Enhanced Interpretation of the Mini-Mental State Examination

Diman Todorov

Cardiff University

School of Engineering

March 2013

This dissertation is submitted for

the degree of Doctor of Philosophy

Acknowledgements

I would like to express my gratitude to Dr. Setchi and Dr. Bayer, my research supervisors, for their guidance and critiques of this thesis. I would also like to thank Ms. Copeland for her practical assistance in collecting the MMSE data. My thanks are extended to Dr. Hicks, Dr. Stankov, Dr. Shi and Mr. Bennasar for stimulating discussions and technical input.

Finally, I wish to thank my parents, Bistra and Valentin Todorovi for their support and encouragement throughout my study.

Abstract

The goal of the research reported in this thesis is to contribute to early and accurate detection of dementia. Early detection of dementia is essential to maximising the effectiveness of treatment against memory loss. This goal is pursued by interpreting the Mini-Mental State Examination (MMSE) in novel ways. The MMSE is the most widely used screening tool for dementia, it is a questionnaire of 30 items. The objectives of the research are as follows:

- to *reduce the dimensions* of the MMSE to the most relevant ones in order to inform a predictive model by using *computational methods* on a data set of MMSE results,
- to *construct a model* predicting a diagnosis informed by the features extracted from the previous step by applying, comparing and combining traditional and novel modelling methods,
- to propose a *semantic analysis* of the *sentence writing* question in the MMSE in order to utilise information recorded in MMS examinations which has not been considered previously.

Traditional methods of analysis are inadequate for questionnaire data such as the MMSE due to assumptions of normally distributed data. Alternative methods for analysis of discrete data are investigated and a novel method for computing information theoretic measures is proposed. The methods are used to demonstrate that an automated analysis of the MMSE sentence improves the accuracy of differentiating between types of dementia. Finally, models are proposed which integrate the semantic annotations with the MMSE data to derive rules for difficult to distinguish types of dementia.

Contents

Appendices 127

List of Figures

List of Tables

Nomenclature

X, Y, Z	multi-dimensional random variables
X_i	ith dimension of X
$X_{i=u}$	X_i taking the value u
X_S	selected dimensions of X
X_{-S}	not selected dimensions of X
$H(X)$	entropy of X
$I(X;Y)$	mutual information between X and Y
$I(X;Y;Z)$	three-way interactions between X,Y, and Z

Chapter 1

Introduction

1.1 Motivation

The goal of the research reported in this thesis is to contribute to early and accurate detection of dementia. Early detection of dementia is essential to maximising the effectiveness of treatment against memory loss.

The probability of developing dementia symptoms becomes significant after the age of 65 (Ott et al. 1995). Because of the dramatically increased life expectancy in developed countries (see Figure 1.1), addressing issues related to dementia is essential in sustaining high quality of life. The 2005 report to the Alzheimer's Society on dementia (Knapp and Prince 2005) estimates that there were over 680,000 (1.1% of the population in the UK) cases of dementia in 2005. It also reports projected increase of 38% until 2021 and an increase of 154% until 2051. The Alzheimer's Society report on dementia for 2012 estimates that there are 800,000 people with dementia in the Britain with another 670,000 family and friends acting as primary carers (Lakey et al. 2012).

Although dementia is very common – 9.4% percent of those aged over 65 suffer from dementia (Ott et al. 1995) – only about half of those affected are diagnosed (Nicholl 2009). For this reason early detection of dementia is one of the key points in the National Dementia Strategy of England (Palethorpe 2009) and the *National Dementia Plan for Wales* (2009).

Dementia is a term describing a set of symptoms which includes loss of cognitive faculties such as memory, communication and abstract thinking or reasoning. The most widely used tool to screen for dementia is the Mini-Mental State Examination (MMSE) (Ismail, Rajji and Shulman 2010). The MMSE is a questionnaire with 30 questions which assess orientation to time and place; productive and receptive language faculties; attention and recall

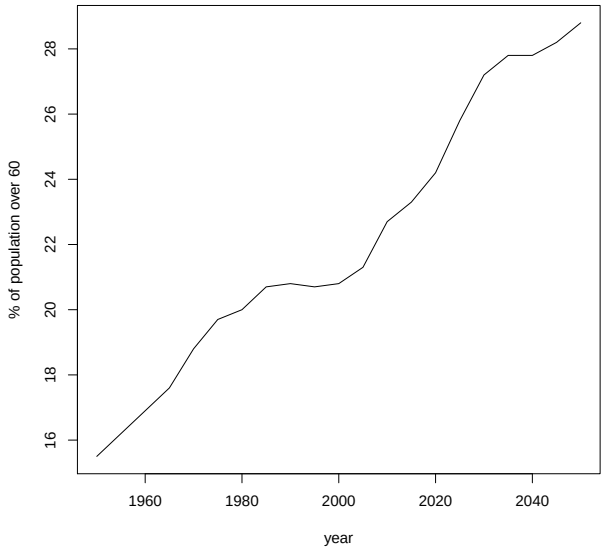


Figure 1.1: Percentage of population over age 60 in the UK. Data sourced from the United Nations Population Division online database.

as well as as visuoconstructive skills. Typically the test is scored by counting one point for each correct answer and using a cut-off point on the total score to decide upon further investigation (O'Bryant et al. 2008). Proposed cut-off points lie between 23 and 27 out of 30. A positive aspect of the MMSE is that it can be administered by practitioners with little or no training as a tool to rule out a diagnosis of dementia. This makes it applicable in community and primary care. Although evidence is limited, provisionally the MMSE is found to be of less value in making a diagnosis of Mild Cognitive Impairment (MCI) (Mitchell 2009).

There is a common agreement, even among lay evaluators, of what comprises an impaired cognitive state. Evidence shows that lay interpretation of the sentence writing question correlates with the overall MMSE score (Shenkin et al. 2008). The goal of formal screening tests however lies in assessing the cognitive state of a patient objectively. The MMSE is an externalisation of what was previously a "vague and subjective impression of cognitive disability" (Folstein, Folstein and McHugh 1975).

The performance of screening tests is a trade-off between accuracy and

applicability. The greatest accuracy is provided by complex screening tests. According to Bush, Kozak and Elmslie (1997) and Haubois et al. (2011) the majority of general practitioners find screening for dementia important however only 24% routinely screen their patients. Reasons cited include lack of time and fear of offending the patients. In this context complex screening tests find less acceptance than tests, which are less accurate but quicker and easier to administer.

To increase the accuracy of the MMSE it is necessary to interpret the MMSE beyond comparing the total score against a cut-off point. However, this additional information needs to be extracted in a manner which does not compromise the simple application of the MMSE. In this thesis an additional layer of interpretation is proposed as an extension to the traditional MMSE scoring method. In this layer, using inductive reasoning, patterns of answers will be identified, which are typical for specific progressions of the future cognitive state of a patient. The challenge of the project is to identify patterns of answers, which predict the type of dementia a patient has with high probability.

Unfortunately the MMSE has become subject to strictly enforced copyright restrictions (Newman and Feldman 2011) allowing the reprint of only up to three MMSE questions. For this reason the individual questions of the MMSE in this thesis are referred to by their number only. An exception is the question which asks the patient to write a sentence which is treated in more detail. The test itself can be found in the seminal paper of Folstein, Folstein and McHugh (1975).

1.2 Aims and Objectives

The scope of this research is the early and accurate detection of dementia by means of the Mini Mental State Examination. Its overall aim is to introduce an automated layer of interpretation to the traditional MMSE scoring process.

In this layer scoring rules will be evaluated which would otherwise compromise the brief nature of the MMSE due to their complexity.

The MMSE will be analysed to extract information which is recorded but not considered in the current evaluation model.

The construction of the proposed interpretation method is pursued in a three staged process. In the first stage the most relevant factors of the MMSE are identified. Two types of methods for dimensionality reduction are being investigated: traditional principal components analysis and an approach with foundations in information theory. In the second stage a predictive model is devised, which associates patterns of answers to MMSE questions with types of dementia. The model will be informed by the factors extracted in the first stage. In the third stage further improvements to the model are attempted by extracting more information than just a binary score from the sentence writing question. More concretely, the reported research pursues the following objectives:

- to *reduce the dimensions* of the MMSE to the most relevant ones in order to inform a predictive model by using *computational methods* on a data set of MMSE results,
- to *construct a model* predicting a diagnosis informed by the features extracted from the previous step by applying, comparing and combining traditional and novel modelling methods,
- to propose a *semantic analysis* of the *sentence writing* question in the MMSE in order to utilise information recorded in MMS examinations which has not been considered previously.

1.3 Outline

The remainder of this thesis is organised as follows: chapter 2 reviews related literature on dimensionality reduction of MMSE data, semantic analysis of the

sentence writing question in the MMSE as well as on general methods suitable for data shaped like the MMSE; chapter 3 gives an overview of the process of collecting and analysing the data; chapters 4, 5 and 6 propose novel methods for dimensionality reduction of discrete data such as the MMSE; chapter 7 proposes linguistic markers on the MMSE sentence suitable for discriminating between diagnoses; and finally, chapter 8 applies the methods proposed in previous chapters to meet the objectives of this research.

Chapter 2

Literature Review

In this chapter an overview of related work on the MMSE data and on analytical methods is given. The chapter is organised as follows: section 2.1 gives an overview of related studies and identifies methodological limitations in reported research; information theory is identified as an adequate foundation for further analysis and a review of concepts is given in section 2.2; section 2.3 gives an overview of methods for analysing covariance with an emphasis on applicability to discrete data, such as the MMSE data; section 2.4 reviews methods for investigating the relevance of individual multivariate dimensions to a class variable using information theory; and finally, section 2.5 reviews research attempting to extract information from the MMSE sentence writing question as well as general research in identifying linguistic markers for dementia.

2.1 Componential Analysis of the MMSE

The purpose of analysing components of the MMSE is to identify groups of questions which account for most of the variation in examination results. Questions which are identified as good predictors for the different types of dementia can then be used to construct models which give more insight into the patient's mental state than just the total score.

Magni et al. (1996) have conducted a factor analysis on MMSE results in an attempt to find factors which discriminate best between two types of dementia, Alzheimer's disease and multi-infarct dementia. The authors have found that the more sophisticated method is necessary, because the total score of the MMSE provided insufficient information to distinguish between both types of dementia. In their analysis they identify two components of the MMSE

which together explain 56.6% of the variance in the data. They found that the component which explains most of the variance is also a good criterion for discriminating between both types of dementia. The questions contained in this component are the same as the ones identified by Wind et al. (1997) with a stepwise logistic regression, namely ones concerning orientation in time and space. The second component was more heterogeneous in terms of question topics. It had a strong correlation with the education level of the patients. The authors conclude that although the MMSE is a simple tool, it provides more information than expected when thoroughly analysed.

Wind et al. (1997) attempted to identify which questions of the MMSE are the most accurate predictor for dementia according to an external criterion. Their external criterion was the outcome of a Geriatric Mental State interview. The authors attempted to identify such questions by constructing a logistic regression model using the MMSE questions as explanatory variables and the external criterion as the predicted value. They applied a stepwise approach in constructing their model by incrementally increasing the number of explanatory variables in the model and only keeping the ones with the largest contribution to the predictive quality of the model. Using this method the authors identified four questions which had the best predictive ability in diagnosing dementia. These questions were about the date, the day of the week, the patient's address and the current prime minister. It should be noted that, the question about the prime minister is not part of the original MMSE. It was added to compensate for the lack of testing of general knowledge and long term memory in the MMSE. A limitation of this study is that the quality of questions as explanatory variables was only assessed for one variable at a time.

Jefferson et al. (2002) proposed an MMSE index which adds up the points of the different domains of the MMSE and scores the intersecting pentagons and the sentence writing questions on scales from 0 to 8 points each. The authors developed these scales deductively, by developing operational definitions

for a wide variety of impairments. To evaluate the index, the score for each domain was statistically compared for groups of patients with different dementia diagnoses. The authors report that taking scores of separate domains of the MMSE into consideration can assist professionals in gaining more information on the patients cognitive impairment. A limitation of this study is that each domain of questions was evaluated separately and combinations of domains were not taken into consideration as a multivariate predictor.

In a more recent publication of Guerrero-Berroa et al. (2009) the use of a similar approach to predict cognitive decline of the elderly was implemented. The authors took into consideration pairwise combinations of four MMSE domains: temporal and spatial orientation, delayed recall and attention. Other domains were excluded either based on previous research suggesting that they are affected only in progressed dementia states or because they only include one question making the scoring range too narrow for statistical analysis. The study shows that the temporal orientation is a strong predictor of cognitive decline.

The domains of questions used in these studies are chosen deductively and empirical studies were only able to verify up to five of them. The classical data driven approach for finding correlated dimensions in multidimensional data is factor analysis (FA) or principal components analysis (PCA). Magni et al. (1996) have used principal components analysis to identify groups of questions through which the groups carry information orthogonal to other groups. The authors report that the component which was most useful in differentiating between Alzheimer's Disease (AD) and Multi-Infarct Dementia (MID) was most influenced by questions from the temporal orientation domain. This result is consistent with other studies (Guerrero-Berroa et al. 2009), which have found that temporal orientation is the MMSE domain impaired earliest in the progression of AD. A limitation of this study is that factor analysis and principal components analysis in their classical form rely strongly on the classical co-

variance measure, which in turn assumes normal distribution. The values of individual dimensions of the MMSE however are either 0 or 1. In this specific case, the covariance measure becomes hard to interpret and other correlation measures are more appropriate (Jolliffe 2002, Kolenikov and Angeles 2004).

Castro-Costa et al. (2009) propose replacing the covariance measure with the tetrachoric correlation coefficient. They have identified five components, which explain most of the variation in MMSE data. Their finding is consistent with other studies (Baos and Franklin 2002) suggesting that the underlying structure of the MMSE is robust. A limitation of this study is that the sentence writing question and the intersecting pentagons question are scored on a 0 to 1 scale although both questions test for several different cognitive impairments.

2.1.1 Conclusions

Although several studies investigate the correlation structure of MMSE data, few of the reported studies employ methods which are adequate for the shape of the data. In this research methods are identified which are adequate for the data, and where such methods are not available new methods are proposed.

A notable feature of the MMSE data, which plays a role in the choice of a method, is that the number of dimensions is relatively low, there are only 30. Many variable selection methods are designed to yield useful results for tens of thousands of variables. Although 30 dimensions are more than can be feasibly analysed with purely explorative approaches, the number is low enough to allow a more exhaustive search than would be feasible with 30000 dimensions for example. A second characteristic of the MMSE data is that its values are binary. A measure particularly suitable for the analysis of discrete data is entropy. Section 2.2 gives a review of concepts in information theory.

Methods which analyse the correlation structure of data are typically methods for dimensionality reduction of which there are two general types: filter approaches and wrapper approaches (Guyon et al. 2006). In filter approaches

the dimensionality reduction process is run before the construction of a predictive model. In wrapper approaches the predictive model is treated as a black box which is used to evalute the quality of subsets of dimensions. In the scope of this thesis dimensionality reduction is considered a preliminary step in devising a predictive model.

Filter approaches for dimensionality reduction can be further classified into feature extraction and variable selection methods. Feature extraction methods attempt to project the original data so that characteristics of the data become more apparent. Variable selection methods on the other hand, attempt to order variables by the information they contribute to the target dimension. Both families of methods will be considered for the analysis of the MMSE data.

A covariance measure based on the assumption of a normal distribution is inadequate for the analysis of MMSE data. In related research(Castro-Costa et al. 2009) the alternative tetrachoric correlation coefficient has been considered. In this research, a departure from traditional covariance measures is proposed and information theory is identified as a measure particularly suitable for the analysis of discrete data. The following section gives a review of concepts in information theory.

2.2 Information Theory

The fundamental definition of information theory is entropy. Entropy is a measure of the uncertainty about a random variable (Cover and Thomas 1991). In statistics derived from the assumption of a normal distribution it is equivalent to the variance of a random variable. Entropy is defined as:

$$H(X) = -\sum_{u} p(X_{X=u}) log(p(X_{X=u})) \tag{2.1}$$

Where X is a random variable, u is a value which X can take and $p(X_{X=u})$

is the probability that X will take the value u. When the logarithm is to the base of 2, entropy is measured in bits. The value of entropy is always greater or equal to zero and its upper limit is bounded by the number of values u which X can take. When the value is high, a lot of information is obtained by observing X. On the other hand, if the value is zero, when there is only one value u with $p(X_{X=u}) = 1$, no information is obtained by observing X.

A closely related concept is conditional entropy. Conditional entropy is an extention of this idea so that instead of a univariate probability a conditional probability is estimated. It is defined as:

$$H(Y|X) = -\sum_{u,v} p(Y_{Y=u}, X_{X=v}) log(p(Y_{Y=u}|X_{X=v})) \tag{2.2}$$

where u are values which Y can take, v are values which X can take and $p(Y_{Y=u}|X_{X=v})$ is the probability that Y will take the value u when X is known to have taken the value v. The magnitude of conditional entropy is measured in bits. Its value is always non-negative and is equal to zero when all probabilities $p(Y|X)$ are equal to 1. In other words, if Y can be predicted with absolute certainty when X is known.

The magnitude of $H(Y|X)$ is dependent on the entropy of $H(Y)$. In order to compare the measure across models, it is necessary to derive a measure for the distance between Y and X which is not dependent on the structure of Y and X themselves. Mutual information $I(X;Y)$ is such a measure, which is interpreted as the amount of uncertainty about Y reduced due to the knowledge of X. It is defined as:

$$I(Y;X) = H(Y) - H(Y|X) \tag{2.3}$$

Mutual information is always non-zero and exactly zero when $H(Y) = H(Y|X)$ or in other words when knowing X does not reduce the uncertainty of Y. The upper bound of the function is $H(Y)$ and is reached when $H(Y|X)$ is zero.

This is the case when Y is functionally dependent on X. Mutual information corresponds to covariance in conventional statistics.

When X_i is a dimension of a multivariate data set X, it is of interest to measure not only the uncertainty X_i reduces about Y but also the redundancy between X_i and and X_j. A measure derived from entropy which takes these two factors into account is conditional mutual information (CMIM). It is calculated as the difference between the entropy of Y conditional on a variable X and the entropy of Y conditional on two variables X_i and X_j.:

$$I(Y; X_i|X_j) = H(Y|X_j) - H(Y|X_i, X_j) \tag{2.4}$$

Intuitively the meaning of CMIM can be interpreted as the amount of uncertainty reduced by X_i about Y when X_j is known. Conditional mutual information has no trivial equivalent in traditional statistics.

2.3 Principal Components Analysis

The most widely used method (Tinklenberg et al. 1990, Hill and Backman 1995, Magni et al. 1996, Brugnolo et al. 2009) for the analysis of components in the Mini-Mental State Examination is factor analysis. Factor analysis is a method closely related to the more common principal component analysis (PCA). Principal component analysis relies on the accuracy of the covariance matrix of the data. However, the variance and covariance measures become less intuitively interpretable when they are estimated from binary data. In literature it is recommended (Jolliffe 2002) to use a tetrachoric correlation coefficient to adjust the measure.

PCA identifies linear combinations of variables with maximal sample variance among all possible linear combinations such that the principal components are minimally correlated with each other. Such combinations are identified by analysing the variance-covariance matrix. Maximising a function of

several variables subject to one or more constraints is done by application of the method of Lagrange multipliers. It can be derived that the solution is equivalent to finding the eigenvectors and eigenvalues of the covariance matrix (equation 2.5).

$$\begin{pmatrix} var(X_1) & cov(X_1,X_2) & \dots & cov(X_1,X_i) & \dots & cov(X_1,X_p) \\ cov(X_2,x_1) & var(X_2) & \dots & cov(X_2,X_i) & \dots & cov(X_2,X_p) \\ \vdots & \vdots & & \vdots & & \vdots \\ cov(X_p,x_1) & cov(X_p,X_2) & \dots & cov(X_p,X_i) & \dots & var(X_p,X_p) \end{pmatrix} \tag{2.5}$$

where X_i are dimensions of the random variable X and $var(X_i)$ and $cov(X_i, X_j)$ are defined as:

$$var(X_i) = \sum_{j=1}^{n} \frac{(X_{i_j} - \bar{X}_i)^2}{n-1} \tag{2.6}$$

$$cov(X_i, X_j) = \sum_{k=1}^{n} \frac{(X_{i_k} - \bar{X}_i)(X_{j_k} - \bar{X}_j)}{n-1} \tag{2.7}$$

Once the eigenvectors are derived, they are sorted in descending order of their eigenvalues. The ith eigenvalue is equal to the variance of the ith principal component. Subsequently it follows that the principal component with the largest eigenvalue corresponds to the variable explaining most of the variance in the data.

The method as it is described above has a bias towards variables on a larger scale. The reason for this is that if x is scaled by a constant, the variance is scaled by the square of the constant. Since the PCA method is looking for variables with large variance, it is biased towards variables on a larger scale. This issue can be resolved by using a correlation matrix instead of a covariance matrix. The correlation is defined as the variance of the data after it has been

transformed so it is on a scale between -1 and 1:

$$\rho(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)var(X_j)}} \tag{2.8}$$

2.3.1 PCA with a Tetrachoric Correlation Coefficient

When the variables of the data are binary, the classical estimation of the correlation coefficient becomes difficult to interpret. A more natural estimation method is using a tetrachoric coefficient. In this method it is assumed that the variables, although they are observed in a binary fashion, have an underlying standard normal distribution. Let X be a normally distributed random variable. If an observation of X is beyond a certain threshold, the value is recorded as 1 or 0 otherwise. The task is to find parameters of a normal distribution which satisfy the frequencies on both sides of the threshold. When calculating a correlation coefficient there are two variables which are split along one threshold each. Assuming that the frequencies in the four quadrants thus formed are known and assuming that the variables are sampled from a bivariate normal distribution, paramaters for such a distribution can be estimated numerically. Similarly if the observations are not binary but are categorical, a polychoric correlation coefficient can be estimated by matching frequencies in more than four sections of a multivariate normal distribution. The resulting estimate for the correlation can be used to build a correlation matrix which is then used to do an otherwise classical principal components analysis.

In figure 2.1 for example, all points in quadrant I might be coded as $(0, 0)$, in quadrant II as $(0, 1)$, in III as $(1, 1)$, and as $(1, 0)$ in quadrant IV. In order to estimate the theoretical bivariate normal distribution these points were drawn from (represented as an ellipse), one would count the number of points in each quadrant and then approximate the parameters of a normal distribution (covariance matrix and a mean) which are most likely to have produced the observed counts in each quadrant.

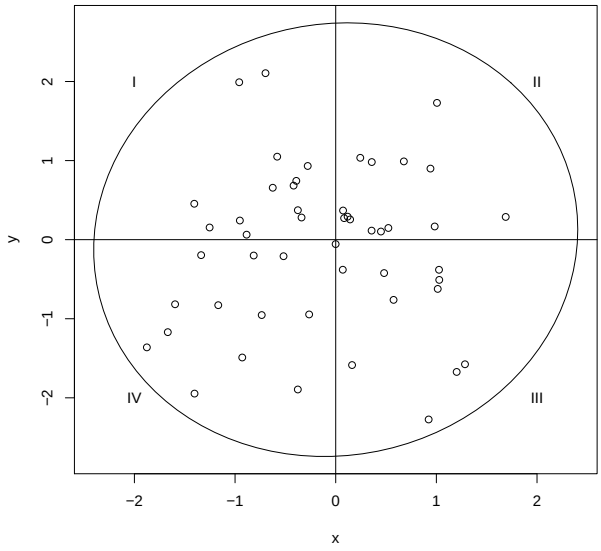


Figure 2.1: Example of an empirical and theoretical bivariate normal distribution.

2.3.2 Kernel PCA

In traditional PCA the goal is to solve the eigenproblem

$$\lambda\alpha = \Sigma\alpha \quad \text{s.t.} \quad \alpha'\alpha = 1 \tag{2.9}$$

where Σ is the covariance matrix, λ are the eigenvalues and α the eigenvectors (Jolliffe 2002). Because the method can be expressed purely in terms of a dot product, it can be reformulated to make use of the kernel trick (Schlkopf, Smola and Müller 1999). Application of the kernel trick is equivalent to projecting the data onto a space with a higher (not necessarily finite) dimensionality before analysis. If the covariance matrix is denoted as:

$$\Sigma = \frac{1}{d}\sum_{j=1}^{d} x_j x_j^T, \tag{2.10}$$

and $\Phi(X)$ denotes a projection of the type:

$$\Phi : \mathbf{R}^N \to F, \tag{2.11}$$

then kernel PCA is equivalent to solving the eigenproblem:

$$C = \frac{1}{d}\sum_{j=1}^{d}\Phi(x_j)\Phi(x_j)^T$$
$$\lambda\alpha = C\alpha. \tag{2.12}$$

A kernel function K is a function which performs in closed form the projection Φ together with the dot product $(\Phi(x_i), \Phi(x_j))$ in such a way, that the evaluation of K is significantly easier than explicitely projecting X before calculating the dot product of x_i and x_j. In cases where the data is projected onto a space with an infinite number of dimensions, an explicite calculation of the dot product is impossible even theoretically. To make the use of a kernel function applicable in PCA, the eigenvalue problem is reformulated as:

$$d\lambda\alpha = K\alpha \tag{2.13}$$

Kernel PCA for Discrete Data

Until this time, no research is available which reports a componential analysis of the Mini-Mental State Examination by means of kernel PCA. However, several researchers report applying kernel PCA on data with binary variables (Wu, Su and Carpuat 2004) and data which includes categorical variables (Rosipal et al. 2001). A limitation found throughout literature reporting kernel PCA of discrete data is that there is no comparison of different kernel functions.

One of the benefits of kernel methods is that kernel functions can be replaced for one another independently of the method. Limited work on the error estimation of kernel PCA models is reported by Twining and Taylor (2003) and Heafield (2005).

	$lake_1$	$lake_2$	$lake_3$
$fish_1$	1	5	4
$fish_2$	12	5	2
$fish_3$	15	8	2

Table 2.1: Fish species in each lake.

2.3.3 Homogeneity Analysis

A less traditional approach to non-linear PCA, and multivariate data analysis in general, is the Gifi system (Gifi 1990, De Leeuw and Mair 2009). The Gifi system is designed for categorical variables and any continuous variables are first converted to categories. Under the Gifi System, observations and measurements are recognised equally. If for example data is collected about the prevalence of different fish species in different lakes (table 2.1) then the data can be interpreted as information about the differenes in lakes but also as information about the differences between the numbers of fish species on a national level..

	$lake_{1;1}$	$lake_{1;12}$	$lake_{1;15}$	$lake_{2;5}$	$lake_{2;8}$	$lake_{3;4}$	$lake_{3;2}$
$fish_1$	1	0	0	1	0	1	0
$fish_2$	0	1	0	1	0	0	1
$fish_3$	0	0	1	0	1	0	1

Table 2.2: Fish species in each lake – indicator matrix.

To represent this relation, the data is coded in the form of an indicator matrix where each column represents the occurence of a specific value in a given row (see table 2.2). Let this binary indicator matrix be denoted as G and the columns corresponding to the ith column in the original matrix be denoted as G_i, then the function to be optimised is defined as:

$$tr\{\sum_{j=1}^{p}(G_jc_j - Y_jb_j)'(G_jc_j - Y_jb_j)\} \tag{2.14}$$

where tr denotes the trace of a matrix, or the sum of its elements on the diagonal. The relation between rows and columns is captured in the matrix

C where each column c_j contains the weights of each category for the jth column of the original matrix. The matrix Y is composed of the scores of p principal components and the matrix B has the coefficient b_j of each principal component in its diagonal. The minimisation of the function takes place with respect to c_j and $Y_j b_j$ at the same time. The difference from linear PCA is that in addition to the optimisation of the scores on the p principal components, there is also an optimisation over the values of the variables. A solution is found by applying an alternating least squares algorithm which fixes c_j and optimises $Y_j b_j$ and then fixes $Y_j b_j$ and optimises c_j until convergence is reached.

This algorithmic optimisation of the function implies that the non-linear projection of the data cannot be expressed in closed form. In other words, the method does not provide a model explaining the data but instead provides insight into the structure of the data.

2.3.4 Entropic PCA

Data projection methods which are based on information theory are surprisingly few. Most of the recent advances in the field (Hild et al. 2006, Qiu and Wu 2009, He et al. 2010) extend seminal work of Torkkola (2003) where an entropy measure is proposed which uses a Parzen window to estimate the probabilities necessary to compute the entropy of a variable. In the Parzen window method a Gaussian kernel is placed ontop of each sample and the probability density function is estimated as a sum of the kernels. While this method yields good results for continuous data, the difficulty of applying it on discrete data begins on a conceptual level. When a variable is nominal, say the profession of a person, estimating probabilities as a multi-modal Gaussian distribution is unreasonable. Such a model would imply not only that there is a measurable distance between the professions “Medical Doctor” and “Lawyer” for example, but also that along this distance there is an infinite number of other professions. For this reason, these methods are excluded from the review.

Although the relation between correlation functions and mutual information has been thoroughly investigated (Cover and Thomas 1991), only one publication reports a principial component analysis based on mutual information (Bollacker and Ghosh 1996) which is applicable to discrete data. The authors propose two covariance measures based on entropy which they use to construct a covariance matrix in a supervised manner. The methods they propose, SMIFE1 and SMIFE2 (Separated Mutual Information Feature Extractor) are a composition of terms with foundations in information theory. Both methods use a term denoted as $I(X_i; X_j; Y)$ named "three-variable mutual information". This term is in concept related but not equivalent to the 3-way interaction gain term proposed by Jakulin and Bratko (2004). Three-variable mutual information is defined as:

$$\begin{aligned} I(X_i; X_j; Y)_{SMIFE1} &= H(X_i, X_j, Y) - H(X_i) - H(X_j) && (2.15) \\ &\quad -H(Y) + I(X_i; Y) + I(X_j; Y) && (2.16) \\ &\quad +I(X_i; X_j) && (2.17) \end{aligned}$$

and is used as a covariance function in the SMIFE1 method without further change. The SMIFE2 method is derived from the above and defined as:

$$I(X_i; X_j; Y)_{SMIFE2} = I(X_i; Y) + I(X_j; Y) - I(X_i, X_j; Y). \quad (2.18)$$

in order to interpret the term it is informative to reformulate SMIFE2 by substitution as:

$$
\begin{aligned}
I(X_i;X_j;Y)_{SMIFE2} &= I(X_i;Y) + I(X_j;Y) - H(X_i,X_j,Y) && (2.19)\\
&\quad -H(X_i) - H(X_j) - H(Y) && (2.20)\\
&\quad +I(X_i;Y) + I(X_j;Y) + I(X_i;X_j) && (2.21)
\end{aligned}
$$

which simplifies to:

$$
\begin{aligned}
I(X_i;X_j;Y)_{SMIFE2} &= H(X_i,X_j,Y) - H(X_i) - H(X_j) && (2.22)\\
&\quad -H(Y) + H(X_i) - H(X_i|X_j) && (2.23)\\
&= H(X_i,X_j,Y) - H(X_j) - H(Y) && (2.24)\\
&\quad -H(X_i|X_j). && (2.25)
\end{aligned}
$$

If multivariable entropy is visualised in terms of set theory for the purpose of understanding (Cover and Thomas 1991), then $H(X_i, X_j, Y)$ can be read as the union of information conveyed by each of the three variables. After subtraction of the terms $H(X_j$ and $H(Y)$, the remainder can be roughly read as $H(X_j|X_i, Y)$ however under violation of the inclusion/exclusion principle since the intersection of $H(X_i)$ and $H(Y)$ is subtracted twice from $H(X_i, X_j, Y)$. From this remainder which approximates $H(X_j|X_i, Y)$ the set $H(X_i)$ under exclusion of $H(X_j)$ is removed. Unfortunately the authors do not provide any reasoning for this unusual treatment of entropy quantities.

