Stony Brook University

Clustering and Classification Methods for Prediction of the risk for Developing Disease

A Dissertation Presented

by

Hyejoo Lee

to

The Graduate School

in Partial Fulillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

May 2016

Stony Brook University

The Graduate School

Hyejoo Lee

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dr.Hongshik Ahn - Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Dr.Stephen Finch – Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Dr.Haipeng Xing - Member
Associate Professor, Department of Applied Mathematics and Statistics

Dr.Sangjin Hong – Outside Member
Professor, Department of Electrical and Computer Engineering

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Clustering and Classification Methods for Prediction of the risk for Developing Disease

by

Hyejoo Lee

Doctor of Philosophy

In Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2016

The purpose of this study is to develop a statistical model to predict the risk for developing disease. In order to enrich our general understanding of schizophrenia disorder, several clustering techniques are used as a preliminary study. Schizophrenia is a heterogeneous decease with great variability in symptoms, cognition, biology and course of illness. Some of this variability may be explained by latent subgroups that differ in etiology and key features. Individuals with paternal age related schizophrenia (PARS) may represent such a subgroup as evidence suggests a distinct

symptom profile. Using K-means and hierarchical clustering on a large sample of schizophrenia patients, this study examines demographic, clinical and the distinctiveness of latent PARS subgroups.

Despite the wide use of K-means clustering, there remain several issues about how best to implement it. One of the main problems in K-means clustering is how to determine the number of clusters in a data set. We propose to develop a method for choosing the optimal number of clusters. The performance of the proposed method is compared to other existing methods by simulation experiments. In this study, the performance of several classification models with the same schizophrenia data set is evaluated. Four predictive classification models including Random Forest (RF), Support Vector Machines (SVM), Linear Discriminant Analysis and Adaboost are trained and their performances are compared. These models are then used to predict a patient who might have more risk of developing schizophrenia. For RF and SVM, adjusted decision threshold is used for a fair comparison.

One of the most critical factors in medical diagnosis is individual's condition to a given disease which varies from one to another. It is difficult to make appropriate medical decision about treatment that works on every patient. This study focuses on to develop a statistical method to classify the data into these two groups: ones that have a risk at potential disease and others who don't. The successful completion of

this study will lead to dramatic improvement in the medical diagnosis which will help the development of decision support system and personalized treatments that focus on specific patient needs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistical pattern recognition or machine learning has been an active research area that applies statistical techniques for studying pattern and regularity of data. There are two main branches of pattern recognition: supervised learning and unsupervised learning approaches. The former is known as classification.

Given a set of data, unsupervised learning analyzes the data where there is no label. The main purpose of unsupervised learning is to find some intrinsic structure or particular input patterns in the data. Clustering, density estimation and learning latent variable models are common unsupervised learning methods.

Supervised learning searches an inferred function, which is called the classifier, from the training data. The classifier is a learning mapping function between input variables and a labeled output variable. The resulting classifier is then used to classify or predict the output for other unlabeled data. There are several approaches for supervised learning such as k-nearest neighbor (k-NN), logistic regression, decision trees, support vector machines (SVM), artificial neural networks (ANN), bagging, boosting and random forest (RF).

Cluster analysis seeks to partition a set of observations into subsets or clusters, so that the objects within each cluster are more closely related to one another and different from the objects in other groups. In general, the clustering algorithms are classified into two types: hierarchical (tree based method) and non-hierarchical (partitioning method) algorithms. To determine appropriate clustering algorithms is a critical factor to the effective use of cluster analysis.

K-means clustering is a partitioning method often used in data mining and machine learning (Huang, 1998; Wagstaff et al., 2001). It aims to partition by minimizing the average squared distance between *n* observations and a cluster centroid, such that each observation is assigned to the cluster with the nearest mean (Hand and Heard, 2005).

Hierarchical clustering techniques are also widely used to find patterns in multi-dimensional data sets. There are two basic strategies in hierarchical clustering, agglomerative and divisive. Agglomerative clustering begins with every observation as an individual cluster in which each step it builds a tree-like structure that merges the closest pair of clusters until only one cluster remains. This requires a choice of a proximity measure. Most statistical packages use an agglomerative method and the most popular agglomerative methods are (1) single linkage (nearest neighbor approach), (2) complete linkage (furthest neighbor), (3) average linkage, (4) Ward's method, and (5) Centroid method (Legendre and

Legendre, 1998). All these methods differ in the definition of the distance measure. Most of the time, the distance is based on Euclidean distance in the sample axes. Although numerous approaches have been developed to decide the number of clusters in a data set, there is no widely used standard criterion. One of the simplest techniques is the elbow method (Thorndike, 1953) which involves graphing the percentage of variance explained on the y-axis and the number of clusters on the x-axis. There have also been several approaches to choosing the number of groups including Gaussian-model-based approaches using an information criterion by Akaike (1974), the Bayesian Information Criterion (BIC) proposed by Schwartz (1978), or the Deviance information criterion (DIC) introduced by Spiegelhalter (2002). Those methods tend to be model based and hence require strong parametric assumptions (Catherine and Gareth, 2003). There is another nonparametric method for determining the number of clusters based on distortion theory which is called the jump method. The distortion is defined as the variance of the distance measures within clusters. This is also known as the average Mahalanobis distance per dimension. The distortion curve is generated by plotting the computed distortion versus K. The largest jump in this plot indicates the optimal number of clusters.

Besides the methods that we mentioned, other widely used methods will be introduced and compared. In this study, improvement of elbow method for choosing the optimal number of clusters is proposed, and the performance of the

new method is evaluated. It was shown that the proposed method performs better than other methods in the simulation studies.

Pattern analysis has been used to improve performance in many fields of practical applications including, but not limited to, the financial industry, medical science, computer science and engineering. In this study, a number of pattern analysis techniques are applied to schizophrenia patient data through the psychiatric clinical assessment. In order to explain variability in schizophrenia symptoms by a distinct subgroup, we used clustering technique which is one of the prominent unsupervised learning approaches. Separately, another analysis is conducted with the same data set using several widely used supervised learning approaches. Several models are examined and their performances are compared. The aim is identifying people at risk of developing schizophrenia.

Schizophrenia is characterized by significant heterogeneity in symptoms, course of illness, and clinical profiles (Tsuang et al., 1990). This heterogeneity complicates the interpretation of research findings and inhibits the discovery of novel treatments for the disorder. Some of the variability in symptoms and illness features among schizophrenia patients may be explained by the presence of latent subgroups that differ in etiology and key neurobiological underpinnings. Identifying these subgroups is important to set the stage for targeted person-specific pharmacological and/or psychological treatments (Jindal et al., 2005).

Advanced paternal age has been associated with the risk for schizophrenia in cohort studies in Israel (Malaspina et al., 2001; Brown et al., 2002), Denmark, (Byrne et al., 2003), Sweden (Zammit et al., 2003; Sipos et al., 2004), Japan (Tsuchiya et al., 2005), and the United States (Torrey et al., 2009). In the Israeli study, a quarter of the risk for schizophrenia was attributable to paternal age and the risk in offspring of fathers aged over 50 at birth was three-fold that of children whose fathers were younger than 25 at birth (Malaspina et al., 2001). Clinical studies have suggested that paternal-age-related schizophrenia (PARS) may be a specific variant of the disease, as symptom and cognitive profiles, regional cerebral metabolism, sex effects, and heart rate variability have been shown to differ from those of other cases (Malaspina, 2001; Malaspina et al., 2001; 2002a; 2005; Rosenfield et al., 2010; Antonius et al., 2011). If these studies are confirmed, then PARS may account for a substantial portion of the disease in clinical treatment.

Currently, however, it is not clear whether PARS explains any of the heterogeneity of schizophrenia. To explore this, we have chosen to use an approach based on clustering analysis in order to generate new hypotheses related to PARS. Two different types of cluster algorithms have been used in our analysis: *K-means* and hierarchical clustering.

In the second study, we applied various supervised learning models in order to classify data into two or more groups as needed for practical applications in schizophrenia research. Several studies have attempted to apply machine learning models to healthcare data in order to predict common disease risks. One of these studies indicated that data mining approach is highly effective for predicting patients who are likely to be at high-risk in the future (Moturu et al., 2007). They used AdaBoost, LogitBoost, Logistic Regression, Logistic Model Trees and SVM. SVM was used to detect a person with diabetes and pre-diabetes (Yu et al., 2010). In cancer research, machine learning approaches have been used (Zhang et al., 2009). They compared ensemble learning approaches with other classification methods in the classification of breast cancer metastasis. Another recent research suggested that machine learning techniques can potentially identify patients at high risk (Wu et al., 2010). They compared the performance of Boosting, SVM and logistic regression for predicting heart failure more than 6 months before clinical diagnosis.

Classification methods are widely used in many applications like risk management, medical diagnosis, decision making and the area of marketing and sales. In statistics, classification is a procedure in which individual items are placed into groups based on quantitative information about one or more characteristics inherent in the items and based on a training set of previously labeled items. These classification tools are supervised learning methods where

the algorithm learns from a training set and establishes a prediction rule to classify new samples using statistical approaches for class prediction.

Suppose that we have clinical characteristics of a patient. Classification methods can make a medical diagnostic decision based on the clinical data and enable appropriate medical treatment for the patient. Various classification models were applied and their performance was compared. A successful completion of this study will lead to improvement in prediction of potential schizophrenia patients among seemingly healthy subjects. There are various algorithms of classification and some of widely used methods are summarized below.

In classification, the *k*-nearest neighbor (*k*-NN) algorithm, originally proposed by Fix and Hodges is one of the fundamental and simplest non-parametric techniques (Fix and Hodges, 1952). The *k*-NN classifies the dataset by finding *k* nearest neighbors from the training dataset. *k* is a user defined constant, typically a small positive integer. For example, in the case of k=1, any training data point is simply assigned to the single class of its nearest neighbor. The value of *k* is found by performing cross-validation. To break ties it is best to use an odd value of *k*. The closest neighbor is defined in terms of a distance measure such as the Euclidean Distance measure. Classification of the given object is made by taking majority voting of its neighbors. Common drawback of majority vote

might be occurred when the distribution of the data set is skewed. One way to overcome the drawback is to assign weights according to relevance of a given object.

Linear Discriminant Analysis (LDA) is a well-known classification technique that separates samples of distinct groups. LDA was formulated by R.A. Fisher (Fisher, 1936). The primary purpose of LDA is to reduce the dimension of dataset by projecting the samples onto a lower dimensional space. LDA is closely related to PCA (Principal Component Analysis). Both algorithms are linear techniques, which reduce the dimension of the data. Unlike PCA which explains the uncertainties or variations of data, LDA tries to separate classes. The LDA algorithm is seeking a linear combination of the variables that maximizes the ratio of the between-class variance to the within-class variance. For example, if the feature space has two dimensions, LDA generates a projection to a line such that classes are well separated. If there are three features, the separator will be a plane. When the number of features is more than three, the separator becomes a hyperplane. However, its main limitation is that LDA assumes normally distributed data. If the distributions are not Gaussian, LDA may not work well as a classifier. Rao extended standard LDA to multiclass LDA in the case of more than two predetermined classes (Rao, 1948). Rao used unweighted covariance matrix of group means for the best separation of groups in multivariate space. Furthermore, a number of alternative LDA techniques have been proposed in

order to overcome the limitations of heteroscedastic data such as NDA, aPAC and minimum Bayes error method.

McCulloch and Pitts introduced the first conceptual model of artificial neural networks (ANN) which is inspired by a biological neural system (McCulloch and Pitts, 1943). ANN is typically organized in different layers. The first layer is made up of hundreds of single input neurons or nodes. Each input layer is transformed to hidden layers via weighted connections. A complex ANN system may have several hidden layers. These hidden layers are linked to the output layers. If there is no hidden layer in the network, ANN reduces to a linear regression model. If there are one or more hidden layers in the network, then ANN is a non-linear generalization of the linear regression model. An ANN is represented as an adaptive learning system. Given the inputs, ANN learns by changing its weight to produce required outputs. ANN has been applied to various research areas that deal with complexity of data such as classification, function approximation, robotics and data processing.

Logistic regression is a special type of ordinary linear regression where the dependent variable is binary (Cox, 1958). Since the dependent variable is binary, logistic regression assumes a Bernoulli distribution of data. In logistic regression, the probability or odds of the outcome is predicted The range of the outcome measure is bounded between 0 and 1. In order to conduct linear regression

approach, logistic regression transforms information of the binary outcome variable to an unbounded continuous variable through the logit transformation, which is called the link function or the sigmoid function. To learn the parameters, the logistic regression model uses maximum likelihood estimation. The Bernoulli distribution is a member the exponential family of probability distributions, and maximum likelihood estimates are typically derived from Newton-Raphson method. After a logistic regression model has been fitted, a global test of goodness of fit of the model is conducted by using the deviance or pseudo-R^2. The significance of each coefficient can be tested by Wald or likelihood ratio test. It is used mainly for binary responses, although there are extensions for multinomial responses as well. Logistic regression or logit regression was originally developed by Cox.

The Support Vector Machine (SVM: Cortes and Vapnik, 1995) is a widely used classification model in data mining and pattern recognition. The standard SVM is a binary classifier which does not directly provide probability estimates. SVM provides a classification mechanism based on finding a hyperplane which divides the data space with a maximum margin (the distance between the hyperplane and the nearest point). The data points holding this hyperplane are defined as support vectors. This is known as the canonical hyperplane. If such a maximal margin hyperplane exists, the linear classifier is defined as the maximum margin classifier, which is also known as perceptron. However, there might not

exist a hyperplane that can separate all data points. In that case, SVM can use a soft margin that minimizes training error. If the data are not linearly separable in the original feature space, they are transformed by applying the kernel trick to a higher dimensional space, where the data become linearly separable. The kernel is a similarity function which operates high dimensional space by computing inner products of all pairs of data. There are commonly used kernels in SVM such as linear, polynomial, Gaussian radial basis function and hyperbolic tangent.

The decision tree is a commonly used tree-like structure model which is a decision support tool which maps an object into possible target classes. The tree consists of root nodes, branches which represent the outcome of the test, and leaves. In a decision tree, each internal node is a test on some attribute value and each leaf node holds a class label. Each leaf node assigns a classification or probability distribution over the class. The decision tree is generated based on the splitting criteria and tree pruning of the data. First, the tree is constructed by recursive splitting (partitioning) of the entire training example. . When the subset node belongs to the same class, the splitting is completed. Tree pruning is the inverse of splitting that is performed in order to identify and remove branches that reflect noise or outliers. The pruning process is conducted to improve accuracy and prevent overfitting. The quality of splits in a decision tree is measured by the Gini index which is one of the most frequently used indices that measure the degree of impurity. Many decision tree implementations have been developed

such as ID3, C4.5, CART, CHAID and MARS. Of these, CART (Classification and Regression Tree) is a widely used data mining technique (Brieman et al, 1984).

Attempting to obtain better performance of classification algorithms, an ensemble learning approach has been developed. Ensemble learning is the process that trains the multiple learning models and then combines the predictions of these base classifiers. According to Biau et al., the ensemble technique combines multiple statistical models to make predictions, which is well known to be a more accurate method than an individual tree model (Biau et al., 2007). Ensemble methods have been studied in various ways of combining base classifiers including parallel combining, stacked combining and weak combining. Of these approaches, weak combining is commonly used ensemble approach which conducts classification on a same training set and combines same type of classifiers. Bagging (bootstrap aggregating), boosting, random space and RF are widely used combining algorithms in many scientific researches.

Bagging (bootstrap aggregating: Breiman, 1996) is a commonly used ensemble algorithm in classification or regression. Bootstrap is a sampling technique that generates random samples with replacement. Given the entire training data set, bagging randomly generates new training data sets by bootstrap samples of the data. Each training data set is used to generate a classifier. The

result is then given as a combination of individual classifiers by taking a simple majority vote of their decisions. For any given instance, the class chosen by a majority of classifiers is the decision of the ensemble.

Boosting (Schapire, 1990) generates multiple base classifiers to form a stronger classifier (Kearns, 1988). In boosting, the base classifiers are weak learners which are slightly better than random guessing. The base classifier is a tree with only one split (stump). In the training process, the weights are adjusted by the misclassification rate. Thus a higher weight is assigned to those observations that have been misclassified more often. Boosting creates three weak classifiers. The first classier is trained on a random subset of the available training data. The second classifier is trained on a training data only half of which is correctly classified by the first classier, and the other half is misclassified. The third classier is trained with instances on which the first classifier and the second classifier disagree. These three classifiers are combined by majority voting. There exist various boosting approaches including AdaBoost, LogitBoost, LPBoost and BrownBoost. The main difference of these boosting approaches is the methods to give weight on the training data. Among these approaches, AdaBoost is one of the most widely used algorithms in machine learning. AdaBoost is a shorter term of Adaptive Boosting (Freund and Schapire, 1996). AdaBoost constructs a strong classifier as linear combination of weak learners by selecting their weight. These weights are updated iteratively and AdaBoost defines a new distribution of

weights over the training data. At each iteration, classifiers are trained by giving adjusted weights and these classifiers are combined by weighted majority voting. AdaBoost selects only known features in the training process so that it is fast, simple and easy to program.

Random Forest (RF: Breiman, 2001) is a classification method with an ensemble of multiple decision trees. In RF, the bagging algorithm uses bootstrap samples to build base trees. Each bootstrap sample is formed by random sampling, with replacement, of the same size as the original data. At each node of a tree, m variables are selected at random ($m < M$, where M is the number of variables in the data) in each node and the best split on these values is used to split the node. The largest possible tree is grown without pruning. The final classification yielded by RF is the class having the most votes across all trees. At each bootstrap iteration, approximately one third of the cases are left and not used in the construction of the tree. These out-of-bag cases can be used to calculate the generalized error rate or the variable importance measure.

RF estimates the variable importance to find which independent variables are important in the classification (Van der Laan, 2006). The measure of variable importance is based on the mean decreased accuracy on the out-of-bag observations. The mean decreased accuracy is the prediction accuracy calculated on the out-of-bag data for the variable which is permuted when all the others

remain the same. After permuting one variable, the prediction accuracy is calculated and compared with the accuracy for the original data. This process is repeated for all the variables. A higher reduction of the accuracy indicates a greater importance of the variable.

In this study, we compared widely used supervised learning techniques and suggested how classification models can utilize potential schizophrenia risk among the people who is in the control group. We compared the performance of well-known classification methods including RF, LDA, AdaBoost and SVM in classifying the schizophrenia data. The models were used to classify the data into different response groups and their accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were compared. We then used these methods to predict the risk for developing schizophrenia. We may encounter imbalance between sensitivity and specificity if the data set is unbalanced. Among the classifiers we considered, RF and SVM were the most sensitive to class imbalance (Blagus and Lusa, 2010). Since the data set we used in this paper is quite unbalanced, we enhanced the RF and SVM models by adjusting the decision threshold. According to our results, RF outperforms other methods for classifying patients in terms of overall accuracy and the area under the ROC curve (AUC). In addition, we found that RF and SVM with adjusted decision threshold show a substantial improvement in balance between sensitivity and specificity over RF and SVM without a threshold adjustment, respectively.

