VU University Amsterdam
Faculty of Sciences

University of Warsaw
Faculty of Mathematics, Computer Science and Mechanics

Master of Science Programme

Beata Roś

Student no. 241862 (UW), 2002493 (VU)

Modeling cancer progression as reflected by DNA copy number data

Master's thesis
in MATHEMATICS

Supervisors:

Wessel van Wieringen
Department of Epidemiology and Biostatistics,
VU University medical center
Mark van de Wiel
Department of Mathematics Faculty of Sciences,
Vrije Universiteit

Second reader:

Marianne Jonker
Department of Mathematics Faculty of Sciences,
Vrije Universiteit

July 2010

Author's statement

Hereby I declare that the present thesis was prepared by me and none of its contents was obtained by means that are against the law.

I also declare that the present thesis is a part of the Joint Master of Science Programme of the University of Warsaw and the Vrije Universiteit in Amsterdam. The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

Background: DNA copy number aberrations are common in cancer cells. DNA copy number data contains information about DNA copy number aberrations. The important property of DNA copy number data is high dimensionality, which means that in the data set there is typically more DNA regions- variables than samples.

Objectives: Investigating the relation between changes in DNA copy number along the genome and cancer progression has been the topic of many research projects. In this study we aim to develop methodology to reconstruct order, according to which genetic aberrations in specific DNA regions are accumulated during cancer progression.

Design and Methods: We build univariate and multivariate continuous time Markov models that describe cancer stage at any time point. The time is measured from the start of the cancer. We make prior assumption about the distribution of time as time does not appear in the data. Our models are used to obtain time estimates for samples. Those estimates are relevant for reconstructing the order of accumulating DNA aberrations.

Results: Several models were built. Estimation procedures were obtained and implemented in R. The accuracy of models' estimators was evaluated on simulated data with a small number of DNA regions. Time estimation procedures were implemented for all multivariate models. We fitted those models to the real data set of breast cancer patients. We found interesting properties of time estimates of advanced models.

Keywords

DNA copy number data, aCGH, cancer progression

Thesis domain (Socrates-Erasmus subject area codes)

11.2 Statistics

Subject classification

92. Biology and other natural sciences
92D. Genetics and population dynamics
92D20. Protein sequences, DNA sequences

Contents

Chapter 1

Introduction

The topic of this thesis requires introducing some basic concepts of molecular biology and cancer. In this chapter it will also be explained what DNA copy number data is. The procedure of collecting the data will be presented.

1.1. Biological background

DNA is a nucleic acid that contains genetic information of living organisms and some viruses. More precisely, information included in DNA is a sequence that codes proteins' structures. This information is needed for protein synthesis, which is essential for any living organism.

Each cell of a living organism contains DNA. There are special molecules in which DNA is kept. Those molecules are called chromosomes. Human somatic cells contain two sex chromosomes and 22 pairs of autosomal (non-sex) chromosomes. In this research we will focus on autosomal DNA.

Gene is a sequence of DNA that codes information of one protein. Each gene is located in a special place of a particular chromosome.

Chromosomal aberrations are disorders of the structure of the set of chromosomes or deletions or multiplications of the part of a chromosome. Some chromosomal aberrations can cause serious diseases.

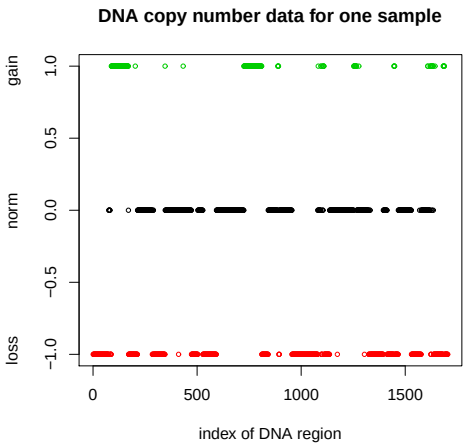
1.2. Cancer

Cancer is a group of diseases. Characteristic feature of those diseases is that there are cells with genetic disorders. By uncontrolled divisions those genetic abnormalities appear in newly created cells. During cell divisions of genetically disordered cells it is likely to happen, that the new genetic modification will appear. Thus cancer cells that were created later tend to have more genetic disorders than those from the earlier stage of cancer.

1.3. DNA copy number

Human somatic healthy cells contain two copies of each autosomal region of DNA. Cancer cells contain genetic disorders, therefore this number can be different for some regions due to deletions or multiplications. DNA copy number data contains information about number of DNA copies for each autosomal region. The DNA copy number data for the i th sample and j th region is one of three possibilities:

Figure 1.1: Data of one cancer sample



- 'normal'- if the number of DNA copies=2 in sample i, region j,
- 'gain'- if the number of DNA copies>2, in sample i, region j,
- 'loss'- if the number of DNA copies<2, in sample i, region j.

Incorrect numbers of DNA copies, so 'gains' and 'losses' are called aberrations.

Figure 1.1 is an example of DNA copy number of one cancer sample. The black points indicate regions with normal DNA copy number, red points- regions with decreased DNA copy number and green points- regions with increased DNA copy number.

It is important that the DNA regions are indexed according to the order in which they are situated on the genome. Therefore regions that are neighbours on the chromosome are also neighbours on the vector containing all regions.

1.4. How is DNA copy number measured

The DNA copy number data is measured during array Comparative Genomic Hybridization (aCGH) experiment. For the experiment a DNA microarray is needed. This is a glass or plastic surface with attached small spots of different sequences of DNA. For applying aCGH one needs to extract autosomal DNA from a sample of tumor and from a control sample. These two DNA samples are differentially labelled with different fluorochromes. They both are hybridized to the DNA microarray. Labelled sequences of DNA bind to complementary sequences from the microarray. The microarray scanner measures the fluorescent signals. For each sequence of DNA from microarray, fluorescence ratio of test sample with respect to control sample is computed. This ratio indicates the relative DNA copy number of tumor sample at this genomic sequence. The set of ordered ratios from the array with corresponding genomic positions gives the experimental data of one tumor sample. Therefore the output of the aCGH experiment is a set of ratios indexed by DNA regions. More detailed description of the aCGH experiment can be found in [1] and [2].

1.5. The data and its preprocessing

Suppose we apply aCGH technique for more tumor samples from different people. We need to preprocess these data. In [3] there are identified steps that are applied during preprocessing stage. These are:

1. Removing useless or poor quality samples.
2. Inserting estimates of missing values.
3. Normalization.
4. Removing outliers.
5. Choosing the levels that will represent DNA copy number.
6. Segmenting regions of DNA with different copy number levels.
7. Classification of segments to copy number levels.

During the first step the data with unknown position information and the data with too many missing values are removed.

For the samples that were not deleted in the first step, but include some missing values, estimates of this missing values are inserted during the second step.

In order to compare the data, normalization is applied. It is done during the third step. In the fourth step outliers are removed from the data.

Although copy number can be any integer, we group it as several levels. Those levels have to be chosen. We need to take at least three levels: one for normal copy number, one indicating 'loss' and one indicating 'gain' of DNA.

In order to assign levels to all regions, first we want to obtain segments of DNA with different copy number levels. There are several algorithms that are used for segmentation. The segmentation is done during the fifth step. Therefore for each sample we obtain a partition of regions into separate intervals. We assume that regions which are located in the same interval have equal DNA copy number.

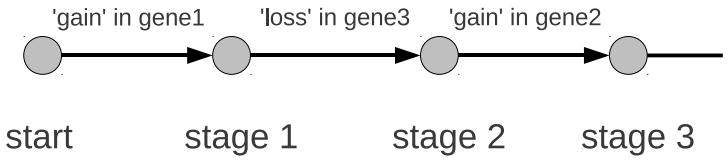
In the last step classification procedure is used to assign copy number levels to segments. This step is called 'calling'. After calling procedure we obtain DNA copy number data.

There is a special R- package: CGHcall, in which those steps are implemented. More information about algorithms used in this package can be found in [4].

1.6. Objective of the thesis

We believe that there exists a pattern, according to which DNA aberrations are accumulated during cancer progression. This pattern is indicative for the cancer progression stage. Figure 1.2 is an example of such a pattern. We assume that aberrations are accumulated according to a linear order. The aim of this thesis is to reconstruct this order. For this purpose we build continuous time Markov chains to describe cancer progression in terms of DNA copy number changes. DNA copy number data of cancer samples is used to estimate the parameters of the Markov chains.

Figure 1.2: Pattern of accumulating aberrations



Chapter 2

General framework for models

In this chapter we present assumptions, that will be used in all models we build in the next chapters.

2.1. Models assumptions

Each of the models is a continuous time Markov chain. We measure the time from the start of the cancer, therefore $t \geq 0$.

2.1.1. States

Denote by J the number of regions we consider in our model. In case we build a model for a region j only, then the state of the chain at time t will be $Z_j^{(t)}$. If we build a multivariate model, then the state of the chain at time t is denoted by $\mathbf{Z}^{(t)} = (Z_1^{(t)}, \ldots, Z_J^{(t)})$. Now we must say what is the state space of the model. We assume that for each region j there are three possibilities for the $Z_j^{(t)}$. These possibilities are: N, G, L. N indicates correct number of DNA copies at region j, while G indicates 'gain' and L 'loss' of DNA at region j. We can have each of this three possibilities at any region at time $t > 0$, therefore the state space at time $t > 0$ is $\{N, G, L\}^J$. This is not the case for the time $t = 0$. This is because we assume that $\mathbf{Z}^{(0)} = (Z_1^{(0)}, \ldots, Z_J^{(0)}) = (N, \ldots, N)$. The reason for stating this assumption is that a healthy cell usually has a correct DNA copy number at all regions. Additionally, we assume that genomic aberrations are accumulated after the start of the cancer.

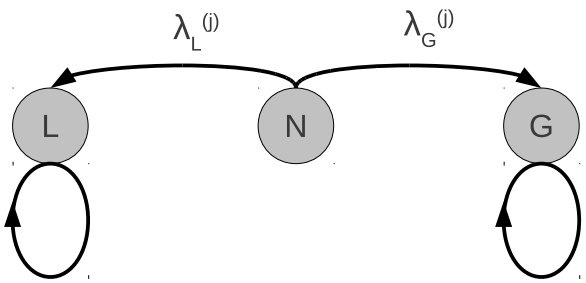
2.1.2. Model parameters

A general assumption is that we consider two parameters for each DNA region. We take $\lambda_G^{(j)}$- a 'gain' parameter for region j and $\lambda_L^{(j)}$- a 'loss' parameter for region j. We need to assume that $\lambda_G^{(j)}, \lambda_L^{(j)} \geq 0$.

2.1.3. Transition probabilities

We state the assumption that for any region j, this region will accumulate a DNA aberration, so 'gain' or 'loss' of DNA copy will appear after some time. After the transition from a state N at region j has already appeared, there will be no more transitions at this region. Therefore G and L are absorbing states for each region. Figure 2.1 shows what are possible transitions within one region.

Figure 2.1: Transition scheme for j th region



We assume that for any fixed time t regions behave independently and the time needed to leave the state N at region j is exponentially distributed with the parameter $\lambda_G^{(j)} + \lambda_L^{(j)}$. Therefore it holds that

$$\mathbb{P}(Z_j^{(t)} = N) = e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}.$$

Because $\lambda_G^{(j)}$ is a 'gain' parameter and $\lambda_L^{(j)}$ a 'loss' parameter for a region j and it must be $\mathbb{P}(Z_j^{(t)} = G) + \mathbb{P}(Z_j^{(t)} = L) = 1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}$, then we assume:

$$\mathbb{P}(Z_j^{(j)} = G) = \frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}\right),$$

$$\mathbb{P}(Z_j^{(j)} = L) = \frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}\right).$$

From the independence between regions at time t we have that for any $u = (u_1, \ldots, u_J)$ in the state space it holds that:

$$\mathbb{P}(\mathbf{Z}^{(t)} = u) = \prod_{j=1}^{J} \mathbb{P}(Z_j^{(t)} = u_j).$$

2.2. The data

We will use DNA copy number data of cancer samples to obtain estimates for parameters of the Markov chain. We consider I samples, each sample consists of J DNA regions.

