2014

Functional Data Analysis and its application to cancer data

Evgeny Martinenko
University of Central Florida

FUNCTIONAL DATA ANALYSIS AND ITS APPLICATION TO CANCER DATA

by

EVGENY MARTINENKO
M.S. University of Central Florida, 2011

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Mathematics
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Summer Term
2014

Major Professor: Marianna Pensky

ABSTRACT

The objective of the current work is to develop novel procedures for the analysis of functional data and apply them for investigation of gender disparity in survival of lung cancer patients. In particular, we use the time-dependent Cox proportional hazards model where the clinical information is incorporated via time-independent covariates, and the current age is modeled using its expansion over wavelet basis functions. We developed computer algorithms and applied them to the data set which is derived from Florida Cancer Data depository data set (all personal information which allows to identify patients was eliminated). We also studied the problem of estimation of a continuous matrix-variate function of low rank. We have constructed an estimator of such function using its basis expansion and subsequent solution of an optimization problem with the Schatten-norm penalty. We derive an oracle inequality for the constructed estimator, study its properties via simulations and apply the procedure to analysis of Dynamic Contrast medical imaging data.

To my beloved wife Penny.

ACKNOWLEDGMENTS

I would like to express my appreciation and gratitude to my professors, family, and friends for their support throughout my graduate studies. I am tremendously thankful for the patient supervision and guidance of Dr.Marianna Pensky. Her insightfulness and constructive comments have been instrumental in the completion of this thesis. Thank you Dr. Ji-Hyun Lee, Dr. Liqiang Ni, and Dr. Jason Swanson for agreeing to serve on my committee. I also would like to thank Dr.Tatiana Zhukov whose idea initiated current work. Finaly I want to thank my good friend Aritra Dutra who was helping me for four last years during good and bad times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The lung cancer is the number one killer for men and women in USA, since year 1987 when it surpassed the breast cancer killing rate in women [1]. The lung cancer causes more death than 3 next most dangerous cancers combained: breasts, colorectal and pancreas. The predicted death rate of lung cancer in year 2014 is almost 160 millions. American Lung Cancer Assosiation reports that the number of deaths due to lung cancer has increased approximately 4.3 percent between 1999 and 2010 from 152,156 to 158,318. The number of lung cancer deaths among men has reached a plateau but the number is still rising among women. In 2010, there were 87,740 deaths due to lung cancer in men and 70,578 in women. The age-adjusted death rate for lung cancer is higher for men (60.3 per 100,000 persons) than for women (38.1 per 100,000 persons). The rate of new lung cancer cases (incidence) over the past 36 years has dropped for men (24% decrease), while it has risen for women (100% increase). In 1975, rates were low for women, but rising for both men and women. In 1984, the rate of new cases for men peaked (102.1 per 100,000) and then began declining. The rate of new cases for women increased further, did not peak until 1998 (52.9 per 100,000), and has now started to decline

Most studies have reported that women receive lung cancer diagnoses at a younger median age, suggesting that they may have an increased susceptibility to the development of lung cancer. Henschke and Miettinen [16] provided evidence that, for a given level of smoking, more women than men develop lung cancer, using baseline CT screening for lung cancer. However, not all studies support this observation and the effects of gender on the lung cancer risk associated with tobacco use remains incompletely resolved [10].

The initial objective of current work was to determine whether observed gender disparity in survival of lung cancer patients can be explained, at least in part, by hormonal differences between

males and females. Gender differences in lung cancer susceptibility, presentation and survival may result from estrogen involvement. In particular, our hypothesis was that females with stage 0, 1 or 2 lung cancer exhibit unusual survival patterns around menopausal age when hormonal instabilities become more pronounced. This conjecture is supported by preliminary data analysis of Florida cancer data base which contains data for cancer patients in the state of Florida between years 1986 and 2005.

Survival analylis is the tool of choise in our investigation. This branch of statistical science deals with finding and analysing relationships between some univariate lifetime variable, which is often referred to as failure time and some set of explanatory variables. To carry out our investigation, we develop a Cox regression model with current age-dependent covariates, which in turn were created as functional expansion of unknow hazard function. It is well known that survival of a lung cancer patient is strongly correlated with age. To account for this fact, patients are usually grouped by their age at diagnosis and the age group numbers act as categorical covariates. This approach works very well for more advanced stages of cancer where survival of a patient beyond 2-3 years is very unlikely. However, for early stages of cancer where average survival time is much longer, a patient might go into remission which is interrupted by an age-dependent event (say, menopause). In order to investigate how survival rates depend on the current age of a patient, one needs to include it into the model together with survival time. Since the current age, the survival time and the age of diagnosis are functionally related, we represent the age dependent component of the hazard function via its expansion over a wavelet basis, in particular, Haar and Mexican hat wavelet bases [15].

The rest of the dissertation is organized as follows. Chapter 2 provides background information. In particular, section 2.1 introduces main notions, such as censoring and truncation, and main techniques in survival analysis, such as the Kaplan-Meier estimator, the Cox proportional hazard model, and accelerated failure time model. Section 2.2.1 reviews wavelet techniques. Section

3.3 presents a preliminary study of gender and age-dependent survival rates. In particular, Section 3.3.1 introduces wavelet model for hazard and survival functtions which allows to take into account current age of a patient while analyzing survival rates. Section 3.3.2 contains comprehensive description of the data. Section 3.3.4 contains results of our study. Section 4 contains father developing of the functional data analysis applied to tensor data. Chapter 4 devoted to another interesting topic - denoising of the medical data. Finally, Section 5.2 concludes the paper with the proposal on future work.

CHAPTER 2: BACKGROUND INFORMATION

2.1 Survival analysis

Survival analysis deals with the death of biological organisms, failure in mechanical systems or generally with timing of events, i.e., the time elapsed from a specific time origin until the event occurs. In practice an event time could be, for example, the time until tumor recurrence(length of remission, the time until a car engine fails, time of death from a particular disease, time for unemployed to find a job, etc. One can see that there is a possibility that a time-to-event will not be observed due to many causes, one of them is that the event has not occur during the time of observation. We call the case of incomplete observation - **censoring**. The necessity of designing methods for analysis that accommodate censoring is the primary reason for developing models and procedures for time-to-event data. Prior to development of these procedures , incomplete data were treated as missing data and omitted from the analysis. This resulted in the loss of information and introducing bias in the estimated quantities, which lowers efficacy of the studies.

2.1.1 Basic definitions

We are following below [28] Assume for the moment that we know the distribution function of events in interest. Let $F(\cdot)$ denote the **distribution function** with corresponding **probability density function** $f(\cdot)$, let T denote a non negative random variable representing the lifetimes of individuals in some population. Then the probability that an individual survives till time t is given by

the **survival function**:

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^{+\infty} f(x)\, dx. \tag{2.1}$$

The **hazard function** specifies the instantaneous rate of failure at $T = t$ given that the individual survived up to time t and is defined as:

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.2}$$

The hazard function is also referred to as **mortality rate**, it can assume values in $[0, \infty)$.It is easily verified that $h(t)$ specifies the distribution of T, since:

$$h(t) = -\frac{dS(t)/dt}{S(t)} = -\frac{d\log(S(t))}{dt}$$

Upon integrating $h(t)$ over $(0, t)$ we obtain another useful quantity- the **cumulative hazard function** $H(t)$:

$$H(t) = \int_0^t h(x)\, dx = -\log(S(t)) \tag{2.3}$$

Therefore knowing the hazard function is enough to find the distribution function $F(t)$ and probability density function $f(t)$ as:

$$F(t) = 1 - \exp(-\int_0^t h(x)\, dx) \quad f(t) = h(t)\exp(-\int_0^t h(x)\, dx)$$

There are couple of useful formulas one can easily derive from the above given information and basis principals: The well known **mean value** of the random variable T , $E(T)$ can be expressed

using the survival function $S(t)$ as :

$$E(T) = \int_0^\infty S(t)\, dt$$

Another interesting function is the **mean residual life** at time T, denoted by $MRL(T)$.For individuals of age T, it is the expected remaining lifetime, defined as:

$$MRL(T) = E(t - T|t > T) = \frac{\int_T^\infty S(t)\, dt}{S(T)} \tag{2.4}$$

It is clear that MRL is the area under survival curve to the right of T divided by $S(T)$.

2.1.1.1 Censoring and truncation

There are three main types of censoring: right, left and interval censoring. The right censoring is the one that occurs most often. This type arises when we do not observe the failure time due to terminating of the experiment or due to withdrawal of the subject from observation. It is very useful to introduce a binary random variable δ which indicates whether we have observed a true failure time:

$$\delta = \begin{cases} 1, & \text{if } T \leq t_c \\ 0, & \text{if } t_c < T \end{cases} \tag{2.5}$$

Thus, in every survival experiment we observe an iid pairs (Y_i, δ_i), where

$$Y_i = \begin{cases} T_i & \text{if } T \leq t_c \\ t_c, & \text{if } t_c < T \end{cases} \tag{2.6}$$

Hence, for the n observations we can construct a joint probability distribution function assuming we know p.d.f. for the one observation $f(y_i)$. That joint p.d.f. is called the likelihood and is denoted by L. The expression for L is easily derived as:

$$L = \prod_{i=1}^{n} [f(y_i)]^{\delta_i} [S(y_i)]^{1-\delta_i}$$

Another type of censoring is the so called left censoring when the event of interest already occurred at the beginning of observation time, but it is not known when exactly. The left censoring is much rare than right one, the examples of the left censoring include:

- the moment of being infected with a sexually-transmitted disease
- time at which teenagers begin to drink alcohol

Finally there is one more type of censoring: interval censoring. The interval censoring occurs if the exact time when event occurs is not known precisely, but an interval bounding this time is known. For example, a study of a light bulb life where we screw bulbs into sockets and then check whether they are on or off once a day is the example of the interval censoring with the interval one day.
One need to distinguish censoring from truncation. Truncation is a procedure where a condition other than the main event of interest is used to screen patients. Here, we cite an example of right truncation [7]. Early in the AIDS outbreak, patients with AIDS were recruited to study the time from being infected with HIV to development of AIDS. At the time of the study, many people were infected with HIV but had not yet developed symptoms of AIDS.
In the research below we shall be concerned only with right censored data.

Table 2.1: Hypothetical drug study (+ indicates a censored value)

Group	**Length of complete remission(weeks)**												
Treated	6	13	13+	15+	19	19+	22	25+	30	44	49+	99	112+
Nontreated	5	8	9+	11	13+	18	18	19+	25	30	37	44	51

2.1.1.2 Nonparametric methods

In real life situation it is a rare case when the distribution function is known a priory. In order to conduct statistical inference one needs methods that make no assumptions on the shape of the underlying distribution function, the so called non parametric methods. It is well understood that non parametric methods are not a panacea -all they are based on some kind of assumptions. The goal on non parametric methods is to retrieve as much information as possible from the given data. We will review here three main non parametric methods ranged by difficulty.

2.1.1.2.1 The Kaplan-Meier estimator

Consider the hypothetical data for cancer remission :

The naive approach would be to treat the data as if there are no censored observations. The **empirical survival function(esf)** denoted as $S_n(t)$ is defined as:

$$S_n(t) = \frac{\#\{\text{of observations} > t\}}{n} = \frac{\#\{t_i > t\}}{n}$$

As we mentioned above, the attempt to ignore censoring leads to biased estimation. To account for the censoring Kaplan and Meier introduced the product-limit estimator is their seminal paper [23]. The following notation will be used:

- n_i -the number of patients alive and not censored just before y_i
- d_i - the number of patients died at time y_i
- $p_i = P(T > y_i | T > y_{i-1})$
- $q_i = 1 - p_i$

Note that the **death** is the generic word for the event of interest. The **Kaplan-Meier estimator** is defined as :

$$\widehat{S(t)} = \prod_{y_i \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$

Figure 2.1. shows the result of application Kaplan-Meier estimator to our toy example data in Table 2.1. One can see that ignoring censored data indeed introduces bias in estimation of the survival function.

