Personalised Health Monitoring for Early Disease Detection

Katharina Proksch[1], Alessandro Di Bucchianico[2], Sandra Keizer[3], Maike de Jongh[4], Marta Regis[5], Roberto Rocchetta[6], Chenyan Huang[7], Larisa Gomaz[8], Aljosa Marjanovic[9], Daphne Nesenberend[10], Marc Paul Noordman[11], Kayode Oshinubi[12], Zohra Rezgui[13] and Oluwatosin Babasola[14]

[1] University of Twente, The Netherlands
[2] Eindhoven University of Technology, The Netherlands
[3] Eindhoven University of Technology, The Netherlands
[4] University of Twente, The Netherlands
[5] Eindhoven University of Technology, The Netherlands
[6] Eindhoven University of Technology, The Netherlands
[7] Eindhoven University of Technology, The Netherlands
[8] University Grenoble Alpes, France
[9] University of Twente, The Netherlands
[10] Leiden University, The Netherlands
[11] University of Groningen
[12] University Grenoble Alpes, France
[13] University of Twente, The Netherlands
[14] University of Bath, United Kingdom

Abstract

Late diagnosis of cancer and cardiovascular diseases often leads to poor chances of cure at high costs. An approach which has the potential to improve the status quo by helping to detect diseases early on, and thereby increase the chances of cure and reduce the costs for treatment, are longitudinal biomarker measurements of microRNA. In this report, we investigate the concept of a personalized baseline based on analysis of variance as well as hierarchical clustering for healthy/sick groups and individual patients in real data. Furthermore, we discuss mathematical models for the detection of illnesses from longitudinal miRNA data. For validation and verification of the proposed methods we develop a data augmentation strategy to generate a large volume of longitudinal miRNA data that can be used and continuously updated.

Keywords: miRNA, sequential detection, biomarkers, clustering

1.1 Introduction

Even with advancing medical treatment, cancer and cardiovascular diseases remain a leading cause of death throughout the world. In Europe alone, cardiovascular diseases claim more than 60 million potential years of life each year Townsend, Kazakiewicz, Lucy Wright, et al. (2022). Early detection of such diseases is very important to improve chances of survival and to decrease medical costs. The company *You2Yourself*[15] (Y2Y) is working on a method to enable early detection of such life-threatening diseases. For this method, urine and blood samples are periodically taken from a large group of initially healthy people over two years in a big study. Based on historical incidence, approximately 7% of the participants of this study are expected to develop a form of cancer, a cardiovascular disease, or a disease of the central nervous system during the two-year duration of the study. The samples are screened for a specific type of biomarker called micro-RNA (miRNA). MiRNAs are small RNAs that play a key role in post-transcriptional gene regulation Lu and Rothenberg (2018).

[15] `https://you2yourself.com/`

Even though different cell types produce the same type of miRNAs, expression profiles vary by tissue type Ludwig et al. (2016). Changes in organs (tumor/inflammation/damage) lead to changes in miRNA profiles, which can be detected in the samples. For example, it has been observed that miRNA patterns change upon tumor formation, suggesting that they might be useful biomarkers for detecting cancer Galvão-Lima, Morais, Valentim, et al. (2021). Taking multiple samples of the same person over time makes it possible to establish a screening procedure based on a personal baseline for the miRNA profile of the blood and urine of the participants. By tracking deviations from that baseline, one could discover the formation of a disease before the onset of clear symptoms. Using a personal baseline instead of the current population based diagnostics is expected to allow for a more sensitive detection, since the biomarker profiles are unique per individual. Figure 1.1 shows a graphical representation of these steps of the study.

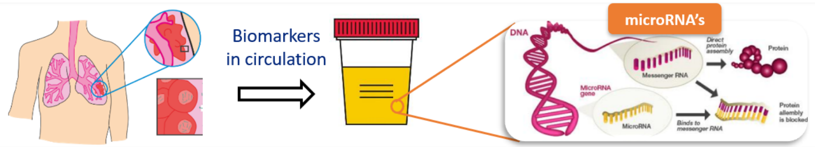

Figure 1.1: **Graphical representation on how deviating miRNA patterns (the biomarkers) end up in a sample.** This patient has a tumor forming in their lung. The miRNA concentrations are different in the tumor microenvironment compared to other parts of the lungs. Some of these biomarkers will end up in circulation, changing the miRNA concentrations in the blood and urine samples as the tumor develops.

The use of miRNAs from urine as biomarkers has the benefit of being patient-friendly and non-invasive when compared to other periodic screening methods or examinations. There are other biomarker options for blood and urine samples, however using miRNA has certain benefits. Figure 1.2 illustrates how deviating biomarker patterns detect formation of illnesses. We stress that this figure is a simplified and idealized representation. In practice, due to the complexity of the biomarker data from the samples, there are several challenges attached to the (statistical) analysis, which we will describe below.

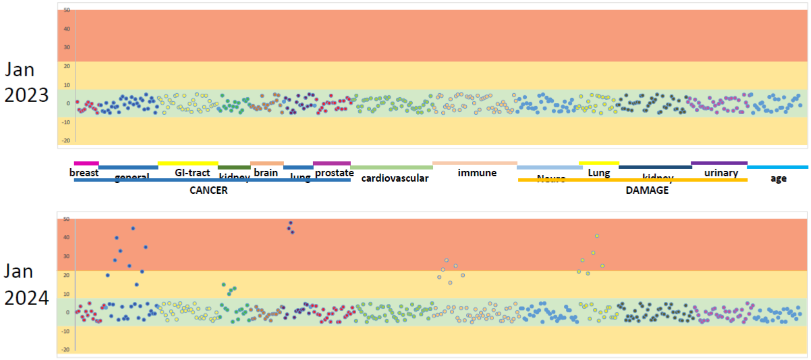

Figure 1.2: **Simplified illustration of how biomarker data could develop over time.** The two figures show the biomarker data of an imaginary patient with one year in between. In January 2023, the biomarker profile is acceptable in reference to the personal baseline. In January 2024 however, a few different miRNAs concentrations are deviating from the personal baseline, indicating that something could be wrong.

1.1.1 Problem statement and organization of the report

One main goal of the research of Y2Y is the extraction of disease detection signatures from complex longitudinal measurements of miRNA profiles. This is difficult because it involves the analysis of longitudinal data in very high dimensions for which no tailor-made methodology exists to the best of our knowledge (see the discussion in Section 1.2.3). A first analysis of the overall project goal led us to define the following three sub-problems, which will be addressed in this report.

1. For a monitoring approach to work, the (stochastic) behavior of the miRNA profile of an average healthy person (a baseline) needs to be explored and put to work as a reference profile. In

Section 1.4.1 we take a closer look at the notion of a baseline and specifically address the question whether a personalized baseline is more suitable than a group or population baseline.

2. Statistical models to detect illnesses and disease progression are needed once a clear notion of a baseline exists. In Section 1.4.3, we discuss the potential of Markov models and mixed effects models in this context.

3. Validation and verification of the proposed analysis tools is important. Since to date only very few longitudinal measurements are currently available, we present a data augmentation strategy in Section 1.4.2, which allows to generate larger samples of synthetic data. The approach is based on the existing data and can be further optimized as soon as more data will become available. Such synthetic data allow to systematically study the performance of any method in the current context in a controlled, yet realistic setting.

This report is organized as follows. We start with a literature review in Section 1.2, followed by a description of the available data and related challenges in Section 1.3. Section 1.4 introduces the approach proposed to tackle some of the challenges. Section 1.5 presents the numerical results and Section 1.6 closes the report with preliminary conclusions and recommendation for future research.

1.2 Literature review

1.2.1 Related work (Medical potential of miRNA for diagnosis)

Since the earliest evidence of miRNA involvement in human cancer was presented in Calin (2002), the topic has been investigated in various studies. A recent review and meta analysis regarding the applicability in the medical field can be found in Condrat et al. (2020), where it is anticipated that monitoring miRNAs will become a routine approach in the development of personalized patient profiles, thus permitting more specific therapeutic interventions as compared to existing, traditional approaches.

One of the two datasets that are analyzed in this report is a cross-sectional data set of 14 controls and 16 patients with stage III and stage IV lung cancer (see Section 1.3 for more details). Related to this, in a meta analysis combining the results of 10 studies, J.-H. Li et al. (2017) investigate the role of miRNAs for the diagnostic and prognostic of lung cancer and the results indicate an excellent overall diagnostic accuracy. Barger and Nana-Sinkam (2015) study miRNAs implicated in lung cancer in general and discuss their usefulness in clinical applications, e.g., as tools for diagnosis, prognosis, and emerging targeted therapeutics.

1.2.2 Related work (Classification and prediction approaches)

Rincon et al. (2019) propose an ensemble feature selection strategy for miRNA signatures for robust cancer classification and detection tasks. They show that a 100-miRNA signature is sufficiently stable to provide nearly the same classification accuracy as the complete Cancer Genome Atlas data set (TCGA, Weinstein, Collisson, and al (2013)). Lopez-Rincon et al. (2020) study a dimensionality reduction and ensemble classification approach for tumor classification from circulating miRNA. Or and Veksler-Lublinsky (2021) recently examined the evolution of miRNA interaction rules and investigate whether these rules are transferable between species using classification methods. Sapre et al. (2016) investigate whether the microRNA (miRNA) profiling of urine could be used to detect urothelial carcinoma of the bladder. Support Vector Machine classifiers with a Student's t-test feature selection procedure is adopted for the detection and the results compared to well-established method (cystoscopy). The authors conclude that miRNA profiling of urine shows promise for the detection of tumour recurrence.