2.2.1. Notation for the data

By $X_{i,j}$ we denote the DNA copy number of i th sample and j th region. We also use notation $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,J})$ for the DNA copy number of i th sample and all regions.

2.2.2. Real data set

In this research we test algorithms and fit our models to the real data set. We use DNA copy number data of breast cancer patients. This data contains 263 samples of 1704 genomic regions. The data is a subset of a publicly available data, which was used in [5].

2.2.3. Relation between the data and a Markov chain

For each model we consider two different states: state of the Markov chain and state of the data. It must be indicated what is the assumed relation between those two states. Let $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,J})$ be the state of the i th sample. If we knew that the time for this sample was T_i, then $\mathbf{X}_i$ would be the point of the trajectory of the process $\mathbf{Z}$ at time T_i. Thus in this case $\mathbf{X}_i$ would be a sample from the distribution of $\mathbf{Z}^{(T_i)}$. From the assumed independence between regions of the Markov chain at a given time point, we can conclude that $X_{i,j}$ would be a sample from the distribution of $Z_j^{(T_i)}$ for $j \in \{1, \ldots, J\}$.

2.2.4. Dealing with time unobserved

The important property of DNA copy number data is that there is no information about time that passed from the start of the cancer until taking sample. Therefore we treat T- time from the start of the cancer as latent random variable. Because we want to use the data to estimate parameters of Markov chain, which involves time, we need to make prior assumption about the distribution of T. For all models we assume that T is exponentially distributed with parameter μ for a chosen $\mu > 0$. $f_T(t)$ denotes the density of the prior time distribution at t, therefore $f_T(t) = \mu e^{-\mu t}$.

2.2.5. Obtaining state distribution of the data

Thanks to assuming prior distribution of T and using the law of total probability we can obtain the distribution of our data. For a state N it holds that:

$$\mathbb{P}(X_{i,j} = N) = \int_0^\infty \mathbb{P}(X_{i,j} = N | T = t) f_T(t) dt = \int_0^\infty \mathbb{P}(Z_j^{(t)} = N) f_T(t) dt.$$

Using marginal state distribution for the Markov chain and applying the same rules for states G and L we obtain marginal state distribution for a region j, which is specified by those expressions:

$$\mathbb{P}(X_{i,j} = N) = \int_0^\infty e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t} f_T(t) dt,$$

$$\mathbb{P}(X_{i,j} = G) = \int_0^\infty \frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}\right) f_T(t) dt,$$

$$\mathbb{P}(X_{i,j} = L) = \int_0^\infty \frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}\right) f_T(t) dt,$$

where $f_T(t) = \mu e^{-\mu t}$. The similar expression holds for the multivariate state distribution. For any $(u_1, \ldots, u_J)$ in the state space we have that

$$\mathbb{P}(\mathbf{X}_i = u) = \int_0^\infty \left(\prod_{j=1}^J \mathbb{P}(Z_j^{(t)} = u_j) \right) f_T(t) dt.$$

Because now we have the state distribution of the data, we are able to obtain expression for the likelihood function, which will be used to obtain parameters' estimators. The second usage of the state distribution is that we can simulate data from this distribution. Data simulation is useful for testing the accuracy of parameters estimation techniques.

Chapter 3

Time estimation

3.1. Time estimation for a general model

Recall that DNA copy number data does not contain information about time from the beginning of cancer progression until taking the sample. We would like to sort samples according to time from the start of the cancer. Because of the lack of time in the data, we will estimate it for each sample. In each model we have a prior assumption that the time is a random variable exponentially distributed with parameter μ. We will use this prior assumption for the estimators.

Suppose we have estimates of $\lambda_G^{(j)}$ and $\lambda_L^{(j)}$ for $j \in \{1, \cdots, J\}$. We will use them to obtain estimates of time for each sample. We consider the following estimator:

$$\hat{T_i} = \mathbb{E}(T|\mathbf{X}_i) = \int_0^\infty t f_T(t|\mathbf{X}_i)dt,$$

where $f_T(t) = \mu e^{-\mu t}$. By Bayes rule, it holds that

$$\hat{T_i} = \frac{\int_0^\infty t\mathbb{P}(\mathbf{X}_i|T=t)f_T(t)dt}{\int_0^\infty \mathbb{P}(\mathbf{X}_i|T=t)f_T(t)dt}.$$

From assumptions stated in the previous chapter we have that

$$\mathbb{P}(\mathbf{X}_i|T=t) = \mathbb{P}((X_{i,1}, \ldots, X_{i,J})|T=t) =$$

$$= \prod_{j=1}^{J} \mathbb{P}(Z_j^{(t)} = G)^{\mathbf{1}_{\{X_{i,j}=G\}}} \mathbb{P}(Z_j^{(t)} = L)^{\mathbf{1}_{\{X_{i,j}=L\}}} \mathbb{P}(Z_j^{(t)} = N)^{\mathbf{1}_{\{X_{i,j}=N\}}},$$

where in the expressions of type $(Z_j^{(t)} = L)^{\mathbf{1}_{\{X_{i,j}=L\}}}$, we state the convention that $0^0 = 1$. From this moment, for all other exponentiations, where the exponent contains an indicator, we assume that $0^0 = 1$.

Let $M_i(t) = \mathbb{P}(\mathbf{X}_i|T=t)f_T(t)$. The aim is to compute $\frac{\int_0^\infty tM_i(t)dt}{\int_0^\infty M_i(t)dt}$. Note that

$$M_i(t) = \mu e^{-\mu t} \prod_{j=1}^{J} \left(\left(\frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \right)^{\mathbf{1}_{\{X_{i,j}=L\}}} \left(\frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \right)^{\mathbf{1}_{\{X_{i,j}=G\}}} \right) \cdot$$

$$\cdot \prod_{j=1}^{J} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}\right)^{\mathbf{1}_{\{X_{i,j}\neq N\}}} e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t \cdot \mathbf{1}_{\{X_{i,j}=N\}}}.$$

It is easily seen that

$$M_i(t) = \mu e^{-\mu t} \prod_{j=1}^{J} \left(\left(\frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \right)^{\mathbf{1}_{\{X_{i,j}=L\}}} \left(\frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \right)^{\mathbf{1}_{\{X_{i,j}=G\}}} \right) \cdot$$

$$\cdot e^{-c_i t} \cdot \prod_{j=1}^{J} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t} \right)^{\mathbf{1}_{\{X_{i,j} \neq N\}}},$$

where $c_i = \sum_{j=1}^{J} (\lambda_G^{(j)} + \lambda_L^{(j)}) \cdot \mathbf{1}_{\{X_{i,j}=N\}}$. Because the part

$$\mu \prod_{j=1}^{J} \left(\left(\frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \right)^{\mathbf{1}_{\{X_{i,j}=L\}}} \left(\frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \right)^{\mathbf{1}_{\{X_{i,j}=G\}}} \right)$$

does not depend on t, we can reduce it from denominator and numerator of $\frac{\int t M_i(t) dt}{\int M_i(t) dt}$, so it holds that $\frac{\int t M_i(t) dt}{\int M_i(t) dt} = \frac{\int t \widetilde{M_i(t)} dt}{\int \widetilde{M_i(t)} dt}$, where

$$\widetilde{M_i(t)} = e^{-\mu t} e^{-c_i t} \cdot \prod_{j=1}^{J} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t} \right)^{\mathbf{1}_{\{X_{i,j} \neq N\}}}.$$

Therefore we can deal with $\widetilde{M_i(t)}$, which is simpler than $M_i(t)$, instead of $M_i(t)$. Note that

$$\widetilde{M_i(t)} = e^{-(c_i + \mu)t} \cdot \prod_{j \in R_i} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t} \right),$$

where $R_i = \{j \in J : X_{i,j} \neq N\}$. Because we can express product of sums as a sum by a generalization of the binomial theorem, we can see that

$$\widetilde{M_i(t)} = \sum_{k=0}^{|R_i|} \left(\sum_{N_k \subseteq R_i : |N_k| = k} (-1)^k e^{-\left(c_i + \mu + \sum_{\tau \in N_k} (\lambda_G^{(\tau)} + \lambda_L^{(\tau)}) \right) t} \right).$$

By partial integration we obtain that

$$\int t e^{-st} dt = -\frac{t e^{-st}}{s} - \frac{e^{-st}}{s^2},$$

and thus

$$\int_0^\infty e^{-st} dt = \frac{1}{s^2}$$

for $s > 0$. Therefore it holds that

$$\int_0^\infty t \widetilde{M_i(t)} dt = \frac{1}{(c_i + \mu)^2} + \sum_{k=1}^{|R_i|} \left(\sum_{N_k \subseteq R_i : |N_k| = k} (-1)^k \frac{1}{\left(c_i + \mu + \sum_{\tau \in N_k} (\lambda_G^{(\tau)} + \lambda_L^{(\tau)}) \right)^2} \right).$$

Because

$$\int e^{-st} dt = -\frac{1}{s} e^{-st},$$

then

$$\int_0^\infty e^{-st}dt = \frac{1}{s}$$

for $s > 0$. Thus it holds that

$$\int_0^\infty \widetilde{M_i(t)}dt = \frac{1}{c_i+\mu} + \sum_{k=1}^{|R_i|}\left(\sum_{N_k\subseteq R_i:|N_k|=k}(-1)^k \frac{1}{\left(c_i+\mu+\sum_{\tau\in N_k}(\lambda_G^{(\tau)}+\lambda_L^{(\tau)})\right)}\right).$$

3.1.1. Removing the dependence on μ

For any region j, we can express parameters describing this region in the form:

$$\lambda_G^{(j)} = \mu\widetilde{\lambda_G^{(j)}},$$

$$\lambda_L^{(j)} = \mu\widetilde{\lambda_L^{(j)}}.$$

Therefore $c_i = \mu\sum_{j=1}^J(\widetilde{\lambda_G^{(j)}}+\widetilde{\lambda_L^{(j)}})\cdot\mathbf{1}_{\{X_{i,j}=N\}}$. Consider $\widetilde{c_i} = \sum_{j=1}^J(\widetilde{\lambda_G^{(j)}}+\widetilde{\lambda_L^{(j)}})\cdot\mathbf{1}_{\{X_{i,j}=N\}}$. Using new parameters, we obtain:

$$\int_0^\infty t\widetilde{M_i(t)}dt = \frac{1}{\mu^2}\left(\frac{1}{(\widetilde{c_i}+1)^2} + \sum_{k=1}^{|R_i|}\left(\sum_{N_k\subseteq R_i:|N_k|=k}(-1)^k\frac{1}{\left(\widetilde{c_i}+1+\sum_{\tau\in N_k}(\widetilde{\lambda_G^{(\tau)}}+\widetilde{\lambda_L^{(\tau)}})\right)^2}\right)\right),$$

$$\int_0^\infty \widetilde{M_i(t)}dt = \frac{1}{\mu}\left(\frac{1}{\widetilde{c_i}+1} + \sum_{k=1}^{|R_i|}\left(\sum_{N_k\subseteq R_i:|N_k|=k}(-1)^k\frac{1}{\left(\widetilde{c_i}+1+\sum_{\tau\in N_k}(\widetilde{\lambda_G^{(\tau)}}+\widetilde{\lambda_L^{(\tau)}})\right)}\right)\right).$$

If for the particular model it holds that estimates of $\widetilde{\lambda_G^{(j)}}$ and $\widetilde{\lambda_L^{(j)}}$ do not depend on μ for all $j\in\{1,\ldots,J\}$, then we have that $\hat{T}_i = \frac{\int t\widetilde{M_i(t)}dt}{\int\widetilde{M_i(t)}dt}$ depends on μ only by factor $\frac{1}{\mu}$. Therefore if we need time estimates only to sort samples, then the choice of μ is not important. As a result, for such models we choose $\mu = 1$.

