9	0	2	677
	5	0	328	8	12	0	7
	6	0	0	307	0	14	185

Table 4.4: Comparison of cluster labels of all migraine ($N = 3{,}410$) cases with the two alternative models. Columns, clusters obtained with the model learned on the ALL dataset. Rows, clusters obtained with the model learned on the OLD dataset.

similarity showed that the discrepancy in the majority cluster when ignoring and using missingness information is not simply due to data collection procedures being different after a certain running participant number.

4.3.5 Methodological implications

The most convincing conclusion from the above results is that the majority cluster obtained is probably based in a large part on patterns of missing data. This is in itself not a surprising, as data is missing in a non-random way, and we explicitly model missing data as an additional value for each variable. Enough missing and enough dependency on missingness patterns between variables implies that the individuals with certain patterns of missing values truly form their own cluster. It is not an indication that the method does not work, as such; rather, it is an indication that the missingness patterns in the data are informative about the individuals. Unfortunately, in this case the information very probably is not of medical interest, but is instead related to data collection procedures.

As the clusters other than the majority cluster seem to reflect clusterings based on actual variable values, it would be tempting to assume that the "true" cluster of the individuals clustered differently when missing data is ignored is the one obtained on observed data alone. The author feels, however, that when the models are so clearly based on the missingness status, and when the missingness is clearly not at random, this would not really be an easily justifiable assumption. Such an assumption would seem to require with confidence ruling out that the missingness pattern is not informative of the "true" cluster of the individuals, which assumption we cannot justify based on the data. Even if the missingness pattern is driven by data collection features, those in itself might be related to the underlying cluster structure. Erring to the side of caution, we ended up recommending this clustering to not be used for medical research purposes.

One possible option for future analysis would have been simply to drop the individuals that change cluster depending on the missing data treatment and concentrate on the individuals in the other, more stable clusters. Unfortunately this would cause the number of individuals with enough genetic data to be too low for the association studies that were the primary interest of the domain experts, and so analysis was not carried further. For some results about migraine on the same dataset, obtained by simpler methods, see the PhD thesis of Verneri Anttila [Ant10].

As a conclusion from this study, we can see that under missing data conditions, it is possible to obtain a clustering that looks convincing regarding to cluster number selection and stability analysis methods, but does not necessarily reflect an interesting underlying structure based on observed data. It is, hence, not enough to rely on the cluster number selection scores to convince oneself of a stable and interesting cluster structure. A cluster structure might exist, but not reflect any underlying phenomenon in the population under study, but rather data features such as coding or missingness.

Stability tests by random dropping can pick up such problems, though one should note that if the missingness mechanism is completely deterministic, this might not work either. Comparing the obtained clusterings by assigning cluster labels with and without utilizing missingness information can give some idea on how dependent the model is on the missingness status.

4.4 Case 4: No clear mixture model clusters in autism data

4.4.1 Background

Autism (from Greek *autos*-, “self”) is a disorder of the development of social interaction and communication, assumed to have a neurological basis.

The so-called “Autism Spectrum Disorders” include conditions of varying symptoms and severity, ranging from sufferers of mild Asperger’s Syndrome able to function as normal adults to deeply affected victims of Infantile autism completely unable to communicate. Autism Spectrum Disorders have a prevalence of 10 to 60 per 10,000, the most severe form of Infantile Autism, or Autistic Disorder as it is also called, comprising 4 to 10 % of that. Boys are at increased risk. [GW99, CF01, Cha02, AG08]

Autism manifests as disturbances of reciprocal social interactions, communication, behavior, and cognitive development. Repetitive actions and interests also occur. Individual symptoms seem to be present in the general population and there is no clear line separating the peculiar from the pathological; we speak of autism when the symptoms appear together and/or as more severe than is typical [Lon07].

Autism has been shown to have a clear genetically heritable basis, but details of genetics remain unclear; the existence of a single causative gene is highly unlikely [AG08]. Finding homogeneous subgroups of sufferers, symptom components with a heritable basis, or predisposing endophenotypes is again needed for further understanding of the heritability mechanics and the disorders in general.

In this study, we attempted a Naïve Bayes mixture model clustering of various measures of Infantile Autism. As this method proved unsuccessful, the research group went on to perform `PCA` and a subsequent k-means clustering of the results (in which work the author of this the present work was only marginally involved).

4.4.2 Data and results

Autism Diagnostic Interview-Revised (ADI-R) is a structured diagnostic interview of the parents or other primary caregivers of individuals with a suspected autism diagnosis. An associated algorithm for calculating scores based on the answers and defining a diagnosis based on the scores is also defined (though was not used in this study) [LRC94]. Most of the questions have 8 possible answers that can be considered ordered.

The data used in the clustering attempt described below consisted of 1395 individuals from the Autism Genetic Research Exchange collection [GSL$^+$01]. Of these, 1075 had received the diagnosis of autism. Number of variables was 210. We attempted a clustering with the Naïve Bayes mixture model procedure, with multinomial distributions, selecting the number of clusters with the `BIC` score. The original clustering attempt produced a clustering of two or three clusters (see Figure 4.13).