1.2.3 Related work (Longitudinal data and mixed models)

The literature on the analysis of high-dimensional longitudinal data is rather scarce as pointed out recently by Zhong, J. Li, and Kokoszka (2021), who consider analysis of variance and change-point detection in such a setting. While the question of detecting changes over time in high-dimensional data is clearly related to our situation, the view-point is an asymptotic one with respect to time (and sample size). The data in our study were measured at only three time points, using a large sample approximation therefore seems unreasonable. The view-point in the latter reference is related to the view-point in time series analysis, where change point detection in high-dimensions has gained attention in recent years (see, e.g., Jirak (2015) or Cho and Fryzlewicz (2015)). However, in the time-series context many measurements over time are considered and often asymptotic results with respect to time are employed. This perspective is in contrast not only to the available measurements to date but also to future monitoring approaches, where the number of time points will be negligible compared to the data dimension or number of subjects.

Mixed models have been successfully utilized in the analysis of longitudinal biometric data and early disease detection. For example, the predictiveness of ovarian cancer (as a bivalent response variable in dependence of a single biomarker, i.e., a one-dimensional measurement per time point) of two linear mixed models and a pattern mixture model based on the linear mixed model have been compared Han et al. (2020). These models could be extended to deal with the data containing multiple biomarkers and outcomes. However, these methods depend heavily on normality assumptions, which are questionable in our context. S. Li, Cai, and H. Li (2021) consider statistical inference for high-dimensional linear mixed-effects models via a quasi-likelihood approach. The approach does not rely on strict normality assumptions, only sub-Gaussian random components are assumed. The method is based on the Lasso under sparsity conditions on the fixed effects. While this seems suitable at the first glance, the setting and viewpoint considered is that of genome-wide association studies, where effects of genetic variants on a measured phenotype is investigated, i.e., additional measurements on the subjects are regressed on the high dimensional measurements.

Furthermore, it is intuitive that the predictiveness of models can be improved by separation of relevant features and their outcomes, one example is given in Blackwell et al. (2020). It is therefore important to distinguish relevant biomarkers and reduce the model dimensions early on, which is a difficult task in high dimensional data analysis, where standard methodology such as principle component analysis (PCA) fails to be valid (see, e.g., Birnbaum et al. (2013), where estimation of the leading eigenvectors of the covariance matrix is studied under additional structural assumptions on the covariance matrix).

1.2.4 Conclusion on the literature review

There is a clear gap in the literature concerning readily usable statistical methodology to analyze longitudinal miRNA data. Derivation of a fully functioning method is clearly beyond the scope of this workshop. However, we discuss the potential of Markov models and mixed effects models in Section 1.4.3 and fit a mixed effects model to a down-scaled data set (see Section 1.5.5 for the results). The discussion of the existing literature shows that statistical learning has been successfully applied in the analysis of miRNA data. To this end, we apply hierarchical clustering methods to investigate the properties of the data in more detail. Since a serious limitation to date is the availability of large scale longitudinal data sets. Therefore, we develop a data augmentation startegy.

1.3 Description of the data

Two datasets were provided for this study and comprise measurements of miRNA concentrations in urine samples for two independent experiments. The tow data set comprise:

(1) Longitudinal miRNA samples from healthy patients.

(2) Cross-sectional miRNA samples from both healthy and unhealthy patients.

The two datasets will be presented in the next sections.

1.3.1 Longitudinal data

The first data set contains *longitudinal data* of 1941 miRNA concentrations for 7 healthy subjects over 3 distinct time points. In the following, these measurements will be denoted by

$$Y^L_{i,t} \in \mathbb{R}^{1941}, \quad i \in \{1, \dots, 7\}, \quad t \in \{1, 2, 3\}, \tag{1.1}$$

where the index i denotes the individual and t denotes the time at which a measurement was made. Since the exact point in time is irrelevant for this study, t is set to l for the l-th measurement in time for each individual. To provide a first idea of our data, Table 1.1 provides a small excerpt of the concentrations of four exemplary miRNAs in two subjects. The unit is ppm, i.e., parts per million.

$Y^L_{1,1}$	$Y^L_{1,2}$	$Y^L_{1,3}$	$Y^L_{2,1}$	$Y^L_{2,2}$	$Y^L_{2,3}$
131.94	21.47	68.35	0.00	53.03	6.26
27.16	8.05	17.09	0.00	13.26	0.00
38.80	17.45	187.97	190.76	92.80	12.53
7.76	0.00	3.42	0.00	0.00	0.00

Table 1.1: Small excerpt of the longitudinal data of four exemplary miRNAs.

The data shown in the table already clearly suggest that we are dealing with a difficult problem. The variation between the concentrations is quite high and miRNAs with low concentrations might not be measured at all for some individuals, resulting in many zeros. For this data set, a healthy state is assumed for all the patients because regular checks did not diagnose major illness, nonetheless, a disease could be (at least in principle) be progressing without being undetected. Note that more time points are to be added during the course of the study.

1.3.2 Cross-sectional data

The second data set comprises *cross-sectional data* of miRNA concentrations from 30 subjects, 16 of whom had been diagnosed with stage

III or IV lung cancer prior to the study, and 14 are healthy. A cut-off removing all zero-measurements (i.e. zero concentrations over all subjects) is introduced in the second dataset, leaving 1400 miRNAs, corresponding to observations

$$(Y_i^{CS}, k_i) \in \mathbb{R}^{1400} \times \{0,1\}, \quad i \in \{1, \ldots, 30\}. \tag{1.2}$$

In the above model, the index i denotes the individual, whereas the variable k_i denotes the state of the i-th individual (0 for healthy, 1 for sick). Note that the cross-sectional data set only has one measurement per subject, making it unsuitable to investigate subject-specific miRNA concentrations over time. Nevertheless, this data set has the advantage of containing both healthy and sick labels and can thus be used to provide a first idea about the kind of changes in the profiles to expect as the result of the onset of a severe disease and which miRNAs are relevant for disease detection in this case.

1.3.3 Difficulties

The use and analysis of miRNA biomarker data comes with manifold challenges. For instance, one main goal of the research of Y2Y is the extraction of disease detection signatures from complex longitudinal data. However, statistical methods to analyse such longitudinal studies are not standard, as our discussion of the related literature in Section 1.2 shows. Moreover, disease signatures are likely to overlap, making the detection and prognostic tasks even harder to tackle. From the description of the data, is apparent that the data dimension is extremely high compared to the sample sizes and therefore, methodology from *classical statistics* may not be applicable and the viewpoint of *high-dimensional statistics* should be assumed (see, e.g., Wainwright (2019) for a comprehensive monograph on high-dimensional statistics). Furthermore, both data sets only contain a small number of samples (7 and 30 subjects, respectively), which makes the results of data exploration analysis only preliminary, requiring further validation when more samples are available.

1.4 The proposed approach

It is a long way to go until a fully developed health monitoring procedure, for which all fundamental and practical issues will be resolved, can be put to use. This work seeks the first step in this direction by investigating a generalized framework for health monitoring from miRNA biomarker samples. Our main contributions to the existing literature can be summarized as follows:

- A preliminary analysis regarding the concept and feasibility of a personal vs. a population or group baseline in Section 1.4.1
- We present a data augmentation strategy to generate artificial data (based on the already existing samples) and test models and methods on larger data sets in Section 1.4.2.
- We examine and discuss modeling options for disease prognostic and health monitoring from longitudinal bio-markers data in Section 1.4.3.

We used various techniques (such as hierarchical clustering, classification, variance decomposition, statistical tests, and more) to study personal and population-based prognostic models applicable to the longitudinal and cross-sectional miRNA data sets. We present the outcomes of the analyses in Section 1.5. Our findings and recommendations for future research are summarized in Section 1.6.

1.4.1 The concept of a baseline

In mathematical terms, health monitoring can be seen as (sequentially) testing for deviations of measured miRNA profiles from a suitable reference profile. In this report, such a reference profile will be referred to as a *baseline* and it corresponds to the “typical miRNA pattern of an average healthy person”.

In this section, we will discuss the concept of a baseline from a statistical perspective to shed more light into the question what a reasonable notion of a baseline could look like. Furthermore, we will provide an exploratory analysis of the data in respect to the question whether or not to use a personal baseline (in contrast to, e.g., a group baseline).

Mathematical concept of personal and group baseline

The concept of a baseline is essential in order to establish a relative rather than absolute meaning of the miRNA data. Measurements of a healthy person will contain measurements of the miRNA concentrations which are typical for this individual in a healthy state. Multiple measurements of the same person over time will show sampling variability due to the measurement process and also due to the constitution of the patient. A baseline needs to take into account both an average profile (i.e., some measure of centrality of each miRNA) and a measure of expected variability, as both are needed to judge whether an observed deviation from the baseline is significant or not.