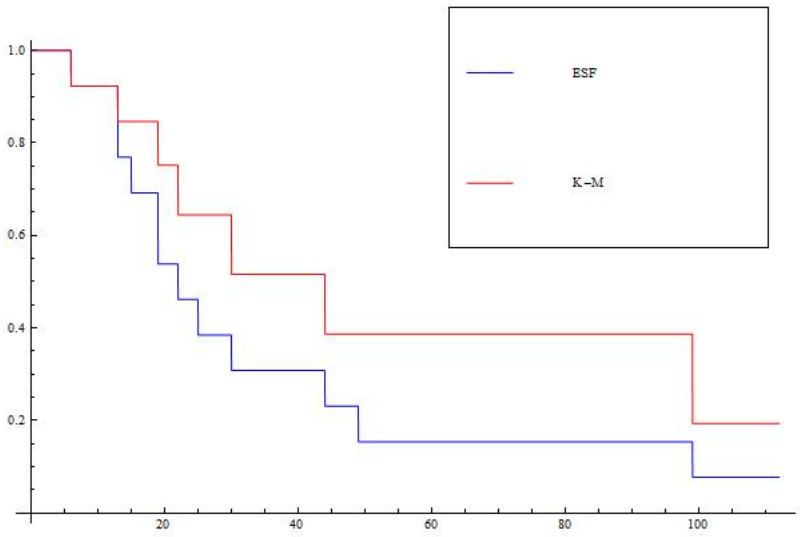


Figure 2.1: Comparison of naive approach (ESF) and Kaplan-Meier (K-M) estimators for the survival data (2.1)

2.1.1.2.2 Cox-Proportional Hazard Model

The Cox-Proportional Hazard model, the Cox model for short, is neither parametric, nor non-parametric model, it is a semi-parametric model. Cox model offers association of hazard function with covariates and, therefore, provides a very, versatile tool for survival analysis. It assumes that the hazard function is of the form

$$h(t|\mathbf{x}) = h_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}, \tag{2.7}$$

where **x(t)** $= (x_1(t), ..., x_n(t))$ is a n-dimensional measurable covariate and β is the n-dimensional regression parameter. Let us consider two sets of predictors from the same model: $\mathbf{X}$ and $\mathbf{X}^*$. Taking the ratio of hazard functions, we obtain the following hazard ratio (HR):

$$HR = \frac{h(t|\mathbf{X}^*)}{h(t|\mathbf{x})} = \exp\left[\sum_{i=1}^{p}(\mathbf{X}^* - \mathbf{X})\,\beta\right]. \tag{2.8}$$

The fact that hazard ratio is independent on time, is the manifestation of the Cox Model and is responsible for the name- Proportional Hazard.

Let T, C and **x** be, respectively, the survival time, the censoring time and their associated covariates. Correspondingly, let Y as in (2.6) be the observed time and $\delta = I(T \leq C)$ be the censoring indicator as in (2.5). It is assumed that T and C are conditionally independent given **x** and the censoring mechanism is noninformative. When the observed data $(\mathbf{x}_i, Y_i, \delta_i), i = 1, ..., n$, forms an i.i.d.random sample from a certain population$(\mathbf{x}, Y, \delta)$, a complete likelihood of the data is given by:

$$L = \prod_{i\in u} f(Y_i|\mathbf{x}_i)\prod_{i\in c}\bar{S}(Y_i|\mathbf{x}_i) = \prod_{u} h(Y_i|\mathbf{x}_i)\prod_{i=1}^{n}\bar{S}(Y_i|\mathbf{x}_i) \tag{2.9}$$

where the sets of indices *c* and *u* refer to the censored and uncensored data respectively, and $\bar{S}(Y_i|\mathbf{x}_i)$, $h(Y_i|\mathbf{x}_i)$ and $f(Y_i|\mathbf{x}_i)$ are the conditional survival function, conditional hazard function and the conditional density function of T given $\mathbf{x}$.

In order to present the likelihood function of Cox's proportional hazards model explicitly, more notation is needed. Let $t_1^0 < \cdots < t_N^0$ denote the ordered observed failure times. Let (j) provide the label for the item falling at t_j^0, so that the covariates associated with N failures are $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}$. Let R_j denote the risk set right before the time t_j^0.

$$R_j = \left\{ i : Y_i \geq t_j^0 \right\}.$$

Using assumption (2.7)of the proportional hazards model , the likelihood in (2.9) becomes:

$$L = \prod_{i=1}^{N} h_0(Y_i) e^{\mathbf{x}_i^T \boldsymbol{\beta}} \prod_{i=1}^{n} \exp(-H_0(Y_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})),$$

where $H_0(\cdot) = \int h_0(t)\, dt$ is the cumulative baseline hazard function. In the Cox proportional hazard model, the baseline is unknown and has not been parametrized. Consider the "least informative" non-parametric modelling for $H_0(\cdot)$, in which $H_0(t)$ has a possible jump h_j at the observed failure time T_j^0. More precisely, let $H_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. Then

$$H_0(Y_i) = \sum_{j=1}^{N} h_j I(i \in R_j) \tag{2.10}$$

Using (2.10), the logarithm of likelihood of (2.9) becomes

$$\sum_{j=1}^{N} \{\log(h_j) + \mathbf{x}_(j)^T \boldsymbol{\beta}\} - h_j I(i \in R_j) - \sum_{i=1}^{n} \left\{ \sum_{j}^{N} h_j I(i \in R_j) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} \tag{2.11}$$

Taking the derivative with respect to h_j and setting it to zero, we obtain:

$$\hat{h}_j = \left\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\}^{-1}$$

Substituting $\hat{h}_j$ back into (2.11) and adding the penalty term, we obtain the penalized Cox likelihood:

$$l = \sum_{j=1}^{N} \left[\mathbf{x}_j^T \boldsymbol{\beta} - \log \left\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} \right] - n \sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{2.12}$$

where d is the dimension of $\boldsymbol{\beta}$. The penalized likelihood estimate of $\boldsymbol{\beta}$ is derived by maximizing (2.12) with respect to $\boldsymbol{\beta}$. With a proper choice of p_λ, many of the estimated coefficients will be zero and hence, the corresponding variables do not appear in the model. This achieves the objectives of the variable selection. To justify the better models, various criteria are used, the most popular of them isAkaike Information criterion(AIC), see e.g. [28].

2.1.1.2.3 Cox Model for Time-Dependent Variables

The great advantage of the Cox model it is its flexibility and expandability. The case of time-dependent covariates is the example of the above statement. Given a situation involving both time-independent and time-dependent predictor variables, we can write an extended Cox model that incorporates both types as shown below:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[\sum_{i=1}^{p_1} \boldsymbol{\beta}_i X_i + \sum_{j=p_1+1}^{p_2} \delta_j X_j(t) \right] \tag{2.13}$$

(see, e.g. [24]). Equation (2.13) is similar to the equation (2.7) in a sense that both contain baseline function $h_0(t)$. Exponential part in equation (2.13) contains, however, both time-independent predictors, as denoted by X_i and time-dependent predictors denoted by X_j. To assess values of δ_j and coefficients $\boldsymbol{\beta}_i$, the maximum likelihood estimation (MLE) procedure is used, which assumes that the hazard at time t depends on the value of $X_i(t)$ at the same time t. However, it is possible to modify the definition of the time-dependent covariates to allow for a "lag-time" effect. MLE's are obtained by maximizing a (partial) likelihood function. However, the computations for the extended Cox model turn out to be more complicated than for the Cox model (2.7), because the risk sets used to form the likelihood function are more complicated with time-dependent variables. The apparent simplicity of the model (2.13) has its drawbacks, the most serious of which is the time dependence of the hazard ratio.

2.1.1.2.4 Cox Likelihood maximization

Let's start with non-penalized version of Cox likelihood. The goal is to find vector $\vec{\boldsymbol{\beta}}$ where:

$$\boldsymbol{\beta} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{j=1}^{N} \delta_j \left[\mathbf{X}_j^T \boldsymbol{\beta} - \log \left(\sum_{i \in R_j} \exp(\mathbf{X}_i^T \boldsymbol{\beta}) \right) \right] \right\}$$

where $\mathbf{X}$ is the matrix of covariates, which can be time dependent. Most common algorithm to solve the maximization problem is the Newton-Raphson algorithm. Starting with some initial guess $\boldsymbol{\beta}^0$ it proceeds iteratively computing:

$$\boldsymbol{\beta}^{i+1} = \boldsymbol{\beta}^i - \mathbf{H}^{-1}(\boldsymbol{\beta}^i)\mathbf{G}(\boldsymbol{\beta}^i) \tag{2.14}$$

where $\mathbf{G}(\cdot)$ is the gradient vector of log-likelihood and $\mathbf{H}(\cdot)$ is the matrix of second derivatives of the log-likelihood. The k^{th} component of the gradient vector $\mathbf{G}(\boldsymbol{\beta})$ is calculated as:

$$\frac{\partial l}{\partial \boldsymbol{\beta}_k} = \sum_{j=1}^{N} \delta_j \left[x_{jk} - \frac{\sum_{i \in R_j} x_{ik} \exp(\mathbf{X}_i \boldsymbol{\beta})}{\sum_{i \in R_j} \exp(\mathbf{X}_i \boldsymbol{\beta})} \right] \tag{2.15}$$

the $p \times p$ matrix $\mathbf{H}$of the second derivative of the log-likelihood , has components given by the following formula:

$$H_{k,m} = \sum_{j=1}^{N} \delta_j \left[\frac{\sum_{i \in R_j} x_{ik} \exp(\mathbf{X}_i \boldsymbol{\beta}) \sum_{i \in R_j} x_{im} \exp(\mathbf{X}_i \boldsymbol{\beta})}{(\sum_{i \in R_j} \exp(\mathbf{X}_i \boldsymbol{\beta}))^2} - \frac{\sum_{i \in R_j} x_{ik} x_{im} \exp(\mathbf{X}_i \boldsymbol{\beta})}{\sum_{i \in R_j} \exp(\mathbf{X}_i \boldsymbol{\beta})} \right] \tag{2.16}$$

The simple observation brought by [37] will allow us to change notation and to simplify equations above. Since we have our design matrix sorted according to survival times, each event time j defines a risk set R_j which is represented by the rows of matrix $\mathbf{X}$ starting with j. We can therefore consider $\frac{\exp(\mathbf{X}_i \boldsymbol{\beta})}{\sum_{i \in R_j} \exp(\mathbf{X}_i \boldsymbol{\beta})}$ as probability $p_i^{(j)}$ from probability distribution P^j. Taking that into consideration we can rewrite equations for gradient and Hessian as:

$$\frac{\partial l}{\partial \boldsymbol{\beta}_k} = \sum_{j=1}^{N} \delta_j \left[x_{j,k} - \langle \mathbf{X}_{\cdot,k} \rangle^{(j)} \right]$$

$$H_{k,m} = \sum_{j=1}^{N} \delta_j \left[\langle \mathbf{X}_{\cdot,k} \rangle^{(j)} \langle \mathbf{X}_{\cdot,m} \rangle^{(j)} - \langle \mathbf{X}_{\cdot,k} \mathbf{X}_{\cdot,m} \rangle^{(j)} \right]$$

where $\langle \cdot \rangle^{(j)}$ is the average over distribution created from the risk set R^j. Properties of that distribution will be described below.

2.1.1.2.5 Accelerated failure time model

Accelerated failure time model is a fully parametric approach. The model assumes that response is log-linear with respect to covariates, i.e.:

$$Y = \log(T) = \vec{X}\beta + Z \tag{2.17}$$

where Z is a random variable which does not appear in the inference. distribution. Then for $T^* = \exp Z$ one has

$$T = \exp(Y) = T^* \exp(\vec{X}\beta) \tag{2.18}$$

Assuming that T^* and T have pdf's $f^*(\cdot)$ and $f(\cdot)$ cdf's $F^*(\cdot)$ and $F(\cdot)$, respectively, define function $g(\cdot)$ such that $T = g(T^*)$. Simple algebra yields:

$$f(t) = P(T < t) = P(T^* < g^{-1}(t)) = F^*(g^{-1}),$$

hence,

$$f(t) = f^*(g^{-1}(t)) \left| \frac{\partial g^{-1}(t)}{\partial t} \right|.$$

If T^* has the hazard function $h_0^*(t^*)$, the above pdf can be rewritten as :

$$f(t) = h_0^*(g^{-1}(t)) \left| \frac{\partial g^{-1}(t)}{\partial t} \right| S^*(g^{-1}(t))$$

Using comparison of equation (2.18) with definition of $g(t)$, the derivative can be evaluated and the following expression for the hazard function obtained:

$$h(t|\vec{X}) = h_0^* \left(t \exp(-\vec{X}\beta) \right) \exp(-\vec{X}\beta) \tag{2.19}$$

The equation (2.19) makes apparent the word "accelerated" in the name of the model.

