UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Analysis of gene expression in adipocytes of individuals with metabolic syndrome by multivariate statistical methods

Jaime Manuel Pinto Combadão

Mestrado em Bioestatística

Trabalho de projeto orientado por:

Prof. Doutora Lisete Maria Ribeiro de Sousa

2017

To my family

"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."

Ronald Fisher
"Statistical methods and scientific induction.", Journal of the Royal Statistical Society, B, 17, 69-78 (1955).

Agradecimentos

Gostaria de agradecer a todo o corpo docente e aos funcionários do Departamento de Estatística e Investigação Operacional (DEIO) com quem eu tive contacto. Sabendo que seria um trabalhador estudante, com menor disponibilidade para aulas presenciais do que aquilo que gostaria, estava algo apreensivo e à espera de algumas dificuldades em seguir os estudos a bom ritmo. Felizmente isso não foi um obstáculo, todos contribuiram para facilitar as dificuldades inerentes a essa situação. Para além disso, o seu comportamento e atitude, sempre próxima e cooperativa permite-me afirmar que o DEIO é um local de ensino estável, agradável, onde realmente se procura a elevação dos estudantes nas áreas de ensino respectivas. Hoje em dia considero o DEIO como uma das minhas casas científicas.

Vários professores tiveram influência na minha formação, mas gostaria de salientar as duas professoras que considero como as minhas mentoras nesta área científica. Ambas são professoras e pessoas excelentes, o que permite aprender e crescer cientificamente num ambiente exigente, mas descontraído e estimulante.

À Prof. Doutora Lisete Sousa, a minha orientadora de mestrado, gostaria de agradecer o seu empenho no sucesso do nosso projecto. É raro encontrar alguém tão dedicado à ciência e ao mesmo tempo aos seus alunos, considero-me felizardo por a ter "encontrado". O seu papel como orientadora excedeu em muito as minhas expectativas e sem dúvida que se tornou um dos meus modelos profissionais.

À Prof. Doutora Marília Antunes, por me ter mostrado, logo no 1º semestre do mestrado, como é o raciocínio estatístico e como um bioestatístico resolve questões e problemas. Era uma das minhas grandes expectativas e foi óptimo ter sido cumprida logo no primeiro semestre. Aquelas aulas de Epidemiologia Estatística ficaram na memória, a sua clareza nos conceitos e a sua capacidade de ouvir e extrair as perguntas relevantes sempre me impressionaram.

Acho quase inacreditável que tenhamos um mestrado tão interessante, com pessoas tão válidas no nosso pequeno país.

Por último, gostaria de agradecer aos colegas, amigos e família. Tornaram tudo mais fácil.

Abstract

Metabolic syndrome is defined as a cluster of cardiovascular risk factors. Its presence is associated with the occurrence of many biologic phenomena, diseases and conditions, as insulin resistance, inflammation, oxidative stress, diabetes, mental diseases and increased severity of health problems. It is also very prevalent in modern societies due to lifestyle choices and due to the ageing of the populations. Due to human variability in behaviors, food choices, chosen environments, genetic and epigenetic traits, amongst other factors, the working definition of metabolic syndrome must be adapted to the population under study. Some previous work from other researchers suggests that a definition of metabolic syndrome as a continuous variable can be better suitable to the clinical and ambulatory settings, to effective interventions in the population and to the progress in the scientific knowledge. Besides that, it is our believe that gene expression studies (and generally genomics studies) can also benefit greatly from this redefinition.

In this work, for a male Finnish population, from whom we have clinical measures, we have redefined the metabolic syndrome as a continuous variable. This result can be used to improve the knowledge in the diagnostics and prognostics of this syndrome, in this population. Even more, with the data of the gene expression in abdominal adipocytes of these men, we have used multivariate statistical methods, as principal component analysis, non-negative matrix factorization and independent component analysis to create components/factors that are associated with the continuous variable mentioned. In this way, by annotation of the genes that have the major contributions in these components/factors, we expect to flag genes as good candidates to further research.

Keywords: Metabolic Syndrome, Confirmatory Factor Analysis, Principal Component Analysis, Non-negative Matrix Factorization, Independent Component Analysis

Resumo

A síndrome metabólica é definida como um agregado de fatores de risco cardiovasculares. Pode ser medida facilmente por um conjunto de medidas em ambiente clínico e/ou ambulatório: a tensão arterial; a frequência cardíaca; o índice de massa corporal ou outra medida antropométrica com funcionalidade semelhante; concentrações séricas de biomoléculas; etc. Por esse motivo, é considerada como facilmente mensurável. A sua importância vem do facto de estar associada ao aparecimento prematuro ou ao aumento da gravidade de várias doenças, condições ou fenómenos biológicos, como a diabetes, várias doenças mentais, a resistência à insulina, inflamação, stress oxidativo. Em alguns estudos transversais e estudos de coorte, a razão de chances e o risco relativo foram estimados, obtendo valores de cerca de 1.2-1.8 para a progressão para a doença cardiovascular e de 4.1 para o risco de progressão para o diabetes tipo 2. Para além da importância clínica, de risco individual, de cada sujeito, a prevalência da síndrome metabólica nas populações europeias modernas é elevada e, por esse motivo, considerada como um problema de saúde pública. Tipicamente encontramos valores na gama de 20% a 80% de prevalência (ajustada para as idades) nas diferentes comunidades europeias.

A definição atual da síndrome metabólica, feita pela pontuação obtida num conjunto de pontos de corte em medidas clínicas já referidas acima é pouco fina, pelo que não permite a desejável discriminação das pessoas afetadas em termos de gravidade da síndrome. Ao longo dos últimos anos na investigação desta síndrome, existe um cada vez maior consenso na necessidade de se evoluir desta situação para uma em que a definição de síndrome metabólica seja baseada numa variável contínua composta. No entanto, algumas das tentativas realizadas foram baseadas apenas na soma de valores standardizados das variáveis referidas, e não tiveram em conta a necessidade de atribuir pesos diferentes às variáveis aquando da construção da variável contínua referente à síndrome metabólica. Mais recentemente, outros autores progrediram para o uso da análise fatorial confirmatória para estimar a diferente contribuição de cada uma das variáveis clínicas referidas. No entanto, esta construção é sempre específica para a população em estudo devido à existência de muitas variáveis que ainda são desconhecidas ou cuja medição é complicada.

Neste trabalho começámos por usar o método de análise fatorial confirmatória para quantificar a gravidade da síndrome metabólica numa população de homens finlandeses. Este modelo, que se apresentou muito adequado aos dados usados, permitirá melhorar o diagnóstico e, assim, melhorar o prognóstico para esta população. A quantificação da gravidade permitirá também melhorar as intervenções populacionais pelo maior esforço nos grupos de maior risco e na melhoria da sensibilização dos utentes dos serviços de saúde. Esta metodologia permite também a utilização do modelo construído através de uma série de expressões fáceis de calcular em ambiente ambulatório e por todos os profissionais de saúde, sem necessidade de se recorrer a *software* estatístico ou a profissionais mais qualificados na área da estatística.

No entanto, as variáveis clínicas são uma consequência da presença da síndrome metabólica e de muitos outros fenómenos sociais e biológicos. Para entendermos esta síndrome é necessário pesquisar as suas causas. Nesta parte do trabalho concentrámo-nos nas possíveis associações genéticas, ou seja, na identificação dos genes cuja expressão esteja associada à gravidade desta síndrome. Mais especificamente, analisámos a expressão génica dos adipócitos abdominais, da população já referida acima, através da análise de dados de *microarrays*.

Existem vários métodos estabelecidos na literatura para a análise de *microarrays*, sendo a maior parte deles da área da estatística multivariada. Um dos métodos utilizados foi o de análise em componentes principais, sendo este um método clássico na área. Em termos biológicos, este método tem sido referenciado como tendo a desvantagem de criar componentes ortogonais. Considera-se que as componentes nestes sistemas de expressão génica não têm que ser necessariamente ortogonais e, talvez por isso, esta técnica seja à partida menos adequada que outros métodos na área. No entanto, é um facto que apresentou resultados que permitiram a evolução da área e continua a ser uma referência.

A factorização matricial não negativa é um método de decomposição matricial utilizado quando as matrizes não apresentam valores negativos. Tem sido utilizada de forma frequente na análise de dados genómicos e apresenta a vantagem de ser mais propensa a criar soluções com maior *sparsity*, levando a interpretações mais simples e mais fáceis de testar biologicamente. Neste trabalho, começámos por analisar os dados de *microarrays* por análise em componentes principais e depois prosseguimos para a decomposição matricial não-negativa e para a análise em componentes independentes. Esta última análise tem como uma das suas propriedades que as componentes criadas serão estatisticamente independentes, o que é considerado uma vantagem em sistemas de regulação genética, já que permite testar a modularidade natural destes sistemas. Estes métodos permitiram criar novas componentes/factores

que, ao mesmo tempo que explicaram uma parte importante da variabilidade dos dados, permitiram a análise dos mesmos num espaço de dimensão muito mais reduzida. Nesse sentido, todas as técnicas referidas tiveram sucesso.

No entanto, um dos objectivos desta parte do trabalho foi o de encontrar componentes/factores que pudessem ter uma associação com a variável contínua de gravidade da síndrome metabólica criada por análise fatorial confirmatória. O que esperamos é que o fenótipo medido pelas variáveis clínicas e codificado pelo modelo construído por análise fatorial confirmatória possa dar-nos pistas e maior conhecimento sobre quais os genes cuja expressão poderá estar mais associada à síndrome. A análise correlacional por coeficiente de Pearson indicou uma das componentes criadas pela análise en componentes independentes como tendo uma correlação razoável (0.63) com a gravidade da síndrome metabólica, sendo esta maior do que as correlações obtidas com outros métodos. No entanto, é interessante referir que com a análise em componentes principais também se obteve uma componente com uma correlação de 0,47. A seguir à identificação das componentes com maior correlação com a gravidade do fenótipo estudado, procedemos à identificação dos genes com maior contribuição para estas componentes. Descobrimos então que estes genes são essencialmente da área metabólica, inflamação e relacionados com o sistema imunitário. Embora o facto de estas áreas serem importantes nesta síndrome seja já do conhecimento generalizado dos investigadores, a pesquisa mais fina permitirá muito provavelmente identificar genes que podem ser alvo de mais pesquisa para aumentar o conhecimento nesta síndrome e nas terapias a ela associadas.

Em sumário, neste trabalho construímos um modelo estatístico para a gravidade da síndrome metabólica, com a ajuda da análise fatorial confirmatória, aplicável à população de onde a amostra foi originada. Cremos que esta foi a primeira vez que um modelo estatístico desta natureza foi construído para uma população de homens da Finlândia. Este modelo permitirá melhorar a qualidade do diagnóstico e do prognóstico nesta população. Para além disso, através dos métodos estatísticos multivariados já referidos, conseguimos encontrar os genes cuja expressão génica tem uma contribuição mais associada a esta síndrome metabólica. Este conhecimento, também criado pela primeira vez por estas técnicas, permitirá aprofundar a pesquisa genómica nesta síndrome levando potencialmente a maior conhecimento científico e a novas terapias.

Palavras-chave: Síndrome Metabólica, Análise Factorial Confirmatória, Análise em Componentes Principais, Análise em Componentes Independentes

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this modern era, the development of technology is driven by discoveries in science. But this development by itself also promotes the enlargement of our scientific and technical methodologies to observe and measure the phenomena that we seek to understand. This virtuous cycle is also accelerating and nowadays there is always something that can be discovered or revisited with a fresh and comprehensive inspection.

This is certainly true in Biology and in Biomedical Sciences, areas that are strongly descriptive by nature, but where quantitative traits and measures were always looked for. In present days, we witness the fast development and use of techniques that allow the measurement of the entire set of genes or transcriptomes, for example, in a given organism. This characteristic, of measuring everything, gave rise to the suffix omics and these areas of knowledge are now called genomics and transcriptomics.

Although Statistics and Biology have a long history of cooperation, never as today has Biology benefited from statistical reasoning and statistical methodologies and never as today has Statistics been challenged and stimulated by Biology. The huge amount of data, produced in a multivariate framework calls for rigorous, advanced and inspired methods that can extract the relevant information from a noisy environment.

So, we should start by a suitable representation of the data, which is vital to the applications as it determines the course of subsequent processing and analysis. This representation should be amenable to interpretation and computationally feasible. One popular process to obtain what was stated is to reduce the dimensionality and, at the same time, denoise the data and increase computational efficiency, interpretability and help visualization.

Although, in theory, we can use nonlinear methods in our data model, in Biology, and particularly in genomics and transcriptomics, they are not so common as we could assume. This is because although we know that some

associations between variables of interest follow nonlinear processes, the type of data in this area, where the sample size is much smaller than the number of parameters to be estimated, are not ideal. Indeed we can easily fall in overfiting data.

Due to the former argument, linear algebra became a key tool in modern analysis of gene expression. Linear models do not represent these biological processes in full rigor, but they are a reasonable approximation.

1.1 Motivation

Methods in Multivariate Statistics are used extensively in gene expression studies. Among others, some of the used ones are principal component analysis (PCA), nonnegative matrix factorization (NMF) and independent component analysis (ICA). All of them start with a linear model that aims to explain the data, but vary in their assumptions. Due to the complex nature of the phenomena studied, usually we do not know what method will be best and which assumptions better fit the data. So, we have to test different methods on the same data and in different data sets to empirically conclude which methods are best and in what conditions.

Most gene expression studies are comparative, in the sense that we have two or more groups of individuals that have some extreme variation in some measured trait. For example, we can compare severe cases of disease with low severity or with controls. Although this is a strategy that is common, it may lead to an incomplete assessment of the variation in gene expression profile and it does not take into account all the samples available.

In this work we intended to study gene expression in the metabolic syndrome, which is commonly defined as a qualitative trait (having or not having the syndrome). So, we started by constructing a quantitative latent variable from the measured clinical variables (representing the phenotype) taking advantage of the *a priori* knowledge in the area by means of confirmatory factor analysis (CFA). After this we pursued by investigating the relationship of this construct with gene expression profiles decomposed by matrix decomposition methods.

In the end, the objective is to identify genes that can be related with this syndrome, so that further experiments in this syndrome can be more specific and productive.