Intuitively, it seems to be obvious that a personal baseline is preferable over a group baseline. However, an important disadvantage of the use of a personal baseline is that several measurements of the miRNA profile of an individual in a healthy state are needed in order to properly capture the individual profile including the natural, personal variations in the measured values. In contrast, a group baseline could profit from many prior measurements, so that already a person's first measured miRNA profile could be used for the detection of a disease. The question is which of these two approaches is more feasible in practice. Also, a hybrid approach, combining both personal information and pooled information across individuals, could be a reasonable approach. Based on the longitudinal data set described in Section 1.3 we will look into these question in more detail.

Classification

Longitudinal data allow for the assessment of within individual variation of the miRNA samples over time. Intuitively, one may think that two profiles of the same person should be closer to each other than two profiles of different individuals. To investigate whether this is the case, an attempt has been made to recognize different groups of subjects using hierarchical (linkage) clustering. The hierarchical clustering algorithm starts with a point-cloud, $\{Y_{i,t}^L\}_{i,t}$, say, where every single measured vector of miRNAs starts as a cluster for itself. In each step of the algorithm the distances between the clusters are being calculated and two clusters with the smallest distance are then merged together. This is done until only one cluster remains. The implementation of the algorithm depends, of course, on the notion of distance between the points (i.e. a metric or some dissimilarity function d between the points in $\mathbb{R}^{1941}$) and a notion of a distance between the clusters, also called the linkage function D. Some common linkage functions include arithmetic, geometric and harmonic averages of distances between singular points in the two clusters and minimum and maximum of all the distances. The last two linkage functions give rise to so-called single and complete linkage methods, respectively. The results of such clustering algorithms are dendrograms representing the merging of the clusters, from which the relevant distances can be read. Besides the Euclidean (l_2) distance or the Manhattan (l_1) distance, there are a variety of other metrics and dissimilarity functions available which may capture the separation of the clusters along selected features better. A notion of distance suitable for this task should not put much emphasis on absolute sizes of the components but rather consider the difference of components relative to their sizes, that is, suitably re-scaled versions of common norms might be more appropriate in our context. One such example is the *Canberra distance*. For vectors $u, v \in \mathbb{R}^p$, the Canberra distance is given as

$$d_{\mathrm{Cb}}(u,v) = \sum_{i=1}^{p} \frac{|u_i - v_i|}{|u_i| + |v_i|}.$$

This distance equalizes the contributions of the smaller and larger components and is upper bounded by the dimension p of the space, i.e., $\|d_{\mathrm{Cb}}(\cdot,\cdot)\|_\infty \leq p$. The Canberra distance between two vectors is large if a sparse vector is compared to a non-sparse vector, regardless of the total size of the components of the non-zero vector. In contrast to the Euclidean distance it does not result in extremely large values if the components of one vector are much larger than the components of the other vector. Table 1.2 and Table 1.3 show distance matrices of the data vectors shown in Table 1.1, based on the Euclidean distance and the Canberra distance, respectively. From this it is already obvious that the notions of nearness are very different between these two distances. The clustering results are presented in Section 1.5.1.

	$Y^L_{1,1}$	$Y^L_{1,2}$	$Y^L_{1,3}$	$Y^L_{2,1}$	$Y^L_{2,2}$	$Y^L_{2,3}$
$Y^L_{1,1}$	0	114.39	162.53	177.11	203.22	174.82
$Y^L_{1,2}$	114.39	0	177.11	174.82	81.86	17.90
$Y^L_{1,3}$	162.53	177.11	0	177.11	96.53	186.92
$Y^L_{2,1}$	177.11	174.82	70.59	0	112.18	178.34
$Y^L_{2,2}$	203.22	81.86	96.53	112.18	0	93.84
$Y^L_{2,3}$	174.82	17.90	186.92	178.34	93.84	0

Table 1.2: Distance matrix corresponding to the data presented in Table 1.1, based on the Euclidean distance.

	$Y^L_{1,1}$	$Y^L_{1,2}$	$Y^L_{1,3}$	$Y^L_{2,1}$	$Y^L_{2,2}$	$Y^L_{2,3}$
$Y^L_{1,1}$	0	2.64	1.59	3.66	2.18	3.42
$Y^L_{1,2}$	2.64	0	2.71	3.78	1.80	2.28
$Y^L_{1,3}$	1.59	2.71	0	3.01	1.59	3.71
$Y^L_{2,1}$	3.66	3.78	3.01	0	3.13	3.75
$Y^L_{2,2}$	2.18	1.80	1.59	3.13	0	3.40
$Y^L_{2,3}$	3.42	2.28	3.71	3.75	3.40	0

Table 1.3: Distance matrix corresponding to the data presented in Table 1.1, based on the Canberra distance.

Analysis of variance

From a statistical viewpoint, a personal baseline is preferable over a group baseline if the within person variation of the profiles is smaller than the between person variation. Figure 1.3 shows the sample mean ± one sample standard deviation for two exemplary miRNAs for each of the seven individuals of the longitudinal data set. While the first miRNA seems to have a high variation between the different individuals, the second seems to be dominated by the between groups variation. A statistical methodology that explores exactly this, is analysis of variance (ANOVA), which deals with the problem of testing whether the means of several populations agree. More precisely, a statistical test

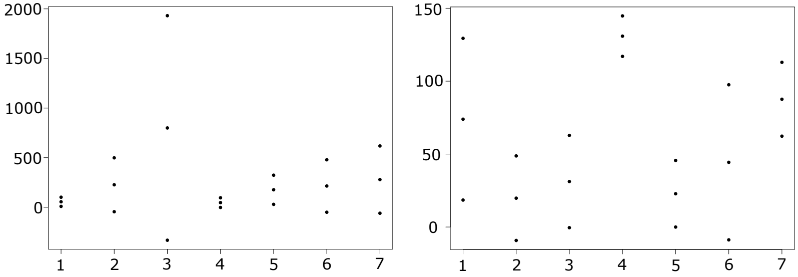

Figure 1.3: sample mean ± one sample standard deviation for two exemplary miRNAs for each of the seven individuals

problem with the following hypotheses is considered:

$$H_0 : \mu_1 = \ldots = \mu_k \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j \quad \text{for at least one pair } i \neq j.$$

In the most basic, one-dimensional setting, observations $Y_{i,j}$, where $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$ are considered and it is assumed that the $Y_{i,j}$ are independent and follow a normal distribution with mean μ_i and variance σ^2. The F statistic is the ratio of the MST and the MSE:

$$F = \frac{MST}{MSE} = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (Y_{i\cdot} - Y_{\cdot\cdot})^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - Y_{i\cdot})^2}.$$

Here, $Y_{i\cdot}$ and $Y_{\cdot\cdot}$ denote the group means and the overall mean respectively. The numerator and the denominator can be interpreted as components of the total variance, the residual sum of squares RSS:

$$
\begin{aligned}
\mathrm{RSS} &= \underbrace{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{i,j} - Y_{\cdot\cdot})^2}_{\text{total variance}} \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{i,j} - Y_{i\cdot})^2}_{\text{within groups variance}} + \underbrace{\frac{1}{n}\sum_{i=1}^{k} n_i \sum_{j=1}^{n_i}(Y_{i,\cdot} - Y_{\cdot\cdot})^2}_{\text{between groups variance}} .
\end{aligned}
$$

This means that the F-test compares the within groups variance to the variance between groups. Under the null hypothesis and the given a assumptions, the F statistic follows an F distribution with $k-1$ and $\sum_{i=1}^{k}(n_i - 1)$ degrees of freedom. We computed the values of the F statistic for all miRNAs. The results of this analysis can be found in Section 1.5.2. Clearly, the the validity of the assumptions is questionable in this context, but all F values in relation to each other can nonetheless be seen an indicator for stability.

1.4.2 Synthetic data generation mechanism

This section presents a data augmentation strategy to simulate a large-scale longitudinal study with several volunteers. Data simulators can support the validation and verification of algorithms and speed up the development of data analysis pipelines. Furthermore, the analysis of synthetic data can support decision-making, future data collection and experiments. Once new empirical evidence is collected, it can tune and improve the simulator's accuracy and adherence to reality.

The proposed simulator generates miRNA concentrations from the empirical marginal distributions conditional to the health state of the subjects (healthy/sick). Mathematically, this corresponds to sampling from $\hat{F}_j(\mathbf{x}|y=0)$ for healthy patients and $\hat{F}_j(\mathbf{x}|y=1)$ for sick patients. For notation convenience, we referred to miRNA profiles as $\mathbf{x}$ and to the health labels as y, where $y=0$ indicates a healthy patient.