3.2. Reconstructing the order of accumulating aberrations

Recall that the aim of the project is to reconstruct the order according to which particular aberrations in particular genes are accumulated within the time. For that purpose we can order samples according to time estimates. Let $\mathbf{X}_i$ for $i\in\{1,\ldots,I\}$ be all samples and $\hat{T}_1,\ldots,\hat{T}_I$ the corresponding time estimates. For $t\geq\min\{\hat{T}_i : i=1,\ldots,I\}$ and fixed region j we construct functions:

$$\text{profile}N_j(t) = \frac{\sum_{i=1}^I\mathbf{1}_{\{X_{i,j}=N,\hat{T}_i\leq t\}}}{\sum_{i=1}^I\mathbf{1}_{\{\hat{T}_i\leq t\}}},$$

$$\text{profile}G_j(t) = \frac{\sum_{i=1}^I\mathbf{1}_{\{X_{i,j}=G,\hat{T}_i\leq t\}}}{\sum_{i=1}^I\mathbf{1}_{\{\hat{T}_i\leq t\}}},$$

$$\text{profile}L_j(t) = \frac{\sum_{i=1}^{I} \mathbf{1}_{\{X_{i,j}=L, \hat{T}_i \leq t\}}}{\sum_{i=1}^{I} \mathbf{1}_{\{\hat{T}_i \leq t\}}}.$$

They describe the pattern, according to which aberrations at region j are accumulated during cancer progression. We are interested in those regions j for which there is a relevant increase in profile$G_j(t)$ or profile$L_j(t)$ when exceeding some particular time point t.

Chapter 4

Univariate models

4.1. Model's assumptions

First, a univariate model that describes the accumulation of chromosomal aberrations is built. This model is appropriate for one region. Suppose then that we fix the region j. According to our framework, we denote by $Z_j^{(t)}$ the state of the Markov chain for j th region at time t. We measure time from the beginning of the development of cancer. Recall that we assume $Z_j^{(0)} = N$ and $Z_j^{(t)} \in \{N, G, L\}$ for $t > 0$, where N, G, L denotes the number of DNA copies equal, bigger or smaller than 2 at region j. The chain starts in a state N and there are two transitions possible. It can go to the state G or L. G and L are absorbing states, therefore after reaching one of them, the chain will not leave it.

In this case the assumption about the distribution of the Markov chain gives the following probabilities:

$$\mathbb{P}(Z_j^{(t)} = N) = e^{-(\lambda_G^{(j)}+\lambda_L^{(j)})t},$$

$$\mathbb{P}(Z_j^{(t)} = G) = \frac{\lambda_G^{(j)}}{(\lambda_G^{(j)}+\lambda_L^{(j)})}\left(1 - e^{-(\lambda_G^{(j)}+\lambda_L^{(j)})t}\right),$$

$$\mathbb{P}(Z_j^{(t)} = L) = \frac{\lambda_L^{(j)}}{(\lambda_G^{(j)}+\lambda_L^{(j)})}\left(1 - e^{-(\lambda_G^{(j)}+\lambda_L^{(j)})t}\right).$$

4.2. The likelihood function

We want to obtain the expression for the likelihood function of the parameters. We consider the data. Suppose we have I samples and $X_{i,j}$ is the state of the j th region of i th sample. If we knew that the time of this sample was T_i, then the likelihood of parameters $(\lambda_G^{(j)}, \lambda_L^{(j)})$ would be

$$L(X_{i,j}; \lambda_G^{(j)}, \lambda_L^{(j)}) = \mathbb{P}(Z_j^{(T_i)} = N)^{\mathbf{1}_{\{X_{i,j}=N\}}} \cdot \mathbb{P}(Z_j^{(T_i)} = G)^{\mathbf{1}_{\{X_{i,j}=G\}}} \cdot \mathbb{P}(Z_j^{(T_i)} = L)^{\mathbf{1}_{\{X_{i,j}=L\}}}.$$

However the time is unobserved, therefore we use the prior distribution of T. Therefore from the law of total probability, the likelihood for j th region and i th sample is

$$L(X_{i,j}; \lambda_G^{(j)}, \lambda_L^{(j)}) =$$

$$= \int_0^\infty \mathbb{P}(Z_j^{(t)} = N)^{\mathbf{1}_{\{X_{i,j}=N\}}} \cdot \mathbb{P}(Z_j^{(t)} = G)^{\mathbf{1}_{\{X_{i,j}=G\}}} \cdot \mathbb{P}(Z_j^{(t)} = L)^{\mathbf{1}_{\{X_{i,j}=L\}}} \mu e^{-\mu t} dt.$$

Now using probabilities computed before one can see that

$$L(X_{i,j};\lambda_G^{(j)},\lambda_L^{(j)}) =$$

$$= \int_0^\infty \left(e^{-(\lambda_G^{(j)}+\lambda_L^{(j)})t}\right)^{\mathbf{1}_{\{X_{i,j}=N\}}} \cdot \frac{\left(\lambda_G^{(j)}\right)^{\mathbf{1}_{\{X_{i,j}=G\}}} \cdot \left(\lambda_L^{(j)}\right)^{\mathbf{1}_{\{X_{i,j}=L\}}}}{\left(\lambda_G^{(j)}+\lambda_L^{(j)}\right)^{\mathbf{1}_{\{X_{i,j}\neq N\}}}} \cdot$$

$$\cdot \left(1-e^{-\left(\lambda_G^{(j)}+\lambda_L^{(j)}\right)t}\right)^{\mathbf{1}_{\{X_{i,j}\neq N\}}} \mu e^{-\mu t}dt.$$

It is easily seen that the likelihood can be written as

$$L(X_{i,j};\lambda_G^{(j)},\lambda_L^{(j)}) =$$

$$= \left(\frac{\lambda_G^{(j)}}{\lambda_G^{(j)}+\lambda_L^{(j)}}\right)^{\mathbf{1}_{\{X_{i,j}=G\}}} \cdot \left(\frac{\lambda_L^{(j)}}{\lambda_G^{(j)}+\lambda_L^{(j)}}\right)^{\mathbf{1}_{\{X_{i,j}=L\}}} \cdot \mu \int_0^\infty \left(1-e^{-(\lambda_G^{(j)}+\lambda_L^{(j)})t}\right) e^{-\mu t}dt$$

if $X_{i,j} \neq N$, and

$$L(X_{i,j};\lambda_G^{(j)},\lambda_L^{(j)}) = \mu \int_0^\infty e^{-(\lambda_G^{(j)}+\lambda_L^{(j)}+\mu)t}dt$$

when $X_{i,j} = N$.

Let $A_j = \int_0^\infty \left(1-e^{-(\lambda_G^{(j)}+\lambda_L^{(j)})t}\right) e^{-\mu t}dt$ and $B_j = \int_0^\infty e^{-(\lambda_G^{(j)}+\lambda_L^{(j)}+\mu)t}dt$. Then we have that $A_j = \int_0^\infty e^{-\mu t}dt - \int_0^\infty e^{-(\lambda_G^{(j)}+\lambda_L^{(j)}+\mu)t}dt$. By elementary calculus we obtain that

$$A_j = \frac{\lambda_G^{(j)}+\lambda_L^{(j)}}{\mu(\lambda_G^{(j)}+\lambda_L^{(j)}+\mu)},$$

$$B_j = \frac{1}{\lambda_G^{(j)}+\lambda_L^{(j)}+\mu}.$$

Then it holds that

$$L(X_{i,j};\lambda_G^{(j)},\lambda_L^{(j)}) =$$

$$= \left(\frac{\lambda_G^{(j)}}{\lambda_G^{(j)}+\lambda_L^{(j)}}\right)^{\mathbf{1}_{\{X_{i,j}=G\}}} \cdot \left(\frac{\lambda_L^{(j)}}{\lambda_G^{(j)}+\lambda_L^{(j)}}\right)^{\mathbf{1}_{\{X_{i,j}=L\}}} \cdot \mu \cdot A_j^{1-\mathbf{1}_{\{X_{i,j}=N\}}} \cdot B_j^{\mathbf{1}_{\{X_{i,j}=N\}}}.$$

We assume that samples are independent. Thus the likelihood for j th region and I samples is the product of likelihoods of each sample:

$$L(X_{i,j} : i = 1,\ldots,I;\lambda_G^{(j)},\lambda_L^{(j)}) = \prod_{i=1}^{I} L(X_{i,j};\lambda_G^{(j)},\lambda_L^{(j)}).$$

Let

$$n_j = \sum_{i=1}^{I} \mathbf{1}_{\{X_{i,j}=N\}},$$

$$l_j = \sum_{i=1}^{I} \mathbf{1}_{\{X_{i,j}=L\}},$$

$$g_j = \sum_{i=1}^{I} \mathbf{1}_{\{X_{i,j}=G\}}.$$

Note that we keep the convention that $0^0 = 1$ if n_j, g_j, l_j is involved in the exponent. Therefore, by what has been shown before, we have that

$$L(X_{i,j} : i = 1, \ldots, I; \lambda_G^{(j)}, \lambda_L^{(j)}) =$$

$$= \left(\frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}}\right)^{g_j} \left(\frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}}\right)^{l_j} \mu^{n_j+g_j+l_j} A_j^{g_j+l_j} B_j^{n_j} =$$

$$= \frac{\mu^{n_j} \cdot \left(\lambda_G^{(j)}\right)^{g_j} \cdot \left(\lambda_L^{(j)}\right)^{l_j}}{\left(\mu + \lambda_G^{(j)} + \lambda_L^{(j)}\right)^{n_j+g_j+l_j}}.$$

4.3. Finding maximum likelihood estimators

The aim is to estimate parameters $\lambda_G^{(j)}$ and $\lambda_L^{(j)}$ by maximizing the likelihood function. Thus we look for $(\lambda_G^{(j)}, \lambda_L^{(j)})$ which maximizes $L(X_{i,j} : i = 1, \ldots, I; \lambda_G^{(j)}, \lambda_L^{(j)})$.

Now we use the above expression for the maximum likelihood function, to obtain estimators of parameters for some special cases. If $g_j = 0$ and $n_j > 0$ then the likelihood function is $\frac{\mu^{n_j} \cdot (\lambda_L^{(j)})^{l_j}}{\left(\mu+\lambda_G^{(j)}+\lambda_L^{(j)}\right)^{n_j+l_j}}$, so it is obvious that $\hat{\lambda}_G^{(j)} = 0$. By computing the derivative of the likelihood function with respect to $\lambda_L^{(j)}$ and solving the equation $\frac{\partial}{\partial \lambda_L^{(j)}} L(X_{i,j} : i = 1, \ldots, I; \hat{\lambda}_G^{(j)}, \lambda_L^{(j)})$ it can be shown that $\hat{\lambda}_L^{(j)} = \frac{l_j \mu}{n_j}$. Similarly if $l_j = 0$ and $n_j > 0$ then $(\hat{\lambda}_G^{(j)}, \hat{\lambda}_L^{(j)}) = (\frac{g_j \mu}{n_j}, 0)$. If $g_j, l_j = 0$, then in particular $n_j > 0$ and using the previous two cases, it holds that $(\hat{\lambda}_G^{(j)}, \hat{\lambda}_L^{(j)}) = (0, 0)$.