Chapter 2

Clustering Analysis

2.1 K-Means Cluster Analysis

2.1.1 Determining the Number of Clusters

Although *K-means* clustering technique is a very powerful statistical tool in many applications, it also has some limitations. One of the main problems in *K-means* clustering is how to determine the number of clusters in a data set. We propose to develop a method for choosing the optimal number of clusters based on the elbow method. We generated *K-means* clustering and found optimal *k* with the 5 methods using the schizophrenia data.

A. Elbow difference

In the graph, a large reduction in WSS indicates an appropriate number of clusters, hence the term "elbow criterion". In some case, however, the graph may have several incremental points indicating that more than one natural set of clusters fit data (Aldenderfer and Blashfield, 1984). The location of the elbow in the resulting plot suggests a suitable number of clusters for the *K-means*.

Figure 1 depicts the graphical representation of WSS over different number of clusters, and the values of WSS for $k = 1$ through $k = 15$ are listed in Table 1. The table also provides WSS, difference of WSS and difference in differences of WSS between clusters. In Figure 1, WSS decreases rapidly in k from 1 to 2 and from 2 to 3. However, WSS decreases slowly after $k = 3$. Table 1 shows that the largest difference in difference of WSS between clusters is 2142.58 between $k = 2$ and $k = 3$. Therefore, $k = 3$ is the optimal number of clusters under this criterion.

B. Proposed elbow ratio

As an alternative approach, we propose a method named elbow ratio for identifying a rapid change in the difference of WSS by examining the reduction rate of the change in WSS. Suppose that we have the difference of WSS between clusters. Then the ratio of two differences of WSS is,

$$\text{Ratio of the differences in WSS between two clusters} = E_k = \frac{D_k}{D_{k-1}}$$

$$\text{where } D_k = WSS_{k-1} - WSS_k$$

The optimal number of clusters is then the value of k with minimum E_k.

Determining the elbow point in the curve by elbow difference may not always be the best measure. For example, the Elbow difference method fails to detect the best location of the elbow in some cases. Suppose that the data points are $WSS_1 = 19{,}727,\ WSS_2 = 15{,}733,\ WSS_3 = 12{,}259,\ WSS_4 = 10{,}691,\ WSS_5 = 10{,}079$ for $k = 1,\ 2,\ 3,\ 4,\ 5$. Then we have $D_1 = 3{,}994,\ D_2 = 3{,}474,\ D_3 = 1{,}568,\ D_4 = 612$ and the differences of difference in WSS are $3994 - 3474 = 520,\ 3474 - 1568 = 1906,\ 1568 - 612 = 956$. Since 1906 is the largest gap change in the difference of WSS, the elbow point in the curve should be 3 if the elbow difference method is used. On the other hand, the ratio of the

differences in WSS, which will be explained in details below, is $3474/3994 = 0.870,\ 1568/3474 = 0.451, 612/1568 = 0.390.$ Since 0.390 is the smallest among these three ratios, based on the elbow ratio, $k = 4$ is the elbow point based on the Elbow ratio method. Since the elbow difference method is affected by the order of magnitude, a rapid change in the difference in WSS is not recognized if the order of magnitude is small even if the elbow has the lowest angle in the figure. However, the Elbow ratio method can recognize it regardless of the order of magnitude.

For the given schizophrenia data, Table 2 shows that, WSS decreases rapidly for $k = 1$ to 3. The corresponding values for ratio of the difference E_k also decrease from 0.73 to 0.45 and E_k reaches the minimum value of 0.45 when $k = 3$. Based on the Elbow ratio criterion, $k = 3$ is the optimal number of clusters.

Figure 1: WSS plot over the number of clusters

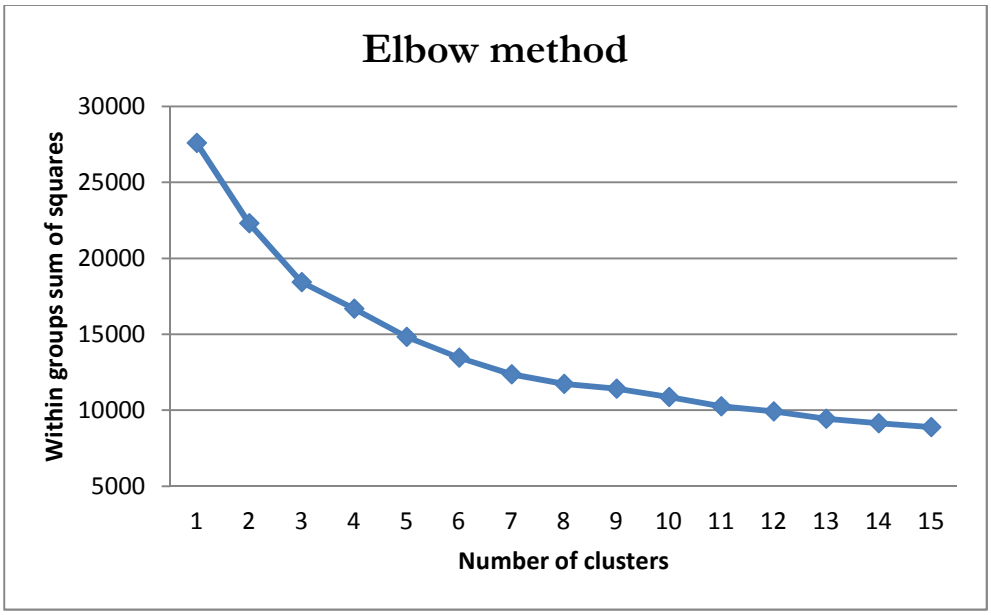


Table 1: Optimal k based on the Elbow difference approach.

#Clusters	WSS	Difference of WSS between clusters $=D_k$	Difference in differences of WSS $=D_{k-1} - D_k$
1	27609.7		
2	22319.43	5290.27	1403.49
3	18432.65	3886.78	2142.58
4	16688.45	1744.2	-104.31
5	14839.94	1848.51	474.32
6	13465.75	1374.19	277.96
7	12369.52	1096.23	463.41
8	11736.7	632.82	321.03
9	11424.91	311.79	-250.72
10	10862.4	562.51	-39.31
11	10260.58	601.82	262.898
12	9921.658	338.922	-141.869
13	9440.867	480.791	182.507
14	9142.583	298.284	44.01
15	8888.309	254.274	

Table 2: Optimal k based on the Elbow ratio approach.

#Clusters	WSS	Difference of WSS between two clusters=D_k	Ratio of WSS between two clusters=D_k/D_{k-1}
1	27609.7		
2	22319.43	5290.27	0.734704
3	18432.65	3886.78	0.448752
4	16688.45	1744.2	1.059804
5	14839.94	1848.51	0.743404
6	13465.75	1374.19	0.797728
7	12369.52	1096.23	0.577269
8	11736.7	632.82	0.492699
9	11424.91	311.79	1.804131
10	10862.4	562.51	1.069883
11	10260.58	601.82	0.563162
12	9921.658	338.922	1.418589
13	9440.867	480.791	0.620403
14	9142.583	298.284	0.852456
15	8888.309	254.274	

C. Average silhouette method

The average silhouette method computes the average distance of observations within clusters of data in order to represent an evaluation of clustering validity. Each data point i has its own average distance from all the other data points within the same cluster. This is the dissimilarity of data point $a(i)$. Let $b(i)$ be the minimum average distance to a cluster which does not contain the object i. In other word, $b(i)$ is defined as the lowest average dissimilarity of i to any other cluster. Then the silhouette coefficient $s(i)$ is

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & if\ a(i) < b\ (i) \\ 0 & if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & if\ a(i) > b(i) \end{cases}$$

$$\text{where } -1 \leq s(i) \leq 1$$

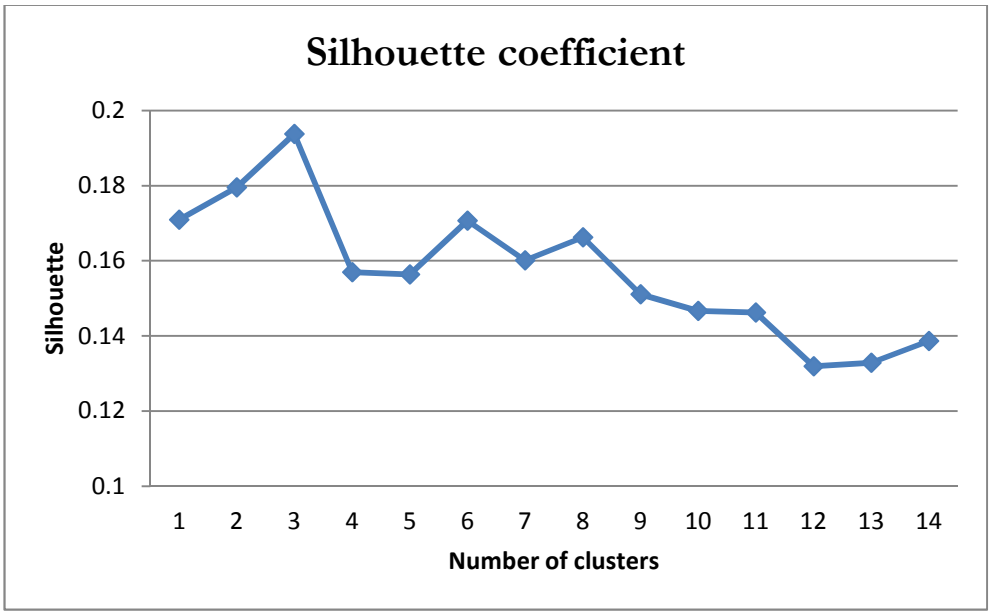
If the value $s(i)$ is close to 1, it means that a sample has been assigned in an appropriate cluster, while the value $s(i)$ near negative one indicates that a sample has been assigned to the wrong cluster. Thus the average silhouette method is used to determine the appropriate number of clusters in *K-means*

clustering. The silhouette computes the silhouette coefficient for different values of k. The optimal number of clusters k is the one that maximizes the silhouette coefficient over a range of possible values for k (Kaufman and Rousseeuw, 1990). Table 4 shows the silhouette coefficients over different values of k. The result indicates that the silhouette coefficient is highest when $k = 3$ suggesting that the optimal number of clusters is 3. The Figure 2 also shows that the silhouette coefficient curve has a local peak at $k = 3$, which then provides the estimated optimal number of k as three.

Table 3: Silhouette coefficients over the number of clusters.

#Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**
silhouette coefficient	0.1709	0.1796	0.1938	0.1570	0.1564	0.1707	0.1601
#Clusters	**8**	**9**	**10**	**11**	**12**	**13**	**14**
silhouette coefficient	0.1662	0.1511	0.1466	0.1462	0.1319	0.1329	0.1386

Figure 2: Silhouette coefficient plot over the number of clusters



D. Gap statistic

The gap statistic is another approach for estimating the optimal number of clusters in a set of data (Tibshirani, 2001). The strategy of the algorithm is to compare the total within cluster dispersion with their expected value. Application of the gap method is a powerful procedure in determining the number of clusters on both *K-means* and hierarchical clustering method.

For each data point i, let $d_{ii'}$ be the distance betweenobservation i' and i. Let D_r be the sum of the pairwise distance for all points within cluster r. Here $d_{ii'}$ can be interpreted as the squared Euclidean distance.

For a fixed value of k, define W_k as

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

where $D_r = \sum d_{ii'}$ and $d_{ii'} = \sum_j (x_{ij} - x_{i'j})^2$

In order to calculate the gap statistic, the expected value of mean dispersion W_k*is computed by the reference dataset which is generated using an appropriate null distribution. The gap statistic is defined as:

$$Gap_n(k) = E_n^*\{log(W_k^*)\} - log(W_k)$$

The optimal number of clusters k is the smallest k such that

$$Gap_n(k) \geq Gap_n(k+1) - s_{k+1}$$

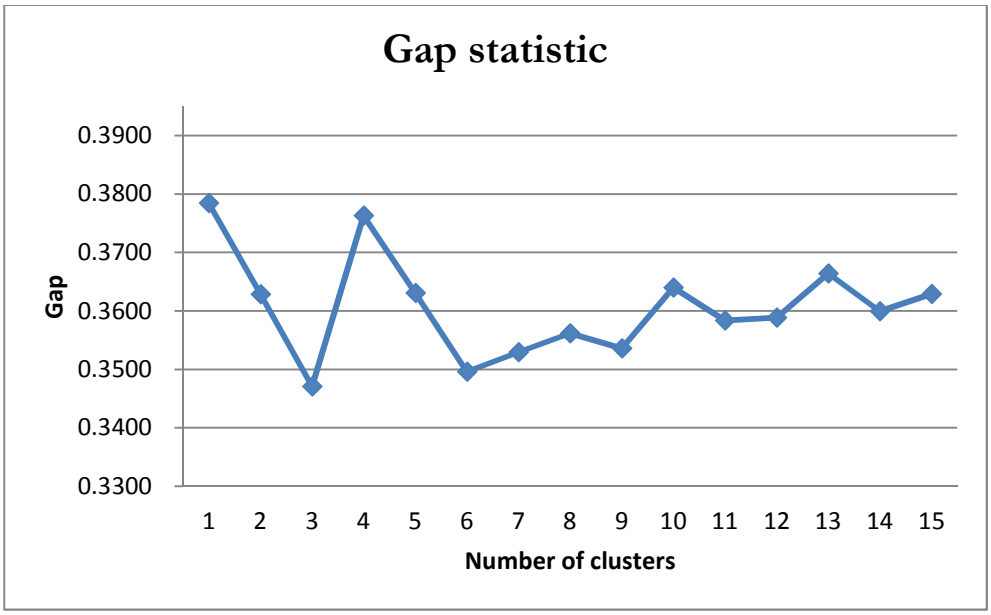
where s_{k+1} is the standard error.

The result of gap statistic approach for different number of clusters is provided in Table 4 and Figure 3. In Table 4, the gap statistic is the lowest when $k = 3$. However, the difference in gap statistic and standard error with $k = 4$ is greater than the gap statistic of $k = 3$. Thus the optimal k based on the gap statistic is 6.

Table 4: Gap statistics over the number of clusters.

#Clusters	logW	E(logW)	Gap= E(logW)-logW	SE
1	6.4785	6.8570	0.3785	0.0168
2	6.3672	6.7301	0.3629	0.0168
3	6.2913	6.6384	0.3471	0.0185
4	6.1919	6.5682	0.3763	0.0164
5	6.1526	6.5157	0.3631	0.0153
6	6.1186	6.4681	0.3496	0.0155
7	6.0778	6.4307	0.3529	0.0157
8	6.0446	6.4007	0.3562	0.0157
9	6.0196	6.3732	0.3536	0.0149
10	5.9848	6.3488	0.3640	0.0146
11	5.9673	6.3257	0.3584	0.0132
12	5.9454	6.3042	0.3588	0.0147
13	5.9181	6.2845	0.3664	0.0146
14	5.9039	6.2638	0.3599	0.0145
15	5.8824	6.2453	0.3629	0.0150

Figure 3: Gap statistic plot over the number of clusters.



E. Bayesian Information Criterion (BIC)

In implementing mixture models for clustering, there are two approaches: One approach is an iterative relocation approach through the expectation maximization (EM) algorithm for maximum likelihood estimation (Celeux and Govaert, 1995). However, if the number of clusters is increasing in the mixture model, then the dimensionality of the model is also increasing and it causes an increase in its likelihood and possible over fitting. In order to avoid such a problem, a BIC criterion is applied that does not depend on the likelihood.

The formula of BIC is defined as

$$\mathrm{BIC} = -2 * \ln(\mathrm{likelihood}) + \ln(\mathrm{N}) * \mathrm{k}$$

where k is degree of freedom of the parameters and N is the number of observations.

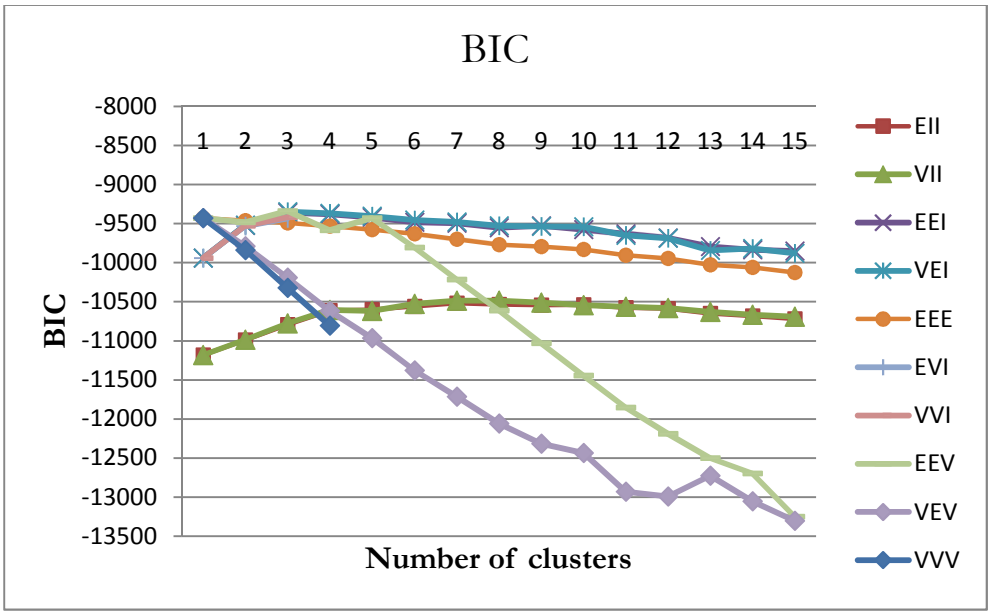
This is the value of the maximized log likelihood measure with a penalty. The package mclust in R with all default settings was used to evaluate BIC. Mclust uses a model option that is represented by an identifier for parameterization of the covariance matrix. The identifier is the geometric characteristic of the models including three letters: E for equal variance, V for varying variance and I for the coordinate axes. For example, EEV denotes a model in which the volumes of all clusters are equal (E), the shapes are equal (E)

and the orientation is varying (V). There are 10 combinations of volume, shape and orientation included in mclust package. Among the available models, the model with the highest BIC value is taken as the best model. Figure 4 shows the BIC values of 10 different models over the number of clusters and EEV is estimated as the best model since it takes the largest BIC value when $k = 3$. Table 5 provides the results of BIC to see the highest value over the number of clusters.

Table 5: BIC value of the model EEV over the number of clusters.

#Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**
BIC	-9428	-9482	-9340	-9583	-9428	-9804	-10215
#Clusters	**8**	**9**	**10**	**11**	**12**	**13**	**14**
BIC	-10614	-11030	-11444	-11852	-12189	-12500	-12696

Figure 4: BIC plot over the number of clusters.