For a given label y, a miRNA concentration x_j is sampled from the empirical marginal distribution $\hat{F}_{X_j}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{x_{i,j}\le x\}}$ and realizations in-between samples are obtained by linear interpolation. The procedure work as described next. Consider a vector of miRNA densities $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, where d is the number of miRNA concentrations. A probability value for each entry can be obtained as follows:

$$(U_1, U_2, \ldots, U_d) = (F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \tag{1.3}$$

where the probability values are uniformly distributed in the unit hypercube $[0, 1]^d$. Our simulator uses a copula model, which defines the dependency between the components of the vector $(U_1, U_2, \ldots, U_d)$:

$$C_\Sigma(u_1, u_2, \ldots, u_d) = \mathbb{P}[U_1 \leq u_1, U_2 \leq u_2, , ..., U_d \leq u_d]. \tag{1.4}$$

A Gaussian copula is used in this work,

$$C_\Sigma = F_G(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_d);\ \Sigma), \tag{1.5}$$

where F_G is the joint Gaussian distribution parameterized by the correlation matrix Σ and $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable. We use an empirical $\hat{\Sigma}$ estimated from data. Once a copula structure is defined, pseudorandom samples are obtained sampling correlated uniform vectors

$$(u_1, ..., u_d) \sim C_{\hat{\Sigma}},$$

and then mapping these realization to the space of miRNA densities. This last step is done by inverting of the empirical distributions evaluated at $(u_1, ..., u_d)$:

$$(x_1, \ldots, x_d) = \left(\hat{F}_1^{-1}(u_1), \ldots, \hat{F}_d^{-1}(u_d)\right). \tag{1.6}$$

An example of the procedure is presented in Figure 1.4, where the copula structure is assumed to be independent of the health state and therefore shared among the two groups. Any distribution family can replace the empirical marginals $\hat{F}_j$ and any copula can replace the heuristic copula for healthy and sick patients. Selecting an appropriate distribution family may be a patient-specific and disease-specific issue that is not further considered in this work. Furthermore, a Gaussian copula family requires a large $d \times d$ correlation matrix as an input and this can complicate numerical tractability given the high dimensionality of the sample space. The selection of a subset of highly correlated miRNAs may be advisable for future developments.

Disease progression model and generation of longitudinal data

Our simulation model samples miRNA concentrations of n_p patients at time fixed time steps $t_1, t_2, ..., t_{n_t}$. A health state index $k_i(t_j), i = 1, .., n_p$ is assigned to to the patients at each time t_j and patients are assumed healthy at the beginning of the longitudinal study, i.e., $k_i(t_1) = 0$ for all i. Several patients will likely develop a sickness during the study. Thus, we introduce a probabilistic transition model to simulate this change in population health over time. The following discrete-time Markov chain defines the transition probabilities:

$$\mathbb{P}[k(t_{j+1}) = 1 | k(t_j) = 0] = P_{H2S}$$

where P_{H2S} is the probability that an healthy patient will develop a sickness in the interval $[t_j, t_{j+1}]$. We assume a sick patient to be unable to recover during the course of the simulation, i.e., $\mathbb{P}[k(t_{j+1}) = 1 | (t_j) = 1] = 1$.

Because a sickness fully develops over some time, we propose a disease progression model that combines the empirical distribution of healthy and sick patients. The proposed mixture distribution model is defined as follows:

$$x_j(t) \sim \rho(t)\hat{F}_j(x|k = 0) + (1 - \rho(t)) \cdot \hat{F}_j(x|k = 1)$$

where $\rho(t) \in [0, 1]$ is a real-valued time-dependent sickness factor quantifying the progression of the disease. A value of $\rho(t) = 0$ indicates an healthy patient at time t whilst $\rho(t) = 1$ indicates a fully developed disease. We assume $\rho(t)$ to be a linearly increasing function in the interval t_s and $t_s + t_{trn}$, where y changes from 0 to 1 at t_s and $\rho(t_s + t_{trn}) = 1$ when the sickness is fully developed. Generally speaking, time t_{trn} is a random time which depends on the individual characteristics of the subject and disease. However, for the sake of simplicity, the transition time t_s is a constant input in our model.

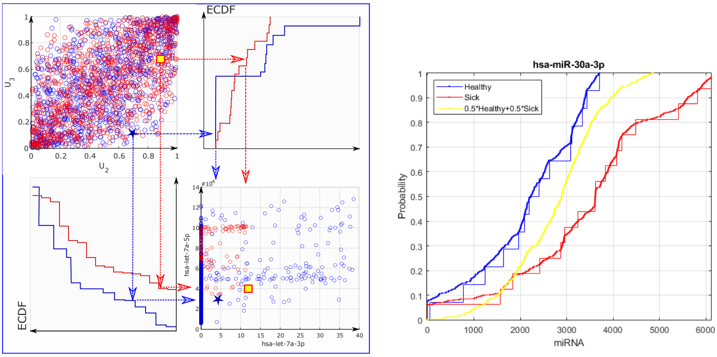

Figure 1.4: Left panel: Correlated samples and inverse empirical CDF transformation for healthy and sick patients. Right Panel: Transition from healthy distribution (blue) to a sick distribution (red).

1.4.3 Detection

This section introduces a mathematical framework for cancer disease predictions that best captures the longitudinal biomarker data. The proposed approach employs mixed effects models and Partially Observable Markov Decision Processes and, due to a lack of time, the latter is only mathematically introduced and not directly applied to the disease prediction problem.

Markov models and POMDPs

In this subsection, we describe a Markovian approach to decision-making problems under uncertainty. Chapter 4 of Poor and Hadjiliadis (2008) gives a detailed description of Markov decision processes applications to sequential detection. Similar ideas have also been explored in the context of maintenance, see e.g., Linderman, McKone-Sweet, and Anderson (2005), Mehrafrooz and Noorossana (2011), and Panagiotidou and Tagaras (2010).

Partially observable Markov decision processes, or POMDPs, provide a formal framework for the interaction of a decision maker (an agent) with a stochastic, partially observable environment. That is, it provides an agent with the capabilities to reason about both action uncertainty, as well as state uncertainty. A POMDP is a discrete time model, in which the agent selects an action at every time step or stage. It extends the regular Markov decision process (MDP) to settings in which the state of the environment cannot be observed. It can be formally defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \Omega, \mathcal{T}, \mathcal{O}, \mathcal{C}, h)$, where $\mathcal{S}$ is a (finite) state space, $\mathcal{A}$ is a (finite) set of action, Ω is the space of observations, $\mathcal{T}(s, a, s') = \mathbb{P}(s'|s, a)$ is a transaction probability function that specifies the probability of a next state s' given a current state s and action a, $\mathcal{O}$ is an observation function, $\mathcal{C}(s, a, s')$ is an immediate cost function for a particular transition s, a, s' and h is the horizon of the problem.

The model $\mathcal{M}$ can help a decision-maker by recommending good actions that maximize the long-run revenue of the monitoring systems or, similarly, that minimize the expected cumulative sum of costs over the horizon h. The rule that dictates which action the agent must take in each state is known as the policy, a map $\pi : \mathcal{S} \to \mathcal{A}$ from the state space to the space of actions. In this work, we wish to maximize the reward generated by a correct prediction of disease from miRNA readings (and minimize costs due to wrong predictions and missed alerts).

States and actions

A state vector $\mathbf{s} \in \mathcal{S}$ consists of three parts, $\mathbf{s} = (\mathbf{x}, y, z)$, where $\mathbf{x}$ is the actual profile of miRNA's, y denotes the actual health status (stage of cancer), and z is a binary variable indicating whether or not cancer has been detected via a traditional diagnostic methods, i.e., an indicator function defined as follows:

$$z = \begin{cases} 0 \text{ if no cancer has been detected.} \\ 1 \text{ if cancer has been detected.} \end{cases} \tag{1.7}$$

We also define action vectors $a \in \mathcal{A}$ as binary indicator of diagnosis based on the miRNA profile:

$$a = \begin{cases} 0 \text{ if diagnosed as healthy based on miRNA profile} \\ 1 \text{ if diagnosed as ill based on miRNA profile.} \end{cases} \tag{1.8}$$

Observations space and transition probability

The observation function $\mathcal{O} : \mathcal{S} \times \mathcal{A} \times \Omega \to [0, 1]$ is a function defining the probabilistic accuracy of the miRNA counts and an observation $\omega \in \Omega$ is a vector containing the miRNA counts. The transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is based on the following probabilities:

- Let $p_1 : \mathcal{X} \times \mathcal{Y} \times \mathcal{X} \to [0, 1]$ denote the probabilistic transition function between miRNA profiles, where $\mathcal{X}$ is the space of possible profiles.

- Let $p_2 : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ denote the probabilistic transition function between different stages of cancer, where $\mathcal{Y}$ denotes the set of possible stages.

- Let $p_3 : \mathcal{Y} \to [0, 1]$ denote the probability that a patient has symptoms severe enough to see a doctor and that cancer is successfully detected given the cancer stage.

Note that $\mathcal{T}$ combines the probability of moving from the present miRNA $\mathbf{x}$ to a new concentration $\mathbf{x}'$ and from a present cancer stage y to a next stage y'. By definition, this is a map from state-to-state $\mathbb{P}(\mathbf{s}'|\mathbf{s})$, and diagnostic actions a have no effect on it.