It was proved in the stability tests, however, that the three-cluster solution was not stable – repeated clusterings by removing even 10 percent of the data did not lead to a similar model. The two-cluster model was stable, but even with a cursory look of the results it proved obvious that the differentiating factor between the clusters was capability of speech, which, while an important divisive factor of autism sufferers, is not the kind of interesting new subgrouping one would hope from a clustering study. Due to

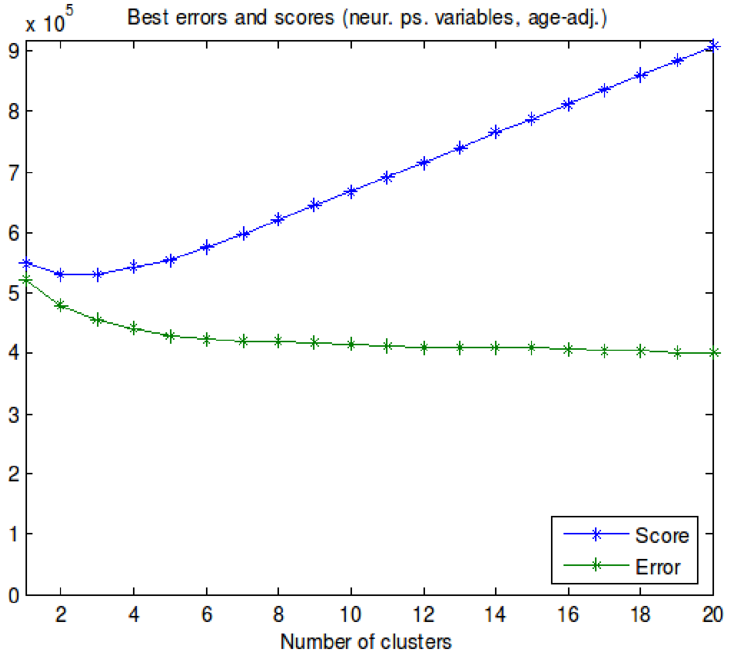

Figure 4.13: Model error and the `BIC` score in the autism study. X-axis, number of clusters. Y-axis, BIC score (blue, the top curve) and model error (green, the lower curve). We can see the score flattening out at 2-3 clusters and then rise steeply.

the instability of the results and the suspicion that the high dimensionality of the data compared to the number of individuals in the sample could be the cause of it, it was decided that this path of analysis be abandoned, and a `PCA`-based study performed instead.

4.4.3 Implications

The failure of mixture model clustering to produce meaningful results was likely due to high dimensionality of the data compared to number of individuals and/or the lack of cluster structure in the data. The author of this book was personally initially reluctant to apply `PCA`, due to not considering the data continuous by nature, but with advice from more experienced members of the team it was decided that the person primarily in charge of the analyses would proceed in that direction instead of e.g. dimensionality reduction by other methods and re-clustering. The author was proved wrong by subsequent results providing insight to Autism genetics (Roine et al., submitted).

Stability tests proved the clustering to be problematic in a relatively early phase of the study, and the line of research was abandoned, so the below should be taken as speculation. It can, however, be relatively easily intuitively understood that in cases where there might not be clear underlying group structure, clustering methods might not provide stable results.

In this case, the author's best guess for the reason behind the unstable structure is that the autism spectrum disorders — just as the name implies — do not constitute of separate subgroups. Rather, the sufferers form a continuum, from mild forms of the disease barely separate from extreme personalities to severely debilitating conditions, *and* everything in between.

Figure 3.3 and Section 3.3 explain in another context how clustering methods adjust to a gradient, resulting in non-hierarchical system when k is increased. Through a similar phenomenon, with a gradient structure in high dimensionality, it is likely that the randomly chosen initial clustering affects the results enough to cause unstable results. The fact that interesting results were obtained by `PCA` speaks for the theory of a gradient instead of cluster structure being present (though by no means confirms it with certainty).

Chapter 5

Conclusions

"A conclusion is just the place where you got tired of thinking."
(Nancy Kress)

In this thesis, we set out on a journey with a hammer, and proceeded to hit some things to find out if they behave like nails. Regardless of the abundance of such old tool-related jokes about it, this kind of basic applicative work is still relatively rare in computer science: there is a gap between algorithm development, where work usually stops when it has been experimentally proved that the method works on some datasets, and medical research, where methods are only rarely used unless they are a part of some relatively easy-to-use software package. This work has been a pebble thrown into that canyon that I hope many will follow.

In the practical real data studies included in this work I have shown that clustering methods can in some cases provide crucial insight into complex diseases. In the schizophrenia family study, our results shed light on the reasons for previous inconclusive results on the association of particular genes to disease. While conclusive proof of the role of these genes still eludes the scientific community, similar conclusions have been reached by other researches via independent methods, confirming our thinking as basically sound. In the the temperament study, we were successful in summarizing the 12-dimensional temperament scale into four groups without losing practically any information about background associations. We showed that males and females have similar basic temperament groups, and that these groups have associations to lifestyle, health, and position in the society.