We will also obtain estimators in case if $g_j, n_j, l_j > 0$. The partial derivatives are given by:

$$\frac{\partial}{\partial \lambda_G^{(j)}} L(X_{i,j} : i = 1, \cdots, I; \lambda_G^{(j)}, \lambda_L^{(j)}) =$$

$$= \frac{\mu^{n_j} \left(\lambda_L^{(j)}\right)^{l_j} \left(\lambda_G^{(j)}\right)^{g_j-1} \left[l_j \left(\mu + \lambda_G^{(j)} + \lambda_L^{(j)}\right) - (n_j + g_j + l_j)\lambda_L^{(j)}\right]}{\left(\mu + \lambda_G^{(j)} + \lambda_L^{(j)}\right)^{n_j+g_j+l_j+1}}$$

and

$$\frac{\partial}{\partial \lambda_L^{(j)}} L(X_{i,j} : i = 1, \cdots, I; \lambda_G^{(j)}, \lambda_L^{(j)}) =$$

$$= \frac{\mu^{n_j} \left(\lambda_G^{(j)}\right)^{g_j} \left(\lambda_L^{(j)}\right)^{l_j-1} \left[g_j \left(\mu + \lambda_G^{(j)} + \lambda_L^{(j)}\right) - (n_j + g_j + l_j)\lambda_G^{(j)}\right]}{\left(\mu + \lambda_G^{(j)} + \lambda_L^{(j)}\right)^{n_j+g_j+l_j+1}}.$$

Because on $\mathbb{R}_+ \times \mathbb{R}_+$ these partial derivatives are continuous, then the gradient exists and it consists of the above partial derivatives.

It is easily seen that for $\min\{\lambda_G^{(j)}, \lambda_L^{(j)}\} = 0$ $L(X_{i,j} : i = 1, \ldots, I; \lambda_G^{(j)}, \lambda_L^{(j)}) = 0$. For $\max\{\lambda_G^{(j)}, \lambda_L^{(j)}\}$ tending to ∞ $L(X_{i,j} : i = 1, \ldots, I; \lambda_G^{(j)}, \lambda_L^{(j)})$ tends to 0. It also holds that

$L(X_{i,j} : i = 1, \ldots, I; \lambda_G^{(j)}, \lambda_L^{(j)}) \geq 0$ for all $(\lambda_G^{(j)}, \lambda_L^{(j)}) \in \mathbb{R}_+ \times \mathbb{R}_+$, therefore the likelihood function has a maximum. Additionally, for $(\lambda_G^{(j)}, \lambda_L^{(j)})$ maximizer of $L(X_{i,j} : i = 1, \ldots, I; \cdot)$ it holds that

$$\frac{\partial\ L(X_{i,j} : i = 1, \ldots, I; \lambda_G^{(j)}, \lambda_L^{(j)})}{\partial \lambda_G^{(j)}} = 0,$$

and

$$\frac{\partial\ L(X_{i,j} : i = 1, \ldots, I; \lambda_G^{(j)}, \lambda_L^{(j)})}{\partial \lambda_L^{(j)}} = 0.$$

It gives two equations to be solved:

$$g_j(\mu + \lambda_G^{(j)} + \lambda_L^{(j)}) - (n_j + g_j + l_j)\lambda_G^{(j)} = 0$$

$$l_j(\mu + \lambda_G^{(j)} + \lambda_L^{(j)}) - (n_j + g_j + l_j)\lambda_L^{(j)} = 0$$

The solution is

$$\lambda_G^{(j)} = \frac{g_j \mu}{n_j},$$

$$\lambda_L^{(j)} = \frac{l_j \mu}{n_j}.$$

The same expressions hold if $g_j = 0$ or $l_j = 0$. Thus if $n_j > 0$, we obtain estimators of $(\lambda_G^{(j)}, \lambda_L^{(j)})$:

$$\hat{\lambda}_G^{(j)} = \frac{g_j \mu}{n_j},$$

$$\hat{\lambda}_L^{(j)} = \frac{l_j \mu}{n_j}.$$

The last case, which was not analysed yet is $n_j = 0$. For this case the likelihood function does not reach the maximum, therefore the parameters estimators do not exist. There is an interpretation of this fact. Recall that parameter $\lambda_G^{(j)} + \lambda_L^{(j)}$ indicates how rapidly the transition from a state N to one from G and L at region j occurs. If the data set consists of samples that have aberration at region j, then we are not able to estimate the speed in which the transition at region j occurs.

However, for a reasonable number of samples, the situation that $n_j = 0$ is very unlikely to happen.

4.3.1. Removing the dependence on μ

Because we obtain estimators:

$$\hat{\lambda}_G^{(j)} = \frac{g_j \mu}{n_j},$$

$$\hat{\lambda}_L^{(j)} = \frac{l_j \mu}{n_j},$$

then in this case it holds that estimators of $\widetilde{\lambda_G^{(j)}}$ and $\widetilde{\lambda_L^{(j)}}$ do not depend on μ. Therefore, according to the analysis from the previous chapter, in this model we can assume that $\mu = 1$.

4.4. Evaluation

In order to evaluate this model, we used publicly available DNA copy number data. This is the data of breast cancer patients. The data consists of 263 samples of 1704 genomic regions. We fitted a univariate model for each DNA region separately. Thus we obtained $(\hat{\lambda}_G^{(j)}, \hat{\lambda}_L^{(j)})$ for $j = 1, \ldots, J$.

4.4.1. Graphics

Several plots were made to summarize those models.

Figure 4.1: λ_G and λ_L for different regions

Figure 4.1 shows that most of λ_G and λ_L are close to zero. There are several outliers. As one can see from the plots, it is very common that for regions which lie next to each other on the plot, we observe similar estimates of the parameters. It can support the thesis that there are strong correlations between features that are situated close to each other on the genome.

Figure 4.2 and 4.3 show that for most regions the parameters are very small. There are only few regions for which at least one parameter's estimate stands out.

Figure 4.2: Histogram of λ_G for all regions

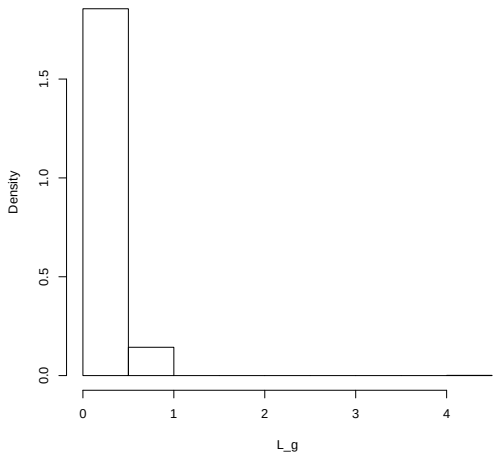


Figure 4.3: Histogram of λ_L for all regions

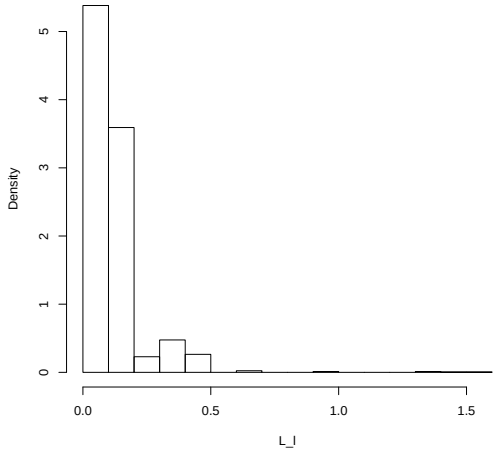
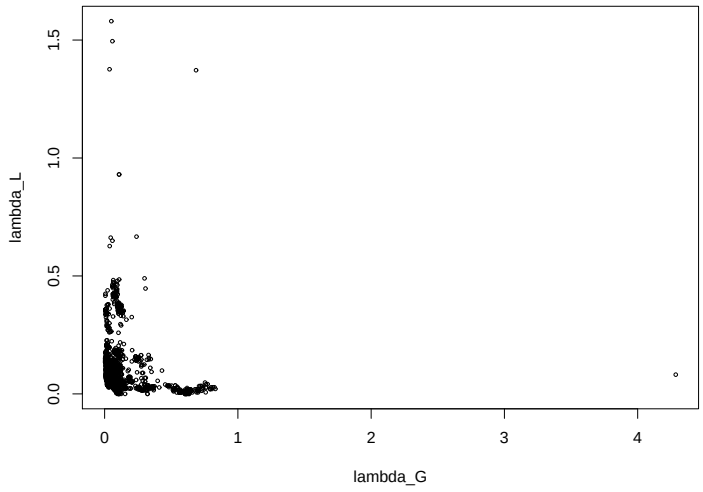


Figure 4.4: Relation between λ_G and λ_L

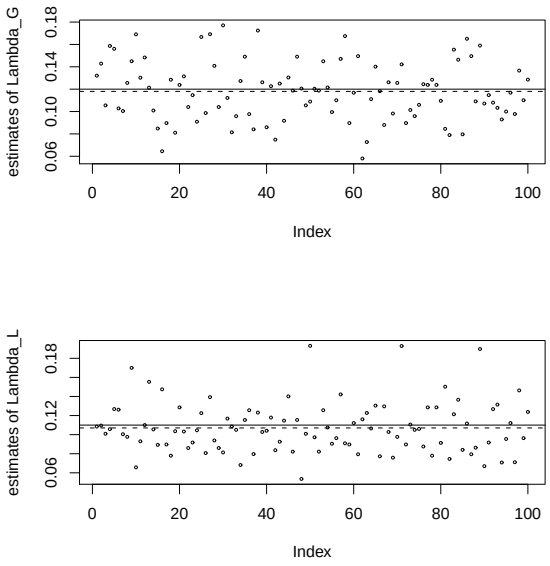


From the Figure 4.4 one can see that for most of regions both $\hat{\lambda}_G^{(j)}$ and $\hat{\lambda}_L^{(j)}$ are close to zero. Also, usually when one of $\hat{\lambda}_G^{(j)}, \hat{\lambda}_L^{(j)}$ is big, then the other is close to zero.

4.4.2. Checking the accuracy of estimators

We decided to use the data simulation technique to check the accuracy of the estimators of the univariate model's parameters. For simulating the data, we need to choose values of the parameters λ_G and λ_L. We took $\lambda_G = 0.12$ and $\lambda_L = 0.11$. We simulated the data of 263 samples from the state distribution, with $(\lambda_G, \lambda_L) = (0.12, 0.11)$, and computed parameters' estimates. We repeated it 100 times. Figure 4.5 shows the results of this procedure. Horizontal continuous lines are the values of a real parameter, dashed lines are the averages of parameters' estimates. It can be inferred, that the estimates of the parameters are consistent and the parameters' estimators for the univariate model have high accuracy.

Figure 4.5: Estimates of parameters for the simulated data



Chapter 5

Simple multivariate model

5.1. Motivation

In this chapter we build a simple multivariate model, which takes all regions into account. The aim of this model is to describe the average behaviour of all regions.

5.2. Model's assumptions and probabilities

For the multivariate case, the state of the chain at time t is $\mathbf{Z}^{(t)} = (Z_1^{(t)}, \ldots, Z_J^{(t)})$. We repeat the assumption that $\mathbf{Z}^{(0)} = (N, \ldots, N)$ and $Z_j^{(t)} \in \{N, G, L\}$ for $t > 0$. So the state of the system is a vector containing copy number information about all regions. We assume that for any fixed time t regions behave independently, thus $Z_1^{(t)}, \ldots, Z_J^{(t)}$ are independent random variables. It was stated that for each region j there are two parameters describing the region: $\lambda_G^{(j)}$ and $\lambda_L^{(j)}$. In this model we add the assumption that $\lambda_G^{(1)} = \ldots = \lambda_G^{(J)}$ and $\lambda_L^{(1)} = \ldots = \lambda_L^{(J)}$, therefore there are two parameters that describe the behaviour of all regions. Thus we can skip the dependence on the region in parameters and consider λ_G and λ_L. We still assume that $\lambda_G, \lambda_L \geq 0$.