2.2 Functional Data Analysis

Functional Data Analysis (FDA) is a way to represent discrete data by continuous functions (see e.g. see [21]) with intention to achieve the following goals:

- reduce the noise
- find the underlying pattern
- predict the outcomes for the new data

Functional Data Analysis tends to treat discrete data as functional entities (see e.g. [21]) by assuming the existence of the intrinsic structure of the data. In practice, functional data are observed discretely as pairs (t_j, y_j) where

$$y_j = f(t_j) + \epsilon_j, \quad j = 1, \cdots, n.$$

is possibly blurred by the measurement errors. We usually wish to declare function $f(t)$ as smooth, i.e., the pair of adjacent data values, y_j and y_{j+1} are linked together and unlikely to be too different from each other. If this property doesn't hold there would be nothing much to be gained by treating the data as functional rather than just multivariate. Assuming that the observed data values are values of some function, the first task is to convert these values to a function, computable for any sensible argument. Due to the error, one can not use an interpolation algorithm and need to employ some kind of smoothing technique. The basic ideas of FDA can be described as follows. Consider the model: $f_i(x) = \sum_{i=1}^{N} \phi_i(x) c_i$. The system $\{\phi(\cdot)_j\}_1^\infty$ is called the basis system for

f. A basis function system is a set of known functions that are linearly independent and can approximate arbitrary well any function by taking a weighted sum of a sufficiently large number of these functions. The most trivial basis function system would be the collection of monomials:

$$1, t, t^2, t^3, \cdots, t^n, \cdots .$$

. In spite of its simplicity, the system above is used very rarely in real application of FDA due to the very high value of the conditional number of the transformation matrix and thus difficulties to invert the transformation. Therefore it is important to select a right basis system. Below we discuss the three most popular bases: the Fourier basis, the B-spline basis and the wavelet basis.

- **Fourier Basis**.

$$\left\{e^{i2\pi\, k\, t}\right\}_{k=-\infty}^{\infty}$$

  - natural representation of periodic functions
  - exelent from the computational point of view is the sample data are equidistant.
  - can be problematic forrepresentation of the non-periodic functions.

  The best known expansion, the basis is periodic and the period is usually equal to the length of interval over which the observation took place. Fourier series generally yields expansions which are uniformly smooth, but they are inappropriate to some degree for data with discontinuities in the function itself or in low order derivatives.

- **B-splines**.

  Consider with a partition of some interval (a, b) We call that partition a knot sequence $\{t_i\}$.

Thus the **B-splines** of the first order will be the indicator function for that partition, i.e. :

$$B_{i1}(x) = \begin{cases} 1,, & \text{if } t \in [t_i, t_{i+1}) \\ (n+1)/2, & \text{otherwise.} \end{cases} \tag{2.20}$$

with the constraint: $\sum_i B_{i1}(t) = 1, \quad$ for all t. Starting from the first order, all following orders of **B-splines** obtained recursively:

$$B_{ik} = \omega_{ik} B_{i,k-1} + (1 - \omega_{i+1,k}) B_{i+1,k-1} \tag{2.21}$$

where

$$\omega_{ik}(t) = \begin{cases} \frac{t-t_i}{t_{i+k-1}-t_i}, & \text{if } t \neq t_{i+k-1} \\ 0, & \text{otherwise.} \end{cases}$$

B-splines constitute a common choice for non-periodic data. To define a spline one needs to divide the interval, over which a function is to be approximated, into L subintervals . Over each subinterval, a spline is a polynomial of a specified order m. Adjacent polynomials joint up smoothly at the breakpoints, moreover, derivatives up to order $(m-2)$ also match up. The advantages of the B-splines are:

- good for non cyclic data
- very effective computationally
- tricky knot placement can be a drawback of B-splines

- **Wavelets**

The wavelet expansion of a function f gives a multiresolution analysis, providing information on a sequence of degrees of locality. It copes very well with discontinuities or rapid changes in the behavior of the function. More information on wavelets are given in the

section below.

2.2.1 Wavelet Basis

We are following below [8]. The fundamental idea behind wavelets is to analyse a function according to scale and location simultaneously. It introduces univariate function ψ, defined on $\mathbb{R}$,which, when subjected to the fundamental operations of shifts and dyadic dilation, yields an orthogonal basis of $L_2(\mathbb{R})$. That is, the functions

$$\psi_{j,k}(\cdot) = 2^{j/2}\psi(2^j \cdot -k), \quad j,k \in \mathbb{Z} \tag{2.22}$$

form a complete orthogonal basis of $L_2(\mathbb{R})$. One can view wavelet as a "bump" with a compact support(though not necessary). Dilation squeezes or expands the "bump" and translation shits it. Thus $\psi_{j,k}$ is a scaled version of ψ centered at the dyadic integer $k/2^j$. The large value of j -the smaller support of $\psi_{j,k}$.

The requirement that the set $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ forms an orthonormal system means that any function $f \in L_2(\mathbb{R})$ can be represented as a series

$$f = \sum_{j,k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle \, \psi_{j,k} \tag{2.23}$$

Wavelets corresponding to small values of j contribute to broad resolution of f; those corresponding to larger values of j give finer details. The known advantage of using wavelet basis versus other type of bases is that it accommodates jump discontinuities in the underlying function. Wavelets allow one to investigate behaviour of a spatially inhomogeneous function without any assumptions on the size or location of the intervals of smooth (or stable) behaviour. The decomposition (2.23) is analogous to the Fourier decomposition of a function f in terms of complex exponents, but there

is important differences:

- The exponential functions have global support. Thus, all terms in the Fourier decomposition contribute to the value of f at a point x. On the other hand, wavelets usually either have compact supports or have fast decay at infinity. Thus, only the terms in (2.23) corresponding to $\psi_{j,k}$ with $k/2^j$ near x make a large contribution at x. In this sense, representation (2.23) is local.

- The coefficients in wavelet decompositions usually encode all information needed to tell whether f is in a smoothness space, such as the Sobolev and Besov spaces. For example, if ψ is smooth enough, then a function f is in the Lipschitz space $\text{Lip}(\alpha, L_\infty(\mathbb{R}))$ iff

$$\sup_{j,k} \left\{ 2^{j(\alpha+1/2)} |\, \langle f, \psi_{j,k} \rangle \,| \right\} < \infty$$

  .

In this work we have used only two types of wavelets: Haar and Mexican Hat wavelets. We give father details below.

2.2.1.1 The Haar Wavelets

The Haar functions are known since 1910. They are the most elementary wavelets. While they have many drawbacks, the main one being lack of smoothness, they illustrate, in the most direct way, some of the main features of wavelet decompositions. For this reason, we shall consider in some detail the properties that make them suitable for numerical applications. We will follow here

[15]. Consider the following subspace V_0 of $L_2(\mathbb{R})$:

$$V_0 = \{f \in L_2(\mathbb{R}) : f \quad \text{is constant on } (k, k+1], k \in \mathbb{Z}\}$$

Thus,

$$f \in V_0 \Leftrightarrow f(x) = \sum_k c_k \phi(x-k)$$

where

$$\phi(x) = I\{x \in (0,1]\} = \begin{cases} 1, & x \in (0,1] \\ 0, & x \notin (0,1] \end{cases} \tag{2.24}$$

The translated function denote as $\phi_{0,k} = \phi(x-k), \quad k \in \mathbb{Z}$. Applying dilation, define a new linear subspace of $L_2(\mathbb{R})$ by

$$V_1 = \{h(x) = f(2x) : f \in V_0\}$$

It is clear that $V_1 \subset V_0$. Upon normalization the system of basis function of V_1 is $\{\phi_{0,k}\}$, where

$$\phi_{0,k}(x) = \sqrt{2}\phi(2x-k), \quad k \in \mathbb{Z}$$

Following the pattern one can continue building spaces:

$$... \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 ...$$

and their basis's:

$$\phi_{j,k}(\cdot) = 2^{k/2}\phi(2^k \cdot -j), \quad j,k \in \mathbb{Z}$$

Since every function in $L_2(\mathbb{R})$ can be approximated by a piecewise constant function, it means that the linear span of the system of functions $\{\{\phi_{0,k}\}, \{\phi_{1,k}\}, ...\}$ is dense in $L_2(\mathbb{R})$. This system

is not an orthogonal basis in $L_2(\mathbb{R})$, but it is possible orthogonalize it. Introducing orthogonal complement of V_0 to V_1:

$$W_0 = V_1 \ominus V_0 \tag{2.25}$$

One can see that W_0 is the linear subspace of $L_2(\mathbb{R})$ spanned by a certain orthogonal basis. The following function:

$$\psi(x) = \begin{cases} -1, & x \in (0, 1/2] \\ 1, & x \notin (1/2, 1] \end{cases} \tag{2.26}$$

starting the basis in W_0, which is the system $\{\psi_{0,k}\}$ where

$$\psi_{0,k} = \psi(x - k), \quad k \in \mathbb{Z}$$

The graph of Haar functions is presented on Fig.(2.2). We will use names: "mother wavelet" for function $\psi(\cdot)$ and " father wavelet" or "scaling function" for $\phi(\cdot)$. From(2.25) we have $V_1 = V_0 \oplus W_0$ the construction can be easily extended to every V_i as:

$$V_{i+1} = W_i \oplus V_i = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_i \tag{2.27}$$

It is also can be proved that $\bigcup_i V_i$ is dense in $L_2(\mathbb{R})$, and

$$L_2(\mathbb{R}) = V_0 \oplus \bigoplus_{i=0}^{\infty} W_i \tag{2.28}$$

The equation(2.28) means the for every function $f(\cdot)$ from spase $L_2(\mathbb{R})$, the following expansion takes place:

$$f(x) = \sum_k \alpha_{0k}\phi_{0k}(x) + \sum_{j=0}^{\infty}\sum_k \beta_{jk}\psi_{jk}(x) \tag{2.29}$$

where α_{0k} and β_{jk} are coefficients of this expansion.

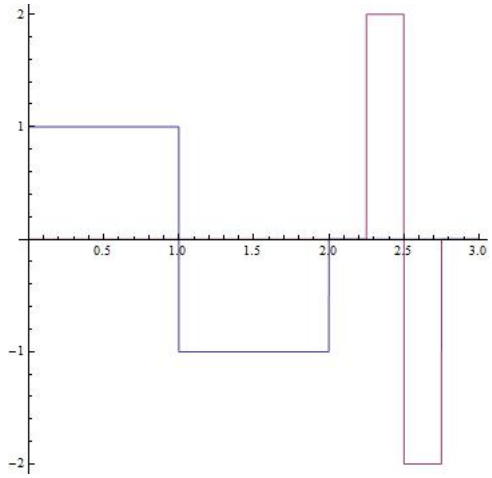


Figure 2.2: Graphs of $\psi_{-1,0}$ and $\psi_{9,2}$

2.2.1.2 Mexican Hat Wavelets

In many numerical applications, the orthogonality of the translated dilates $\psi_{j,k}$ is not vital. Mexican Hat wavelet(MHW) is the one of the non orthogonal wavelets. MH originates from the negative , normalized second derivative of the Gaussian function.The formula for MHW is

$$\psi(t,\sigma) = \frac{2}{\sqrt{3\sigma\sqrt{\pi}}}\left(1-\frac{t^2}{\sigma^2}\right)\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{2.30}$$

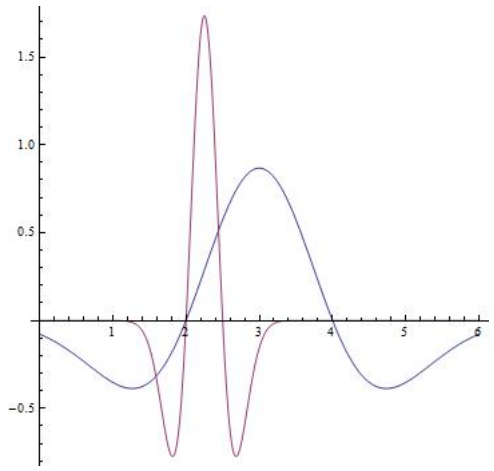


Figure 2.3: Graphs of the MHW functions $\psi_{0,3}(t,1)$ and $\psi_{9,2}(t,1)$

The graph of Mexican Hat Wavelet functions is presented on Fig.(2.3).