We also found that not all nails are made equal for our hammer. Clustering methods are not always applicable, or easily applicable, and the reasons can be data-dependent (as was the case in the migraine study) or result from the actual phenomenon under study (as might have been the case in the autism work). The conclusion must be drawn that clustering is unlikely to be the “press a button and it goes” solution to this kind of data analysis, but instead great care and a lot of effort must go into curating and analyzing the data, and tailoring method selection and missing data handling, for each dataset separately. The cleaner and simpler the data, the easier it is to apply clustering methods, but unfortunately for those who love their hammer, it also holds that the cleaner and simpler the data is, the less need there is to go for complex analysis methods, instead of just looking at simple correlations.

The use of these clustering methods is not as such very complicated, as far as the required mathematical understanding and computer programming skills go. However, many parts of the procedure are data-dependent and data-driven, and interesting datasets tend to be complicated and noisy. For these reasons, we do not consider it likely that the methods would ever be easily usable without at least some programming skills. Even where data allows the use of commercialized or otherwise packaged programs, the interpretation of the results requires solid understanding of both the processes and the peculiarities of the data at hand.

Especially the evaluation of the stability and validity of the clusterings requires some care. As we have shown both in the simulations and in the studies with real data, patterns of missing data and coding decisions can affect the outcomes of the clustering algorithm, and detecting these artefacts requires careful analysis of the stability of the clustering solutions. We propose the use of random dropping of individuals and variables as one excellent means to detect such artifacts. Moreover, we suggest that replication in a separate sample should be held as the golden standard of validation also for clustering studies, even though we have to acknowledge the difficulties involved in replicating this kind of complex datasets.

In addition to these practical observations, various observations of the behavior of these clustering methods were reported and confirmed on artificial data. We compared 10-fold cross-validation and Bayesian information criterion in the selection of cluster number, and found them close to equal for realistic N in the presence of a cluster structure. The BIC score has the tendency to exaggerate the number of clusters in the absence of one, though, and the 10-fold cross-validation procedure to underestimate the number of clusters for small N (in the order of hundreds).

We also observe in simulations and in real data that in the presence of a true clustering structure in the data, non-hierarchical clustering methods tend to produce hierarchical clustering models for subsequent k, and that replication in a new sample can also confirm or deny the presence of a cluster structure. For missing data, we show that for the missing data handling procedures used in the real-life studies, data needs not to be missing at random, and that even datasets with fairly large numbers of missing data can still produce the clusterings obtained in full data. Finally, we show that also randomly dropping rows from data the data matrix and re-clustering is a good way to explore whether clusters are real.

To conclude, I consider clustering methods a viable alternative for this kind of medical data analysis, given that there is a research group that includes expertise both on the methods and on the domain. Good practical programming skills on the method side and clinical experiences from the disease under study on the medical side are a big bonus. It must be stressed though that these methods are only *an* alternative. No exploratory data analysis tool fits every data set, and sometimes exploration is not the best alternative: for example, if you have a clear hypothesis, you should test it, instead of explore in the hopes of landing on a proof.

References

[ADHP09] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75:245–249, 2009.

[AG08] Brett S Abrahams and Daniel H Geschwind. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet*, 9(5):341–355, May 2008.

[Ame00] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision)*. American Psychiatric Publishing, Washington, DC, 4th edition, July 2000.

[Ant10] Timo Verneri Anttila. *Identification of Genetic Susceptibility Loci For Migraine*. PhD thesis, University of Helsinki, Faculty of Medicine, 2010.

[AOY$^+$02] Juko Ando, Yutaka Ono, Kimio Yoshimura, Naoko Onoda, Manabu Shinohara, Shigenobu Kanba, and Masahiro Asai. The genetic structure of Cloninger's seven-factor model of temperament and character in a Japanese sample. *J Pers*, 70(5):583–609, Oct 2002.

[BHEG02] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

[BPBC96] M. Battaglia, T. R. Przybeck, L. Bellodi, and C. R. Cloninger. Temperament dimensions explain the comorbidity of psychiatric disorders. *Compr Psychiatry*, 37(4):292–298, 1996.

[BQMR97] R. Bakeman, V. Quera, D. McArthur, and B.F. Robinson. Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2:357–370, 1997.

[BSH+07] Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh, and Edward R. Dougherty. Model-based evaluation of clustering validation measures. *Patter Recognition*, 40(3):807–824, March 2007.

[CBS+08] J. E. Chubb, N. J. Bradshaw, D. C. Soares, D. J. Porteous, and J. K. Millar. The DISC locus in psychiatric illness. *Mol Psychiatry*, 13(1):36–64, Jan 2008.

[CDN+08] Aiden Corvin, Gary Donohoe, Jeanne Marie Nangle, Siobhan Schwaiger, Derek Morris, and Michael Gill. A dysbindin risk haplotype associated with less severe manic-type symptoms in psychosis. *Neurosci Lett*, 431(2):146–149, Jan 2008.

[CF01] S. Chakrabarti and E. Fombonne. Pervasive developmental disorders in preschool children. *JAMA*, 285(24):3093–3099, Jun 2001.

[CG92] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14:315 – 332, 1992.