1: EII = equal volume, VII = unequal volume, EEI = equal volume and shape, VEI = unequal volume and equal shape, EEE = equal volume, shape, and orientation, EVI = equal volume and varying shape, VVI = varying volume and shape, EEV = equal volume, shape and varying orientation, VEV = varying volume, equal shape and varying orientation, VVV = varying volume, shape, and orientation.

2.1.2 Comparison of the methods for estimation of k

First, the data set was randomly divided into two groups with a ratio of 2:1 for finding optimal *k*. The two-thirds of the data are chosen as a training set and the remaining subset is used as a test data set. The K-means clustering is trained on the training data in order to find the optimal number of *k* by 5 different methods (Elbow Difference, Elbow Ratio, Average silhouette method, Gap statistic, and Bayesian Information Criterion). One hundred repetitions of K-means clustering were conducted. In each of the repetitions, the same training and test set pair was used to evaluate the performance of 5 methods. The optimal *k* obtained for each method was then tested on the test data. Table 6 shows the frequency that the same optimal *k* obtained in the training set was obtained in the test set. The percentage of the same optimal *k* in the test set as in the training set is provided in Table 7.

The result in the training set indicates that all of the five methods detect k=3 as the dominant optimal number of clusters. The BIC method estimated the number of clusters as 3 for 94 times. On the other hand, the Gap method determined k as 3 for only 77 times. In Table 7, BIC method reveals the highest

percentage of finding the same optimal k in the test set as in the training set. The original elbow difference method as well as the proposed elbow ratio method give similar estimating results in finding the optimal number of clusters.

Table 6: Frequency of finding the same optimal k in the test set as in the training set. The frequency for the training set and the number in parentheses is the frequency in the test set.

k	**Elbow Diff**	**Elbow Ratio**	**Silhouette**	**Gap**	**BIC**
2	13 (0)	3 (0)	7 (1)	12 (1)	0 (0)
3	85 (75)	86 (75)	85 (61)	77 (63)	94 (78)
4	2 (0)	11 (1)	1 (0)	11 (1)	5 (1)
5	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
6	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)

Table 7: Percentage of finding the same optimal k in the test set as in the training set.

Method	Frequency
Elbow Diff	75%
Elbow Ratio	76%
Silhouette	62%
Gap	65%
BIC	79%

2.1.3 Simulation Study

A simulation study was conducted to examine the performance of the five different methods for finding optimal number of clusters. We generated 15 variables and 120 subjects. Two simulation experiments were conducted. Each variable was divided into three classes of sizes 40 each. The first 10 variables were randomly generated from a normal distribution with different means and variance σ^2 of 1 or a uniform distribution with different means and variance of 0.75. The remaining 5 variables were generated from a normal distribution with mean 1 and variance 1.

A. Simulation Experiment 1

Figure 5 displays the data design for the simulation experiment. Among the first 10 variables, the first 2 variables were generated from *N* (1, 1), for 40 samples in class 1, *N* (3, 1) for 40 samples in class 2 and *N* (5, 1) for 40 samples in class 3. The next 2 variables were generated from *N* (3, 1) for class 1, *N* (5, 1) for class 2 and *N* (1, 1) for class 3. The next 2 variables were generated from *N* (5, 1) for class 1, *N* (3, 1) for class 2 and *N* (1, 1) for class 3. Next 2 variables were generated from *U* (0, 3) for class 1, *U* (2, 5) for class 2 and *U* (4, 7) for class 3.

The remaining 2 variables were generated from U (4, 7) for class 1, U (2, 5) for class 2 and U (0, 3) for class 3. The remaining 5 variables were generated from N (1, 1). For each simulated data set, the optimal number of k was estimated by 5 different methods. The entire process was repeated by generating 100 simulation data from the above design. The correct number of clusters should be $k = 3$. The estimated optimal number of k is provided in Table 8.

We generated simulated data sets from 5 different models that contain clear cluster structure with known number of clusters. The efficiency of the methods is measured by the frequency of the data sets for which the number of clusters is correctly estimated. Overall, the simulation study shows that all of the 5 methods appeared to be efficient in estimating the optimal number of clusters. The methods selected 3 clusters from 85 to 96 times. Among the 5 methods, Elbow Ratio and BIC performed the best. The BIC model selected $k = 3$ clusters 96 times, while Elbow Ratio selected 93 times. However, the BIC had a wider range of k than Elbow Ratio. In terms of consistency, Elbow Ratio performed better than BIC. The mean of k in Elbow ratio (2.99) was the closest to 3 among the 5 methods. The standard deviation of k in Elbow Ratio was far less than that of the other 4 methods. The estimated k in Elbow Ratio ranged from 2 to 4, while it ranged from 3 to 7 in BIC.

Figure 5: Data design for simulation experiment 1.

120 samples	10 variables					5 variables
	2	2	2	2	2	
Class 1 : 40	$N(1,1)$: 40	$N(3,1)$: 40	$N(5,1)$: 40	$U(0,3)$	$U(4,7)$	$N(1,1)$
Class 2 : 40	$N(3,1)$: 40	$N(5,1)$: 40	$N(3,1)$: 40	$U(2,5)$	$U(2,5)$	
Class 3 : 40	$N(5,1)$: 40	$N(1,1)$: 40	$N(1,1)$: 40	$U(4,7)$	$U(0,3)$	

Table 8: Frequency of the optimal k in the simulated data set

k	**Elbow Diff**	**Elbow Ratio**	**Silhouette**	**Gap**	**BIC**
2	7	4	9	3	0
3	85	93	87	90	96
4	6	3	3	5	2
5	2	0	1	0	1
6	0	0	0	2	0
7	0	0	0	0	1
mean	3.03	2.99	2.96	3.06	3.08
sd	0.4596	0.2657	0.4000	0.4221	0.4645

B. Simulation Experiment 2

In the second simulation experiment, the means in the three clusters were more close to each other than those of the first experiment. Among the first 10 variables, the first 2 variables were generated from $N(1.5, 1)$, for 40 samples in class 1, $N(3, 1)$ for 40 samples in class 2 and $N(4.5, 1)$ for 40 samples in class 3. Next 2 variables were generated from $N(3, 1)$ for class 1, $N(4.5, 1)$ for class 2 and $N(1.5, 1)$ for class 3. Next 2 variables were generated from $N(4.5, 1)$ for class 1, $N(3, 1)$ for class 2 and $N(1.5, 1)$ for class 3. Next 2 variables were generated from $U(0, 3)$ for class 1, $U(1.5, 4.5)$ for class 2 and $U(3, 6)$ for class 3. The remaining 2 variables were generated from $U(3, 6)$ for class 1, $U(1.5, 4.5)$ for class 2 and $U(0, 3)$ for class 3. The remaining 5 variables were generated from N (1, 1).

Figure 6 shows the data design for this simulation experiment 2. All five methods showed lower accuracy than simulation experiment 1 in determining the number of clusters, as expected. The results had the same pattern as in experiment 1. BIC chose k = 3 slightly more (73) frequently than Elbow Ratio (71), but it was less consistent in selecting k than Elbow Ratio. In this experiment, the Elbow method showed

better performance than the other three methods. Between the two Elbow methods, Elbow Ratio performed better. As in Experiment 1, the mean of k (3.07) was the closest to 3 for the Elbow Ratio than for the other methods. The difference was more than 10% in Silhoutte (mean 3.31), Gap (mean 3.35) and BIC (3.36). The standard deviation in Elbow Ratio was the smallest (0.5366), while it was over 0.80 in those three methods. The standard deviations of k in the Elbow methods were far less than that in the other 3 methods. The estimated k in the Elbow methods ranged from 2 to 4, while it ranged from 2 to 7 in BIC.

Figure 6: Data design for simulation experiment 2.

120 samples	10 variables					5 variables
	2	2	2	2	2	
Class 1 : 40	$N(1.5,1)$: 40	$N(3,1)$: 40	$N(4.5,1)$: 40	$U(0,3)$	$U(3,6)$	$N(1,1)$
Class 2 : 40	$N(3,1)$: 40	$N(4.5,1)$: 40	$N(3,1)$: 40	$U(1.5,4.5)$	$U(1.5,4.5)$	
Class 3 : 40	$N(4.5,1)$: 40	$N(1.5,1)$: 40	$N(1.5,1)$: 40	$U(3,6)$	$U(0,3)$	

Table 9: Frequency of the optimal k in the simulated data set

k	Elbow Diff	Elbow Ratio	Silhouette	Gap	BIC
2	10	11	8	7	3
3	63	71	67	68	73
4	27	18	14	15	11
5	0	0	9	3	12
6	0	0	1	7	0
7	0	0	1	0	1
8	0	0	0	0	0
mean	3.17	3.07	3.31	3.35	3.36
sd	0.5870	0.5366	0.8610	0.9252	0.8105

2.1.4 Exhaustive search of k

Following MacQueen's (1967) K-means methodology, we used an algorithm in which each item is assigned to the cluster having the nearest centroid (mean). This nonhierarchical method initially takes the number of components of the population equal to the final required number of clusters. The final required number of clusters is chosen such that the points in different clusters are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters. The process is composed of the following three steps: 1) partition the items into initial clusters; 2) proceed through the list of items, assigning an item to the cluster whose centroid is nearest (we used Euclidean distance as the measure of distance). Then, recalculate the centroid for the cluster receiving the new item and for the cluster losing the item; 3) repeat Step 2 until no more reassignment takes place. In this study, we ran numerous analyses with various values of K (from K=2 to K=12 for the above described two sets of variables), with the goal of finding clusters with high concentrations of PARS subjects.

A. Application

To understand the characteristics of schizophrenia related to the paternal age, we conducted a clustering analysis (Lee et al., 2011). For our analyses we were interested in a set of core factors consisting of demographic, clinical and cognitive variables. Thus, we included in our analyses only cases that had the following variables: age of onset of psychosis, sex, family history of schizophrenia, age of the father at the case's birth (paternal age), diagnosis, severity of psychopathological symptoms, and neuropsychological function.

This study relies on cases with 284 schizophrenia or schizo-affective disorder patients recruited at the New York State Psychiatric Institute (NYSPI) Schizophrenia Research Unit (SRU) in 1992-2007. The study was approved by the Institutional Review Board at NYSPI and all patients provided written informed consent. For our analyses we were interested in a set of core factors consisting of demographic, clinical and cognitive variables. Thus, we included in our analyses only cases on whom we had the following variables: age of onset of psychosis, sex, family history of schizophrenia, age of the father at the case's birth (paternal age), diagnosis, severity of psychopathological symptoms, and neuropsychological function. We operationally defined PARS as the absence of any family history of schizophrenia among first- and second-degree relatives and for cases whose fathers' age at birth was >35 years; all other cases were

considered non-PARS (based on Malaspina et al., 2002). All cases were taking medication at the time of assessment.

Diagnosis was obtained using the Diagnostic Interview for Genetic Studies (DIGS; Nurnberger et al., 1994). In addition to these measures, we included symptoms (Positive and Negative Syndrome Scale; PANSS), cognitive tests (Wechsler Adult Intelligence Scale—Revised; WAIS-R) and olfaction (University of Pennsylvania Smell Identification Test; UPSIT). The Wechsler Adult Intelligence Scale--Revised (WAIS-R; Wechsler, 1981) was used to obtain the following neuropsychological factors: full scale intelligence quotient (FIQ), verbal IQ (VIQ) and performance IQ (PIQ), as well as the verbal subtests (arithmetic, digit span, information, vocabulary, comprehension, similarities) and the performance subtests (object assembly, picture arrangement, picture completion, digit symbol, and block design). We also obtained a verbal-performance differential score (VIQ-PIQ). The UPSIT is a standardized, multiple-choice, scratch-and-sniff test with the maximum score of 40 (perfect identification) that is found to be stable and reliably measured in schizophrenia patients (Malaspina et al., 1994).

The descriptive data for the demographic, clinical and neuropsychological variables are presented in Tables 10 and 11. The above described the *K-means*

clustering algorithm was run using various combinations of variables in order to identify latent subgroups of PARS.

The clusters were generated using the following core variables: age of onset of psychosis, sex (males = 0; females = 1), family history (no family history of schizophrenia = 0; family history of schizophrenia = 1) and paternal age (age of the father at the case's birth). Together with these variables we added one of two sets of variables, or a combination of them. These two sets of variables were: the WAIS-R FIQ, PIQ, VIQ, VIQ-PIQ and the verbal and performance subtests; and the PANSS scores from the standard model (positive, negative and general psychopathology subscale scores).

B. Results

Two of our *K-means* clustering analyses produced clusters with high PARS concentration. We defined PARS as not having any family history of schizophrenia among first and second-degree relatives and fathers' age at birth ≥35 years (PARS = 1; non-PARS = 0). Each of these two clustering analyses generated seven clusters (K=7) and yielded some prominent features related to the PARS subjects. The first analysis included the 11 WAIS-R subtests in addition to the four core demographic variables (age of onset of psychosis, sex, paternal age, family history of schizophrenia). The second analysis included the VIQ-PIQ variable and the PANSS factors from the standard model (positive, negative and general psychopathology symptoms) together with the same four core demographic variables used in the first analysis.

For each variable included in the clustering analysis, we conducted a two-sample *t*-test for continuous variables and a chi-square test for categorical variables for a comparison between a specific cluster and the rest of the data. The data are expressed as the mean ± standard error of the mean. The means (and standard deviation; SD) are presented in Tables 3 and 4. A two-sided test is used for all the analyses.

Table 12 shows the first analysis, which included 136 cases, 34 (25.0%) of which were PARS. Among the clusters, Cluster 3 (N=24) contained 20 PARS cases (83%). From Figure 7, this cluster had a higher average differential score between VIQ and PIQ (VIQ-PIQ) than the rest of the sample (12.9±2.3 vs. 6.3±1.0, p=0.009). Also, the mean paternal and maternal ages were relatively high at 41 and 33 years, respectively as shown in Figure 8.

Figure 7: First cluster analysis: scatter diagram for VIQ-PIQ variable according to PARS and cluster.

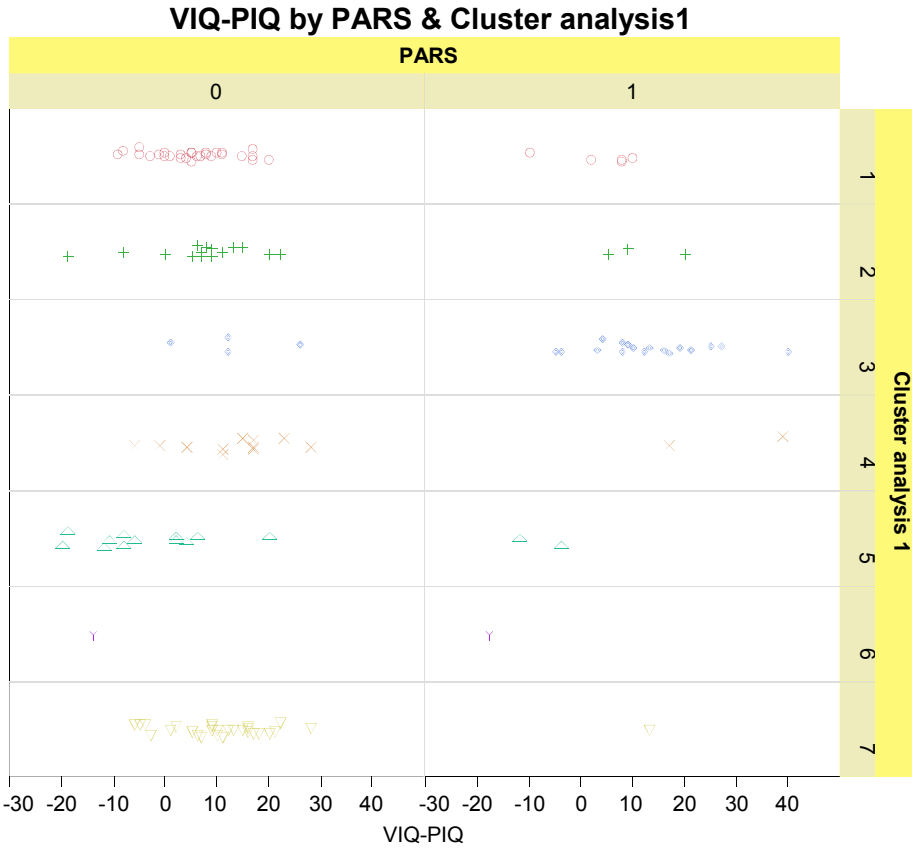
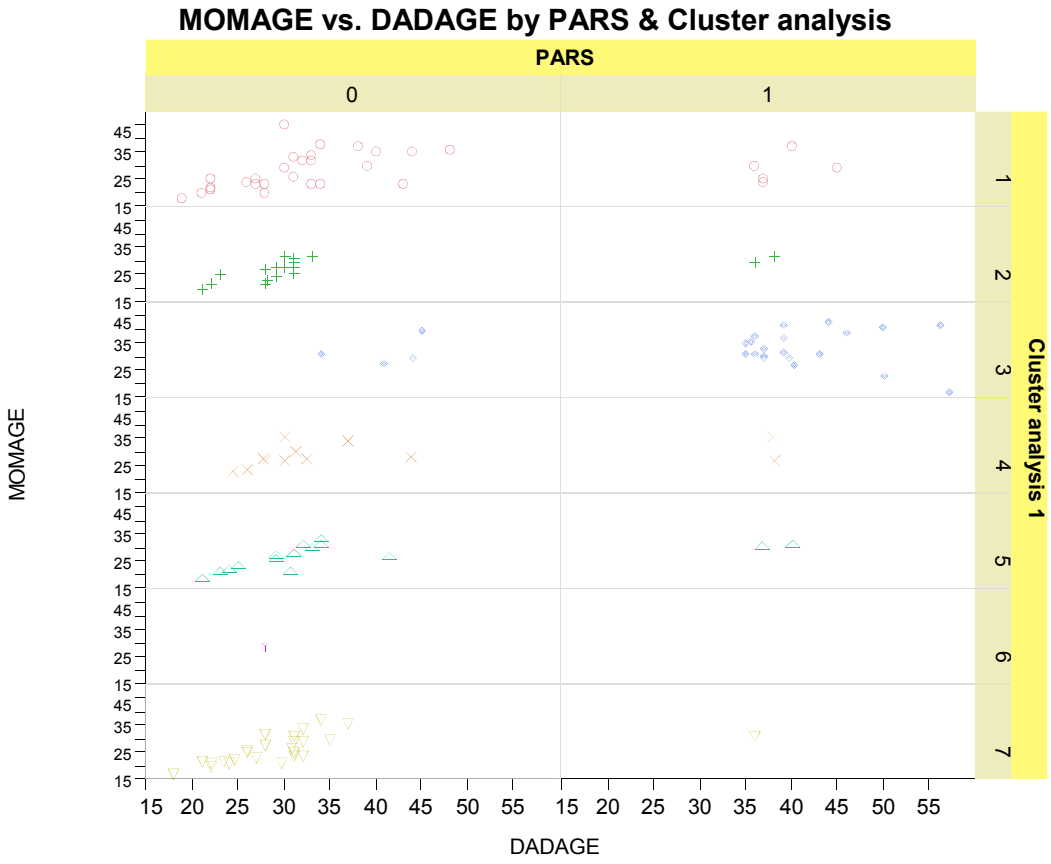


Figure 8: First cluster analysis: scatter diagram for maternal age and paternal age according to PARS and cluster.