2.3.5 Conclusion

While projection methods for analysing covariance structure in discrete data are available, the MMSE has not been analysed with them – a notable exception is reported by Castro-Costa et al. (2009). A research gap is identified in projection methods which are based on entropic covariance estimates.

2.4 Variable Selection with Information Theoretic Criteria

While the previous section introduced dimensionality reduction by means of projection, the following section discusses dimensionality reduction by variable selection. The goal of variable selection is to select a subset of variables of a high dimensional dataset so that the subset predicts the class of an observation with as little redundancy as possible. Information theoretic methods for variable selection are particularly well suited for the analysis of discrete data. This type of methods are computationally intensive and have recently gained in popularity due to the increased availability of powerful hardware.

A considerable research effort on information theoretic methods for variable selection with continuous variables is reported in related literature (Torkkola 2003). However in this section only discrete variables are considered.

In the remainder of the thesis Y denotes the class of an observation; X_i denotes the ith dimension of a multi dimensional random variable X; $X_{i=u}$ denotes the random variable X_i taking the value u; X_S denotes a projection of X such that only dimensions are included which are selected according to some criterion; X_{-S} is a projection of X such that only dimensions are included which have not been selected but are potential candidates for inclusion into X_S; d denotes the number of dimensions of X; n denotes the number of observations of X.

2.4.1 Goal Function

Dimensionality reduction is usually treated as an optimisation problem. To solve or approximate an optimisation problem, a precise definition is needed of what is being optimised. In variable selection a function is maximised which estimates the amount of information about the class of observations contained in a subset of dimensions. One extremum of this function is reached when the class variable is functionally dependent on the subset of dimensions. The other extremum is reached when the subset of dimensions is independent from the

class variable.

2.4.2 Naive Approach

Mutual information $I(Y;X)$ (see section 2.2) can be used for variable selection without alteration. Let X be a p dimensional multivariate random variable and $X_{1...k}$ be a random variable composed by a subset of $k < p$ dimensions of X. The measure $I(Y;X_{1...k})$ will then be interpretable as the uncertainty about Y which is reduced by the knowledge of $X_{1...k}$. If this quantity is estimated for all subsets of dimensions of X with size k, then the subset with the largest value of the metric can be considered the subset of variables which best explain Y.

Several problems are inherent to this naive approach. Firstly, the metric needs to be estimated for $\binom{p}{k}$ subsets of dimensions. With increasing p and k, this requirement leads to computational infeasibility due to combinatorial explosion. Secondly, to estimate $H(Y|X_{1...k})$, one needs an estimate of the probability distribution $P(Y|X_{1..k})$. The amount of data required to estimate this function becomes prohibitive with a k larger than 3-4 (Battiti 1994). Consider that even if enough data were available, the probability of each value of Y conditional on any combination of values of $X_1, ..., X_k$ needs to be estimated. The number of probabilities to be estimated grows exponentially in the number of values that the variables can take. If $Y, X_1, ..., X_k$ are binary variables, then 2^{k+1} probabilities have to be estimated for each of the $\binom{p}{k}$ subsets of dimensions. A third problem with the naive approach is that although the ideal subset of variables can be determined, the metric does not give information about the quality of each of the selected variables.

2.4.3 Classical Method

The classical method, Information Gain (IG) (Manning, Raghavan and Schtze 2008), to address the difficulties of the naive approach is to consider the amount of uncertainty about the target variable that is reduced by selecting a single

dimension X_i of X. To select a subset of k dimensions of X, $I(Y;X)$ is evaluated for all X_i and Y pairs and k dimensions are chosen with the highest mutual information with Y. Note that the function requires the estimation of at most two dimensional probabilities and that it is evaluated p times, where p is the number of dimensions of X. This brings the computational complexity and the data sparsity problems under control. In addition, the contribution of each individual variable to the explanation of Y is measured. A limitation of this method is that it disregards redundancy in the data. Let X_i and X_j be two variables which explain Y very well, but which are also functionally dependent on each other (there is an $f(x)$ such that $f(X_i) = X_j$). If X_j is included in the model, when X_i has already been chosen, X_j contributes no additional information. The goal function will however still score X_j as high as X_i.

2.4.4 Conditional Mutual Information

An approach which is a compromise between the naive and classical method is to use conditional mutual information $I(Y;X_i|X_j)$ (see section 2.2) as the goal function. When using this goal function, at most three dimensional probability functions need to be estimated. The complexity of an exhaustive search using this approach is $\binom{d}{2}$ where d is the number of dimensions. The reason for this is that the conditional mutual information has to be evaluated for all pairs of dimensions X_i and X_j $(i \neq j)$. While this complexity is a lot more manageable than the exponential complexity of the naive method it can still become hard to compute for even simple problems. For example, a relatively small image of 100 by 100 pixels will have 10000 dimensions, if each pixel is treated as a dimension. Even for an image of this size the function would need to be evaluated for close to 50×10^6 pairs of dimensions. If the evaluation of the function could be performed in 1 millisecond (which is an unrealistically low estimate) the computation would still take almost 14 hours to complete.

Fleuret (2004) propose an algorithm which makes variable selection with a conditional mutual information criterion computationally feasible for very high-dimensional data as typically found in image recognition. Their method relies on two crucial ideas: bit counting using look-up tables and a heuristic forward search through the space of subsets of dimensions.

The most time-consuming part of estimating entropy based measures is estimating point-wise probabilities. In the context of this thesis they are computed by counting the number of observations, in which the variable X takes a given value u, and dividing this number by the total number of observations. While related literature proposes other methods for estimating CMI under assumptions of probability distributions (Jung, Seo and Kang 2011, Tsimpiris, Vlachos and Kugiumtzis 2012), such methods are designed for continuous data and have limited applicability in the discrete case.

In the estimation method proposed by Fleuret (2004) the data is arranged in such a way, that each 32 values of one column of the data are stored in one of the 32 bits of an integer variable. A look-up table provides information on how many of the 32 bits of the integer value obtained that way are set to one.

The combinatorial explosion of the search space is dealt with by using a directed search rather than an exhaustive computation. Variables are first sorted by their mutual information with the class. In the first step, the variable with the highest score is selected. In subsequent steps variables are selected so that the minimal conditional mutual information $I(Y; X_j|X_s)$ is maximal, where Y is the class, X_j is the candidate variable and X_s is each of the selected variables.

$$X_{S_1} = \max_{X_i \in X}(I(Y; X_i)) \tag{2.26}$$

$$X_{S_{>1}} = \max_{X_j \in X_{-S}}\left(\min_{X_i \in X_S}(I(Y; X_j|X_i))\right) \tag{2.27}$$

where X_S denotes the set of selected variables, X_{-S} denotes the set of

candidate variables, X_{S_1} denotes the first selected variable and $X_{S_{>1}}$ denotes all variables selected after the first.

The procedure is repeated until the most informative variables are selected. The benefit of this approach is that 32 (or 64, depending on the platform) rows of data can be examined at a time. However, because the values are stored in single bits, the method is limited to analysing only binary data in not more than two classes.

2.4.5 Approximating CMI

An approach to estimating CMI while avoiding computational complexity is to avoid calculating the exact CMI estimate and instead approximate it with terms which are easier to calculate.

One such approach is the Mutual Information Feature Selection under Uniform Information (MIFS-U) method proposed by Kwak and Choi (2002). The authors recognise that direct computation of CMI is expensive and propose a form of CMI which is equivalent under certain assumptions but only requires calculating MI. The assumption they make is that conditioning on the class variable Y does not alter the ratio of information between the candidate variable X_i and the selected variable X_j:

$$\frac{H(X_j|Y)}{H(X_j)} = \frac{I(X_j; X_i|Y)}{I(X_j; X_i)} \tag{2.28}$$

.

Under this assumption the following equivalency can be used to approximate CMI:

$$I(Y; X_i|X_j) = I(Y; X_i) - \frac{(I(Y; X_i) * I(X_i; X_j))}{H(X_j)} \tag{2.29}$$

Although this method can be be used as a goal function in a greedy forward search, the authors propose using the Taguchi method, a form of wrapper

method which uses training neural networks to guide the search.

Another approach which avoids direct computation of CMI is the Maximal Relevance Minimal Redundancy (MRMR) method proposed by (Peng, Long and Ding 2005). The authors propose a measure which maximises mutual information but which also contains a term to adjust for redundancy within the subset of selected variables. The criterion for selecting X_i with the MRMR method is formulated as:

$$X_{MRMR} = \max_{X_i \notin S}(I(Y; X_i) - \frac{1}{|S| - 1} \sum_{X_j \in S} I(X_i; X_j)) \tag{2.30}$$

,

where S is the subset of selected variables and X_i is a candidate variable not in S. The method proceeds to select variables from the remaining candidates with a greedy forward search.

2.4.6 Search Strategies

Typically variable selection using information gain is done using sequential forward search. The forward search can be greedy, as proposed by Quinlan (1993), or guided by a meta-heuristic approach (Huang, Cai and Xu 2007). One difficulty with forward search methods is that they are biased to overestimate the quality of variables which are selected in later iterations. Methods based on the mutual information criterion deal with this issue by altering the goal function with a penalising term, which grows as a function of the number of selected variables. A recent example of such a method is the normalized mutual information criterion which is reported to yield significantly better results than other methods when 10 or more variables have been selected (Estévez et al. 2009). The authors define normalized mutual information $NI(X_1, X_2)$ as:

$$NI(X_1, X_2) = \frac{I(X_1; X_2)}{\min\{H(X_1), H(X_2)\}} \tag{2.31}$$

This term is averaged over the selected variables to yield the goal function G_{NI}:

$$G_{NI}(X_i) = I(Y; X_i) - \frac{1}{|X_s|} \sum_{X_j \in X_s} NI(X_i; X_j) \tag{2.32}$$

where X_i is the variable which is a candidate for selection.

Liu et al. (2009) propose an algorithmic adjustment of the mutual information criterion to account for the bias toward lately chosen variables. Instead of relying on a penalising term, their algorithm partitions the data after each iteration. The partitioning is done in such a way, that observations, which predict the class without ambiguity by using only the selected variables, are not considered for the estimation of probabilities in subsequent iterations. The authors report good results when comparing their algorithm with classical methods. They do not compare their method with conditional mutual information based goal functions as currently the only feasible method for this computation is limited to binary variables and two classes. This limitation is satisified only by few well known data sets and thus makes it difficult to compare the method with other methods.

2.4.7 Three-way interactions

A recently published method for variable selection with information theoretic ranking criteria is based on three-way interaction gain (3IG) (Akadi, Ouardighi and Aboutajdine 2008). This measure is defined by Jakulin and Bratko (2004) as:

$$I(X_i; X_j; Y) = I(X_i, X_j; Y) - I(X_i; Y) - I(X_j; Y) \tag{2.33}$$

It can be interpreted as a measure for the reduction of uncertainty caused by joining the variables X_1 and X_2 in a Cartesian product. It is important to note that this measure can be negative when X_1 and X_2 carry the same information.

Akadi, Ouardighi and Aboutajdine (2008) propose a variable selection goal function based on 3IG. This function incorporates mutual information, 3IG and an adjustment to devalue variables which are selected in later iterations. The Interaction Gain Feature Selection (IGFS) is defined as:

$$X_{IGFS} = \max_{X_i \in X_{-S}} I(X_i; Y) + \frac{1}{|S|} \sum_{X_j \in X_S} I(X_i; X_j; Y) \tag{2.34}$$

The authors report better results than other state-of-the-art information theoretic methods for variable selection including the conditional mutual information criterion.

A limitation of the IGFS goal function is that it requires the computation of three-way interactions which is a time consuming task. It can be shown, that IGFS can be transformed into an equivalent form in which the hardest to compute term is an estimate of conditional mutual information. A proof can be found in section 6.1.

2.4.8 Clustering

The methods presented above demonstrate that sequential forward search does not yield the best results. The attempts to resolve this issue range from algorithmic solutions (Kwak and Choi 2002) to introducing corrective terms (Estévez et al. 2009). A theoretical treatise of forward and backward search bias is given by Van Dijck and Van Hulle (2010). One approach which circumvents the issue of sequential searching is Minimal Relevant Redundancy Clustering (mRRC) proposed by Martínez Sotoca and Pla (2010). In this approach the authors propose a distance metric based on CMI. A complete distance matrix is calculated for each pair of variables after which the matrix is used to perform a clustering of variables. The distance measure is defined

as:

$$D(X_i; X_j) = I(Y; X_i|X_j) + I(Y; X_j|X_i) \tag{2.35}$$

and the clustering algorithm used is Ward's method. The reasoning behind using a clustering approach is that variables which carry similar information will be close to each other within the proposed metric and thus will be in the same cluster. The number of clusters required of the clustering method is equal to the number of desired variables. Once each variable has been labelled, one variable per cluster is chosen for the final subset of selected variables. Since it can be assumed that variables X_i belonging to a cluster C_j carry the same information, the variable X_i with the maximal mutual information $I(Y; X_i)$ can be chosen as representative for the cluster.

2.4.9 Conclusion

In order to alleviate the difficulties of the naive approach, several methods have been proposed. To achieve computational feasibility, (i) the search space is limited by reducing the dimensionality of the goal function, (ii) a guided forward search through the problem space is used instead of an exhaustive search and (iii) the domains of application are restricted, for example by only considering binary data. One method which does not make any of those compromises is mRRC, however mRRC runs prohibitavely long on large data based on several thousands of observations.

Works cited in this thesis consistently report an improvement of results by using conditional mutual information instead of just mutual information. This is interpreted as evidence that increasing the dimensionality of the goal function reduces redundancy in the resulting model.

2.5 Sentence Writing Analysis

Previous research has investigated the answer to the question of the MMSE asking a patient to write a sentence for additional information. Shenkin et al. (2008) compared several metrics (including word count, frequency, first person use, time orientation and letter case). In addition lay raters were asked to assess the state of patients by reading the MMSE sentence. The authors found that although the objective criteria they proposed showed no association with variables of the MMSE or cognitive impairments, the naive rating was a a good estimate of the test score. The conclusion of this study is that criteria exist which can differentiate between different states of cognitive impairment and furthermore the challenge is to identify these. Press et al. (2012) measure the number of words per sentence in the MMSE as well as positive or negative emotional polarity. The authors find that the number of words correlates well with the degree of cognitive impairement while polarity can be associated with depression.

2.5.1 Linguistic Markers for Dementia

Several authors have researched the difference in language use between people with and without dementia.

Deleon et al. (2012) prompt testees to produce sentences with varying syntactic complexity. They then draw conclusions from the number of successfully produced sentences. The study targets 11 sentence structures with two items for each of them. For example, a test administrator will prompt: “She owes her friend a dollar. She goes to see her friend. She takes out a dollar. What next?” expecting the response: “She gives her the dollar.” The goal is to use the test results to discriminate between different types of neurodegenerative dementias and normal aging. The authors report that patients with non-fluent agrammatic PPA produce fewer correct sentences when compared to the norm or other forms of dementia. However, due to the relatively small sample of

58 individuals and the imbalance of group sizes it stands to reason that the reported outcome may be the result of sample bias.

Gross et al. (2012) sentence processing rather than producing. They present testees with a sentence and ask them to match the sentence to one of two pictures. The result of the test is used to discriminate between patients with Parkinsons Disease and patients Parkinson Dementia (PD) or Dementia with Lewy Bodies (DLB). The authors report that sentences lenghtened with a prepositional phrase as well as sentences with a centre embedded clause were good discriminators. The result is consistent with other research (Roark et al. 2011) which argues that deviations from the typical parse tree of the english language, which tends to be right-heavy, indicate increased syntactic difficulty.

These findings are of limited value in investigating the MMSE as the researchers assume an imposed linguistic difficulty. In the MMSE on the other hand, the testees are free to vary the complexity of the sentence, typically choosing to write a sentence of up to 9 words (see section 7.1). This length is too short to demonstrate complex sentence structure.

A more directly applicable finding is reported by Vigliocco et al (Vigliocco et al. 2011) – the report reviews whether nouns and verbs are processed in separate neural networks in the brain. The outcome is that the neural network is shared implicating that grammatical class is not an organizational principle in the brain. Nevertheless the authors argue that verbs demand higher cognitive load than nouns because verbs have necessarily more participants than nouns and because verbs are more complex morphologically. They find that while the count of verbs and nouns may give indication to impairment, it does not point at disease in specific brain regions.

A further question, investigated by Bencini et al. (Bencini et al. 2011), is wether linguistic markers for cognitive impairment are language specific. The authors compare the underuse of subject phrases between 12 Italian and 10 English speaking patients with dementia and conclude that Italian speakers

omit more subjects than English speakers. This can be explained with the grammatical structure of the italian language which allows subject omission.

2.5.2 Automated Assessment of Linguistic Markers for Dementia

While there is ample research reporting on sentence structure in language used by patients with dementia, the availability of automated sentence analysis tools has been largely neglected with some noteable exceptions.

Roark et al. (2011) derive measures for detecting Mild Cognitive Impairment (MCI) in spoken language. To produce a corpus of data, the authors ask 74 testees (37 with MCI and 37 at norm) to retell a story. The audio is recorded, transcribed and analysed with natural language processing tools based on the Stanford language parser (Klein and Manning 2003). The hypothesis of the analysis is that when english sentences are parsed, the parse tree tends to branch to the right. Three metrics are applied to measure the right tendency of sentences: Yingve and Frazier scores as well as dependency distance. In addition, the number of propositions and content density are also measured. The authors report that the number of words per sentence clause and the Yingve score are highly significant in identifying MCI. The other measures are beneath the significance threshold. The results are consistent with the findings of Ahmed et al. (2013) in their study of 18 cases of autopsy confirmed dementia at an early stage.

The research conducted on the largest number of sentences per patient however also on the smallest number of patients is reported by S et al. (2011). The authors analyse 4 books written by Iris Murdoch, an irish author with Alzheimer's Dementia. Murdoch was known to prohibit editing of his works prior to publishing and is thus particularly suitable for an analysis of the progression of language impairment as the disease progresses. The metrics with significant decline between books are words per sentence as well as Yingve and Frazier score.

2.5.3 Conclusion

The research reviewed in this section is conducted on spoken language and on connected text. This differs from the MMSE where a testee is asked to write a sentence with a pen. While the findings are not directly applicable to the MMSE, the research demonstrates that by using automated sentence parsing, the number of sentences and the number of indicators which can be tested can be dramatically increased in comparison to the manual approach.

2.6 Findings of the Literature Review

The remainder of this thesis is strongly guided by the conclusions drawn from the literature reviewed in this chapter:

1. Related research does not report analysis of the MMSE with non-linear methods or methods which do not assume a normal distribution (see section 2.1) despite the shape of MMSE data which is inherently non-normally distributed.

2. Measures derived from Information Theory are a suitable alternative to assuming a normal distribution (see section 2.2).

3. The MMSE has been analysed componentially, however there are no reports of applying variable ranking methods to the MMSE. The result of variable ranking methods (typically variable selection methods) are more easily interpretable than componential analysis methods.

4. There is a research gap in projection methods for discrete data using information theory as their foundation (see section 2.3).

5. Measures from information theory are computationally intensive and thus a time-efficient method to estimate them is needed (see section 2.4).

6. Although related research reports evidence for value of the MMSE sentence writing question beyond a binary score, no concrete findings applicable in practice are reported (see section 2.5).

These issues are addressed in the following chapters. A time-efficient computational method for estimating information theory measures is proposed in chapter 4; a method suitable for componential analysis of the MMSE is proposed in chapter 5; a method suitable for variable ranking (or selection) of discrete data, such as the MMSE data, is proposed in chapter 6; the MMSE is analysed with the proposed methods and a predictive model transferrable into practice is proposed in chapter 8.

Chapter 3

Research Methodology

The study is designed to systematically investigate the relation between answers to questions of the MMSE and the diagnosis of patients. Once identified, this relation is used to predict the type of dementia a patient may have.

3.1 MMSE Data

The data analysed in this study were provided by the Memory Clinic at University Hospital Llandough (Great Britain). It was collected between the years 1991 and 2009 from 778 patients and includes answers given to each MMSE question, patient age, gender and diagnosis. Of the 778 records analysed, 472 are of female patients and 316 are of male patients. Patient age varies between 35 and 96 years, with the vast majority of patients being between 60 and 95 years old. The imbalance in gender can be explained with the higher life expectancy of women.

The study includes patients who were referred to the Memory Centre at Llandough hospital. Patients who exhibit dementia symptoms but are not in treatment are excluded. These criteria select the population which would benefit most from an improved MMSE. To avoid selection bias patients were not asked for consent to participate in this project. This decision raises two major ethical issues: the issue of asking patients for consent and the issue of handling confidential data. The decision not to ask participants for consent is merited for two reasons: first, the project pursues a goal of recognised benefit for the population considered in the study – namely early and accurate diagnosis of dementia; second, the study does not expose patients to any potential risks. The second issue is addressed by observing only data which is not person

identifiable.

The size of the sample is dictated by the size of readily available data at Llandough Hospital. Its size guarantees a power level for multivariate correlation very close to one, even though up to 45 regression variables are taken into consideration. One difficulty with the shape generally found in medical data is that different classes of observations usually vary in size. Data mining methods, such as neural networks or decision tree induction, are sensitive towards unequal sizes of classes in the training data and usually bias towards the larger classes.

The score of the MMSE has been found to correlate with age, sex and educational background of patients (Crum et al. 1993). To account for these biases age and sex are included in the data as additional variables. Educational background has not been considered in this study as this variable is not recorded as part of the routine screening procedure.

Twenty-seven different diagnoses are found within the dataset. All diagnoses were made using standard criteria. Many of these 27 types of dementia have been observed too infrequently to allow quantitative analysis. Before analysis, the types of dementia were grouped by difference in clinical management. The frequencies of different types of dementia and the grouping of the dementia types is listed in table 3.1.

A second clinically motivated transformation of the data is the interpretation of missing answers. In this study missing answers are interpreted as wrong. Some more severly impaired patients are not asked every question if it is clear that they would not be able to provide an answer, for example in the case of comprehension problems. Another example are questions the correct answer of which depends on the correct answer of previous questions. For instance, if the patient struggled with the previous question, the administrator of the test has to skip any questions which build on it.

Group	Diagnosis	Frequency	Total
1	Alzheimer's Disease	316	316
2	Vascular Dementia	129	
	Vascular Cognitive Impairment	38	
	Mixed Dementia	57	224
3	Cognitive Impairment no Dementia	15	
	Mild Cognitive Impairment	40	55
4	Depression	39	
	Anxiety	14	53
5	Parkinson's Disease	10	
	Dementia with Lewey Bodies	19	29
6	Alcoholism	16	
	Fronto-temporal Dementia	13	
	Stroke	25	
	Brain Tumour	4	
	Epilepsy	2	
	Head Injury	3	
	Cerebellar Ataxia	1	
	low B12	1	
	low education	1	
	post-encephalitis	1	
	post neurosurgery	1	
	Primary Progressive Aphasia	1	
	side effects of drugs	1	
	Subarachnoid Haemorrhage	1	71
7	Norm	40	40
	Total		778

Table 3.1: Absolute frequencies of diagnostic groups.

3.2 Analysis

The study is conducted in two phases. First, a pilot study is performed using 100 observations taken from the entire sample. The pilot study ensures that the data preparation procedure is robust and that the analysis methods are appropriate and integrate well with each other. Upon the successful completion of the pilot study, the entire data set is used to formulate and validate a predictive model.

The private nature of the data has repercussions on the data preservation policy for this project. After completion of the research, the digital as well as the paper form of the data will be returned to Llandough hospital where it can be made available to future research projects subject to ethical approval by the NHS research ethical committee.

The work flow of the project, depicted in Figure 3.1, is designed to make the process easily reproducible so that results can be re-evaluated quickly if one of the steps is modified. After the data has been prepared for analysis, it is run through a dimensionality reduction process in order to identify groups of questions best suited to be used as indicators in models. The semantic analysis of the sentence writing question is parallel to the rest of the work flow - its outcome is necessary but not critical to the outcome of the study. As additional potential predictor variables are identified in the semantic analysis step, the variable selection process is re-evaluated with the new input variables. The predictive modeling step critically depends on the outcome of the dimensionality reduction step as a simple model is more readily transferable into clinical practice.

3.3 Methods

A traditional approach to feature extraction is principal components analysis (PCA). PCA with a tetrachoric corellation coefficient has been previously

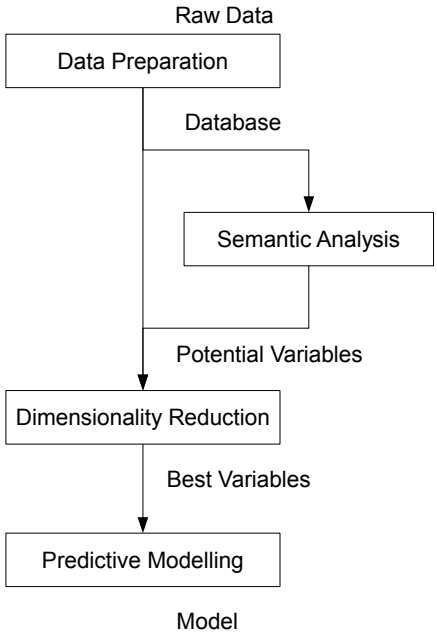


Figure 3.1: Work flow diagram.

applied to analyse the structure of the MMSE (Castro-Costa et al. 2009). Kernel PCA, homogeneity analysis, random orthogonal projections, principal component grid search, random rotation forests and independent component analysis (ICA) are evaluated as alternative methods and are compared with a novel method proposed in chapter 5.

In chapter 6 a novel entropic (see section for related work 2.4) variable selection method is proposed. 14 variations of the proposed method are compared with 36 state-of-the-art and classical variable selection methods are compared in the same chapter.

In chapter 8 the optimal subset of variables identified in chapter 6, a version of the data rotated using the method proposed in chapter 5 and the unmodified data are used to train a predictive model. The models which were considered are: C5.0 tree induction, random forests, K-nearest neighbours, naive Bayes and support vector machines with a radial basis function as a kernel. Random forests yield a good result which however is not easily interpretable due to the ensamble nature of the method. With the findings gained from the random

forest model, trees were induced from subsets of the data using the C5.0 tree induction algorithm in an effort to induce rules transferable into practice.

Using automated semantic annotation, the MMSE data is used to extract linguistic cues aiding in the screening for dementia. The annotation is performed using the method proposed by Klein and Manning (2003). Parts of speech (such as noun, verb etc.) and grammatical dependencies such as (subject to object) are considered as predictors.

The technologies, which are used in this project were chosen using the three guiding criteria:

1. When available, use software packages which are being maintained by a scientific community.

2. Integrate different software packages so that quality does not have to be sacrificed for compatibility.

3. Experiments should be easily reproducible so that after introduction of changes to the modelling process, the results can be easily re-evaluated.

PCA, kernel PCA, homogeneity analysis, random orthogonal projections and PCA grid search are carried out in the R[1] statistical package. The implementation used for random rotation forests is available in Weka[2] and is integrated with R using the *RWeka* package. Conditional mutual information estimation is implemented in C++ and utilises the CUDA[3] parallel processing framework. The CPU implementation used as a comparison benchmark is available in the R package (package *inftheo*) The CMIM algorithm (described in chapter 4) is integrated in the R environment. The stanford lexical parser [4] is used to annotate the MMSE data with semantic information. The annotation itself is automated with the Ruby scripting language [5] and the resulting

[1] http://www.r-project.org
[2] http://weka.sourceforge.net
[3] http://www.nvidia.com/object/cuda_home_new.html
[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] http://www.ruby-lang.org/en/

annotated data is imported in the R statistical environment. The prediction methods C5.0, random forests, support vector machines, naive Bayes and K-nearest neighbours are used in implementations available with the R software package.

3.4 Reference Data

To demonstrate the generalisability of the methods proposed in the chapters 5 and 6, use was made of eight different data sets from various domains. All data is publically available from the UCI machine learning repository [6]. Table 3.2 lists characteristic properties of the different data. In the table, the dimension count includes the class label.

	dimensions	observations	classes
Zoo	17	101	7
SPECT	23	266	2
Dermatology	35	366	6
Soybean	36	307	19
Sonar	61	208	2
KR vs KP	37	3196	2
Mushroom	23	8124	2
Splice	61	3190	3
Simulated	2000	61	2

Table 3.2: Summary of the data used in the evaluation.

Zoo

The zoo dataset is made up of 16 variables, each describing an animal attribute. The class variable is the animal type. The classification goal is to identify an animal type (e.g. “gorilla”) using the attributes (e.g. “airborne”).

SPECT Image Analysis

SPECT imaging is a tool for diagnosing myocardial perfusion. The original data (Kurgan et al. 2001) was collected as two sets of three-dimensional im-

[6] http://archive.ics.uci.edu/ml/

ages. The data used in this evaluation only contains 22 variables which were extracted from the original images.

Dermatology

The dermatology data (Güvenir, Demiröz and Ilter 1998) was collected to identify rules for differential diagnosis of erythemato-squamous diseases. The difficulty in analysing the data is that many of the diseases share histopathological features and that a disease may have its characteristic features only in late stages.

King Rook VS King Pawn

The King Rook vs King Pawn (KRvsKP) data (Bain and Muggleton 1995) is an analysis of the chess endgame with these figures. This setup is of fundamental importance in the game as many other endgames can be reduced to this one. Different board configurations are described with 37 binary attributes and the class label denotes whether white can win.

Mushroom

The mushroom data (Duch et al. 1997) is used to deduce rules for distinguishing between poisonous and edible mushrooms. The variables are various characteristics of the plant. Although there are no simple rules to distinguish between the two classes, a 99.9% correct classification was reported using 4 out of 22 variables (Schlimmer 1987).