[Cha02] Tony Charman. The prevalence of autism spectrum disorders. Recent evidence and future challenges. *Eur Child Adolesc Psychiatry*, 11(6):249–256, Dec 2002.

[CKL+98] T. D. Cannon, J. Kaprio, J. Lönnqvist, M. Huttunen, and M. Koskenvuo. The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. *Arch Gen Psychiatry*, 55(1):67–74, Jan 1998.

[Clo87] C. R. Cloninger. A systematic method for clinical description and classification of personality variants. A proposal. *Arch Gen Psychiatry*, 44(6):573–588, Jun 1987.

[Coh60] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

[COO06] Nick Craddock, Michael C O'Donovan, and Michael J. Owen. Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. *Schizophr Bull*, 32(1):9–16, Jan 2006.

[CPvW94] C.R. Cloninger, T.R. Przybeck, D.M. Švrakić, and R.D. Wetzel. *The Temperament and Character Inventory (TCI): A Guide* *to Its Development and Use.* Washington University, Center for Psychobiology of Personality, St. Louis, 1994.

[CSWM94] D. J. Castle, P. C. Sham, S. Wessely, and R. M. Murray. The subtyping of schizophrenia in men and women: a latent class analysis. *Psychol Med*, 24(1):41–51, Feb 1994.

[CT91] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley, 1991.

[CvP93] C. R. Cloninger, D. M. Švrakić, and T. R. Przybeck. A psychobiological model of temperament and character. *Arch Gen Psychiatry*, 50(12):975–990, Dec 1993.

[DLR77] A. P. Dempster, N. M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).*, 1:1–38, 1977.

[DP97] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103 – 130, 1997.

[EBS+04] Christer Engström, Sven Brändström, Sören Sigvardsson, Robert Cloninger, and Per-Olof Nylander. Bipolar disorder: I. Temperament and character. *J Affect Disord*, 82(1):131–134, Oct 2004.

[Ems87] H. E. Emson. Health, disease and illness: matters for definition. *CMAJ*, 136(8):811–813, Apr 1987.

[FADA+02] Secondo Fassino, Giovanni Abbate-Daga, Federico Amianto, Paolo Leombruni, Sonia Boggio, and Gian Giacomo Rovera. Temperament and character profile of eating disorders: a controlled study with the Temperament and Character Inventory. *Int J Eat Disord*, 32(4):412–425, Dec 2002.

[FFP+04] Birgit Funke, Christine T. Finn, Alex M. Plocik, Stephen Lake, Pamela DeRosse, John M. Kane, Raju Kucherlapati, and Anil K. Malhotra. Association of the DTNBP1 locus with schizophrenia in a U.S. population. *Am J Hum Genet*, 75(5):891–898, Nov 2004.

[FK08] Ayman H. Fanous and Kenneth S. Kendler. Genetics of clinical features and subtypes of schizophrenia: a review of the recent literature. *Curr Psychiatry Rep*, 10(2):164–170, Apr 2008.

[GB03] Simon Günter and Horst Bunke. Validation indices for graph clustering. *Patter Recognition Letters*, 24(8):1107–1113, May 2003.

[GCHM03] Nathan A. Gillespie, C. Robert Cloninger, Andrew C. Heath, and Nicholas G. Martin. The genetic and environmental relationship between Cloninger's dimensions of temperament and character. *Personality and Individual Differences*, 35(8):1931 – 1946, 2003.

[GG03] Irving I. Gottesman and Todd D. Goul. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry*, 160(4):636–645, Apr 2003.

[GJ94] Z. Ghahramani and M. I. Jordan. Learning from incomplete data. Technical report, MIT Center for Biological and Computational Learning, 1994.

[GLF02] Peter J. Goadsby, Richard B. Lipton, and Michel D. Ferrari. Migraine — current understanding and treatment. *New England Journal of Medicine*, 346:257–270, 2002.

[GSL$^+$01] D. H. Geschwind, J. Sowinski, C. Lord, P. Iversen, J. Shestack, P. Jones, L. Ducat, S. J. Spence, and A. G. R. E. Steering Committee. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am J Hum Genet*, 69(2):463–466, Aug 2001.

[GW99] C. Gillberg and L. Wing. Autism: not an extremely rare disorder. *Acta Psychiatr Scand*, 99(6):399–406, Jun 1999.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193 – 218, 1985.

[HbJ$^+$09] Mirka Hintsanen, Laura Pulkki-Råback, Markus Juonala, Jorma S. A. Viikari, Olli T Raitakari, and Liisa Keltikangas-Järvinen. Cloninger's temperament traits and preclinical atherosclerosis: the Cardiovascular Risk in Young Finns study. *J Psychosom Res*, 67(1):77–84, Jul 2009.

[HBV01] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.

[HCM94] A. C. Heath, C. R. Cloninger, and N. G. Martin. Testing a model for the genetic structure of personality: a comparison of the personality systems of Cloninger and Eysenck. *J Pers Soc Psychol*, 66(4):762–775, Apr 1994.

[HJ03] Lynette Hunt and Murray Jorgensen. Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis*, 41(3–4):429–440, January 2003.

[HK06] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 2006.