Recall that the chain starts at $(N, \ldots, N)$. Now consider $Z_j^{(t)}$ a coefficient of $\mathbf{Z}^{(t)}$. $Z_j^{(t)}$ is a univariate Markov chain. We have that $Z_j^{(0)} = N$ and Z_j can leave the state N and go either to G or L. After one of this transitions happens, the state of the chain $Z_j^{(t)}$ will not change anymore. Therefore for any coefficient of $\mathbf{Z}^{(t)}$ we have two possible transitions to absorbing states G and L.

The multivariate model is also a Markov chain. From the general assumptions we have that:

$$\mathbb{P}(Z_j^{(t)} = N) = e^{-(\lambda_G + \lambda_L)t},$$

$$\mathbb{P}(Z_j^{(t)} = G) = \frac{\lambda_G}{\lambda_G + \lambda_L}\left(1 - e^{-(\lambda_G + \lambda_L)t}\right),$$

$$\mathbb{P}(Z_j^{(t)} = L) = \frac{\lambda_L}{\lambda_G + \lambda_L}\left(1 - e^{-(\lambda_G + \lambda_L)t}\right).$$

Therefore the above expressions indicate marginal state distribution at time t. Now we need to obtain multivariate state distribution. Because for any time t we assume independence

between regions, then the multivariate state distribution is a product of marginal distributions. Thus for any $(u_1, \ldots, u_J)$ in the state space we have that

$$\mathbb{P}(\mathbf{Z}^{(t)} = (u_1, \ldots, u_J)) =$$

$$= \prod_{j=1}^{J} \left(\mathbb{P}(Z_j^{(t)} = G)^{\mathbf{1}_{\{u_j = G\}}} \mathbb{P}(Z_j^{(t)} = L)^{\mathbf{1}_{\{u_j = L\}}} \mathbb{P}(Z_j^{(t)} = N)^{\mathbf{1}_{\{u_j = N\}}} \right).$$

Using expressions for probabilities for one region, we obtain:

$$\mathbb{P}(\mathbf{Z}^{(t)} = (u_1, \ldots, u_J)) =$$

$$= \left(\frac{\lambda_G}{\lambda_G + \lambda_L} \right)^g \left(\frac{\lambda_L}{\lambda_G + \lambda_L} \right)^l \left(1 - e^{-(\lambda_G + \lambda_L)t} \right)^{g+l} e^{-n(\lambda_G + \lambda_L)t},$$

where

$$n = \sum_{j=1}^{J} \mathbf{1}_{\{u_j = N\}},$$

$$g = \sum_{j=1}^{J} \mathbf{1}_{\{u_j = G\}},$$

$$l = \sum_{j=1}^{J} \mathbf{1}_{\{u_j = L\}}.$$

5.3. The likelihood function

Now we consider the data. Denote by $X_{i,j}$ the DNA copy number of the j th region of i th sample. Thus we have $X_{i,j} \in \{N, G, L\}$. By $\mathbf{X}_i$ we denote the information about DNA copy number for all regions of i th sample, so $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,J})$. We state the same assumption about prior distribution of time, so $T \sim Exp(\mu)$. If the data contained the information about time, and T_i was time for i th sample, then $\mathbf{X}_i$ would be the sample from the distribution of $\mathbf{Z}^{(T_i)}$. Using the law of total probability, the state distribution of the Markov chain at time t, the assumed relation between the data and Markov chain we obtain the likelihood function for i th sample.

$$L(\mathbf{X}_i; \lambda_G, \lambda_L) =$$

$$= \int_0^\infty \left(\frac{\lambda_G}{\lambda_G + \lambda_L} \right)^{g^{(i)}} \left(\frac{\lambda_L}{\lambda_G + \lambda_L} \right)^{l^{(i)}} \left(1 - e^{-(\lambda_G + \lambda_L)t} \right)^{g^{(i)} + l^{(i)}} e^{-n^{(i)}(\lambda_G + \lambda_L)t} \mu e^{-\mu t} dt,$$

where

$$n^{(i)} = \sum_{j=1}^{J} \mathbf{1}_{\{X_{i,j} = N\}},$$

$$g^{(i)} = \sum_{j=1}^{J} \mathbf{1}_{\{X_{i,j} = G\}},$$

$$l^{(i)} = \sum_{j=1}^{J} \mathbf{1}_{\{X_{i,j} = L\}}.$$

Then, because of the independence between samples, the likelihood for I samples, $L(\mathbf{X}_i : i = 1, \ldots, I)$, will be:

$$\prod_{i=1}^{I} \int_0^\infty \left(\frac{\lambda_G}{\lambda_G + \lambda_L}\right)^{g^{(i)}} \left(\frac{\lambda_L}{\lambda_G + \lambda_L}\right)^{l^{(i)}} \left(1 - e^{-(\lambda_G+\lambda_L)t}\right)^{g^{(i)}+l^{(i)}} e^{-n^{(i)}(\lambda_G+\lambda_L)t} \mu e^{-\mu t} dt,$$

where $n^{(i)}, g^{(i)}, l^{(i)}$ are defined above.

5.4. Finding maximum likelihood estimators

We are interested in finding maximum likelihood estimators of parameters λ_G and λ_L. In this case it is convenient to change variables and consider $p = \frac{\lambda_L}{\lambda_G+\lambda_L}$ and $s = \lambda_G + \lambda_L$. For each pair $(\lambda_G, \lambda_L) \in \mathbb{R}_+ \times \mathbb{R}_+$ there is exactly one pair (p, s) such that $p, s \geq 0$ and $p \leq 1$. Note that $\lambda_L = ps$ and $\lambda_G = s(1-p)$. Therefore by finding optimal (p, s), we can also obtain optimal (λ_G, λ_L).

We change a variable with respect to which we integrate. We take $u = st$. Therefore the likelihood function for new parameters will be

$$L(\mathbf{X}_i : i = 1, \ldots, I; p, s) = \prod_{i=1}^{I} (1-p)^{g^{(i)}} p^{l^{(i)}} \int_0^\infty \frac{\mu}{s} \left(1 - e^{-u}\right)^{g^{(i)}+l^{(i)}} e^{-n^{(i)}u} e^{-\frac{\mu u}{s}} du.$$

We can write it as

$$(1-p)^{\sum_{i=1}^{I} g^{(i)}} p^{\sum_{i=1}^{I} l^{(i)}} \prod_{i=1}^{I} \int_0^\infty \frac{\mu}{s} \left(1 - e^{-u}\right)^{g^{(i)}+l^{(i)}} e^{-n^{(i)}u} e^{-\frac{\mu u}{s}} du.$$

This expression is very useful, because it shows that the likelihood function is a product of two functions:

$$f_1(\mathbf{X}_i : i = 1, \ldots, I; p) = (1-p)^{\sum_{i=1}^{I} g^{(i)}} p^{\sum_{i=1}^{I} l^{(i)}},$$

$$f_2(\mathbf{X}_i : i = 1, \ldots, I; s) = \prod_{i=1}^{I} \int_0^\infty \frac{\mu}{s} \left(1 - e^{-u}\right)^{g^{(i)}+l^{(i)}} e^{-n^{(i)}u} e^{-\frac{\mu u}{s}} du.$$

f_1 depends only on the parameter p and the data, f_2 depends only on s and the data. Additionally, f_1 and f_2 are non-negative. Therefore maximizing the likelihood function with respect to parameters p and s is equivalent to maximizing the function f_1 with respect to p and f_2 with respect to s.

We will start with maximizing f_1 with respect to parameter p. Therefore we look for p such that $\frac{\partial}{\partial p} f_1(\mathbf{X}_i : i = 1, \ldots, I; p) = 0$. The derivative of f_1 with respect to p is:

$$\frac{\partial f_1(\mathbf{X}_i : i = 1, \ldots, I; p)}{\partial p} =$$

$$= (1-p)^{\sum_{i=1}^{I} g^{(i)} - 1} p^{\sum_{i=1}^{I} l^{(i)} - 1} \left(-p \sum_{i=1}^{I} g^{(i)} + (1-p) \sum_{i=1}^{I} l^{(i)}\right).$$

If $p \in (0, 1)$ and $\frac{\partial}{\partial p} L(\mathbf{X}_i : i = 1, \ldots, I; p, s) = 0$, then

$$-p \sum_{i=1}^{I} g^{(i)} + (1-p) \sum_{i=1}^{I} l^{(i)} = 0,$$

which implies that

$$\hat{p} = \frac{\sum_{i=1}^{I} l^{(i)}}{\sum_{i=1}^{I} g^{(i)} + \sum_{i=1}^{I} l^{(i)}}.$$

The same expression holds for the case if $\sum_{i=1}^{I} l^{(i)} = 0$ or $\sum_{i=1}^{I} g^{(i)} = 0$. There is no maximizer of f_1 in the case if $\sum_{i=1}^{I} \left(l^{(i)} + g^{(i)}\right) = 0$. But in this case we would observe normal DNA copy number in all samples and regions. This is very unlikely to happen for data of cancer patients with a reasonable number of samples and regions. Therefore for any reasonable data we know which p is optimal.

Therefore we only need to find $\hat{s}$ such that $\hat{s}$ maximizes the function f_2. Because parameter μ also appears in the function f_2, then we need to take it into account. Suppose we take $\mu = 1$ and find a maximizer $\hat{s} = c^*$ of the function f_2. Then if we want to find $\hat{s}$ for arbitrary $\mu > 0$, then from the expression of $f_2\left(\mathbf{X}_i : i = 1, \ldots, I; s\right)$ we can conclude that $\hat{s} = c^*\mu$. Therefore we can focus on finding $\hat{s}$ when $\mu = 1$. Therefore from this point in our computations we assume $\mu = 1$. But in the results we will obtain a method of computing $\hat{s}$ for any $\mu > 0$.

In order to obtain $\hat{s}$ we computed $\frac{\partial f_2(\mathbf{X}_i : i=1,\ldots,I;s)}{\partial s}$ and $\frac{\partial \ln f_2(\mathbf{X}_i : i=1,\ldots,I;s)}{\partial s}$ and tried to solve $\frac{\partial f_2(\mathbf{X}_i : i=1,\ldots,I;s)}{\partial s} = 0$ or $\frac{\partial \ln f_2(\mathbf{X}_i : i=1,\ldots,I;s)}{\partial s} = 0$. But both $\frac{\partial f_2(\mathbf{X}_i : i=1,\ldots,I;s)}{\partial s}$ and $\frac{\partial \ln f_2(\mathbf{X}_i : i=1,\ldots,I;s)}{\partial s}$ are very complex, thus it is impossible to find direct solution. Therefore we had to apply a numerical maximization procedure to find $\hat{s}$.