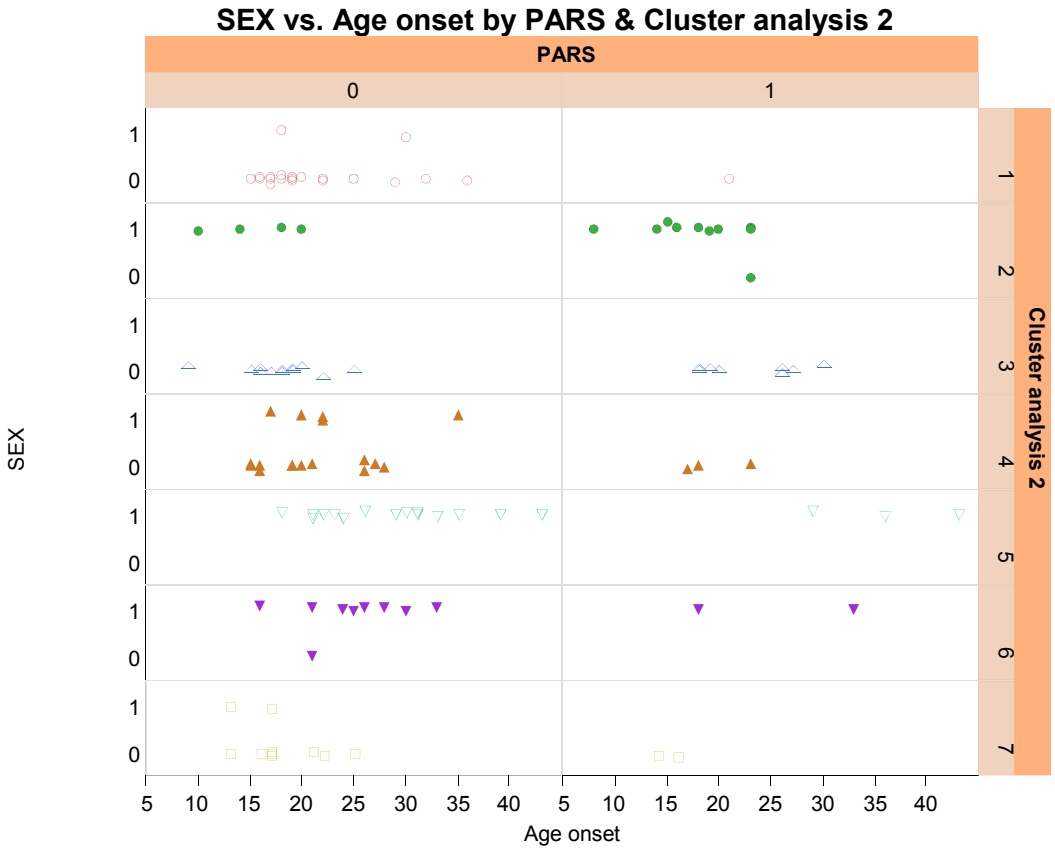


Although less relevant to our main focus on PARS, Table 12 displays some other significant results. Compared to the rest of the data, Cluster 1 had more familial cases (p=0.004), lower average UPSIT score (p=0.039), higher proportion of deficit syndrome cases (p=0.009), lower neuropsychological scores (p<0.0001), and higher PANSS (standard model) subscale scores (all p<0.005). Cluster 2 showed higher scores on block design, object assembly and digit symbol scores (p=0.003, 0.013 and 0.001, respectively) compared to the rest of the data. Cluster 4 and Cluster 5 demonstrated higher FIQ, VIQ, PIQ and most of the WAIS-R subtest scores than the rest of the data. The standard model PANSS subscales were lower in Cluster 5 than the rest of the data (all p<0.05), but no difference was observed in Cluster 4. Clusters 5 and 6 consisted of only males with early age onset of psychosis and these cases had higher PIQ scores than VIQ scores

The second analysis yielded 123 cases of which 30 (24.4%) were PARS. This analysis is shown in Table 13 Cluster 2 produced the highest proportion of PARS with 10 out of 14 cases (71%; these were all sporadic patients). Inspection of the data revealed that the majority of the cases in this cluster were also in cluster 3 in the first analysis discussed above. Interestingly, compared to the other clusters, the cluster 2 cases were mostly females (93% vs. 35% females, p<0.0001) with earlier age of onset of psychosis (17.2±1.7 vs. 22.4±0.6, p=0.006) as shown in Figure 9. The mean paternal age was higher in the Cluster 2 group (39.3;

SD=8.8) than the rest of the data. None of the Cluster 2 cases had the deficit syndrome, and the neuropsychological scores and the PANSS scores for this group were not significantly different from those in the other groups.

Figure 9: Second cluster analysis: scatter diagram for gender and age onset of psychosis according to PARS and cluster.



From the other significant results in Table 13, Cluster 1 had a higher proportion of males ($p=0.001$), more schizophrenia than schizo-affective cases ($p=0.018$), lower mean UPSIT score ($p=0.006$), higher proportion of deficit syndrome cases ($p<0.0001$), lower main WAIS-R scores (FIQ: $p=0.025$; PIQ: $p=0.030$; VIQ: $p=0.031$), lower standard scale PANSS positive symptoms ($p=0.0005$), and higher standard scale PANSS negative symptoms ($p<0.0001$) compared to the rest of the data. Cluster 3 contained 100% male cases with a larger difference between VIQ and PIQ scores than the rest of the data ($p<0.0001$). This is due to the low mean PIQ score, which is the lowest among the 7 clusters ($p=0.019$). Cluster 4 demonstrated lower paternal age ($p=0.003$), higher FIQ, VIQ and PIQ scores (all $p\leq0.02$), and lower standard PANSS negative and general psychopathology scores ($p<0.0001$) than the rest of the data.

Additionally, this is the only group with lower mean VIQ than PIQ. Cluster 5 consisted of 100% females with a later onset of psychosis ($p<0.0001$), higher UPSIT score ($p=0.007$), more familial cases ($p=0.0003$), and lower standard PANSS symptom subscales (all $p<0.01$) than the rest of the data. Cluster 6 had a higher proportion of females than the rest of the data ($p=0.0005$). Cluster 7 showed earlier onset of psychosis ($p=0.008$), higher proportion of males ($p=0.018$), more familial cases ($p=0.003$), and lower FIQ ($p=0.010$) and VIQ ($p=0.001$) than the rest of the data. Both Cluster 6 and Cluster 7 had no deficit

syndrome cases, and they had higher PANSS standard positive and general symptoms scales ($p<0.0001$) than the rest of the data.

It should be noted that in the first cluster analysis, the 34 PARS cases were distributed as follows: Cluster 1 (N=5), Cluster 2 (3), Cluster 3 (20), Cluster 4 (2), Cluster 5 (2), Cluster 6 (1) and Cluster 7 (1). In the second cluster analysis, the 30 PARS patients were distributed as follows: Cluster 1 (N=1), Cluster 2 (10), Cluster 3 (9), Cluster 4 (3), Cluster 5 (3), Cluster 6 (2) and Cluster 7 (2). Additionally, the 20 PARS cases in the Cluster 3 group of the first analysis were 10 males and 10 females. Among these, three cases were excluded in the second analysis due to missing PANSS scores. Of the remaining 17 PARS cases, most belong to Cluster 2 and Cluster 3, depending on their sex. Cluster 3 contained six male cases, whereas five females and one male were in Cluster 2. The remaining five PARS cases were distributed as follows: Cluster 1 (one male), Cluster 5 (two females) and Cluster 6 (two females).

Due to missing data points on some of the core measures, certain cases were only captured in the first clustering analysis (N=16; 8 males and 8 females), and some cases were only captured in the second analysis (N=3; 2 males and 1 female). 120 (70 males; 50 females) of the cases were captured in both clustering analyses. See supplementary data for the descriptive data for the cases that

overlap (are both in analysis 1 and 2) and for the cases that are either only in analysis 1 or 2.

Our first cluster analysis, which considered demographic variables and neuropsychological test score variables, showed a cluster containing 83% PARS cases. It was characterized by relatively high paternal age (mean age = 41 years). The mean maternal age (33 years) was also relatively high in this cluster group. Interestingly, the cases in this group demonstrated a significant difference between the WAIS-R verbal and performance intelligence, with verbal functioning being on average 12.97 points better.

The verbal versus performance IQ decrement is notable. The result is driven by better performance on all the verbal subtasks (arithmetic, digit span, information, vocabulary, comprehension, similarities) compared to the performance subtasks (object assembly, picture arrangement, picture completion, digit symbol, block design). This result supports previous findings in other populations showing a strong relationship between older fathers and human intelligence (Auroux et al., 1989; Malaspina et al., 2005). These data also replicate previous findings from the same population (Malaspina et al., 2002b), despite using a different method of analysis, suggesting reductions in non-verbal intelligence compared to verbal intelligence in cases with older paternal age. The fact that paternal age has a larger effect on performance intelligence than verbal

intelligence may be of interest. Future research that examines intelligence together with other cognitive and neurological symptoms would be valuable in further determining the distinctiveness of PARS as a subgroup of schizophrenia.

Our second k-means cluster analysis, which included the VIQ-PIQ discrepancy score, the PANSS standard subscale scores and the same demographic variables, revealed a cluster consisting of 71% PARS cases. This cluster had a significantly higher concentration of females and demonstrated an earlier age of psychosis onset. These findings show that later paternal age may have a particularly strong influence on the symptoms and clinical characteristics of female PARS cases in comparison with other female cases. A high risk of schizophrenia for females of older fathers was recently reported by Perrin et al (2010). Females of an affected sister born to fathers 35 years and older had a fourfold greater risk of schizophrenia than females with an affected sister born to fathers <35 years at time of birth. By contrast the risk of schizophrenia in males with an affected brother was only doubled for older versus younger fathers. They proposed that paternally expressed genes on the X chromosome could play a role in the risk associated with females of an affected sister born to older fathers.

A non-significant discrepancy (6.21 points) between the verbal and performance intelligence was evident in the high-concentration PARS group in the second analysis. The reduction in discrepancy from the first analysis may not

be surprising as we included the VIQ-PIQ discrepancy score as a factor in the second cluster analysis. The reduction may also be partly explained by gender separation in this cluster. As discussed above, 12 of the PARS cases from Cluster 3 in the first analysis were sorted into Cluster 2 (five females and one male) and Cluster 3 (six males) in the second analysis. The mean VIQ-PIQ of these 12 subjects was 11.8, and it was reduced in Cluster 2 (mean VIQ-PIQ of the six patients was 6.17), while it was increased in Cluster 3 (mean VIQ-PIQ of the six patients was 17.5). The PANSS scores revealed similar results as found in the first cluster analysis.

In summary, 284 cases with schizophrenia or schizo-affective disorder were included and the clusters were generated using different combinations of demographic, symptom and cognitive variables based on PARS relevance (high concentration of PARS cases grouping in the same cluster). The result of analyses indicated that some of the variability in schizophrenia symptoms can be explained by PARS cases that tend to "cluster" in groups with particular characteristics. Of particular note, a subgroup of largely female cases was identified as having separate features in association with later paternal age. This finding supports PARS as a distinct subgroup.

Table 10: K-means clustering: Descriptive statistics of nominal variables in Cluster 1 and Cluster 2

	Cluster 1	Cluster 2
Sex (male/female)	78/58	72/51
Diagnosis (schizophrenia/schizo-affective disorder)	103/33	92/31
Family history of schizophrenia (yes/no)	41/95	40/83
PARS (yes/no)	34/102	30/93
Deficits Syndrome (yes/no)	15/102	13/93

Table 11: K-means clustering: Demographic, clinical, neuropsychological and olfaction data in Cluster1 and Cluster 2

		Cluster 1			Cluster 2		
		N	Mean	SD	N	Mean	SD
Age of onset of psychosis	In years	136	21.90	6.62	123	21.77	6.70
Maternal age at the case's birth	In years	129	28.39	5.95	117	28.36	5.86
Paternal age at the case's birth	In years	136	32.49	7.40	123	33.00	7.42
PANSS[1]	Total	122	52.52	13.53	123	52.23	13.26
PANSS standard model	Positive symptoms	123	12.72	5.58	123	12.55	5.48
	Negative symptoms	122	13.94	5.59	123	13.96	5.56
	General psychopathology symptoms	123	25.89	6.83	123	25.74	6.72
PANSS 5-factor model	Positive symptoms	123	9.89	4.52	123	9.72	4.44
	Negative symptoms	123	16.74	6.40	123	16.73	6.41
	Activation	123	8.37	3.13	123	8.46	3.22
	Dysphoric mood	123	8.98	3.57	123	8.96	3.50
	Autistic preoccupation	123	10.34	3.48	123	10.23	3.32
WAIS-R[2]	Full scale intelligence (FIQ)	134	90.34	13.79	123	90.23	14.96
	Verbal intelligence (VIQ)	134	94.13	14.70	123	93.97	14.72
	Performance intelligence (PIQ)	134	86.73	15.42	123	86.74	15.68
	VIQ-PIQ	134	7.40	11.09	123	7.23	11.15
WAIS-R, Verbal subtests	Arithmetic	136	8.13	2.87	123	8.11	2.79
	Digit span	136	8.88	2.93	123	8.80	2.94
	Information	136	9.49	3.29	123	9.47	3.24
	Vocabulary	136	9.40	3.39	123	9.30	3.43
	Comprehension	136	8.26	3.26	123	8.26	3.34
	Similarities	136	8.90	2.90	123	8.98	2.95
WAIS-R, Performance subtests	Object assembly	136	7.69	3.31	120	7.72	3.40
	Picture Arrangement	136	7.60	2.93	123	7.59	2.96
	Picture completion	136	7.34	3.20	123	7.43	3.28
	Digit symbol	136	6.74	2.58	122	6.80	2.70
	Block design	136	8.29	3.30	122	8.29	3.33
Smell Identification (UPSIT[3])	Total	69	32.01	4.75	66	32.21	4.70

1: PANSS: Positive and Negative Syndrome Scale (Kay et al., 1989); 2: WAIS-R: Wechsler Adult Intelligence Scale--Revised (Wechsler, 1981); 3: UPSIT: University of Pennsylvania Smell Identification Test (Doty et al., 1984)

Table 12: First k-means cluster analysis[1]: means (and standard deviations) of the demographic, clinical, neuropsychological and olfaction variables according to cluster.

Cluster	N	Age onset	Sex	Diagnosis	Maternal age	Paternal age	PARS	Family history	UPSIT	Deficit syndrome
1	34	20.12	0.35	0.79	28.12	32.53	0.15	0.50	30.11	2.73
		(6.03)	(0.49)	(0.41)	(6.57)	(7.17)	(0.36)	(0.51)	(5.24)	(0.45)
2	19	20.11	0.53	0.74	27.08	29.86	0.16	0.05	34.17	2.94
		(5.18)	(0.51)	(0.45)	(3.86)	(4.43)	(0.37)	(0.23)	(4.55)	(0.25)
3	24	22.63	0.50	0.71	33.23	41.47	0.83	0.13	32.56	2.89
		(8.02)	(0.51)	(0.46)	(6.36)	(6.46)	(0.38)	(0.34)	(4.93)	(0.32)
4	13	24.08	0.62	0.69	29.13	31.87	0.15	0.38	33.25	2.92
		(3.93)	(0.51)	(0.48)	(4.31)	(5.68)	(0.38)	(0.51)	(3.41)	(0.28)
5	15	19.80	0.00	0.73	26.60	30.93	0.13	0.20	30.00	2.92
		(5.88)	(0.00)	(0.46)	(4.61)	(5.98)	(0.35)	(0.41)	(5.83)	(0.28)
6	2	17.00	0.00	1.00	29.00	32.50	0.50	0.00	31.00	3.00
		(1.41)	(0.00)	(0.00)	.	(6.36)	(0.71)	(0.00)	.	(0.00)
7	29	25.03	0.55	0.79	25.95	27.82	0.03	0.41	32.67	2.92
		(7.15)	(0.51)	(0.41)	(5.11)	(5.04)	(0.19)	(0.50)	(3.64)	(0.28)

	PANSS positive symptoms	PANSS negative symptoms	PANSS general symptoms	PANSS-5, negative symptoms	PANSS-5, positive symptoms	PANSS-5, activation	PANSS-5, dysphoric mood	PANSS-5, autistic preoccup.	PIQ	VIQ	VIQ-PIQ
1	15.16	17.84	28.84	20.56	11.31	10.22	9.34	12.63	71.32	76.62	5.29
	(6.95)	(5.13)	(7.13)	(6.20)	(5.08)	(4.05)	(3.53)	(3.23)	(6.13)	(6.98)	(7.67)
2	12.56	12.75	25.75	16.44	10.13	7.88	9.69	10.00	92.00	98.89	6.89
	(4.59)	(6.89)	(6.16)	(9.64)	(3.86)	(2.16)	(3.93)	(2.61)	(7.23)	(7.27)	(10.21)
3	12.36	13.27	25.50	16.05	9.50	7.82	8.82	10.18	83.35	96.22	12.87
	(5.23)	(5.29)	(6.92)	(5.52)	(4.32)	(2.15)	(3.40)	(3.46)	(5.97)	(6.32)	(10.42)
4	10.50	11.67	23.75	14.17	8.67	7.42	9.67	9.17	104.77	119.54	14.77
	(3.83)	(4.23)	(5.40)	(4.67)	(3.65)	(1.68)	(4.19)	(2.95)	(8.02)	(8.20)	(11.81)
5	9.85	11.00	20.08	13.77	7.38	7.00	6.54	7.54	106.40	102.13	-4.27
	(2.97)	(4.28)	(3.43)	(4.78)	(2.79)	(1.68)	(2.03)	(1.76)	(8.20)	(7.52)	(10.49)
6	13.50	12.50	23.00	14.00	10.50	7.00	8.50	10.00	142.50	126.50	-16.00
	(0.71)	(2.12)	(2.83)	(0.00)	(0.71)	(1.41)	(0.71)	(5.66)	(14.85)	(12.02)	(2.83)
7	12.54	13.04	26.81	15.69	10.08	8.12	9.19	9.85	81.75	92.04	10.29
	(5.60)	(4.66)	(7.08)	(4.67)	(5.16)	(3.44)	(3.75)	(3.67)	(5.82)	(6.22)	(8.61)

	Info	Picture comp	Digit span	Picture arrang	Vocab	Block design	Arith	Object assembly	Comp	Digit symbol	Similarities
1	5.91	5.21	6.41	5.15	5.62	5.21	5.59	4.91	4.74	4.53	5.68
	(2.47)	(2.28)	(2.27)	(1.56)	(1.84)	(1.45)	(1.92)	(2.02)	(1.78)	(1.54)	(1.51)
2	10.74	7.47	10.05	8.68	10.58	10.37	8.79	9.42	9.42	8.53	9.63
	(2.00)	(1.87)	(2.12)	(1.63)	(2.43)	(2.39)	(2.90)	(2.78)	(2.36)	(2.37)	(1.89)
3	9.67	6.42	9.21	7.25	9.79	7.42	9.42	7.17	8.50	6.63	9.13
	(2.35)	(2.15)	(2.17)	(2.01)	(2.17)	(2.06)	(2.19)	(2.97)	(2.13)	(1.88)	(2.03)
4	14.15	11.15	10.85	11.23	14.54	10.62	11.08	10.00	14.00	9.08	13.69
	(1.46)	(2.97)	(2.30)	(2.24)	(2.26)	(1.50)	(1.61)	(2.65)	(2.24)	(2.84)	(1.70)
5	11.40	11.60	11.93	10.47	10.40	12.40	9.13	10.87	9.53	8.53	10.07
	(1.92)	(1.50)	(3.13)	(1.77)	(2.47)	(2.38)	(2.50)	(2.85)	(1.85)	(2.50)	(1.49)
6	13.00	15.00	12.50	14.50	14.00	19.00	13.50	16.50	11.50	9.50	15.00
	(2.83)	(4.24)	(0.71)	(4.95)	(2.83)	(0.00)	(3.54)	(0.71)	(3.54)	(0.71)	(2.83)
7	9.41	6.07	8.03	6.45	9.59	7.34	7.38	6.97	8.00	6.07	8.86
	(2.31)	(1.75)	(2.13)	(2.31)	(2.34)	(1.84)	(2.13)	(1.74)	(1.71)	(1.67)	(1.55)

1: Variables used in generating the clusters: age of onset of psychosis, sex, paternal age, family history of schizophrenia, WAIS-R verbal subtests (arithmetic, digit span, information, vocabulary, comprehension, similarities) and WAIS-R performance subtests (object assembly, picture arrangement, picture completion, digit symbol, block design).