Soybean Disease Diagnosis

The soybean data (Tan and Eshelman 1988) is used to diagnose diseases typical for the plant. Several attributes were collected from the sampled plant itself as well as its enviroment. Each plant is labelled with a diagnosis and the goal is to deduce the diagnosis from the collected attributes.

Splice Data

This data (Noordewier, Towell and Shavlik 1990) was collected to analyse splice-junction gene sequences. When proteins are formed from DNA strands, splice junctions are the points at which DNA is removed. The challenge in this data is to determine the boundaries between DNA which is retained and DNA which is removed after splicing. The data, which consists of DNA subsequences, is labelled in three classes: a boundary between retained and removed DNA (EI), a boundary between removed and retained DNA (IE), as well as spurious sequences (N).

Sonar Data

The sonar data has 61 variables. The first 60 are reflected energy in different frequency bands and last variable is the class label – rock or metal cylinder. The version used in the comparison is a discrete version with 2 values for each variable.

Simulated Data

In the analysis described in the following chapters, many of the compared methods perform similarly on the reference data chosen in this project. To demonstrate the difference of methods, data was generated which has a large number of observations but also has a structure which makes it challenging for methods which do not take redundancy into account.

The simulated data has 60 variables and a response variable which contains a class label. The 60 variables are structured in 10 groups. The variables within each group are correlated with a correlation coefficient of 0.85 but are independent from any of the variables outside the group. For example $\rho(X_i, X_j) = 0.85$ for $1 \leq i, j \leq 6$ and $\rho(X_k, X_l) = 0$ for $1 \leq k < 6$ and $6 < l \leq 60$. Each of the variables has a domain of 5 values. The class label Y is a function of 10 of the

variables with one variable from each group of correlated variables:

$$Y = \begin{cases} 0 & \text{if } \sum_i X_i < 25 \\ 1 & \text{otherwise,} \end{cases} \tag{3.1}$$

where i goes over the variables characteristic for each group.

Chapter 4

Parallel Computation Method for Estimating Conditional Mutual Information

In chapter 2 the problem of dimensionality reduction of discrete multivariate data is discussed and entropic methods are identified as the most suitable approach. The main issue with this family of methods is that reduction of redundancy in the model is limited by computational feasibility in addition to availability of data.

In this chapter, an algorithm is proposed which leverages modern graphics card hardware to push the boundaries of computational feasibility of conditional mutual information. By making the process of conditional mutual information faster, the severity of the trade-offs made can be reduced. The algorithm can process data with discrete variables, rather than only data with binary variables. Finally, the algorithm can estimate conditional mutual information with more than one conditional variable. This makes the availability of data rather than computational infeasibility the main challenge in practical applications.

4.1 Problem Definition

The computational complexity of estimating conditional mutual information is dominated by the estimation of point-wise probabilties. Conditional mutual information is defined in terms of entropy. Entropy, $H(X)$, is defined as the sum of l terms where l is the number of values X can take – in other words, the length of the alphabet U of X. Each of the l terms is an operation on

the occurance probability of a value $u \in U$. This probabilty is estimated from data by counting the number of observations taking the value u dividing by the total number of observations. The question that needs to be answered is how many duplicates of the value u were encountered in the series of observations that make up the data. A method which increases the speed of duplicate counting,can be applied directly to speed up the estimation of conditional mutual information.

In section 2.4.2 it is argued that estimating CMI for variable selection with $k-1$ conditional variables, where k is the number of selected variables, is infeasible for two reasons: first, the amount of required data unrealistic for large k, and secondly, estimating conditional mutual information $\binom{p}{k}$ times is not computationally feasible. A design requirement of the proposed algorithm is that the algorithm completes computation in reasonable time for a k, which is limited not by computation speed but by the amount of available data.

4.2 Parallel Computing Methods

To solve the problem defined in the previous section, it is crucial that the proposed algorithm counts duplicates efficiently with respect to time. Serial algorithms for duplicate counting typically operate in two steps: first the data are sorted, then, in a second phase, the number of consecutive repetitions of each value in the sorted sequence is counted. Although the process can be formulated in a similar way for parallel computation, the building blocks of the algorithm are different from their sequential counterparts.

Counting duplicates is done most efficiently by sorting the entire sequence and then iterating through it while incrementing a counter, which is reset every time the successor of an element is not equal to the current element. This process has complexity $O(n + nlog(n))$ and its performance is limited by the efficiency of the sorting algorithm. In related literature, several sorting algorithms have been proposed which, by using the parallel hardware on graphics

adapters, are significantly faster than sequential sorting algorithms running on a CPU (Satish, Harris and Garland 2009).

The sorting algorithm used in this research (Merrill and Grimshaw 2011). is reported to reach up to a tenfold increase in speed when comparing a readily available GPU to a 32-core CPU.

To count duplicates in a sorted sequence, the method proposed in this research employs a parallel prefix sum (scan) algorithm (Reif 1993). A scan is defined as a procedure which takes a binary associative operator $\bigoplus$ and an ordered list of n elements

$$\{a_0, a_1, ..., a_n\} \tag{4.1}$$

as input and returns the ordered list

$$\{a_0, (a_0 \bigoplus a_1), ..., (a_0 \bigoplus a_1 \bigoplus ... \bigoplus a_n)\} \tag{4.2}$$

For example, if the operator is addition, and the array is $1, 4, 2, 3, 5$, the result of the procedure would be $1, 5, 7, 10, 15$.

The sorting algorithm (Merrill and Grimshaw 2011) used in the proposed method relies on efficient scan algorithms to achieve its high performance.

While the parallel sorting algorithm is the source of the largest performance gain in the proposed algorithm over equivalent sequential algorithms, additional steps in the process of counting duplicates have been parallelised to further reduce the computation time. The most notable extension of the proposed algorithm allows it to run most efficiently when the conditional mutual information $I(Y; X_i|X_j)$ is estimated for thousands of $< Y, X_i, X_j >$ triplets at the same time.

4.3 Data Model

Two crucial data structures are used in the proposed parallel duplicate counting algorithm: a *sort buffer* and an *index buffer*. The sort buffer is used to store a sorted array which is an intermediate result for counting duplicates. The parallel nature of the algorithm allows counting the duplicates in more than one array at the same time. In this case index buffer is used to map each element to its corresponding array.

Parallel sorting algorithms operate most efficiently on 32 bit integers. The method would remain unchanged if the data model is extended to 64 bits, which may be required in case compromising computation time for expressiveness is acceptable. The decision to represent the data data in terms of 32 bit integers is based on two reasons: first, the inadequate tool support for 64 bit GPU enabled software across platforms would severely limit the applicability of the method; and secondly, sorting 32 bit sequences is currently faster when compared to sorting 64 bit sequences of the same length. Making the algorithm available to a large audience requires the implementation to work on a Windows operating system, depending only on freely available software. However, on a Windows operating system, the NVidia GPU compiler supports 64 bit applications only for the commercially, but not the freely available version of the Visual Studio development environment.

The algorithm stores the entire data in form of an array in the RAM of the graphics card. Each row of the array represents one row of the original data. To ensure efficient processing, the rows are padded in order to correspond to the number of elements which the hardware can process in parallel. For example, a GeForce 580GTX graphics card can process 512 elements concurrently. If the data has 60 dimensions then the first 60 elements of the data representation in the RAM contain the cells of the first row. The second row is stored at offset 512. When evaluating the goal function for a set of dimensions $X_1, ..., X_k, Y$, only those dimensions are copied into a separate buffer array for

1. reduce the row to the variables of interest e.g. $Y, X_1, ..., X_p \rightarrow Y, X_1, X_2$
2. represent the values of the reduced variables in base 2 e.g. $2 \rightarrow 10$
3. concatenate the base 2 digits starting from right to left e.g. $1, 2, 0 \rightarrow 011000$ – in this example each of the variable occupies 2 bits.
4. if Y is one of the variables of interest, prepend the digit 1 to the thus obtained number e.g. $011000 \rightarrow 1011000$. If Y is not one of the variables of interest, the digit 0 is prepended.

Figure 4.1: Data projection algorithm.

further processing.

Each observation is projected from $k + 2$ dimensions onto a single, 32 bit wide, whole-numbered dimension. This projection is only applicable under the assumption that all the dimensions being evaluated are narrow enough to fit into 31 bits (the remaining bit is used for a different purpose). While this restriction may seem severe, it is justifiable: first, the wider the range of a dimension, the larger the training data needs to be to achieve an accurate probability estimation; secondly, if more than 4 to 5 dimensions are analysed at the same time, the problem of sparse data arises again. The data is projected by using the following algorithm listed in figure 4.1.

Table 4.1 shows an illustrative example of the projection algorithm on data. To ensure optimal use of the 32 bits, each column is allocated with the minimal number of bits required to encode all values as radix 2 numbers. For example, if a variable X_1 can take the values $1, 2, 3, 4, 5$, the minimal number of bits required to represent X_1 is $\lceil log_2(5) \rceil = 3$. In this example, the value 1 of X_1 would be encoded as 001.

The algorithm processes variables of different width by using the minimal required width for each variable. For example if Y requires 2 bits, X_1 requires 4 bits, X_2 requires 5 bits and X_3 requires 1 bit, the algorithm would use 12 out of 32 bits ($2 + 4 + 5$ and one additional bit for distinguishing between Y, X_1, X_2 and X_1, X_2). In theory, if Y, X_1, and X_3 are considered, only 8 bits

Y	X1	X2	X3	Y,X1,X2	X1,X2
0	0	2	2	1000010	0000010
2	2	2	1	1101010	0001010
0	0	1	2	1000001	0000001
2	1	0	0	1100100	0000100
0	2	1	2	1001001	0001001
1	2	0	2	1011000	0001000
1	1	2	0	1010110	0000110
2	2	2	0	1101010	0001010
2	1	2	0	1100110	0000110
2	1	2	1	1100110	0000110

Table 4.1: Example data and projections.

are required. However, due to the parallel nature of the algorithm, the cases Y, X_1, X_2 and Y, X_1, X_3 may be considered in the same iteration. This limits the minimal number of bits required to the sum of the bits of the widest $k+1$ variables, where k is the number of conditional variables, plus the width of the class label. In this example, the limit would be the combined width of X_2, X_1 and Y.

The last step of the algorithm – setting the leftmost bit to 1 if Y is one of the variables – ensures that any value which includes Y in its projection is always larger than any value which does not include Y. The advantage of this property is explained explained by observing that a conditional probability can be estimated as a result of two non-conditional probabilities:

$$P(Y|X_1, X_2) = \frac{P(Y, X_1, X_2)}{P(X_1, X_2)} \tag{4.3}$$

Consider a situation where the projections of the tuples $< Y, X_1, X_2 >$ and $< X_1, X_2 >$ are in one array, allowing that after the sorting procedure each value can be mapped to one of the two original sets. Only one duplicate counting iteration is necessary to estimate a value for the probability. The data in table 4.1 represented as described would be an array $A(X_1|X_2)$ with

the following values:

$$[\underbrace{0000010, ..., 0000110}_{X_1,X_2}, \underbrace{1000010, ..., 1100110}_{Y,X_1,X_2}] \tag{4.4}$$

This array is used to estimate $I(Y; X_1|X_2)$ from equation 2.4. If this estimation were to be calculated on a state-of-the-art graphics processing unit with several hundred processing cores, many of the processing cores would idle while a small number of cores processed the 10 rows of data. Although this example is presented only for illustrative purposes, the issue of maximising the utilisation of processing power remains when working with real data. In order to reduce the idle time between iterations of the duplicate counting algorithm, the data representation needs to ensure that the amount of computation carried out in each step is limited by the hardware capabilities and not the algorithm itself.

High occupancy of the hardware is achieved by processing the duplicate counting step for several combinations of variables at the same time. For example, to assess the quality of X_1, the algorithm needs to estimate:

$$I(Y; X_1|X_2), I(Y; X_1|X_3), I(Y; X_1|X_4), ..., I(Y; X_1|X_p) \tag{4.5}$$

where p is the total number of variables. The duplicate counting algorithm then receives an array as input which is the concatenation of

$$A(X_1|X_2), A(X_1|X_3), A(X_1|X_4), ..., A(X_1|X_k) \tag{4.6}$$

where k is the index of the last array that can be fitted into the memory of the graphics card. The array thus produced is the *sort buffer* denoted as S.

As noted in section 4.2, duplicate counting is achieved by sorting the array S and counting subsequent repetitions of its elements. However, after S has been sorted, elements from different arrays A may be next to each other. In order to map individual array elements to a specific array A_i, each of the arrays

```
in: X ... p x q - 1 data vectors
    Y ... q element target vector
out: weights ... q element vector

 1  queue = {}, allweights = {}, result = {}
 2
 3  for each subset U of X of size k
 4    queue = queue + <Y, U>
 5    for each X_i in X / U
 6       queue = queue + <Y, X_i, U>
 7       if queue full
 8         process(allweights)
 9      end
10    end
11 end
12 return allweights
```

Figure 4.2: Search space traversal.

A is assigned an index i. A second array, M, is created such that the element $m_k \in M$, corresponding to the element $s_k \in S$, has the value i – the index of the projection A_i by means of which s_k was produced. In the example above, the element s_k for $k = 15$ is produced from the second projection in S, which is $A(X_1|X_3)$, and thus $m_k = 2$. The array M is the *index buffer*.

4.4 Control Flow

The parallel CMI algorirthm consists of two modules: a top-level module which directs the search space traversal and evaluates the results, a scheduling module executing iteration batches queued for parallel processing and a duplicate counting module making use of parallel hardware.

4.4.1 Search Space Traversal

Figure 4.2 lists the algorithm module which iterates through all combinations of variables of size k. The calculation of weights is deferred to the scheduling module listed in figure 4.3. The iteration over k-sized subsets is done in lexical

order using the LEXSUB algorithm (Nijenhuis and Will 1978). Each subset is queued for processing and, once the queue is full, the scheduling module is invoked.

The size of the queue is determined by the amount of physical memory available on the GPU. A rough estimate of the memory occupancy of the queue is the number of iterations in the queue multiplied by the size of the data times two (once for the sort buffer and once for the index buffer) and times 4 (4 bytes in an integer). For example, if a GPU has a total of 1024 megabytes of RAM and a given data-set has 2000 observations, the maximal queue size can be estimated in bytes as $1048576/(2000 * 4 * 2) = 65536$. It is advisable to choose a queue size at least 10-20% smaller than this number. For one, the computation itself requires some working memory. Furthermore, if the GPU is used for display purposes as well as computation a queue size which is approaching the physical amount of memory may render the system unstable. The algorithm is distributed with a conservative queue size per default and the option to set the queue size manually. Note that the version of the algorithm discussed in this thesis is built for 32 bit platforms and thus can only address up to 4Gbyte of working memory imposing a further constraint on the queue size.

4.4.2 Scheduling

The second module, listed in figure 4.3, mediates between the parallel and sequential components. The first task of this module is to invoke the duplicate counting module described in section 4.4.3. This module returns a data structure containing a set of tuples $< t, d >$ for each component array A of the *sort buffer*, where t is the projection of a data-point and d is its number of repetitions in A.

For example, the data point $X_1 = 1, X_2 = 2, X_3 = 0$ in table 4.1 would be represented as $< 6, 2 >$ – 6 being the integer equivalent of 0000110 and 2

```
 1 begin process(out allweights)
 2    H(Y|U), H(Y|X,U), I(Y;X|U)
 3    dall, dcond = count_duplicates(queue)
 4    for each array A_i  in queue
 5       if(X not in A_i)
 6          H(Y|U) = dall[A_i] / n * log(dall[A_i] / dcond[A_i])
 7       else
 8          H(Y|X,U) = dall[A_i] / n * log(dall[A_i] / dcond[A_i])
 9          I(Y;X|U) = H(Y|U) - H(Y|X,U)
10        allweights[A_i] =  I(Y;X|U)
11       end
12    end
13 end
```

Figure 4.3: Scheduler

being the number of occurances of the value 0000110 in the last column of the table.

This data structure is described as *dall, dcond* in the algorithm listing for brevity. If $< Y, X_i, X_{j_1}, ..., X_{j_k} >$ is a projection of a data-point, then *dall* represents the number of its repetitions and *dcond* represents the number of repetitions of the projection $< X_i, X_{j_1}, ..., X_{j_k} >$. If, for example, the queue contained the sets of variables $\{Y, X_1, X_2\}$ and $\{X_1, X_2\}$ at the time of invocation of the duplicate counter, then *dall* would be $\{< 1010110, 1 >, < 1100110, 1 >\}$ and *dcond* would contain $\{< 0000110, 2 >\}$. These tuples are needed to estimate the conditional entropy as:

$$H(Y|X_i, U) = \frac{dall}{n} log(\frac{dall}{dcond}), \tag{4.7}$$

where $U = \{X_{j_1}, ..., X_{j_k}\}$.

The structure of the top-level component ensures that an iteration which estimates $H(Y|X_{j_1}, ..., X_{j_k})$ is always followed by q iterations which estimate $H(Y|X_i, X_{j_1}, ..., X_{j_k})$ (one for each X_i). This assumption is used in lines 5 to 11 of the scheduling module. In these lines the algorithm estimates $H(Y|X_{j_1}, ..., X_{j_k})$ only once for all $H(Y|X_i, X_{j_1}, ..., X_{j_k})$ with $1 \leq i \leq q$. The

```
 1 begin count_duplicates for all p in queue
 2   sort_buffer = {}, index_buffer = {}, h_counts = {}
 3   begin in parallel
 4      for each p in queue
 5         A_i = project data along p
 6         sort_buffer append A_i
 7         index_buffer append n repetitions of i
 8      end
 9      stable sort index_buffer using sort_buffer as key
10      stable sort sort_buffer using index_buffer as key
11   end
12
13   for each A in sort_buffer
14      begin in parallel
15         stencil = { A[i-1] == A[i] ? 0 : 1 }
16         scanned_counts = { stencil[i] = 1 ? 0 :
17                                scanned_counts[i-1] + }
18         h_counts[tuple] = { stencil[i] = 1 ?
19                                tuple = A[i],
20                                count = scanned_counts[i] }
21      end
22   end
23   return h_counts
24 end
```

Figure 4.4: Duplicate counting.

algorithm evaluates

$$I(Y; X_i|X_{j_1}, ..., X_{j_k}) = H(Y|X_{j_1}, ..., X_{j_k}) - H(Y|X_i, X_{j_1}, ..., X_{j_k}) \qquad (4.8)$$

for all variable subsets of size k. The results are stored in *allweights* and returned to the top-level component.

4.4.3 Duplicate Counting

The core of the algorithm is the parallel duplicate counting module listed in figure 4.4. In the notation of this listing *begin in parallel* denotes the beginning of a section in which each statement is processed in parallel. The statements are invoked sequentially however. For example, lines 9 and 10 sort arrays, and

although the sorting is done in parallel, line 9 is guaranteed to complete before line 10 begins.

The module can be subdivided into three larger steps. In the first step, the contents of the arrays A_i are calculated by applying step four of the algorithm in figure 4.1 for each subset of variables in the queue. Following this, the *sort buffer* and *index buffer* are initialised.

In the second step, the arrays A_i are sorted so that duplicate elements are adjacent. The sorting takes place in two steps: first, the *index buffer* is sorted using the *sort buffer* as an index. Although this ensures that duplicates are adjacent, it does not ensure that all elements of A_i are placed before the elements of A_{i+1}. To ensure that elements from different arrays A are separate, the *sort buffer* is sorted, this time using the *index buffer* as a key. Using a stable sorting algorithm ensures that if an element a_i is placed before a_{i+j} in the *sort buffer* as ordered after the execution of line 9, then the same order persists after the execution of line 10 assuming both elements are in the same sequence A_i. After line 10, the *sort buffer* contains the sequences $A_1, ...A_n$ guaranteeing that the sequence A_i is adjacent to A_{i+1} and the elements of each sequence are sorted in ascending order.

In the third step of the duplicate counting algorithm, the number of repetitions of each element in the *sort buffer* is counted. The structure of this step is strongly governed by the architecture of the parallel hardware it is designed for. For this reason, an indepth discussion is not possible without a thorough review of the parallel architecture. To avoid such a review, which would go beyond the scope of the thesis, this part of the algorithm is discussed only briefly.

To count the duplicates, first a *stencil* is calculated. The stencil is an array of the same length as the *sort buffer*, which contains 0 at the index i if the element a_i equals a_{i+1} in the *sort buffer* and 1 if the two elements differ. The last element of the *stencil* is always 1. For example, if the *sort buffer* contains

the values $\{1, 1, 2, 3, 3\}$, the *stencil* would contain the values $\{0, 1, 1, 0, 1\}$.

The *scanned counts* array is calculated by incrementing a running variable and resetting it to 1 at the indices where the stencil contains a 1 at the following index. This is achieved by using a prefix sum (described in section 4.2) with the binary operator:

$$a_i \bigoplus a_{i+1} = \begin{cases} i \leftarrow i + 1 & \text{if } a_i \neq a_{i+1} \\ i \leftarrow 1 & \text{otherwise} \end{cases} \tag{4.9}$$

The *scanned counts* array for the values $\{1, 1, 2, 3, 3\}$ computed in this way would be $\{1, 2, 1, 1, 2\}$. Finally, the elements of the *scanned counts* buffer and the elements of the *sort buffer* coinciding with elements of the *stencil* that are equal to 1 are returned as results of the duplicate counting phase. In the above example, the second, third and last element of the *scanned counts* array are used in combination with the *sort buffer* to return a data structure of the form $\{< 1, 2 >, < 2, 1 >, < 3, 2 >\}$.

The counting step is parallel for each component A_i of the *sort buffer* but is invoked n times, where n is the number of components A of the *sort buffer*. To further improve the speed of the algorithm, this step can be redesigned to count duplicates in parallel for the whole sort buffer rather than just for one component at a time.

4.5 Comparison With Sequential Method

The GPU imlementation of the algorithm was compared against a CPU implementation of the same method using a state-of-the-art information theoretic estimator (Meyer 2008). To perform the comparison, the metric proposed by Martínez Sotoca and Pla (2010) was calculated on CPU and on GPU. The runtime for both approaches is shown in table 4.2.

Two conclusions can be deduced from this comparison. First, the speed of

	CPU	GPU
Zoo	0.37	0.95
SPECT	0.72	1.0
KR vs KP	7.4	4.13
Mushroom	4.2	3.26
Simulated	15.0	7.24

Table 4.2: Comparison of runtimes for CPU and GPU implementation in seconds with one conditional variable.

	CPU	GPU
Zoo	3.04	1.72
SPECT	12.05	5.14
KR vs KP	357.0	12.82
Mushroom	63.47	37.91
Simulated	2103.70	156.0

Table 4.3: Comparison of runtimes for CPU and GPU implementation in seconds with two conditional variables.

the GPU method is not a linear function of the number of samples. For small data of a few hundred observations such as Zoo and Spect, the CPU method is more appropriate. However, as the complexity of the data and the sample size grows, the GPU method can be up to 50% faster. While this does not have a large impact in the metric of this comparison, it yields large benefits when CMIM with more than one conditional variable is estimated.

The non-linear increase in speed of the GPU method becomes much more significant when probabilities with more than one conditional are estimated (see table 4.3). When $I(Y; X_i|X_j, X_k)$ is estimated on large data with many variables, such as KR vs KP or the simulated data, the speed gain can be more than tenfold.

4.6 Conclusion

In this chapter an algorithm is proposed which leverages parallel graphics hardware to explore the boundaries of computational feasibility in variable selection.

The merit of the PCMIM algorithm is two-fold: First, its efficiency makes it

feasible to experiment with using more than one conditional dimension. This is achieved by pushing back infeasibility caused by combinatorial explosion. Secondly, the proposed algorithm expands the application domains of the current state-of-the-art CMIM algorithm by relaxing the requirement for binary data to discrete data.

A limitation of the algorithm is that it is constrained to discrete data with a limited domain. As such, the algorithm is not suitable for continuous data or for application domains such as image recognition or bio-informatics where data typically has several thousand variables. The method is designed largely with applications in medical diagnostic support in mind, but, as demonstrated in chapters 5 and 6, it is equally well applicable to other domains with data that is discrete and that has a medium number of variables (up to 100). The method is particularly advantegeous for data with upwards of 1000 observations.

Additional performance improvement of the proposed parallel algorithm is achievable with a parallel reformulation of lines 13 to 22 in figure 4.4 such that more than one variable at a time is processed for every iteration of this phase of the algorithm. The proposed implementation of the algorithm assumes that the entire data fits in the RAM of the graphics card. To make the algorithm applicable for very large data, it would be beneficial to extend it so that partitions of a data can be analysed allowing the results to be recombined at a later stage. Lastly, for data with a large number of dimensions an exhaustive search through the problem space is not computationally feasible with currently available hardware.

Chapter 5

PCA for Discrete Data

In this chapter a method for supervised entropic PCA – mRRPCA – is proposed. In addition, 4 types of principal component analysis are compared for their applicability to discrete data as it is found in the Mini-Mental State Examination. The methods compared are: classical (linear) PCA, kernel PCA, homogeneity analysis and PCA using entropic covariance measures.

5.1 Entropic Covariance Measures

In linear PCA, use is made of the fact that the covariance and in turn the correlation matrix of a data is symmetrical and full rank (nonsingular). The latter assumption is not strict; if the correlation matrix has less than full rank, then decomposition methods exist which can extract less eigenvectors than the matrix has columns. To motivate the formulation of the mRRPCA method it is necessary to highlight the properties of an eigenvalue decomposition of a matrix which is square, symmetrical and nonsingular (Shifrin and Adams 2011):

Theorem 5.1. *Let A be a symmetric $n \times n$ matrix. Then*

1. *The eigenvalues of A are real.*

2. *There is an orthonormal basis $\{v_1, ..., v_k\}$ for $\mathbb{R}^n$ consisting of eigenvectors of A. That is, there is an orthogonal matrix Q so that $Q^{-1}AQ = \Lambda$ is diagonal.*

where an orthonormal basis of A is defined as:

Definition 5.1. *Let $v_1, ..., v_k \in \mathbb{R}^k$. We say $\{v_1, ..., v_k\}$ is an* orthognoal set *of vectors provided $v_i \dot{v}_j = 0$ whenever $i \neq j$. We say $\{v_1, ..., v_k\}$ is an orthogonal basis for a subspace V if $\{v_1, ..., v_k\}$ is a basis for V and an orthogonal*

set. Moreover, we say $\{v_1, ..., v_k\}$ *is an* orthonormal basis *for* V *if it is an orthogonal basis consisting of unit vectors.*

a basis of A is defined as:

Definition 5.2. *Let* $V \subset \mathbb{R}^n$ *be a subspace. The set of vectors* $\{v_1, ..., v_k\}$ *is called a* basis *for* V *if:*

1. $V = Span(v_1, ..., v_k)$ *and*
2. $\{v_1, ..., v_k\}$ *are linearly independent.*

and the span of $\{v_1, ..., v_k\}$ is defined as:

Definition 5.3. *Let* $v_1, ..., v_k \in \mathbb{R}^k$*. The set of all linear combinations* $\{v_1, ..., v_k\}$ *is called their* span*, denoted* $Span(v_1, ..., v_k)$*. That is,*

$$Span(v_1, ..., v_k) = \{v \in \mathbb{R}^k = c_1v_1 + ... + c_kv_k | c_i \in \mathbb{R}^1\} \tag{5.1}$$

In other words, because a covariance matrix is square, symmetrical and usually nonsingular, a corollary from definition 5.1 is that it can be decomposed in the form of:

$$AQ = \Lambda \tag{5.2}$$

where A is the covariance matrix, Q is an *orthonormal basis* of A (see definition 5.1) and Λ is a diagonal matrix of scaling coefficients. Or less formally, finding an eigendecomposition of A is the same as finding directions in the data which maximise variance but are at the same time orthogonal to each other (for a more formal treatise refer to Jolliffe (2002)).

If the covariance matrix is replaced with a different matrix which is square, symmetrical, nonsingular and describes the data in some measure, its eigenvalue decomposition will still be an *orthonormal basis* of the new matrix. Classically, use is made of this fact by replacing the covariance matrix with the

correlation matrix and more recently with a matrix estimating covariance in a higher dimensional projection space as described in section 2.3. The latter is more commonly known as kernel PCA.

A less commonly accepted idea is to estimate the covariance between columns in the data using information theory (see section 2.3.4 for a review of related work). The adoption of such non-linear covariance measures has been slow because the classical method typically performs well and most non-linear methods add significant computational load. This fact is investigated further in the following section 5.2.

The availability of the algorithm for estimating conditional mutual information proposed in chapter 4 alleviates the issue of computational feasibility thus allowing the exploration of new covariance estimates. The covariance measure considered in this chapter has been proposed by Martínez Sotoca and Pla (2010) for variable selection and is calculated as:

$$I(X_i;Y|X_j) + I(X_j;Y|X_i) = 2I(X_i,X_j;Y) - I(X_i;Y) - I(X_j;Y) \tag{5.3}$$

which is rewritten as:

$$H(Y|X_i) + H(Y|X_j) - H(Y|X_i,X_j) \tag{5.4}$$

In spite of the efficient form, the estimation of this function on CPU becomes infeasible for data with more than 1000 observations. With the method described in chapter 4 the function can be estimated within seconds for large data. More direct evidence for the speed comparison of Sotoca's method implemented on CPU and GPU is presented in chapter 6. In the following, the metric proposed by Sotoca is considered as an alternative covariance metric for finding principal components in the data. The function is estimated without

reformulation in the more intuitively interpretable form:

$$a_{ij} = I(Y; X_i|X_j) + I(Y; X_j|X_i) \tag{5.5}$$

where a_{ij} is the cell at the ith column and the jth row of the covariance matrix estimate. For $i = j$ the cell contains an estimate of $2I(Y; X_i)$.

An advantage of estimating the function without reformulating it is that each rewriting of the function makes impractical assumptions about the structure of the data. The strongest evidence in support of this claim is the structure of the function itself. In information theory $I(Y; X_i|X_j)$ is equal to $I(Y; X_j|X_i)$. However when X_i and X_j are not linearly independent, the Bayes theorem for rewriting probabilities does not apply. When investigating principal components, assumption of linear independency goes against all reason as the purpose of PCA is to identify redundancy (linear dependency) in the data. When estimated on real data the equality $I(Y; X_j|X_i) = I(Y; X_i|X_j)$ does not necesserily hold, which is the reason why both directions are taken into account when estimating the covariance of the variables X_i and X_j.