2.2.1.3 Wavelet thresholding

Sometimes wavelet estimators may produce a bit of spikes. This reflects the fact that unnecessary high oscillations are included. Therefore, it is natural to introduce the selection procedure for the coefficients in wavelet expansion(2.29). Most natural would be to suppress the smallest coefficient by introducing a threshold. There are several thresholdning procedures, the most common among them are two: the *soft* and the *hard* thresholdings.

- Soft thresholding is when we are replacing all β_{jk} in(2.29) by:

$$\hat{\beta^S_{jk}} = (|\hat{\beta_{jk}}| - t)_+ \text{ sign}(\hat{\beta_{jk}}) \tag{2.31}$$

where t is some threshold. This estimator has also name:*wavelet shrinkage* estimator.

- Hard thresholding is when we are replacing all β_{jk} in(2.29) by:

$$\hat{\beta_{jk}^{H}} = \hat{\beta_{jk}} I\{|\hat{\beta_{jk}}| > t\} \tag{2.32}$$

The common problem of thresholding is how to chose a proper value of threshold t ([26]). The threshold t must be chosen just above the maximum level of noise. Assuming the Gaussian white noise with variance σ^2, it can be proved that the maximum amplitude of the noise has a very high probability of being just below

$$T = \sigma\sqrt{2\log_e N}$$

, where N is the number of coefficients. The fact that the threshold increases with N is due to the tail of Gaussian distribution and the increasing of the amplitudes of the coefficients for large values of N.

CHAPTER 3: METHODS AND RESULTS

3.1 Regularization

Consider regression problem

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \tag{3.1}$$

where $\mathbf{X} \in \mathbf{R}^{n \times m}$ is a design matrix, $\mathbf{y} \in \mathbf{R}^n$ is the data we want to fit and $\boldsymbol{\beta} \in \mathbf{R}^m$ is the vector of coefficients. The solution is the well known least square estimator$\hat{\boldsymbol{\beta}}$ and is given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

It is also known (see, e.g., [36]) that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, thus

$$E||\hat{\boldsymbol{\beta}}||^2 = ||\boldsymbol{\beta}||^2 + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Therefore, in the case of ill-conditioned $\mathbf{X}^T\mathbf{X}$ one will have inflated estimation error, and as consequence, the poor prediction properties. There are several methods of regularization designed to mitigate the problem, from which we will consider Ridge regression and Lasso.

3.1.1 Ridge Regression

Ridge regression has been used as an alternative estimation method in multiple linear regression models when there is multicollinearity among the covariates. With the multicollinearity, the ridge type estimator is suggested because it has a smaller total mean square error (MSE) than the maximum likelihood estimator. When the multicollinearity is large, the reduction in MSE can be

significant. The ridge regression estimator was first proposed by Hoerl and Kennard [17] who replaced problem (3.1) by the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \tag{3.2}$$

The solution is found by the method of Lagrange multipliers and yelds the following rigde estimator:

$$\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

The "ridge" addition to the diagonal of $\mathbf{X}^T\mathbf{X}$ clearly reduces the high conditional numbers of the matrix $\mathbf{X}^T\mathbf{X}$. It can be proved(see, e.g., [36] or [14]) that the quantity $\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}$ is always invertable and there is always a unique solution $\hat{\boldsymbol{\beta}}^{ridge}$. Introducing the linear estimator:

$$\mathbf{H}_{ridge} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$$

which is analogous to "hat" matrix in ordinary regression, the degrees of freedom can be defined as $df = \text{Tr}(H_{ridge})$. The degrees of freedom df are used in calculation of the Akaike information criterion (AIC).

$$AIC = n\log(RSS) + 2df$$

the minimum of which helps to choose the value of λ.

The question now is how to apply the formalism above to our case of time-dependent Cox log-likelihood? Ridge regression is a special case of the penalized likelihood approach was studied by Huang and Harrington [19]. The most elegant solution is to use the following trick from [38]. Consider the following Cholesky decomposition of the negative second derivative of the log-likelihood:

$$\nabla^2 l(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{X} \tag{3.3}$$

where $\nabla = -\frac{\partial}{\partial \boldsymbol{\beta}}$ and $\nabla^2 = -\frac{\partial}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ and set up the pseudo response vector:

$$\mathbf{Y} = (\mathbf{X}^T)^{-1} \left\{ \nabla^2 l(\boldsymbol{\beta}) \boldsymbol{\beta} - \nabla l(\boldsymbol{\beta}) \right\} \tag{3.4}$$

Now we can rewrite Cox proportional hazards penalized optimization problem similar to equation (3.2) :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \tag{3.5}$$

where **X** and **Y** are understood in terms of (3.3) and (3.4). That method can be applied not only to ridge but also to another types of penalties. The next problem is the selection of the thresholding parameter λ.

3.1.2 Lasso

By the Cox proportional hazards model, the hazard function for the patient i is given as

$$\lambda_i(t) = \lambda_0(t) \exp(x_i^T \beta)$$

Then the partial log-likelihood can be expressed as:

$$l(\beta) = \sum_{i=1}^{N} \delta_i (x_i^T \beta - \log(\sum_{j \in R_i} \exp(x_i^T \beta))$$

where R_i is the set of indices of the patients in the risk set at time t_i. For the cases that impose serious collinearity problem, instead of maximizing the partial likelihood directly, Tibshirani [33]

proposed to estimate β under the L_1-constraint subject to $||\beta||_1 \leq s$:

$$\hat{\beta} = \arg\max l(\beta)$$

or, equivalently by the Lagrange multiplier version:

$$\hat{\beta} = \arg\min\{-l(\beta) + \lambda||\beta||_1\} \quad \text{with } \lambda > 0.$$

for $\lambda > 0$.

We will consider in detail the shooting algorithm for lasso, reported first by Fu [11] The description will follows the lines of the tutorial of Pendse [29] Since we are able to reduce the log-likelihood optimization problem to the least squares problem (see equations (3.4) and (3.5)). Without loss of generality the following objective function will suffice:

$$h(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \tag{3.6}$$

Denote $textbfX = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$, $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_p]$, $\mathbf{X}^{(-i)} = [\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_p]$ and $\boldsymbol{\beta}^{(-i)} = [\beta_1, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_p]$. Denoting $\mathbf{y}_i = \mathbf{y} - \mathbf{X}^{(-i)}\boldsymbol{\beta}^{(-i)}$, the following representation will take place:

$$h(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 = \frac{1}{2}||\mathbf{y}_i - \mathbf{x}_i\beta_i||_2^2 + \lambda||\boldsymbol{\beta}^{(-i)}||_1 + \lambda|\beta_i|$$

Replace problem (3.6)by the series of one-dimensional optimization problems (3.7) and carry out minimization for $i = 1, \cdots, p$, in a loop till process converges.

$$\beta_i = \underset{\beta_i}{\text{argmin}} \left\{ \frac{1}{2}||\mathbf{y}_i - \mathbf{x}_i\beta_i||_2^2 + \lambda||\boldsymbol{\beta}^{(-i)}||_1 + \lambda|\beta_i| \right\} \tag{3.7}$$

.

3.2 Denoising and optimization

Another application of FDA which was under study in current work, is the low rank tensor denoising. The recent progress in technology has made possible a multitude of novel applications, which typically require large amount of multidimensional data, such as large-scale images, 3D video sequences, and neuroimaging data (see, e.g., [4], [32]). To match the data dimensionality, tensors (also called multiway arrays) have been proven to be a natural and efficient representation for such massive data. There is a difficulty in a regression setup, of treating the clinical outcome as a response and treating images as the covariates. Indeed, classical regression methods use covariates presented in a vector form. Naively turning an image array into a vector is clearly an unsatisfactory solution. For instance, (see, e.g., [39]), with a typical anatomical MRI image of size $256 \times 256 \times 256$, it implicitly requires $256^3 = 16,777,216$ regression parameters. Both computability and theoretical properties of the classical regression models are compromised by this ultra-high dimensionality. It is natural to attempt the solutions which adpplies the FDA (see, e.g.,f Reiss and Ogden [30])

Our study is motivated by the analysis of Dynamic Contrast medical imaging data when one has to observe time changes of some organ recorded as a series of images. From mathematical point of view this problem can be viewed as a tensor recovery problem (see, e.g., [32]) for an extensive description. We will follow notation, from [12] which also contains an appropriate algorithm for the problem in hand.

3.3 Study of gender and age-specific survival rates

3.3.1 Wavelet model for hazard and survival function

The objective of our research is to check a hypothesis that the hazard functions for male and female cancer patients differs from each other, depending on the ages of the individuals. It is well known that probability of survival decreases when one's age increases, however, this general phenomenon may not be valid for relatively young individuals in the presence of a life-threatening disease like lung cancer. We attempt to use wavelets to separate time-dependent part of the hazard function is (2.13).

Wavelets have been successfully applied in the areas of signal and image processing, physics, etc. They also proved to be useful in nonparametric statistics (see, e.g. [1]:[3]) The known advantage of using wavelet basis versus other type of bases is that it accommodates jump discontinuities in the underlying function. Wavelets allow one to investigate behaviour of a spatially inhomogeneous function without any assumptions on the size or location of the intervals of smooth (or stable) behaviour.

Using Haar wavelet basis, generated by a scaling function $\phi(x)$ given by equation (2.24) and by shifts and dilations of the mother wavelet $\psi(x)$ given by equation (2.26), we obtain:

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \qquad j = 0, 1, ..., \infty,$$
$$k = 0, 1, ..., 2^{j-1}; \qquad x \in [0, 1] \tag{3.8}$$

In order to use wavelet basis for representation of a function, one has to scale the argument of the function to the interval (0, 1). For example, if we considered survival times of patients at ages between 40 and 75 , so that $(t/12 + A_i) \in (0, 1)$, we need to scale $t/12 + A_i$ to $(0, 1)$ by applying

transformation

$$x = [t/12 + A_i - 40]/35.$$

Here A_i is the age of i^{th} patient. In order to account for time-independent covariates as well as for dependence of the hazard rate on the current age of individual, we represented individual hazard functions as

$$h_i(t) = h_0(t) \exp(f(\frac{t}{12} + A_i)) \exp(\vec{\mathbf{X}}_i \vec{\beta}) \tag{3.9}$$

where A_i the age of individual i at the time of diagnosis, t is survival time in months, $\vec{\mathbf{X}}_i$ is the row vector of time independent covariates of an individual i and $\vec{\beta}$ is a column vector of coefficients. Note that $\frac{t}{12} + A_i$ is the age of individual i when his survival time is t months, so that function $f(\frac{t}{12} + A_i)$ represents the relationship between survival time and current age of a patient. Since, the functional form of $f(\cdot)$ is unknown, we estimate function $f(\cdot)$ non-parametrically. In particular, to take advantage of both modern techniques of non parametric statistics and software which allows to estimate coefficients in the Cox regression model (2.12), we represent $f(\cdot)$ using wavelet basis. Using wavelet basis (3.8) and equation(2.29), any function f(x) can be written as

$$f(x) = a\phi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j - 1} b_{jk} \psi_{jk}(x) \tag{3.10}$$

however, since the resolution of the data is limited, one usually applies

$$f(x) = a\phi(x) + \sum_{j=0}^{J} \sum_{k=0}^{2^j - 1} b_{jk} \psi_{jk}(x) \tag{3.11}$$

where J is the maximum resolution level. The average survival time for our patients was 28 months. Based on this, we have chosen $J = 4$, which gives us 31 time dependent coefficients.