1.2 Objectives

The technical objectives in this work are:

1. Construct a quantitative variable that represents metabolic syndrome severity, by using clinical variables and confirmatory factor analysis
2. Decompose the gene expression data matrix by principal component analysis, non-negative matrix factorization and independent component analysis. Compare the results obtained by these three methods.
3. Study the relationship between the metabolic syndrome variable and the components and factors from the linear algebra methods

1.3 Reading guide

In the next chapter, the second, we introduce the biological framework in which these phenomena occur. In the third chapter, the statistical methods used are introduced and explained. In the fourth chapter, the main results of the various methods are presented. Finally, in the fifth chapter, we disccuss the main findings.

Chapter 2

Biological Background

2.1 Metabolic Syndrome

Metabolic Syndrome [Grundy et al., 2005, Arnlöv et al., 2011], here abbreviated as MS, is defined as a cluster of cardiovascular risk factors. It is associated with insulin resistance [DeBoer et al., 2011] and has correlation with biologic phenomena such as inflammation [Rizza and Federici, 2011] and oxidative stress [Onat and Hergenç, 2011, Gagliardi et al., 2009]. This combination is also associated with a rising risk for type 2 diabetes (T2D) and disease progression. This risk has been quantified and individuals with MS, compared with individuals without MS, have an odds ratio of 1.2-1.8 [Dekker et al., 2005, Vinluan et al., 2012] in the progression to cardiovascular disease and of 4.1 in the risk of progression to T2D [Hanley et al., 2005]. As such, the development of instruments to measure this syndrome more robustlly and the understanding of its biological background, are a priority in world health.

Human populations are complex and vary in biological, sociological and in lifestyle choices. These factors have a strong influence in the impact and working definition of the MS. Due to this, it is necessary to model this syndrome in each ethnic group/population and, besides that, the history of the individuals in the population is also important [Gaillard et al., 2010, Sumner and Cowie, 2008, Walker et al., 2012, Lee et al., 2006]. There is disparity in the common prevalence of MS in different populations and if one of the reasons arise from the factors cited, another important contribution is the fact that MS is defined as a binary variable. This MS status definition, as having or not having the syndrome, is accomplished by measuring biochemical/antrhopometric variables and noting if a given individuals is below or above some given cutoffs. Another problem with this crude definition of the

MS variable is that it does not measure severity of the syndrome, even knowing that severity is important in terms of prognostics and occurrence of future disease. Because of these lackings, some have proposed the use of a continuous variable for the MS, producing an individual score [Kahn et al., 2005].

In this work, a model in the confirmatory factor analysis framework was produced, that constructs a continuous variable named MS and validates this model by goodness of fit measures, for the Finnish male population under study.

2.2 The technology of microarrays

Microarray technology was developed based on pioneer experiments in the 1970s, where the use of labelled nucleic acids attached to a solid support allowed monitoring of expression of nucleic acids. It was only in 1995 that Patrick Brown and colleagues at Stanford University published a paper thoroughly describing how DNA microarray technology could be used in expression studies [Schena et al., 1995]. Since its early days, the technology has evolved and improved tremendously to become a high-throughput and versatile technique. The multiple companies now providing sample processing and analysis packages, have made this tool most popular among the scientific community. It is widely acknowledged that this technique, partly due to its broad spectrum features, provides insight and information difficultly accessible through conventional molecular biology approaches, where only what is known or expected is analysed.

2.3 What is a Microarray?

Microarray studies allow the parallel gene expression analysis of thousands of known genes of known and unknown function, as well as detection of mutations or polymorphisms through DNA homology analysis. In more physical terms, a microarray consists of an orderly arrangement of hundreds to thousands of identified and sequenced genes which are immobilized to a solid support through printing, known as Spotting (robotic printing) (see figure 2.1) or through Photolithography (synthesised in situ), that allows obtaining ultra-high density microarrays (up to 1016 probes per cm^2), presented with an average size array of $1.28cm^2$ (see figure 2.1).

When performing expression analysis, the information is obtained from the RNA, which is extracted from the samples in study and converted into a stable nucleic acid with the corresponding complementary information,

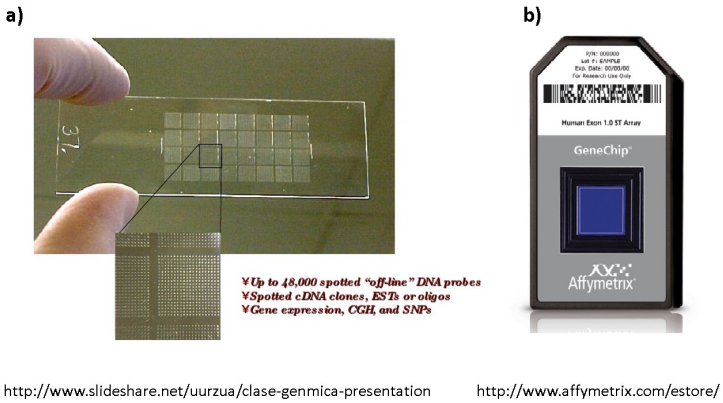

Figure 2.1: Physical structure of a microarray. a) Solid glass slide microarray prepared through Spotting (robotic printing). b) GeneChip version of a microarray produced through in situ oligonucleotide synthesis.

cDNA. The probes that will hybridize to the samples through complementary base-pairing, allowing scanning of the genetic information, are labelled with molecules that possibilitate its detection, typically Cy3 (green) and Cy5 (red) (see figure 2.2). Each of the probes detects a particular RNA species (transcript) and quantitative measurements are made possible due to integration of the signal from each probe, proportionally to the amount of hybridised RNA. This design permits the inspection of the genome of interest fitted into the slide in a single experiment.

So, in a nut shell, the premise of microarrays is that it allows comparison of gene expression between groups and differentially expressed genes may provide some biological insight. This is done in parallel for a myriad of genes, so that a single reaction combines swiftness and efficiency of analysis. Presently, there are several technologies and platforms in this field. The bioinformatical and statistical methods used in this thesis are appropriate for the Illumina platform, for one channel measurements.

2.4 DNA, the backbone of genetic information

DNA, or deoxyribonucleic acid, is the hereditary material present in all pluri- and uni-cellular organisms and ensures the long-term storage of the genetic information. It contains all that is pivotal to instruct the cell on how to construct other components of the cell, namely RNA and proteins. In eukaryotes, animals and plants, it is confined mostly to the cell nucleus, and

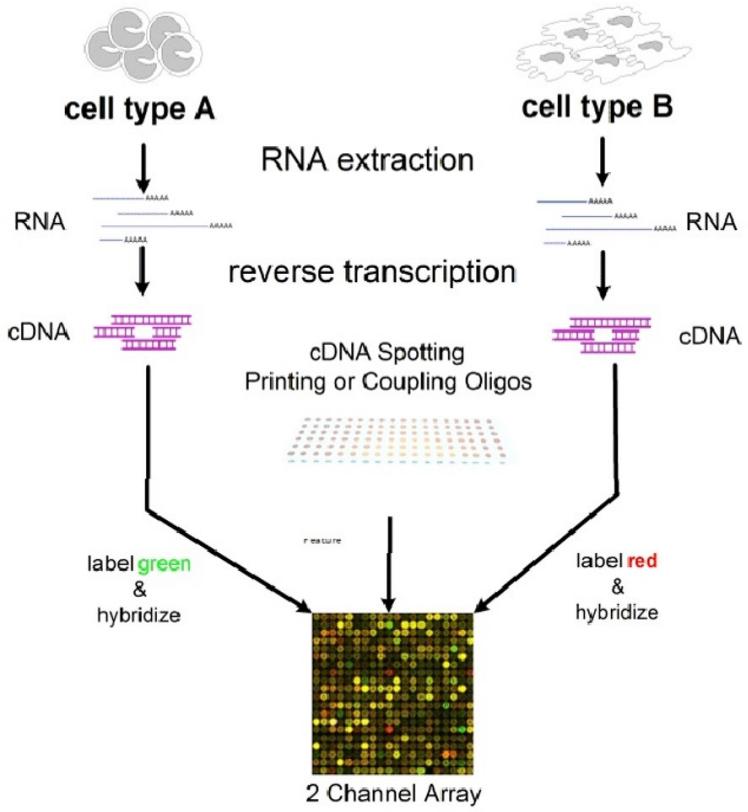

Figure 2.2: Basic protocol of a microarray comparative experiment. The scheme describes the steps involved in an expression microarray experiment, namely the extraction of RNA and conversion into cDNA, followed by probe labelling and hybridization, to produce a 2 channel (2 labelled probes) array.

it can also be found in mitochondria and chloroplasts, in the case of plant cells. In contrast, in prokaryotes it is scattered in specialized areas of the cytoplasm, the nucleoid. The informational units or DNA segments that carry genetic information are designated as genes, corresponding to coding DNA, which may specify proteins and protein subunits or functional RNAs, such as transference RNAs and ribosomal RNAs; other non-coding sequences can have essential structural functions, to allow DNA integrity and stability, and also regulatory functions, that allow the fine-tuning of genetic expression. The DNA molecule is composed of two long complementary polynucleotide chains or strands that are formed by nucleotide subunits, composed of a

five-carbon sugar, deoxyribose, with one phosphate group and a nitrogen-containing base, which may be Adenine (A), Cytosine (C), Guanine (G) or Thymine (T). The nucleotides are covalently linked through the sugars and phosphates to form a chain or strand. Therefore, the only difference in the four types of nucleotide building blocks of the DNA molecule is the nitrogen-containing base. As the subunits combine themselves to form a "beads on a necklace-like" structure, all of the subunits are aligned in the same orientation, with one end of the strand bearing a 3'-hydroxyl group and the other a 5' phosphate group at its terminus. This defines the so called polarity or 5' to 3' orientation of the DNA molecule (see a) in figure 2.3).

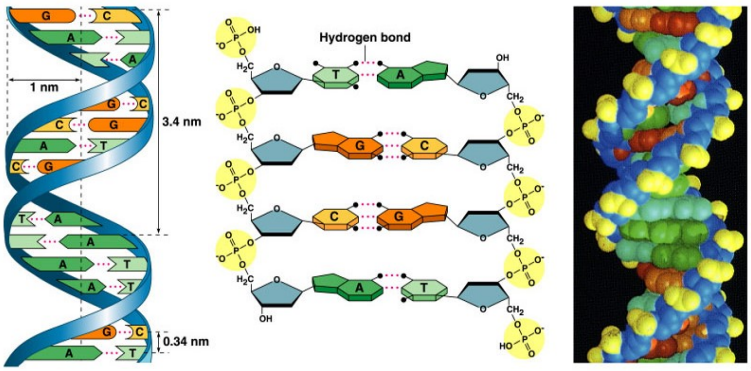

Figure 2.3: Structure of the DNA molecule. a) The DNA molecule is composed of two complementary strands of nucleic acids paired to each other and wrapped around a central axis in the form of a double helix. b) The complementary strands of nucleic acid pair through hydrogen bonds between the complementary bases, Thymine and Adenine and Cytosine and Guanine. c) A 3D filled model of the DNA molecule

The DNA molecule results from the combination, through complementary association between the bases, of 2 DNA strands that wrap in a helix-like manner around a central axis. This complementarity is defined by the chemical and structural features of the polynucleotide chains. The purines, the two-ring bases Adenine or Guanine, always pair with a Pyrimidine, a single-ring base, Thymine or Cytosine, so that an adenine and a guanine always pair with a thymine or a cytosine, respectively. This pairing is achieved through the creation of hydrogen bonds, so that this structural organization confers the DNA molecule with the most energetically favourable arrangement of the interior components of the helix (Figure Zb and Zc). The helix is only formed between two complementary and antiparallel strands of DNA, that is when the polarity of one chain is opposite to that of the other.

2.5 The central dogma of molecular biology: the expression of genetic information

The central dogma of genetics was introduced by Francis Crick in 1959 and later published in Nature in 1970 and very simply states that "DNA Encodes RNA and RNA Encodes Protein." [Crick, 1970]. In a very simplistic view, the flow of information from DNA to RNA can be bi-directional, but unidirectional when converted from RNA into proteins: that is to say that genetic information cannot be retrieved from proteins (see figure 2.4).

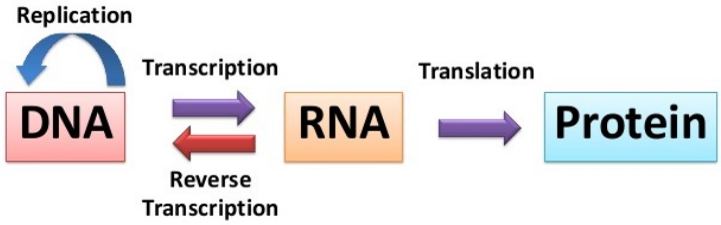

Figure 2.4: The Central Dogma of Molecular Biology.

The order in which the four nucleotide bases are organized in the DNA strands defines the information there contained. This information is read in basic units of 3 nucleotides, designated as codons. The functional unit of DNA defined as gene above is composed of a variable number of these codons or nucleotide triplets. The genetic code has a total of 64 codons which will transform into the building blocks of proteins, aminoacids. Proteins are the product of gene expression. However, the conversion of any given block of informational DNA into protein is extremely complex and tightly regulated at multiple levels and is divided into distinct biological processes.

2.6 The unravelling of genetic information: transcription and translation

As stated, the genetic code uses a four letter alphabet defined by the four different nucleotide bases, A, C, G and T. The combination of these bases in the form of triplets or codons results in the formation of 64 codons that will be used to translate into the building blocks of the effectors molecules of gene expression, proteins. Transcription is the process by which one strand of DNA is copied into a complementary strand of ribonucleic acid (RNA).

It is an essential step in using the information from genes in DNA to make proteins, and relies on the activity of a pivotal enzyme, RNA polymerase. Only one of the DNA strands is used as a template, the so called template strand. In order for transcription to be initiated, the DNA molecule must unwind in the vicinity of the gene to be transcribed, through the action of topoisomerases, allowing access of the transcription bubble to RNA polymerase. This enzyme binds to a promoter region, called the initiation site, near the beginning of the gene, and the new nucleic acid so formed will be complementary to the DNA template, with the important difference that all the thymines are replaced with the uracyl nucleotide, represented by the letter U. Relevant of note, in eukaryotes crucial proteins are involved in the process of transcription, the so called transcription factors, without which transcription would be a much less efficient process. The process proceeds through elongation, during which the RNA polymerase walks along the template strand in the 3'to 5' direction, adding matching complementary bases. Some codons, called termination codons, signal the polymerase to stop transcribing, in a process known as termination, producing a pre-mRNA or pre-messenger RNA (see figure 2.5). In eukaryotes, the RNA transcript must undergo additional processing steps in order to become a mature messenger RNA (mRNA), with the functional capacity to code for a polypeptide chain or protein (see figure 2.5), following which the mRNA is prepared to leave the nucleus and move into the cytoplasm to be translated into protein.