A final adjustment which needs to be made to the goal function for it to be an adequate replacement for variance is to formulate it as a minimisation problem. As elaborated in section 2.2, conditional mutual information (CMIM) is the amount of uncertainty reduced about Y by knowing X_i and X_j. It is desirable that this measure be high so that a large amount of uncertainty is reduced. However in a covariance matrix a large value is interpreted as high correlation of X_i and X_j and is undesirable when the goal is to minimise redundancy. The function is reformulated into a minimisation problem by subtracting all values in the matrix A from the maximum value in a cell of A:

$$A = max(A) - A \tag{5.6}$$

This method of constructing an entropic covariance matrix and analysing

its eigenvalue decomposition is denoted as mRRPCA (maximum Relevance, minimum Redundancy Principal Component Analysis).

Since the metric cannot be feasibly compared for speed against an equivalent CPU implementation, a simpler version of the metric is formulated instead:

$$a_{ij} = I(X_i; X_j) + I(X_j; X_i). \tag{5.7}$$

The main difference to the previous method is that this simpler version is not supervised – it only considers the explanatory variables in the data to uncover structure. Along the diagonal a_{ii} the matrix contains a multiple of the estimate of the entropy of the ith column $2H(X_i)$. This latter method is also not reported in related literature. It is denoted as MIPCA (Mutual Information Principal Component Analysis) in the rest of the thesis.

5.2 Comparison of PCA Methods for Discrete Data

In this section 9 PCA methods are compared for their performance on discrete data. The methods are compared for amount of variation explained by the first few components and computation speed. The variation explained by a subset of the components is compared in terms of relative magnitude of component coefficients and by class separability after rotating the original data along the directions of maximal variance. Class separability is compared by splitting the rotated data in a training and test set and evaluating the accuracy of a naive Bayes classifier. The stability of the accuracy is gauged with tenfold cross-validation.

All methods were compared on 5 of the data described in section 3.4: Zoo, SPECT, KR vs KP, Mushroom as well as the simulated data. Further, the methods are compared on the MMSE data with semantic annotations as described in chapter 7 reduced to only patients diagnosed with either Alzheimer's Disease or Vascular Dementia (mmsesub). The Zoo data was chosen for its

relatively small size and simplicity. The remaining 4 data were chosen because they group the observations in two classes. When plotted onto two-dimensional space, the separability of two classes can easily be inspected visually. To avoid undue verbosity, the separability of the data is presented visually only in cases where it yields insight into the comparison of two of the methods.

The methods which are compared with the two methods proposed in the previous section are described in section 2.3. Those methods are: the classical method (*PCA*), kernel PCA with an RBF kernel (*KPCA RBF*), homogeneity analysis (*HOMALS*), the SMIFE2 method (*SMIFE2*), independent components analysis in the fastICA variation (*fastICA*) (Hyvrinen, Karhunen and Oja 2001) as well as two projection methods which look for projections algorithmically rather than analytically: *PCAgrid* (Croux, Filzmoser and Oliveira 2007) and *RPorth* (Varmuza, Filzmoser and Liebmann 2010). Those methods are compared against the supervised method (*mRRPCA*) and the unsupervised method (*MIPCA*) from the previous section. In the list of methods there are two notable omissions: tetrachoric PCA and SMIFE1. Both of these methods were considered for inclusion however the tetrachoric (or polychoric in the case of multinomial distributions) method failed to converge to a solution for any of the data and the SMIFE1 method gave results which were worse and slower than SMIFE2.

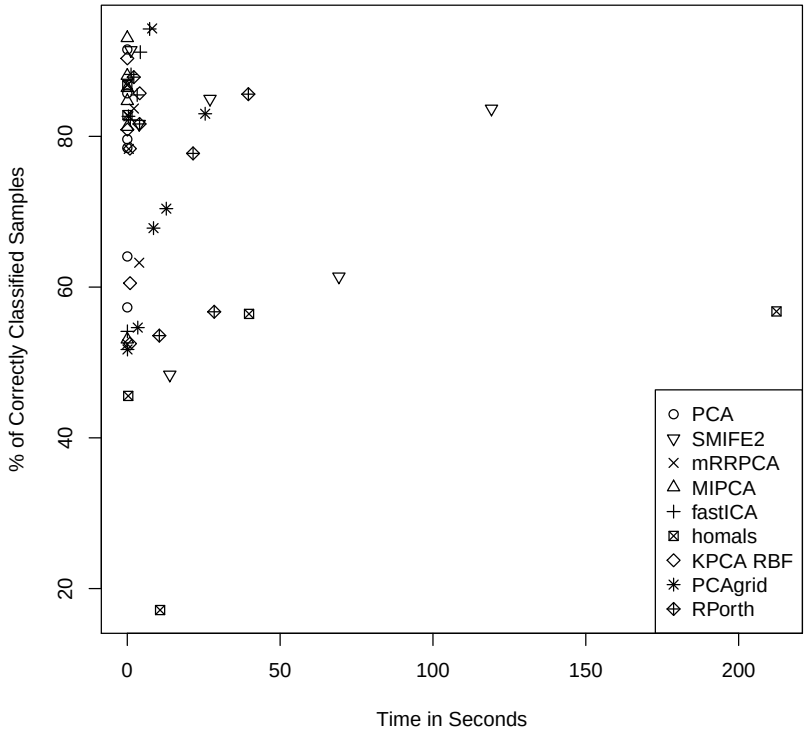


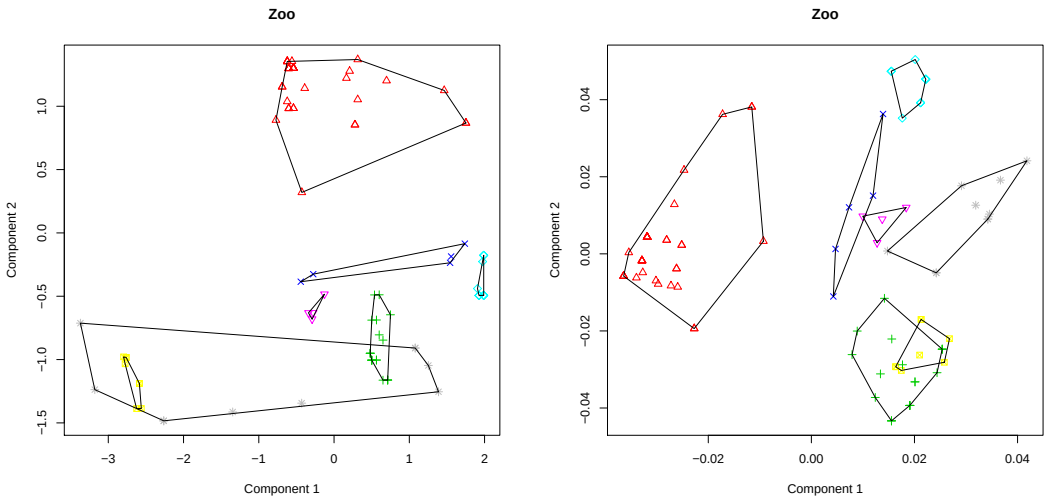
Figure 5.1: Classification rate versus running time for different PCA methods.

	PCA	SMIFE2	mRRPCA	MIPCA	fastICA	homals	KPCA RBF	PCAgrid	RPorth
krvskp	0.64	0.88	0.91	0.63	0.61	0.61	0.56	0.70	0.57
mmsesub	0.57	0.53	0.54	0.52	0.52	0.48	0.46	0.55	0.54
mushroom	0.78	0.85	0.85	0.84	0.78	0.85	0.57	0.68	0.78
simulatedNumeric	0.92	0.93	0.94	0.94	0.86	0.84	0.17	0.83	0.86
SPECT	0.80	0.81	0.82	0.78	0.81	0.82	0.83	0.83	0.82
zoo	0.86	0.86	0.88	0.87	0.90	0.91	0.87	0.52	0.88

Table 5.1: Comparison of classification accuracy for different PCA methods.

	PCA	SMIFE2	mRRPCA	MIPCA	fastICA	homals	kpca RBF	PCAgrid	RPorth
krvskp	0.01	0.01	4.26	3.90	0.91	69.23	39.83	12.76	28.44
mmsesub	0.01	0.01	0.04	0.01	0.89	13.91	0.31	3.43	10.48
mushroom	0.01	0.01	3.24	2.16	0.86	27.06	212.38	8.58	21.52
simulatedNumeric	0.01	0.01	7.31	8.13	4.06	119.15	10.73	25.48	39.61
SPECT	0.01	0.01	1.00	0.50	0.05	3.88	0.05	0.38	4.01
zoo	0.01	0.01	1.23	0.22	0.02	1.17	0.01	0.06	2.13

Table 5.2: Comparison of running time in seconds for different PCA methods.



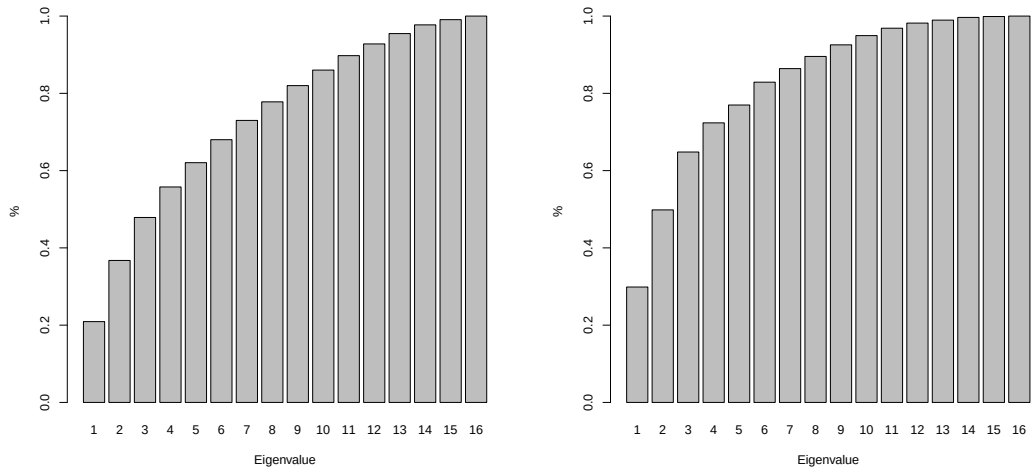
(a) PCA (b) HOMALS

Figure 5.2: Class separation along first 2 PCs for a) PCA and b) HOMALS methods on Zoo data.

The classification accuracy was tested on data which was reduced to three dimensions projected along the first 3 principal components. Therefore the classification accuracy can be viewed as a gauge of how well the different projection methods eliminate redundancy in the data. The accuracy results are displayed in table 5.1 with the best result for each data printed in bold face. The methods each exhibit a difference in time-efficiency for data processing, with kernel PCA methods being particularly computationally intensive. An overview of the durations for each experiment is presented in table 5.2.

The data listed in table 5.1 and 5.2 is shown as a graph in figure 5.1 in which the duration of each experiment is plotted in relation to its outcome. For example, the point in the far right of the graph shows that while the HOMALS method has given a competitive result for this experiment, it took disproportionately long to calculate the solution. Kernel PCA with an RBF kernel on the other hand finishes its calculations in more reasonable time frames but it also has yielded some of the worst results.

The spirit of the comparisons presented in the following is demonstrated with the example of the result produced by classical PCA projection and by homogeneity analysis. First, visual separation of the different classes in the



(a) PCA (b) HOMALS

Figure 5.3: Cumulative explained variance for a) PCA and b) HOMALS methods on Zoo data.

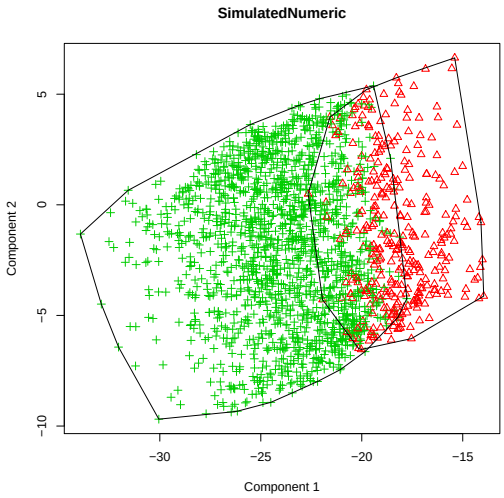
data is considered (see figures 5.2a and 5.2b). The plot is generated by plotting the data against the first two principal components computed by the corresponding PCA method and by adding a convex hull around all classes. The convex hull should not be interpreted as the boundary of the class, it serves merely as a visual aid in gauging the overlap of classes after the data is projected. For example, the triangle and diamond class in the Zoo data are clearly well separated when linear PCA is used as the projection method (figure 5.2a). If the convex hull is used as a gauge, one can count that when linear PCA is used, 8 data points cannot be classified unambiguously while when homogeneity analysis is used only 6 points lie in more than one convex hull.

In this case, the hull plot does not go very far in clearing up the question of which method gives a more adequate insight into the data. To further investigate the difference between the two methods, the relative magnitude of the component coefficients can be compared. The magnitude of the component coefficient is representative for the amount of variation explained by its corresponding component. The coefficients cannot be compared between data or between methods without normalisation. The normalisation chosen for this

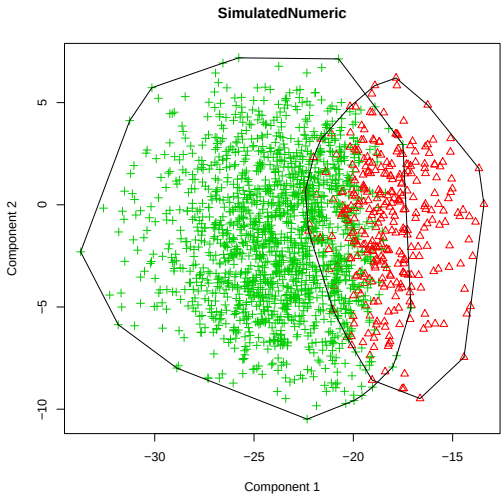
comparison is to assume that when all components are taken into account, all variation in the data is explained. One can then easily calculate what percentage of variation each component explains by dividing by the sum of all coefficients. This correspondence is visualised in a cumulative form. For example, the 3rd bar in figure 5.3a can be traced to 0.5, this is interpreted as the first 3 components, explaining 50% of the variance in the data. Using this visualisation method and comparing figure 5.3a with figure 5.3b, it can be observed that the first 3 components of the HOMALS method explain over 60% in the data while the first 3 components in the of the PCA projection only explain around 50% of the data. These findings suggest that while at first glance the PCA method projects the data in a direction in which the classes can be separated well by the classifier, thus the HOMALS method is superior in explaining the structure of the data.

The results for the SPECT data are similar for all methods suggesting that the covariance structure in the data is not very complex. An experimental outcome which provides strong evidence in favor of the mRRPCA method are the results for the KR vs KP data. While the performance for all other methods breaks down in this data, mRRPCA yields a high accuracy in a fraction of the time that any of the other non-linear PCA methods takes. On the Mushroom data, the best result is produced by the HOMALS method. Regardless of the accuracy of the HOMALS method, the MIPCA and mRRPCA methods may be preferable under certain circumstances for two reasons: first, computation is significantly faster and more importantly, the HOMALS method does not give a projection which can be reproduced with unseen data while MIPCA and mRRPCA methods return a rotation matrix which can be used to model or explain new data. The mRRPCA method is likely preferable over the MIPCA method as it yields the same accuracy in less than half the time.

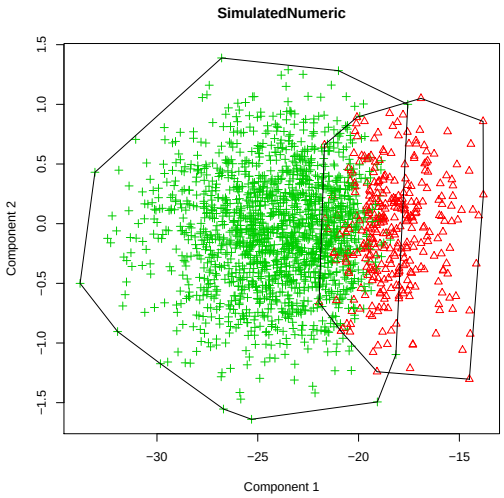
The most informative result is the classification accuracy on the synthetic data. The highest accuracy is achieved by the entropic methods, again provid-



(a) MIPCA

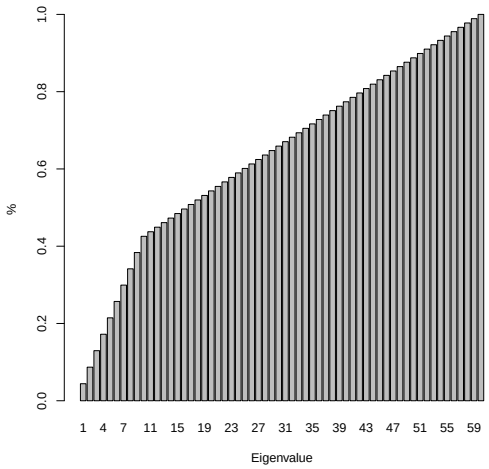


(b) mRRPCA

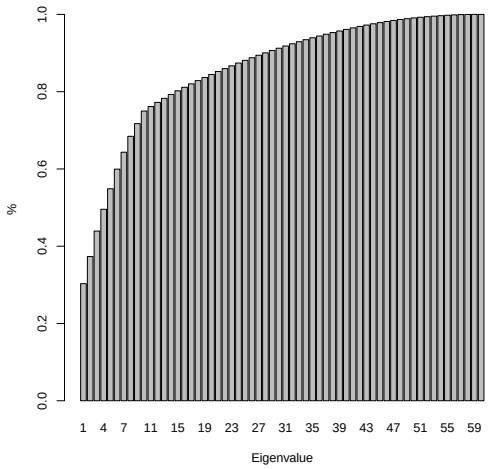


(c) SMIFE2

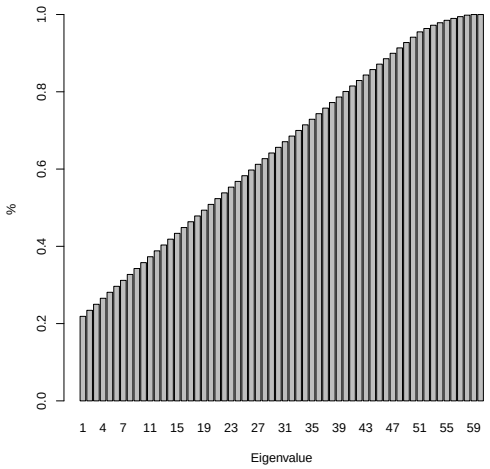
Figure 5.4: Class separation along the first 2 PCs with the a) MIPCA, b) mRRPCA and c) SMIFE2 methods on simulated data.



(a) MIPCA



(b) mRRPCA



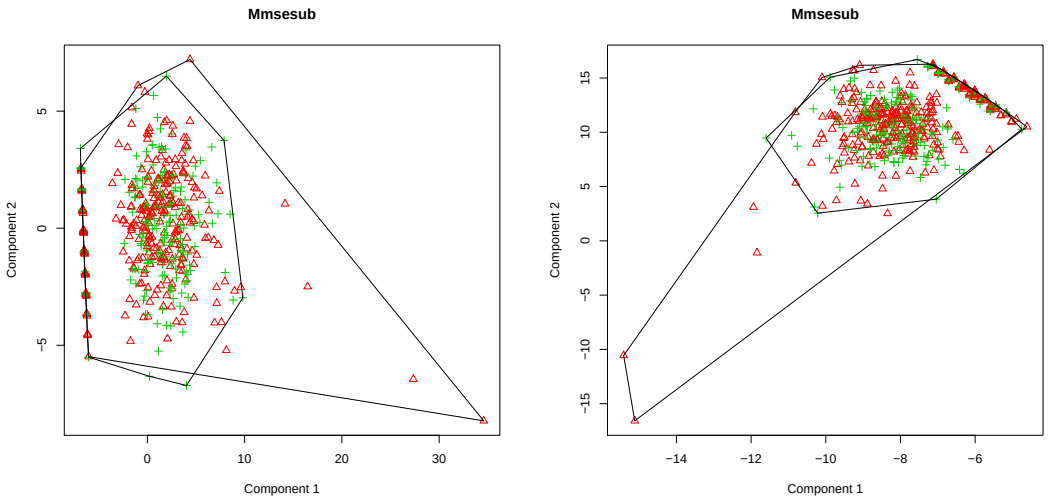
(c) SMIFE2

Figure 5.5: Cumulative explained variance for the a) MIPCA, b) mRRPCA and c) SMIFE2 methods on simulated data.

ing evidence in support of considering entropic PCA for discrete data. When comparing class separability along the first 2 principal components for MIPCA, mRRPCA and SMIFE2 (see figure 6.1) little can be said about how the three methods compare beyond what is already known from their classification accuracy. The most convincing gauge in support of mRRPCA is the plot comparing the cumulative explained variance for the three methods. The SMIFE2 method fails to uncover any of the structure used to generate the data. The MIPCA and mRRPCA methods uncover the structure of 10 highly correlated components. This result is more pronounced with mRRPCA where the first 10 components (corresponding to the 10 components used to generate the data) explain almost 80% of the variance in the rotated data as opposed to just over 40% when using MIPCA.

	PCA	SMIFE2	mRRPCA	MIPCA	fastICA	homals	kpca RBF	PCAgrid	RPorth	RF
krvskp	0.98	0.99	0.99	0.99	0.98	0.98	0.97	0.98	0.99	0.99
mmsesub	0.61	0.62	0.67	0.66	0.65	0.68	0.66	0.66	0.71	0.66
mushroom	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
simulatedNumeric	0.90	0.88	0.87	0.90	0.91	0.89	0.91	0.88	0.89	0.90
SPECT	0.86	0.82	0.82	0.81	0.84	0.81	0.81	0.87	0.76	0.80
zoo	0.97	1.00	0.93	0.93	0.93	0.93	0.97	0.93	0.93	0.97

Table 5.3: Comparision of accuracy of rotation forests and pca with random forest combinations.

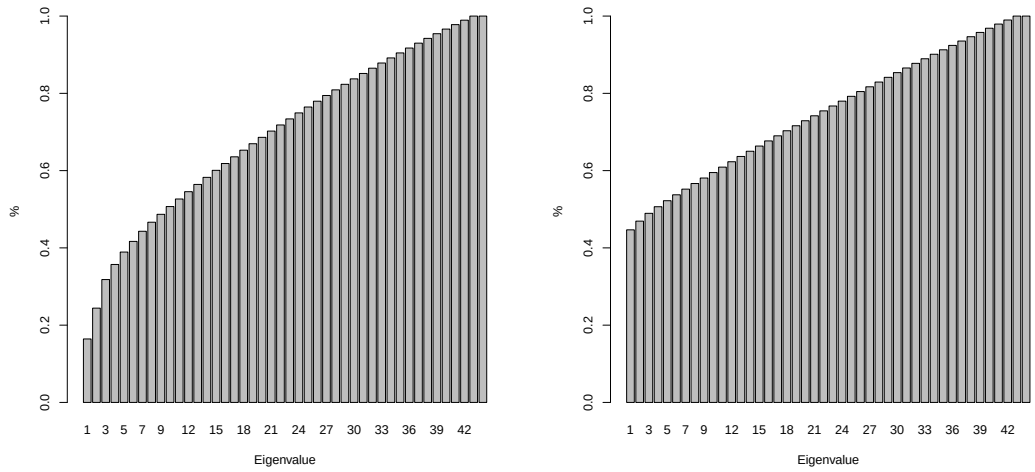


(a) PCA (b) mRRPCA

Figure 5.6: Class separation along first 2 PCs for a) PCA and b) mRRPCA methods on mmsesub data.

While table 5.1 yields insights into the performance of different PCA approaches on discrete data it is difficult to neglect the poor performance on the MMSE data. One possible conclusion of this result is that the MMSE data cannot be modelled with a linear projection model. This supposition is supported by figures 5.6a, 5.6b, 5.7a and 5.7b. While figures 5.7a and 5.7b show that the mRRPCA method explains more variance in fewer principal components, figures 5.6a and 5.6b show that the projection does not actually improve classification accuracy visibly. For this reason, a second comparison of methods was undertaken using random forest induction (Breiman 2001) as the classifier. Random forests can be extended to rotation forests (Rodriguez, Kuncheva and Alonso 2006) by projecting the subset of rows and columns for each subtree before inducing the tree. The drawback of this rotation method is that it gives no information about the correlation structure of the data or redundant variables in the data.

Accuracy of rotation forests and PCA with random forests as a classifier are presented in tables 5.3 . In this experiment setup the cross validation is not repeated as the random forest classifier is already an ensemble learning method which averages error rates. The runtime for the rotation forest method



(a) PCA (b) mRRPCA

Figure 5.7: Cumulative explained variance for a) PCA and b) mRRPCA methods on mmsesub data.

is less than 2 seconds on any of the data. The runtime for the combinations of projection methods with random forests are bound by the duration of the PCA computations in table 5.2.

Overall random forests perform better on the data chosen for benchmarking than naive Bayes classification. However, the added improvement seems to come from the change of the classification method rather than from rotating each sub-tree individually. Rotating the entire data with classical PCA first and then classifying performs worse only for the MMSE data. An interesting result is the 71% classification accuracy achieved with random orthogonal projections (*RPorth*) in combination with random forests. Unfortunately this result could not be reproduced in subsequent runs due to the non-deterministic nature of the rotation algorithm.

5.3 Conclusion

In this chapter two methods for principal components analysis were proposed, MIPCA and mRRPCA. Evidence is provided using known data and ensuring stability with cross-validation which encourages the use of the proposed

methods for classification problems with discrete data. A limitation of the mRRPCA is that in its proposed form it requires graphics hardware capable of parallel computation. When such hardware is not available, the MIPCA method is a viable alternative which gives results competitive with the SMIFE2 method in a fraction of the time it takes to calculate the SMIFE2 matrix.

Chapter 6

Variable Selection with Entropic Criteria

In this chapter two novel methods are proposed: a goal function for forward search in variable selection, which is based on the three-way interaction gain method proposed by Akadi, Ouardighi and Aboutajdine (2008), and a parallel algorithm for conditional mutual information maximisation based variable selection which allows pushing back the boundaries of computational feasibility for information theoretic feature selection methods.

The chapter is structured into two parts: first, the goal function is derived; second, the new method is experimentally compared with state-of-the-art variable selection methods.

6.1 Goal Function

As noted above a goal function for variable selection will be proposed which is derived from interaction gain feature selection. Interaction gain feature selection (IGFS) is defined by Akadi, Ouardighi and Aboutajdine (2008):

$$X_{IGFS} = \max_{X \in X_{-S}} (I(X_i; Y) + \frac{1}{d} \sum_{X_j \in X_S} I(X_i; X_j; Y)) \tag{6.1}$$

The authors report that this goal function performs better than the conditional mutual information criterion. In the following it is argued that interaction gain feature selection is equivalent to conditional mutual information algorithms with an added adjustment for devaluing dimensions selected in later iterations. The argument of equivalency is based on reformulating the IGFS

goal function:

$$X_{IGFS_{goal}} = I(X_i;Y) + \frac{1}{d}\sum_{j=1}^{|X_s|} I(X_i;X_j;Y) \tag{6.2}$$

Firstly, it is important to note, that the term being reformulated is the goal function of the IGFS method (equation 6.2). In this equation, d refers to the number of selected variables. In the following it will be written as $|X_S|$ for clarity.

$$\begin{aligned} X_{IGFS_{goal}} &= I(X_i;Y) + \frac{1}{|X_S|}\sum_{j=1}^{|X_s|} I(X_i;X_j;Y) = \\ &= I(X_i;Y) + \frac{1}{|X_S|}\sum_{j=1}^{|X_S|} I(X_i,X_j;Y) - I(X_i;Y) - I(X_j;Y) \end{aligned} \tag{6.3}$$

In the first step (equation 6.3) the 3IG part of the term is replaced with it is definition (equation 2.33). After this first step, the equation is in a form that only requires the computation of the traditional mutual information (defined in equation 2.3). This has the added benefit that its structure can be analysed with a multitude of equivalency theorems available in literature (most notably Cover and Thomas (1991)).

$$\begin{aligned} X_{IGFS_{goal}} &= I(X_i;Y) + \frac{1}{|X_S|}\sum_{j=1}^{|X_S|} I(X_i,X_j;Y) - I(X_i;Y) - I(X_j;Y) = \\ &= \frac{|X_S|I(X_i;Y)}{|X_S|} - \frac{|X_s|I(X_i;Y)}{|X_S|} + \frac{1}{|X_S|}\sum_{j=1}^{|X_S|} I(X_i,X_j;Y) - I(X_j;Y) \end{aligned} \tag{6.4}$$

The second step (equation 6.4) makes use of the fact that the sum is over the index j. This implies, that the term $-I(X_i;Y)$ remains constant for all summands. The term is taken out of the sum by multiplying it by the number of summands, $|X_S|$, and by the scaling factor $\frac{1}{|X_S|}$. The term $I(X_i;Y)$, which

was originally infront of the sum, is multiplied and divided by $|X_S|$ to allow the combination of it with the term that was just taken out of the sum.

$$\begin{aligned} X_{IGFS_{goal}} &= \frac{1}{|X_S|} \sum_{j=1}^{|X_S|} I(X_i, X_j; Y) - I(X_j; Y) = \\ &= \frac{1}{|X_S|} \sum_{j=1}^{|X_S|} H(Y) - H(Y|X_i, X_j) - H(Y) + H(Y|X_j) \end{aligned} \tag{6.5}$$

In step three (equation 6.5) the two terms infront of the sum eliminate each other. The more crucial rewriting in this step is the replacement of the mutual information terms by the conditional entropy terms they are based on. See also equation 2.4 for the definition of mutual information in terms of entropy. Note that mutual information is symmetrical which implies several possible expansions in terms of entropy. The expansion in this transformation was chosen because it allows the elimination of $H(Y)$ easily by showing that it is added and subtracted to the same term.

$$X_{IGFS_{goal}} = \frac{1}{|X_S|} \sum_{j=1}^{|X_S|} H(Y|X_j) - H(Y|X_i, X_j) \tag{6.6}$$

In step four (equation 6.6), $H(Y)$ is eliminated from the sum because it is added as well as subtracted from the term.

$$X_{IGFS_{goal}} = \frac{1}{|X_S|} \sum_{j=1}^{|X_S|} I(Y; X_i|X_j) \tag{6.7}$$

And finally, it is observed that the term which remains inside the sum, $H(Y|X_j) - H(X_i|X_j)$, is in a form which corresponds to the definition of conditional mutual information (see equation 2.4). After rewriting it as $I(Y; X_i|X_j)$, it is demonstrated, that $X_{IGFS_{goal}}$ can be rewritten into an equivalent form

which only requires the computation of conditional mutual information.

In this form the function is immediately reminiscent of the feature selection method proposed by Fleuret (2004). The main difference is that the new function reduces the weight of dimensions chosen under a large cardinality of X_S. In literature it is consistently reported that forward search methods are biased (Van Dijck and Van Hulle 2010) towards dimensions which were chosen in late iterations and that introducing an adjustment for this bias improves performance. It can therefore be argued that this difference is what contributes to the performance improvement of IGFS over Fleuret's CMIM method.