Cost function

The Cost function $\mathcal{C} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a fundamental component of POMPDs and must be carefully defined. The cost function assigns to any transition, e.g. from a state-action pair to new state, a cost/reward function. In this detection problem, a cost. The lower the cost, the higher is the value of the action taken in a specific state. In this cancer prediction problem, costs can arise due to delays in detecting (changes in) the health status of a patients from miRNA counts, costs for false diagnosis (false positives), and from missed diagnosis (false negatives). Hence, the selected cost function include three terms:

$$C(\mathbf{s}, a, \mathbf{s}') = \begin{cases} C_1(y) \text{ if } a = 1 \text{ and } z = 1 \text{ (true positive)} \\ C_2 \text{ if } a = 1 \text{ and } z = 0 \text{ (false positive).} \\ C_3(y) \text{ if } a = 0 \text{ and } z = 1 \text{ (missed detection)} \end{cases} \tag{1.9}$$

where $C_1 : \mathcal{Y} \to \mathbb{R}$ and $C_3 : \mathcal{Y} \to \mathbb{R}$ are increasing functions of the stage of cancer y. The cost C_1, is cost associated to correct prediction of illnesses, this quantity should be negative (a reward) and should have a large in absolute value in the early stages (higher rewards for an early detection). The cost C_2 is associated to false positive events whilst cost C_3 arise when cancer is not predicted from the miRNA profile but by other means, e.g., the cancer has detected due to symptoms, but not detected from the miRNA profile. The cost C_3 should be high, especially for later stages.

Remarks and challenges

Numerical analysis of POMDPs generally assumes a finite horizon for the analysis and computation of optimal decision-making policies. In this work, we assume that a person is tested for cancer if either he develops symptoms severe enough to see a doctor or his/hers miRNA profile indicates the potential presence of cancer and we define the end of the horizon as the moment a person is diagnosed with cancer ($z = 1$). Unfortunately, the proposed cost function does not take into account finite horizon and if the disease is not detected within a certain time frame, it may result too late in magnitude. This consideration is similar to other discussions on the performance of control charts in statistical process control, where it has been argued that instead of looking at average time to detection (called ARL = average run length), it is more relevant to consider as performance the probability of successful detection with a certain time frame (called PSD = probability of successful detection). The interested reader is reminded to e.g. Kenett and Pollak (2012) for a detailed discussion on this topic. Another issue concerns the definition of cancer stage y and its relationship with the diagnostic outcome z of established screening tests. To increasing the predictive power of this framework, it would be advisable to study a suitable quantifier for y. Moreover, it would be useful to study a function $y = \psi(\mathbf{x})$ that maps actual miRNA structure/changes to this health state, i.e, a model for the minimal change in miRNA counts that will cause a person transitioning from a healthy to sick state.

Mixed effect models

Monitoring multiple patients over time on a series of miRNAs leads to a high-dimensional dataset, multivariate and longitudinal. It can be large in the number of patients (n), in the number of outcomes (m miRNA) in the number of time points (T), all at the same time. Multivariate longitudinal data come with the challenge of correlations and heterogeneity: the clustering at individual level, different miRNAs can have different variances, measurements can be correlated at each time point for different markers, and counts from the same miRNA can be correlated in time. Even picturing an idea of such correlations in the whole dataset is challenging due to the high dimensionality. On the other hand, the complexity and multidimensionality of the data offers a wide choice of models and possible methods to detect changes. For example, beside looking at the shifts and changes in trend of single microRNAs, it is possible to study how the multiple markers vary together in the healthy status, and use this to detect changes.

Mixed models offer a flexible framework to capture different forms of correlation in the data, and to choose the most suitable covariance structure. The general linear mixed model equation is given by

$$Y_i = X_i\beta + Z_i u_i + e_i \tag{1.10}$$

Y_i is the matrix of observations for the i^{th} individual, X_i a matrix of covariates of interest and β the corresponding matrix of coefficients to be estimated. This term is defined as *fixed* and captures the trend over the whole population, as opposed to u_i that are called *random effects* and model individual-specific characteristics. Normality is often assumed for the random effects ($u_i \sim N(0, \tau^2)$). The design matrix for the random effects Z_i can be a subset of X_i but does not have to be. Finally there is the error term e_i that captures the correlations within individual. It's often assumed to be normal, but extensions exists. The two random effects u_i and e_i are often assumed to be independent, but through their covariances it is often possible to investigate the complex dependency structure of the data. There exist different methods to accommodate the structure of multivariate longitudinal data. One option is to include a Kronecker product covariance $V \otimes \Sigma$ for the repeated measurements - repeated for each subject on the different miRNAs and for multiple points in time, where V models the inter-miRNAs correlations between multiple markers measured at the same time point, and Σ the intra-marker correlation at different time points (again the same for all miRNAs). One reasonable structure could involve unstructured V and autoregressive Σ

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \dots & \dots & \dots & \dots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \dots & \end{pmatrix}.$$

Such models can easily be fitted to a large number of study participants (n) and/or to a long time period (large T). Most likely the implementation of mixed models is not easily scalable to a large number of outcomes (i.e. to the whole set of microRNAs), but they can already be used to study how a number of biomarkers vary together, for example a selection that is of particular interest. We will illustrate this on a set of biomarkers in Section 1.5.5.

Generalized linear mixed models extend the linear mixed model in (1.10) by introducing a link function $g(x)$ that connects the linear predictor $\eta = X_i\beta + Z_iu_i$ to the observed outcome

$$g(E(Y_i)) = \eta, \tag{1.11}$$

so that also outcomes with distributions other than normal can be modelled (for example binary or count outcomes).

For the current purpose, generalized linear mixed model would be suitable - since the data at hand are count data. However, given the time constraints for this initial investigation we have decided to use the linear mixed model on the log-transformed variable. The main reason concerns the available implementation of multivariate models for continuous outcomes with complex covariance structures, but also the fact that the actual data we are using is *derived from* count data. The major limitation of this approach on the other hand is due to the zeros present in the data (about 30% across the five modelled microRNAs), that are lost with the log-transformation. Future research should focus on adapting existing methods to model the original counts with an appropriate generalized linear mixed model. Our results can however be an indication of *how* these methods could be used and *what* they could provide - to enable informed decisions.

1.5 Results

1.5.1 Clustering based on available data

Due to the high dimensionality of the original data, we propose to apply a dimensionality reduction technique using Kolmogorov-Smirnov (KS) tests, which are based on the marginal empirical cdfs of the healthy vs. diseased cohorts. Specifically, we tested the hypotheses

$$H_{0,j} : F_{\mathrm{miRNA}_j}(\cdot \,|\, \text{healthy}) = F_{\mathrm{miRNA}_j}(\cdot \,|\, \text{lung cancer})$$

and we included all miRNAS with p-values below 0.05 in the lower dimensional space. A subset of 27 miRNAs was selected in this manner:

$$Y_{\mathrm{KS},i,t} \in \mathbb{R}^{27}, \quad i = 1, \ldots, 7,\ t = 1, 2, 3. \tag{1.12}$$

Eight of these are also included in the 100 miRNA signature found by Rincon et al. (2019). Figure 1.5 and Figure 1.6 show two exemplary outcomes of linkage clusterings of the longitudinal data.

Figure 1.5 shows an arithmetic linkage clustering approach to the data (1.12), where the Euclidean distance is used to measure the distance between the vectors $Y_{\mathrm{KS},i,t}$ and $Y_{\mathrm{KS},k,s}$. We clearly see that the nearest neighbors of the measurement $Y_{\mathrm{KS},i,t}$ is typically one of the $Y_{\mathrm{KS},k,s}$ with $k \neq i$, i.e., measurements between individuals may very well be closer to each other than measurements of the same individual at different time points. The outcome of the clustering algorithm suggests forming three clusters, whose elements are listed in Table 1.4.

Only for person 3 and person 4, all measurements are in the same cluster. For person 1 the three measurements over time are even assigned to three different clusters. If seven clusters are formed, ideally,

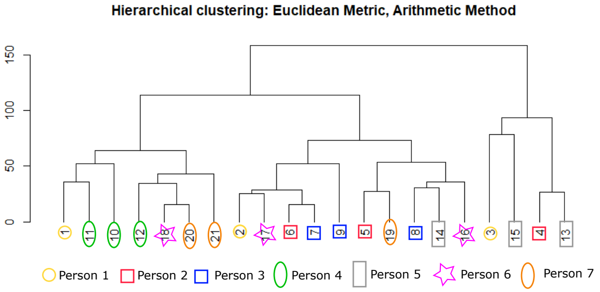

Figure 1.5: Arithmetic linkage clustering approach to the data (1.12), where the Euclidean distance is used to measure the distance between the vectors $Y_{KS,i,j}$ and $Y_{KS,k,l}$.