For applying numerical optimization procedures, we need to have a numerical representation of a function we want to maximize. Note that there are integrals in the function f_2. It is possible to compute these integrals by rewriting $(1 - e^{-u})^{g^{(i)}+l^{(i)}}$ as a sum, using binomial theorem. Then we obtain

$$\int_0^\infty \frac{1}{s} \sum_{k=0}^{g^{(i)}+l^{(i)}} \binom{g^{(i)} + l^{(i)}}{k} (-1)^k e^{-\left(k+n^{(i)}+\frac{1}{s}\right)u} du =$$

$$= \sum_{k=0}^{g^{(i)}+l^{(i)}} \binom{g^{(i)} + l^{(i)}}{k} (-1)^k \frac{1}{\left(k + n^{(i)}\right) s + 1}.$$

Here we assume that $\binom{n}{0} = 1$ for any $n \in \mathbb{N}$. Then it holds that

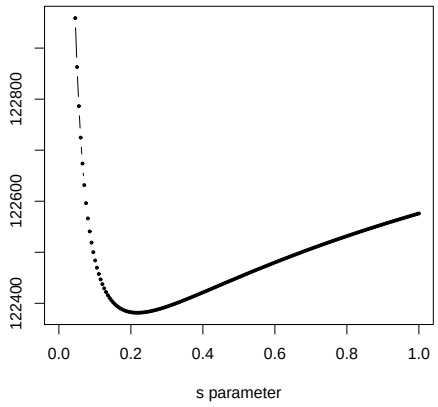
$$f_2\left(\mathbf{X}_i : i = 1, \ldots, I; s\right) = \prod_{i=1}^{I} \sum_{k=0}^{g^{(i)}+l^{(i)}} \binom{g^{(i)} + l^{(i)}}{k} (-1)^k \frac{1}{\left(k + n^{(i)}\right) s + 1}$$

Unfortunately we cannot use this representation of the function f_2 for numerical optimization. The reason is that the part $\sum_{k=0}^{g^{(i)}+l^{(i)}} \binom{g^{(i)}+l^{(i)}}{k} (-1)^k \frac{1}{\left(k+n^{(i)}\right)s+1}$ has bad numerical properties. Therefore we will use a quadrature for representing integrals in the expression $f_2\left(\mathbf{X}_i : i = 1, \ldots, I; s\right)$. Note that

$$\int_0^\infty \frac{1}{s} \left(1 - e^{-u}\right)^{g^{(i)}+l^{(i)}} e^{-n^{(i)}u} e^{-\frac{u}{s}} du = \int_0^\infty h_i(u) \cdot w(u) du,$$

where $h_i(u) = \frac{1}{s} \left(1 - e^{-u}\right)^{g^{(i)}+l^{(i)}} e^{-\left(n^{(i)}-1\right)u} e^{-\frac{u}{s}}$ and $w(u) = e^{-u}$. Therefore, we will use the Laguerre quadrature. It means that we will approximate $\int_0^\infty h_i(u) \cdot w(u) du$ by $\sum_{k=1}^{K} h_i(u_k) \cdot w_k$, where u_k are nodes and w_k weights of the Laguerre quadrature, K is the number of nodes we use in the quadrature.

Figure 5.1: minus logarithm of the representation of the function f_2



With the help of the Laguerre quadrature, we obtained a good numerical representation of the likelihood function. Note that maximizing the function f_2 is equivalent to minimizing the minus logarithm of the function f_2. We plotted the numerical representation of minus logarithm f_2 for the breast cancer data of 263 samples and 1704 regions.

From the Figure 5.1 we can conclude that this representation is smooth and the function $-\log f_2$ has indeed an optimum. Those are good prognostics for using the numerical optimization algorithm.

We applied optimization method which is a combination of golden section search and successive parabolic interpolation.

5.4.1. Checking the accuracy of the estimator of parameters p and s

We decided to check if the optimization procedure for finding estimate of the parameter s works well. At the same time we will check the accuracy of the estimator of the parameter p. For this purpose we choose real values of the parameters $(\lambda_G, \lambda_L) = (1, \frac{1}{2})$. Therefore $(p, s) = (\frac{1}{3}, \frac{3}{2})$.

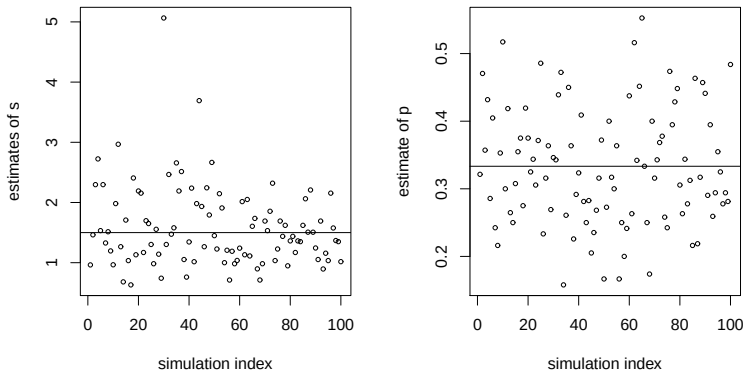
We simulated 100 times the data of 263 samples and 1704 regions with the real parameters as mentioned above.

The Figure 5.2 shows that estimated values are in principle spread around the real value of the parameter s. There are possible two outliers

For the estimator of the parameter p we observe that estimates are spread around the real value of p and it does not seem that there are any outliers. One can say that the variance of the estimator of p is much smaller than of estimator of s.

But it is good to realize that the parameter p can take values only from the unit interval, therefore the variance of the estimator of the parameter p cannot be bigger than $\frac{1}{4}$. We will prove this fact. Let Y be the random variable such that $\mathbb{P}(Y \in [0,1]) = 1$. First, we will show that $\mathbb{E}Y \in [0,1]$. The definition of the expectation is $\mathbb{E}Y = \int_\Omega Y d\mathbb{P}$. Because Y with probability 1 takes only non-negative values and $\mathbb{P}$ is a probabilistic measure, then $\int_\Omega Y d\mathbb{P} \geq \int_\Omega 0 d\mathbb{P} = 0$. Therefore $\mathbb{E}(Y) \geq 0$. Similarly, because $\mathbb{P}(Y \leq 1) = 1$, then $\int_\Omega Y d\mathbb{P} \leq \int_\Omega 1 d\mathbb{P} = 1$. So $\mathbb{E}Y \leq 1$. It shows that $\mathbb{E}Y \in [0,1]$.

Figure 5.2: Estimates of parameters s and p while $(p, s) = (\frac{1}{3}, \frac{3}{2})$



Now consider $Var(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2$. Because $\mathbb{P}(Y \leq 1) = 1$ and $\mathbb{P}(Y \geq 0) = 1$, then $\mathbb{E}Y^2 \leq \mathbb{E}(Y \cdot 1) = \mathbb{E}Y$. Therefore it holds that $Var(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 \leq \mathbb{E}Y - (\mathbb{E}Y)^2 = \mathbb{E}Y(1 - \mathbb{E}Y)$. The function $\mathbb{E}Y(1 - \mathbb{E}Y)$ takes its maximum if $\mathbb{E}Y = \frac{1}{2}$. The maximal value is then $\frac{1}{4}$. Therefore we have that $Var(Y) \leq \mathbb{E}Y(1 - \mathbb{E}Y) \leq \frac{1}{4}$.

Because the s parameter can take any positive value, we do not have the upper bound for the variance of estimator for s parameter.

5.5. Time estimation

In this section we apply time estimation for this model. Recall that $\hat{T}_i = \frac{\int_0^\infty t\widetilde{M_i(t)}dt}{\int_0^\infty \widetilde{M_i(t)}dt}$. Therefore we need to compute $\int_0^\infty t\widetilde{M_i(t)}dt$ and $\int_0^\infty \widetilde{M_i(t)}dt$. For this model we have $c_i = \sum_{j=1}^J (\lambda_G^{(j)} + \lambda_L^{(j)}) \cdot \mathbf{1}_{\{X_{i,j}=N\}} = \sum_{j=1}^J (\lambda_G + \lambda_L) \cdot \mathbf{1}_{\{X_{i,j}=N\}} = (\lambda_G + \lambda_L) \sum_{j=1}^J \mathbf{1}_{\{X_{i,j}=N\}} = (\lambda_G + \lambda_L) n^{(i)} = n^{(i)} \hat{s} = n^{(i)} c^* \mu$. It also holds that

$$\int_0^\infty \widetilde{tM_i(t)}dt = \frac{1}{\mu^2 \left(n^{(i)} + 1\right)^2} + \sum_{k=1}^{|R_i|} (-1)^k \binom{|R_i|}{k} \frac{1}{(c_i + \mu + k\hat{s})^2} =$$

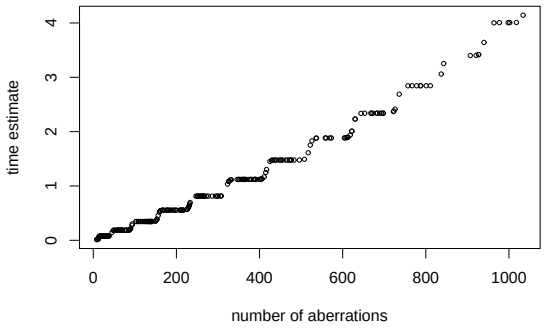
$$= \frac{1}{\mu^2 \left(n^{(i)} + 1\right)^2} + \sum_{k=1}^{|R_i|} (-1)^k \binom{|R_i|}{k} \frac{1}{\left(n^{(i)} c^* \mu + \mu + k c^* \mu\right)^2} =$$

$$= \frac{1}{\mu^2} \left(\frac{1}{\left(n^{(i)} + 1\right)^2} + \sum_{k=1}^{|R_i|} (-1)^k \binom{|R_i|}{k} \frac{1}{\left(n^{(i)} c^* + 1 + k c^*\right)^2} \right).$$

Similarly it can be shown that

$$\int_0^\infty \widetilde{M_i(t)}dt = \frac{1}{\mu} \left(\frac{1}{n^{(i)} + 1} + \sum_{k=1}^{|R_i|} (-1)^k \binom{|R_i|}{k} \frac{1}{n^{(i)} c^* + 1 + k c^*} \right).$$

Figure 5.3: Time estimates for different number of aberrations



Therefore the time estimate for the i th sample is

$$\hat{T}_i = \frac{\int_0^\infty t\widetilde{M_i(t)}dt}{\int_0^\infty \widetilde{M_i(t)}dt} = \frac{1}{\mu} \cdot \left(\frac{1}{\left(n^{(i)}+1\right)^2} + \sum_{k=1}^{|R_i|} (-1)^k \binom{|R_i|}{k} \frac{1}{\left(n^{(i)}c^* + 1 + kc^*\right)^2} \right) \cdot$$

$$\cdot \left(\frac{1}{n^{(i)}+1} + \sum_{k=1}^{|R_i|} (-1)^k \binom{|R_i|}{k} \frac{1}{n^{(i)}c^* + 1 + kc^*} \right)^{-1}.$$

Note that $\hat{T}_i$ depends on μ only by a factor $\frac{1}{\mu}$. Therefore, as in the previous model, we can assume that $\mu = 1$.

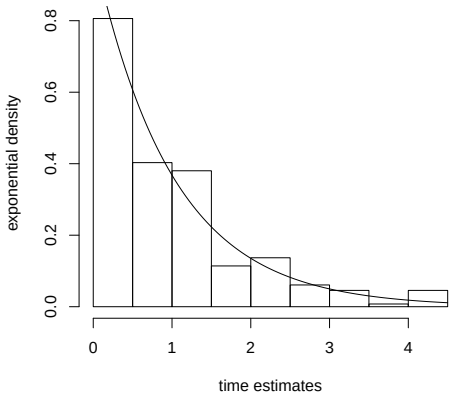
It must be stated that we did not use analytical expressions for computing time estimates. This is because of bad numerical properties of the expressions: $\sum_{k=1}^{|R_i|} (-1)^k \binom{|R_i|}{k} \frac{1}{n^{(i)}c^*+1+kc^*}$. Instead of using analytical expressions, we used Laguerre quadrature to replace integrals in $\int_0^\infty t\widetilde{M_i(t)}dt$ and $\int_0^\infty \widetilde{M_i(t)}dt$ by sums.