Table 13: Second k-means cluster analysis[1]: means (and standard deviations) of the demographic, clinical, neuropsychological and olfaction variables according to cluster.

Cluster	N	Age onset	Sex	Diagnosis	Maternal age	Paternal age	PARS	Family history	UPSIT	Deficit syndrome
1	21	21.24	0.10	0.95	27.28	30.41	0.05	0.52	29.44	2.47
		(5.84)	(0.30)	(0.22)	(5.25)	(4.74)	(0.22)	(0.51)	(3.85)	(0.51)
2	14	17.21	0.93	0.57	29.18	39.32	0.71	0.00	35.40	3.00
		(4.68)	(0.27)	(0.51)	(6.78)	(8.83)	(0.47)	(0.00)	(2.70)	(0.00)
3	22	19.82	0.00	0.73	29.74	34.05	0.41	0.05	31.00	2.90
		(4.68)	(0.00)	(0.46)	(6.33)	(7.24)	(0.50)	(0.21)	(7.95)	(0.30)
4	21	21.33	0.24	0.81	26.75	28.70	0.14	0.19	32.90	3.00
		(5.16)	(0.44)	(0.40)	(4.65)	(5.21)	(0.36)	(0.40)	(2.88)	(0.00)
5	19	29.95	1.00	0.74	29.16	35.18	0.16	0.68	35.31	2.88
		(7.41)	(0.00)	(0.45)	(5.86)	(7.80)	(0.37)	(0.48)	(2.93)	(0.33)
6	11	25.00	0.91	0.55	29.97	31.95	0.18	0.09	31.67	3.00
		(5.71)	(0.30)	(0.52)	(4.04)	(4.08)	(0.40)	(0.30)	(4.23)	(0.00)
7	15	17.53	0.13	0.73	26.83	33.25	0.13	0.67	31.50	3.00
		(3.38)	(0.35)	(0.46)	(7.32)	(9.02)	(0.35)	(0.49)	(4.66)	(0.00)

	PANSS positive symptoms	PANSS negative symptoms	PANSS general symptoms	PANSS-5, negative symptoms	PANSS-5, positive symptoms	PANSS-5, activation	PANSS-5, dysphoric mood	PANSS-5, autistic preoccup.	PIQ	VIQ	VIQ-PIQ
1	8.86	23.19	27.19	26.29	6.33	9.52	7.62	10.24	80.00	87.67	7.67
	(1.85)	(4.03)	(4.73)	(6.34)	(1.74)	(2.11)	(3.46)	(2.62)	(12.84)	(15.13)	(7.36)
2	13.43	13.29	25.21	16.43	10.29	8.14	9.14	10.29	84.00	90.21	6.21
	(4.16)	(4.30)	(4.17)	(5.09)	(3.71)	(2.63)	(3.03)	(2.70)	(8.73)	(13.24)	(8.59)
3	10.50	12.00	23.55	14.55	8.18	6.95	8.68	9.50	79.68	97.00	17.32
	(3.38)	(2.81)	(4.87)	(3.46)	(2.94)	(1.46)	(3.30)	(2.54)	(9.62)	(11.84)	(9.55)
4	10.90	9.71	20.57	11.90	8.71	7.19	7.33	8.43	105.33	100.76	-4.57
	(2.98)	(1.98)	(3.67)	(2.17)	(3.00)	(1.54)	(2.80)	(2.46)	(17.79)	(13.55)	(8.92)
5	9.42	10.53	21.95	13.11	7.95	6.74	8.37	8.58	88.74	99.32	10.58
	(3.13)	(3.08)	(4.78)	(3.35)	(3.15)	(1.45)	(3.44)	(2.48)	(14.37)	(15.43)	(9.25)
6	20.45	13.18	35.91	16.91	16.18	10.55	13.55	13.27	85.18	98.27	13.09
	(5.47)	(2.48)	(6.59)	(3.83)	(4.05)	(4.93)	(3.27)	(4.17)	(12.25)	(15.34)	(7.06)
7	20.40	15.40	32.00	18.07	15.07	11.93	10.73	13.60	81.67	82.40	0.73
	(4.10)	(3.81)	(6.07)	(5.24)	(3.77)	(4.80)	(1.87)	(3.42)	(12.81)	(9.93)	(9.16)

	Info	Picture comp	Digit span	Picture arrang	Vocab	Block design	Arith	Object assembly	Comp	Digit symbol	Similarities
1	7.90	6.52	8.24	6.95	7.86	6.76	7.62	6.65	6.90	5.90	7.29
	(3.92)	(2.86)	(3.03)	(2.69)	(3.20)	(2.51)	(3.17)	(3.70)	(3.21)	(2.21)	(2.72)
2	8.29	6.79	8.07	7.79	9.07	7.57	7.79	7.93	7.93	7.71	8.36
	(3.29)	(2.04)	(3.17)	(2.19)	(2.87)	(2.38)	(2.69)	(3.36)	(2.50)	(3.02)	(2.65)
3	10.14	6.18	9.32	6.32	9.59	7.68	8.86	6.59	8.59	5.18	9.50
	(2.53)	(2.56)	(2.93)	(2.28)	(3.14)	(3.08)	(2.64)	(2.56)	(3.49)	(1.56)	(2.24)
4	10.48	10.76	10.90	9.81	10.10	11.95	9.33	10.45	9.33	8.86	10.38
	(2.48)	(3.36)	(2.74)	(2.86)	(2.76)	(3.46)	(2.56)	(3.55)	(2.78)	(2.90)	(2.75)
5	10.21	7.74	8.47	8.26	11.32	8.05	8.53	7.37	10.32	7.63	10.05
	(3.29)	(3.45)	(2.20)	(3.25)	(3.23)	(3.27)	(2.52)	(3.15)	(3.84)	(2.54)	(3.01)
6	11.00	6.45	8.73	7.55	11.09	8.00	7.36	7.91	9.00	6.36	10.18
	(3.16)	(2.73)	(3.26)	(3.47)	(4.21)	(2.79)	(3.11)	(3.24)	(2.32)	(2.01)	(3.16)
7	8.33	6.80	7.07	6.27	6.13	7.29	6.33	7.21	5.33	5.86	7.00
	(2.99)	(3.08)	(2.02)	(2.63)	(2.36)	(2.40)	(2.02)	(2.78)	(2.13)	(2.60)	(2.56)

1: Variables used in generating the clusters: age of onset of psychosis, sex, paternal age, family history of schizophrenia, VIQ-PIQ, PANSS subscales (standard model: positive, negative and general psychopathology subscale scores).

2.2 Hierarchical Cluster Analysis

2.2.1 Methods

Hierarchical methods start with the individual observation, and these initial objects are merged according to their similarity (distance matrix) until the observations become one cluster. A dendrogram displaying the result of clustering is given in Figure 11. Among the various linkage criteria for determining the distance between clusters, average linkage methods were used for generating clusters in this study. Other distance measures will be tested in our future work. Average linkage treats the distance between two clusters given below as the average distance between elements of each cluster.

Distance between two clusters A and B = $\frac{1}{|A|\times|B|}\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} d(x,y)$

where, |A| is the cardinality of the cluster A which is the number of elements of the set A

|B| is the cardinality of the cluster B which is the number of elements of the set B

Unlike *K-means* clustering, there are criteria for finding an optimal number of clusters for hierarchical clustering. The pseudo *F*, pseudo t^2 statistics along with CCC (Cubic Clustering Criterion) are used for determining the number of clusters. It has been recommended to look for consensus among these statistics. We look for local peaks of the pseudo *F* combined with a small value of the pseudo t^2 statistic and a larger pseudo t^2 for the next cluster fusion. A high CCC value is preferred.

2.2.2 Application

The goal of this study is to identify a readily usable biomarker associated with core features of PARS for use in treatment studies. This sample included 114 psychiatric cases diagnosed with schizophrenia (n=60), schizoaffective disorder (n=19), bipolar disorder (n=20) or major depressive disorder (MDD; n=14) recruited from a large urban state psychiatric facility from 1992 to 2007. Additionally, 51 controls with no history of mental illness were recruited from the community; one participant had missing diagnostic information. All participants signed informed consent forms and research procedures were approved by the local Institutional Review Board. Among them, only 10 patients were from the PARS group. The data contain demographic information, Neuropsychological Test Scores, PANSS scales, odor detection score and global assessment of function scores (GAF). We have here used hierarchical cluster analysis with the average linkage method to predict association of odor acuity and symptoms of schizophrenia subgroups. Other statistical analysis such as regression and correlation were also used to support our findings.

In the cluster analysis, 2 demographic measures with UPSIT (smell identification test score), MNTHRES (Mean OLF Threshold) and neuropsychological domains are used to generate the clusters. As we highlighted in Figure 11, Cluster 13 shows relatively high pseudo *F,* large jump in pseudo t^2 statistic and positive CCC number. Based on the criterion, 13 clusters have been chosen for detecting characteristics among the schizophrenia patients. Our choice of cluster 13 is also represented as the vertical line

through the 3 different plots in Figure 10 and the horizontal line separates the tree chart in Figure 11.

Table 14: Hierarchical clustering: Cluster criterions for the number of clusters.

NCL	Cluster Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Dist
15	CL24	CL55	6	0.0047	0.836	0.827	1.35	33.8	4	0.5011
14	CL19	11023	9	0.0031	0.833	0.82	1.9	36	2.1	0.5025
13	CL16	CL35	41	0.0097	0.823	0.813	1.48	36.8	5.4	0.5388
12	CL17	CL14	46	0.0168	0.806	0.805	0.16	36.3	9.5	0.5444
11	CL13	CL15	47	0.0112	0.795	0.796	-0.21	37.6	5.6	0.5503
10	CL32	10024	3	0.0034	0.791	0.787	0.56	41.3	2.3	0.5629
9	CL12	5593	47	0.0038	0.788	0.777	1.38	45.9	1.8	0.5631
8	CL11	CL18	50	0.0089	0.779	0.765	1.59	50.3	4.1	0.5881
7	1014	CL26	3	0.0039	0.775	0.752	2.42	57.9	2.4	0.6009
6	CL9	CL10	50	0.0094	0.765	0.737	2.68	66.6	4.3	0.6152
5	CL8	CL33	53	0.0119	0.753	0.72	2.78	78.7	5.2	0.6297
4	CL6	CL7	53	0.0112	0.742	0.698	1.81	99.9	4.8	0.6571
3	CL4	11018	54	0.0068	0.736	0.657	3.16	146	2.7	0.7069
2	CL3	1001	55	0.0092	0.726	0.55	7.47	281	3.6	0.799
1	CL5	CL2	108	0.7263	0	0	0	.	281	1.3079

1: NCL: Number of cluster used in the clustering analysis. FREQ: frequency, SPRSQ: semi partial R square, RSQ: R square, ERSQ: expected value of R square, CCC: cubic clustering criterion, PSF: pseudo F, PST2: pseudo T square statistics and Dist: distance between two clusters.

Figure 10: Graphical view of CCC, pseudo F and pseudo T-square statistics for the number of clusters.

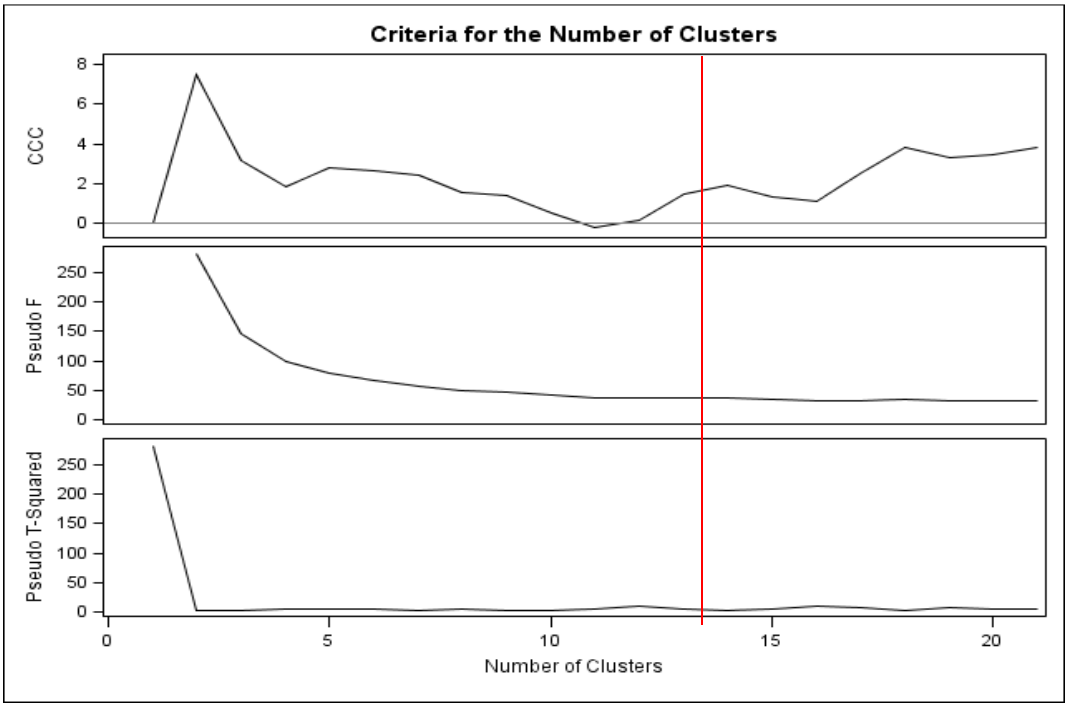
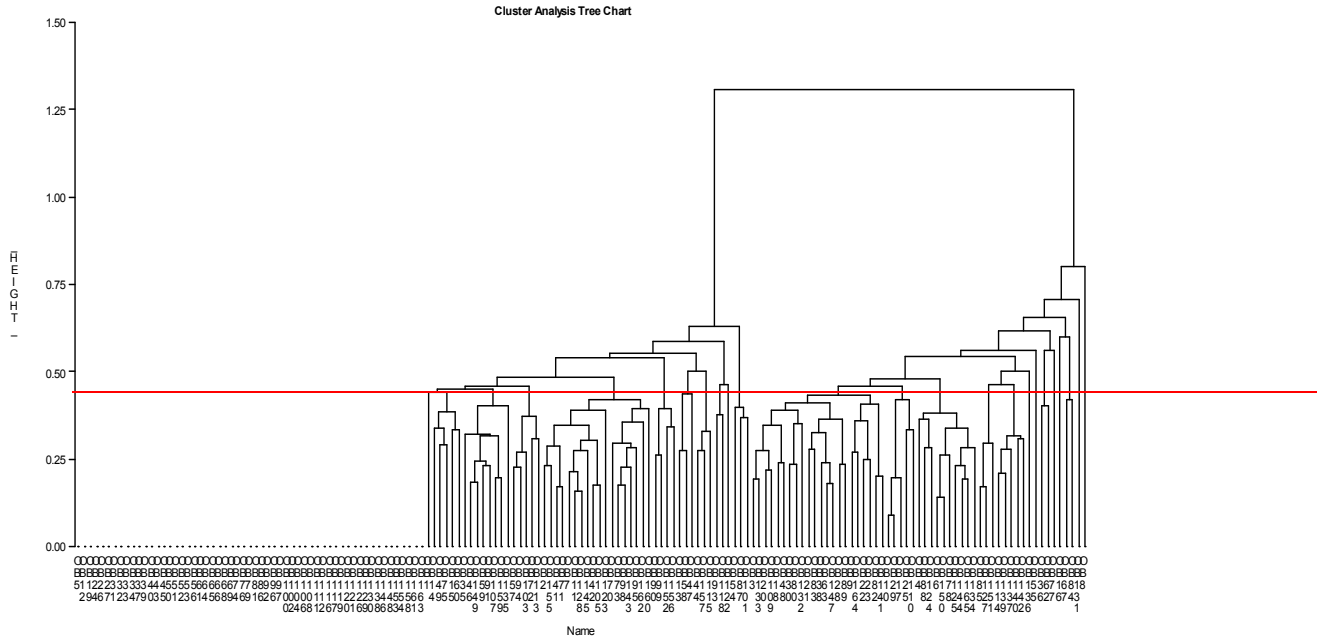


Figure 11: Tree chart for hierarchical clustering



2.2.3 Results

Table 15 shows the cluster analysis with the 13 clusters, which includes 73 psychotic cases (65.2%), 16 schizo-affective, 37 schizophrenia, 11 bipolar and 9 MDD cases. It is interesting to see that Cluster 1 only contains male patients with large proportion of schizophrenia cases and Cluster 2 contains only female cases. The average of the core factors is provided in Table 16. Patients in Cluster 1 were mostly schizophrenia and schizo-affective patients with high positive symptom scale (Mean: 12.83, SD: 5.56) and high difference in GAF score. These patients also show extremely low scores in all Wechsler Adult Intelligence Scale (WAIS). Cluster 2 contains only female patients with 63% bipolar and more than half of MDD cases. These patients have shown high negative symptom scores (NTOT) and high GAF current scores (Mean: 37.64, SD: 15.98). Perhaps analogously, we found increased acuity predicted more depression but better work function, whereas lesser odor detection acuity was associated with mania, social fear and avoidance (Hardy, in press). In this cluster, POI (Perceptual organization index) score is relatively higher whereas WMI (Working memory index) score is lower than overall cases.