Cox model is very popular also because of its simplicity - the time dependent parts of the hazard function are canceled from the partial likelihood. However, using the two dimensional time vari-

able allows us to keep them. We will illustrate that with a simple example. Consider an artificial sample of data:

Table 3.1: Artificial Data for time dependent likelihood

Survival Time	**Dead(1) or Censored(0)**	**Age of Diagnostics**
1	1	50
2	1	45
3	0	61
5	1	51
6	1	50

Assume that we will represent time-dependent part of hazard function using the following model:

$$\exp\left(\sum_{i=1}^{3}\beta_i\phi_i(T)\right) = \exp\left(\beta\cdot\mathbf{X}(T)\right)$$

Here T is the current age of the patient. Then the partial likelihood can be easily written as:

$$L(\beta) = \frac{\exp\left(\beta\cdot\mathbf{X}(51)\right)}{\exp\left(\beta\cdot\mathbf{X}(51)\right)+\exp\left(\beta\cdot\mathbf{X}(46)\right)+\exp\left(\beta\cdot\mathbf{X}(62)\right)+\exp\left(\beta\cdot\mathbf{X}(52)\right)+\exp\left(\beta\cdot\mathbf{X}(51)\right)}\cdot$$

$$\frac{\exp\left(\beta\cdot\mathbf{X}(47)\right)}{\exp\left(\beta\cdot\mathbf{X}(47)\right)+\exp\left(\beta\cdot\mathbf{X}(63)\right)+\exp\left(\beta\cdot\mathbf{X}(53)\right)+\exp\left(\beta\cdot\mathbf{X}(52)\right)}\cdot$$

$$\frac{\exp\left(\beta\cdot\mathbf{X}(56)\right)}{\exp\left(\beta\cdot\mathbf{X}(56)\right)+\exp\left(\beta\cdot\mathbf{X}(55)\right))}$$

It is clear that there are no cancellation of time-dependent functions.

3.3.2 Data description

For the project we used partial records from the Florida cancer data base. The data base contains data for cancer patients in the state of Florida between years 1986 and 2005. The lung cancer patients data contains almost 220,000 records. For the sake of privacy we only used information about age at diagnosis, stage and type of cancer, smoking patterns and treatment information for each of the patients. The age distribution of the patients is shown in the Figure3.1.

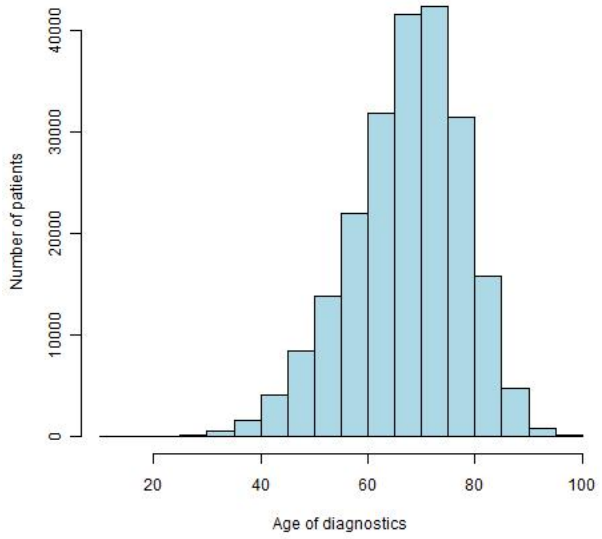


Figure 3.1: Overall age distribution of cancer patients

One can see that the distribution is heavily biased toward upper ages: 60-80 years old. Since our goal is to study survival of women of menopausal age ($40-61$ y.o.), we restricted our studies to the ages 35-65 years old which left us with 82049 cases from which 35858 records are for women and 46189 records are for men. The rationale for this data reduction is that survival patterns in children

and young adults and old people differ significantly from those for mature adults. In the course of the data analysis, we discovered that survival patterns vary greatly with the type of cancer.Type of cancer was encoded using "Histology" variable denoted as $H_i, i = 1 \cdots 8$. In the course of the data analysis, we discovered that survival patterns vary greatly for the earlier (Stages 0, 1 and 2) and the later (Stages 3-9) of cancer and they depend on the type (histology) of cancer as well. Thus, from the start we have split out data into four categoriess:

- Early Stages, females - 8943 records.
- Earlier stages, males - 9448 records.
- Advanced Stages, females - 26915 records
- Advanced Stages, males - 36741 records

3.3.3 Preliminary results

To simplify analysis we've introduced new categorical variable "AgeGrade" with the help of which we divided patients onto 8 groups according age of diagnostics. The coding details are given in the Table 3.2.

At early phase of analysis we fit Cox proportional hazards model with the help of proprietary software SPSS, using the following model:

$$SurvMonths = SurvMonths[Status = Dead](Race, Age, YBirth, Stage, Tobacco, AgeGrade)$$

The Chi-square value is 8650 for 6 degrees of freedom which makes the p-value close to zero which also indicates the validity of assumptions of the Cox model. We have plotted cumulative hazards for all for groups of patients separated by age groups (see Figures 3.3, 3.4, 3.5, 3.6) . For

earlier stages of cancer, the dependance on age is pretty regular and intuitive, with the exception that the 4^{th} age group patients (ages 45-50) have higher survival probability than 3^{d} age group (ages 40-45) for both sexes. Taking age group 1 (age 35) out from consideration the following dependence of the cumulative hazards on age is observed for both sexes:

$$2 < 4 < 3 < 5 < 6 < 7$$

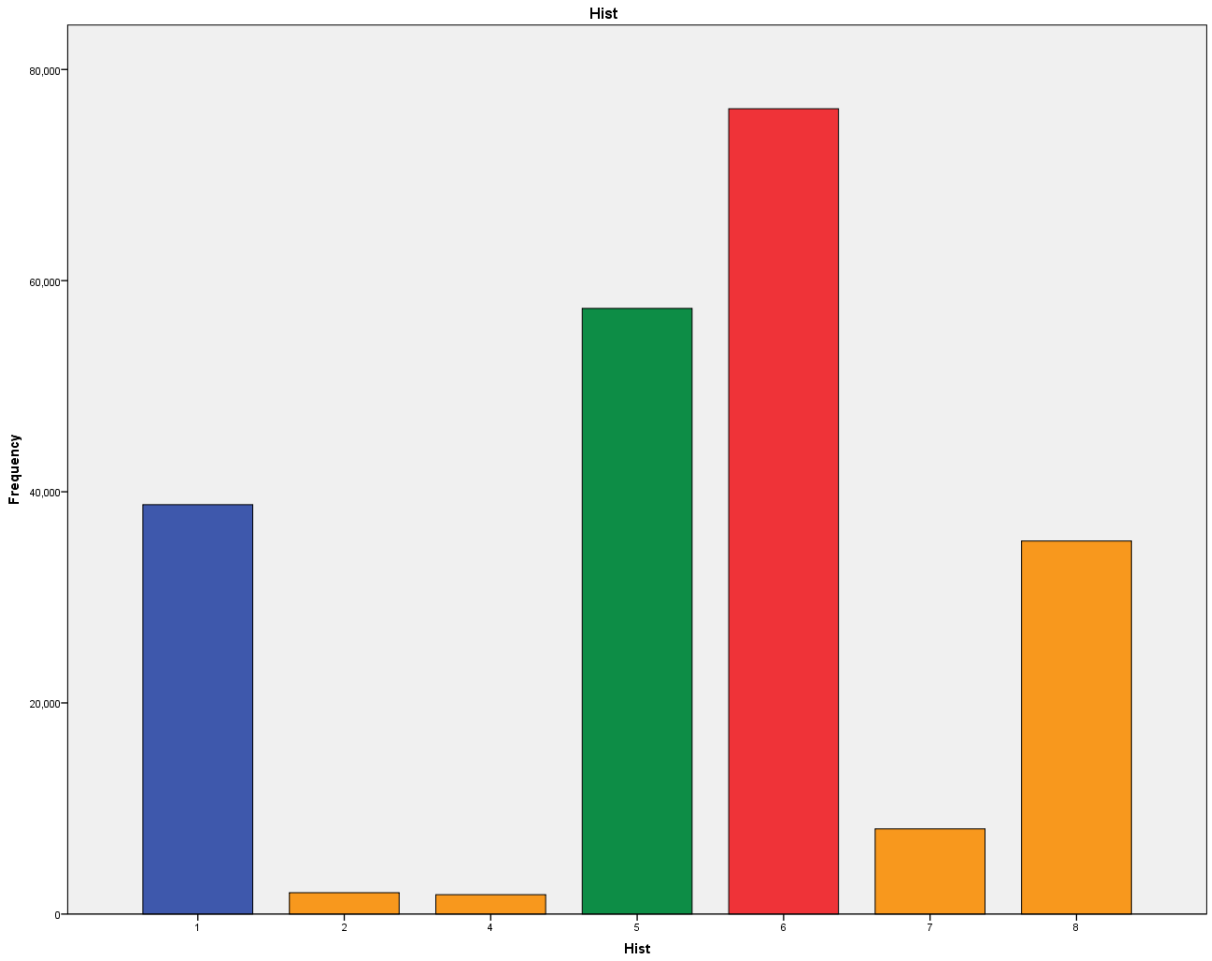


Figure 3.2: Distribution according type of cancer

Table 3.2: Age groups coding

Coding	Age of Diagnostics
Age grade1	age ≤ 35
Age grade2	$35 <$ age ≤ 40
Age grade3	$40 <$ age ≤ 45
Age grade4	$45 <$ age ≤ 50
Age grade5	$50 <$ age ≤ 55
Age grade6	$55 <$ age ≤ 60
Age grade7	$60 <$ age ≤ 65

As for advanced stages of cancer (Stages 3-9), one can see that contrary to common understanding of hazard rate increasing with age female patients show quite irregular dependence on age, but male patiens show the same pattern as for earlier stages. (see Figures 3.5 and 3.6 below). The list of age groups ordered according to increasing hazard rate for females:

$$3 < 2 < 6 < 5 < 4 < 7$$

and for males:

$$2 < 4 < 3 < 5 < 6 < 7$$

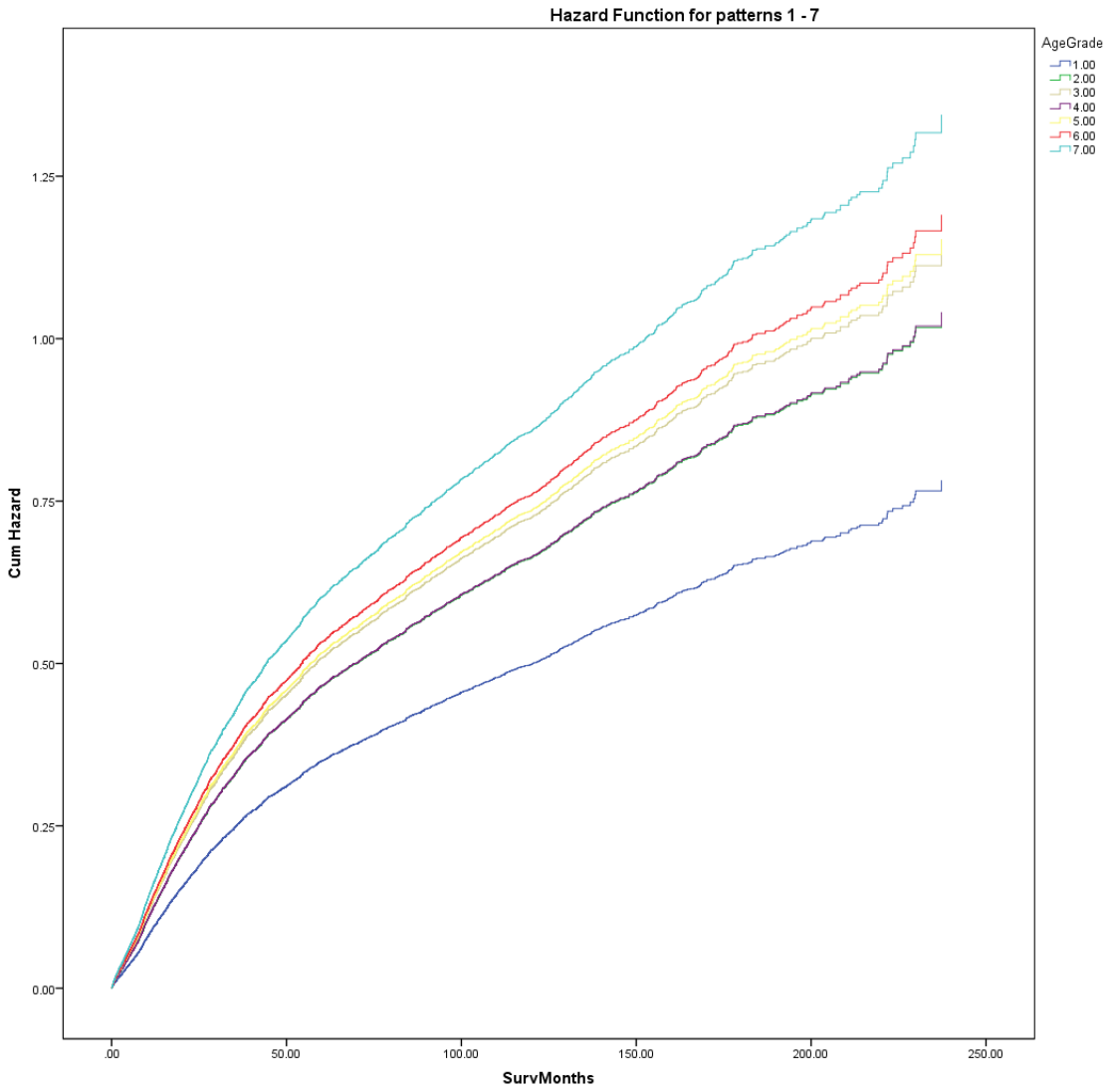


Figure 3.3: Hazard For Early Stages of Cancer (females)