5.5.1. Graphics for time estimation

We obtained time estimates for samples of breast cancer data. Figure 5.3 indicates that an estimate of time is a non-decreasing function of the number of aberrations among all regions.

Figure 5.4 shows how the time estimates are spread. The histogram resembles density of the prior assumption of time distribution.

Figure 5.4: Histogram of time estimates and the prior time density



Chapter 6

Final multivariate model

6.1. Motivation

In this step we build the full multivariate model of cancer progression. We build this model for two possible applications.

One is to have a multivariate model that will be used for low dimensional case. If one has a small number of important DNA regions selected from all regions, it can be useful to fit the full dimensional model to the low dimensional data.

The model we build in this chapter can also be used for the high dimensional data. In this case the idea behind is to divide regions into two categories. In the first category there will be regions which behaviour is not particular, so those regions do not stand out from the set of all regions. The group of regions in this category will be described by only two parameters. Therefore the number of parameters needed to describe the behaviour of all regions in the category one is 2. To the second category belong regions that stand out from the set of all regions. Therefore each region from the second category will be described by two parameters. Thus number of parameters needed to describe the behaviour of all regions from the category two is $2 \times$ number of regions in the second category. Therefore in this case we consider a sparse model.

6.2. Model's assumptions

Now we state clear assumptions of this model. The model we are building is a multivariate continuous time Markov chain. The state of the chain at time t is $\mathbf{Z}^{(t)} = (Z_1^{(t)}, \ldots, Z_J^{(t)})$. The start is at $\mathbf{Z}^{(0)} = (N, \ldots, N)$, which is a state of the healthy cell. We assume that the time needed to leave N at region j is exponentially distributed with parameter $\lambda_G^{(j)} + \lambda_L^{(j)}$. After the chain leaves N at region j, it will not be possible to change the state at region j anymore. Therefore at each region j, there will be one transition from a state N either to G or L. We assume that for any fixed time t regions behave independently. We have the same expressions of marginal probabilities as in the previous model:

$$\mathbb{P}(Z_j^{(t)} = N) = e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t},$$

$$\mathbb{P}(Z_j^{(t)} = G) = \frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \cdot \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}\right),$$

$$\mathbb{P}(Z_j^{(t)} = L) = \frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_L^{(j)}} \cdot \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t}\right).$$

Because of the assumption about independence between regions at any time t, the state probability for any state $(u_1, \ldots, u_J)$ at time t is described by:

$$\mathbb{P}(\mathbf{Z}^{(t)} = (u_1, \ldots, u_J)) = \prod_{j=1}^{J} \left(\mathbb{P}(Z_j^{(t)} = u_j) \right).$$

6.3. The likelihood function

We want to obtain parameters estimators of the Markov chain. For this we use our data. We hold the same notation for the data as in the previous models, so $X_{i,j}$ is the DNA copy number of i th sample and j th region. $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,J})$ is the vector describing the i th sample with respect to DNA copy number. We state the same assumption about the prior distribution of time, so $T \sim Exp(\mu)$, $f_T(t) = \mu e^{-\mu t}$ is a density of prior time distribution at time t. The same as in the previous models, we assume in this model that $\mu = 1$.

Let $\underline{\lambda} = (\lambda_G^{(1)}, \ldots, \lambda_G^{(J)}, \lambda_L^{(1)}, \ldots, \lambda_L^{(J)})$. From assumptions of the model and the law of total probability, the likelihood function is

$$L(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}) = \prod_{i=1}^{I} \int_0^{\infty} \left(\prod_{j=1}^{J} \mathbb{P}\left(Z_j^{(t)} = X_{i,j} \right) \right) f_T(t) dt.$$

Now the choice of $\mu = 1$ will be justified. We can express probabilities for the Markov chain by replacing parameters $\lambda_G^{(j)}$, $\lambda_L^{(j)}$ by parameters $\widetilde{\lambda_G^{(j)}}$, $\widetilde{\lambda_L^{(j)}}$ and time t by μt:

$$\mathbb{P}(Z_j^{(t)} = N) = e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t} = e^{-(\widetilde{\lambda_G^{(j)}} + \widetilde{\lambda_L^{(j)}})\mu t},$$

$$\mathbb{P}(Z_j^{(t)} = G) = \frac{\lambda_G^{(j)}}{\lambda_G^{(j)} + \lambda_G^{(j)}} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t} \right) = \frac{\widetilde{\lambda_G^{(j)}}}{\widetilde{\lambda_G^{(j)}} + \widetilde{\lambda_G^{(j)}}} \left(1 - e^{-(\widetilde{\lambda_G^{(j)}} + \widetilde{\lambda_L^{(j)}})\mu t} \right),$$

$$\mathbb{P}(Z_j^{(t)} = L) = \frac{\lambda_L^{(j)}}{\lambda_G^{(j)} + \lambda_G^{(j)}} \left(1 - e^{-(\lambda_G^{(j)} + \lambda_L^{(j)})t} \right) = \frac{\widetilde{\lambda_L^{(j)}}}{\widetilde{\lambda_G^{(j)}} + \widetilde{\lambda_G^{(j)}}} \left(1 - e^{-(\widetilde{\lambda_G^{(j)}} + \widetilde{\lambda_L^{(j)}})\mu t} \right).$$

By changing a variable with respect to which we integrate in the likelihood function and considering $u = \mu t$, we can conclude that the likelihood function of parameters $\widetilde{\lambda_G^{(1)}}, \ldots, \widetilde{\lambda_G^{(J)}}$ and $\widetilde{\lambda_L^{(1)}}, \ldots, \widetilde{\lambda_L^{(J)}}$ does not depend on μ. Therefore the choice $\mu = 1$ is now justifiable.

Now we consider the natural logarithm of the likelihood function $\log L(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda})$. We will approximate the integral in this function by the Laguerre quadrature with H nodes. It will be convenient to introduce some notation. Let

$$\mathbb{P}_{i,j}(t) = \mathbb{P}(Z_j^{(t)} = X_{i,j}).$$

Therefore we have:

$$\log L(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}) \approx lL(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}) = \sum_{i=1}^{I} log \left(\sum_{h=1}^{H} \prod_{j=1}^{J} \mathbb{P}_{i,j}(th) \cdot w_{th} \right),$$

where th are nodes and w_{th}- weights of the Laguerre quadrature.

Because in the second application of the model the aim is to obtain equal estimates for many regions, we will add an L^2 penalty to the lL function. Let

$$\text{Pen}(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}) = \sum_{j=1}^{J} \left(\left(\lambda_G^{(j)} - \hat{\lambda}_G \right)^2 + \left(\lambda_L^{(j)} - \hat{\lambda}_L \right)^2 \right)$$

be this penalty. $\hat{\lambda}_G, \hat{\lambda}_L$ are estimates from the simple multivariate model. We use this particular penalty to obtain estimates of regions that follow the typical behaviour equal to estimates from the model that describes the typical behaviour. Therefore we will maximize the function

$$lL(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}) - \gamma \text{Pen}(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma),$$

for fixed $\gamma \geq 0$. The bigger γ we take, then for more regions it should be $\hat{\lambda}_G^{(j)} = \hat{\lambda}_G$ and $\hat{\lambda}_L^{(j)} = \hat{\lambda}_L$. For the first application, we take $\gamma = 0$ in order to avoid penalization.

6.4. Maximization procedure

Let $F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma) = lL(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}) - \gamma \text{Pen}(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda})$. F is an approximated penalized log-likelihood. We will maximize the function F with respect to $\underline{\lambda}$. For this we will look for $\underline{\lambda}$ solving the set of equations:

$$\frac{\partial}{\partial \lambda_G^{(j)}} F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma) = 0,$$

$$\frac{\partial}{\partial \lambda_L^{(j)}} F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma) = 0,$$

for $j \in \{1, \ldots, J\}$.

For this purpose we will use the expectation-maximization algorithm. In the M- step we will use Newton-Raphson algorithm. Newton-Raphson algorithm uses the gradient and the hessian of the function one wants to maximize. Therefore we need to compute first and second order derivatives of the function F.

Of course $\mathbb{P}_{i,j}(t)$ depends on $\lambda_G^{(j)}$ and $\lambda_L^{(j)}$ and does not depend on $\lambda_G^{(k)}$ and $\lambda_L^{(k)}$ for $k \neq j$. Therefore for any $i \in \{1, \ldots, I\}$ and $t > 0$

$$\frac{\partial}{\partial \lambda_G^{(j)}} \prod_{k=1}^{J} \mathbb{P}_{i,k}(t) = \frac{1}{\mathbb{P}_{i,j}(t)} \cdot \frac{\partial \mathbb{P}_{i,j}(t)}{\partial \lambda_G^{(j)}} \cdot \prod_{k=1}^{J} \mathbb{P}_{i,k}(t),$$

$$\frac{\partial}{\partial \lambda_L^{(j)}} \prod_{k=1}^{J} \mathbb{P}_{i,k}(t) = \frac{1}{\mathbb{P}_{i,j}(t)} \cdot \frac{\partial \mathbb{P}_{i,j}(t)}{\partial \lambda_L^{(j)}} \cdot \prod_{k=1}^{J} \mathbb{P}_{i,k}(t).$$

Therefore we obtain expressions for derivatives of F with respect to $\lambda_G^{(j)}$ and $\lambda_L^{(j)}$:

$$\frac{\partial F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma)}{\partial \lambda_G^{(j)}} = \sum_{i=1}^{I} \sum_{h=1}^{H} \frac{\frac{1}{\mathbb{P}_{i,j}(th)} \cdot \frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_G^{(j)}} \cdot \prod_{k=1}^{J} \mathbb{P}_{i,k}(th) \cdot w_{th}}{\sum_{h=1}^{H} [\prod_{k=1}^{J} \mathbb{P}_{i,k}(th) \cdot w_{th}]} - 2\gamma \left(\lambda_G^{(j)} - \hat{\lambda}_G \right),$$

$$\frac{\partial F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma)}{\partial \lambda_L^{(j)}} = \sum_{i=1}^{I} \sum_{h=1}^{H} \frac{\frac{1}{\mathbb{P}_{i,j}(th)} \cdot \frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_L^{(j)}} \cdot \prod_{k=1}^{J} \mathbb{P}_{i,k}(th) \cdot w_{th}}{\sum_{h=1}^{H} [\prod_{k=1}^{J} \mathbb{P}_{i,k}(th) \cdot w_{th}]} - 2\gamma \left(\lambda_L^{(j)} - \hat{\lambda}_L \right).$$

Let

$$W_{i,h} = \frac{\prod_{k=1}^{J} \mathbb{P}_{i,k}(th) \cdot w_{th}}{\sum_{h=1}^{H} [\prod_{k=1}^{J} \mathbb{P}_{i,k}(th) \cdot w_{th}]}.$$

Then it holds that

$$\frac{\partial F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma)}{\partial \lambda_*^{(j)}} = \sum_{i=1}^{I} \sum_{h=1}^{H} \frac{1}{\mathbb{P}_{i,j}(th)} \cdot \frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_*^{(j)}} \cdot W_{i,h} - 2\gamma(\lambda_*^{(j)} - \hat{\lambda}_*),$$

where $* \in \{G, L\}$.