We found that the sensitivity for odor detection (acuity) decreases as paternal age advances in both male and female controls ($r=-.40$, $n=37$, $p=.01$); whereas it conversely increases in schizophrenia ($r= .243$, $n=57$, $p= .067$). We next looked at the association of odor acuity with course and symptoms in schizophrenia, finding that lesser acuity predicted declining function, assessed as changes in global assessment of function scores (GAF) between the current interview and previous month ($r=-.432$, $n=50$, $p=. 002$). We explored

the symptoms in the PANSS to examine which features explained the association between acuity and functional decline. Backward elimination from a full regression model showed increased positive symptoms predicted the relationship between less acuity and greater GAF decline.

Descriptive Statistics (N=165)

Variable	N	Mean	SD
Age	165	35.1	11.0
Onset	103	22.5	7.3
Mom age	122	27.8	6.4
Dad age	115	31.4	6.9
MNTHRES	128	4.5	1.5
UPSIT	144	32.0	4.4
PTOT	136	10.4	5.6
NTOT	138	11.7	5.5
GTOT	138	23.9	8.8
GAF_CUR	66	35.7	14.5
GAF_PAS	67	46.0	12.5
VIQ	132	102.6	15.6
PIQ	132	96.1	15.7
VIQ-PIQ	132	6.4	12.3
VCI	133	105.2	16.2
POI	132	96.8	15.8
WMI	133	98.2	15.0
PSI	131	95.1	14.7

Table 15: Hierarchical clustering: frequency table of DIAGALL and gender by cluster.

DIAGALL by CLUSTER

	1	2	3	4	5	6	7	8	9	10	11	12	13	total
Schizo-aff	7	6	0	1	0	0	0	0	1	1	0	0	0	16
Schizo	17	9	1	2	3	0	1	2	0	0	1	0	1	37
Bipolar	1	7	1	1	0	1	0	0	0	0	0	0	0	11
MDD	1	5	1	1	0	0	1	0	0	0	0	0	0	9
total	26	27	3	5	3	1	2	2	1	1	1	0	1	73

SEX by CLUSTER

Frequency	1	2	3	4	5	6	7	8	9	10	11	12	13	total
Male	26	0	3	0	0	0	2	2	1	1	1	0	1	37
Female	0	27	0	5	3	1	0	0	0	0	0	0	0	36
total	26	27	3	5	3	1	2	2	1	1	1	0	1	73

Table 16: Hierarchical clustering: means (and standard deviations) of the demographic, clinical, neuropsychological and olfaction variables according to cluster.

CLUSTER		AGE	MOM AGE	DADA GE	ONSET	MNTHRES	UPSIT	PTOT	NTOT	GTOT	GAF_CUR	GAF_PAS
1	N	26	17	17	22	26	26	24	24	24	20	20
	MEAN	39.42	26.33	29.87	22.95	**3.78**	32.04	**12.83**	12.92	25.58	**30.15**	**46.45**
	STD	10.20	7.56	7.53	5.35	0.99	3.47	5.56	4.39	7.08	14.44	12.11
2	N	27	20	17	26	27	27	25	25	25	14	14
	MEAN	37.24	25.82	30.57	22.35	4.57	33.37	9.56	**13.20**	26.20	**37.64**	45.93
	STD	10.28	6.39	7.59	8.74	1.01	2.94	3.62	5.59	6.06	15.98	13.31
3	N	3	2	2	2	3	3	3	3	3	1	1
	MEAN	27.54	31.34	34.34	17.50	4.67	31.33	8.67	12.00	23.33	60.00	60.00
	STD	3.06	0.95	3.29	0.71	0.45	4.04	1.53	3.61	4.73		
4	N	5	4	5	4	5	5	4	4	4	3	3
	MEAN	36.41	23.34	27.87	27.25	4.08	31.40	11.50	9.50	21.75	25.33	44.33
	STD	10.21	5.27	9.11	11.32	1.64	2.61	3.32	3.11	6.65	4.51	15.04
5	N	3	3	3	3	3	3	3	3	3	1	1
	MEAN	27.74	28.97	36.49	20.67	9.10	33.00	7.67	11.00	21.33	45.00	25.00
	STD	5.03	5.54	7.46	2.31	0.78	4.36	1.15	0.00	3.21		
6	N	1	1	1	1	1	1	1	1	1	0	0
	MEAN	23.00	27.24	31.39	19.00	3.58	26.00	7.00	9.00	20.00		
	STD											
7	N	2	1	1	2	2	2	2	2	2	2	2
	MEAN	47.83	21.13	30.69	27.50	5.71	26.50	14.50	18.50	35.00	39.00	43.00
	STD	12.98			12.02	0.30	4.95	10.61	2.12	16.97	1.41	7.07
8	N	2	0	1	2	2	2	2	2	2	0	1
	MEAN	35.58		37.00	21.50	5.38	33.50	10.50	10.50	21.50		50.00
	STD	14.60			4.95	0.86	3.54	4.95	4.95	4.95		
9	N	1	1	1	1	1	1	1	1	1	0	0
	MEAN	38.69	26.18	41.49	22.00	7.73	23.00	12.00	15.00	28.00		
	STD											
10	N	1	0	0	1	1	1	1	1	1	1	1
	MEAN	21.72			21.00	4.35	29.00	12.00	18.00	27.00	25.00	32.00
	STD											
11	N	1	1	1	1	1	1	1	1	1	1	1
	MEAN	23.00	31.00	40.00	18.00	2.00	21.00	18.00	7.00	22.00	40.00	61.00
	STD											
12	N											
	MEAN											
	STD											
13	N	1	1	0	1	1	1	0	0	0	1	1
	MEAN	48.00	18.00		25.00	2.50	11.00				21.00	55.00
	STD											

CLUSTER	viq	piq	viqp	Fsiq	vci	poi	wmi	psi
1	26	26	26	26	26	26	26	26
	93.08	**87.62**	5.46	**89.88**	**95.19**	**90.58**	**90.88**	**86.00**
	15.28	11.28	11.45	13.36	16.22	11.14	13.26	11.19
2	27	27	27	27	27	27	27	27
	101.56	98.85	2.70	100.41	106.15	**99.41**	**94.78**	95.04
	14.58	13.50	12.52	13.39	15.61	12.83	10.99	14.41
3	3	3	3	3	3	3	3	3
	111.67	118.67	-7.00	115.67	106.67	113.33	123.33	109.00
	8.50	15.53	17.06	9.29	4.93	15.63	11.68	11.36
4	5	5	5	5	5	5	5	5
	99.20	77.40	21.80	88.60	107.80	73.40	82.80	78.60
	18.43	17.10	3.03	19.11	17.98	9.66	16.30	6.95
5	3	3	3	3	3	3	3	3
	105.00	103.00	2.00	104.33	111.00	105.33	96.33	102.00
	5.00	11.53	8.54	7.51	5.20	10.97	8.02	15.59
6	1	1	1	1	1	1	1	1
	107.00	99.00	8.00	104.00	109.00	95.00	102.00	128.00
7	2	2	2	2	2	2	2	2
	115.50	107.00	8.50	113.00	111.50	116.00	124.50	92.50
	12.02	2.83	9.19	8.49	9.19	2.83	7.78	9.19
8	2	2	2	2	2	2	2	2
	134.00	139.00	-5.00	139.50	140.50	141.50	113.50	106.00
	1.41	19.80	18.38	9.19	6.36	12.02	7.78	0.00
9	1	1	1	1	1	1	1	1
	100.00	119.00	-19.00	108.00	103.00	123.00	97.00	108.00
10	1	1	1	1	1	1	1	1
	75.00	80.00	-5.00	76.00	72.00	84.00	78.00	63.00
11	1	1	1	1	1	1	1	1
	130.00	121.00	9.00	128.00	131.00	133.00	113.00	108.00
12								
13	1	1	1	1	1	1	1	1
	79.00	72.00	7.00	74.00	84.00	72.00	82.00	76.00

Chapter 3

Classification Analysis

3.1 Methods

RF

RF models consisting of 500 trees were used. The analysis was performed with the R statistical software package RandomForest (Liaw and Wiener, 2002) using all the default settings. We used two different decision thresholds, 0.5 and 0.69 (114/165, the proportion of patients in the data), to improve the balance between sensitivity and specificity (Ahn et al., 2007).

SVM

R package e1071 was used with all default settings. We considered both the radial basis and linear kernels, but only the linear basis kernel was reported because it showed a higher accuracy. The adjusted decision threshold was also used in SVM with linear kernel.

LDA, AdaBoost

A package MASS in R was employed for LDA in the comparison. We used the default options. Among various boosting algorithms, AdaBoost was adopted using the R package adabag with all default settings. The number of boosting iterations was set as the default of *mfinal*=100.

Decision Threshold

When a data set has highly unbalanced class sizes, the classification result tends to be biased toward the majority class. In order to resolve this problem, we assigned an adjusted decision threshold by using the proportion of the sizes of the majority class as the decision threshold (see Ahn et al., 2007). For the schizophrenia data, the decision threshold becomes the proportion of patients (114/165). This threshold is applied to RF and SVM model in hope for improved balance between sensitivity and specificity.

The efficiency of these models on these data was evaluated by overall accuracy, sensitivity, specificity, PPV and NPV. Overall accuracy was obtained by the total number of correct predictions divided by the total number of predictions. Sensitivity measures the proportion of positive predictions among actual positives

and specificity is the proportion of negative predictions among actual negatives. PPV is the proportion of true positives among the positive predictions and NPV is the proportion of true negatives among the negative predictions. In our study, patients with disease are regarded as people in the positive class, and the control group is regarded as the negative class. In this study, we conducted 20 repetitions of 10-fold cross-validation using the same random data sets for different models. For cross-validation, 10-fold cross-validation is widely used although computation power allows using more folds (Kohavi 1995).

McNemar's test was used for comparison of the significance of difference in prediction measures between two models. We also used the area under the ROC curve (AUC) to compare the performance of the classification algorithms. The ROC curve is a graphical plot which represents a trade-off between sensitivity and specificity for every possible cut off.

3.2 Example

3.2.1 Participants

This study is based on cases with schizophrenia, schizo-affective, bipolar or major depressive disorder (MDD) recruited at the New York State Psychiatric Institute (NYSPI) Schizophrenia Research Unit (SRU) from 1992 to 2007. The data contain 165 human subjects. Among them, 51 were healthy controls, 19 were schizo-affective, 60 were schizophrenia patients, 20 were bipolar patients and 14 were MDD patients. We had one missing observation. The data set contains demographic information, neuropsychological test scores, PANSS (Positive and Negative Syndrome: Kay et al., 1987) scales and global assessment function of current and past scores. A summary of the data is given in Table 17. In our analysis, the missing observations were imputed by the mean for continuous variables and mode for categorical variables.

3.2.2 Measures

The following variables are included in the data:

DIAGALL: Diagnosis, 5 categories

BRAGE_M: Mother's age at birth

BRAGE_F: Father's age at birth

SEX: male, female

FAMHXANY: Family history of schizophrenia

MNTHRES: Mean Olfactory Threshold

UPSIT: University of Pennsylvania Smell Identification Test

VIQ: Verbal IQ

PIQ: Performance IQ

FSIQ: Full Scale Intelligence Quotient

VCI, POI, WMI, PSI, WSINFS, WSPIXCS, WSDSS, WSPIXAS, WSVOCS, WSBDS, WSARITHS, WSOBASSS, WSCOMPS, WSDS YMS, WSSIMILS, WSMATRS, WSLETNMS, WSSYMSES, WSSYMBS: verbal subtests (arithmetic, digit span, information, vocabulary, comprehension, similarities) and the performance subtests (object assembly, picture arrangement, picture completion, digit symbol, block design)

PTOT: Positive Syndrome Scale

NTOT: Negative Syndrome Scale

GTOT: General Syndrome Scale

GAF-CUR: Global Assessment Function, current

GAF-PAS: Global Assessment Function, past

Table 17: Classification: Descriptive statistics of the NYSPI data

Overall (N=165)

	Age	Onset	BAGE_M	BAGE_F	MNTH-RES	UPSIT	FSIQ	PTOT	NTOT	GTOT	GAF-CUR	GAF-PAS
N	165	103	122	115	128	144	132	136	138	138	59	61
μ	35.1	22.5	27.8	31.4	4.5	32.0	100.6	10.4	11.7	23.9	37.3	46.0
SD	11.0	7.3	6.4	6.9	1.5	4.4	17.6	5.6	5.5	8.8	16.1	12.5

Diagnosis (DIAGALL, N=164): control (0), schizo-affective (1), schizo (2), bipolar (3), MDD (4)

		Onset	AGE_M	AGE_F	MNTH-RES	UPSIT	FSIQ	PTOT	NTOT	GTOT	GAF-CUR	GAF-PAS
0	N	0	40	37	41	44	37	40	41	41	0	0
	μ		29.2	32.2	4.5	33.0	106.0	7.0	7.6	16.8		
	SD		6.3	6.7	1.2	4.0	13.0	0.0	1.6	1.7		
1	N	17	11	10	17	17	17	15	15	15	12	12
	μ	22.1	24.3	27.1	4.3	31.9	95.6	12.4	13.7	27.4	32.3	39.8
	SD	7.7	5.0	7.2	1.3	4.1	19.7	4.4	4.9	8.1	9.6	8.7
2	N	58	46	46	46	54	50	57	57	57	32	34
	μ	23.2	27.8	32.6	4.4	31.0	96.7	12.9	13.6	26.5	31.0	43.7
	SD	6.1	6.5	6.3	1.9	5.1	20.1	7.0	5.8	10.0	12.0	10.4
3	N	16	16	15	12	16	18	12	13	13	9	9
	μ	21.2	27.1	30.7	4.6	32.9	100.4	8.4	11.9	25.8	55.2	56.3
	SD	8.0	7.6	8.1	0.9	3.2	13.9	1.7	5.8	4.2	14.3	12.8
4	N	12	9	7	12	12	10	12	12	12	6	6
	μ	21.6	26.7	27.1	4.5	31.8	109.0	9.8	14.1	29.1	54.5	55.8
	SD	10.8	4.7	5.9	1.0	2.9	15.7	4.6	5.9	9.0	17.9	16.9

3.2.3 Variables Included in the Analysis

In our analysis, the missing observations were imputed by the mean for continuous variables and mode for categorical variables.

In this study, we compared the performance of RF, SVM, LDA and AdaBoost with each of different response variables. The response variable was divided into two groups: schizophrenia patients and the healthy control group. We then used different symptoms of the patients as a response variable. The sets of different responses are as follows:

1) Control=0, Patient=1

2) Schizophrenia=0, other patients=1 (Schizo-affective, Bipolar & MDD)

3) Schizophrenia & Schizo-affective=0 Bipolar & MDD=1

Predictor variables included in the full model were Paternal age, Maternal age, Family history, Mean OLF Threshold, SEX, UPSIT, VIQ, PIQ, FSIQ, VCI, POI, WMI, PSI, WSINFS, WSPIXCS, WSDSS, WSPIXAS, WSVOCS, WSBDS, WSARITHS, WSOBASSS, WSCOMPS, WSDSYMS, WSSIMILS, WSMATRS, WSLETNMS, WSSYMSES, WSSYMBS, PTOT, NTOT, GTOT, GAF_CUR and GAF_PAS.

3.3 Evaluation of the Methods

Table 18 summarizes the results from each model for classifying the data into patients and the control group. We also compared the results with the decision threshold of 0.5 and the results with the threshold of 0.69 (the proportion of patients in the data). P-values of the McNemar's test for the difference in overall accuracy between RF and other methods are given in Table 18.

Based on the results in Table 18, RF shows the highest overall accuracy of 0.856. For RF, the overall accuracy with 0.69 threshold (0.856) was improved from the accuracy of 0.837 with 0.5 threshold. The variable importance ranking obtained by RF (see Table 21) indicates that the PANSS score is important in classifying the data into patients and control. When the decision threshold was 0.5, both RF and SVM gave high sensitivity (0.89 for RF and 0.87 for SVM), but they showed very low specificity. When the decision threshold of RF and SVM was changed from 0.5 to 0.69, the overall accuracy improved and the sensitivity and specificity became more balanced. Accuracy of LDA, on the other hand, was the lowest among all classifiers. AdaBoost performed well in terms of accuracy, but it showed imbalance between sensitivity and specificity. Figure 12 shows ROC of different models. ROC of RF with 0.69 threshold was above the ROC of the other models indicating that it

performed better than the other models. RF with 0.69 threshold had the highest AUC of 89% than the other models as shown in Table 18.

Table 19 shows the performance of the classification models in classifying the patients into schizophrenia and other symptoms (Schizo-affective, Bipolar and MDD). The overall prediction accuracy of RF was 0.72, which was the best among the 4 models. RF appeared to perform well in all other measures as well. The overall prediction accuracies of LDA, AdaBoost and SVM were below 0.7 in this analysis. The overall accuracy of RF was compared to that of the other models. The result indicates that the overall accuracy obtained by RF was significantly higher than that of SVM and LDA. Family history of schizophrenia was highly ranked in RF variable importance ranking. The current global assessment function score also played an important role in classifying the schizophrenia and the other group (see Table 22). Figure 13 shows ROC of each model for the classification shown in Table 19. ROC for RF and AdaBoost were very close (AUC was 74.8% for RF and 74.1% for AdaBoost). LDA and SVM also showed similar patterns in ROC.

We then divided the patients differently into two groups: one group consisted of schizophrenia and schizo-affective patents, and the other group consisted of bipolar and MDD cases (Table 20). In overall accuracy, both SVM and RF showed better performance than the other models. The RF variable importance ranking (Table

23) again showed that family history of schizophrenia is an important factor in this analysis. LDA had the lowest accuracy among all the models. The differences in overall accuracy between SVM and the other models were significant (p-values less than 0.05). AUC obtained for SVM was 76.2% (see the ROC curve in Figure 14).

In the next analysis, we performed a multi-class classification with RF, SVM, AdaBoost and LDA. First, the response was divided into three levels which were control, schizophrenia and other diseases. Next, we conducted a classification into five classes by dividing the other disease group into schizo-affective, bipolar and MDD. Table 24 provides the results of the three-way classification and Table 25 shows the results of the five-way classification. Based on the average of 20 repetitions of 10-fold cross-validation, the overall accuracy of RF performed the best in the three-way classification. For the five-way classification, AdaBoost achieved the highest prediction accuracy. Overall, our results indicate that multi-way classifications did not perform well compared to a binary classification.