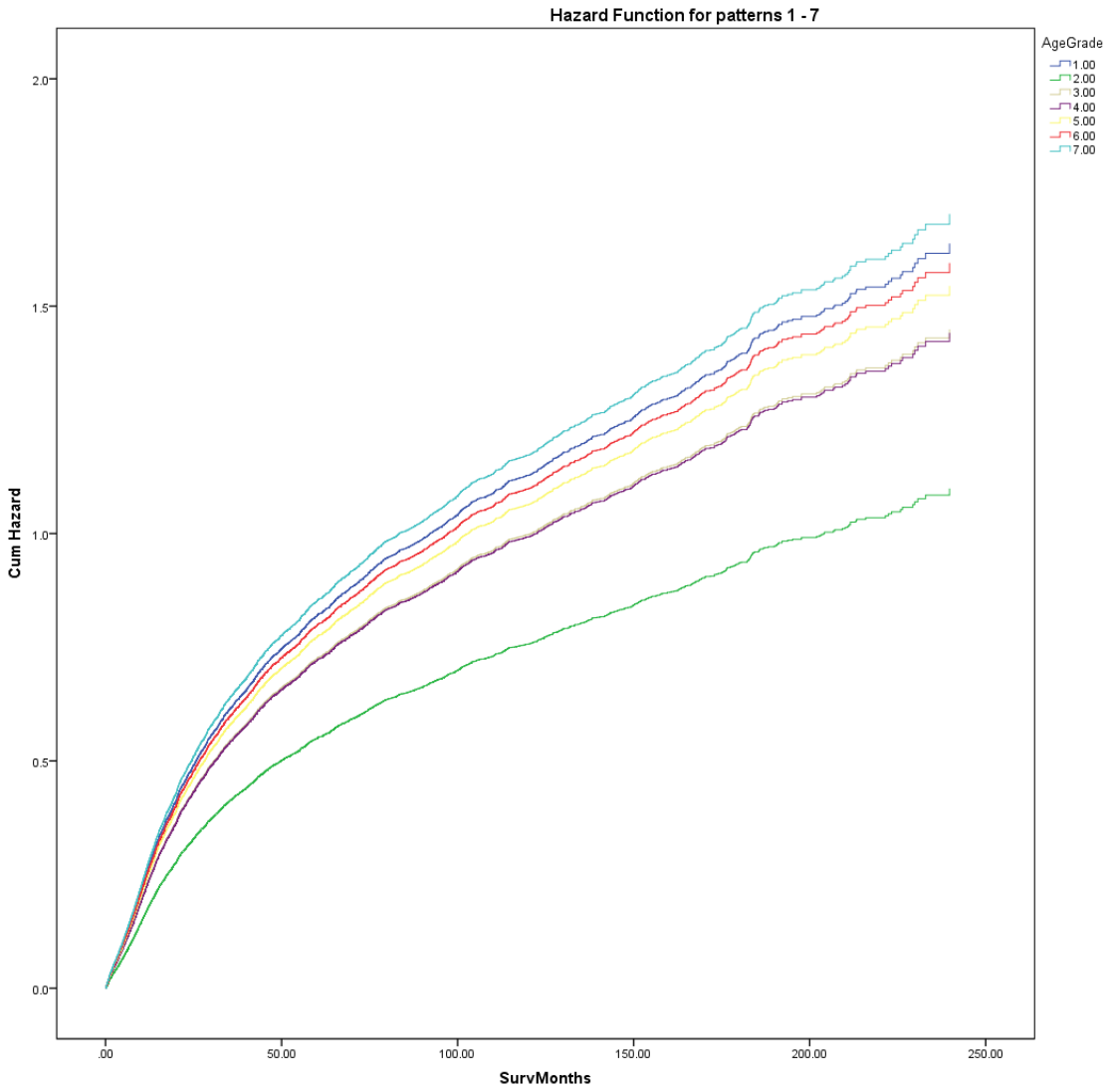


Figure 3.4: Hazard For Early Stages of Cancer (males)

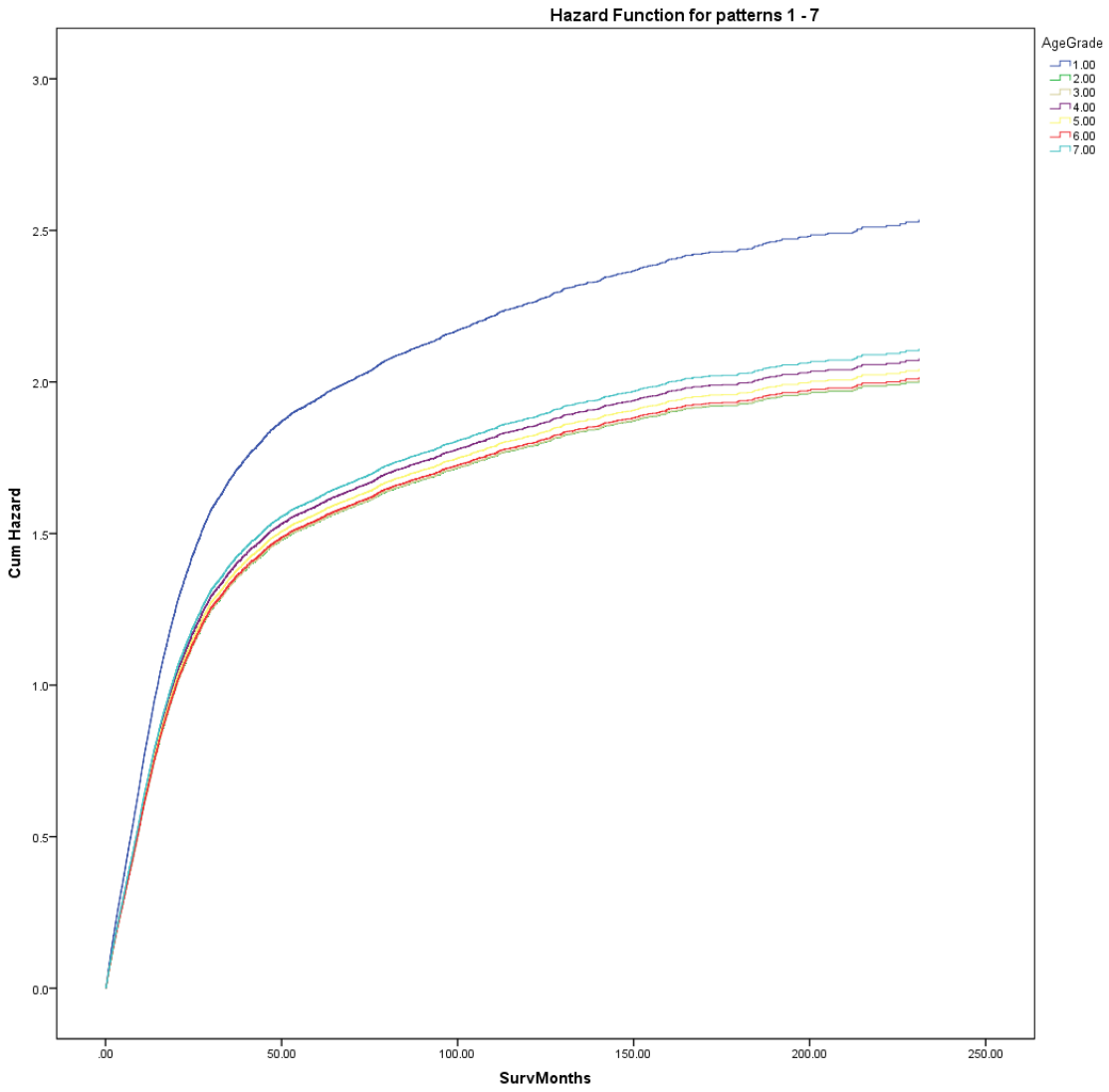
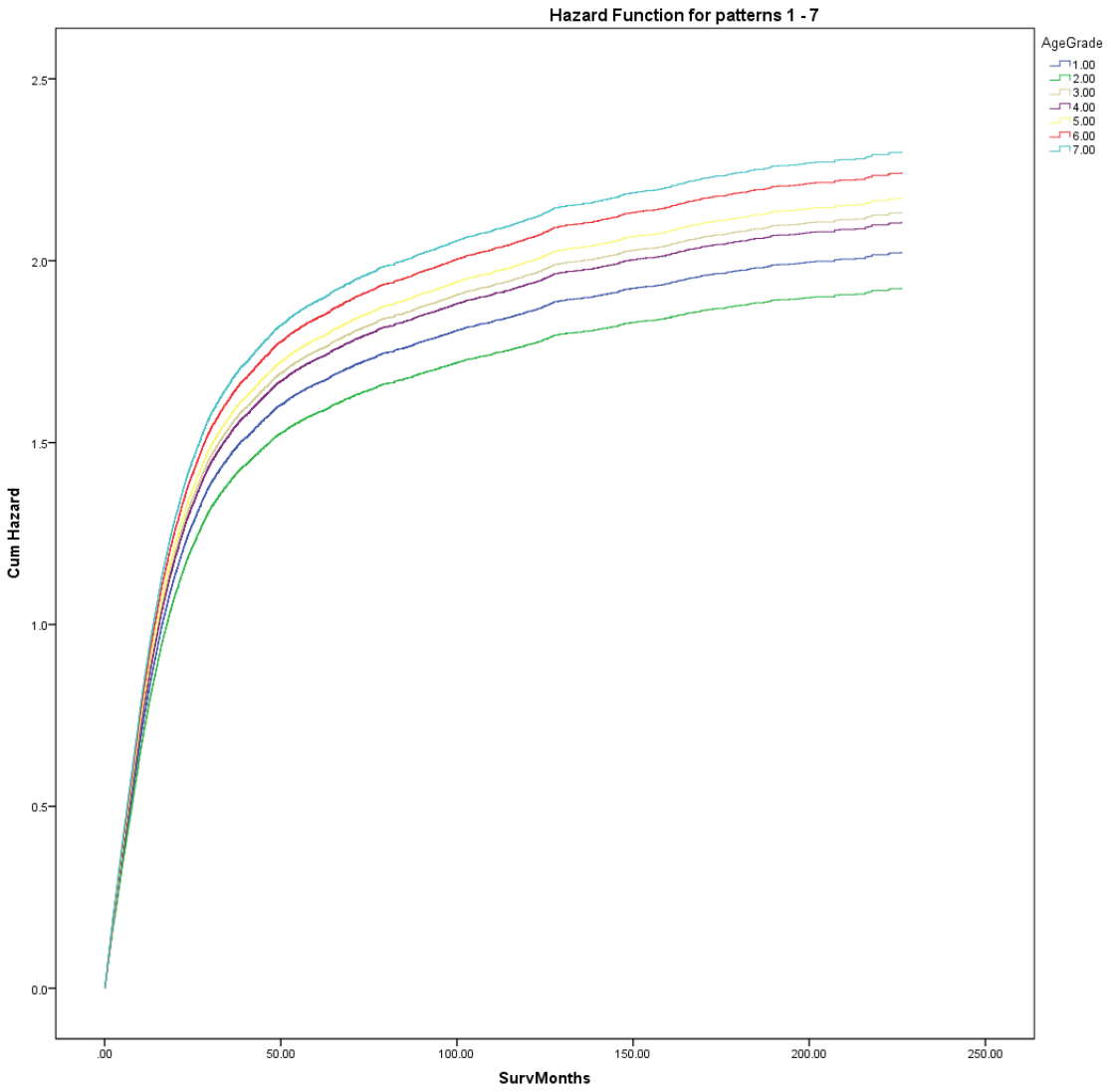


Figure 3.5: Hazard For Advanced Stages of Cancer(F)



]

Figure 3.6: Hazard For Advanced Stages of Cancer(M)

The coding is decrypted in the Table 3.3. The distribution of patients according to the type of cancer is shown in the Figure 3.2. The majority of data corresponds to cancer types: $H_6(35\%)$, $H_5(25\%)$, $H_1(18\%)$ and $H_8(16\%)$ or together for 94% of all data.