Cluster 1:	$Y_{\mathrm{KS},1,1}, Y_{\mathrm{KS},4,2}, Y_{\mathrm{KS},4,1}, Y_{\mathrm{KS},4,3}, Y_{\mathrm{KS},6,3}, Y_{\mathrm{KS},7,2}, Y_{\mathrm{KS},7,3}$
Cluster 2:	$Y_{\mathrm{KS},1,2}, Y_{\mathrm{KS},6,2}, Y_{\mathrm{KS},2,3}, Y_{\mathrm{KS},3,1}, Y_{\mathrm{KS},3,3}, Y_{\mathrm{KS},2,2}, Y_{\mathrm{KS},7,1},$ $Y_{\mathrm{KS},3,2}, Y_{\mathrm{KS},5,2}, Y_{\mathrm{KS},6,1}$
Cluster 3:	$Y_{\mathrm{KS},1,3}, Y_{\mathrm{KS},5,3}, Y_{\mathrm{KS},2,2}, Y_{\mathrm{KS},5,1}$

Table 1.4: Elements of the clusters as formed via average linkage clustering using the Euclidean distance, when three clusters are formed.

one would see that each person forms their own cluster. Instead, we see that the measurements of no person stay in the same cluster (see table 1.5).

Arguably, the Euclidean distance might not be the best distance measure when comparing miRNA profiles. However, while the clustering results using other metrics look different, the general tendency of measurements of the same person over time end up in different clusters, remains. To showcase this, Figure 1.6 shows the outcome of the average linkage clustering based on the Canberra metric. The same can be observed for different types of metrics and different types of linkage functions. Furthermore, we performed the same clustering algorithms for different sub-selections of miRNA, always yielding comparable results. In particular, we used a selection of five miRNA, which had been

Cluster 1:	$Y_{\mathrm{KS},1,1}, Y_{\mathrm{KS},4,2}, Y_{\mathrm{KS},4,1}$
Cluster 2:	$Y_{\mathrm{KS},4,3}, Y_{\mathrm{KS},6,3}, Y_{\mathrm{KS},7,2}, Y_{\mathrm{KS},7,3}$
Cluster 3:	$Y_{\mathrm{KS},1,2}, Y_{\mathrm{KS},6,2}, Y_{\mathrm{KS},2,3}, Y_{\mathrm{KS},3,1}, Y_{\mathrm{KS},3,3}$
Cluster 4:	$Y_{\mathrm{KS},2,2}, Y_{\mathrm{KS},7,1}, Y_{\mathrm{KS},3,2}, Y_{\mathrm{KS},5,2}, Y_{\mathrm{KS},6,1}$
Cluster 5:	$Y_{\mathrm{KS},1,3}$
Cluster 6:	$Y_{\mathrm{KS},5,3}$
Cluster 7:	$Y_{\mathrm{KS},2,2}, Y_{\mathrm{KS},5,1}$

Table 1.5: Elements of the clusters as formed via average linkage clustering using the Euclidean distance, when seven clusters are formed.

previously found in an unpublished data set via differential expression analysis, corresponding to the measurements

$$Y_{\mathrm{DE},i,t} \in \mathbb{R}^5, \quad i = 1, \ldots, 7, \; t = 1, 2, 3, \tag{1.13}$$

for the longitudinal data set and

$$Y^{\mathrm{CS}}_{\mathrm{DE},i} \in \mathbb{R}^5, \quad i = 1, \ldots, 30, \tag{1.14}$$

for the cross-sectional data.

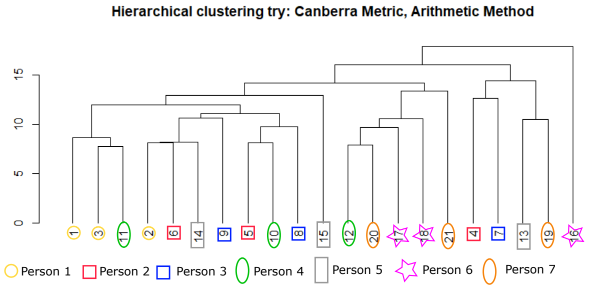

Figure 1.6: Arithmetic linkage clustering approach to the data (1.12), where the Canberra distance is used to measure the distance between the vectors $Y_{\mathrm{KS},i,t}$ and $Y_{\mathrm{KS},k,s}$.

In order to investigate whether such a clustering method can produce useful results in the context of miRNA analysis, we applied it to the cross-sectional data set $\{Y^{\text{CS}}_{\text{DE},i} \mid \quad i = 1, \ldots, 30, \}$ as well. The result (Figure 1.7) clearly shows that the observations of sick patients seem to be comparable and close to each other. In particular, one big cluster of mainly sick individuals and one big cluster of mainly healthy individuals and two smaller clusters are suggested. This shows that, up to fine tuning, such a clustering approach can yield quite reasonable results.

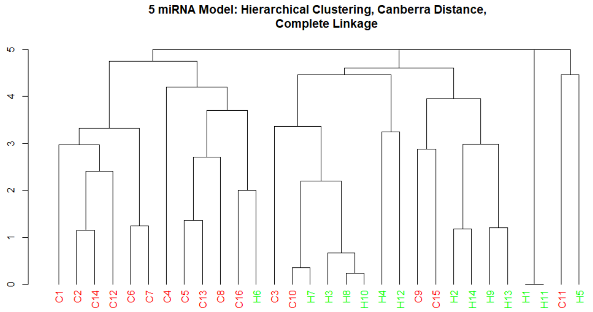

Figure 1.7: Complete linkage clustering approach to the data (1.14), where the Canberra metric is used to measure the distance between the vectors $Y^{\text{CS}}_{\text{DE},i,j}$ and $Y^{\text{CS}}_{\text{DE},k,l}$.

1.5.2 Analysis of variance

Table 1.6 contains values of the F-statistic and corresponding p-values for seven exemplary miRNAs for the longitudinal data of the seven individuals. The two miRNAs from Figure 1.4 are highlighted in blue. While certainly the normal assumption and the independence assumption are highly questionable for our data, these values give a first indication of how difficult the problem is.

A histogram of all p-values is shown in Figure 1.8. Clearly, the region from 0.45 to 0.7 is overpopulated. Most of the miRNAs that have a p-value in this region are zero for most of the measurements. If these are filtered out in a pre-processing step, the histogram would

F-value	1.432	0.870	1.947	3.864	5.661	1.222
p-value	0.271	0.541	0.143	0.017	0.0036	0.352

Table 1.6: Values of the F-statistic and corresponding p-values for seven exemplary miRNAs for the longitudinal data of the seven individuals.

look uniform with a slight elevation in the first bin, indicating that indeed, several miRNAs differ substantially between individuals. In fact, 11, 61 and 108 miRNAs have a p-value of less than $0.01, 0.05$ or 0.1, respectively. Selecting the most significant miRNAs according to this criterion, i.e., the most individually different ones, yields a selection of 11 miRNAs. We applied the complete linkage clustering algorithm to this sub-selection of miRNAs as well. The results are shown in Figure 1.9 for the Euclidean distance (right) and the Canberra distance (left).

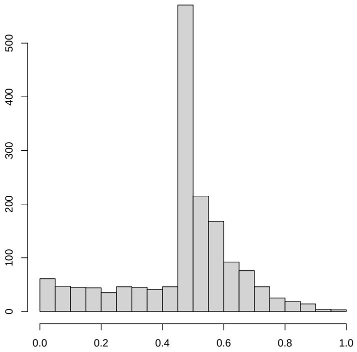

Figure 1.8: Histogram of all p-values from the F-tests.

While the clustering based on the Euclidean distance does not see any strong within person similarities as compared to between person similarities, the Canberra distance clearly does. When 7 clusters are formed, six are person specific. Only one cluster consists of all three

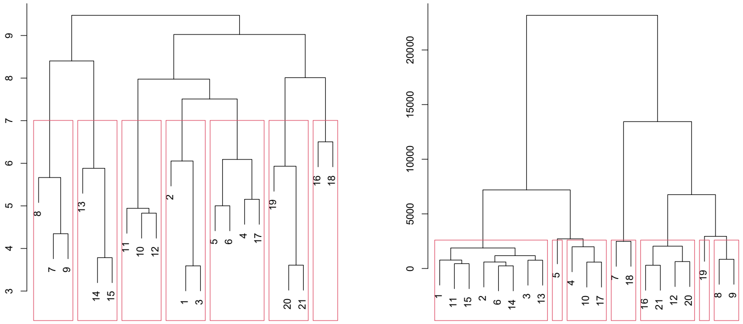

Figure 1.9: complete linkage clustering algorithm to the ANOVA subselection of miRNAs with the Euclidean distance (right) and the Canberra distance (left).

measurements of one subject and one additional measurement of another subject. This indicates once more that when comparing miRNA measurements via their distances, the Canberra distance might be a suitable measure of proximity.

Our first exploratory data analysis clearly suggests that some miRNAs might be better suited for a group baseline, whereas others require a personal baseline. This is certainly an important topic for future research.

1.5.3 Generation of synthetic data

Algorithmic details

The data generating mechanism has been coded within the MATLAB environment and in the *'Simulator_miRNA_LongDataGenMech.m'* function. The DGM takes as input the number of synthetic patients $N_{patients}$, number of time steps N_t (number of longitudinal measurements), and a structure containing options and additional parameters for the transition model. The simulator generates n_t longitudinal samples of $N_{miRNA} = 1421$ miRNA concentrations, health indices (0 health, 1 sick), and disease progression coefficients $\rho(t)$ for each patient. The option input structure contains three fields:

1. *Option.Tran_prob_Healthy2Sick* that represents the healthy to sick transition probability P_{H2S}.