Table 18: Performance (SD in parentheses) of classification algorithms into patients and control. Twenty repetitions of 10-fold CV were used for each method.

Models	Accuracy	P-value*	Sensitivity	Specificity	PPV	NPV	AUC
RF	0.837 (0.014)		0.894 (0.010)	0.710 (0.036)	0.875 (0.014)	0.751 (0.020)	0.874
RF w/0.69th	0.856 (0.009)	0.083	0.854 (0.012)	0.861 (0.028)	0.932 (0.013)	0.726 (0.015)	0.890
SVM	0.816 (0.018)	0.052	0.870 (0.019)	0.693 (0.030)	0.864 (0.013)	0.706 (0.034)	0.833
SVM w/0.69th	0.799 (0.018)	<0.001*	0.803 (0.019)	0.789 (0.040)	0.895 (0.018)	0.642 (0.025)	0.835
LDA	0.773 (0.015)	<0.001*	0.833 (0.012)	0.637 (0.038)	0.837 (0.015)	0.631 (0.022)	0.822
AdaBoost	0.812 (0.016)	0.031	0.877 (0.020)	0.669 (0.034)	0.856 (0.013)	0.709 (0.033)	0.829

*P-values of the McNemar test for the difference in overall accuracy between RF and the given method

Table 19: Performance (SD in parentheses) of classification algorithms into schizophrenia and other symptoms. Twenty repetitions of 10-fold CV were used for each method.

Models	Accuracy	P-value*	Sensitivity	Specificity	PPV	NPV	AUC
RF	0.724 (0.017)		0.745 (0.018)	0.701 (0.041)	0.743 (0.022)	0.710 (0.013)	0.748
SVM	0.674 (0.019)	<0.001*	0.643 (0.025)	0.701 (0.028)	0.656 (0.023)	0.690 (0.017)	0.697
LDA	0.674 (0.027)	<0.001*	0.652 (0.035)	0.693 (0.035)	0.653 (0.032)	0.693 (0.025)	0.691
AdaBoost	0.696 (0.027)	0.0176	0.794 (0.057)	0.609 (0.092)	0.646 (0.035)	0.775 (0.034)	0.747

*P-values of the McNemar test for the difference in overall accuracy between RF and the given method

Table 20: Performance (SD in parentheses) of classification algorithms into (Schizophrenia & Schizo-affective) and (Bipolar & MDD). Twenty repetitions of 10-fold CV were used for each method.

Models	Accuracy	P-value*	Sensitivity	Specificity	PPV	NPV	AUC
RF	0.720 (0.018)	0.0046	0.738 (0.023)	0.701 (0.035)	0.743 (0.022)	0.706 (0.017)	0.744
SVM	0.749 (0.027)		0.760 (0.039)	0.738 (0.038)	0.720 (0.030)	0.778 (0.030)	0.762
LDA	0.665 (0.028)	<0.001*	0.650 (0.048)	0.678 (0.036)	0.641 (0.029)	0.688 (0.030)	0.727
AdaBoost	0.681 (0.026)	<0.001*	0.716 (0.063)	0.649 (0.081)	0.647 (0.039)	0.724 (0.029)	0.718

*P-values of the McNemar test for the difference in overall accuracy between SVM and the given method

Table 21: Variable importance ranking for the RF result shown in Table 18

Rank	Variable	Mean Decrease Accuracy	Rank	Variable	Mean Decrease Accuracy
1	GTOT	0.05674	18	wsbds	0.00110
2	PTOT	0.03881	19	wsinfs	0.00088
3	NTOT	0.03298	20	vci	0.00076
4	FAMHXANY	0.01507	21	wspixas	0.00062
5	wsdsyms	0.00711	22	wsmatrs	0.00049
6	psi	0.00686	23	upsittot	0.00028
7	wsariths	0.00399	24	BRAGE_F	0.00016
8	viq	0.00277	25	wspixcs	0.00016
9	piq	0.00249	26	wssimils	0.00009
10	wssymses	0.00245	27	wsobasss	0.00001
11	wmi	0.00223	28	wssymbs	-0.00010
12	poi	0.00162	29	BRAGE_M	-0.00026
13	wsvocs	0.00157	30	wsdss	-0.00052
14	fsiq	0.00152	31	MNTHRES_A	-0.00185
15	SEX	0.00150	32	wspixas	0.00062
16	wscomps	0.00131	33	wsmatrs	0.00049
17	wsletnms	0.00116			

Table 22**:** Variable importance ranking for the RF result shown in Table 19

Rank	Variable	Mean Decrease Accuracy	Rank	Variable	Mean Decrease Accuracy
1	FAMHXANY	0.03086	18	viq	0.0027
2	GAF_CUR	0.00989	19	wsariths	0.002
3	fsiq	0.00776	20	wssymbs	0.00144
4	BRAGE_F	0.00756	21	poi	0.00105
5	vci	0.00672	22	NTOT	0.00089
6	MNTHRES_A	0.00635	23	wssimils	0.00078
7	GTOT	0.00603	24	wspixas	0.00063
8	wsletnms	0.00538	25	wssymses	0.00026
9	piq	0.00452	26	upsittot	0.00022
10	wsinfs	0.00436	27	wsdss	-0.00003
11	wsdsyms	0.00415	28	wscomps	-0.00034
12	GAF_PAS	0.0038	29	wsobasss	-0.00044
13	wmi	0.00356	30	wsmatrs	-0.00078
14	psi	0.00341	31	wsbds	-0.00092
15	PTOT	0.00324	32	SEX	-0.00098
16	wsvocs	0.00301	33	BRAGE_M	-0.00108
17	wspixcs	0.00284			

Table 23: Variable importance ranking for the RF result shown in Table 20

Rank	Variable	Mean Decrease Accuracy	Rank	Variable	Mean Decrease Accuracy
1	FAMHXANY	0.02986	18	wssimils	0.00205
2	BRAGE_F	0.01105	19	wspixcs	0.0019
3	wsvocs	0.01083	20	wsbds	0.00176
4	fsiq	0.00976	21	wssymbs	0.00152
5	GAF_CUR	0.00869	22	poi	0.0013
6	wsletnms	0.00637	23	NTOT	0.0013
7	vci	0.00402	24	wscomps	0.00127
8	wmi	0.004	25	wsariths	0.00107
9	GTOT	0.00393	26	wsdss	0.00037
10	GAF_PAS	0.00337	27	wssymses	0.00024
11	MNTHRES_A	0.00328	28	SEX	-0.0002
12	piq	0.00319	29	upsittot	-0.00021
13	viq	0.00299	30	wspixas	-0.00045
14	psi	0.00292	31	wsmatrs	-0.00091
15	wsinfs	0.00262	32	wsobasss	-0.00202
16	PTOT	0.00261	33	BRAGE_M	-0.00254
17	wsdsyms	0.00256			

Table 24: Performance of classification algorithms into control, Schizophrenia and other diseases. Twenty repetitions of 10-fold CV were used for each method.

Models	Accuracy	SD	Min	Max
RF	0.6484	0.0149	0.6181	0.6667
SVM	0.6387	0.0181	0.6181	0.6727
LDA	0.606	0.0178	0.5757	0.6424
AdaBoost	0.6336	0.0248	0.5939	0.6787

Table 25: Performance of classification algorithms into control, Schizophrenia, Schizo-affective, Bipolar and MDD. Twenty repetitions of 10-fold CV were used for each method.

Models	Accuracy	SD	Min	Max
RF	0.5642	0.0112	0.5454	0.5878
SVM	0.5457	0.0131	0.5333	0.5757
LDA	0.5018	0.0168	0.4787	0.5272
AdaBoost	0.5909	0.0132	0.5636	0.6121

Figure 12: ROC curve: classification algorithms into patients and control

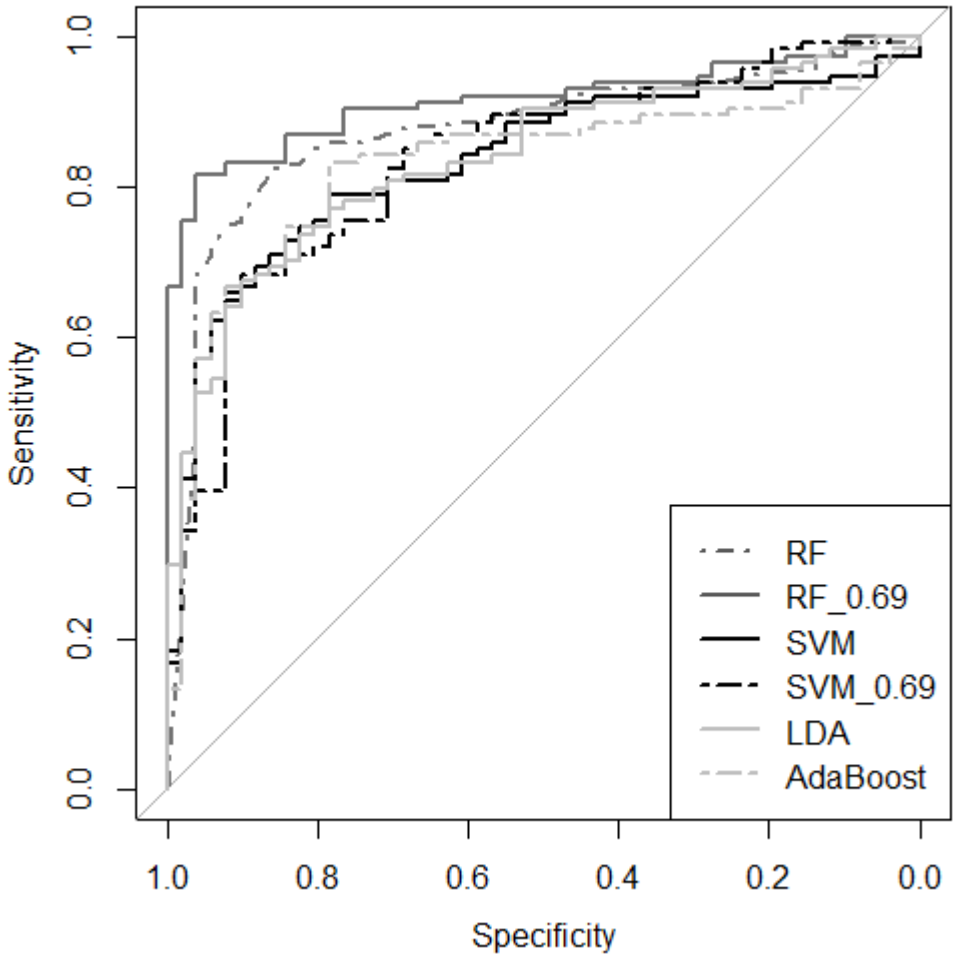


Figure 13: ROC curve: classification algorithms into Schizophrenia and other symptoms.

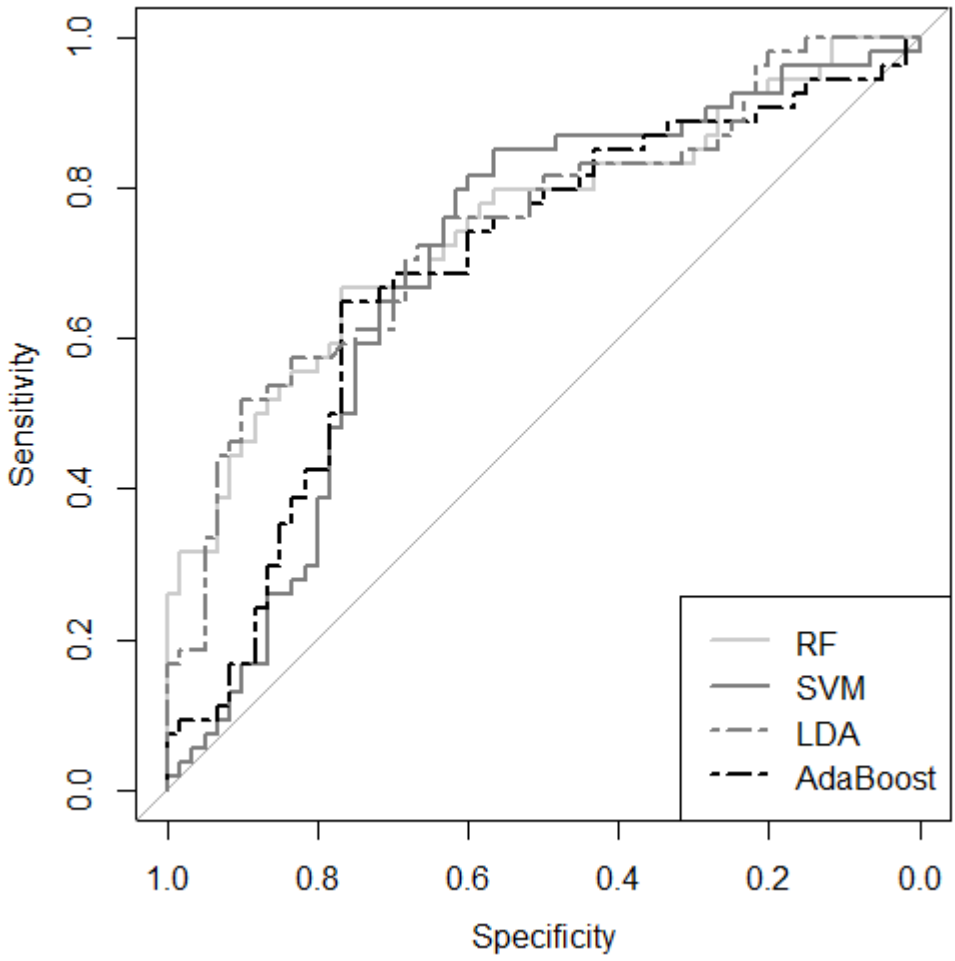
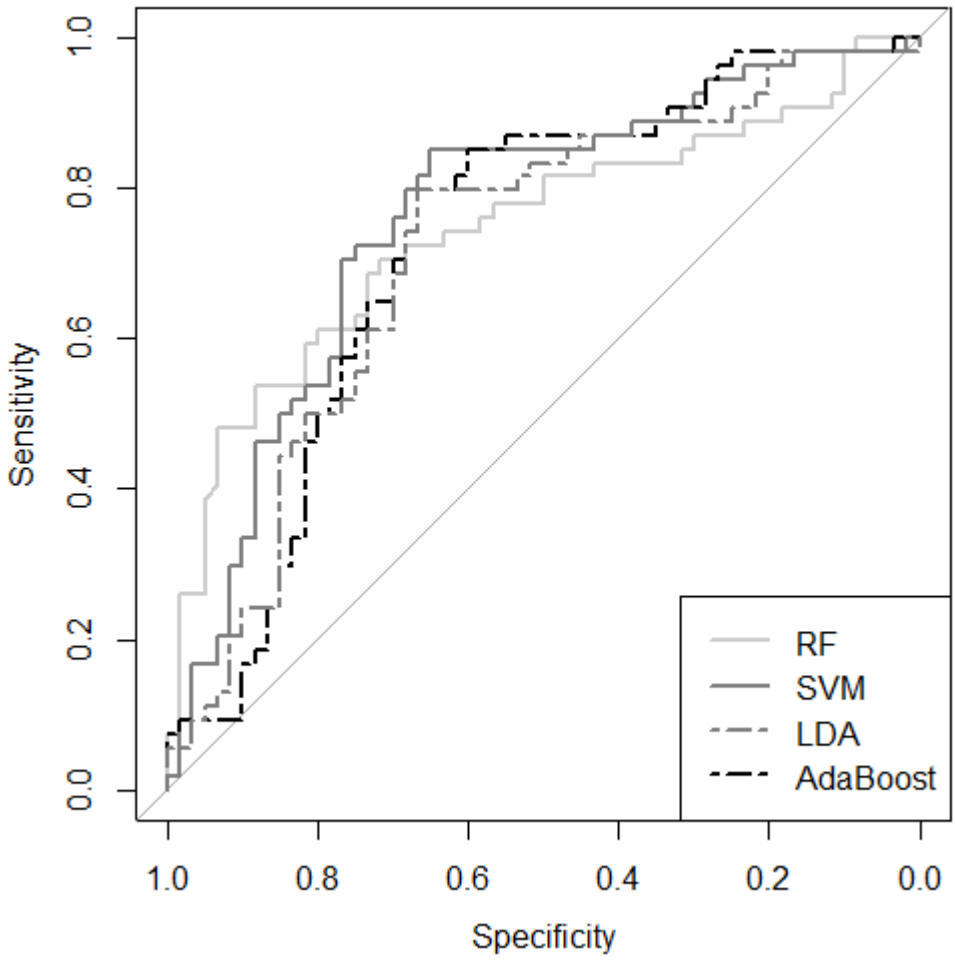


Figure 14: ROC curve: classification algorithms into (Schizophrenia & Schizo-affective) and (Bipolar & MDD)



3.4 Prediction of Potential Schizophrenia Cases

In order to detect the risk of developing schizophrenia, RF, SVM, LDA and AdaBoost were used to predict the occurrence of schizophrenia among the 51 subjects in the healthy control group. In this analysis, the data were divided into training and test sets. First, the control group was randomly divided into 3 groups: G1, G2 and G3. In the first analysis, G1 and G2 plus the 113 patients formed the learning set and G3 served as the test set. In the second analysis, G1 and G3 plus the 113 patients formed the learning set and G2 served as the test set. In the last analysis, G2 and G3 plus the 113 patients were in the learning set and G1 served as the test set.

Table 29 provides a list of subjects who were predicted as patients. Among the 51 controls, RF predicted 6 subjects, SVM and LDA predicted 7 subjects, and AdaBoost predicted 9 subjects as potential patients. The subjects who were predicted as potential schizophrenia patients by different models were not the same, but there was an overlap of prediction across the models. For example, the person with ID 11003 was classified as patient by all four methods.

In the next analysis, we used the same procedure as above, but used the four classification models to distinctively predict two different groups: schizophrenia or the other symptoms. RF predicted that 4 people potentially have a disease. Among

them, the model predicted 3 of them might have schizophrenia and one person might have other disease. The results show that the majority of the subjects at risk of the disease in this RF analysis were also predicted as potential patients in the previous analysis. LDA predicted 5 people as potential schizophrenia patients and one person as a patient of other disease. AdaBoost predicted 8 subjects and SVM predicted 4 subjects as potential schizophrenia patients and no other subject as a patient of other disease.