In translation, the basic triplets or codons in the mRNA molecule are read or translated from the 5' to the 3' end by molecules transfer RNAs or tRNAs (see figure 2.5). The triplets are recognized by a complementary base pairing anti-codon in the tRNA molecule, which also carries attached to its other end the corresponding building block of proteins, the corresponding coded aminoacid. The bininding of tRNAs to mRNAs occurs inside a specialized structure known as ribosome, and as the codon-anti-codon binding takes place, an additional aminoacid is covalently linked to the growing polypeptide chain being synthesized. Similarly to transcription, translation is also divided in initiation, defined by the initiating codon AUG recognized by the matching initiator anticodon for methionine, elongation and termination, defined by one of three codons, UAA, UAG or UGA. In eukaryotes, these steps rely on the action of translation factors, proteins without which translation would be a lot less efficient. The genetic code is degenerate because 61 codons encode only 22 amino acids.

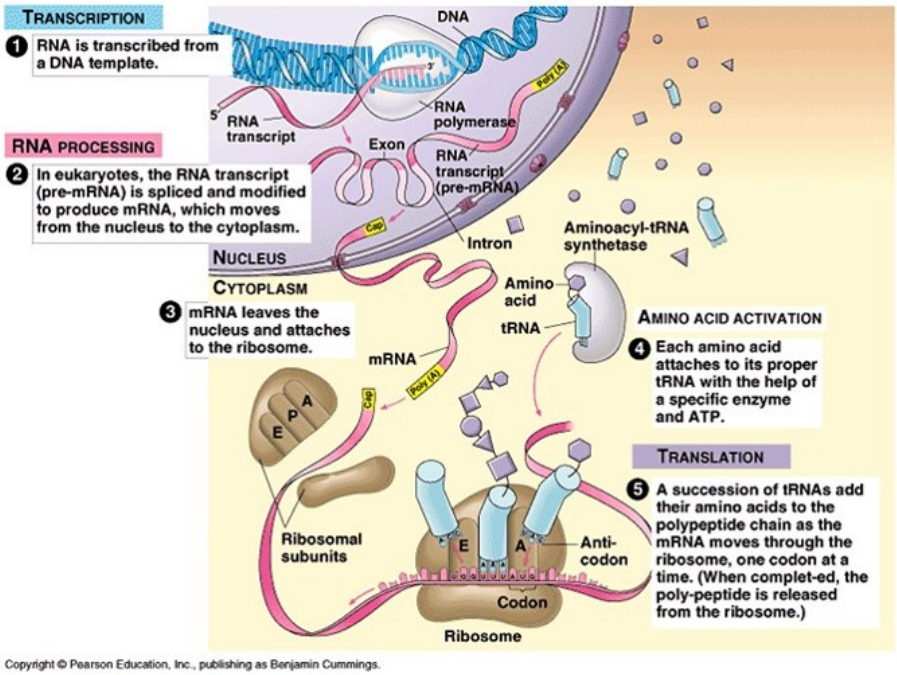

Figure 2.5: Transcription and Translation. Schematic representation of the principal steps involved in Transcription and Translation.

Chapter 3

Statistical Background

This chapter introduces some of the most important methods in multivariate statistics used in modern data analysis and particularly in microarray data analysis. In the following pages we will discuss the concepts and results more specific to the microarray analysis will be left for the last chapter, where the discussion of the results is made. In general, we use these techniques to summarize the information in the data by means of reducing the dimensionality of the data set. These data sets have the characteristic of collecting data on several variables for each unit of observation (in the case of this work, for each individual). Due to the underlying interdependence of the data, no univariate analysis is considered adequate.

3.1 Basic concepts

This work depends heavily on some basic concepts in statistics, like the covariance matrix, correlation matrix, eigenvectors and eigenvalues. As such, these former concepts will be briefly described to the benefit of the reader.

3.1.1 Covariance Matrix

When analysing a multivariate quantitative data set, the arithmetic means, standard deviation and variance of the variables are commonly used. Nonetheless, to measure the association between variables we need to measure the covariance between pairs of variables. For example, for any two random variables, from the random vector $\mathbf{X} = (X_1, \cdots, X_p)$, the covariance will be:

$$\text{cov}(X_i, X_j) = \text{E}\big[(X_i - \text{E}(X_i))(X_j - \text{E}(X_j))\big], \text{ for } i, j = 1, ..., p$$

Where E(.) is the expected value of the random variable. For more than two random variables, we can calculate the covariance between each pair of random variables. A practical representation of these relationships between pairs of variables is the covariance matrix, a symmetric matrix, which represents all possible covariance between all possible pairs of variables. In the following we show an example for p variables, where the main diagonal contains the variance of the variables.

$$\Sigma = \begin{bmatrix} \operatorname{var}(X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_p) \\ \operatorname{cov}(X_2, X_1) & \operatorname{var}(X_2) & \cdots & \operatorname{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p, X_1) & \operatorname{cov}(X_p, X_2) & \cdots & \operatorname{var}(X_p) \end{bmatrix}$$

This matrix also serves as the starting point for more elaborate statistical tecniques that aim in infering the structure of the phenomena studied.

3.1.2 Correlation Matrix

For two quantitative variables we can measure the strenght of the linear association between them by using Pearson's correlation coeficient ρ. For example, if we have a data matrix $X_{m\times p} = [\mathbf{x_1 x_2} \cdots \mathbf{x_p}]$, where $\mathbf{x_i} = (x_{1i}, x_{2i}, \cdots, x_{mi})$ is a column vector with $i = 1, \cdots, p$ and where m is the number of samples, $\rho_{i,j}$ will be:

$$\operatorname{corr}(X_i, X_j) = \rho_{i,j} = \frac{\mathrm{E}\big[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))\big]}{\sigma_i \sigma_j}, \text{ for } i, j = 1, ..., p$$

where σ_i and σ_j are the standard deviation of the respective variable. This coefficient varies between -1 and 1. The sign points to the direction of the association and its absolute value the strength of that association.

As for the covariance matrix, we can construct a correlation matrix with all the possible pairs of variables. In this matrix, the main diagonal is always composed of ones and it is also a symmetric matrix.

3.1.3 Eigenvectors and eigenvalues

If A is a square matrix $(n \times n)$, λ a scalar and $\mathbf{x}$ a non zero column vector, so that $A\mathbf{x} = \lambda\mathbf{x}$, then λ is the eigenvalue of A and $\mathbf{x}$ is the associated eigenvector. Importantly, eigenvectors will be orthogonals, which makes possible the decomposition of the data as a base formed by eigenvectors.

3.2 Principal component analysis

Principal component analysis (PCA) is a classical method in multivariate analysis [Jolliffe, 2002] and, as such, it is widely used since the beginnings of the genomic era. It is implemented as a method to reduce dimensionality and to create new variables that may show a visual sumarization of the original data. As such, helping in the interpretation of the relationships between the measured variables, while retaining as much as possible of the variation present in the data. This is accomplished by creating new uncorrelated variables, the principal components, which have an order, beginning by the component that retains the most variation to the one that retains the least. Specifically, PCA projects the data to a new coordinate system by determining the eigenvectors and eigenvalues of the covariance or correlation matrix.

The mathematical model for PCA can be described as follows:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = \mathbf{a}_j^{'}\mathbf{X}, \text{ for } j = 1, ..., p$$

where Y_1, ..., Y_p refers to the new components (uncorrelated between themselfs), with weights $\mathbf{a}_j^{'} = (a_{1j}, ..., a_{pj})$ and X_1, ..., X_p refers to the original random variables. $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_p$ are the p eigenvectors associated to the eigenvalues of the covariance matrix. So, the new components are linear combinations of the observed variables and are determined by the use of the covariance or the correlation matrix. Commonly the correlation matrix is used when we want to give the same importance to the scale of all variables in the data set. In this determination, the eigenvalues of the correlation matrix are assessed, followed by the corresponding eigenvectors. In the end, each principal component uses a eigenvector, with the first being the one that is constructed with the eigenvector corresponding to the greatest eigenvalue, and so on.

The weights represent the relative importance of the observed variables in the composition of the components. But what is commonly analised are the correlation between the observed variables and the components, by what is called the loadings. These loadings are defined as $l_{ij} = \frac{a_{ij}}{s_j}\sqrt{\lambda_i}$, where j refers to the observed variable, a_{ij} to the weight of that j variable in the component and s_j is the standard deviation of that variable. Finally, the resulting values (y_{ij}), for each observation i on the component Y_j, are called scores. Being A the orthogonal matrix whose columns are the p eigenvectors, $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_p$, and Y the scores matrix, we will have $X = YA'$.

3.2.1 Choosing the number of components

We start with a $n \times p$ data matrix, where n is the number of observations and p is the number of variables, where the sample means are all zeros. Then, $\sum = \frac{1}{n} X^T X$ is the $p \times p$ sample covariance matrix, where T is the transpose of the respective matrix. So, for any vector $\mathbf{v} = (v_1, ..., v_p)^T \in \Re^p$ with $l_2 - norm$ of $\mathbf{v}$ as $\|\mathbf{v}\|_2 = (\sum_{I=1}^{p} v_i^2)^{\frac{1}{2}}$, the coefficient vector $\mathbf{v} \in \Re^p$ of the first principal component will be the solution to:

$$\max_{\mathbf{v} \in \Re^p, \|\mathbf{v}\|_2 = 1} \frac{\mathbf{v}^T \Sigma \mathbf{v}}{\|\mathbf{v}\|_2^2} = \max_{\mathbf{v} \in \Re^p, \|\mathbf{v}\|_2 = 1} \mathbf{v}^T \Sigma \mathbf{v}$$

Due to limitation of the analysis to just a few of the new components, linear combinations of the original variables, this method leads easily to a visual inspection of the data. As such, it has been very useful in cluster analysis. When the data seems to support this type of analysis, the number of principal components can be selected by choosing the ones that explain most variation and stopping with at least one of the following criteria :

1. the components selected explain a fixed part of the total variance (usually 70%)

2. the components selected have all eigenvalues greater than one (for the correlation matrix)

3. observe the scree plot and choose components until the last sharp decline in *eigenvalues*

4. test the null hypothesis that the last eigenvalues are equal, by a suitable statistics

In practice, the aim is to satisfy a high number of criteria and empirically, in microarray data, the first and third criteria are the ones most often applied. In this work, we will follow the third criteria more closely, although with the goal of following the first criteria, when possible.

3.3 Factor analysis and confirmatory factor analysis

Factor analysis (FA) can be understood as part of the greater framework of structural equation modeling [Kline, 2016], it takes the assumption that the phenomena that is the object of study, and from which we have data, can be

easier to explain by the existence of latent variables. These latent variables will determine the observations, as they are functions of the latent variables.

More explicitly, factor analysis describes the observed variables as linear combinations of new variables that are not observed, known as latent variables [Brown, 2015]. The frequent goal in this analysis is to explain the correlation between observed variables by a less numerous number of latent variables (in comparison with the number of observed variables). In this way, this approach tries to reduce the dimensionality of the data without losing too much information.

The statistical model for factor analysis is:

$$X_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{ir} f_r + \boldsymbol{\epsilon_j}, \text{ for } i = 1, ..., p$$

or, in matricial form:

$$X = \Lambda \mathbf{f} + \boldsymbol{\epsilon}$$

where X_i is one of p variables of the random vector $\mathbf{X} = (X_1, ..., X_p)$, $\mathbf{f}$ is a column vector with r rows and $\boldsymbol{\epsilon}$ is a column vector with p rows. The matrix Λ is a $p \times r$ matrix, which is a matrix of the coefficients of the linear combinations of the factors that compose the observed variables, known as loadings. The error term $\boldsymbol{\epsilon}$ has $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \Psi$, this latter matrix is a diagonal matrix where each element of the main diagonal is the variance of the error term of the respective variable. Besides this, $\mathbf{f}$ and $\boldsymbol{\epsilon}$ are assumed independent between each other, meaning that $\mathrm{cov}(\mathbf{f}, \boldsymbol{\epsilon}) = \mathbf{0}$. Before using the model we should also assess if it is appropriate to the data in use. In this work, we used one measure of sampling adequacy, both overall and for each variable, the KMO statistic [Cerny and Kaiser, 1977]. Finally, this model also has the following properties:

1. covariance Matrix $\Sigma = \Lambda\Lambda^T + \Psi$
2. the factor loadings represent the covariances of the variables with the factors and are designated as λ_{ij}, where i is the index for the variable and j for the factor
3. the variance is decomposed between variance due to the common factors (communality, h) and a variable specific component
$$\mathrm{var}(X_i) = (\lambda_{i1}^2 + ... + \lambda_{ir}^2) + \Psi_i = h_i^2 + \Psi_i$$
4. there is not uniqueness, the factors are identifiable only for a specific orthogonal transformation ($\mathbf{O}$)
$$\Lambda \mathbf{f} + \boldsymbol{\epsilon} = \Lambda \mathbf{O}\ \mathbf{O}^T \mathbf{f} + \boldsymbol{\epsilon}$$

In general, there are two methods that are usually applied to estimate the factors. One is the principal components method and the other is the maximum likelihood method, which is the one followed in this work. In the maximum likelihood method we assume that data comes from a population with Normal distribution, $N(\mu, \Sigma)$.

The criteria used to choose the number of factors use at least one of the following conditions:

1. the factors selected explain a fixed part of the total variance (70%)

2. the factors selected have eigenvalues greater than one (using the correlation matrix)

3. observe the scree plot and choose factors until the last sharp decline in eigenvalues

4. test the null hypothesis that there are a given number of factors, by a suitable statistic (not used in this work)

Finally, there are various methods that by changing the orthogonal transformations (rotations) can give a solution with a better interpretation. This is indeed an important aspect of factor analysis, but in this work it is not followed because we make use of the confirmatory factor analysis, where this is not possible.