2. *Option.Sick_progression_interval* that defines the number of longitudinal measurements needed for the disease to fully develop.

3. *Option.UseCorrelation* a Boolean index defining weather or the empirical correlation $\hat{\Sigma}$ has to be used when sampling $miRNA$ profiles.

Six are the outputs of the data simulator:

- i *miRNA*: cell array $[1 \times N_{patients}]$ with the miRNA samples (grouped by patients), where each elements is a $[N_{miRNA} \times N_t]$ matrix.
- ii *Time vector*: Vector of time indices $(1, 2, ..., N_t)$.
- iii *Health indicators*: $[N_{patients} \times N_t]$ matrix of Boolean health indicators.
- iv *miRNA names*: $[N_{miRNA} \times 1]$ string containing the names of the miRNAs.
- v *Sick Percentage*: $[N_{patients} \times N_t]$ matrix. Each element in the matrix defines the sick percentage indicator $\rho(t)$.
- vi *Data per miRNAType*: cell array $[1 \times N_{miRNA}]$ with miRNA samples (grouped by miRNA type), where each element is a $[N_{patients} \times N_t]$ matrix.

The data generating mechanism runs very efficiently.
If *Option.UseCorrelation* is set to FALSE, the function took 6.5 seconds to generate data for $N_{patients} = 1000$ over a 2 years longitudinal study ($N_t = 8$). On the other hand, 16 seconds were needed to generate $N_{patients} \times N_t = 1000 \times 8$ applying the correlation structure.

Figure 1.10 presents an example of conditional marginal CDF $F_X(x|y)$ for healthy and sick patients and compare simulated data and real measurements. Note that the simulated marginal CDFs are very similar to the empirical marginal density and, thus, the probabilistic behaviour of

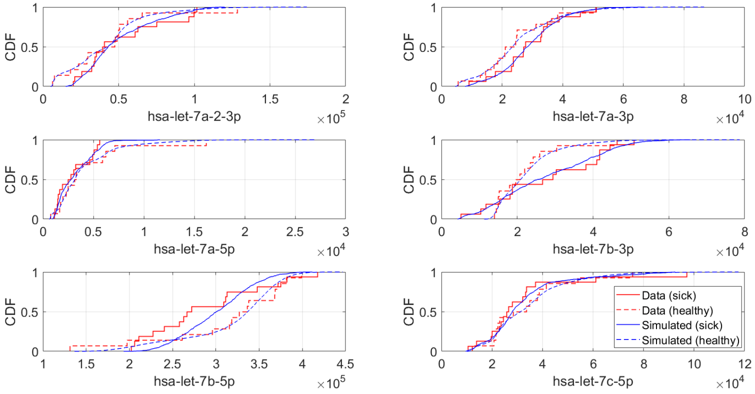

Figure 1.10: A comparison between simulated and experimental F_x for six miRNA types. Solid and dashed red lines display, respectively, the empirical CDFs for sick and healthy patients. The distributions of the simulated data miRNA are presented by blue CDFs.

the real data (at least the behaviour of the marginals) is well-captured by the simulator. Figure 1.11 shows correlated samples for the simulated (blue) vs experimental data (red markers). Qualitatively the simulated samples display overall a reasonable trend, although sub-optimal fitting can be observed for some of the miRNA intances. As example, ’hsa-let-7a-2-3p’ shows to be strongly correlated (linearly) with ’hsa-let-7b-3p’, see red markers in the top right panel of Figure 1.11. Unfortunately this strong correlation is partially lost in the synthetic data, i.e., the blue markers (synthetic samples) are still positively correlated but with larger dispersion. Despite these limitations, the proposed data simulation tools offer a valuable contribution and can be used to design, test and verify predictive models before expensive data collection is carried out. This can speed up algorithmic developments, inform further data collection, and improve the overall effectiveness of the study.

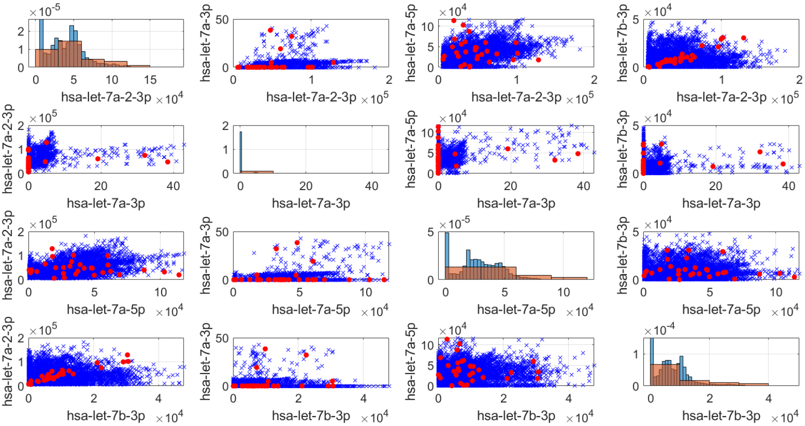

Figure 1.11: An example of correlated synthetic miRNA samples (blue markers) versus the experimental data (red markers). The off-diagonal panels present pairs-wise comparison of four selected miRNA concentrations and the panels on the diagonal compare the marginal distribution of the data (red histograms) and the simulated samples (blue histograms).

1.5.4 Clustering based on synthetic data

The results for the clustering based on the simulated data are comparable to what we obtained for the real data for the longitudinal data sets. However, the separation in healthy versus sick seams is slightly clearer for the real data.

1.5.5 POMDP and mixed effects model

We have fitted a mixed model jointly to a selection of five miRNAs, $Y_{\mathrm{DE},i,t}$, that were indicated as informative by the problem owner, and were also in large part found again in the baseline analysis (cf. Section 1.4.1). Among the fixed effect we included a distinct intercept (microRNA-specific average), and a distinct effect for time (taken as categorical, so that no trend was imposed a priori) for each of the microRNAs. The model has no random effects, and a Kronecker product

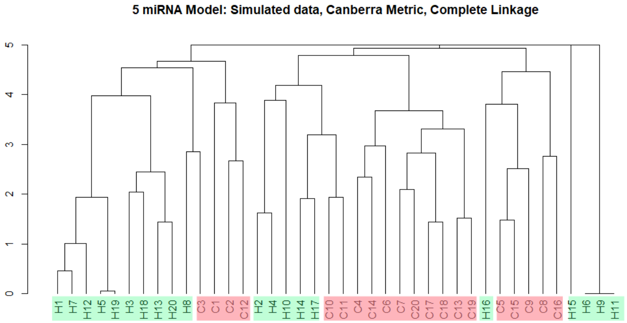

Figure 1.12: Complete linkage clustering of a synthetic data set.

for the covariance matrix as illustrated in Section 1.4.3. The estimates of the fixed effect can be found in Table 1.7. For each of the modelled microRNAs, we report the estimates of the intercept and the estimate at two time points. The third (last) is taken as reference (= 0). Beside the estimate we report the standard error and the corresponding p-value.

More of interest for the current analysis are the estimated variance-covariance parameters between microRNAs (Table 1.8), and the estimated autoregressive coefficient for the correlation in time. These can be found in Table 1.8, together with their standard errors and p-values. None of the covariances between microRNAs is estimated to be significantly different from zero.

1.6 Conclusions and Recommendations

1.6.1 Pros and cons of the proposed approaches

We explored the notion of a baseline using a classification approach and ANOVA. This exploratory data analysis suggests that some miRNAs might be better suited for a group baseline, whereas others might be better suited for a personal baseline. Therefore, a hybrid version might be a good solution and is certainly a direction to think about more in

Intercept				
microRNA		estimate	std. error	pvalue
$Y_{\mathrm{DE}}(1)$		2.85	0.50	<.0001
$Y_{\mathrm{DE}}(2)$		2.95	0.39	<.0001
$Y_{\mathrm{DE}}(3)$		3.04	0.31	<.0001
$Y_{\mathrm{DE}}(4)$		4.01	0.33	<.0001
$Y_{\mathrm{DE}}(5)$		2.63	0.62	0.0178
Effect of time				
microRNA	time	estimate	std. error	pvalue
$Y_{\mathrm{DE}}(1)$	1	1.63	0.70	0.0335
	2	0.44	0.61	0.4877
$Y_{\mathrm{DE}}(2)$	1	0.11	0.60	0.8564
	2	-0.44	0.49	0.3889
$Y_{\mathrm{DE}}(3)$	1	0.35	0.56	0.5469
	2	-0.28	0.38	0.4867
$Y_{\mathrm{DE}}(4)$	1	0.47	0.58	0.4356
	2	-0.20	0.41	0.6239
$Y_{\mathrm{DE}}(5)$	1	0.62	1.10	0.6089
	2	0.22	0.76	0.7870

Table 1.7: Fixed effect estimates

the future. As a general finding it seems that complete linkage clustering based on the Canberra metric seems to suitable to find patterns in our data, whereas other metrics and dissimilarity functions as well as other linkage functions could not provide convincing results.