Finally, the 51 subjects in the control group were used to predict patients of one of the 4 different diseases and the result is shown in Table 31. In this analysis, SVM predicted the same 4 subjects predicted in the previous analysis as potential schizophrenia patients. Unlike the prediction of potential schizophrenia patients done in the beginning of this section, the predictions of SVM and RF were very different. RF classified the person with ID 11051 who was predicted as a patient with other disease in the previous analysis as a potential patient of bipolar disease. LDA also predicted the person with ID 11043 who was predicted as a patient with other disease in the previous analysis as a potential patient of bipolar disease. LDA classified two people as potential bipolar patients and additional 3 people as potential schizophrenia patients. AdaBoost predicted 7 people as potential schizophrenia patients. All four models predicted the person with ID 11053 as a potential schizophrenia patient.

Table 26: Variable importance ranking for the prediction of schizophrenia among controls

Rank	Variable	Mean Decrease Accuracy	Rank	Variable	Mean Decrease Accuracy
1	GTOT	0.05442	17	BRAGE_M	0.00123
2	PTOT	0.04126	18	wspixcs	0.00083
3	NTOT	0.03239	19	poi	0.00082
4	FAMHXANY	0.0122	20	wscomps	0.0007
5	psi	0.00698	21	wsletnms	0.00059
6	wsdsyms	0.00468	22	wsinfs	0.00048
7	wsariths	0.00438	23	wsvocs	0.00037
8	wssymses	0.00413	24	wspixas	0.00033
9	piq	0.00332	25	wssimils	0.00025
10	viq	0.00324	26	wsmatrs	0.00007
11	wmi	0.00309	27	wsdss	-0.00035
12	fsiq	0.00287	28	wsobasss	-0.0005
13	vci	0.00216	29	wssymbs	-0.00096
14	upsittot	0.00211	30	BRAGE_F	-0.0014
15	wsbds	0.00177	31	MNTHRES_A	-0.00144
16	SEX	0.00158			

Table 27: Variable importance ranking for the prediction of schizophrenia (schizophrenia vs. other diseases) among controls

Rank	Variable	Mean Decrease Accuracy	Rank	Variable	Mean Decrease Accuracy
1	GTOT	0.04597	17	wspixcs	0.00212
2	PTOT	0.04316	18	wscomps	0.00189
3	NTOT	0.03862	19	BRAGE_F	0.00166
4	FAMHXANY	0.01143	20	wssymbs	0.00163
5	psi	0.00781	21	viq	0.00138
6	wsdsyms	0.00759	22	wmi	0.00136
7	wsvocs	0.00729	23	wsbds	0.00118
8	fsiq	0.00497	24	SEX	0.00104
9	vci	0.00455	25	wssimils	0.00101
10	wsinfs	0.00447	26	MNTHRES_A	0.00093
11	wssymses	0.00389	27	wspixas	0.00033
12	wsariths	0.00357	28	wsobasss	0.00012
13	piq	0.00319	29	wsmatrs	0.00005
14	upsittot	0.00254	30	wsdss	-0.00069
15	wsletnms	0.00237	31	BRAGE_M	-0.001
16	poi	0.00215			

Table 28: Variable importance ranking for the prediction of schizophrenia (schizophrenia, schizo-affective, bipolar and MDD) among controls

Rank	Variable	Mean Decrease Accuracy	Rank	Variable	Mean Decrease Accuracy
1	PTOT	0.04972	17	poi	0.00234
2	GTOT	0.04491	18	wspixcs	0.00234
3	NTOT	0.04422	19	wsariths	0.00234
4	FAMHXANY	0.01266	20	wsbds	0.0023
5	wsdsyms	0.00982	21	Vci	0.00183
6	psi	0.00776	22	SEX	0.0018
7	piq	0.00611	23	Wspixas	0.00109
8	wsinfs	0.00561	24	Wsmatrs	0.00108
9	wscomps	0.00427	25	BRAGE_F	0.00104
10	wsvocs	0.00401	26	Wmi	0.00091
11	wssymses	0.00381	27	Upsittot	0.00036
12	fsiq	0.00347	28	MNTHRES_A	0.00008
13	wssimils	0.00314	29	Wssymbs	0.00002
14	viq	0.00275	30	Wsdss	-0.00058
15	wsobasss	0.00272	31	BRAGE_M	-0.00248
16	wsletnms	0.00251			

Table 29: Subjects in the control group who were predicted as potential patients by the classification methods

	Scizophrenia vs control			
id	RF	SVM	LDA	Ada
11001			O	O
11002		O	O	O
11003	O	O	O	O
11019		O	O	
11020			O	
11021	O	O		
11026				O
11033		O		O
11040				O
11043		O	O	
11046	O			O
11048	O			O
11051	O			O
11053	O	O	O	

Table 30: Subjects in the control group who were predicted as potential schizophrenia (schizophrenia vs. other diseases).

	Schizophrenia vsother vs control			
id	RF	SVM	LDA	Ada
11001		scz		
11002				scz
11003			scz	
11019			scz	
11020			scz	scz
11021	scz			
11026				scz
11028		scz		
11033	scz			scz
11040				scz
11043		scz	other	scz
11046	scz			
11051	other		scz	scz
11053		scz	scz	scz

Table 31: Subjects in the control group who were predicted as potential schizophrenia (schizophrenia, schizo-affective, bipolar and MDD).

	Five classes			
id	RF	SVM	LDA	Ada
11001		scz		scz
11002		scz		scz
11003			scz	
11020			scz	scz
11021	scz			
11028			bip	
11040	scz			scz
11043		scz	bip	scz
11048				scz
11051	bip			
11053	scz	scz	scz	scz

Chapter 4

Discussion and Conclusion

In contrast to other clustering methods (e.g., hierarchical clustering) there is no standard way to find the optimal number of clusters. Thus we proposed Elbow Ratio method for finding an optimal number of k in the K-means cluster analysis. In the first simulation study, Elbow Ratio and BIC showed better performance than other methods in determining the optimal number of clusters. In terms of consistency, Elbow Ratio showed better performance than BIC. In the second simulation study, the accuracy of all five methods was lower than the first experiment. However as in experiment 1, Elbow Ratio performed the best in estimating the number of clusters.

The process of our two cluster analysis and classification aim at the same goal, that is, of developing predictive model. In order to make a good prediction when solving almost any type of complicated problem, it is important to enrich our understanding of the general phenomena, and the hidden facts among the data for long term development. We defined PARS subgroup and explored the heterogeneous nature of schizophrenia symptoms. The results of clustering provide further evidence

that the genetic and neurobiological underpinnings of schizophrenia associated with the illness risk attributable to later paternal age may be different than that of other cases. Our prior findings of clustering in regards to importance of later paternal age are also confirmed through the variable importance ranking for the RF result. The RF utilized the variables paternal age and family history of schizophrenia as important predictor variables.

This study employed cluster analyses to examine if specific illness features of schizophrenia are associated with later paternal age. We identified PARS cases that clustered in groups with particular characteristics. One group was characterized by a greater differential between verbal and performance intelligence, and the other group showed a high concentration of female cases and significant early onset of psychosis.

In two different k-means clustering analyses we demonstrated, for perhaps the first time, that some of the variability in schizophrenia symptoms can be explained by PARS cases that tend to "cluster" in groups with particular characteristics. Of particular note, a subgroup of largely female cases was identified as having separate features in association with later paternal age. This finding supports PARS as a distinct subgroup. K-means clustering analysis may be a useful statistical method to examine the relationships between etiological and other symptoms in schizophrenia; to explore latent subgroups with distinct features. By overcoming the heterogeneity in

schizophrenia, we may advance our understanding of the decease and lay the groundwork for the development of new treatments.

In hierarchical clustering, Cluster 1 only contains male patients with large proportion of schizophrenia cases, which was characterized by relatively high positive symptom scale and extremely low scores in Wechsler Adult Intelligence Scale (WAIS). We found that a cluster consisting of only female cases, of which 63% were bipolar and high negative symptom scores (NTOT) and high GAF current scores. The result revealed that the sensitivity of odor detection increases as paternal age advances in schizophrenia cases.

The strengths of this study include the use of a statistical method that is particularly germane for resolving the heterogeneity of schizophrenia. The clustering model requires variability in numerous factors to generate separate independent clusters that share common attributes. Moreover, cluster analysis assures minimal variation within the clusters, but maximum variation between the clusters, creating more homogeneous subgroups. Our results showed that etiological data and clinical information can both be considered in the procedures.

Next, we considered the issue of developing predictive model by employing several classification models. Classification is an important supervised learning

method that builds predictive model based on the prior knowledge learned from the training data set. Currently, disease risk prediction has played an important role in health care research and clinical practice. Attempting prediction of disease risk has been recently studied by various machine learning techniques.

In this study, we evaluated the performance of 4 different classification models such as RF, SVM, LDA and Adaboost to predict occurrence of schizophrenia or other mental diseases. Chen et al. (2004) noted that most of the classification algorithms tend to be biased for unbalanced data because they focus on minimizing the overall error rate. Classifying unbalanced data set can be improved by adjusting the decision threshold. Our results showed that both RF and SVM with adjusted decision threshold improved the balance of sensitivity and specificity. Among these four models, RF consistently performed well in overall accuracy and AUC. Both SVM and AdaBoost showed higher performance than LDA in classification into patients and control.

We further examined the potential risk of schizophrenia or other diseases among the subjects in the control group. We hypothesized that the person who was predicted as a patient frequently might have more risk of developing schizophrenia or other mental disorders in the future. According to our study, there was an overlap of

prediction across the models, and some of the subjects were predicted as a potential patient by all four models.

In conclusion, we have attempted to explain and compare the performance of different classification models that are being applied to schizophrenia prediction. Specifically, we predicted the potential risk of individuals to develop schizophrenia by generating different classification models. We believe that the predictive rule produced by our process can be an effective channel to predict potential schizophrenia patients among seemingly healthy individuals. Overall, our studies may be particularly helpful in the development of appropriate clinical treatment, early diagnosis and will enable individuals to prepare their potential disease risk. A follow up study is needed to test and verify the result obtained in this study.

Chapter 5

Future Study

- In this study, we studied only well separated data sets in the simulation experiments. We are planning to apply our proposed method to data set containing more complicated cluster structure. In addition to this, we will work on the different data design to compare the performance of Elbow Ratio method. Since the optimal number of cluster can vary from data set to data set, various types of data design would help find the characteristic of Elbow Ratio method.

- It should be noted that our total PARS sample size for the clustering analysis was relatively small. In the first analysis we had a total of 34 PARS cases, with 20 PARS cases belonging to the high PARS concentration cluster. The second analysis only included 30 PARS cases, with 10 cases being in the high PARS concentration cluster. The hierarchical clustering also contained only 10 patients from the PARS group. Future studies that screen for higher inclusion rates of PARS cases may be successful in recruiting larger PARS

samples. Such studies may also want to include variables not examined here but found to be of etiological significance.

- For the future work, we will study other widely used classification methods and compare their performance with the four classification methods used in this study. It is expected to develop our predictive rule produced by this study. We may also consider various ensemble methods.

References

Antonius, D., Kimhy, D., Harkavy-Friedman, J., Crystal, S., Goetz, R. and Malaspina, D. Paternal age related schizophrenia and cardiac autonomic regulation profiles. *Schizophrenia Research*, 127(1):273-275, 2011.

Ahn, H., Moon, H., Fazzari, MJ., Lim, N., Chen, JJ. and Kodell, RL. Classification by ensembles from random partitions of high-dimensional data. *Computational Statistics and Data Analysis,* 51:6166-6179, 2007.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19(6):716–723, 1974.

Aldenderfer, MS. and Blashfield, RK. Cluster analysis. *Sage University Paper Series on Quantitative Applications in the Social Sciences,* 07-044, 1984

Biau, G., Devroye, L. and Lugosi, G. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2007.

Blagus, R. and Lusa, L. Class prediction for high-dimensional class-imbalanced data, *BMC Bioinformatics,* 11:523, 2010.

Breiman, L. Bagging predictors. *Machine Learning*, 24:123-140, 1996.

Breiman, L. Random forests. *Machine Learning*, 45(1):5-32, 2001.

Breiman, L, Friedman, JH., Olshen, RA. and Stone, CJ. *Classification and Regression Trees*. The Wadsworth Statistics/Probability, Belmont, CA, 1984.

Brown, AS., Schaefer, CA., Wyatt, RJ., Begg, MD., Goetz, R., Bresnahan, MA., Harkavy-Friedman, J., Gorman, JM., Malaspina, D. and Susser, ES. Paternal age and risk of schizophrenia in adult offspring. *American Journal of Psychiatry*, 159:1528-1533, 2002.

Byrne, M., Agerbo, E., Ewald, H., Eaton, WW. and Mortensen, PB. Parental age and risk of schizophrenia: a case-control study. *Archives of General Psychiatry,* 60(7):673-678, 2003.

Catherine, AS. and Gareth, MJ. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98(463): 750-763, 2003.

Chen, C., Liaw, A. and Breiman, L. Using random forests to learn imbalanced data, *Technical Report #666, Statistics Department of Statistics, University of California, Berkeley*, 2004.

Cortes, C. and Vapnik, VN. Support vector networks, *Machine Learning*, 20:273-297, 1995.

Cox, DR. The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B (Methodological),* 20(2):215–242, 1958.

Doty, RL., Shaman, P., Kimmelman, CP. and Dann, MS. University of Pennsylvania Smell Identification Test: A rapid quantitative olfactory function test for the clinic. *Laryngoscope*, 94:176–178, 1984.

Fisher, RA. The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* 7(7):179-188, 1936.

Fix, E. and Hodges, JR. Discriminatory analysis-nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique,* 57(3):238-247, 1951.

Freund, Y. and Schapire, RE. A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences,* 55(1):119-139, 1995.

Hand, DJ. and Heard, NA. Finding groups in gene expression data. *BioMed Research International,* 2:215-225, 2005.

Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2:283-304, 1998.

Jindal, R., MacKenzie, EM., Baker, GB. and Yeragani, VK. *Cardiac risk and schizophrenia. Journal of Psychiatry and Neuroscience,* 30(6):393, 2005.

Kay, SR., Fiszbein, A. and Opler, LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletine*, 13(2):261-276, 1987.

Kearns, M. Thoughts on hypothesis boosting. Unpublished manuscript, 1988.

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In C. S. Mellish (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* 6(4):1137–1143, 1995.

Lee, H., Malaspina, D., Ahn, H., Perrin, M., Opler, MG., Kleinhaus, K., Harlap, S., Goetz, R. and Antonius, D. Paternal Age Related Schizophrenia (PARS): Latent subgroups detected by k-means clustering analysis. *Schizophrenia Research,* 128:143-149, 2011.

Legendre, P. and Legendre, L. *Numerical Ecology*: Second English Edition. Developments in Environmental Modelling, 20, *Elsevier*, Netherlands 1998.

Liaw, A. and Wiener, M. Classification and regression by randomforest, *R News*, 2(3):18-22, 2002.

MacQueen, JB. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* 1(14):281-297, 1967.

Malaspina, D. Paternal factors and schizophrenia risk: de novo mutations and imprinting. *Schizophrenia Bulletin*, 27(3):379-393, 2001..

Malaspina, D., Brown, A., Goetz, D., Alia-Klein, N., Harkavy-Friedman, J., Harlap, S. and Fennig, S. Schizophrenia risk and paternal age: a potential role for de novo mutations in schizophrenia vulnerability genes. *CNS Spectrums*, 7(01):26-29, 2002.

Malaspina, D., Harlap, S., Fennig, S., Heiman, D., Nahon, D., Feldman, D. and Susser, ES. Advancing paternal age and the risk of schizophrenia. *Archives of General Psychiatry,* 58(4):361-367, 2001.

Malaspina, D., Reichenberg, A., Weiser, M., Fennig, S., Davidson, M., Harlap, S., Wolitzky, R., Rabinowitz, J., Susser, E. and Knobler, HY. Paternal age and intelligence: implications for age-related genomic changes in male germ cells. *Psychiatric Genetics,* 15(2):117-125, 2005.

McCulloch, WS. and Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

Moturu, ST., Johnson, WG. and Liu, H. Predicting future high-cost patients: a real-world risk modeling application, *IEEE International Conference on Bioinformatics and Biomedicine*, 202-208, 2007.

Nurnberger, JI. Jr., Blehar, MC., Kaufmann, CA., York-Cooler, C., Simpson, SG., Harkavy-Friedman, J., Severe, JB., Malaspina, D. and Reich, T. Diagnostic interview for genetic studies. Rationale, unique features, and training. *Archives of General Psychiatry,* 51(11):849-859, 1994.

Rao, CR. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B (Methodological),* 10(2): 159–203, 1948.

Rosenfield, PJ., Kleinhaus, K., Opler, M., Perrin, M., Learned, N., Goetz, R., Stanford, A., Messinger, J., Harkavy-Friedman, J. and Malaspina, D. Later paternal age and sex differences in schizophrenia symptoms. *Schizophrenia Research*, 116(2):191-195, 2010.

Schapire, RE. The strength of weak learnability. *Machine Learning*, 5(2):197-227, 1990.

Sipos, A., Rasmussen, F., Harrison, G., Tynelius, P., Lewis, G., Leon, DA. and Gunnell, D. Paternal age and schizophrenia: a population based cohort study. *BMJ,* 329(7474):1070, 2004.

Spiegelhalter, DJ., Best, NG., Carlin, BP., and Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583-639, 2002

Torrey, EF., Buka, S., Cannon, TD., Goldstein, JM., Seidman, LJ., Liu, T., Hadley, T., Rosso, IM., Bearden, C. and Yolken, RH. Paternal age as a risk factor for schizophrenia: how important is it?. *Schizophrenia Research*, 114(1):1-5, 2009.

Tsuang, MT., Lyons, MJ. and Faraone, SV. Heterogeneity of schizophrenia. Conceptual models and analytic strategies. *The British Journal of Psychiatry*, 156(1):17-26., 1990.

Tsuchiya, KJ., Takagai, S., Kawai, M., Matsumoto, H., Nakamura, K., Minabe, Y., Mori, N. and Takei, N. Advanced paternal age associated with an elevated risk for schizophrenia in offspring in a Japanese population. *Schizophrenia Research,* 76(2):337-342, 2005.

Vapnik, VN. The Nature of Statistical Learning Theory. *Springer Science & Business Media*, 1995.

Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. Constrained k-means clustering with background knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning,* 1: 577-584, 2001.

Wechsler, D. WAIS-R manual: Wechsler adult intelligence scale-revised. *Psychological Corporation*, 1981.

Wu, J., Roy, J. and Stewart, WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches, *Medical Care*, 48(6):106-113, 2010.

Yu, W., Liu, T., Valdez, R., Gwinn, M. and Khoury, MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, 10(1):16, 2010.

Zammit, S., Allebeck, P., Dalman, C., Lundberg, I., Hemmingson, T., Owen, MJ. and Lewis, G. Paternal age and risk for schizophrenia. *The British Journal of Psychiatry*, 183(5):405-408, 2003.

Zhang, W., Zeng, F., Wu, X., Zhang, X. and Jiang, R. A comparative study of ensemble learning approaches in the classification of breast cancer metastasis. *IEEE International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing,* 242-245, 2009.