Table 3.3: Types of cancer and their code names

Coding	Type of Cancer
H1	Neuroendocrine/small cell carcinoma
H2	Neuroendocrine/carcinoid tumor
H3	Neuroendocrine/atypical carcinoma
H4	Neuroendocrine/mixed tumors
H5	Squamous cell carcinoma
H6	Adenocarcinoma
H7	Bronchido-alveolar carcinoma
H8	Other NSCLC

The graphs of cumulative hazard rates have one thing in common: each curve has the negative second derivative. Since hazard function is just the derivative of corresponding cumulative hazard, hazard function is decreasing. Since hazard can't be negative, the safe assumption would be that there exists a limiting value of hazard which can be different for each age group. Based on this heuristic arguments, we have a justification for fitting the right end of the cumulative hazard graph with a linear function. The left side of each graph of a cumulative hazard function is increasing fast initially with the rate of growth decreasing afterwards. For this reason, the following approach was employed: cumulative hazard function was fit with a linear function for the larger survival times and with a Michaelis-Menten function of the form ($\frac{c \cdot t}{d+t}$) for the smaller times. Therefore we fit the cumulative hazard function by $\Lambda(t)$ of the following form:

$$\Lambda(t) = \begin{cases} a \cdot t + b, & \text{if } t \leq t_{st} \\ \frac{c \cdot t}{d+t}, & \text{if } t > t_{st} \end{cases} \tag{3.12}$$

where the a, b, c, d and t_{st} are positive constants satisfying the following conditions:

$$\begin{aligned} c &= \frac{(b+a\cdot t_{st})^2}{b} \\ d &= \frac{a\cdot t_{st}^2}{b} \end{aligned}$$

There is the additional rationale behind choosing the Michaelis-Menten function for fitting: the equation of Michaelis-Menten is the basic equation of the enzymatic kinetics. It describes the dependence of the rate of catalyzed reaction on the concentration of the substrate. That equation has two parameters: the saturation value - which we call "assumed cumulative hazard", and "half-way" constant which we call "the critical time" and interpret as the time which it takes to decrease the mortality rate from the initial value to some saturated value, which is more descriptive for the advanced stages of cancer.

Fitting the cumulative hazard function by $\Lambda(t)$ allows us to estimate the critical time at which survival patterns change as well as the parameters of the survival time. The fitting is carried out using the least squares and produces very accurate results for all types of cancer and for all age groups. See the Figure(3.3.3) The very first look on the results reveals the basic features of the survival. Females have in general lower values of hazard than males. The transition process is more abrupt in the case of advanced cancer that in earlier stages. Somehow counter intuitive results indicates that saturated hazard rate is higher for earlier stages, while assumed cumulative hazard shows the opposite tendency. The "critical time" parameter appears to indicate the duration of the renewal process and it is in inverse relations with a risk to die. For advanced stages it is about one year and 3-4 years for early stages.

The survival patterns differ significantly for the types of cancer but is is hard to see any gender related differences so far.

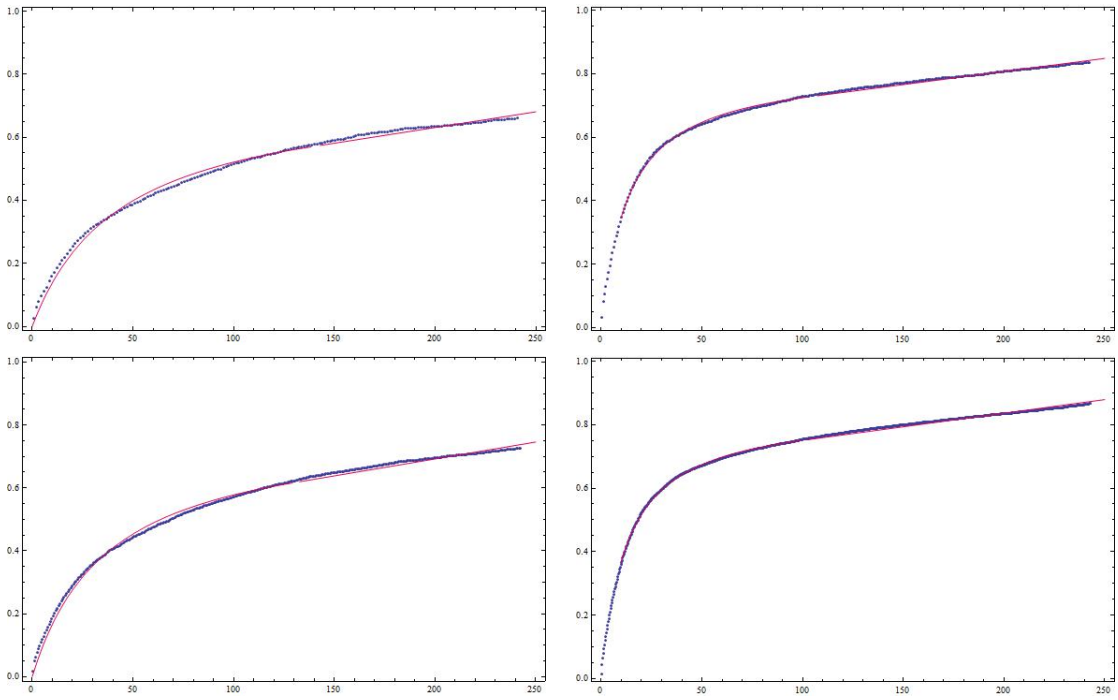


Figure 3.7: Cumulative hazard rate function fit for the squamous cell carcinoma patients. Left: Early Stages. Right: Advanced Stages. Top: Females. Bottom: Males.

3.3.4 Results of the analysis using the time-varying Cox regression model.

Since the advanced stages of lung cancer are associated with the very short survival time, we restricted our attention to patients who have stages 1 or 2 of cancer. This operation left us with

Table 3.4: Results of fit of squamous cell carcinoma patients

Gender	Stage	$a \cdot 10^{-4}$	critical time	assumed cum.hazard
F.	Early	9.95	44.3	0.75
F.	Advanced	8.25	13.5	0.82
M.	Early	10.76	37.8	0.80
M.	Advanced	8.56	12.7	0.85

only 26% of the original data (57147 of the 219685 records).

We also separated our data according to the sex and the type of cancer. In the dissertation we are presenting three most common cancer types: Neuroendocrine/small cell carcinoma(H1), Squamous cell carcinoma (H5) and Adenocarcinoma (H6).

Using Haar wavelets as our basis we've constructed the following representation of the hazard function:

$$f(t) = \sum_{j=0}^{5} \sum_{k=0}^{2^j-1} \boldsymbol{\beta}_{j,k} \textbf{HaarPsi} \left(\frac{2^j(t - Age_{min})}{Age_{max} - Age_{min}} - k \right)$$

where **HaarPsi**$(\cdot)$ is given by formula (3.8). We have generated vectors of functional values for all 6 categories and then Cox proportional hazard regression was carried out. That procedure let us have estimations of coefficients $\boldsymbol{\beta}_{j,k}$. After removing non significant coefficients the hazard functions were reconstructed. The results are shown in the Tables (3.4-3.9).

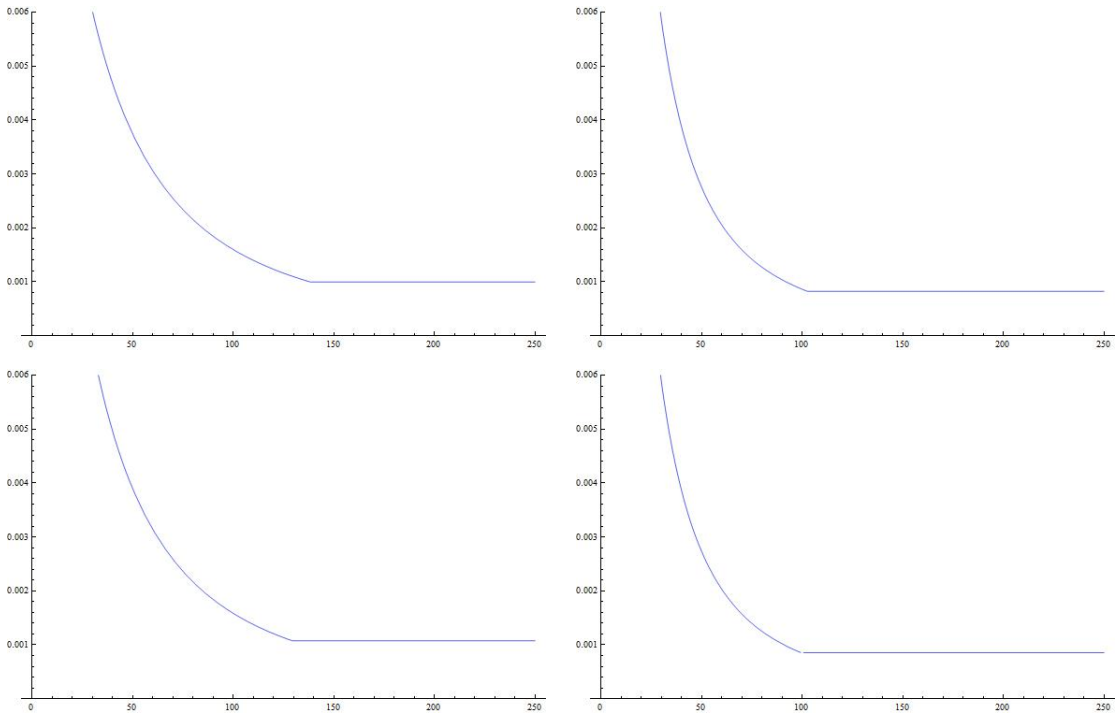


Figure 3.8: Hazard of squamous cell carcinoma patients. Left: Early Stages. Right: Advanced Stages. Top: Females. Bottom: Males.