3.3.1 Confirmatory factor analysis

In this method, the structure, meaning the relationship between the measured variables and the latent variables, is already qualitatively known. Or, as an alternative, we have some hypothesized structured models and compare the fit between them to try to decide which model best fits the data. So, in this approach we use measures of goodness of fit to make a decision on the adequacy of the models to the data. In the next chapter, we mention some of these indices and the values that are empirically determined to be reasonable cutoffs for them.

In practice, this method is used when we know much about the phenomena under study, but for some reason we believe that some variable exists (or can be conceived) that helps to explain the phenomena and that can not be measured directly. Other situation is when we want to have a construct, a composite variable or indicator, that helps to summarize the relevant information.

In this work, confirmatory factor analysis is used to create a composite variable, or indicator, as commonly used in the social and health sciences, to create a continuous score in the metabolic syndrome. This method is ideal in this situation because all measured variables related to this syndrome are continuous and because there is a clear advantage in modeling the severity of the syndrome as a continuous variable, instead of a binary variable, as is more common, which merely indicates the presence or absence of the syndrome in a particular individual.

3.4 Non-negative matrix factorization

The main feature in non-negative matrix factorization (NMF), compared with other methods in the area, is that the attribute values of the data are never negative [Cichocki, 2009]. The most common type of data are counts, intensities or quantities, of words, images or chemical reactions, amongst others. This imposes an important property, that the decomposition can only add components together and can never subtract them. For this reason, this method must be applied with common sense and with some knowledge about the phenomena under study. Besides these characteristics, in practice, most algorithms were implemented with sparseness in mind, meaning that the researchers tried to obtain solutions with a minimal number of relevant factors. This latter fact was due to the type of problems that they were trying to address, in theory there is not any association between non-negativity and sparseness.

Before delving further in this method, it is important to note that, in our datasets, the measured gene expressions (observed variables) will be in rows and the samples in columns. The data matrix X that we are defining is the concatenation of the column vector of observed variables $\mathbf{x}$ refered in earlier sections, for the various n samples. So, the definition of the data matrix X, in the NMF framework is:

$$X = WH + \epsilon$$

where X is $m \times p$, W is $m \times r$ and H is $r \times p$, with $r \leq m$. As described earlier, W and H contain only non-negative values. W is the matrix of factors (or basis components) and H is the mixing matrix (or mixture coeficients). Some empirical quantitative rules have been proposed to determine a reasonable value for r, although, typically, all imply that $r \ll \min(m, p)$, there is little theoretical support for them. In this work, the visualization method implemented to choose r, also known in NMF literature as rank, is explained in the next chapter.

The algorithms used for the optimization problem are very different, the most simple ones start by finding the solution to the minimization of the next formula, by iterating steps.

$$\|A - WH\|_F^2$$

More recent literature has given importance to algorithms that enforce sparsity, using the general concept of regularization. In the area of microarray analysis [Carmona-Saez et al., 2006] have proposed the addition of a third matrix:

$$A = WSH + \epsilon$$

where S is a smoothing $r \times r$ matrix, with the formulation

$$S = (1 - \theta)I + \theta\frac{II^T}{r}$$

By changing the control parameter (θ) between 0 and 1, the smothing matrix varies between the identity matrix and the value $1/r$ in its entries. This subject will not be followed any further here because the default and more conservative approach has attained the results intended.

3.5 Independent component analysis

Independent component analysis (ICA) starts by assuming that the factors into each the method decomposes the data matrix are independent. Besides this strong assumption, one other characteristic of ICA is that all but one of the distributions of the objects along the axes of the factors must be non-Gaussian. These two assumptions come from the goal this method was created for. ICA was developed to do blind source separation, when we typically have more than a source of signal and when these sources are independent. For example, conversation in a crowded room, the travel of light in a gravitational field and image decomposition. Care must be taken to be certain that ICA makes sense for the data at hand, taken into consideration that statistical independence is not always easy to assess.

To better understand the stronger assumption of statistical independence, compared with the assumption of uncorrelation, let's briefly consider the two. Two attributes (A, A^*) are uncorrelated if:

$$E[A]E[A^*] = E[AA^*]$$

but statistical independence requires:

$$E[g(A)]E[h(A^*)] = E[g(A)h(A^*)]$$

where g and h are non-linear functions.
The statistical model for ICA is the following:

$$\mathbf{x_j} = a_{1j}\mathbf{s}_1 + a_{2j}\mathbf{s}_2 + \cdots + a_{rj}\mathbf{s}_r + \boldsymbol{\epsilon}_j, \text{ for } j = 1, ..., p$$

or, in matricial form:

$$X = AS + \epsilon$$

Where $\mathbf{x_j}$ is the vector of m observed values for variable X_j, a_{ij} is a mixing parameter and s_i is a source of the signal. By definition all sources are mutually independent. In this problem, the mixing parameters and the sources of the signal are unknown. In the matricial form, X is the $m \times p$ data matrix, A is the $m \times r$ mixing matrix, and S is the $r \times p$ source matrix, with $r \leq p$, ϵ is a matrix $m \times p$. The independent components are r. We should note that to attain more biological meaning, the microarray data is usually provided with genes in rows and samples in columns, which would correspond to the transpose matrix of the more common form of the data matrix.

To estimate the ICA model, the following assumptions and restrictions must be made [Stone, 2004]:

1. The independent components are assumed statistically independent
2. The independent components must have nongaussian distributions
3. The mixing matrix is a square matrix

This method did not have the goal of reducing the dimensionality of the data, but recent software reorders the components according to the largest coeficients and non-Gaussian distribution behaviour, which can be used to reduce the dimensionality of the problem.

The algorithms to find the solutions to the optimization problem function by measuring the deviation from the Gaussian distribution. This is called the objective or contrast function and varies with the algorithm. They commonly use two non-linear functions (like g and h in our example above) to determine when components are statistically independent. They are all iterative and update the cited matrices until there is convergence. In this work, the algorithm used was FastICA.

Chapter 4

Implementation and Results

In this chapter, the description of the study population and the data is made and the main results are presented. For easier reading, only the relevant analysis are described, within each method used. In reality, the analysis was more comprehensive and more complex, with some results in latter methods leading to changes in the implementation of the methods that are presented first. Almost all this work was done in R [R Core Team, 2016], and for the microarray data, we worked in the Bioconductor framework. A good reference for this software is the book by Draghici [Draghici, 2012]. The workflow in figure 4.1 gives a general overview of this study. In the next sections we will give detailed information about each of the steps in the figure.

4.1 Study Population

All data in this cross-sectional analysis comes from the Metabolic Syndrome In Men study (METSIM), a cohort study implemented from 2005 onwards, that includes 10,197 men, aged from 45 to 73 years, randomly selected from the population of Kuopio (95,000 habitants), in Eastern Finland (Stancakova2009). This cross-sectional and longitudinal study has the goal of quantifying genetic and non-genetic factors associated with the ocurrence of type 2 diabetes, cardiovascular disease and insulin-resistance related traits. To achieve these objectives, the clinical history, lifestyle and anthropometric characteristics of the participants were measured and some questionnaires were implemented.

In a subset of the participants (200 individuals), a sample of the abdominal subcutaneous adipose tissue was surgically extracted [Stancakova et al., 2012]. From this sample, total RNA was isolated and RNA integrity was assessed. High quality samples were hybridized, with the help of 29,487 probes, and

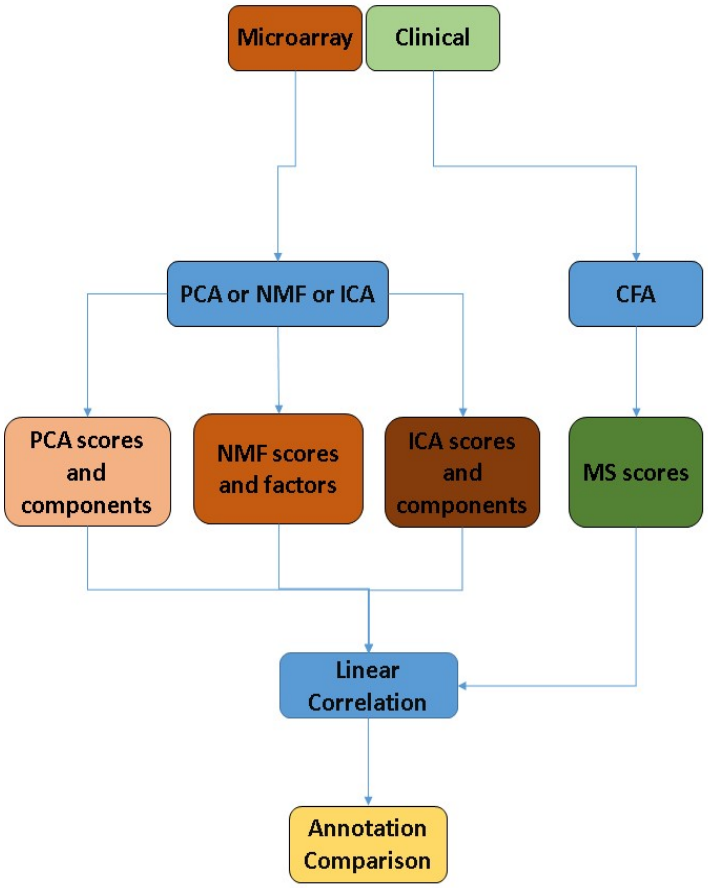

Figure 4.1: The general workflow of this study. Two data sets were used, one from microarray data and other from clinical data. The clinical data were used for the construction of the Metabolic Syndrome (MS) variable, by confirmatory factor analysis (CFA).The microarray data of mRNA expression was submited to principal component analysis (PCA), non-negative matrix factorization (NMF) and independent component analysis (ICA) for dimensionality reduction and feature extraction.Associations between the MS and the components from the other methods was assessed by linear correlation. In the end, after annotation, the genes more associated with MS were identified.

the data was processed by non-parametric background correction, followed by quantile normalization, using the *negc* function in the `limma` package (R

v2.13.0) (this was done by the original authors).

4.2 The Gene Expression Omnibus (GEO)

The data used in this work was downloaded from the Gene expression Omnibus (GEO) database. GEO is a public functional genomics data repository, that archives and distributes array and sequence-based data submitted by researchers. This portal also offers tools and allows users to query its database of experiments and datasets of curated gene expression profiles.

Two datasets from GEO were used in this analysis, the GSE32512, our main dataset, and the GSE45159, these two are from the Illumina platform and have one channel measurements. The first dataset used the platform GPL6947 Illumina HumanHT-12 V3.0 expression beadchip and the second the GPL11154 Illumina HiSeq 2000 (Homo sapiens). In the former, an experiment of the type expression profiling by array, the mRNA expression profile for the 200 individuals was extracted. These were the data used in the PCA, ICA and NMF and the format used in these analyses was as in table 4.1. In the later experiment, the biochemical, anthropometric and clinical measurements, for the same individuals were extracted. We used a data format similar to table 4.1, where the values for the probes were substituted by the values for these clinical variables. These later data were used to construct the latent variable metabolic syndrome by a CFA. All the GEO data were loaded in RStudio [RStudio Team, 2016] with the help of the GEOquery package from the Bioconductor framework. Due to the large amount of information on the dataset, a working subset was constructed where only the 2000 probes with the highest inter quartile range were used. With this approach, and without loosing relevant information, the computational requirements were maintained within reasonable technical requisites.

Table 4.1: Table of the first ten individual samples and the results for some probes

	$ILMN_1651285$	$ILMN_1651343$	$ILMN_1651354$	$ILMN_1651358$	$ILMN_1651496$
GSM804886	7.472886	8.675003	7.933583	5.715654	6.759096
GSM804887	7.389020	8.258173	8.614554	7.364000	7.673838
GSM804888	7.351523	8.700251	9.948965	6.726036	7.024794
GSM804889	7.058686	7.472886	10.613547	5.069931	7.161656
GSM804890	7.676957	7.517674	7.415332	5.045486	7.448686
GSM804891	8.189751	8.846231	7.044753	6.697839	7.574495
GSM804892	7.882517	8.019152	9.266485	6.117614	7.654056
GSM804893	6.631730	8.258173	10.176672	6.267422	8.128527
GSM804894	6.976386	8.924730	11.135864	7.055182	6.777712
GSM804895	7.809580	8.402534	9.466393	5.579977	8.265623

4.3 Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) was used to construct a continuous latent variable named Metabolic Syndrome (MS), from well established components of the syndrome. This was accomplished with the software R (R code team, 2014) and the packages lavaan, psy, psych, sem, e1071. The identification of the variables for the data model was done according to the Metabolic Syndrome international definition. Firstly, all variables were logarithmized, centered and scaled (functions *log* and *scale* in R), so that their distribution would approach a normal distribution. The variables fasting glucose (FP), High-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TCOL) and triglycerides (TGS) were connected imediately with the latent variable MS. For the anthropometric variables fat mass (FM) and body mass index (BMI), as they are closely related, a latent variable dyslipidemia (DLP), that is a composition of these two, was created. After that, DLP was connected with MS. In the model no correlations between the error terms were allowed.

The adjustment of the model to the data is done by iteratively minimizing a fit function that measures the difference between the observed covariance matrix and the one obtained by the model, by a method using a maximum likehood approach [Luo, 2009]. To assess the quality of the adjustment of the model to the data, and because there is not a single index that indicates without doubt the quality of the model, several fit indices were used. Assessment began by the χ^2 statistics value (Bollen, 1989), but we also used: the root mean square error of approximation (RMSEA), using the limit of 0.06 or lower as indicative of good/adequate fit (Steiger, 1990; Schermelleh-Engel, Moosbrugger, Müller, 2003); the comparative fit index (CFI), which must be higher than 0.95 (Bentler, 1990, Schermelleh-Engel et al., 2003); a lower value than 0.08 in the Standardized Root Mean Square Residual (SRMR) as adequate fit (Hu Bentler, 1999) and AIC - Akaike Information Criteria (Akaike, 1987).

After maximum likelihood estimation, we assessed the fit of the CFA model built with a sample size of 362 individual datapoints. The χ^2 statistics observed value was 18.451, for 4 degrees of freedom, with a $p = 0.001$ for the estimated against the null model. The root-mean-square error of approximation was $RMSEA = 0.100$, with p-value = 0.03, $95\%CI[0.057, 0.148]$; the Standardized Root Mean Square Residual was $SRMR = 0.050$; the comparative fit index was $CFI = 0.954$ and $AIC = 4709.687$. By the analysis of these indices (see last paragraph) we can conclude that the model has an adequate fit to the data, altough RMSEA is higher than desirable.