We also need to compute second order derivatives of $F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma)$. Because

$$\frac{\partial W_{i,\tilde{h}}}{\partial \lambda_*^{(j)}} = W_{i,\tilde{h}} \cdot \left(\frac{1}{\mathbb{P}_{i,j}(t\tilde{h})} \cdot \frac{\partial \mathbb{P}_{i,j}(t\tilde{h})}{\partial \lambda_*^{(j)}} - \sum_{h=1}^{H} \left(\frac{1}{\mathbb{P}_{i,j}(th)} \cdot \frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_*^{(j)}} \cdot W_{i,h} \right) \right),$$

then by direct computation and some transformations we obtain that

$$\frac{\partial^2 F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma)}{\partial \lambda_+^{(j_2)} \partial \lambda_*^{(j_1)}} =$$

$$= \sum_{i=1}^{I} \sum_{h=1}^{H} \left(\frac{1}{\mathbb{P}_{i,j_1}(th)} \cdot \frac{\partial \mathbb{P}_{i,j_1}(th)}{\partial \lambda_*^{(j_1)}} \cdot W_{i,h} \cdot \left(\frac{1}{\mathbb{P}_{i,j_2}(th)} \cdot \frac{\partial \mathbb{P}_{i,j_2}(th)}{\partial \lambda_+^{(j_2)}} - \sum_{h=1}^{H} \left(\frac{1}{\mathbb{P}_{i,j_2}(th)} \cdot \frac{\partial \mathbb{P}_{i,j_2}(th)}{\partial \lambda_+^{(j_2)}} \cdot W_{i,h} \right) \right) \right),$$

where $*, + \in \{G, L\}$ and j_1, j_2 are different regions.

If we consider parameters for the same region j, then the second order derivative is

$$\frac{\partial^2 F(\mathbf{X}_i : i = 1, \ldots, I; \underline{\lambda}, \gamma)}{\partial \lambda_*^{(j)} \partial \lambda_+^{(j)}} =$$

$$= \sum_{i=1}^{I} \sum_{h=1}^{H} \left(\frac{W_{i,h}}{\mathbb{P}_{i,j}(th)} \left(\frac{\partial^2 \mathbb{P}_{i,j}(th)}{\partial \lambda_*^{(j)} \partial \lambda_+^{(j)}} - \frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_+^{(j)}} \cdot \sum_{h=1}^{H} \left(\frac{1}{\mathbb{P}_{i,j}(th)} \cdot \frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_*^{(j)}} \cdot W_{i,h} \right) \right) \right) - 2\gamma \cdot \mathbf{1}_{\{*=+\}}.$$

6.4.1. Expectation-maximization algorithm

The expectation-maximization (E-M) algorithm is an iterative procedure of finding maximum likelihood estimates of parameters in models, where there is a latent variable present in the data. Examples of using E-M algorithm can be found in [6] and [7]. We will use the E-M procedure to obtain parameter estimates in the final multivariate model. For the initial estimate of the vector of parameters we use estimates from the univariate models. Therefore we initially take $\underline{\lambda} = (\frac{g_1}{n_1}, \ldots, \frac{g_J}{n_J}, \frac{l_1}{n_1}, \ldots, \frac{l_J}{n_J})$. We will use the following procedure to obtain better estimate of $\underline{\lambda}$:

- E-step: Use current estimate of $\underline{\lambda}$ to compute values of $W_{i,h}$ for $i \in \{1, \ldots, I\}$ and $h \in \{1, \ldots, H\}$.
- M-step: Use $W_{i,h}$ computed in the E-step to obtain representation of functions gradient and hessian. Use those representations to apply Newton-Raphson maximization algorithm. It will find better estimate of $\underline{\lambda}$.
- Update estimates of $\underline{\lambda}$ and go back to the E-step.

Those iterations are repeated until convergence.

6.4.2. Implementation of the E-M algorithm

There are a number of issues that are important in the implementation of this expectation-maximization algorithm. One should be aware that the gradient of the F function has length $2J$, the hessian of F is a $2J \times 2J$ matrix. J is the number of DNA regions, which is typically several hundreds to hundred thousand. This number is very big if we realize that the value of gradient and hessian must be computed for each step of the Newton-Raphson algorithm. Therefore the code of obtaining values of gradient and especially hessian must be optimized such that the time needed to compute those values is as short as possible.

6.4.3. Testing the estimation procedure on the simulated data

In order to check the accuracy of the estimation algorithm, we repeated 20 times the following procedure:

- sample 4 regions from the data
- compute univariate estimates of parameters of those 4 regions
- simulate the data of 300 samples and parameters equal to univariate estimates
- apply the E-M procedure without penalty to compute parameters' estimates

We compared estimates with the real parameters. Figure 6.1 shows that estimates do not deviate much from the real parameters. Thus the estimating procedure gives good results in the low-dimensional case.

Additionally, we computed log-likelihood for the real parameters and for the estimates. Figure 6.2 shows that likelihood for the estimated parameters- red crosses and the log-likelihood for the real parameters- black circles. It can be seen that the log-likelihood of the estimated parameters is slightly bigger than for the real parameters. This fact supports the thesis that the estimation procedure find maximizers of the log-likelihood function.

Figure 6.1: Comparison of real parameters and estimates

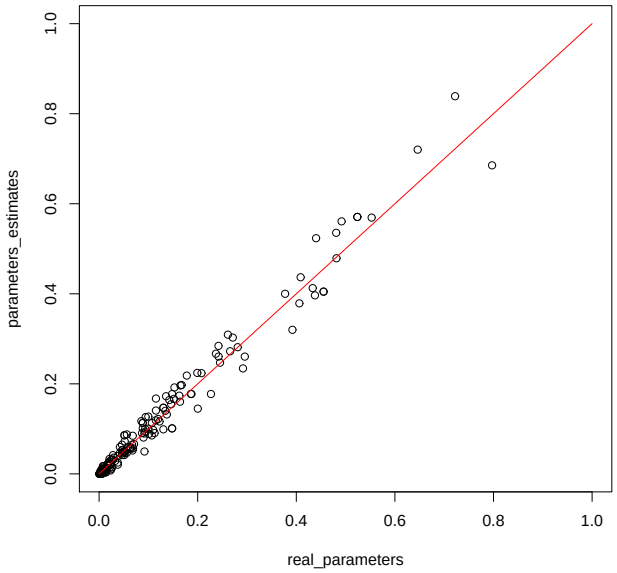
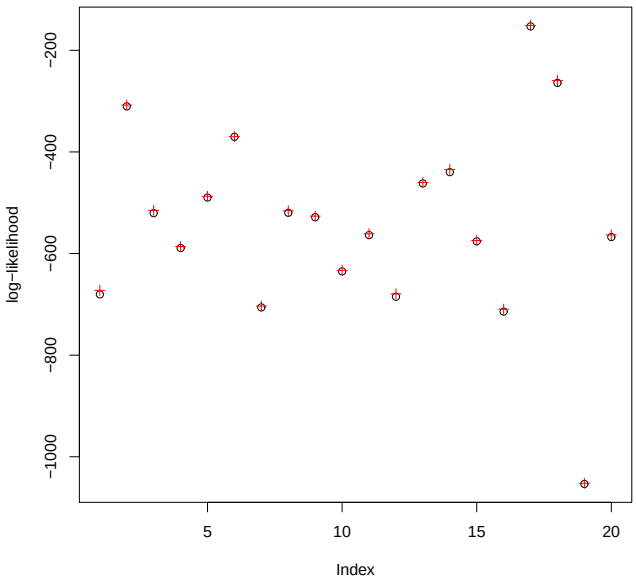


Figure 6.2: Comparison of the log-likelihood for real and estimated parameters



Chapter 7

Conclusions and future work

7.1. Conclusions

In this thesis we have developed a methodology for reconstructing the order, according to which DNA copy number aberrations are accumulated during cancer progression. Three different Markov chains were built to describe cancer progression. In this research, time estimation is essential for reconstructing the order of accumulating aberrations. Each model can be used for time estimation.

The simplest is the univariate model. We can use it only for one DNA region. The model has the advantage that there exist analytical solutions for the maximum likelihood estimators. If the data set consists of I samples and J genomic regions, we can fit a univariate model to each region separately, thus we will have J univariate models. For each model we obtain two parameters' estimates. Thus we can easily compute $2J$ estimates. We can use them as starting values for the algorithm in the final multivariate model.

The second model is a simple multivariate Markov chain with only two parameters. It describes the average behaviour of all regions in time. The estimation strategy in this model combines analytical solution and numerical optimization techniques. Estimates of the parameters of this model are present in the penalized log-likelihood function of the final multivariate model.

The third model is the final multivariate model. We designed the estimation strategy for this model. This is the expectation-maximization algorithm that uses parameters' estimates from the previous models. There are two methods of using this model. First is to fit the low-dimensional multivariate model. For this purpose we use unpenalized log-likelihood. The second method is to fit the model to the high-dimensional data. In this case we add a penalty to the log-likelihood function. The penalty contains estimates of the simple multivariate model, therefore the regions that show the average behaviour will inherit parameters from the simple multivariate model.

7.2. Future work

There are some upgrades that can be added to the project in order to improve the performance of the programmes and reduce the computation time.

7.2.1. Parallel programming

The expectation-maximization algorithm is time consuming when applied to a big data set. If the algorithm needs to compute the value of a gradient and a hessian matrix for the parameter $(\lambda_G^{(1)}, \ldots, \lambda_G^{(J)}, \lambda_L^{(1)}, \ldots, \lambda_L^{(J)})$, then the following components have to be computed:

- $\mathbb{P}_{i,j}(th)$,
- $W_{i,h}$,
- $\frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_G^{(j)}}$,
- $\frac{\partial \mathbb{P}_{i,j}(th)}{\partial \lambda_L^{(j)}}$,
- $\frac{\partial^2 \mathbb{P}_{i,j}(th)}{\partial \lambda_G^{(j)2}}$,
- $\frac{\partial^2 \mathbb{P}_{i,j}(th)}{\partial \lambda_L^{(j)2}}$,
- $\frac{\partial^2 \mathbb{P}_{i,j}(th)}{\partial \lambda_G^{(j)} \partial \lambda_L^{(j)}}$,

for $i \in \{1, \ldots, I\}$, $j \in \{1, \ldots, J\}$, $h \in \{1, \ldots, H\}$. If the data consists of many regions or samples, then the expectation-maximization algorithm becomes time consuming.

There is a solution of this problem. For this solution it is important that the above components have the following important property. For example, for computing $\mathbb{P}_{i,j}(th)$ for all i, j, th, we can build a three- dimensional array and it holds that for calculating a single element of this array, we do not need to know values of any other elements. Therefore computations of this array could be done in parallel. It would speed up the process. Thus we propose parallel programming for computing elements of the above arrays.

7.2.2. Fitting the model to the real data set

Assuming we have the efficient programme for the E-M estimation, we can fit the final multivariate model to the high-dimensional data set for different values of γ. First we choose big γ, so that we should obtain many parameters equal to those from the simple multivariate model. We decrease γ until we reach the point in which we obtain a reasonable number of different parameters. We use the obtained parameters' estimates to calculate time estimates. Samples with corresponding time estimates are used to obtain values of functions: profileN_j, profileG_j and profileL_j. Those functions describe the pattern of accumulating aberration at region j.

Bibliography

[1] Kallioniemi A. CGH microarrays and cancer. *Current Opinion In Biotechnology* 2008;19:36–40. doi: 10.1016/j. copbio.2007.11.004.

[2] Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* 37(Suppl): S11–S17.

[3] van de Wiel MA, Picard F, van Wieringen WN, Ylstra B; Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform* 2010, (Epub ahead of print).

[4] van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B; CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics* 2007, in press.

[5] Zhang Y, Martens JWM, Yu JX et al.: Copy number alterations that predict metastatic capability of human breast cancer. *Cancer Res.* 69(9),3795–3801 (2009).

[6] Muraki E, A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, v16 n2 p159-76 Jun 1992.

[7] Bock, R. D, Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm *Psychometrika* 46(4), 443-459 Jan 1981.