[HKB$^+$05] Joachim F Hallmayer, Luba Kalaydjieva, Johanna Badcock, Milan Dragovic, Sarah Howell, Patricia T Michie, Daniel Rock, David Vile, Rachael Williams, Elizabeth H Corder, Kate Hollingsworth, and Assen Jablensky. Genetic evidence for a distinct subtype of schizophrenia characterized by pervasive cognitive deficit. *Am J Hum Genet*, 77(3):468–476, Sep 2005.

[HL01] T. Hill and P Lewicki. Electronic statistics textbook, 2001.

[HMS01] D. J. Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining.* The MIT Press, 2001.

[Hun05] David J. Hunter. Gene-environment interactions in human diseases. *Nature Reviews Genetics*, 6:287–298, April 2005.

[HVS$^+$99] I. Hovatta, T. Varilo, J. Suvisaari, J. D. Terwilliger, V. Ollikainen, R. Arajärvi, H. Juvonen, M. L. Kokko-Sahin, L. Väisänen, H. Mannila, J. Lönnqvist, and L. Peltonen. A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am J Hum Genet*, 65(4):1114–1124, Oct 1999.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[Kii08] Harri T. Kiiveri. A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. *BMC Bioinformatics*, 9:195, 2008.

[KMG$^+$93] K. S. Kendler, M. McGuire, A. M. Gruenberg, A. O'Hare, M. Spellman, and D. Walsh. The Roscommon family study. I. Methods, diagnosis of probands, and risk of schizophrenia in relatives. *Arch Gen Psychiatry*, 50(7):527–540, Jul 1993.

[KNW95] K. S. Kendler, M. C. Neale, and D. Walsh. Evaluating the spectrum concept of schizophrenia in the Roscommon family study. *Am J Psychiatry*, 152(5):749–754, May 1995.

[KS96] David J Ketchen and Christopher L Shook. The appplication of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17:441–458, 1996.

[KSB$^+$04] Kelly L Klump, Michael Strober, Cynthia M Bulik, Laura Thornton, Craig Johnson, Bernie Devlin, Manfred M Fichter, Katherine A Halmi, Allan S Kaplan, D. Blake Woodside, Scott Crow, James Mitchell, Alessandro Rotondo, Pamela K Keel, Wade H Berrettini, Katherine Plotnicov, Christine Pollice, Lisa R Lilenfeld, and Walter H Kaye. Personality characteristics of women before and after recovery from an eating disorder. *Psychol Med*, 34(8):1407–1418, Nov 2004.

[KWF01] M. Kallela, M. Wessman, and M. Färkkilä. Validation of a migraine-specific questionnaire for use in family studies. *Eur J Neurol*, 8(1):61–66, Jan 2001.

[LK77] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[Llo82] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[LMR07] Mark F. Lenzenweger, Geoff McLachlan, and Donald B. Rubin. Resolving the latent structure of schizophrenia endophenotypes using expectation-maximization-based finite mixture modeling. *J Abnorm Psychol*, 116(1):16–29, Feb 2007.

[Lon07] Eric London. The role of the neurobiologist in redefining the diagnosis of autism. *Brain Pathol*, 17(4):408–411, Oct 2007.

[LR02] J. A. Little and Donald B. Rubin. *Statistical analysis of missing data.* Wiley, 2002.

[LRBB04] Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Comput*, 16(6):1299–1323, Jun 2004.

[LRC94] C. Lord, M. Rutter, and A. Le Couteur. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord*, 24(5):659–685, Oct 1994.

[LSS05] Kenneth Lange, Janet S Sinsheimer, and Eric Sobel. Association testing with Mendel. *Genet Epidemiol*, 29(1):36–50, Jul 2005.

[Luc01] C. Lucas. *Computing nearest covariance and correlation matrices. MS Thesis.* University of Manchester, 2001.

[Mac67] J. B MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Mei05] Marina Meilă. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd international conference on Machine learning*, volume 119 of *ACM International Conference Proceeding Series*, pages 577 – 584. ACM, 2005.

[MH01] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1-2):9 – 29, 2001.

[MKE$^+$04] J. Miettunen, L. Kantojärvi, J. Ekelund, J. Veijola, J. T. Karvonen, L. Peltonen, M. R. Järvelin, N. Freimer, D. Lichtermann, and M. Joukamaa. A large population cohort provides normative data for investigation of temperament. *Acta Psychiatr Scand*, 110(2):150–157, Aug 2004.

[MLK$^+$08] Jouko Miettunen, Erika Lauronen, Liisa KantojÃ¤rvi, Juha Veijola, and Matti Joukamaa. Inter-correlations between Cloninger's temperament dimensions– a meta-analysis. *Psychiatry Res*, 160(1):106–114, Jul 2008.

[MMM$^+$05] V. Murray, I. McKee, P. M. Miller, D. Young, W. J. Muir, A. J. Pelosi, and D. H. R. Blackwood. Dimensions and classes of psychosis in a population cohort: a four-class, four-dimension model of schizophrenia and affective psychoses. *Psychol Med*, 35(4):499–510, Apr 2005.