This CFA model (see figure 4.2) was built with the purpose of construct-

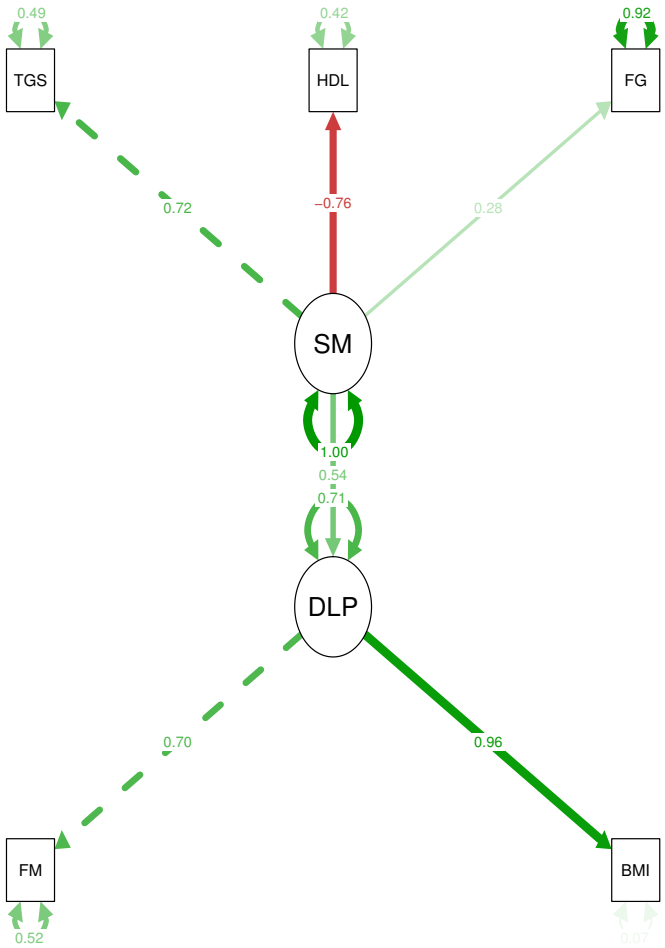

Figure 4.2: The CFA model, the standardized loadings are the straigth arrrows, green arrows for positive loadings and red arrow for negative loading. The standardized variances are the curved arrows. Metabolic Syndrome (MS), fasting glucose (FP), High-density lipoprotein (HDL), triglycerides(TGS), fat mass (FM), body mass index (BMI) and dyslipidemia (DLP)

ing a continuous latent variable named MS. Due to this objective, the next step was to assess a score in this MS variable to each individual, which was ac-

complished with the *predict* function. These individual scores will be used in the next sections when necessary, mainly to assess correlations between these scores and other components of gene expression. For future applications, this model can be presented as a mathematical expression that can be used to calculate the individual MS scores, for example in a clinical environment.

4.4 Principal Component Analysis

Principal component analysis (PCA) was performed for the data described in the first section of this chapter, which was scaled and centered. The gene expression profile data was analised in R, with the packages `FactoMineR` [Lê et al., 2008] and `factoextra` [Kassambara and Mundt, 2016]. Altough these two packages were enough for the goals intended, some confirmatory analysis was done with the package `psych` [Revelle, 2016], but it is not shown in this text. The computation of the PCA resulted in all components having *eigenvalues* higher than one, but besides the first three components, each one of the others were responsible for a very small percentage of total variance, as can be seen in the scree plot (see figure 4.3) and a very small *eigenvalue*, compared with the first eigenvalue. This situation is not unexpected, because of the high number of parameters and a somewhat small sample size, for the system in analysis. Besides this statistical reason, biology tells us that these genetic systems are complex and have a great amount of interactions, some with mechanistical explanations and others as collateral effects.

The first component had an eigenvalue of 386.4 and it was responsible for 19.32% of the total variance. The second component had an eigenvalue of 171.6 and it represented 8.58% of the total variance. The third component had 115.1 as the eigenvalue and explained 5.75% of the total variance. These three components assembled 33.65% of the total variance. At component number 13, 50% of the total variance was reached and 70% of the total variance was achieved with 54 components. This is a typical scenario in these type of data and reflects one of the difficulties in assessing the quality of the methods and in concluding strong inferences from these type of data.

After choosing these components, the individual scores in each of the three components mentioned were correlated with the MS variable constructed in the previous section. The correlation, measured by the Pearson correlation coefficient, was 0.37, 0.47 and 0.07, correspondingly, with p-values of < 0.01, < 0.01 and 0.32 (calculated with *rcorr* function of package `Hmisc`). The first two have a low correlation but significantly different from zero and the third has a non-significant correlation at usual significance levels.

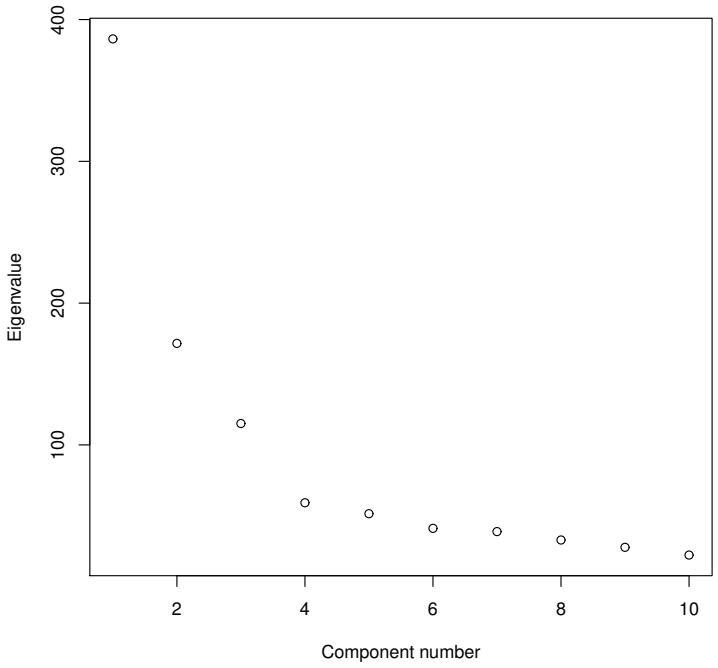

Figure 4.3: Scree plot. As can be seen, the first three components have a very large eigenvalue and after the third component there is a very gradual decrease.

4.5 Non-negative matrix factorization

Next, non-negative matrix factorization (NMF) was applied to the same dataset previously analyzed, using the Brunet algorithm, based in Kullback-Leibler divergence [Yu et al., 2014]. As in PCA, the aim was to find at least a factor (or component) that best explains the total variance in the sample and that has a high correlation with the MS latent variable constructed with the CFA.

For the NMF estimation, 100 runs were made, with the number of factors varying from 2 to 10. From the analysis of the results, and particularly by noting that only 2 and 3 factors solutions have a high value in the cophenetic plot and that between these two, the 3 factors solution has lower residuals, the 3 factors solution was chosen as the best one in our estimation procedure (see figure 4.4).

By observing the consensus map (see figure 4.5) for the solutions with two to four factors, we can also reach the same conclusion. If the number of

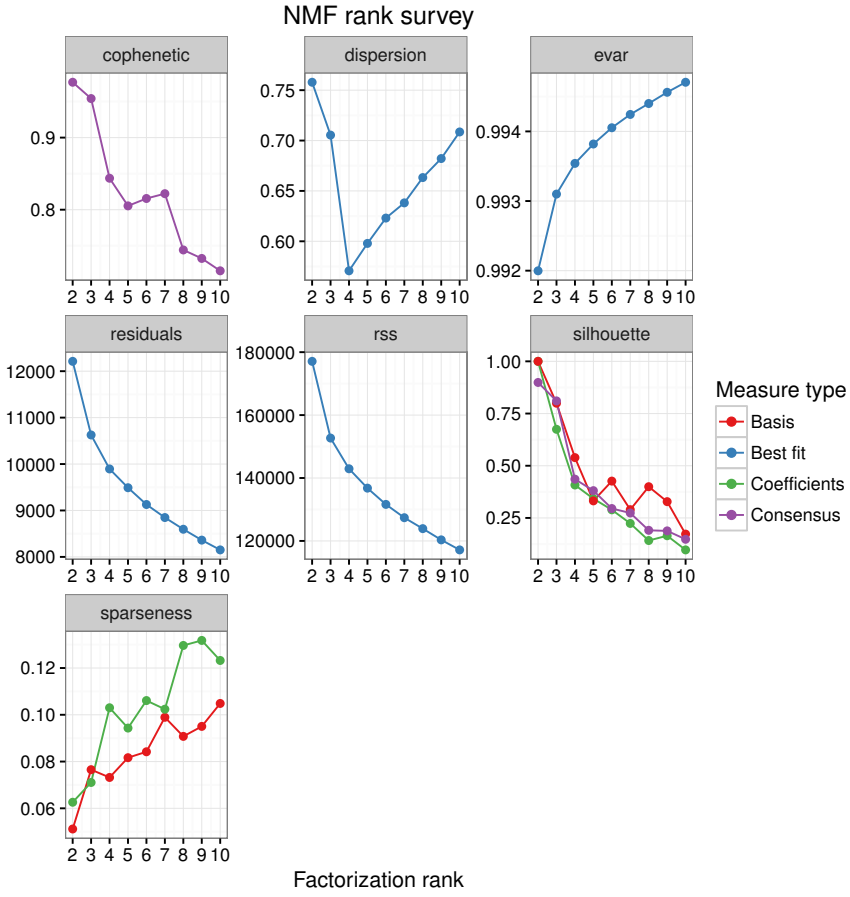

Figure 4.4: Plots used in the estimation of the best number of factors in the NMF method. In this case, the 3 factors solution was chosen.

classes can be acceptable for ranks 2 and 3, in rank 4, and taking into account the dimensionality that this dataset appears to have, the fragmentation is clearly excessive.

Having chosen the rank, we then estimated the solution with a higher number of runs (200) and proceeded to the visualization of the basis map and the coefficient map (see figure 4.6). The basis map clearly shows the partition of the data in three metagenes (3 classes, number of columns) across the gene expression (rows). In the coefficient map, the amount of each metagene that each sample has can be observed (columns).

As in the previous section, the correlation with the MS latent variable was assessed. The individual Pearson correlation coefficients in each of the three metagenes mentioned are the following: 0.50, 0.15 and -0.50, correspondingly, with p-values of < 0.01, 0.03 and < 0.01.

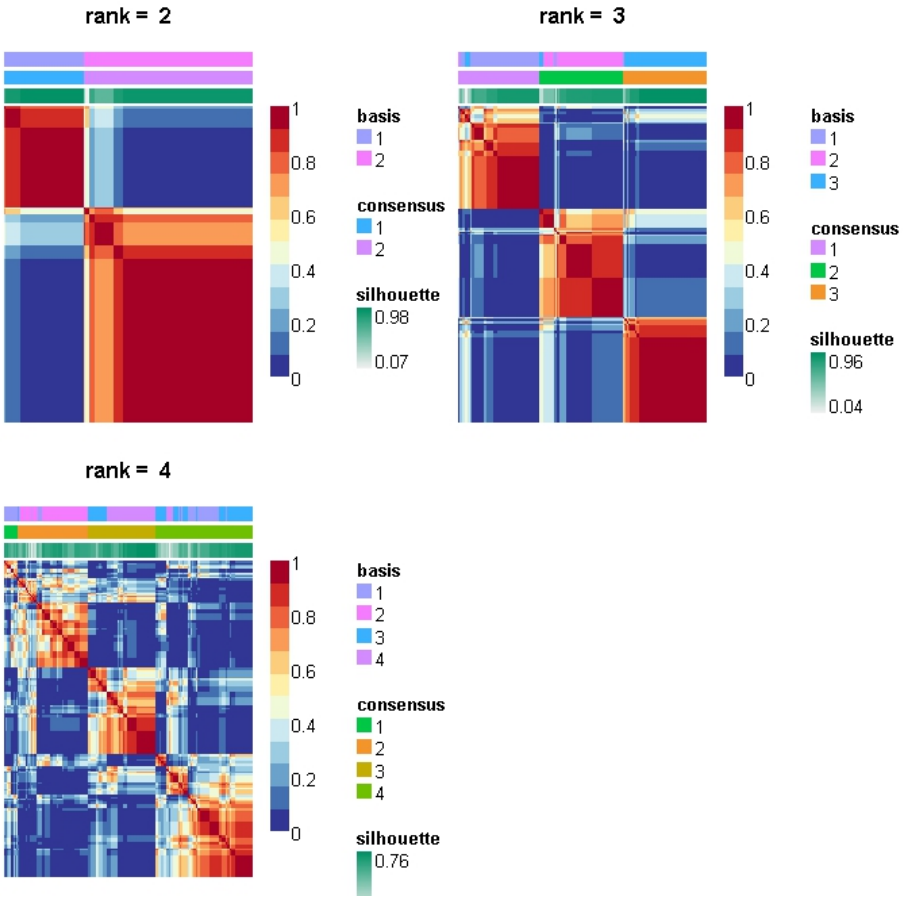

Figure 4.5: Consensus map for ranks 2 to 4. For rank 4, the number of classes is already very high. The consensus map (or matrix) provides information about the stability of the solutions, it is a square matrix determined by the number of samples. It distinguishes the samples that are difficult to classify from those that are consistent. The ideal would be a map without intermediate colors between dark blue and dark red.