Further possible directions for performance improvement of this new goal function are motivated by the difficulties of the naive method for information theoretic variable selection. In the naive variable selection method, the goal is to optimise $H(Y|X_S)$ (see equation 2.2).

The reasons for which this method is difficult to compute are in the necessity to iterate over all possible subsets of dimensions X_S as well as the estimation of high dimensional probability distributions. In IGFS and CMIM the first difficulty is dealt with by applying a forward search strategy. The second is dealt with by calculating the conditional mutual information which only requires the estimation of up to 3-dimensional distributions. There are thus two options for approaching the theoretical performance of the naive method: to conduct a more exhaustive search in the subset space or to consider more dimensions when estimating the goal function.

The search strategy is hard to improve beyond meta-heuristics due to the non-linearity of the goal function. Meta-heuristic methods have been considered for improving the performance of information theoretic feature selection (Huang, Cai and Xu 2007). Such methods are applicable regardless of the shape of the goal function and are not investigated in the scope of this research.

This leaves the second approach which is to increase the number of con-

ditional variables for the conditional mutual information. Up until now only one conditional variable was taken into consideration (in this example X_1 is evaluated as conditional on X_2):

$$I(Y; X_1|X_2) = H(Y|X_2) - H(Y|X_1, X_2) \tag{6.8}$$

The conditional variable can be replaced with a multi-dimensional variable (in this example denoted as X_j) without loss of generality:

$$I(Y; X_1|X_{j_1}, ..., X_{j_k}) = H(Y|X_{j_1}, ..., X_{j_k}) - H(Y|X_1, X_{j_1}, ..., X_{j_k}) \tag{6.9}$$

where $X_{j_1}, ..., X_{j_k}$ are the dimensions of X which are already selected. However if the number k of conditional variables is not limited, the method will quickly run into problems similar to those found in the naive approach – computational complexity and sparsity of the data. To avoid these problems k can be limited to a constant. If k is chosen to be large (more than 3) then a large data set is needed to reliably estimate the required probabilities. In the original CMIM method $\binom{n}{2}$ evaluations are needed to estimate the theoretical information contributed by a candidate dimension. If k conditional variables are taken into account, $\binom{n}{k}$ evaluations are needed. While this number can grow quickly, it is a lot more manageable than the 2^n evaluations in the naive approach.

A version of the transformed IGFS goal function (equation 6.7) which takes these considerations into account can be written as:

$$X_{IGFS_{NEW}} = \frac{1}{\binom{|X_s|}{k}} \sum_{j=1}^{\binom{|X_s|}{k}} I(Y; X_i|X_j, ..., X_k) \tag{6.10}$$

In the version with multiple conditional variables (equation 6.10) there are $\binom{|X_S|}{k}$ summands. For this reason the sum needs to be scaled proportionally

and is divided by $\binom{|X_s|}{k}$ rather than just $|X_S|$.

6.2 Evaluation

The PCMIM algorithm proposed in chapter 4 can be plugged into any feature selection strategy which requires an exhaustive estimate of CMI. For the purpose of demonstrating the advantage of the proposed algorithm, its output was used to calculate a distance matrix for the mRRC method as described in section 2.4.8. A minor change in the method, which is responsible for the improvemed performance of mRRC based on PCMIM in comparison to a CPU implementation of mRRC, is that the function is calculated in its verbatim form:

$$D(X_i; X_j) = I(Y; X_i|X_j) + I(Y; X_j|X_i) \tag{6.11}$$

rather than being rewritten to:

$$D(X_i; X_j) = H(Y|X_i) + H(Y|X_j) - H(Y|X_i, X_j). \tag{6.12}$$

an argument in support of this claimed advantage is given in section 5.1. Both versions of the goal function are implemented and compared

The performance of 51 classifiers was compared on the 10 data. Of these, one is MRMR (Peng, Long and Ding 2005), 36 are state-of-the-art and classical feature selection methods implemented in R[1] (Robnik-ikonja and Kononenko 2003, Robnik-ikonja 2003, Kononenko 1995) including variants of Relief and ReliefF, minimum description length (MDL), DKM as well as standard methods such as information gain, gini, euclidian distance and others. Of the remaining 14 one mRRClust algorithm as originally proposed by Martínez Sotoca and Pla (2010) (mRRCPU). An alteration of his method is that it is calculated with a state-of-the-art implementation of a CMIM estimator (Meyer 2008) as

[1] `http://cran.r-project.org/web/packages/CORElearn/`

the author's implementation is prohibitavely slow. Another 6 are variations of this method with different clustering methods implemented in the hclust R package. The remaining 7 are the mRRClust method with the exact goal function implemented on GPU with the 7 clustering methods available in hclust (mRRGPU). The $X_{IGFS_{NEW}}$ method was not considered because in section 2.4 ample evidence is given that forward search strategies are sub-optimal. For brevity the results of the experiments are sumarised and only interesting outcomes are discussed in detail.

6.2.1 Data

The performance of the proposed algorithm comes at the cost of flexibility. The implementation imposes some constraints on which data can be analysed using the parallel CMIM algorithm:

First, the data needs to be discrete. A further, stricter constraint is imposed on the number of values a dimension may take. This constraint stems from the fact that the algorithm is based on sorting 32 bit integers. If the CMIM criterion is evaluated for 2 conditional dimensions, then every subset of dimensions with cardinality 4 (1 dimension for the class, 1 for the dimension that is being evaluated and 2 conditional dimensions) must be representable in 31 bits (the 32nd bit is used by the sorting algorithm). For example, a dimension that can take 12 values can be represented in 4 bits. If the four dimensions in a data set which can take the largest number of different values all are limited to 12 values, then they can be represented in $4 \times 4 = 16$ bits which makes the data suitable for the proposed algorithm.

The second constraint is computational; because the algorithm is exhaustive, care needs to be taken to a k so that $\binom{d}{k}$ remains managable where d is the total number of dimensions and k is the number of dimensions for which a multi-dimensional probability is estimated. For example, if a data has 60 dimensions and CMIM is to be calculated with 2 conditional dimensions, the

	dimensions	observations	classes	imbalance	max. bits
Zoo	17	101	7	36.63	9
SPECT	23	266	2	58.8	5
Dermatology	35	366	6	25.1	15
Soybean	36	307	19	12.7	14
Sonar	61	208	2	6.73	5
KR vs KP	37	3196	2	4.44	6
Mushroom	23	8124	2	3.59	20
Splice	61	3190	3	27.8	15
Simulated	2000	61	2	66.6	15
mmsesub	45	540	2	17.03	19

Table 6.1: Summary of data

algorithm would evaluate $\binom{60}{4} = 487635$ combinations of dimensions. Evaluating CMIM with 3 conditional dimensions for the same data would take 11 times as much time.

To demonstrate the generalisability of the PCMIM algorithm, the comparison was carried out using eight different data sets from various domains as well as simulated data. All data is publically available from the UCI machine learning repository [2] and is included in the R package which implements the algorithm. A more detailed description of the data is found in section 3.4. A more thorough investigation is carried out on the on the MMSE data with semantic annotations as described in chapter 7 reduced to only patients diagnosed with either Alzheimer's Disease or Vascular Dementia.

For different data it is reasonable to expect different performance depending on factors such as number of observations, number of dimensions and number of classes. These factors are listed for all considered data in table 6.1. In the table the dimension count includes the class label. The last column, max. bits, shows how many bits are needed to encode the 5 widest dimensions of the data – for the PCMIM algorithm to be applicable these dimensions need to fit in 31 bits (see also section 4.3).

Another factor which has been taken into account as a possible indicator of the performance of CMIM feature selection is the number of bits required

[2] `http://archive.ics.uci.edu/ml/`

to represent the 5 dimensions with a widest domain. The rationale for this is that the number of observations needs to be large to estimate a large number of pointwise probabilities accurately. Conversely, if the domains of the dimensions are narrow, less data is needed for an accurate estimation. An intuitive justification is to consider the case of a loaded coin and a loaded die. One would need more experiments to identify a loaded six-sided die by esimating its probabilities from observations than to identify a loaded coin which only has two possible outcomes.

One factor which has been measured but not discussed in the description of the algorithm is class imbalance. In machine learning literature, class imbalance is recognised as a source of bias in many algorithms (Japkowicz and Stephen 2002).

The class imbalance measure used here is the difference of the most prevalent class in the data versus the rarest class in the data. It is defined as CI:

$$CI = \frac{100 \times (|C_{max\_obs}| - |C_{min\_obs}|)}{n} \tag{6.13}$$

where C_{max_obs} is the class which was observed most frequently in the data, C_{min_obs} is the class which was observed least frequently and n is the total number of observations. The unit of the measure is a percentage of the total number of observations.

For example a data with two classes, the first of which has been observed 50 times and the second of which has been observed 100 times has a class imbalance coefficient of:

$$\frac{100}{150} - \frac{50}{150} = \frac{1}{3} = 33\% \tag{6.14}$$

It is important to note that this coefficient gives only a rough picture of the spread of the number of observations over classes. While it measures the distance between the classes with the largest and smallest number of observations

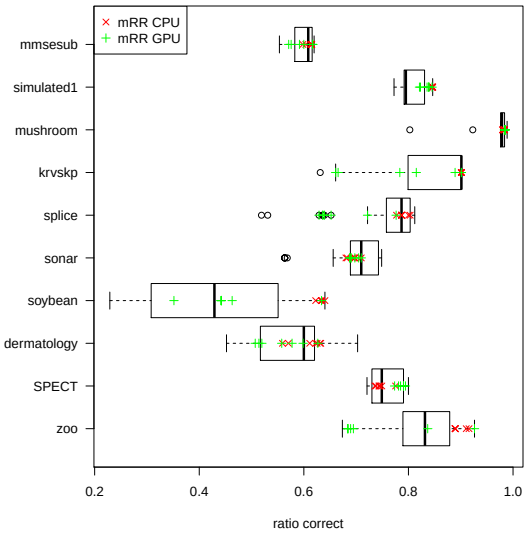
it gives no information on the distribution of the remaining classes.

6.2.2 Results

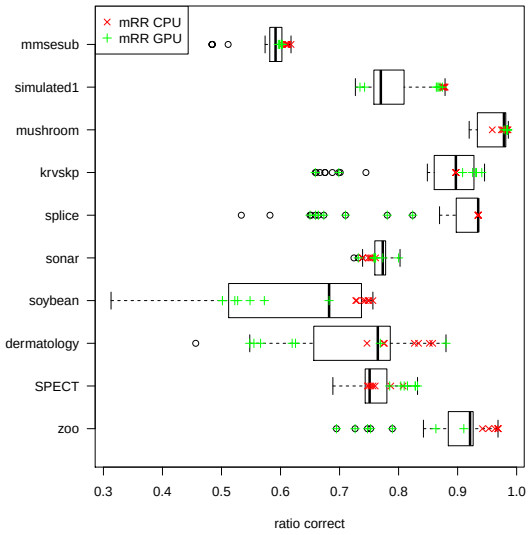
Because of the large number of feature selection methods, it is difficult to visualise all results for all methods for all data. For this reason, the classification accuracy of the different methods is visualised as a box and whiskers plot for each data. The performance of the proposed methods is highlighted with a cross for the mRRClust variants based on a CPU approximation and with an x for the methods based on the exact computation on GPU. Figure 6.1a shows the performance of a naive Bayes classifier with cross validation on the best three variables with each method. Figures 6.1b and 6.1c show the same comparison with 6 and 10 variables respectively. The accuracy variance across cross validation iterations was estimated but as it was less than one percent in all cases, it is not reported in detail.

Overall it can be said, that the Zoo, KR vs KP, mushroom data are easy as most of the methods achieve more than 80% classification accuracy regardless of how many variables are selected. The simulated data is a good demonstration for the merit of the proposed exact mRRClust method – the classification performance of all methods but the GPU based ones decreases as more variables are selected. Because of the redundancy introduced in the data with its generation, selecting new variables which do not contain information already present in the model becomes more difficult as the number of selected variables increases. Regardless, some of the clustering methods seem to be inadequate as a replacement for the ward method which overall achieved the best and most consistent results across the different data. An interesting result is the performance of mRRCWard on the Soybean data and on the Dermatology data which demonstrate that there is merit in estimating the mRRC matrix without rewriting the goal function.

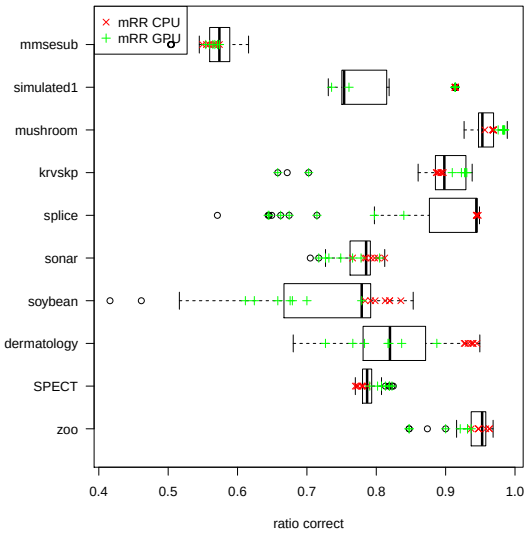
Figure 6.2 summarises the running time for all feature selection methods



(a) 3 variables



(b) 6 variables



(c) 10 variables

Figure 6.1: Classifcation performances with the 3, 6 and 10 best variables using various methods.

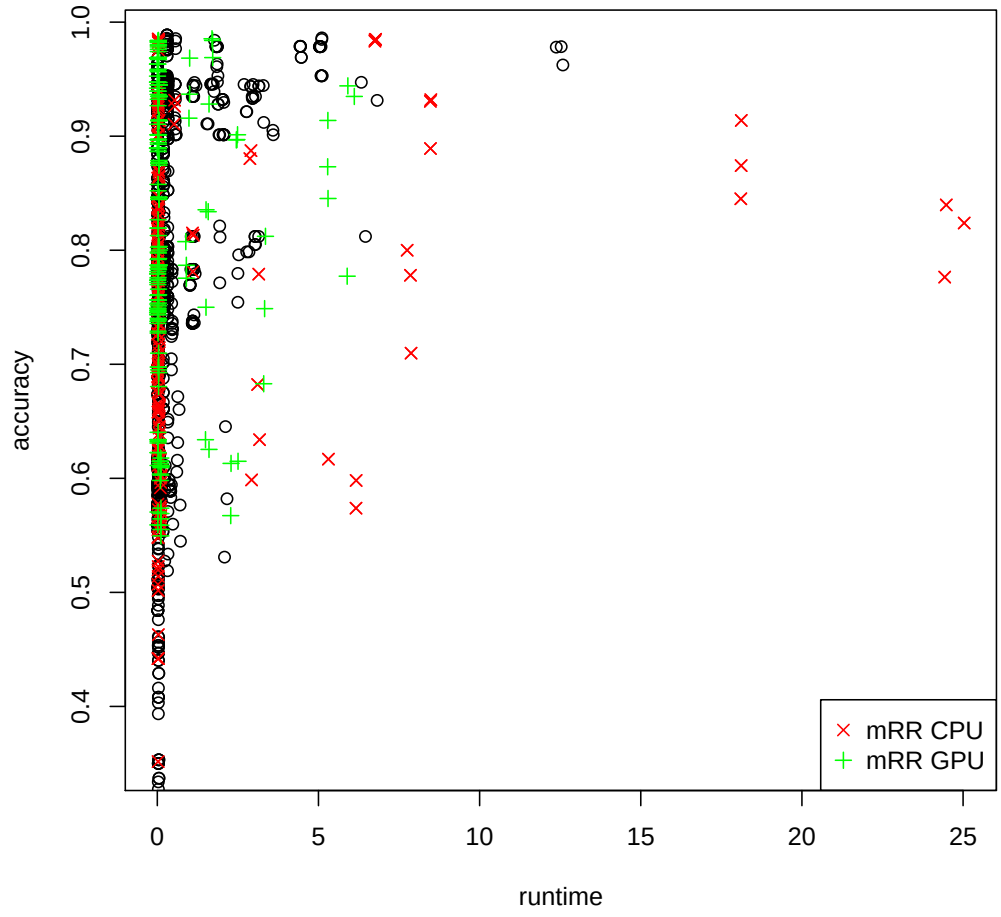


Figure 6.2: Runtime versus accuracy for all experiments.

and plots them against the achieved performance. The duration of calculations for the CPU imlpementation of the goal function are marked with an x and the runtimes for the GPU imlpementation with a cross. Overall, the GPU computation is competitive with the other methods although in some cases the computation takes up to 6 seconds. Table 6.2 shows a summary of the GPU versus the GPU mRR implementations. This summary suggests that while for the majority of cases the computation times of both methods are similar, for the 25% most difficult cases in the 4th quartile the CPU implementation can take up to five times longer.

Similarly to the PCA methods in chapter 5, all methods seem to perform poorly on the reduced, semantically annotated MMSE data. In the previous

	GPU	CPU
Min.	0.0050	0.0080
1st Qu.	0.0250	0.0250
Median	0.0330	0.0355
Mean	0.4064	1.1681
3rd Qu.	0.0580	0.0600
Max.	6.1120	25.0340

Table 6.2: Summary of runtimes for GPU and CPU mRR implementations in seconds.

chapter a random forest (Breiman 2001) classifier showed significanntly better results on the MMSE data than the naive Bayes classifier used in the above experiment. For this reason, a second experiment was conducted in which all 51 methods were used to select the best 3, 4, 5, up to 40 variables of the MMSE data. The quality of the subsets was gauged by the ratio of correctly classified instances on test set made up of 20% of the original data. Figure 6.3 summarises the result of this experiment. Each of the bars is represents the best classification accuracy on the particular subset achieved with any of the 51 feature selection methods. The best result, just over 70% correctly classified instances, is achieved with 20 variables. Several methods achieved this accuracy with this subset within a small margin of each other.

The subset of variables is listed in table 6.3. The score is given a high priority because in this version of the data it is a numeric variable where each value is treated separately. To achieve a more realistic result, a startification by ranges of score may be necessary. Such issues are investigated further in chapter 8.

6.3 Conclusion and Future Work

In this chapter two major propositions were made: a reformulation of a state-of-the-art feature selection criterion in terms of conditional mutual information maximisation; a family of efficient algorithms based on the mRRC method and the PCMIM algorithm proposed in chapter 4. The presented research

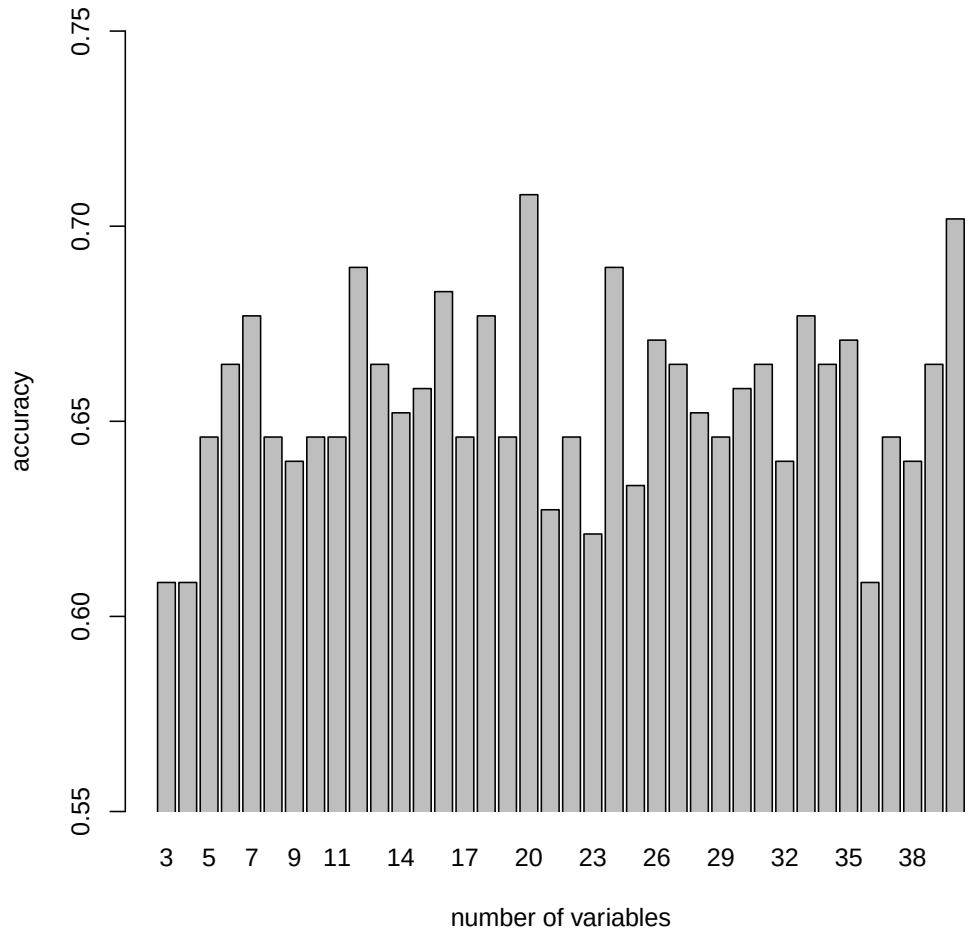


Figure 6.3: Classification accuracy of random forest classifier on MMSE data with 3 to 40 out 45 selected variables.

demonstrates that shifting from a sequential to a parallel paradigm, can extend the applicability of existing variable selection strategies to larger data.

The merit of reformulating X_{IGFS} is that by analysing equivalent forms, its structure can be directly compared to other state-of-the-art methods, most prominently it can be directly compared to conditional mutual information maximisation. This comparison yields an understanding of the theoretical nature of the performance improvement of IGFS when compared to CMIM as proposed by Fleuret (2004).

In future work a metric for the mRRC method will be designed which takes more than one conditional variable into account.

1	sex
2	score
3	word_count
4	A11
5	A8
6	auxiliary
7	adjectives
8	A3
9	A4
10	A21
11	nouns
12	A26
13	subjects
14	adverbs
15	A9
16	determiners
17	A17
18	A7
19	A30
20	modifiers

Table 6.3: Best subset of variables of the MMSE data for discrimination between Alzheimer's Disease and Vascular Dementia using random forests.

Chapter 7

Semantic Analysis of the MMSE Sentence

One of the MMSE questions prompts the testee to write a sentence. The question is scored with 1 point for a successful attempt and correct sentence and 0 points otherwise. In this chapter, more diagnostic cues are extracted from the sentence writing question than just a binary score. Although language impairment in dementia patients is well researched (Kempler and Zelinski 1994), little literature reports success in identifying linguistic markers applicable to the MMSE. Nevertheless, researchers report indirect evidence for the relevance of the sentence in the MMSE (Shenkin et al. 2008, Press et al. 2012).

In this study 101 grammatical and syntactical cues are evaluated for their contribution to discriminating between Alzheimer's (AD) and Vascular Dementia (VaD) using the MMSE.

The remainder of the chapter is structured as follows: section 7.1 describes the automated linguistic analysis method used in this study, section 7.2 presents results and discusses selected linguistic markers indepth, and section 7.3 concludes by putting the presented research in a larger context and outlining future work.

7.1 Linguistic Processing

The sentences in the data were parsed using the Stanford semantic parser (Klein and Manning 2003). The Stanford parser is reported to have high accuracy when matched against annotations made by linguists. The accuracy of automated parsing in detecting MCI has been confirmed by Roark et al. (2011), who report coincident significance levels for data produced manually

and automatically. Automatic parsing allows annotating a larger corpus of text with more semantic information than is feasible with manual annotation.

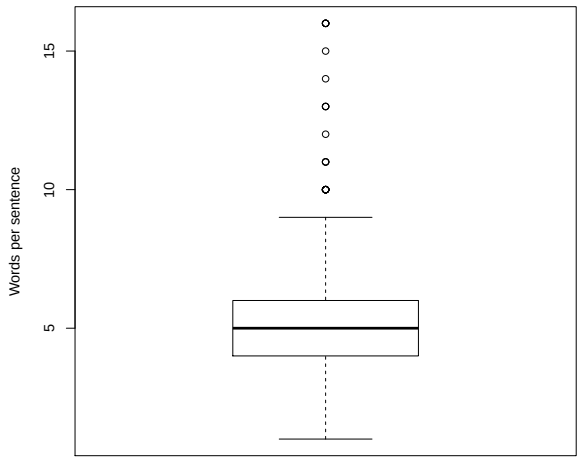


Figure 7.1: Distribution of the number of words per sentence in the MMSE data.

The 418 sentences were annotated with 48 types of parts of speech (for example noun, verb, adjective...) and 53 types of grammatical dependencies (for example subject to object, auxiliar verb to verb...). The median number of words per sentence is 5 with 75% of the sentences being shorter than 6 words (see figure 7.1). Such relatively short sentences do not allow for complex grammatical structures to emerge. For this reason, complexity measures scoring the shape of the parse tree of a sentence were not considered.

Out of the 101 assessed measures, only those are reported in the scope that carry more information about the diagnosis than the least informative MMSE question. These measures are (listed in descending order of information):

1. words per sentence
2. maximal word length
3. adjectives (adjectives in the comparative and superlative form are included)
4. nouns (singular and plural forms of nouns and proper nouns are counted)

5. determiners (e.g. which, the, a)
6. auxiliar verb relationships (e.g. should do)
7. adjective complements (e.g in "She looks very beautiful." the adjective "beautiful" complements the verb "looks" - it acts as an object of the verb)
8. verbs (all verb inflections are counted)
9. prepositions
10. adverbs (includes comparative and superlative forms)
11. adjectival, temporal and noun compound modifiers
12. coordinating conjunctions
13. occurences of the word "to"
14. subject clauses

The relevance of the rules matching for nouns, verbs and number of words per sentence has been confirmed in related research (Vigliocco et al. 2011, Roark et al. 2011). The rule counting the number of subject clauses has been investigated by Bencini et al. (2011), who found that, compared to Italian native speakers, English native speakers do not omit subject clauses.

7.2 Results

To order the measured variables (101 linguistic markers, the 30 MMSE questions as well as the gender of the patient) in decreasing order according to their contribution to the diagnosis, an entropic dimensionality reduction method is chosen. Classical dimensionality reduction methods, such as factor analysis or principial components analysis, are avoided due to their bias when applied to

discrete data (Kolenikov and Angeles 2004). The method used in this study is the parallel mRRCWard method proposed in chapter 6.

After the linguistic markers listed in section 7.1 are assessed and the most significant markers are identified, the relation between each marker and the diagnosis needs to be determined. This relationship is reported in an exploratory manner for two reasons: first, the discrete distribution of the variables adveresly affects the practicality of classical statistical tools. A χ^2 test could be used to compare the histograms of, for example, the number of words per sentence for AD and VaD patients, however such a test would yield potentially misleading results because of the long distributional tails. The second reason is that the relationship between the variables and the diagnosis is not linear in all cases. Providing a linear model model would obscure information and a non-linear model would be more difficult to interpret intuitively.

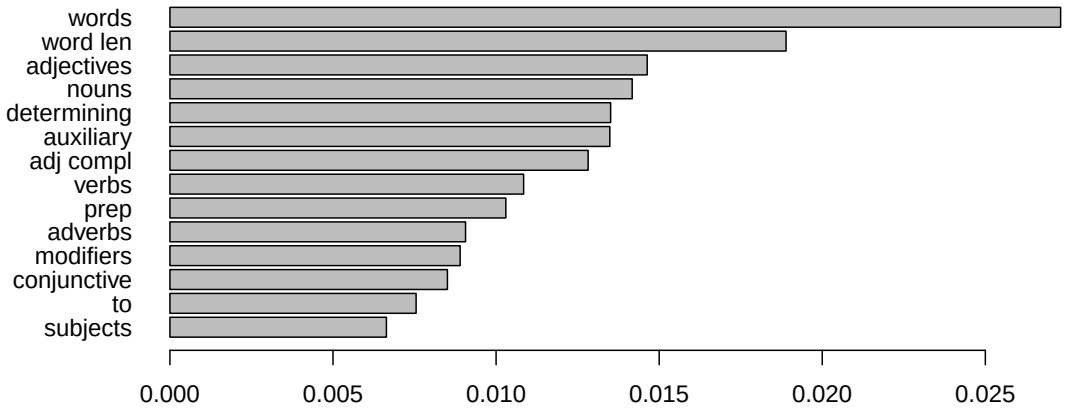


Figure 7.2: Information of linguistic markers about diagnosis measured in bits.

The amount of information each of the linguistic markers contributes to the final diagnosis is shown in figure 7.2. Although there are more linguistic markers which are scoring better than the least informative MMSE question in discriminating between groups, only the markers with information within the margins of the best 10 MMSE questions are presented in order to reduce noise.

Following this, the relation between the best 3 markers (words per sentence, length the longest word and number of adjectives) and the diagnosis are discussed more closely. Additionally the results are put in context with related work by discussing the number of nouns and verbs in sentences (Vigliocco et al. 2011) as well as the number of subject clauses (Bencini et al. 2011).

The relation between each variable and the class is reported with a graph containing two histograms, one obtained from the AD patients in the data and one from VaD patients, and a third graph depicting the difference of both histograms calculated by subtracting the corresponding variable values for AD and VaD patients. For example, figure 7.3 depicts the difference in the number of words per sentence for both patient groups. By examining the bar representing sentences with 3 words in histogram a) and histogram b) little difference between patient groups is observed. Graph c) shows that the difference between both histograms is indeed close to 0.

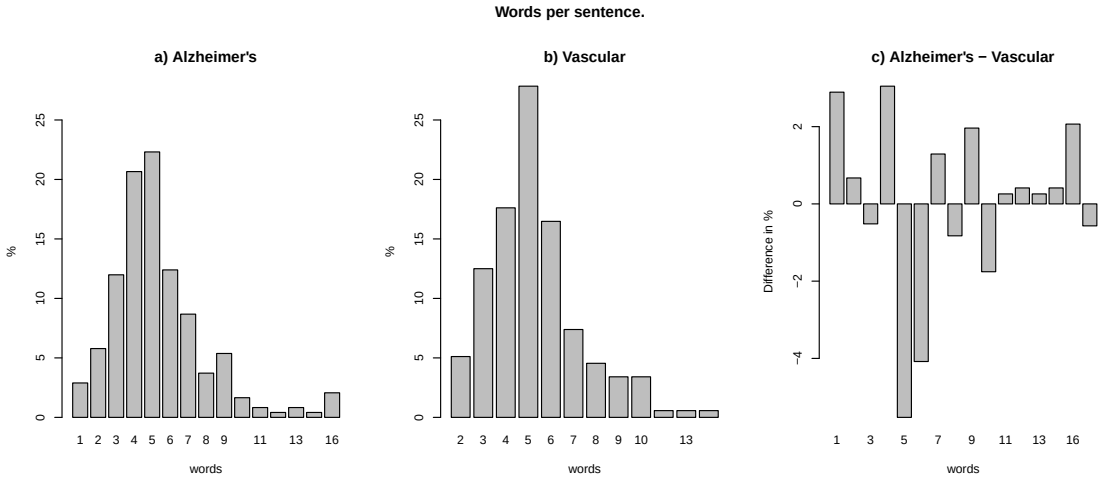


Figure 7.3: Comparison of words per sentence in different patient groups.