[MNL$^+$04] John A McGrath, Gerald Nestadt, Kung-Yee Liang, Virginia K. Lasseter, Paula S. Wolyniec, M. Danielle Fallin, Mary H. Thornquist, James R. Luke, and Ann E. Pulver. Five latent factors underlying schizophrenia: analysis and relationship to illnesses in relatives. *Schizophr Bull*, 30(4):855–873, 2004.

[MP00] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, Inc., 2000.

[MVL$^+$07] Jouko Miettunen, Juha Veijola, Erika Lauronen, Liisa Kantojärvi, and Matti Joukamaa. Sex differences in Cloninger's temperament dimensions–a meta-analysis. *Compr Psychiatry*, 48(2):161–169, 2007.

[Nab01] Ian Nabney. *Netlab: Algorithms for Pattern Recognition*. Advances in Pattern Recognition. Springer, 2001.

[PTHH$^+$04] Tiina Paunio, Annamari Tuulio-Henriksson, Tero Hiekkalinna, Markus Perola, Teppo Varilo, Timo Partonen, Tyrone D. Cannon, Jouko Lönnqvist, and Leena Peltonen. Search for cognitive trait components of schizophrenia reveals a locus for verbal learning and memory on 4q and for visual working memory on 2q. *Hum Mol Genet*, 13(16):1693–1702, Aug 2004.

[Ran69] P. Rantakallio. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr Scand*, 193:Suppl 193:1+, 1969.

[Ran71] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistican Association*, 66(336):846 – 850, 1971.

[RJR$^+$08] Olli T Raitakari, Markus Juonala, Tapani Rönnemaa, Liisa Keltikangas-Järvinen, Leena Räsänen, Matti Pietikäinen, Nina Hutri-Kähönen, Leena Taittonen, Eero Jokinen, Jukka Marniemi, Antti Jula, Risto Telama, Mika Kähönen, Terho Lehtimäki, Hans K Akerblom, and Jorma S. A. Viikari. Cohort profile: the cardiovascular risk in Young Finns study. *Int J Epidemiol*, 37(6):1220–1226, Dec 2008.

[RLBB02] Volker Roth, Tilman Lange, Mikio Braun, and Joachim Buhmann. A resampling approach to cluster validation. In *In Intl. Conf. on Computational Statistics*, pages 123–128, 2002.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[RS04] Michael Ritsner and Ehud Susser. Temperament types are associated with weak self-construct, elevated distress and emotion-oriented coping in schizophrenia: evidence for a complex vulnerability marker? *Psychiatry Res*, 128(3):219–228, Oct 2004.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[SDL$^+$08] Alan R. Sanders, Jubao Duan, Douglas F. Levinson, Jianxin Shi, Deli He, Cuiping Hou, Gregory J. Burrell, John P. Rice, Deborah A. Nertney, Ann Olincy, Pablo Rozic, Sophia Vinogradov, Nancy G. Buccola, Bryan J. Mowry, Robert Freedman, Farooq Amin, Donald W. Black, Jeremy M. Silverman, William F. Byerley, Raymond R. Crowe, C. Robert Cloninger, Maria Martinez, and Pablo V. Gejman. No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. *Am J Psychiatry*, 165(4):497–506, Apr 2008.

[SG03] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583 – 617, 2003.

[SHC$^+$96] M. C. Stallings, J. K. Hewitt, C. R. Cloninger, A. C. Heath, and L. J. Eaves. Genetic and environmental structure of the tridimensional personality questionnaire: three or four temperament dimensions? *J Pers Soc Psychol*, 70(1):127–140, Jan 1996.

[SKM$^+$07] Ulla Sovio, Vanessa King, Jouko Miettunen, Ellen Ek, Jaana Laitinen, Matti Joukamaa, Juha Veijola, and Marjo-Riitta Järvelin. Cloninger's temperament dimensions, socio-economic and lifestyle factors and metabolic syndrome markers at age 31 years in the Northern Finland birth cohort 1966. *J Health Psychol*, 12(2):371–382, Mar 2007.

[SM97] Robert L. Souhami and John Moxham, editors. *Textbook of Medicine*. Churchill Livingstone, 1997.

[SM06] Timothy W. Smith and Justin MacKenzie. Personality and risk of physical illness. *Annu Rev Clin Psychol*, 2:435–467, 2006.

[Smy96] P. Smyth. Clustering using Monte Carlo cross-validation. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-2)*, 1996.

[SSQG06] Fiona M. Shrive, Heather Stuart, Hude Quan, and William A. Ghali. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Med Res Methodol*, 6:57, 2006.

[TCS^{+}01] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, Jun 2001.

[Tho53] Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[TSF00] M. T. Tsuang, W. S. Stone, and S. V. Faraone. Toward reformulating the diagnosis of schizophrenia. *Am J Psychiatry*, 157(7):1041–1050, Jul 2000.

[TSK06] Pan-Ning Tan, Michael Steinach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[Tuk80] John W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, February 1980.

[TWM^{+}94] P. Tienari, L. C. Wynne, J. Moring, I. Lahti, M. Naarala, A. Sorri, K. E. Wahlberg, O. Saarento, M. Seitamaa, and M. Kaleva. The Finnish adoptive family study of schizophrenia. implications for family research. *Br J Psychiatry Suppl*, 23(23):20–26, Apr 1994.