4.6 Independent Component Analysis

For independent component analysis (ICA), the gene expression matrix was decomposed using the Jade package. One crucial step in the ICA is in choosing the number of components. There is not a single strategy around this decision, but we chose to follow the strategy that chooses the number of components as being near the value, but higher, than the value obtained for the dimensionality of this phenomenon in the PCA (which is three). We expect that the components are nongaussian and because of that that the distribu-

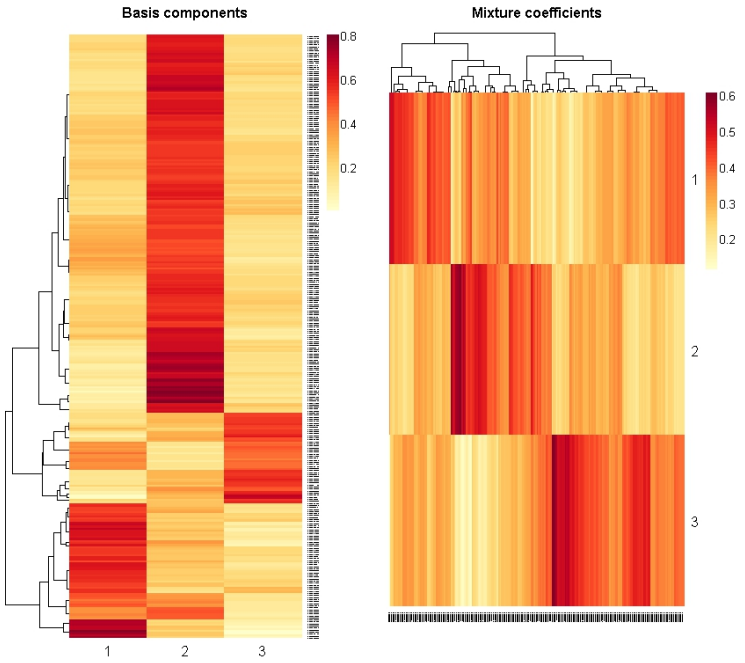

Figure 4.6: Basis and coefficient maps for the rank 3 NMF solution. On the side the hierarchical clustering. The basis has as many rows as probes for the genes and the coefficient map has as many columns as there are samples.

tion will have a non zero kurtosis, but in microarray data a negative kurtosis is more relevant [Scholz et al., 2004]. To be certain of this late decision, we calculated the kurtosis of the components, obtaining the values 1.128697, -1.097101, -0.4313413, 0.9890764, 0.3953929 for the first to the fifth component, respectively. Only in the sixth component do we have another negative value for kurtosis, which is already a high value for the dimensionality estimated. So, the decision was made to proceed with the first five components. After the exploration, the number of components was set in five and the A and S matrices assessed (see figures 4.7 and 4.8).

The individual Pearson correlation coefficients between MS and each of the five components were: -0.43, 0.63, 0.19, -0.20 and 0.06, correspondingly, with p-values of < 0.01, < 0.01, < 0.01, < 0.01 and 0.42.

4.7 Annotation

To conclude a comparison of some of the components/factors that were obtained in the last three methods was performed. We showed that PCA, NMF

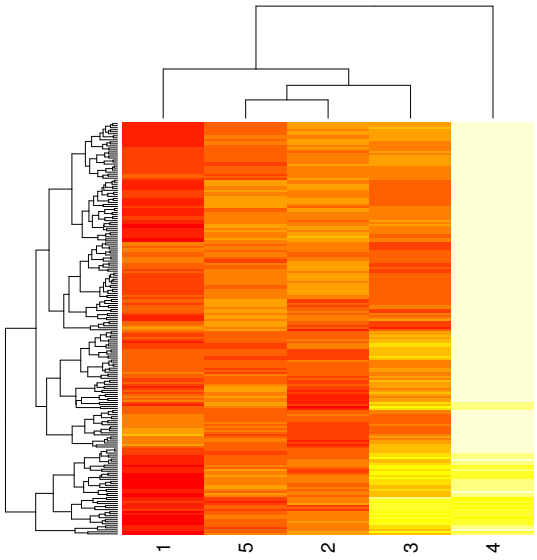

Figure 4.7: The latent variable matrix (A) in ICA. It has 204 rows, as the number of samples and 5 columns, as the number of components

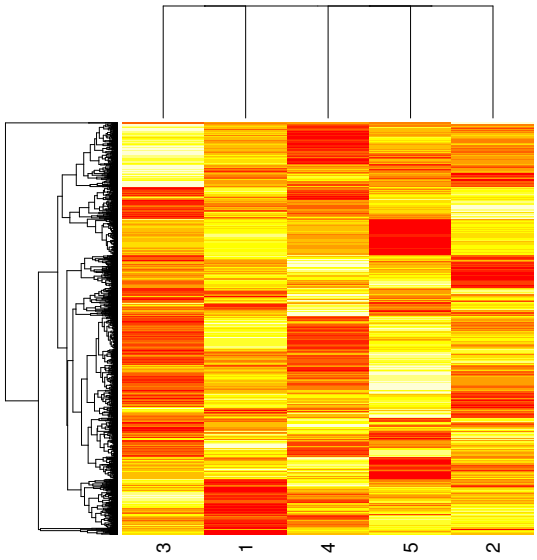

Figure 4.8: The gene signature matrix (S) in ICA, transposed for easier visualization. The original matriz has 2000 columns (gene probes).

and ICA were all able to decompose the data in a way that resulted in having some component/factor with a reasonable correlation with MS (for this

type of problem). Nevertheless, ICA had a component with a higher correlation (0.63). Comparing some of the higher correlated components (see figure 4.9), it can be observed that not only does the ICA component mentioned have a higher correlation with MS, but also that the spreading of the data is narrower in the upper part of the graph, where the score in the latent variable MS indicates a higher severity of the syndrome. This emphasizes the apparent importance of this component.

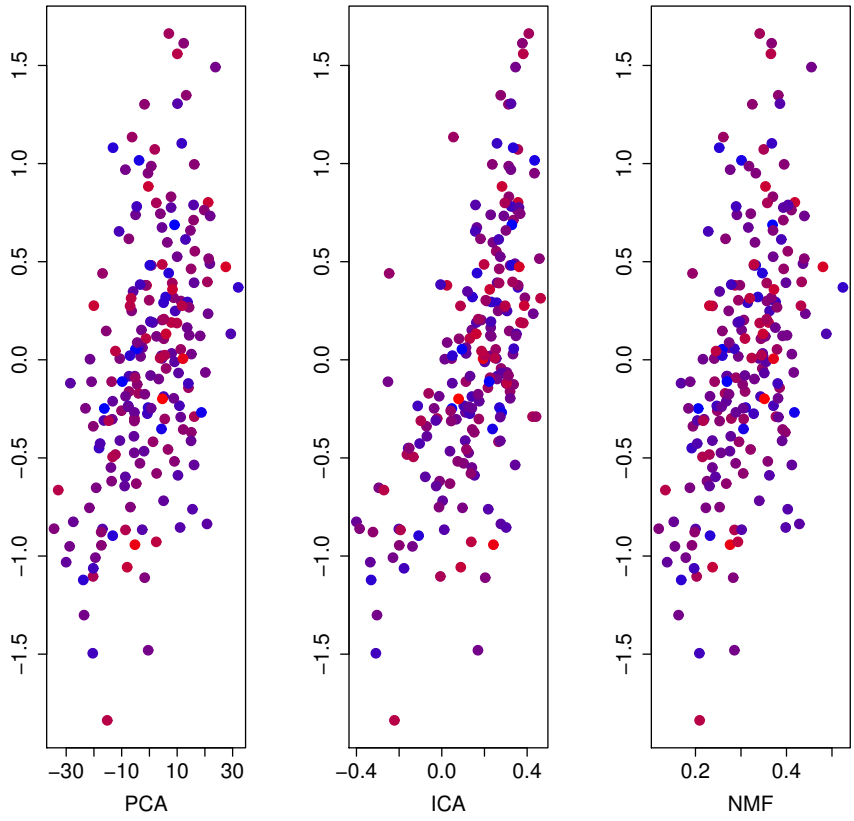

Figure 4.9: The plot of the values in the latent variable MS and the scores of the higher correlated components for each of the three methods used. The colors are just for visual aid.

The analysis of the results was continued by noticing which genes were associated with the ICA component with the higher correlation. Although the full list of genes will not be shown here (see figure A.1), it suffices to say that the 82 genes identified were the ones mostly associated with proliferation, metabolic, immune and inflammatory functions. The importance of

Table 4.2: Table of the first ten genes with higher absolute contribution in the second component of ICA

Gene	Contribution	Notes
EGFL6	7.379127	Epidermal growth factor-like 6
SLC27A2	6.857326	Fatty Acid Transporter
SPX	5.695257	Spexin Hormone
AGPAT9	4.855926	1-Acylglycerol-3-Phosphate - O - Acyltransferase
FHOD3	4.398952	Formin Homology - 2 Domain containing Protein 3
SFRP4	4.163057	Secreted Frizzled-related Protein 4
NPR3	3.916920	Atrial Natriuretic Peptide Clearance Receptor
AZGP1	3.884464	Alpha-2-Glycoprotein, Zinc
UCHL1	3.813064	Ubiquitin Carboxyl-terminal Esterase L1
TUBB2B	3.759544	Tubulin, Beta-2B

these three areas is well established in the study of the metabolic syndrome [Lam, 2015]. In the next table, as an example, we show the genes that had the higher contribution on the second component of ICA, the component with higher correlation with the severity of the metabolic syndrome (see 4.2 table). For this same component, we can see the contribution profiles of the genes with higher contribution (more than two standard deviations from the mean), see figure 4.10.

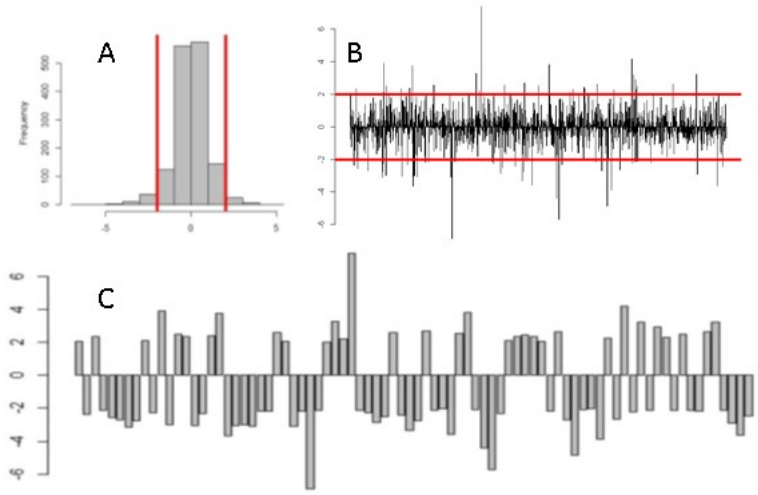

Figure 4.10: In A we can see the histogram of the individual gene contributions to the second component of ICA with the limits corresponding to two standard deviations from the mean of all contributions, in color red. In B the individual gene contributions in the same second component. Finally, in C the individual gene contributions that have contribution values above or below the limits shown in inlet A and B.

Chapter 5

Discussion of Results and Conclusions

In the very first part of this work, we made, with the confirmatory factor analysis methodology, a model that can be used to construct a continuous MS score, that takes into account the relationship between the defining variables. With this model and the individual scores, one can assess the severity of the syndrome in each of the Finnish males. This model was validated by measures commonly used in assessing the quality of this type of models. This is an important accomplishment and, to our knowledge, it is the first time that such a score is created in the Finnish population. By using this score and with the information about the occurrence of disease that exists for this cohort, researchers can improve the odds ratio estimation, which was somewhat confounded, leading to more precise estimates.

For the microarray data of the gene expression in adipocytes of males, PCA, NMF and ICA methodologies were applied. These three methodologies were successful in decomposing the data and in creating a few number of new components or factors with the capacity to explain a reasonable amount of total variation. In this sense, and having no more measure of success, it is difficult to argue that there is a better methodology. As cited in the literature, perhaps one of the measures of adequacy of these methodologies to the data can be the sparsity and the less number of relevant contributions of genes presented by NMF and ICA, which are commonly defined as a good method. In this work, another objective function was pursued as a measure of success, this was the correlation of the components/factors created by these methodologies to the continuous variable MS created by the CFA. With this in mind, and although all methods showed components/factors that had a reasonable amount of correlation (for this field of genomics), there was a clear wining method, which was ICA. To our knowledge, this was also the

first time that a continuous score of MS was correlated with any results in the area of genomics.

Going further with the annotation of the genes that have a relevant contribution in the component with a higher correlation with the continuous variable MS, one can find new clues about the biological background of the MS and acknowledge possible new targets for research that, hopefully, can bring novel knowledge and even new therapies.

In summary, we have constructed and validated a continuous score of the metabolic syndrome, that allows the identification of the severity of the syndrome in each individual. This was accomplished by CFA of the common variables measured in clinical and ambulatory settings, which is specific for the population from where the sample comes from, the Finnish male population. So, this score can be frequently and rapidly assessed. The association of this score with the results from the data analysis that comes from microarray data allowed to choose the component to which the gene expression data is more associated. This acts as a tool to the identification of the genes that may be responsible for the metabolic syndrome and permits greater insight in the genetic mechanisms and its regulation.

Bibliographical References

[Arnlöv et al., 2011] Arnlöv, J., Sundström, J., Ingelsson, E., and Lind, L. (2011). Impact of BMI and the metabolic syndrome on the risk of diabetes in middle-aged men. *Diabetes care*, 34(1):61–5.

[Brown, 2015] Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. The Guilford Press.

[Carmona-Saez et al., 2006] Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J., and Pascual-Montano, A. (2006). bioNMF: a versatile tool for non-negative matrix factorization in biology. *BMC Bioinformatics*, (7):366.

[Cerny and Kaiser, 1977] Cerny, B. A. and Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate behavioral research.*

[Cichocki, 2009] Cichocki, A. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley.

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

[DeBoer et al., 2011] DeBoer, M. D., Dong, L., and Gurka, M. J. (2011). Racial/ethnic and sex differences in the ability of metabolic syndrome criteria to predict elevations in fasting insulin levels in adolescents. *The Journal of pediatrics*, 159(6):975–81.e3.

[Dekker et al., 2005] Dekker, J. M., Girman, C., Rhodes, T., Nijpels, G., Stehouwer, C. D. A., Bouter, L. M., and Heine, R. J. (2005). Metabolic syndrome and 10-year cardiovascular disease risk in the Hoorn Study. *Circulation*, 112(5):666–73.

[Draghici, 2012] Draghici, S. (2012). *Statistics and data analysis for microarrays using R and Bioconductor*. CRC Press.

[Gagliardi et al., 2009] Gagliardi, A. C. M., Miname, M. H., and Santos, R. D. (2009). Uric acid: A marker of increased cardiovascular risk. *Atherosclerosis*, 202(1):11–7.