We developed a numerical simulator to generate large amount of synthetic longitudinal miRNA data. The model was used to efficiently simulate a 2-years long longitudinal study with 10^3 - 10^4 volunteers and only took a few minuets to generate this large amount of labelled longitudinal samples. We captured correlation between miRNA concentrations within the simulated data and the marginal distributions of the synthetic miRNA realizations well-mimic the probabilistic behaviour of the empirical data. Because the simulator provides data for sick patients and for the disease progression (although only artificial), it can be conveniently used for the numerical validation and verifica-

	estimate	std. error	pvalue
VAR($Y_{\mathrm{DE}}(1)$)	1.58	0.56	0.0024
VAR($Y_{\mathrm{DE}}(2)$)	0.74	0.41	0.0374
VAR($Y_{\mathrm{DE}}(3)$)	0.59	0.32	0.0355
VAR($Y_{\mathrm{DE}}(4)$)	0.76	0.30	0.0057
VAR($Y_{\mathrm{DE}}(5)$)	2.15	1.91	0.1310
COV($Y_{\mathrm{DE}}(1), Y_{\mathrm{DE}}(2)$)	0.40	0.40	0.3156
COV($Y_{\mathrm{DE}}(1), Y_{\mathrm{DE}}(3)$)	-0.02	0.25	0.9393
COV($Y_{\mathrm{DE}}(1), Y_{\mathrm{DE}}(4)$)	-0.10	0.29	0.7357
COV($Y_{\mathrm{DE}}(1), Y_{\mathrm{DE}}(5)$)	-0.13	0.52	0.8046
COV($Y_{\mathrm{DE}}(2), Y_{\mathrm{DE}}(3)$)	0.11	0.22	0.6103
COV($Y_{\mathrm{DE}}(2), Y_{\mathrm{DE}}(4)$)	0.33	0.31	0.2827
COV($Y_{\mathrm{DE}}(2), Y_{\mathrm{DE}}(5)$)	-0.46	0.58	0.4305
COV($Y_{\mathrm{DE}}(3), Y_{\mathrm{DE}}(4)$)	0.50	0.27	0.0696
COV($Y_{\mathrm{DE}}(3), Y_{\mathrm{DE}}(5)$)	-0.12	0.45	0.7910
COV($Y_{\mathrm{DE}}(4), Y_{\mathrm{DE}}(5)$)	-0.87	0.78	0.2598
AR(1)	0.24	0.22	0.2741

Table 1.8: Covariance paramter estimates.

tion of the data analysis tools and to speed up the construction of data analysis pipelines. For instance, it could be used to test the accuracy of classification and disease detection methods and models to define a baseline for healthy patients before more data is collected. Another advantage of the proposed method is that it is possible to scale up and tune the simulator with new experimental evidence, e.g., new miRNA samples and discovered relationships between miRNA concentrations and specific diseases and illness progressions. The simulation model has however some limitations, specifically, (i) lack of data and knowledge in literature makes it difficult to define a realistic model; (ii) the model is relatively simple and for the time being only incorporates one illness, neglects time-correlations (auto correlation and correlations between miRNAs) and neglects population heterogeneity; (iii) because synthetic samples are generated from the empirical marginals, this artificially reduces the uncertainty in the distribution of healthy and sick patients'; (iv) transition from healthy assumed linear;(v) censoring and study dropouts neglected; (vi) non-disease related changes not taken into account; (vii) no personalized baseline.

1.6.2 Future research and recommendations

This study shows that it is difficult to prescribe personalized solutions from uncertain low-density mRNA. More samples are needed to get reliable results, e.g. though the URIMON study. In the meantime, our synthetic data can be used to test and validate data analysis and prediction methods. Conceptual approaches for timely detection of disease based on temporal evolution of miRNA counts have been discussed and reviewed (POMDP,mixed effect models). These need to be further developed. Possibly, a combination of POMDP and mixed effect models could be a feasible solution. In order to understand the properties of the data better, quantification of the uncertainty in the measurement process would be very helpful, e.g., repeated measurements on the same urine sample. in order to be able to characterize the variability in the mRNA density per-patient, 1 sample per day (say) for 10 healthy persons for a week or two could be collected and analyzed. Research directions for the future are Statistical Process Control theory and concepts, such as self-starting control charts to automatically obtain personalized baselines of healthy patients. Combinations of SPC and Markov Decision Processes and SPC and mixed effect models exist in other application areas than health. Therefore, we believe that the combination of POMDP, mixed effect models and SPC is a good strategy to profit from the best of the three worlds.

References

Barger, Jennifer F and S Patrick Nana-Sinkam (2015). “MicroRNA as tools and therapeutics in lung cancer”. In: *Respiratory medicine* 109.7, pp. 803–812.

Birnbaum, Aharon et al. (2013). “Minimax bounds for sparse PCA with noisy high-dimensional data”. In: *Ann. Statist.* 41.3, pp. 1055–1084.

Blackwell, Jacob N. et al. (2020). “Early Detection of In-Patient Deterioration: One Prediction Model Does Not Fit All”. In: *Critical Care Explorations* 2.

Calin, G.A. et al. (2002). “Frequent deletions and down-regulation of micro-RNA genes miR-15 and miR-16 at 13q14 in chronic lymphocytic leukemia.” In: *Proc. Natl. Acad. Sci. U. S. A.* 99, pp. 15524–15529.

Cho, Haeran and Piotr Fryzlewicz (2015). “Multiple-change-point detection for high-time series via sparsified binary segmentation”. In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 77.2, pp. 475–507.

Condrat, Carmen Elena et al. (2020). “miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis”. In: *Cells* 9.2, p. 276.

Galvão-Lima, L.J., A.H.F. Morais, R.A.M. Valentim, et al. (2021). “MiRNAs as biomarkers for early cancer detection and their application in the development of new diagnostic tools”. In: *BioMed Eng OnLine* 20.21.

Han, Y. et al. (2020). “Statistical approaches using longitudinal biomarkers for disease early detection: A comparison of methodologies.” In: *Stat Med.* 39, pp. 4405–4420.

Jirak, Moritz (2015). “Uniform change point tests in high dimension”. In: *Ann. Statist.* 43.6, pp. 2451–2483.

Kenett, R.S. and M. Pollak (2012). “On assessing the performance of sequential procedures for detecting a change”. In: *Quality and Reliability Engineering International* 28.5, pp. 500–507.

Li, Jing-Hua et al. (2017). “MiR-205 as a promising biomarker in the diagnosis and prognosis of lung cancer”. In: *Oncotarget* 8.54, p. 91938.

Li, Sai, T. Tony Cai, and Hongzhe Li (2021). “Inference for High-Dimensional Linear Mixed-Effects Models: A Quasi-Likelihood Approach”. In: *Journal of the American Statistical Association* 0.0, pp. 1–12. DOI: 10.1080/01621459.2021.1888740.

Linderman, K., K.E. McKone-Sweet, and J.C. Anderson (2005). “An integrated systems approach to process control and maintenance”. In: *European Journal of Operational Research* 164.2, pp. 324–340.

Lopez-Rincon, Alejandro et al. (2020). “Machine Learning-Based Ensemble Recursive Feature Selection of Circulating miRNAs for Cancer Tumor Classification”. In: *Cancers* 12.7.

Lu, Thomas X. and Marc E. Rothenberg (2018). "MicroRNA". In: *Journal of Allergy and Clinical Immunology* 141.4, pp. 1202–1207.

Ludwig, Nicole et al. (Feb. 2016). "Distribution of miRNA expression across human tissues". In: *Nucleic Acids Research* 44.8, pp. 3865–3877.

Mehrafrooz, Z. and R. Noorossana (2011). "An integrated model based on statistical process control and maintenance". In: *Computers & Industrial Engineering* 61.4, pp. 1245–1255.

Or, Gilad and Isana Veksler-Lublinsky (Mar. 2021). "Comprehensive machine-learning-based analysis of microRNA-target interactions reveals variable transferability of interaction rules across species". In: *BMC BioInformatics* 22.

Panagiotidou, S. and G. Tagaras (2010). "Statistical process control and condition-based maintenance: A meaningful relationship through data sharing". In: *Production and Operations Management* 19.2, pp. 156–171.

Poor, H.V. and O. Hadjiliadis (2008). *Quickest Detection.* Cambridge University Press.

Rincon, Alejandro Lopez et al. (2019). "Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection". In: *BMC Bioinformatics* 20.1.

Sapre, Nikhil et al. (2016). "A urinary microRNA signature can predict the presence of bladder urothelial carcinoma in patients undergoing surveillance". In: *British Journal of Cancer* 114, pp. 454–462.

Townsend, N., D. Kazakiewicz, F. Lucy Wright, et al. (2022). "Epidemiology of cardiovascular disease in Europe." In: *Nat Rev Cardiol* 19, pp. 133–143.

Wainwright, Martin J. (2019). *High-dimensional statistics – A non-asymptotic viewpoint.* Vol. 48. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Weinstein, J.N., E.A. Collisson, and et al (2013). "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nat Genet.* 45.10, pp. 1113–1120.

Zhong, Ping-Shou, Jun Li, and Piotr Kokoszka (2021). "Multivariate analysis of variance and change points estimation for high-dimensional longitudinal data". In: *Scand. J. Stat.* 48.2, pp. 375–405.