[VCS^{+}08] Elisabet Vilella, Javier Costas, Julio Sanjuan, Miriam Guitart, Yolanda De Diego, Angel Carracedo, Lourdes Martorell, Joaquin Valero, Antonio Labad, Rosa De Frutos, Carmen Nájera, M. Dolores Moltó, Ivette Toirac, Roser Guillamat, Anna Brunet, Vicenç Vallès, Lucía Pérez, Melquìades Leon, Fernando Rodríguez de Fonseca, Christopher Phillips, and María Torres. Association of schizophrenia with DTNBP1 but not with DAO, DAOA, NRG1 and RGS4 nor their genetic interaction. *J Psychiatr Res*, 42(4):278–288, Mar 2008.

[VEB09] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, volume 382 of *ACM International Conference Proceeding Series*, pages 1073 – 1080, 2009.

[VLH$^+$00] T. Varilo, M. Laan, I. Hovatta, V. Wiebe, J. D. Terwilliger, and L. Peltonen. Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet*, 8(8):604–612, Aug 2000.

[Wag01] Kiri Wagstaff. Clustering with missing values: No imputation required. In *Proceedings of the Meeting of the International Federation of Classification Societies*, 2001.

[WPTH$^+$09] Jaana Wessman, Tiina Paunio, Annamari Tuulio-Henriksson, Mikko Koivisto, Timo Partonen, Jaana Suvisaari, Joni A. Turunen, Juho Wedenoja, William Hennah, Olli P. H. Pietiläinen, Jouko Lönnqvist, Heikki Mannila, and Leena Peltonen. Mixture model clustering of phenotype features reveals evidence for association of DTNBP1 to a specific subtype of schizophrenia. *Biol Psychiatry*, 66(11):990–996, Dec 2009.

[WSM$^+$12] Jaana Wessman, Stefan Schönauer, Jouko Miettunen, Hannu Turunen, Pekka Parviainen, Jouni K. Seppänen, Eliza Congdon, Susan Service, Markku Koiranen, Jesper Ekelund, Jaana Laitinen, Anja Taanila, Tuija Tammelin, Mirka Hintsanen, Laura Pulkki-Råback, Liisa Keltikangas-Järvinen, Jorma Viikari, Olli T. Raitakari, Matti Joukamaa, Marjo-Riitta Järvelin, Nelson Freimer, Leena Peltonen, Juha Veijola, Heikki Mannila, and Tiina Paunio. Temperament clusters in a normal population: Implications for health and disease. *PLoS ONE*, in press, 2012.

[ZB08] Marchel Zentner and John E. Bates. Child temperament: An integrative review of concepts, research programs and measures. *European Journal of Developmental Science*, 2:7 – 37, 2008.

TIETOJENKÄSITTELYTIETEEN LAITOS
PL 68 (Gustaf Hällströmin katu 2 b)
00014 Helsingin yliopisto

JULKAISUSARJA **A**

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 68 (Gustaf Hällströmin katu 2 b)
FIN-00014 University of Helsinki, FINLAND

SERIES OF PUBLICATIONS **A**

Reports may be ordered from: Kumpula Science Library, P.O. Box 64, FIN-00014 University of Helsinki, FINLAND.

A-2004-5 J. Viljamaa: Applying Formal Concept Analysis to Extract Framework Reuse Interface Specifications from Source Code. 206 pp. (Ph.D. Thesis)

A-2004-6 J. Ravantti: Computational Methods for Reconstructing Macromolecular Complexes from Cryo-Electron Microscopy Images. 100 pp. (Ph.D. Thesis)

A-2004-7 M. Kääriäinen: Learning Small Trees and Graphs that Generalize. 45+49 pp. (Ph.D. Thesis)

A-2004-8 T. Kivioja: Computational Tools for a Novel Transcriptional Profiling Method. 98 pp. (Ph.D. Thesis)

A-2004-9 H. Tamm: On Minimality and Size Reduction of One-Tape and Multitape Finite Automata. 80 pp. (Ph.D. Thesis)

A-2005-1 T. Mielikäinen: Summarization Techniques for Pattern Collections in Data Mining. 201 pp. (Ph.D. Thesis)

A-2005-2 A. Doucet: Advanced Document Description, a Sequential Approach. 161 pp. (Ph.D. Thesis)

A-2006-1 A. Viljamaa: Specifying Reuse Interfaces for Task-Oriented Framework Specialization. 285 pp. (Ph.D. Thesis)

A-2006-2 S. Tarkoma: Efficient Content-based Routing, Mobility-aware Topologies, and Temporal Subspace Matching. 198 pp. (Ph.D. Thesis)

A-2006-3 M. Lehtonen: Indexing Heterogeneous XML for Full-Text Search. 185+3 pp. (Ph.D. Thesis)

A-2006-4 A. Rantanen: Algorithms for ^{13}C Metabolic Flux Analysis. 92+73 pp. (Ph.D. Thesis)