The most significant linguistic marker for distinguishing between AD and VaD is also the best predictor among the 30 MMSE questions, the 101 linguistic markers and knowledge of the gender of the patient. In related literature a lower number of words per sentence is associated with Alzheimer's disease (Bencini et al. 2011, S et al. 2011). This result, while confirmed, only seems to apply to sentences of up to 6 words. Although this is true for the majority of

sentences (see also figure 7.1), AD patients have written majority of the longest sentences. A possible explanation for this difference is that the patients were asked to write a sentence. Stopping an action is an ability which is affected in patients with AD.

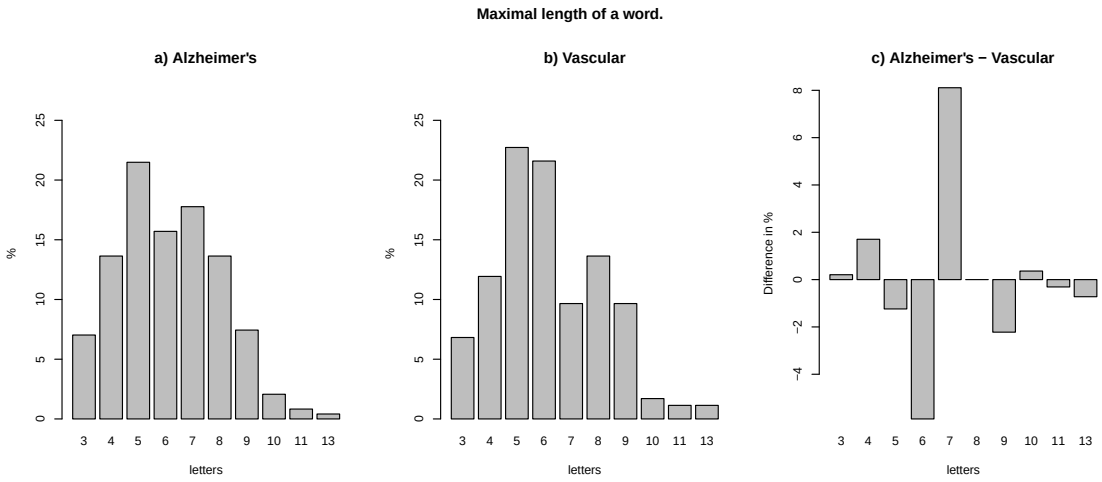


Figure 7.4: Comparison of the length of the longest word in a sentence for different patient groups.

The second most informative linguistic marker, the length of the longest word in a sentence, is preceded by 6 questions of the original MMSE. From the histograms in figure 7.4 it can be concluded that there is little difference between different groups of patients for sentences with the longest word no more than 5 letters long. However, a significant difference can be observed for sentences in which the longest word is 6 or 7 letters long. Patients with AD write fewer sentences in which the longest word is 6 letters long than sentences in which it is 7 letters long. In VaD patients, this relation is reversed - the relative difference between patients who prefer 6 letter words and those who prefer 7 letter words is more than 10%. A preliminary investigation of differences in the longest words of sentences does not yield additional insight and further investigation is warranted.

The information content of the maximal word length variable is immediatly followed by the variable counting the number of adjectives in a sentence. In figure 7.5 it can be seen that AD patients are more likely to use an adjective

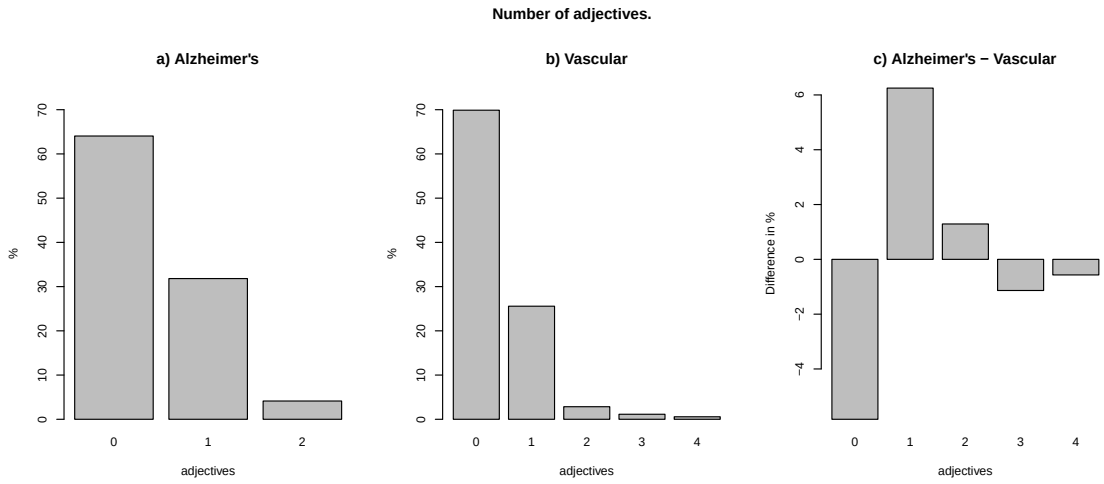


Figure 7.5: Comparison of adjectives per sentence in different patient groups.

in a sentence than VaD patients. In the english language adjectives are often associated with overly dramatic, less informative language. Considering this aspect, these findings are consistent with results reported in related work (Roark et al. 2011) which associate reduced idea density in use of language with AD.

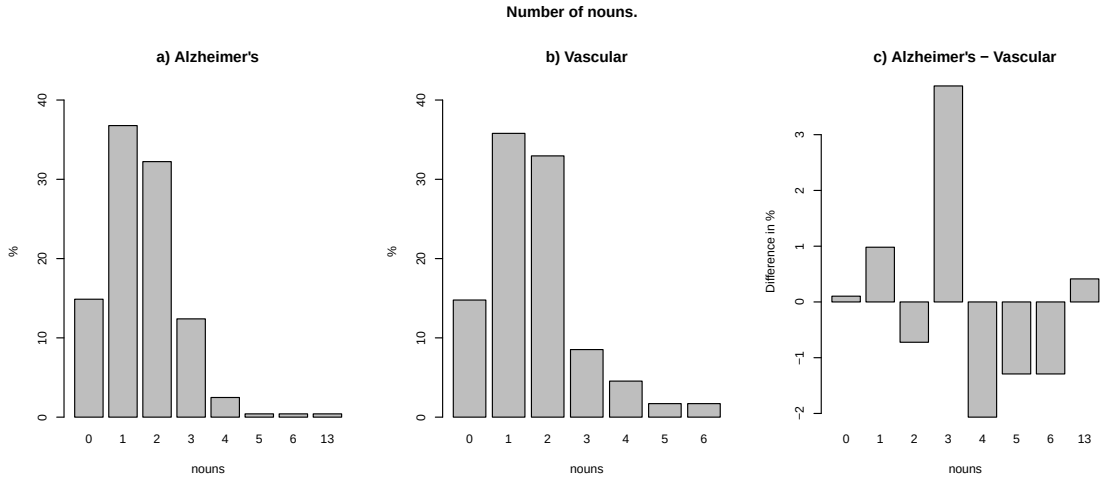


Figure 7.6: Comparison of nouns per sentence in different patient groups.

The comparison of the number of nouns (figure 7.6) and the number of verbs (figure 7.7) yields less information about the difference between groups. However, the results found in the data confirm the findings reported by Vigliocco et al. (2011): verbs place a higher load on working memory - which is impaired in AD patients, and the use of nouns and verbs differ in a similar way

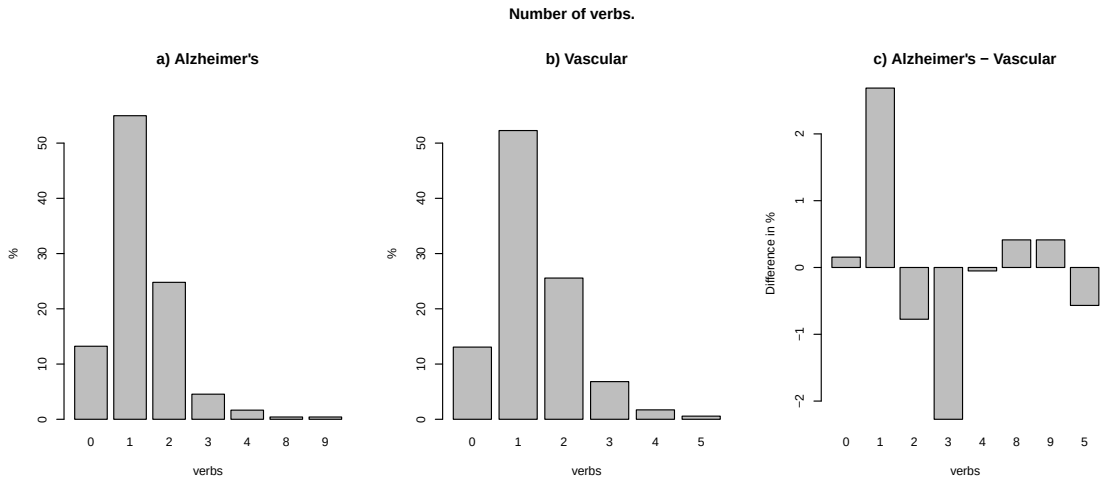


Figure 7.7: Comparison of verbs per sentence in different patient groups.

between groups, which can be explained with the shared brain region used for producing both types of words. The number of nouns as well as the number of verbs per sentence is lower in AD patients. The higher difficulty of verbs over nouns is reflected in the lower overall number of verbs compared to nouns in all sentences.

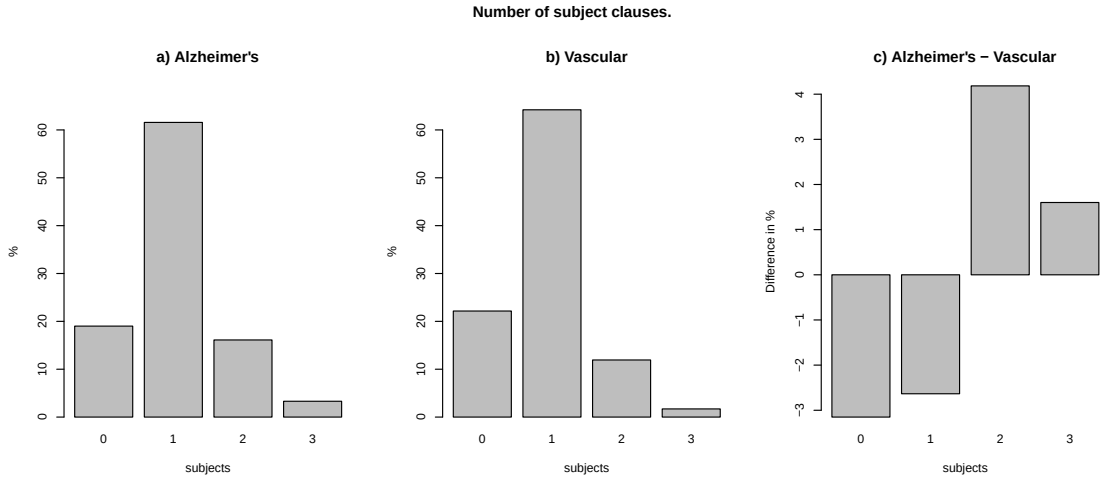


Figure 7.8: Comparison of subject clauses per sentence in different patient groups.

Bencini et al. (2011) report an affected use of subject clauses in AD patients. The authors report that English native speakers with AD do not omit subject clauses while Italian native speakers do. A trend for subject clause omission however is clearly visible in figure 7.8: AD patients write more sentences with 0 or 1 subject clauses and fewer sentences with more than one subject clause

than VaD patients. While these findings are consistent with the research of Bencini et al. (2011), it implies that subject clause usage may be affected across languages with the difference occuring in the shape of the affection.

Additional experiments providing supporting evidence for the merit of semantic annotations for discriminating between types of dementia are reported in chapter 8. More specifically: section 8.2 argues against considering MMSE data without semantic annotation for analysis with additional results presented in tables A.1 to A.24; section 8.3 discusses the accuracy of discriminating between all types of dementia listed in table 3.1; table 8.5 reports evidence supporting the results obtained in this chapter; section A reports accuracy results using the annotated MMSE data with classification methods, different dimensionality reduction methods (feature extraction and variable selection) and considering all 1 vs 1 and 1 vs all combinations of dementia groups.

7.3 Conclusion

This research proposes an automated method for identifying linguistic markers for discerning between AD and VaD. The automation of the task allows testing more markers on a larger text corpus than is feasible by manual annotation. Related research using automated methods to quantify language for the purpose of diagnostic support for dementia is scarce (Roark et al. 2011, S et al. 2011, Ahmed et al. 2013). The findings of comparable related work yield results which are not readily transferable to screening for dementia in clinical practice.

In the scope of this research 14 markers are identified which provide more information about the type of dementia a patient may have than 2 thirds of the original MMSE questions. The identified markers can be tested for by test administrators without expert knowledge of linguistics (number of words, number of adjectives, nouns...). In future work, the significance of adjectives in distinguishing between AD and VaD patients will be investigated.

A non-linear model which integrates the MMSE and the identified linguistic markers for the purpose of predicting the type of dementia is proposed in chapter 8.

Chapter 8

Predictive Model

In this chapter, the MMSE data is analysed for its applicability beyond a screening tool for dementia. The data is analysed for answer patterns and linguistic cues typical for types of dementia. Each of the patients is diagnosed with one of 24 types of dementia or is classed as norm. The 24 types of dementia are collected in diagnostic groups according to table 3.1.

The remainder of this chapter is organised as follows: section 8.1 gives a descriptive overview of the main features of the data; in section 8.2 the data is analysed in an exploratory manner, without formulating a model; section 8.3 formulates models for the data using the methods proposed in chapters 6 and 5.

8.1 Descriptive Statistics

Traditionally, the main features of a data are described using the central moments of the data - mean, variance, skewness and kurtosis. Since the MMSE data is discrete, and in the case of gender nominal, these measures, which assume a normal distribution, will not be considered. Instead, the data is presented in relative frequencies and, where appropriate, distributional shapes are depicted with a boxplot graph (Tukey 1977). A boxplot graph shows the inter-quartile range (IQR) as the bottom and top edge (hinge) of its box. The line accross the box is at the median of the data and the top and bottom end of the whiskers are at the 5th and 95th percentile. In addition, the boxplot graphs contain notches whose top and bottom ends are calculated as $+ - IQR/\sqrt{n}$ where n is the number of observations. These notches overlap when the medians of two samples are not significantly different.

In the MMSE data, there are roughly 20% more female than male patients

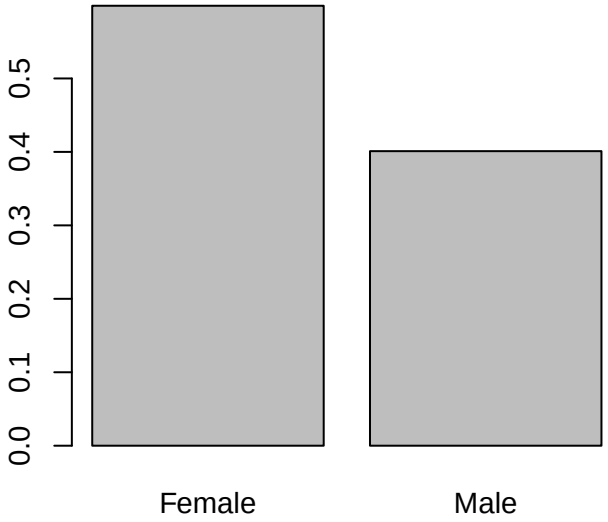


Figure 8.1: Gender ratio in the MMSE data.

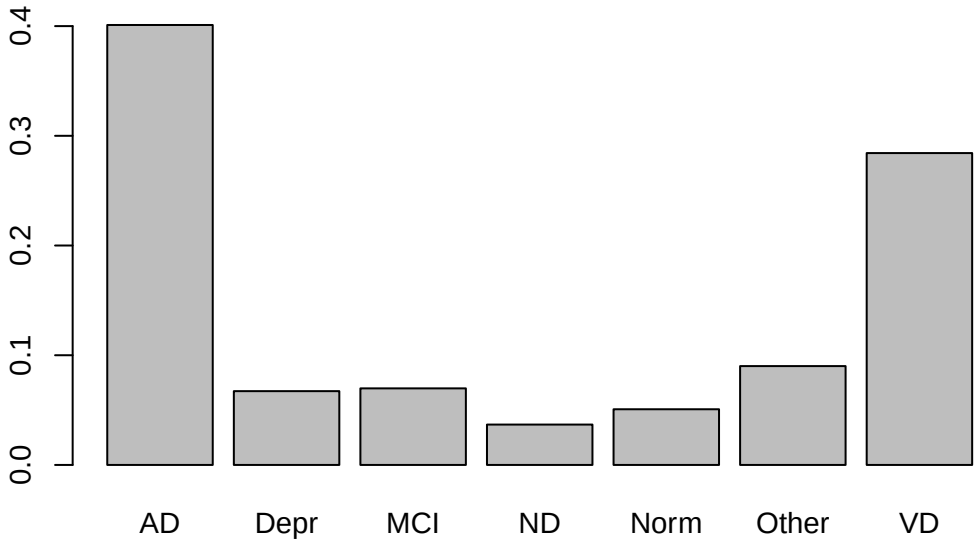


Figure 8.2: Diagnostic group ratios in the MMSE data.

Group	Abbreviation	Full Name
1	*AD*	Alzheimer's Disease
2	*Depr*	Depression
3	*MCI*	Mild Cognitive Impairment
4	*ND*	Neuro Degenrative Dementia
5	*Norm*	Norm
6	*Other*	Other
7	*VD*	Vascular Dementia

Table 8.1: Abbreviations used in graphs and tables.

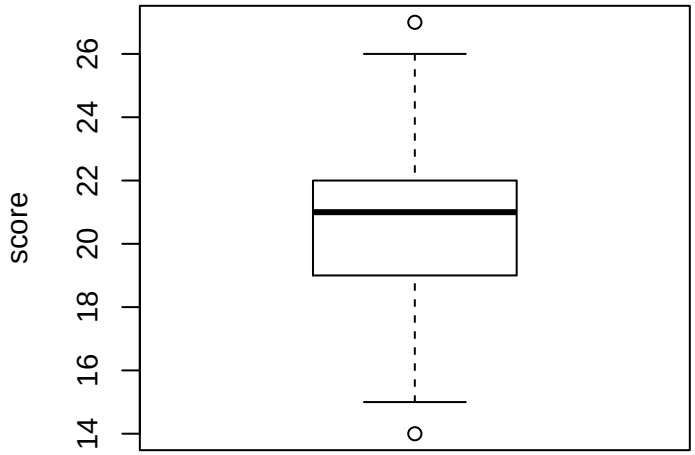


Figure 8.3: Total MMSE score frequencies.

(figure 8.1). This can be explained with the generally higher life-expectancy of women. The two most prevalent diagnoses in the data are Alzheimer's Disease and Vascular Dementia (figure 8.2). To keep the width of the graphs reasonable, the diagnostic groups are abbreviated (see table 8.1).

The median of the MMSE scores is 21 with the IQR between 19 and 22 – in other words, 50% of the patients had a score between 19 and 22. There is no significant difference in the spread of scores between male and female patients as can be seen from the overlapping notches of the boxplots in figure 8.4.

Although with 50% of the patients being scored with one of only 4 values it is difficult to draw further conclusions from the spread of scores, it is in-

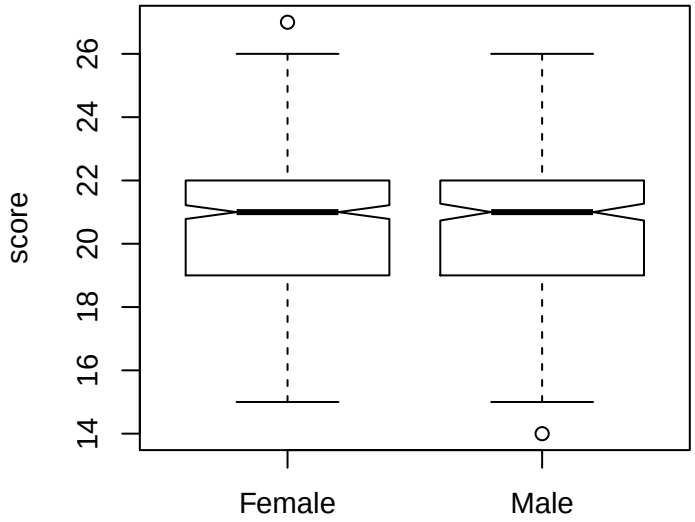


Figure 8.4: Total MMSE score for both genders.

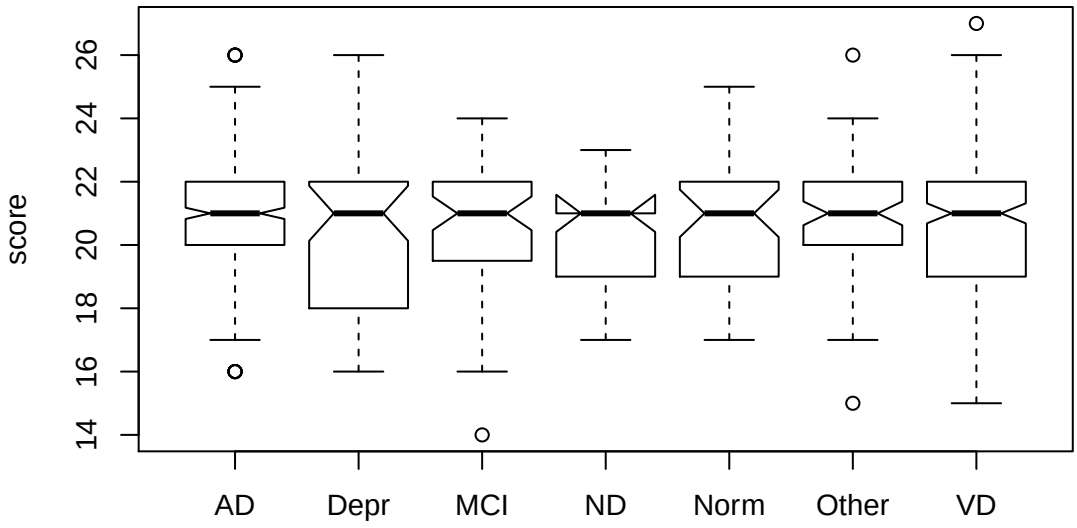


Figure 8.5: Total MMSE score for each diagnostic group.

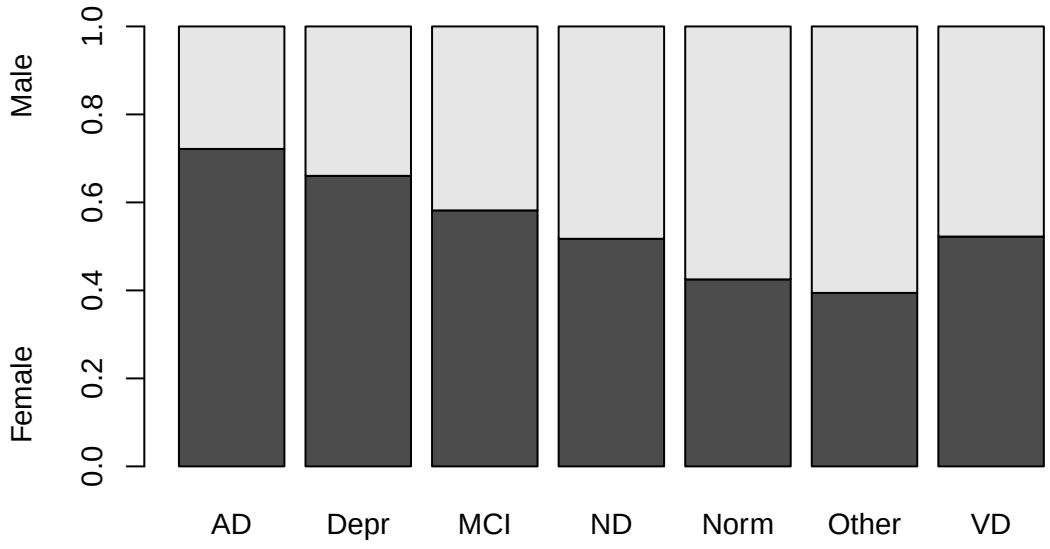


Figure 8.6: Gender ratios for each diagnostic group.

formative to look at the difference in the scores between different diagnoses (see figure 8.5). The notches of all boxplots overlap, which strongly suggests that there is no significant difference between the scores for different diagnostic groups. Curiously, the median score for the patients at norm is indistinguishable from the patients with a cognitive impairment. An explanation for this can be found in the population from which the data was drawn. All MMS Examinations which were considered in this study were completed by patients referred by their physician to the Memory Centre at Llangdough hospital. The general practitioners who referred the patients to the hospital have adherred to the guideline of the National Institute for Clinical Excellence which suggests an MMSE score of 26 as a cut-off point for dementia screening.

Finally, structural information can be gleaned by examining the ratio of genders in each of the diagnoses (see figure 8.6). The most significant finding is the difference in the gender ratio between Vascular Dementia (VD) and Alzheimer's Disease (AD). This is a well known correlation associated with the higher prevalence of vascular disease in the male population. The over-

representation of male patients in the group Other can be explained by referring to table 3.1 – this group is dominated by stroke and alcoholism patients, both of which are prevalently male.

8.2 Exploratory Analysis

The MMSE data, even in its transformed version, is split into 7 diagnostic groups. A priori, little is known about how the 7 groups relate to each other. The relation between groups can be modelled by asking the question: How well can group A and group B be separated by a statistical model? There are 23 such combinations of pairs of groups. This question can also be asked in the form: How well can diagnostic group A be separated from all other groups? Those two relatively simple inquiries already yield 56 possible hypotheses about the data. However, this number grows further if relations between more than two groups of patients are considered, for example: how well can the group of patients at norm be distinguished from the patients diagnosed with either AD or VaD?

A method for guiding the formulation of hypotheses is exploratory analysis using homogeneity analysis (HOMALS) as proposed by Gifi (1990) and described in section 2.3.3. The HOMALS method rotates data in such a way that variations in the data are maximised considering observations against variables as well as variables against observations. Because the optimisation of the rotation is performed iteratively, the rotation cannot be generalised for new data. On the other hand, the rotation is performed without distributional assumptions about the data, which yields better results than methods making such assumptions (see chapter 5 for a comparison).

In this study two sets of data were considered: the original MMSE data which includes a score for each of the 30 questions and the gender of a patient, and a subset of the original MMSE data limited to those patients who attempted the sentence writing question. The linguistic markers identified in

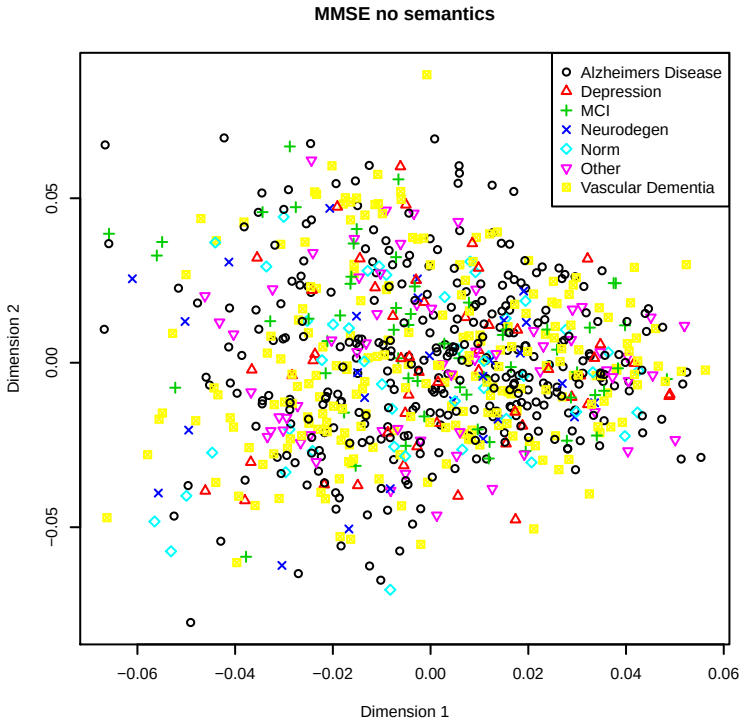


Figure 8.7: Scatter plot of HOMALS rotated MMSE data without semantic analysis.

chapter 7 are included in the second data. The reason for the decision to analyse both sets of data is twofold: First, for some patients who attempted to write a sentence, this sentence is not available; secondly, integrating the semantic analysis with the original data would result in 16 columns full of 0 values for all patients who either did not attempt to write a sentence or for who a sentence is not available. This concentration of equal values in 16 of the dimensions would skew the analysis.

The first aspect discussed is whether the 7 groups can be separated visually after rotation with the HOMALS method and projection onto two dimensions. Figure 8.7 depicts the data cloud produced in this manner from the MMSE data under exclusion of the semantic analysis. There is large overlap between the different diagnostic groups. While it may be possible to achieve some separation by looking at pairs of diagnoses individually, for example by reducing the data to the two largest diagnostic groups, AD and VaD, the visual separation of groups discourages further investigation. In light of the results achieved

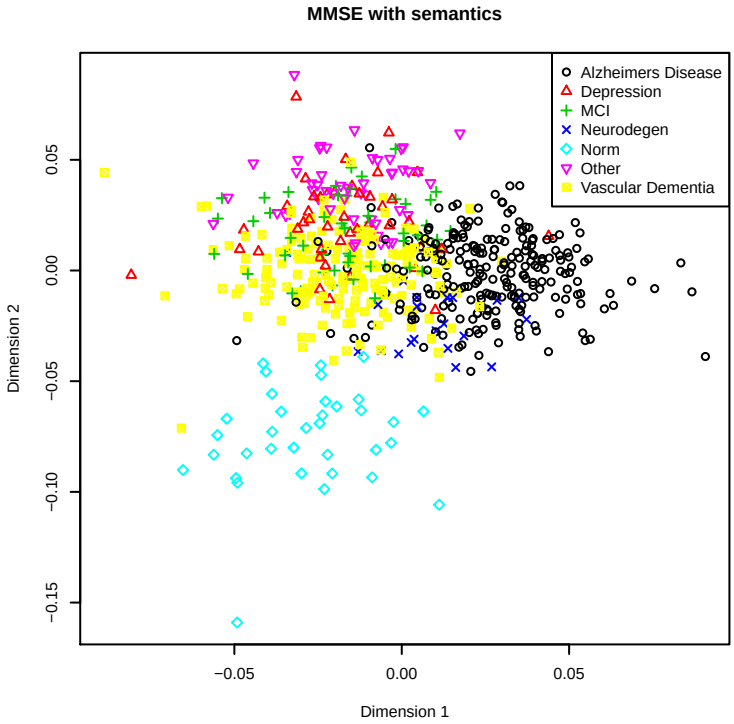


Figure 8.8: Scatter plot of HOMALS rotated MMSE data with semantic analysis.

after the inclusion of linguistic markers from the MMSE sentence, it is deemed unnecessary to pursue an analysis of the original MMSE data.