[Gaillard et al., 2010] Gaillard, T., Schuster, D., and Osei, K. (2010). Differential impact of serum glucose, triglycerides, and high-density lipoprotein cholesterol on cardiovascular risk factor burden in nondiabetic, obese African American women: implications for the prevalence of metabolic syndrome. *Metabolism: clinical and experimental*, 59(8):1115–23.

[Grundy et al., 2005] Grundy, S. M., Cleeman, J. I., Daniels, S. R., Donato, K. A., Eckel, R. H., Franklin, B. A., Gordon, D. J., Krauss, R. M., Savage, P. J., Smith, S. C., Spertus, J. A., Costa, F., American Heart Association, and National Heart, Lung, and Blood Institute (2005). Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. *Circulation*, 112(17):2735–52.

[Hanley et al., 2005] Hanley, A. J. G., Karter, A. J., Williams, K., Festa, A., D'Agostino, R. B., Wagenknecht, L. E., and Haffner, S. M. (2005). Prediction of type 2 diabetes mellitus with alternative definitions of the metabolic syndrome: the Insulin Resistance Atherosclerosis Study. *Circulation*, 112(24):3713–21.

[Jolliffe, 2002] Jolliffe, I. T. (2002). *Principal component analysis*. Springer-Verlag.

[Kahn et al., 2005] Kahn, R., Buse, J., Ferrannini, E., Stern, M., American Diabetes Association, and European Association for the Study of Diabetes (2005). The metabolic syndrome: time for a critical appraisal: joint statement from the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes care*, 28(9):2289–304.

[Kassambara and Mundt, 2016] Kassambara, A. and Mundt, F. (2016). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.3.

[Kline, 2016] Kline, R. B. (2016). *Principles and practice of structural equation modeling*. The Guilford Press.

[Lam, 2015] Lam, D. W. (2015). Metabolic syndrome. *Endotext [Internet]*.

[Lê et al., 2008] Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

[Lee et al., 2006] Lee, S., Bacha, F., Gungor, N., and Arslanian, S. A. (2006). Racial differences in adiponectin in youth: relationship to visceral fat and insulin sensitivity. *Diabetes care*, 29(1):51–6.

[Luo, 2009] Luo, H. (2009). *Confirmatory factor analysis of ordinal variables with misspecified models*. Statistiska institutionen, Uppsala universitet.

[Onat and Hergenç, 2011] Onat, A. and Hergenç, G. (2011). Low-grade inflammation, and dysfunction of high-density lipoprotein and its apolipoproteins as a major driver of cardiometabolic risk. *Metabolism: clinical and experimental*, 60(4):499–512.

[R Core Team, 2016] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Revelle, 2016] Revelle, W. (2016). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.6.6.

[Rizza and Federici, 2011] Rizza, S. and Federici, M. (2011). Cytokines and metabolic syndrome: the perfect storm for arterial aging. *Atherosclerosis*, 215(2):284–5.

[RStudio Team, 2016] RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470.

[Scholz et al., 2004] Scholz, M., Gibon, Y., Stitt, M., and Selbig, J. (2004). Independent component analysis of starch deficient pgm mutants. *Proceedings of the German Conference on Bioinformatics*.

[Stancakova et al., 2012] Stancakova, A., Civelek, M., Saleem, N. K., Soininen, P., Kangas, A. J., Cederberg, H., Paananen, J., Pihlajamaki, J., Bonnycastle, L. L., Morken, M. A., and et al. (2012). Hyperglycemia and a common variant of gckr are associated with the levels of eight amino acids in 9,369 finnish men. *Diabetes*, 61(7):1895–1902.

[Stone, 2004] Stone, J. V. (2004). *Independent component analysis: a tutorial introduction*. MIT Press.

[Sumner and Cowie, 2008] Sumner, A. E. and Cowie, C. C. (2008). Ethnic differences in the ability of triglyceride levels to identify insulin resistance. *Atherosclerosis*, 196(2):696–703.

[Vinluan et al., 2012] Vinluan, C. M., Zreikat, H. H., Levy, J. R., and Cheang, K. I. (2012). Comparison of different metabolic syndrome definitions and risks of incident cardiovascular events in the elderly. *Metabolism: clinical and experimental*, 61(3):302–9.

[Walker et al., 2012] Walker, S. E., Gurka, M. J., Oliver, M. N., Johns, D. W., and DeBoer, M. D. (2012). Racial/ethnic discrepancies in the metabolic syndrome begin in childhood and persist after adjustment for environmental factors. *Nutrition, metabolism, and cardiovascular diseases : NMCD*, 22(2):141–8.

[Yu et al., 2014] Yu, X., Hu, D., and Xu, J. (2014). Non-negative matrix factorization algorithms and applications. *Blind Source Separation*, page 245–311.

Appendix A

R Codes and Outputs

A.1 R codes for CFA

```
#project CFA in  MS, METSIM study

#working directory
#setwd("D:/Dropbox/Projects/MineICA2")
#from file with clinical measurments
# from datau data

E.GEOD.45159.sdrf <- read.delim(".../E-GEOD-45159.sdrf.txt")
datau<-E.GEOD.45159.sdrf

# [5] "Characteristics..fat.mass..."
df<-log10(datau[,5])

# [12] "Characteristics..log10.body.mass.index."
df<-cbind(df,datau[,12])

# [15] "Characteristics..log10.hdl.cholesterol.mmol.l."
df<-cbind(df,datau[,15])

# [22] "Characteristics..log10.ldl.cholesterol.mmol.l."
df<-cbind(df,datau[,22])
# [24] "Characteristics..log10.ogtt.fasting.plasma.insulin.mu.l."
df<-cbind(df,datau[,24])

# [31] "Characteristics..log10.total.cholesterol.mmol.l."
```

```
df<-cbind(df,datau[,31])

# [32] "Characteristics..log10.total.triglycerides.mmol.l."
df<-cbind(df,datau[,32])


# [45] "Characteristics..ogtt.fasting.plasma.glucose.mmol.l."
df<-cbind(df,datau[,45])


df<-scale(df, center = TRUE, scale = TRUE)

######################
#clean of empty data
######################
dfcomplete<-na.omit(df)

#check dimensions
dim(dfcomplete)

#correct names
#colnames(dfcomplete)<-c("fatmass","l10IMC","l10creatina","l10hdl",
"l10dld","l10tcol","l10tgs","l10FG")
colnames(dfcomplete)<-c("l10fatmass","l10IMC","l10hdl",
"l10ldl","l10ins","l10tcol","l10tgs","l10FG")

#in case we want to use in Amos
write.csv(dfcomplete, file = "data_for_amos.csv")


######################
# see correlation and KMO
####################
#library(psych,quietly=TRUE)
foo_corps <- cor(dfcomplete, method = "pearson")
cor.plot(foo_corps)
hist(foo_corps,xlab="Pearson correlation coefficient",main="")
KMO(foo_corps)


######################
```

```
#build scree plot with parallel analysis and some exploratory analysis
#it was not included in the final version of the work
######################

fa.parallel(dfcomplete,fa="fa")

#fit a EFA model
scree(dfcomplete,factors=TRUE,pc=TRUE,main="Scree plot",hline=NULL,add=FALSE)

fafit3<- fa(dfcomplete, nfactors =3, n.obs=dfcomplete,
fm = 'minres', rotate='oblimin')


fa.diagram(fafit3,cut=.1,size(40,30),digits=1)  #construct diagram


######################
######################
#fit a CFA model
######################
######################

##duplicate set to keep dfcomplete intact
alternative<-dfcomplete
colnames(alternative)<-c("FM","BMI","HDL","LDL","INS","TCol","TGS","FG")

#variables names
#"l10fatmass","l10IMC","l10hdl","l10dld","l10tcol","l10tgs","l10FG"

mod.cfaalt <- '# latent variables
DLP =~ FM+BMI
SM =~ TGS+HDL+FG+DLP
'
#fiting function
fitalt <- cfa(mod.cfaalt, data=alternative,estimator = "ML")

#results
summary(fitalt,standardized=TRUE, fit.measures=TRUE, rsq=TRUE)#, modindices=TRUE)


#plots with library(semPlot)
```

```
semPaths(fitalt, "std", edge.label.cex = 0.75, curvePivot = TRUE)
semPaths(fitalt, title = FALSE, curvePivot = TRUE)
```

A.2 R codes for ICA

```
### R code for work with MineICA package

### adapted and expanded from MineICA manual and vignette


#############################################################
### code chunk : extra lib
#############################################################
#source("http://bioconductor.org/biocLite.R")
#biocLite('Biobase')
#install.packages('Rcpp')
#install.packages('plyr')
#install.packages('ggplot2')
#install.packages('foreach')
#install.packages('plyr')
#install.packages('xtable')
#biocLite('biomaRt')

library(Biobase)
library(plyr)
library(ggplot2)
library(foreach)
library(xtable)
library(biomaRt)
###install.packages('GOstats')
#install.packages('Matrix')
#biocLite('GOstats')
library(GOstats)
#biocLite('cluster')
library(cluster)
#biocLite('marray')
library(marray)
#biocLite('mclust')
library(mclust)
library(RColorBrewer)
```

```
#install.packages('igraph')
library(igraph)
#biocLite('Rgraphviz')
library(Rgraphviz)
library(graph)
library(colorspace)
library(annotate)
library(scales)
#biocLite('gtools')
library(gtools)


################################################################
### code chunk : lib
################################################################

#biocLite('MineICA')
library(MineICA)

################################################################
### code chunk : load main data
################################################################

#examples of the loading of datasets

#biocLite('GEOquery')
library(GEOquery)
gds2<-getGEO(filename='GDS3678.soft.gz')
gse<-getGEO(filename='GSE45159.soft.gz')
gse32512 <- getGEO('GSE32512')#,GSEMatrix=TRUE)

#gse2<-getGEOSuppFiles("GSE45159")
gse2 <- getGEO(filename='GSE32512_family.soft.gz')

################################################################
#see description of gse set
# verification of data
################################################################

#visual inspection of the data
head(Meta(gse2))
```

```
# names of all the GSM objects contained in the GSE
names(GSMList(gse2))
# and get the first GSM object on the list
class(GSMList(gse2)[[1]])
head(Meta(GSMList(gse2)[[1]]))
# and the names of the GPLs represented
names(GPLList(gse2))



##############################################################
# load an example of expression data
##############################################################
show(gse)
show(pData(phenoData(gse))[1:5,c(1,6,8)])
head((phenoData(gse)))
varMetadata(phenoData(gse))

##############################################################
# run ICA
##############################################################
library(JADE)

#retain the features with the largest 2000 IQR
#we called main data as mainz
mainz <- selectFeatures_IQR(gse,2000)

# Run ICA
## Features are mean-centered before ICA computation
exprs(mainz) <- t(apply(exprs(mainz),1,scale,scale=FALSE))

##chech dimensions
dim(exprs(mainz))

##correct names
colnames(exprs(mainz)) <- sampleNames(mainz)

## run ICA-JADE
resJade <- runICA(X=exprs(mainz), nbComp=5, method = "JADE", maxit=10000)
```

```
## Create a MineICAParams object, function buildMineICAParams
## build params
params <- buildMineICAParams(resPath="", selCutoff=3, pvalCutoff=0.05)

#############################################################
# Create an IcaSet instance, function buildIcaSet
#############################################################

## load annotation package
library(illuminaHumanv3.db)


ls("package:illuminaHumanv3.db")
mart <- useMart(biomart="ensembl", dataset="hsapiens_gene_ensembl")


#biocLite("lumiHumanIDMapping")
library(lumiHumanIDMapping)

## Define typeID, Mainz data originate from affymetrix HG-U133a microarray
## and are indexed by probe sets.
## The probe sets are annotated into Gene Symbols
typeIDmainz <- c(geneID_annotation="SYMBOL", geneID_biomart="hgnc_symbol",
featureID_biomart="illumina_humanht_12_v3")

 ## define the reference samples if any, here no normal sample is available
refSamplesMainz <- character(0)
resBuild <- buildIcaSet(params=params, A=data.frame(resJade$A),
S=data.frame(resJade$S),dat=exprs(mainz), pData=pData(mainz),
refSamples=refSamplesMainz, annotation="illuminaHumanv3.db",
typeID= typeIDmainz, runAnnot=TRUE,chipManu = 'illumina',
chipVersion="HumanHT12_V3_0_R3_11283641_A", mart=mart)

icaSetMainz <- resBuild$icaSet
params <- resBuild$params


#############################################################
# IcaSet basics
#############################################################
```

```
icaSetMainz

#phenotype data
annot <- pData(icaSetMainz)

#retrieve titles of labels
varLabels(icaSetMainz)

#assesing variable
icaSetMainz$characteristics_ch1.2 #age

#probe set ids
featureNames(icaSetMainz)[1:5] # probe set ids

#gene symbols
geneNames(icaSetMainz)[1:5] #gene symbols

#sampleNames
sampleNames(icaSetMainz)[1:5]

#data in IcaSet
head(dat(icaSetMainz)) #probe set level
head(datByGene(icaSetMainz)) #gene level

# values from ICA "parts"
A(icaSetMainz)
S(icaSetMainz)
SByGene(icaSetMainz)
nbComp(icaSetMainz)
compNames(icaSetMainz)
indComp(icaSetMainz)

################################################################
## Extract the contributing genes
################################################################

contrib <- selectContrib(icaSetMainz, cutoff=0, level="genes")
## Show the first contributing genes of the first and third components
sort(abs(contrib[[1]]),decreasing=TRUE)[1:10]
sort(abs(contrib[[3]]),decreasing=TRUE)[1:10]
## One can also want to apply different cutoffs depending
```

```
## on the components
## for example using the first 4 components:
contrib <- selectContrib(icaSetMainz[,,1:4], cutoff=c(4,4,4,3),
level="genes")




#################################################################
##Extract data of a specific component
## extract sample contrib and gene projections of the 2nd component
#################################################################


comp2 <- getComp(icaSetMainz, level="genes", ind=2)
## access the sample contributions
comp2$contrib[1:5]
comp2$proj[1:5]


#################################################################
##Run global analysis
#################################################################

## select the annotations of interest
varLabels(icaSetMainz)

#characteristics_ch1.1 is ID in METSIM
ch1.1<-as.character(icaSetMainz$characteristics_ch1.1)
ch1.1<-substr(ch1.1,start=12,stop=nchar(ch1.1))
ch1.1<-as.numeric(ch1.1)

#create variable MS in icaSet
icaSetMainz$MS<-rep(0,length(ch1.1))
for (i in 1:length(ch1.1))
{
    for (j in 1:length(prev2[,1]))
    {
        if (ch1.1[i]==prev2[j,1])  (icaSetMainz$MS[i]<-prev2[j,3])
    }
}
```

```
# restrict the phenotype data to the variables of interest
keepVar <- c("MS")#,"er","grade")

## Run the analysis of the ICA decomposition
# only enrichment in KEGG gene sets are tested
runAn(params=params, icaSet=icaSetMainz, writeGenesByComp = TRUE,
       keepVar=keepVar, dbGOstats = "KEGG")


#Plot heatmaps of the contributing elements
resH <- plot_heatmapsOnSel(icaSet = icaSetMainz, selCutoff = 3, level = "gene
 keepVar = keepVar, doSamplesDendro = TRUE, doGenesDendro = TRUE, keepComp =
 heatmapCol = maPalette(low = "blue", high = "red", mid = "yellow", k=44),
 file = "heatmapWithDendro", annot2col=annot2col(params))

##Gene enrichment analysis, function runEnrich
resEnrich <- runEnrich(params=params,icaSet=icaSetMainz[,,1:5],
                        dbs=c("GO"), ontos="BP")

## see plot
head(resEnrich$GO$BP[[1]]$both)
head(resEnrich$GO$BP[[2]]$both)## correlates with MS score
head(resEnrich$GO$BP[[3]]$both)
head(resEnrich$GO$BP[[4]]$both)
head(resEnrich$GO$BP[[5]]$both)



##Association with sample variables
### Quantitative variables
## Compute pearson correlations between variables and the sample contrib


resQuant <- quantVarAnalysis(params=params, icaSet=icaSetMainz,
keepVar=keepVar, typeCor="pearson", cutoffOn="cor", cutoff=0.01,
adjustBy="none", path="quantVarAnalysis/", filename="quantVar")
```

```
##Clustering of the samples according to each component
resmix <- plotAllMix(A=A(icaSetMainz), nbMix=2, nbBreaks=50)


plotPosAnnotInComp(icaSet=icaSetMainz, params=params, keepVar=keepVar,
keepComp=1, funClus="Mclust")


########################################
## heatmaps
########################################
#install.packages('gplots')
library(gplots)

AA<-as.matrix(A(icaSetMainz))
SS<-as.matrix(S(icaSetMainz))

nba_heatmap <- heatmap(as.matrix(xx))
nba_heatmap <- heatmap(A(icaSetMainz), Rowv=NA, Colv=NA,
col = heat.colors(256), scale="column", margins=c(5,10))
nba_heatmap <- heatmap(as.matrix(xx), Rowv=NA, Colv=NA,
col = cm.colors(256), scale="column", margins=c(5,10))
nba_heatmap <- heatmap(AA, Rowv=NA, Colv=NA,
col = heat.colors(256), scale="column", margins=c(5,10))
nba_heatmap <- heatmap(SS, Rowv=NA, Colv=NA,
col = heat.colors(256), scale="column", margins=c(5,10))

heatmap.2(AA, col=redgreen(75), scale="row", ColSideColors=col,
key=TRUE, symkey=FALSE, density.info="none",cexRow=1,
cexCol=1,margins=c(6,11),  trace="none",srtCol=45)
heatmap.2(mostVariable,trace="none",col=greenred(10))
heatmap(AA,col=colors,breaks=breaks,scale="none",Colv=NA,
Rowv=dendrogram,labRow=NA, reorderfun=reorderfun)


par(mfrow=c(1,2))
heatmap(AA,labRow=NA)
heatmap(SS,labRow=NA)
```

```
############################
## phenotypic datA
## choose only 1st value
#########################

# ID of individuals and chosen variables from individual data from datau
dataPhenoInd<-cbind(ind,dfcomplete)

dataPhenoInd<-unique(dataPhenoInd)

############################
## transcriptomic datA
## choose only 1st value
#########################

dataTransInd<-cbind(ch1.1,t(exprs(mainz)))

dataTransInd <- subset(dataTransInd, !duplicated(dataTransInd[,1])
```