A-2006-5 E. Terzi: Problems and Algorithms for Sequence Segmentations. 141 pp. (Ph.D. Thesis)

A-2007-1 P. Sarolahti: TCP Performance in Heterogeneous Wireless Networks. (Ph.D. Thesis)

A-2007-2 M. Raento: Exploring privacy for ubiquitous computing: Tools, methods and experiments. (Ph.D. Thesis)

A-2007-3 L. Aunimo: Methods for Answer Extraction in Textual Question Answering. 127+18 pp. (Ph.D. Thesis)

A-2007-4 T. Roos: Statistical and Information-Theoretic Methods for Data Analysis. 82+75 pp. (Ph.D. Thesis)

A-2007-5 S. Leggio: A Decentralized Session Management Framework for Heterogeneous Ad-Hoc and Fixed Networks. 230 pp. (Ph.D. Thesis)

A-2007-6 O. Riva: Middleware for Mobile Sensing Applications in Urban Environments. 195 pp. (Ph.D. Thesis)

A-2007-7 K. Palin: Computational Methods for Locating and Analyzing Conserved Gene Regulatory DNA Elements. 130 pp. (Ph.D. Thesis)

A-2008-1 I. Autio: Modeling Efficient Classification as a Process of Confidence Assessment and Delegation. 212 pp. (Ph.D. Thesis)

A-2008-2 J. Kangasharju: XML Messaging for Mobile Devices. 24+255 pp. (Ph.D. Thesis).

A-2008-3 N. Haiminen: Mining Sequential Data – in Search of Segmental Structures. 60+78 pp. (Ph.D. Thesis)

A-2008-4 J. Korhonen: IP Mobility in Wireless Operator Networks. (Ph.D. Thesis)

A-2008-5 J.T. Lindgren: Learning nonlinear visual processing from natural images. 100+64 pp. (Ph.D. Thesis)

A-2009-1 K. Hätönen: Data mining for telecommunications network log analysis. 153 pp. (Ph.D. Thesis)

A-2009-2 T. Silander: The Most Probable Bayesian Network and Beyond. (Ph.D. Thesis)

A-2009-3 K. Laasonen: Mining Cell Transition Data. 148 pp. (Ph.D. Thesis)

A-2009-4 P. Miettinen: Matrix Decomposition Methods for Data Mining: Computational Complexity and Algorithms. 164+6 pp. (Ph.D. Thesis)

A-2009-5 J. Suomela: Optimisation Problems in Wireless Sensor Networks: Local Algorithms and Local Graphs. 106+96 pp. (Ph.D. Thesis)

A-2009-6 U. Köster: A Probabilistic Approach to the Primary Visual Cortex. 168 pp. (Ph.D. Thesis)

A-2009-7 P. Nurmi: Identifying Meaningful Places. 83 pp. (Ph.D. Thesis)

A-2009-8 J. Makkonen: Semantic Classes in Topic Detection and Tracking. 155 pp. (Ph.D. Thesis)

A-2009-9 P. Rastas: Computational Techniques for Haplotype Inference and for Local Alignment Significance. 64+50 pp. (Ph.D. Thesis)

A-2009-10 T. Mononen: Computing the Stochastic Complexity of Simple Probabilistic Graphical Models. 60+46 pp. (Ph.D. Thesis)

A-2009-11 P. Kontkanen: Computationally Effcient Methods for MDL-Optimal Density Estimation and Data Clustering. 75+64 pp. (Ph.D. Thesis)

A-2010-1 M. Lukk: Construction of a global map of human gene expression - the process, tools and analysis. 120 pp. (Ph.D. Thesis)

A-2010-2 W. Hämäläinen: Efficient search for statistically significant dependency rules in binary data. 163 pp. (Ph.D. Thesis)

A-2010-3 J. Kollin: Computational Methods for Detecting Large-Scale Chromosome Rearrangements in SNP Data. 197 pp. (Ph.D. Thesis)

A-2010-4 E. Pitkänen: Computational Methods for Reconstruction and Analysis of Genome-Scale Metabolic Networks. 115+88 pp. (Ph.D. Thesis)

A-2010-5 A. Lukyanenko: Multi-User Resource-Sharing Problem for the Internet. 168 pp. (Ph.D. Thesis)

A-2010-6 L. Daniel: Cross-layer Assisted TCP Algorithms for Vertical Handoff. 84+72 pp. (Ph.D. Thesis)

A-2011-1 A. Tripathi: Data Fusion and Matching by Maximizing Statistical Dependencies. 89+109 pp. (Ph.D. Thesis)

A-2011-2 E. Junttila: Patterns in Permuted Binary Matrices. 155 pp. (Ph.D. Thesis)

A-2011-3 P. Hintsanen: Simulation and Graph Mining Tools for Improving Gene Mapping Efficiency. 136 pp. (Ph.D. Thesis)

A-2011-4 M. Ikonen: Lean Thinking in Software Development: Impacts of Kanban on Projects. 104+90 pp. (Ph.D. Thesis)

A-2012-1 P. Parviainen: Algorithms for Exact Structure Discovery in Bayesian Networks. 132 pp. (Ph.D. Thesis)