Figure 8.8 shows the same graph as before however, for the semantically annotated MMSE data. Although a large overlap between classes remains, some interesting separations can be seen. For example, the groups of patients at norm, patients with MCI and AD seem separable. Distinguishing MCI from norm is a typically difficult task in clinical practice. A second interesting observation is that depression can be separated from AD and neuro-degenerative disorders.

The relations between diagnostic groups can also be observed in the category quantification graph of the outcome variable of the data (figure 8.9). While this graph can be shown for each of the variables in the data, the most informative graph for the purposes of this analysis is the graph for the diagnosis. In a category quantification graph, the centroid of each category is plotted onto two-dimensional space. The further apart two centroids are on the graph,

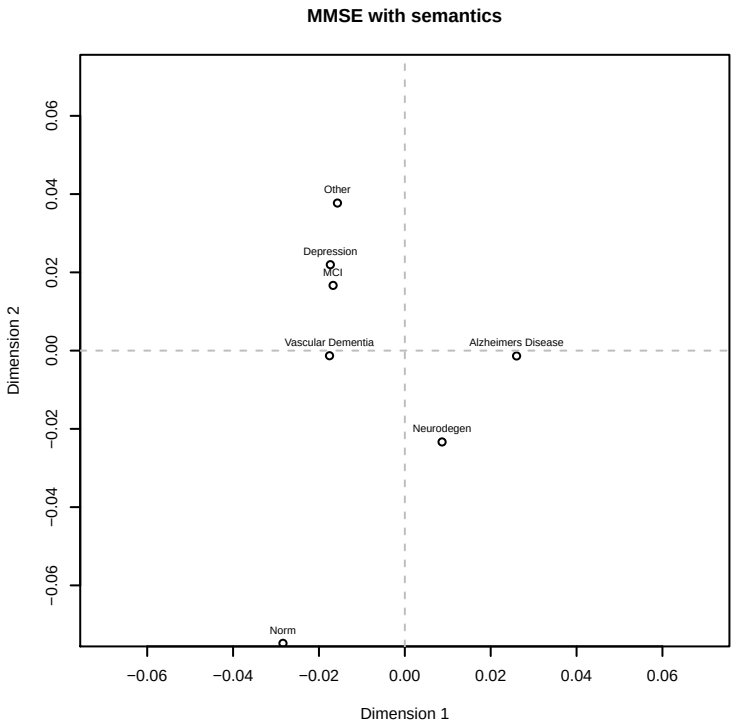


Figure 8.9: Category quantification for plot MMSE data with semantic analysis.

the larger the between groups variance. The graph supports the observations made by rough investigation of the scatter plot: the Norm category is far from all others; depression and MCI are close to each other but far from Neurodegen and Alzheimer's Disease. The distance between AD and VaD is larger than to the other impairments but not as large as to Norm.

Finally, it is informative to investigate the relations of dimensions to each other. The HOMALS method is equivalent to linear decomposition under optimal scaling. Although optimal scaling is not generalisable, the linear decomposition can be interpreted analogous to traditional methods. In linear decomposition, the data, or an indicator of its structure, such as the covariance matrix, is decomposed into components and coefficients. If the value of a cell of the first component (corresponding to a column in the original matrix) is large, this signifies a large contribution to the rotation of the data along the first component by this column of the original matrix. In a loading plot, the first two components of the decomposition are plotted as two-dimensional

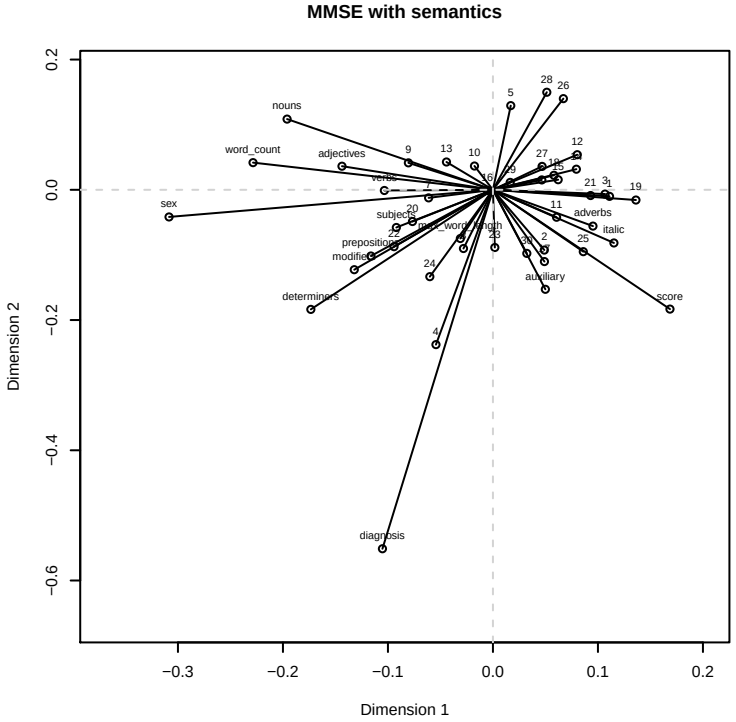


Figure 8.10: Loading plot for MMSE data with semantic analysis.

data. If two points on the plot are close, the two columns corresponding to the points are interpreted as strongly correlated. Two columns are strongly negatively correlated in the projected space if they are symmetric around the origin of the graph.

Figure 8.10 shows a loading plot of the MMSE data with semantic analysis. The numbers from 1 to 30 correspond to one of the MMSE questions. An observation is that the linguistic markers have much larger contributions to the rotation than many of the original MMSE questions. For example, the point corresponding to the number of determiners in a sentence is much further from the origin than the points corresponding to questions 10, 13, 16, 29 and etc. This observation is further evidence supporting the value of semantic analysis of the MMSE data.

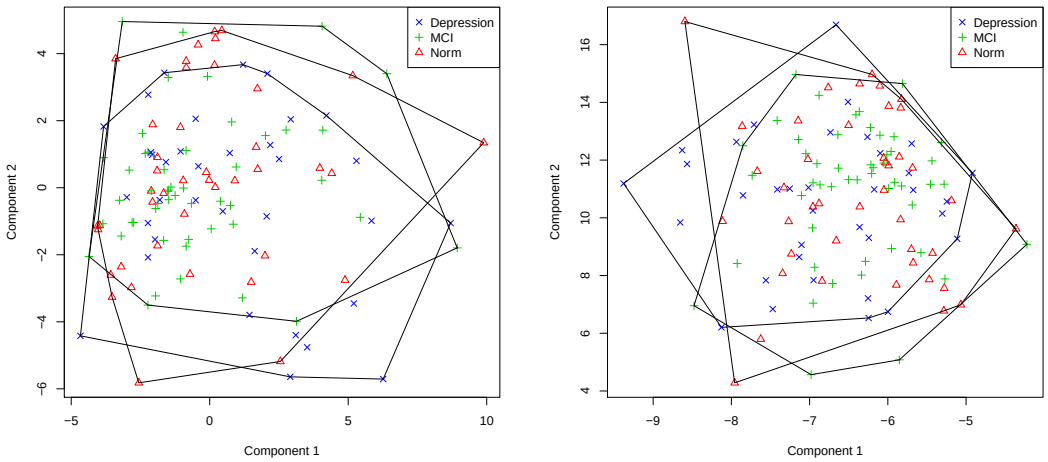
8.3 Predictive Model

In addition to the exploratory analysis, the class imbalance of the data was addressed by training 1 vs 1 models, in which each pair of diagnoses is used to train a model discriminating between them and 1 vs all models, in which each diagnosis is discriminated against all others. Adjusting the models with a-priori knowledge of the class-sizes was attempted however with poor results. A possible explanation for this is that adjusting a-priori weights balances the error between classes which is beneficial when the large classes are recognised well. However as can be seen from the tables in section A, the large classes are discriminated against others poorly.

Discrimination was performed with the following classification methods: naive Bayes, C5.0, random forest, K-nearest neighbours and support vector machines with a radial basis kernel. The results are presented in tabular form in section A. The data was used for discrimination in its original and in its semantically annotated form. Both data were transformed by reducing them to the 20 most informative variables as discussed in chapter 6, rotated along the first 20 mRRPCA directions as discussed in chapter 5 as well as along the first 20 components extracted with the HOMALS method. In addition, models were trained to discriminate each class against all others.

The results are consistent with the exploratory analysis in the previous section. Evidence for this is the performance of random forests on the mRRPCA transformed data and naive Bayes using transformed with HOMALS. In both cases, the traditionally difficult to distinguish classes Norm and Depression are discriminated against each significantly better than with other method combinations. However the best result in discriminating between patients at norm and patients with MCI or depression is achieved reducing the data with entropic variable selection (as in chapter 6) and using naive Bayes as the predictive model.

In the following three of the hypotheses are chosen for closer investigation:



(a) PCA (b) mRRPCA

Figure 8.11: Class separation along first 2 PCs for a) PCA and b) mRRPCA methods on MMSE data.

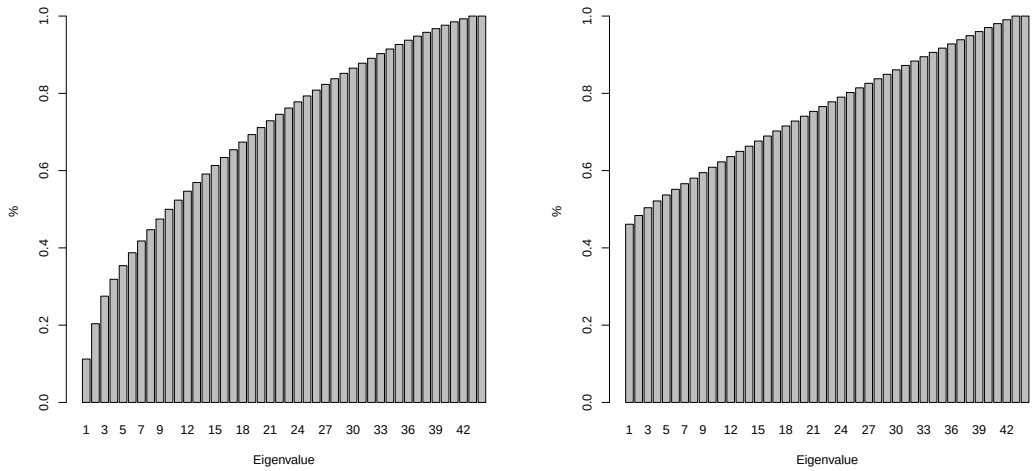
1. Patients at norm are distinguishable from patients with depression or MCI.

2. Patients with AD are distinguishable from patients with VaD.

3. Patients at norm are distinguishable from patients with Parkinson's dementia or dementia with Lewy bodies.

First, the relation of patients at norm with patients with MCI or Depression are be investigated. As a first step, a linear projection is sought, using the method proposed in chapter 5, in which the two groups of patients are separable. Figure 8.11 depicts this attempt which proved unsuccessful. Examining the cumulative variance explained by adding more principal components to the model (see figure 8.12), it is clear that the mRRPCA model retains more information from the original data in fewer components. However, the amount of explained variance is insufficient for our purposes (see section 5.2 for a detailed explanation of the graph).

The investigation is undertaken with a data set reduced with variable selection rather than rotation. This is done because in the previous chapters both methods performed comparably well however results from unrotated data are



(a) PCA (b) mRRPCA

Figure 8.12: Cumulative explained variance for a) PCA and b) mRRPCA methods on MMSE data.

	Female	Male
Depression	0.6923077	0.3076923
Norm	0.4062500	0.5937500

Table 8.2: Conditional probabilities of sex versus depression or norm.

easier to interpret. The best results in discriminating between Norm, MCI and Depression are achieved with entropic dimensionality reduction (table A.30). In the following a single model was trained with this combination which achieves the same accuracy as the one estimated with cross-validation. The MMSE questions and the sex of the patient are treated as categorical variables. The score and the semantic annotations on the other hand are treated as numeric variables.

First, the discrimination between Norm and Depression is investigated. The a-priori probability for the two classes is 0.54 and 0.45 respectively, hence the number of patients in each class is balanced. A good predictor for discrim-

	0	1
Depression	0.2820513	0.7179487
Norm	0.4687500	0.5312500

Table 8.3: Conditional probabilities of score on question 12 versus depression or norm.

	0	1
Depression	0.3076923	0.6923077
Norm	0.5625000	0.4375000

Table 8.4: Conditional probabilities of score on question 28 depression or norm.

inating between both diagnoses is the sex of a patient (see table 8.2). While only 30% of the male patients who took the test suffer from depression, 70% of the patients with this diagnosis are female. The MMSE score is not a good discriminator between the diagnoses with only half a point difference between the means of the two groups. The MMSE questions which are in this reduced data are: 1, 7, 9, 12, 14, 16, 20, 26 and 28. Question 12 is a moderate predictor (see table 8.3) and question 28 is a more decisive predictor (see table 8.4). The two questions are relatively easy and patients are expected to answer them correctly unless they are severly impaired. However, patients at norm seem to get the answers wrong more often than patients diagnosed with depression. On the other hand, patients at norm use more complicated sentence structures than patients with depression. This is concluded from the larger number of auxiliary verbs, determiners, modifiers and overall number of words in their sentences. This trend of patients at norm answering easy questions wrongly but using more complicated sentence structures is also apparent when investigating the difference between patients at norm and patients diagnosed with MCI. The only difference is that in discrimination between Norm and MCI sex does not play a significant role.

The second hypothesis postulated is: patients with AD are distinguishable from patients with VaD. The best achieved result is achieved on the data with the best selected variables using a support vector machine classifier (see table A.34). Unfortunately, the parameters of a trained SVM model are difficult to interpret. The result is very close to the one achieved with random forests leading to conclude that good discrimination can also be achieved with a tree induction approach. Some of the trees induced with random forests were investigated for information transferable into clinical practice. A single model was

trained which achieves 70% accuracy on the semantically annotated MMSE data after variable selection.

	Mean Decrease Gini
score	23.59
word_count	19.51
nouns	14.84
verbs	11.70
determiners	9.33
subjects	8.65
prepositions	8.50
adverbs	7.88
sex	7.65
auxiliary	7.01
modifiers	6.87
A1	6.01
A7	5.81
A20	5.47
A28	5.46
A14	5.14
A26	5.07
A9	4.89
A12	4.59
A16	4.54

Table 8.5: Variable importance of variables for discrimination between AD and VaD in decreased Gini coefficient.

The random forest method was parametrised to induce 500 trees from the data which makes it difficult to extract explicit rules from the model. However, the model does make it possible to assess which variables are most important in the classification of an instance. The score and word count should not be interpreted in this table as the random forest method treats these variables as categorical although they are numerical. These variables aside, the model strongly suggests that linguistic markers and the sex of a patient are more important in discerning between AD and VaD than the original MMSE questions (see table 8.5).

The third hypothesis tested in this study is the ability to distinguish between patients at norm and patients with a neuro-degenerative disease such as Parkinson's dementia or dementia with Lewey bodies. Using the C5.0 al-

adverbs	> 0			
	Q12	0	Norm	
	Q12	1	Neurodegen	
adverbs	≤ 0			
	Q20	1	Norm	
		nouns	≤ 3	Neurodegen
		nouns	> 3	Norm

Table 8.6: Decision tree which dinstiguishes between norm and neurodegenerative types of types of dementia.

gorithm, a model is constructed which separates the two groups with 60% certainty (see table 8.9). The tree is simple simple and only has 5 nodes. The variables used are the number of adverbs and nouns as well as the answers to questions 12 and 20. The question 12 is relatively simple and question 20 is considered difficult. Similarly to the discrimination between patients at norm, MCI or depression, if a patient answers an easy question wrongly but a difficult question correctly and if a patient uses more complicated sentence structure with a larger number of nouns, the patient is more likely to be at norm.

Another widely used classification method using an entropic criterion is C4.5 (Quinlan 1996). Here, an extended version, C5.0 is used to induce rules that discriminate between groups.

With the C5.0 algorithm, the groups Norm and MCI or Depression can be separated with 86.7% certainty. These rules are listed in table 8.7 (the confidence of each rule is the number in brackets next to the predicted group). An interesting result is that 97.66% of the model's accuracy is achieved by distinguishing between patients who use more than 1 auxiliary verb and patients who use 1 or no auxiliary verbs.

The second hypothesis postulated is: patients with AD are distinguishable from patients with VaD. Applying the C5.0 algorithm on the semantically annotated MMSE data returns 29 rules which together classify 88% of the data correctly. However, the number of rules is fairly large and the rules themselves depend on many variables. To reduce the complexity of the model, the 44 variables are reduced to 10 using the algorhithm proposed in chapter

Rule 1	Q28	1 not Norm (0.798)
Rule 2	auxiliary	≤ 1 not Norm (0.746)
Rule 3	Sex Q6 Q14 Q28	Male 1 1 0 Norm (0.889)
Rule 4	Q28 auxiliary	0 > 1 Norm (0.857)
Rule 5	Sex Q8 Q28	Male 1 0 Norm (0.833)
Rule 6	Sex Q6 Q30	Female 0 1 Norm (0.800)
Rule 7	Q30 determiners	1 > 2 Norm (0.750)

Table 8.7: Rules which dinstiguish between Norm and MCI or Depression with 86.7% certainty.

Rule 1	Sex adverbs	Male > 1 AD (0.900)
Rule 2	Sex word count max word length adverbs	Male > 7 > 6 > 0 AD (0.857)
Rule 3	Q3 determiners max word length adverbs	0 0 > 7 > 0 VaD (0.833)
Rule 4	determiners word count max word length adverbs	0 > 8 ≤ 7 > 0 VaD (0.833)

Table 8.8: Rules which dinstiguish between AD and VaD.

Rule 1	score	≤ 23
	Q1	1
	Q7	1
	Q26	1
	determiners	> 0
		Neurodegen (0.800)
Rule 2	score	≤ 23
	Q7	0
		Neurodegen (0.650)
Rule 3	Q1	1
	Q7	1
	Q26	0
		Norm (0.889)
Rule 4	score	> 23
		Norm (0.889)
Rule 5	Q7	1
	determiners	< 1
		Norm (0.857)
Rule 6	Q1	0
	Q7	1
		Norm (0.857)

Table 8.9: Rules which dinstiguish between norm and neurodegenerative types of types of dementia.

6. This reduces the overall accuracy to 70% but produces simpler rules which distinguish between the two patient groups with high confidence. A selection of the rules is listed in table 8.8. It is noteworthy that all of the rules strongly rely on linguistic markers to distinguish betwen patients with AD and patients with VaD.

The third hypothesis tested in this study is the ability to distinguish between patients at norm and patients with a neuro-degenerative disease such as Parkinson's dementia or dementia with Lewey bodies. Using the C5.0 algorithm, a model can be constructed which separates the two groups with 83.9% certainty (see table 8.9). The model consists of 6 rules using only 5 variables from the data: questions 1, 7 and 26 as well as the total score and the number of determiners.

8.4 Conclusion

In this chapter the structure of the MMSE data is analysed using descriptive, explorative and modelling methods. Although distinguishing MCI from Norm using the MMSE is dismissed in related literature (Mitchell 2009), a model is proposed which distinguishes between these two groups of patients with 71% accuracy (see table A.30)

A limitation of the study is that, other than the groups of patients with AD and VaD, the sample sizes are not large enough to ensure statistical stability of the proposed models. However, some rules of thumb can be derived from these seductively precise models. First, a patient who is at norm rather than with MCI or depression will answer seemingly simple questions wrongly, but difficult questions correctly. This patient is also likely to use more complex sentence structure than a patient with depression or MCI. When deciding between AD and VaD, linguistic markers are a stronger predictor than MMSE questions.

Chapter 9

Conclusions and Future Work

This chapter reviews the objectives of the reported research; the achievement of those objectives is summarised, limitations of the results are outlined and directions for future work are set.

The objectives of this thesis are driven by the need for an analysis of the MMSE by using methods appropriate for the data. This need is in part due to the lack of suitable methods. The research objectives as listed in chapter 1.2 are:

- to *reduce the dimensions* of the MMSE to the most relevant ones in order to inform a predictive model by using *computational methods* on a data set of MMSE results,
- to *construct a model* predicting a diagnosis informed by the features extracted from the previous step by applying, comparing and combining traditional and novel modelling methods,
- to propose a *semantic analysis* of the *sentence writing* question in the MMSE in order to utilise information recorded in MMS examinations which has not been considered previously.

Two general approaches to dimensionality reduction are identified: feature extraction and variable selection (see sections 2.3 and 2.4 for a review of both types of methods). Feature extraction methods are considered since the vast majority of related work conducts componential analysis of MMSE data using feature extraction methods. Variable selection methods are largely neglected in MMSE research although their output is more intuitively interpretable than the output of feature extraction methods.

Information theory is investigated closely as a foundation of an alternative method for both types of analysis because of its inherent suitability for discrete data. This investigation poses several questions: there is a research gap in methods for feature extraction from discrete data using information theory (see section 2.3.4); methods for variable selection using information theory are often designed with continuous data in mind and are often not proposed and evaluated in a way which enables an informed choice of a method (see section 2.4); methods based on information theory are computationally intensive when the data is large.

The issue of computational intensity is addressed by proposing a parallel computation method for estimating measures from information theory. The parallel algorithm, proposed in chapter 4, leverages parallel graphics hardware to explore the boundaries of computational feasibility in variable selection. A limitation of the proposed method is its requirement for discrete data with a discrete range of values. This makes the algorithm unsuitable for continuous data. Regardless of this limitation, the algorithm is applicable to various research domains and is adaptable to any methods which require an estimate of conditional mutual information (see chapter 6 for examples).

The gap in feature extraction methods using entropic criteria is addressed by proposing two methods for PCA of discrete data were proposed: MIPCA and mRRPCA. Evidence is provided for the superiority of the proposed methods over general state-of-the-art PCA methods and over entropic PCA methods specifically (see chapter 5).

A limitation of the proposed PCA method as well as the variable selection method is that the metric which is considered only takes one conditional variable into account. In chapter 4 however it is demonstrated that the parallel CMIM algorithm brings the largest benefit in models taking two or more conditional variables into account.

The third objective is achieved using the linguistic analysis method pro-

posed in chapter 7. The method allows testing more markers on a larger text corpus than is feasible by manual annotation. Fourteen linguistic markers are proposed which provide more information about the patient and which specifically help to discriminate between the two most prevalent types of dementia, Alzheimer's Disease and Vascular Dementia (see also chapter 8 for evidence). A limitation of the study is that the markers were induced from data of patients with either AD or VaD.

The objective of proposing a predictive model is met in chapter 8. The findings of the other chapters are consolidated into a predictive model which distinguishes between types of dementia. Although the profile of different types of dementia is not clearly separable, a model is proposed which distinguishes between mental states previously thought indistinguishable solely on the basis of the MMSE. The most notable separation of groups of patients (Mitchell 2009) is between patients at norm and patients with mild cognitive impairment or depression. A limitation of the study is that other than the groups of patients with AD and VaD, the sample sizes are not large enough to ensure statistical stability of the proposed models.

The contributions of this thesis are:

- A *parallel algorithm* for estimating conditional mutual information.
- A *feature extraction method* suited for discrete data.
- An *extension of a state-of-the-art variable selection method* making it applicable to larger data.
- *14 linguistic markers* which are more informative about the diagnosis of a patient than the 10 most informative MMSE questions.
- A *predictive model* for differentiating between types of dementia impairments and most notably between patients and norm and patients with depression or MCI.

9.1 Future Work

The parallel algorithm proposed in chapter 4 will be utilised to its full potential by proposing variable selection metrics with more than one conditional variable. Considering that mutual information is computed by summing terms, a starting point for such a metric could be:

$$I(Y; X_i, X_j, X_k) = I(Y; X_i) + I(Y; X_i, X_j) + I(Y; X_k|X_i, X_j) \tag{9.1}$$

where X_k is the candidate variable. Investigations are necessary to determine the amount of data required for an accurate estimate, and the number of variables with which estimating this goal function is feasible.

Measures from information theory will be considered as a replacement for the traditional maximum-likelihood estimates. The convincing results produced with mRRPCA encourage extending the method to a comprehensive framework for analysis of multi-variate discrete data by investigating the applicability of the proposed concepts to linear regression and linear discriminant analysis. A model for extending the method is provided by principal component regression (PCR). In PCR, the data is projected along the principal components, a regression model is found in the projected data and the prediction of the regression model is projected back to the original space. This approach of conducting analysis in the projected space and then returning it to the original space is a promising venue for further investigation of the mRRPCA method.

In chapter 7 linguistic markers for distinguishing between AD and VaD are identified. Further investigation is necessary to understand the information such indications represent about patients when no prior information on their potential impairment is available. One of the identified markers for distinguishing diagnoses is the number of adjectives in a sentence. This marker has not been previously reported as a discriminator between AD and VaD.

Investigation is necessary to determine whether the significance of the marker is due to the specific sample of the population or whether a change in the use of adjectives is typical for patients with AD.

Chapter 8 discusses the MMSE data using a descriptive as well as exploratory approach. The findings of these two steps are used to formulate hypotheses for constructing a predictive model. The predictive model is enriched with information from analysing a qualitative component of the MMSE in a quantitative manner. In future work the process of analysis used in devising a predictive model will be investigated for its transferrability to other domains where data similar to the MMSE is studied. Typically such data is found in psychology, sociology, but also in economy and politology.