A.3 R codes for PCA

```
##Install factoextra package as follow:

if(!require(devtools)) install.packages("devtools")
devtools::install_github("kassambara/factoextra")


##FactoMineR can be installed as follow:

#install.packages("FactoMineR")

##Load the packages:

library(factoextra)
library(FactoMineR)

#install.packages("psych")
library(psych)
```

```
#######
# Compute principal component analysis

res.pca.trans <- PCA(dataTransInd[,2:2001])

#loadings
#eigen
#%variance
#%cumulative variance
res.pca.trans$eig

#choose components:
res.pca.trans2 <- PCA(dataTransInd[,2:2001],ncp=3)

#Scores:
res.pca.trans2$ind$coord

#factor plots
plot(res.pca.trans2$ind$coord[,c(1,2)])
plot(res.pca.trans2$ind$coord[,c(1,3)])
plot(res.pca.trans2$ind$coord[,c(2,3)])

#### color plot

#Create a function to generate a continuous color palette
rbPal <- colorRampPalette(c('red','blue'))

#This adds a column of color values
# based on the y values
datCol <- rbPal(40)[as.numeric(cut(prev2[,3],breaks = 40))]

plot(dat$x,dat$y,pch = 20,col = dat$Col)
plot(res.pca.trans2$ind$coord[,c(1,2)],pch = 20,col = datCol)
plot(res.pca.trans2$ind$coord[,c(1,3)],pch = 20,col = datCol)
plot(res.pca.trans2$ind$coord[,c(2,3)],pch = 20,col = datCol)


# 3D Scatterplot
library(scatterplot3d)
```

```
scatterplot3d(res.pca.trans2$ind$coord[,1],res.pca.trans2$ind$coord[,2]
,prev2[,3], main="3D Scatterplot")

# Spinning 3d Scatterplot
library(rgl)

plot3d(res.pca.trans2$ind$coord[,1],res.pca.trans2$ind$coord[,2],
prev2[,3],col = datCol, main="3D Scatterplot")
plot3d(res.pca.trans2$ind$coord[,1],res.pca.trans2$ind$coord[,2]
,icaSetMainz$MS,col = datCol, main="3D Scatterplot")
plot3d(res.pca.trans2$ind$coord[,1],res.pca.trans2$ind$coord[,3]
,prev2[,3],col = datCol, main="3D Scatterplot")
plot3d(res.pca.trans2$ind$coord[,2],res.pca.trans2$ind$coord[,3]
,prev2[,3],col = datCol, main="3D Scatterplot")


#Pearson correlation with results from CFA
cor(res.pca.trans2$ind$coord[,1],icaSetMainz$MS[1:200])
cor(res.pca.trans2$ind$coord[,2],icaSetMainz$MS[1:200])
cor(res.pca.trans2$ind$coord[,3],icaSetMainz$MS[1:200])

#to get p values for the pearson coeff
rcorr(cbind(res.pca.trans2$ind$coord[,1],icaSetMainz$MS[1:200]),
type="pearson")
rcorr(cbind(res.pca.trans2$ind$coord[,2],icaSetMainz$MS[1:200]),
type="pearson")
rcorr(cbind(res.pca.trans2$ind$coord[,3],icaSetMainz$MS[1:200]),
type="pearson")

#scatter plot
plot(res.pca.trans2$ind$coord[,2],icaSetMainz$MS[1:200],pch = 20,
col = datCol)


#### end of color plot


#loadings
sweep(res.pca.trans2$var$coord,2,
sqrt(res.pca.trans2$eig[1:ncol(res.pca.trans2$var$coord),1]),FUN="/")
```

```
######################
#build scree plot
####################

plot(res.pca.trans$eig[1:10,1], xlab="Component number", ylab="Eigenvalue")
```

A.4 R codes for NMF

```
#########################
## NMF
#########################

#install.packages("NMF")
library(NMF)

# perform 100 runs for each value of r in range 2:10
estim.r <- nmf(t(dataTransInd[,2:2001]), 2:10, nrun = 100, seed = 123456)

plot(estim.r)

consensusmap(estim.r,  labCol = NA, labRow = NA)

# subset to make graph in report
estim.r2 <- nmf(t(dataTransInd[,2:2001]), 2:4, nrun = 100, seed = 123456)
plot(estim.r2)
consensusmap(estim.r2,  labCol = NA, labRow = NA)


#now with the number of factors chosen
res.multirun <- nmf(t(dataTransInd[,2:2001]), 3, nrun = 200)


layout(cbind(1, 2))
# basis components
basismap(res.multirun, subsetRow = TRUE,tracks=NA)
# mixture coefficients
```

```
coefmap(res.multirun, tracks=NA)


cor(coef(res.multirun)[1,],icaSetMainz$MS[1:200])
cor(coef(res.multirun)[2,],icaSetMainz$MS[1:200])
cor(coef(res.multirun)[3,],icaSetMainz$MS[1:200])

rcorr(cbind(coef(res.multirun)[1,],icaSetMainz$MS[1:200]), type="pearson")
rcorr(cbind(coef(res.multirun)[2,],icaSetMainz$MS[1:200]), type="pearson")
rcorr(cbind(coef(res.multirun)[3,],icaSetMainz$MS[1:200]), type="pearson")
```

A.5 Annotation results

In the following figure A.1 we show an example page of the full annotation results for the genes analised.

This table describes the genes contributing to at least one component, given the threshold(s) 3.
The genes are ranked according to the standard deviation of their expression profiles.
Click on the gene id to access to its GeneCards page.
Click on the components index to access the scaled projections of the complete list of genes.

gene	nbOcc	components	sd_expr	1	2	3	4	5
HLA-DRB5	3	3,4,5	3.09	-0.03528	2.829	**-3.132**	**-3.965**	**-3.83**
HLA-DRB1	1	4	2.69	-2.384	-0.5041	-2.329	**-4.681**	-0.9036
HBG2	1	1	2.13	**-10.61**	-2.526	1.788	0.534	1.338
HBG1	1	1	2.08	**-10.39**	-2.474	1.911	0.2381	1.674
HBD	1	1	2.04	**-12.43**	-1.285	-0.411	0.5075	1.396
S100A8	1	1	1.86	**-11.76**	-1.645	1.263	-0.802	-1.739
SPP1	1	2	1.79	1.797	**12.09**	0.7135	0.2734	-1.509
TUBB1	1	1	1.77	**-10.44**	-2.016	-0.7946	-0.05993	-0.03165
TM4SF19	1	2	1.76	1.742	**12.6**	-0.08242	-1.057	-2.336
HBM	1	1	1.76	**-10.73**	-0.567	0.6624	-0.04097	0.6318
EGFL6	2	2,5	1.76	1.892	**10.64**	-0.02412	0.8823	**5.362**
MMP9	1	2	1.76	0.6999	**12.68**	-0.5314	-0.68	-1.238
NFE2	1	1	1.76	**-11.54**	-1.718	0.03787	0.5722	0.6564
FPR1	1	1	1.74	**-11.38**	-1.268	0.2636	0.5934	-2.12
AHSP	1	1	1.72	**-10.79**	-1.44	0.1892	0.7857	1.415
S100A9	1	1	1.71	**-10.72**	-0.6002	0.9169	-0.2588	-2.218
PLA2G7	1	2	1.62	0.5122	**11.91**	-0.002988	-0.5172	-2.158
FCGBP	1	2	1.60	1.321	**12.23**	-0.2564	-0.4488	1.452
GP9	1	1	1.59	**-9.261**	-0.7987	-0.901	0.7236	0.9405
MMP25	1	1	1.58	**-10.29**	-1.314	-0.153	-0.9838	-1.878
SPX	1	2	1.57	-0.2817	**-10.2**	1.513	0.07624	-1.872
PROK2	1	1	1.56	**-10.19**	-1.325	0.203	0.3278	-1.598
DEFB1	1	4	1.55	0.4339	0.1043	2.828	**5.334**	1.507
AQP9	1	1	1.54	**-9.186**	1.984	0.7638	-0.446	-2.31
RGS18	1	1	1.54	**-9.354**	-1.059	0.624	1.081	-0.1712
HBQ1	1	1	1.52	**-9.652**	-1.537	-0.1531	-0.06214	1.389
NRGN	1	1	1.51	**-8.65**	-0.9813	-0.5694	-0.1845	0.7963
MYL4	1	1	1.50	**-9.393**	-0.8683	-0.4939	0.3142	1.443
PADI4	1	1	1.49	**-9.679**	-0.7929	-0.2477	0.3435	-1.405
CMTM2	1	1	1.46	**-9.633**	-1.071	0.1844	0.06865	-0.5955
SLC7A10	1	2	1.43	0.603	**-7.578**	2.527	1.673	-1.289
SLC27A2	2	2,5	1.42	-0.6446	**-8.736**	0.9504	-1.962	**-7.403**
FCN1	1	1	1.42	**-8.816**	0.459	0.9756	0.8975	-0.2766
SELL	1	1	1.41	**-9.343**	-0.6656	0.699	1.667	-0.5232
SERPINA1	1	1	1.41	**-9.298**	-0.3742	0.3444	-0.3181	-0.2334
PI3	1	1	1.40	**-8.449**	-0.4268	-0.6257	0.9481	0.2022
DEFB132	1	2	1.40	0.2228	**-6.41**	0.7951	1.778	-1.14
S100P	1	1	1.39	**-8.428**	-0.7943	0.02677	-0.3898	-2.081
ELOVL6	2	2,4	1.38	-0.8479	**-6.369**	2.007	**3.089**	0.2926
PRKAR1A	1	4	1.37	0.3856	1.066	1.851	**-2.99**	-2.738
MNDA	2	1,5	1.36	**-7.616**	2.613	0.5626	1.44	**-3.033**
IL7R	1	1	1.35	**-8.558**	-0.08113	-0.9146	0.0478	0.5902
SNCA	1	1	1.34	**-7.872**	1.526	0.08175	0.8879	0.8946

Figure A.1: The first page of the annotation results. The columns indicate the gene, the number of observations in the components to which the gene contributes, in what components does the gene has a contribution, the standard deviation of the expression profile of the gene and the projections of each of the components in ICA.