Appendices

Appendix A

Classification Accuracy Results

A.1 Results for MMSE data without semantic annotation

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.82	0.82	0.91	0.85	0.77	0.56
Depression			0.44	0.57	0.59	0.47	0.72
MCI				0.49	0.56	0.56	0.74
Neurodegen					0.55	0.59	0.86
Norm						0.60	0.77
Other							0.69
Vascular Dementia							

Table A.1: Classification 1vs1 classification accuracy of naiveBayes on mmsepure

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.84	0.93	0.88	0.81	0.51
Depression			0.50	0.65	0.48	0.48	0.80
MCI				0.56	0.54	0.57	0.75
Neurodegen					0.52	0.70	0.89
Norm						0.58	0.82
Other							0.69
Vascular Dementia							

Table A.2: Classification 1vs1 classification accuracy of train.kknn on mmsepure

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.85	0.92	0.87	0.76	0.57
Depression			0.47	0.59	0.54	0.57	0.77
MCI				0.57	0.58	0.53	0.73
Neurodegen					0.48	0.63	0.90
Norm						0.51	0.82
Other							0.68
Vascular Dementia							

Table A.3: Classification 1vs1 classification accuracy of C5.0 on mmsepure

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.86	0.85	0.93	0.89	0.82	0.60
Depression			0.44	0.65	0.51	0.52	0.81
MCI				0.65	0.58	0.55	0.80
Neurodegen					0.55	0.73	0.90
Norm						0.62	0.85
Other							0.76
Vascular Dementia							

Table A.4: Classification 1vs1 classification accuracy of randomForest on mmsepure

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.86	0.85	0.93	0.89	0.82	0.59
Depression			0.40	0.64	0.59	0.46	0.81
MCI				0.69	0.59	0.50	0.80
Neurodegen					0.54	0.74	0.90
Norm						0.64	0.85
Other							0.76
Vascular Dementia							

Table A.5: Classification 1vs1 classification accuracy of svm on mmsepure

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.58	0.93	0.93	0.95	0.94	0.88	0.67
train.kknn	0.53	0.93	0.93	0.97	0.95	0.90	0.64
C5.0	0.53	0.94	0.93	0.97	0.95	0.91	0.63
randomForest	0.58	0.94	0.93	0.97	0.95	0.91	0.71
svm	0.60	0.94	0.93	0.97	0.95	0.91	0.71

Table A.6: Classification 1 vs all classification accuracy of mmsepure

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.82	0.82	0.87	0.86	0.82	0.58
Depression			0.48	0.68	0.49	0.57	0.77
MCI				0.65	0.58	0.55	0.72
Neurodegen					0.61	0.67	0.86
Norm						0.57	0.82
Other							0.71
Vascular Dementia							

Table A.7: Classification 1vs1 classification accuracy of naiveBayes on mmsepureinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.83	0.79	0.92	0.85	0.80	0.58
Depression			0.52	0.67	0.55	0.56	0.79
MCI				0.71	0.52	0.57	0.72
Neurodegen					0.57	0.69	0.89
Norm						0.60	0.81
Other							0.75
Vascular Dementia							

Table A.8: Classification 1vs1 classification accuracy of train.kknn on mmsepureinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.81	0.92	0.85	0.79	0.51
Depression			0.51	0.56	0.54	0.59	0.78
MCI				0.62	0.56	0.52	0.72
Neurodegen					0.57	0.62	0.89
Norm						0.53	0.82
Other							0.75
Vascular Dementia							

Table A.9: Classification 1vs1 classification accuracy of C5.0 on mmsepureinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.91	0.87	0.81	0.55
Depression			0.53	0.67	0.53	0.51	0.80
MCI				0.71	0.52	0.62	0.77
Neurodegen					0.59	0.69	0.89
Norm						0.65	0.83
Other							0.76
Vascular Dementia							

Table A.10: Classification 1vs1 classification accuracy of randomForest on mmsepureinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.87	0.82	0.57
Depression			0.43	0.70	0.55	0.56	0.81
MCI				0.73	0.60	0.58	0.78
Neurodegen					0.65	0.72	0.90
Norm						0.66	0.83
Other							0.78
Vascular Dementia							

Table A.11: Classification 1vs1 classification accuracy of svm on mmsepureinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.58	0.93	0.92	0.95	0.94	0.90	0.68
train.kknn	0.57	0.93	0.91	0.96	0.94	0.91	0.68
C5.0	0.56	0.93	0.92	0.97	0.94	0.92	0.64
randomForest	0.58	0.93	0.92	0.97	0.94	0.92	0.69
svm	0.60	0.93	0.92	0.97	0.94	0.92	0.71

Table A.12: Classification 1 vs all classification accuracy of mmsepureinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.82	0.77	0.90	0.87	0.80	0.54
Depression			0.49	0.68	0.45	0.50	0.76
MCI				0.65	0.59	0.43	0.72
Neurodegen					0.62	0.64	0.85
Norm						0.60	0.78
Other							0.71
Vascular Dementia							

Table A.13: Classification 1vs1 classification accuracy of naiveBayes on mmsepuremrrpca

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.81	0.92	0.86	0.80	0.53
Depression			0.44	0.63	0.43	0.52	0.78
MCI				0.65	0.49	0.50	0.77
Neurodegen					0.51	0.67	0.88
Norm						0.46	0.82
Other							0.70
Vascular Dementia							

Table A.14: Classification 1vs1 classification accuracy of train.kknn on mmsepuremrrpca

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.79	0.90	0.86	0.79	0.55
Depression			0.50	0.57	0.54	0.57	0.80
MCI				0.66	0.51	0.43	0.76
Neurodegen					0.53	0.71	0.88
Norm						0.53	0.79
Other							0.73
Vascular Dementia							

Table A.15: Classification 1vs1 classification accuracy of C5.0 on mmsepuremrrpca

A.2 Results for MMSE data with semantic annotation

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.81	0.92	0.87	0.82	0.55
Depression			0.48	0.66	0.44	0.57	0.82
MCI				0.64	0.52	0.48	0.77
Neurodegen					0.61	0.73	0.90
Norm						0.55	0.84
Other							0.78
Vascular Dementia							

Table A.16: Classification 1vs1 classification accuracy of randomForest on mmsepuremrrpca

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.87	0.82	0.56
Depression			0.47	0.66	0.47	0.50	0.81
MCI				0.70	0.56	0.49	0.78
Neurodegen					0.65	0.70	0.90
Norm						0.58	0.83
Other							0.78
Vascular Dementia							

Table A.17: Classification 1vs1 classification accuracy of svm on mmsepuremrrpca

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.57	0.92	0.91	0.96	0.93	0.91	0.69
train.kknn	0.52	0.93	0.92	0.97	0.94	0.91	0.65
C5.0	0.60	0.93	0.91	0.97	0.94	0.92	0.70
randomForest	0.61	0.93	0.92	0.97	0.94	0.92	0.70
svm	0.60	0.93	0.92	0.97	0.94	0.92	0.71

Table A.18: Classification 1 vs all classification accuracy of mmsepuremrrpca

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.84	0.93	0.88	0.81	0.60
Depression			0.39	0.62	0.48	0.48	0.78
MCI				0.51	0.54	0.47	0.77
Neurodegen					0.60	0.62	0.89
Norm						0.56	0.83
Other							0.73
Vascular Dementia							

Table A.19: Classification 1vs1 classification accuracy of naiveBayes on mmsepurehomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.84	0.92	0.88	0.81	0.56
Depression			0.54	0.59	0.42	0.52	0.80
MCI				0.59	0.48	0.48	0.75
Neurodegen					0.64	0.71	0.89
Norm						0.60	0.83
Other							0.68
Vascular Dementia							

Table A.20: Classification 1vs1 classification accuracy of train.kknn on mmsepurehomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.85	0.92	0.89	0.78	0.55
Depression			0.48	0.62	0.47	0.53	0.81
MCI				0.57	0.48	0.51	0.79
Neurodegen					0.47	0.71	0.89
Norm						0.63	0.83
Other							0.72
Vascular Dementia							

Table A.21: Classification 1vs1 classification accuracy of C5.0 on mmsepurehomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.86	0.85	0.93	0.89	0.82	0.60
Depression			0.43	0.59	0.54	0.51	0.81
MCI				0.59	0.54	0.48	0.80
Neurodegen					0.59	0.71	0.90
Norm						0.61	0.85
Other							0.76
Vascular Dementia							

Table A.22: Classification 1vs1 classification accuracy of randomForest on mmsepurehomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.86	0.85	0.93	0.89	0.82	0.59
Depression			0.41	0.64	0.53	0.49	0.81
MCI				0.68	0.55	0.51	0.80
Neurodegen					0.59	0.74	0.90
Norm						0.63	0.85
Other							0.76
Vascular Dementia							

Table A.23: Classification 1vs1 classification accuracy of svm on mmsepurehomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.59	0.94	0.93	0.96	0.95	0.91	0.70
train.kknn	0.55	0.93	0.93	0.97	0.95	0.91	0.65
C5.0	0.59	0.94	0.93	0.97	0.94	0.90	0.71
randomForest	0.60	0.94	0.93	0.97	0.95	0.91	0.71
svm	0.59	0.94	0.93	0.97	0.95	0.91	0.71

Table A.24: Classification 1 vs all classification accuracy of mmsepurehomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.77	0.82	0.83	0.67	0.57
Depression			0.67	0.67	0.47	0.61	0.72
MCI				0.57	0.35	0.55	0.58
Neurodegen					0.73	0.64	0.82
Norm						0.41	0.81
Other							0.69
Vascular Dementia							

Table A.25: Classification 1vs1 classification accuracy of naiveBayes on mmsesent

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.90	0.87	0.82	0.55
Depression			0.50	0.58	0.53	0.78	0.77
MCI				0.71	0.41	0.60	0.76
Neurodegen					0.45	0.71	0.90
Norm						0.65	0.79
Other							0.73
Vascular Dementia							

Table A.26: Classification 1vs1 classification accuracy of train.kknn on mmsesent

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.85	0.84	0.46
Depression			0.39	0.75	0.53	0.72	0.74
MCI				0.57	0.53	0.55	0.73
Neurodegen					0.64	0.50	0.90
Norm						0.53	0.79
Other							0.64
Vascular Dementia							

Table A.27: Classification 1vs1 classification accuracy of C5.0 on mmsesent

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.87	0.82	0.56
Depression			0.44	0.58	0.47	0.61	0.81
MCI				0.71	0.53	0.65	0.78
Neurodegen					0.55	0.71	0.90
Norm						0.47	0.81
Other							0.78
Vascular Dementia							

Table A.28: Classification 1vs1 classification accuracy of randomForest on mmsesent

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.87	0.82	0.52
Depression			0.17	0.67	0.53	0.39	0.81
MCI				0.71	0.65	0.55	0.78
Neurodegen					0.64	0.71	0.90
Norm						0.71	0.83
Other							0.78
Vascular Dementia							

Table A.29: Classification 1vs1 classification accuracy of svm on mmsesent

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.76	0.81	0.88	0.85	0.75	0.56
Depression			0.44	0.67	0.67	0.67	0.79
MCI				0.64	0.71	0.60	0.76
Neurodegen					0.45	0.71	0.77
Norm						0.47	0.74
Other							0.56
Vascular Dementia							

Table A.30: Classification 1vs1 classification accuracy of naiveBayes on mmsesentinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.90	0.81	0.75	0.57
Depression			0.56	0.67	0.40	0.44	0.74
MCI				0.71	0.53	0.60	0.71
Neurodegen					0.73	0.79	0.90
Norm						0.65	0.81
Other							0.69
Vascular Dementia							

Table A.31: Classification 1vs1 classification accuracy of train.kknn on mmsesentinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.84	0.92	0.87	0.82	0.54
Depression			0.39	0.67	0.53	0.67	0.77
MCI				0.64	0.24	0.55	0.73
Neurodegen					0.55	0.71	0.90
Norm						0.47	0.83
Other							0.67
Vascular Dementia							

Table A.32: Classification 1vs1 classification accuracy of C5.0 on mmsesentinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.84	0.82	0.92	0.87	0.81	0.59
Depression			0.44	0.75	0.40	0.50	0.81
MCI				0.64	0.59	0.40	0.78
Neurodegen					0.45	0.71	0.92
Norm						0.76	0.86
Other							0.80
Vascular Dementia							

Table A.33: Classification 1vs1 classification accuracy of randomForest on mmsesentinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.87	0.82	0.60
Depression			0.33	0.50	0.53	0.78	0.81
MCI				0.71	0.59	0.45	0.78
Neurodegen					0.73	0.79	0.90
Norm						0.59	0.83
Other							0.78
Vascular Dementia							

Table A.34: Classification 1vs1 classification accuracy of svm on mmsesentinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.82	0.82	0.88	0.85	0.74	0.56
Depression			0.61	0.58	0.67	0.56	0.65
MCI				0.71	0.53	0.60	0.73
Neurodegen					0.55	0.50	0.82
Norm						0.53	0.83
Other							0.62
Vascular Dementia							

Table A.35: Classification 1vs1 classification accuracy of naiveBayes on mmsesentmrrc

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.81	0.92	0.83	0.77	0.61
Depression			0.72	0.67	0.47	0.61	0.79
MCI				0.64	0.59	0.50	0.76
Neurodegen					0.55	0.71	0.90
Norm						0.71	0.86
Other							0.73
Vascular Dementia							

Table A.36: Classification 1vs1 classification accuracy of train.kknn on mmsesentmrrc

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.84	0.79	0.92	0.87	0.70	0.52
Depression			0.44	0.50	0.33	0.44	0.72
MCI				0.79	0.47	0.70	0.64
Neurodegen					0.45	0.86	0.90
Norm						0.71	0.83
Other							0.78
Vascular Dementia							

Table A.37: Classification 1vs1 classification accuracy of C5.0 on mmsesentmrrc

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.89	0.82	0.56
Depression			0.56	0.75	0.53	0.44	0.81
MCI				0.64	0.65	0.50	0.78
Neurodegen					0.45	0.79	0.90
Norm						0.41	0.83
Other							0.73
Vascular Dementia							

Table A.38: Classification 1vs1 classification accuracy of randomForest on mmsesentmrrc

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.87	0.82	0.59
Depression			0.50	0.75	0.47	0.67	0.81
MCI				0.71	0.53	0.50	0.78
Neurodegen					0.55	0.71	0.90
Norm						0.65	0.83
Other							0.78
Vascular Dementia							

Table A.39: Classification 1vs1 classification accuracy of svm on mmsesentmrrc

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.74	0.92	0.85	0.77	0.56
Depression			0.39	0.67	0.40	0.67	0.74
MCI				0.57	0.53	0.50	0.58
Neurodegen					0.45	0.71	0.90
Norm						0.65	0.83
Other							0.60
Vascular Dementia							

Table A.40: Classification 1vs1 classification accuracy of naiveBayes on mmsesenthomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.84	0.82	0.90	0.87	0.81	0.40
Depression			0.61	0.67	0.47	0.67	0.77
MCI				0.50	0.35	0.65	0.78
Neurodegen					0.36	0.71	0.92
Norm						0.71	0.81
Other							0.76
Vascular Dementia							

Table A.41: Classification 1vs1 classification accuracy of train.kknn on mmsesenthomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.80	0.87	0.82	0.56
Depression			0.44	0.58	0.13	0.67	0.81
MCI				0.64	0.53	0.55	0.78
Neurodegen					0.64	0.71	0.82
Norm						0.47	0.79
Other							0.78
Vascular Dementia							

Table A.42: Classification 1vs1 classification accuracy of C5.0 on mmsesenthomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.84	0.81	0.94	0.87	0.81	0.48
Depression			0.56	0.58	0.53	0.50	0.81
MCI				0.71	0.59	0.50	0.78
Neurodegen					0.36	0.71	0.87
Norm						0.65	0.83
Other							0.78
Vascular Dementia							

Table A.43: Classification 1vs1 classification accuracy of randomForest on mmsesenthomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
Alzheimers Disease		0.85	0.82	0.92	0.87	0.82	0.51
Depression			0.44	0.67	0.40	0.61	0.81
MCI				0.71	0.53	0.50	0.78
Neurodegen					0.64	0.71	0.90
Norm						0.59	0.83
Other							0.78
Vascular Dementia							

Table A.44: Classification 1vs1 classification accuracy of svm on mmsesenthomals

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.59	0.88	0.86	0.93	0.86	0.87	0.59
train.kknn	0.64	0.93	0.91	0.97	0.94	0.90	0.61
C5.0	0.55	0.93	0.92	0.97	0.94	0.92	0.65
randomForest	0.64	0.93	0.92	0.97	0.94	0.92	0.71
svm	0.60	0.93	0.92	0.97	0.94	0.92	0.71

Table A.45: Classification 1 vs all classification accuracy of mmsesent

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.60	0.91	0.91	0.94	0.92	0.89	0.62
train.kknn	0.64	0.93	0.91	0.97	0.93	0.91	0.64
C5.0	0.58	0.93	0.92	0.97	0.94	0.92	0.67
randomForest	0.64	0.93	0.92	0.97	0.94	0.91	0.69
svm	0.63	0.93	0.92	0.97	0.94	0.92	0.71

Table A.46: Classification 1 vs all classification accuracy of mmsesentinfgain

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.57	0.90	0.85	0.95	0.89	0.83	0.70
train.kknn	0.55	0.93	0.92	0.97	0.93	0.89	0.74
C5.0	0.59	0.93	0.92	0.97	0.94	0.92	0.63
randomForest	0.65	0.93	0.92	0.97	0.94	0.92	0.71
svm	0.62	0.93	0.92	0.97	0.94	0.92	0.71

Table A.47: Classification 1 vs all classification accuracy of mmsesentmrrc

	Alzheimers.Disease	Depression	MCI	Neurodegen	Norm	Other	Vascular.Dementia
naiveBayes	0.50	0.93	0.92	0.96	0.91	0.91	0.60
train.kknn	0.64	0.93	0.92	0.97	0.94	0.91	0.60
C5.0	0.61	0.93	0.92	0.97	0.94	0.92	0.71
randomForest	0.64	0.93	0.92	0.97	0.94	0.92	0.71
svm	0.59	0.93	0.92	0.97	0.94	0.92	0.71

Table A.48: Classification 1 vs all classification accuracy of mmsesenthomals

References

Ahmed, Samrah, Celeste A. de Jager, Anne-Marie Haigh and Peter Garrard. 2013. "Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease." *Neuropsychology* 27(1):79–85.

Akadi, Ali El, Abdeljalil El Ouardighi and Driss Aboutajdine. 2008. "A Powerful Feature Selection approach based on Mutual Information." *International Journal of Computer Science and Network Security* 8:116–121.

Bain, M. and S. Muggleton. 1995. Learning optimal chess strategies. In *Machine intelligence 13*. Oxford University Press, Inc. pp. 291–309.

Baos, James H. and Lucy M. Franklin. 2002. "Factor Structure of the Mini-Mental State Examination in Adult Psychiatric Inpatients." *Psychological Assessment* 14(4):397–400.

Battiti, Roberto. 1994. "Using mutual information for selecting features in supervised neural net learning." *IEEE Transactions on Neural Networks* 5:537–550.

Bencini, Giulia M.L., Lucia Pozzan, Roberta Biundo, William J. McGeown, Virginia V. Valian, Annalena Venneri and Carlo Semenza. 2011. "Language-specific effects in Alzheimers disease: Subject omission in Italian and English." *Journal of Neurolinguistics* 24(1):25 – 40.

Bollacker, Kurt D. and Joydeep Ghosh. 1996. Linear Feature Extractors Based on Mutual Information. In *In Proceedings of the 13th International Conference on Pattern Recognition*. pp. 720–724.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Brugnolo, A., F. Nobili, M. P. Barbieri, B. Dessi, A. Ferro, N. Girtler, E. Palummeri, D. Partinico, U. Raiteri, G. Regesta, G. Servetto, P. Tan-

ganelli, V. Uva, D. Mazzei, S. Donadio, F. De Carli, G. Colazzo, C. Serrati and G. Rodriguez. 2009. "The factorial structure of the mini mental state examination (MMSE) in Alzheimer's disease." 49(1):180–185.

Bush, Clarissa, Jean Kozak and Tom Elmslie. 1997. "Screening for cognitive impairment in the elderly." *Canadian Family Physician* 43:1763–1863.

Castro-Costa, Erico, Cintia Fuzikawa, Cleusa Ferri, Elizabeth Uchoa, Joselia Firmo, Maria Fernanda Lima-Costa, Michael E. Dewey and Robert Stewart. 2009. "Dimensions Underlying the Mini-Mental State Examination in a Sample With Low-Education Levels: The Bambui Health and Aging Study." *American Journal of Geriatric Psych* 17(10):863–872.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of information theory.* New York, NY, USA: Wiley-Interscience.

Croux, C., P. Filzmoser and M.R. Oliveira. 2007. "Algorithms for ProjectionPursuit robust principal component analysis." *Chemometrics and Intelligent Laboratory Systems* 87(2):218 – 225.
URL: `http://www.sciencedirect.com/science/article/pii/S016974390700007X`

Crum, Rosa M., James C. Anthony, Susan S. Bassett and Marshal F. Folstein. 1993. "Population-Based Norms for the Mini-Mental State Examination by Age and Educational Level." 269(18):2386–2391.

De Leeuw, Jan and Patrick Mair. 2009. "Gifi Methods for Optimal Scaling in R: The Package homals." *Journal of Statistical Software, forthcoming* pp. 1–30.

Deleon, J, B Gesierich, M Besbris, J Ogar, ML Henry, BL Miller, ML Gorno-Tempini and SM Wilson. 2012. "Elicitation of specific syntactic structures in primary progressive aphasia." *Brain and Language* 123(3):183–90.

Duch, W, R Adamczak, K Grabczewski, M Ishikawa and H Ueda. 1997. Extraction of crisp logical rules using constrained backpropagation networks - comparison of two new approaches. In *Proceedings of the European Symposium on Artificial Neural Networks*. Bruge, Belgium: pp. 16–18.

Estévez, Pablo A., Michel Tesmer, Claudio A. Perez and Jacek M. Zurada. 2009. "Normalized mutual information feature selection." *Trans. Neur. Netw.* 20:189–201.

Fleuret, Francois. 2004. "Fast Binary Feature Selection with Conditional Mutual Information." *Journal of Machine Learning Research* 5:1531–1555.

Folstein, Marshal, Susan Folstein and Paul McHugh. 1975. "Mini-mental State: A practical method for grading the cognitive state of patients for the clinician." *Journal of Psychiatric Research* 12(3):189 – 198.

Gifi, Albert. 1990. *Nonlinear Multivariate Analysis*. Wiley-Blackwell.

Gross, Rachel G., Corey T. McMillan, Keerthi Chandrasekaran, Michael Dreyfuss, Sharon Ash, Brian Avants, Philip Cook, Peachie Moore, David J. Libon, Andrew Siderowf and Murray Grossman. 2012. "Sentence processing in Lewy body spectrum disorder: The role of working memory." *Brain and Cognition* 78(2):85 – 93.

Guerrero-Berroa, E., X. Luo, J. Schmeidler, M. A. Rapp, K. Dahlman, H. T. Grossman, V. Haroutunian and M. S. Beeri. 2009. "The MMSE orientation for time domain is a strong predictor of subsequent cognitive decline in the elderly." *International Journal of Geriatric Psychiatry* 24(12):1429–1437.

Güvenir, H. Altay, Gülsen Demiröz and Nilsel Ilter. 1998. "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals." *Artificial Intelligence in Medicine* 13(3):147–165.

Guyon, Isabelle, Steve Gunn, Masoud Nikravesh and Lotfi A. Zadeh. 2006. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc.

Haubois, Gladys, Cdric Annweiler, Cyrille Launay, Bruno Fantino, Laure de Decker, Gilles Allali and Olivier Beauchet. 2011. "Development of a short form of Mini-Mental State Examination for the screening of dementia in older adults with a memory complaint: a case control study." *BMC Geriatrics* 11(1):59.

He, Ran, Baogang Hu, XiaoTong Yuan and Wei-Shi Zheng. 2010. "Principal component analysis based on non-parametric maximum entropy." *Neurocomput.* 73(10-12):1840–1852.

Heafield, Kenneth. 2005. Detecting Network Anomalies With Kernel Principal Component Analysis. Technical report Caltech, Netlab.

Hild, K.E., D. Erdogmus, K. Torkkola and J.C. Principe. 2006. "Feature extraction using information-theoretic learning." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28(9):1385–1392.

Hill, Robert and Lars Backman. 1995. "The Relationship between the Mini-Mental State Examination and Cognitive Functioning in Normal Elderly Adults: A Componential Analysis." *Age Ageing* 24(5):440–446.

Huang, Jinjie, Yunze Cai and Xiaoming Xu. 2007. "A hybrid genetic algorithm for feature selection wrapper based on mutual information." *Pattern Recogn. Lett.* 28:1825–1844.

Hyvrinen, Aapo, Juha Karhunen and Erkki Oja. 2001. *Independent Component Ana.* Wiley.

Ismail, Z., T. Rajji and K. Shulman. 2010. "Brief cognitive screening instruments: an update." *International Journal of Geriatric Psychiatry* 25(2):111–120.

Jakulin, Aleks and Ivan Bratko. 2004. Testing the Significance of Attribute Interactions. In *In Proc. of 21st International Conference on Machine Learning (ICML)*. ACM Press pp. 409–416.

Japkowicz, Nathalie and Shaju Stephen. 2002. "The class imbalance problem: A systematic study." *Intell. Data Anal.* 6:429–449.

Jefferson, Angela L., Stephanie A. Cosentino, Susan K. Ball, Bruce Bogdanoff, Norman Leopold, Edith Kaplan and David J. Libon. 2002. "Errors Produced on the Mini-Mental State Examination and Neuropsychological Test Performance in Alzheimer's Disease, Ischemic Vascular Dementia, and Parkinson's Disease." *The Journal of Neuropsychiatry & Clinical Neurosciences* 14(3):311–320.

Jolliffe, I. T. 2002. *Principal Component Analysis*. Springer.

Jung, Chi-Sang, Hyunson Seo and Hong-Goo Kang. 2011. "Estimating redundancy information of selected features in multi-dimensional pattern classification." *Pattern Recogn. Lett.* 32(4):590–596.

Kempler, DANIEL and ELIZABETH M Zelinski. 1994. "Language in dementia and normal aging." *Dementia and normal aging* pp. 331–365.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. ACL '03 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 423–430.

Knapp, Martin and Martin Prince. 2005. *Dementia UK - The Full Report*. Alzheimer's Society.

Kolenikov, Stanislav and Gustavo Angeles. 2004. The Use of Discrete Data in Principal Component Analysis: Theory, Simulations, and Applications to Socioeconomic Indices. In *Proceedings of the American Statistical Association*.

Kononenko, Igor. 1995. On Biases in Estimating Multi-Valued Attributes. Morgan Kaufmann pp. 1034–1040.

Kurgan, L. A., K. J. Cios, R. Tadeusiewicz, M. Ogiela and L. S. Goodenday. 2001. "Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis." *Artificial Intelligence in Medicine* 23(2):149–169.

Kwak, N. and Chong-Ho Choi. 2002. "Input feature selection by mutual information based on Parzen window." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24(12):1667 – 1671.

Lakey, Louise, Karishma Chandaria, Chris Quince, Martina Kane and Tess Saunders. 2012. *Dementia 2012: A national challenge*. Alzheimer's Society.

Liu, Huawen, Jigui Sun, Lei Liu and Huijie Zhang. 2009. "Feature selection with dynamic mutual information." *Pattern Recogn.* 42:1330–1339.

Magni, Eugenio, Giuliano Binetti, Alessandro Padovani, Stefano F. Cappa, Angelo Bianchetti and Marco Trabucchi. 1996. "The Mini-Mental State Examination in Alzheimer's Disease and Multi-Infarct Dementia." *International Psychogeriatrics* 8(01):127–134.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. 1 ed. Cambridge University Press.

Martínez Sotoca, José and Filiberto Pla. 2010. "Supervised feature selection by clustering using conditional mutual information-based distances." *Pattern Recogn.* 43(6):2068–2081.

Merrill, Duane and Andrew S. Grimshaw. 2011. "High Performance and Scalable Radix Sorting: a Case Study of Implementing Dynamic Parallelism for GPU Computing." *Parallel Processing Letters* 21(2):245–272.

Meyer, Patrick Emmanual. 2008. Information-theoretic variable selection and network inference from microarray data PhD thesis Universit Libre de Bruxelles.

Mitchell, Alex J. 2009. "A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment." 43(4):411–431–.

National Dementia Plan for Wales. 2009. Welsh Assembly Government.

Newman, John and Robin Feldman. 2011. "Copyright and Open Access at the Bedside." *The New England Journal of Medicine* 365(26):2447–2449.

Nicholl, Claire. 2009. "Diagnosis of dementia." 338:b1176–.

Nijenhuis, Albert and Herbert S. Will. 1978. *Combinatorial Algorithms: For Computers and Hard Calculators*. 2nd ed. Orlando, FL, USA: Academic Press, Inc.

Noordewier, Michiel O., Geoffrey G. Towell and Jude W. Shavlik. 1990. Training knowledge-based neural networks to recognize genes in DNA sequences. In *Proceedings of the 1990 conference on Advances in neural information processing systems 3*. NIPS-3 San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 530–536.

O'Bryant, Sid E., Joy D. Humphreys, Glenn E. Smith, Robert J. Ivnik, Neill R. Graff-Radford, Ronald C. Petersen and John A. Lucas. 2008. "Detecting Dementia With the Mini-Mental State Examination in Highly Educated Individuals." *Arch Neurol* 65(7):963–967.

Ott, Alewijn, Monique M B Breteler, Frans van Harskamp, Jules J Claus, Tischa J M van der Cammen, Diederick E Grobbee and Albert Hofman. 1995. "Prevalence of Alzheimer's disease and vascular dementia: association with education. The Rotterdam study." 310:970–973–.

Palethorpe, A J. 2009. *Impact Assessment of National Dementia Strategy.* Department of Health.

Peng, Hanchuan, Fuhui Long and Chris H. Q. Ding. 2005. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Trans. Pattern Anal. Mach. Intell.* 27(8):1226–1238.

Press, Yan Y, Natalia N Velikiy, Alex A Berzak, Howard H Tandeter, Roni R Peleg, Tamar T Freud, Boris B Punchik and Tzvi T Dwolatzky. 2012. "A retrospective analysis of the sentence writing component of the mini mental state examination: cognitive and affective aspects." *Dementia and Geriatric Cognitive Disorders* 33(2-3):125–31.

Qiu, Xipeng and Lide Wu. 2009. "Info-margin maximization for feature extraction." *Pattern Recognition Letters* 30(16):1516 – 1522.

Quinlan, J. R. 1996. Bagging, Boosting, and C4.5. In *In Proceedings of the Thirteenth National Conference on Artificial Intelligence.* AAAI Press pp. 725–730.

Quinlan, J. Ross. 1993. *C4.5: programs for machine learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Reif, John H. 1993. *Synthesis of Parallel Algorithms.* 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. chapter Prefix Sums and Their Applications, pp. 35–60.

Roark, B., M. Mitchell, J. Hosom, K. Hollingshead and J. Kaye. 2011. "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment." *Audio, Speech, and Language Processing, IEEE Transactions on* 19(7):2081–2090.

Robnik-ikonja, Marko. 2003. Experiments with Cost-Sensitive Feature Evaluation. In *Machine Learning: ECML 2003*. Vol. 2837 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 325–336.

Robnik-ikonja, Marko and Igor Kononenko. 2003. "Theoretical and Empirical Analysis of ReliefF and RReliefF." *Machine Learning* 53(1-2):23–69.

Rodriguez, J.J., L.I. Kuncheva and C.J. Alonso. 2006. "Rotation Forest: A New Classifier Ensemble Method." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28(10):1619–1630.

Rosipal, R., M. Girolami, L.K. Trejo and A. Cichocki. 2001. "Kernel PCA for Feature Extraction and De-Noising in Nonlinear Regression." *Neural Computing & Applications* 10(3):231–243.

S, Pakhomov, Chacon D, Wicklund M and Gundel J. 2011. "Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing." *Behavior Research Methods* 43(1):136–144.

Satish, Nadathur, Mark Harris and Michael Garland. 2009. "Designing efficient sorting algorithms for manycore GPUs." *Parallel and Distributed Processing Symposium, International* 0:1–10.

Schlimmer, Jeffrey Curtis. 1987. Concept acquisition through representational adjustment PhD thesis University of California, Irvine. AAI8724747.

Schlkopf, Bernhard, Alexander J. Smola and Klaus R. Müller. 1999. "Kernel principal component analysis." *Advances in kernel methods: support vector learning* pp. 327–352.

Shenkin, S. D., J. M. Starr, J. M. Dunn, S. Carter and I. J. Deary. 2008. "Is there information contained within the sentence-writing component of the mini mental state examination? A retrospective study of community dwelling older people." *International Journal of Geriatric Psychiatry* 23(12):1283–1289.

Shifrin, Theodore and Malcolm R. Adams. 2011. *Linear Algebra – A Geometrical Approach*. W.H. Freeman and Company.

Tan, Ming and Larry J. Eshelman. 1988. Using Weighted Networks to Represent Classification Knowledge in Noisy Domains. In *ML*. pp. 121–134.

Tinklenberg, Jared, John O. Brooks, Elizabeth Decker Tanke, Kausar Khalid, Sarah L. Poulsen, A. B. Helena Chmura Kraemer, Dolores Galagher, Joe E. Thornton and Jerome A. Yesavage. 1990. "Factor Analysis and Preliminary Validation of the Mini-Mental State Examination from a Longitudinal Perspective." *International Psy* 2:123–134.

Torkkola, Kari. 2003. "Feature extraction by non parametric mutual information maximization." *J. Mach. Learn. Res.* 3:1415–1438.

Tsimpiris, Alkiviadis, Ioannis Vlachos and Dimitris Kugiumtzis. 2012. "Nearest neighbor estimate of conditional mutual information in feature selection." *Expert Syst. Appl.* 39(16):12697–12708.

Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley.

Twining, C.J. and C.J. Taylor. 2003. "The use of kernel principal component analysis to model data distributions." *Pattern Recognition* 36(1):217 – 227.

Van Dijck, Gert and Marc M. Van Hulle. 2010. "Increasing and Decreasing Returns and Losses in Mutual Information Feature Subset Selection." *Entropy* 12(10):2144–2170.

Varmuza, Kurt, Peter Filzmoser and Bettina Liebmann. 2010. "Random projection experiments with chemometric data." *Journal of Chemometrics* 24(3-4):209–217.
URL: `http://dx.doi.org/10.1002/cem.1295`

Vigliocco, Gabriella, David P. Vinson, Judit Druks, Horacio Barber and Stefano F. Cappa. 2011. "Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies." *Neuroscience & Biobehavioral Reviews* 35(3):407 – 426.

Wind, Annet W., Fran?ois G. Schellevis, Gerrit van Staveren, Rob J. P. M. Scholten, Cees Jonker and Jacques Th. M. van Eijk. 1997. "Limitations of the Mini-Mental State Examination in Diagnosing Dementia in General Practice." 12:101–108–.

Wu, Dekai, Weifeng Su and Marine Carpuat. 2004. A Kernel PCA Method for Superior Word Sense Disambiguation. In *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.