Table 3.5: Wavelet coefficients for female patients of Neuroendocrine/small cell carcinoma

Earlier			Advanced		
$\boldsymbol{\beta}$	Value	Ages covered	$\boldsymbol{\beta}$	Value	Ages covered
$\boldsymbol{\beta}_{3,1}$	-0.267	$(42 \cdots 48)$	$\boldsymbol{\beta}_{1,0}$	-0.093	$(36 \cdots 60)$
$\boldsymbol{\beta}_{4,6}$	0.183	$(54 \cdots 57)$	$\boldsymbol{\beta}_{3,2}$	-.090	$(48 \cdots 54)$
$\boldsymbol{\beta}_{4,10}$	0.521	$(66 \cdots 69)$	$\boldsymbol{\beta}_{3,4}$	-0.071	$(60 \cdots 66)$
$\boldsymbol{\beta}_{5,10}$	-0.517	$(51 \cdots 52.5)$	$\boldsymbol{\beta}_{4,1}$	-0.262	$(39 \cdots 42)$
$\boldsymbol{\beta}_{5,23}$	0.990	$(70.5 \cdots 72)$	$\boldsymbol{\beta}_{4,2}$	-0.355	$(42 \cdots 45)$
			$\boldsymbol{\beta}_{4,3}$	-0.157	$(45 \cdots 48)$

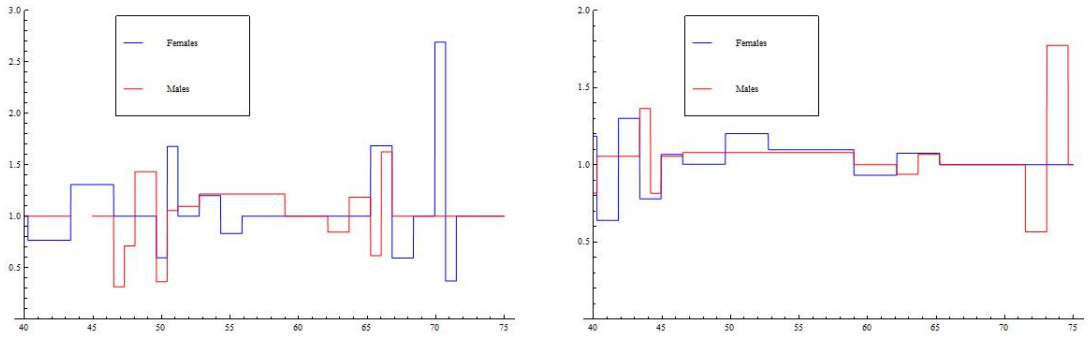


Figure 3.9: Haar wavelets fit of time dependent part of hazard function of Neuroendocrine/small cell carcinoma patients. Left: Early Stages. Right: Advanced Stages

Table 3.6: Wavelet coefficients for male patients of Neuroendocrine/small cell carcinoma

Earlier			Advanced		
β	Value	Ages covered	β	Value	Ages covered
$\beta_{2,1}$	-0.195	$(48 \cdots 60)$	$\beta_{1,0}$	-.075	$(36 \cdots 60)$
$\beta_{4,4}$	-0.554	$(48 \cdots 51)$	$\beta_{2,0}$	-.128	$(36 \cdots 48)$
$\beta_{4,5}$	-0.286	$(51 \cdots 54)$	$\beta_{4,9}$	0.065	$(63 \cdots 66)$
$\beta_{4,9}$	-0.168	$(63 \cdots 66)$	$\beta_{4,12}$	-0.572	$(72 \cdots 75)$
$\beta_{5,1}$	-0.989	$(37.5 \cdots 39)$	$\beta_{5,6}$	0.257	$(45 \cdots 46.6)$
$\beta_{5,6}$	0.579	$(45 \cdots 46.6)$			
$\beta_{5,8}$	-0.988	$(48 \cdots 49.6)$			
$\beta_{5,10}$	-0.533	$(51 \cdots 52.5)$			
$\beta_{5,20}$	-0.485	$(66 \cdots 67.5)$			

Table 3.7: Wavelet coefficients for female patients of Squamous cell carcinoma

Earlier			Advanced		
$\boldsymbol{\beta}$	Value	Ages covered	$\boldsymbol{\beta}$	Value	Ages covered
$\boldsymbol{\beta}_{2,1}$	.140	$(48 \cdots 60)$	$\boldsymbol{\beta}_{3,0}$	.530	$(36 \cdots 42)$
$\boldsymbol{\beta}_{2,2}$	.160	$(60 \cdots 72)$	$\boldsymbol{\beta}_{4,11}$	.376	$(69 \cdots 72)$
$\boldsymbol{\beta}_{3,5}$	.250	$(66 \cdots 72)$	$\boldsymbol{\beta}_{5,10}$	-.216	$(51 \cdots 52.5)$
$\boldsymbol{\beta}_{5,13}$	-.194	$(55.5 \cdots 57)$	$\boldsymbol{\beta}_{5,14}$	.125	$(57 \cdots 58.5)$
$\boldsymbol{\beta}_{5,22}$	-.435	$(69 \cdots 70.5)$	$\boldsymbol{\beta}_{5,15}$	.137	$(58.5 \cdots 60)$
$\boldsymbol{\beta}_{5,25}$	.638	$(73.5 \cdots 75)$	$\boldsymbol{\beta}_{5,19}$	.109	$(64.5 \cdots 66)$

Table 3.8: Wavelet coefficients for male patients of Squamous cell carcinoma

Earlier			Advanced		
$\boldsymbol{\beta}$	Value	Ages covered	$\boldsymbol{\beta}$	Value	Ages covered
$\boldsymbol{\beta}_{2,2}$	-.145	$(60 \cdots 72)$	$\boldsymbol{\beta}_{2,2}$	-.147	$(60 \cdots 72)$
$\boldsymbol{\beta}_{3,1}$	-.213	$(42 \cdots 48)$	$\boldsymbol{\beta}_{3,3}$	-.037	$(54 \cdots 60)$
$\boldsymbol{\beta}_{3,4}$	-.049	$(60 \cdots 66)$	$\boldsymbol{\beta}_{4,3}$	-.101	$(45 \cdots 48)$
$\boldsymbol{\beta}_{3,5}$	-.153	$(66 \cdots 72)$	$\boldsymbol{\beta}_{4,5}$	.068	$(51 \cdots 54)$
$\boldsymbol{\beta}_{4,8}$	.071	$(60 \cdots 63)$	$\boldsymbol{\beta}_{4,12}$	.510	$(72 \cdots 75)$
$\boldsymbol{\beta}_{4,10}$	-.171	$(66 \cdots 69)$	$\boldsymbol{\beta}_{5,6}$	-.173	$(45 \cdots 46.5)$
			$\boldsymbol{\beta}_{5,14}$	-.096	$(57 \cdots 58.5)$
			$\boldsymbol{\beta}_{5,26}$	.629	$(75 \cdots 76.5)$

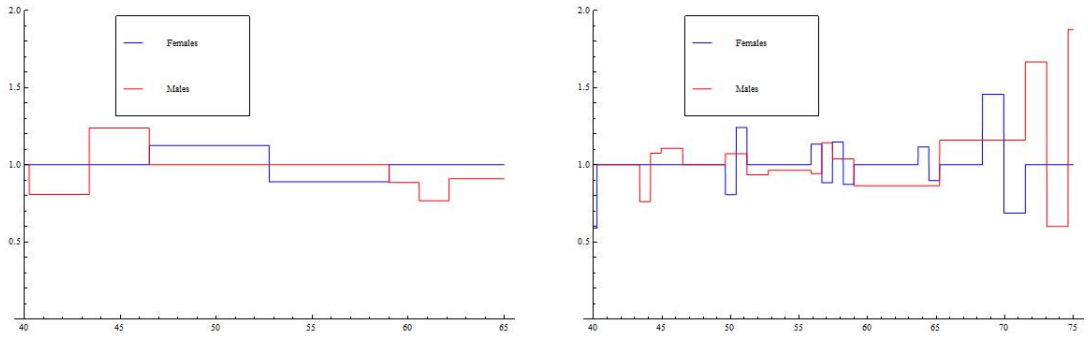


Figure 3.10: : Haar wavelets fit of time dependent part of hazard function of squamous cell carcinoma patients. Left: Early Stages. Right: Advanced Stages

Table 3.9: Wavelet coefficients for female patients of Adenocarcinoma

Earlier			Advanced		
β	Value	Ages covered	β	Value	Ages covered
$\beta_{3,3}$	-.099	$(54 \cdots 60)$	$\beta_{2,1}$	-.036	$(48 \cdots 60)$
$\beta_{3,4}$	-.138	$(60 \cdots 66)$	$\beta_{3,1}$	-.070	$(42 \cdots 48)$
$\beta_{4,4}$	.253	$(48 \cdots 51)$	$\beta_{4,11}$	.355	$(69 \cdots 72)$
$\beta_{4,9}$	-.085	$(63 \cdots 66)$	$\beta_{5,11}$	-.124	$(52.5 \cdots 54)$
$\beta_{4,10}$	.114	$(66 \cdots 69)$	$\beta_{5,13}$	.093	$(55.5 \cdots 57)$
$\beta_{5,2}$	1.254	$(39 \cdots 40.5)$	$\beta_{5,22}$	.236	$(69 \cdots 70.5)$
$\beta_{5,12}$	-.283	$(54 \cdots 55.5)$	$\beta_{5,25}$	-.721	$(73.5 \cdots 75)$

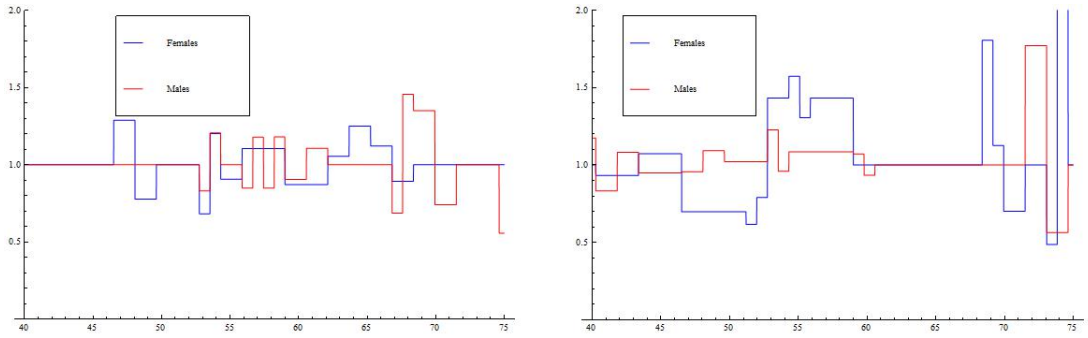


Figure 3.11: : Haar wavelets fit of time dependent part of hazard function of Adenocarcinoma patients. Left: Early Stages. Right: Advanced Stages

Table 3.10: Wavelet coefficients for male patients of of Adenocarcinoma

Earlier			Advanced		
β	Value	Ages covered	β	Value	Ages covered
$\beta_{4,8}$	-.101	$(60 \cdots 63)$	$\beta_{1,0}$	-.052	$(36 \cdots 60)$
$\beta_{4,11}$	.300	$(69 \cdots 72)$	$\beta_{2,1}$	-.030	$(48 \cdots 60)$
$\beta_{5,12}$	-.187	$(54 \cdots 55.5)$	$\beta_{4,1}$	-.213	$(39 \cdots 42)$
$\beta_{5,14}$	-.164	$(57 \cdots 58.5)$	$\beta_{4,2}$	-.131	$(42 \cdots 45)$
$\beta_{5,15}$	-.165	$(58.5 \cdots 60)$	$\beta_{4,4}$	-.066	$(48 \cdots 51)$
$\beta_{5,21}$	-.376	$(67.5 \cdots 69)$	$\beta_{4,12}$	.572	$(72 \cdots 75)$
$\beta_{5,26}$	-.586	$(75 \cdots 76.5)$	$\beta_{5,12}$	.123	$(54 \cdots 55.5)$
$\beta_{5,27}$	.706	$(76.5 \cdots 78)$	$\beta_{5,16}$	.069	$(60 \cdots 61.5)$

Since the wavelet coefficients are time dependent, the computation of Cox regression appears to be involved. The attempt to use already available software was successful for Haar wavelets. We used SPSS feature of introducing time-dependent covariates into the model. The attempt to do the

same with the Mexican Hat wavelets which are more appealing for Functional Data Analysis was not successful - since Mexican Hat wavelets are not orthogonal, we have encountered situation of highly correlated variables which leads to crashing of the algorithm. The problem can be avoided by using some kind of regularization. We were using two types of regularization: the ridge (L2 penalty) and the lasso (L1 penalty).

3.3.4.1 The ridge penalty

Algorithm 1 L2 penalized Cox regression

$\boldsymbol{\beta}$ = initial $\boldsymbol{\beta}_0$ and $\epsilon = 0.001$

while $||\boldsymbol{\beta}^{i+1} - \boldsymbol{\beta}^i|| > \epsilon$ **do**

calculate l; $\quad \nabla = -\frac{\partial l}{\partial \boldsymbol{\beta}}$; $\quad \nabla^2 = -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$

calculate $\mathbf{X}$ $\quad \nabla^2 = \mathbf{X}^T\mathbf{X}$

$\mathbf{Y} \leftarrow (\mathbf{X}^T)^{-1} \left\{ \nabla^2 l(\boldsymbol{\beta}^i)\boldsymbol{\beta}^i - \nabla l(\boldsymbol{\beta}^i \right\}$

$\lambda^{i+1} \leftarrow \frac{\text{trace}((\nabla^2 - \lambda^i)^{-1}\nabla^2)}{||\boldsymbol{\beta}^i||_2^2}$

$\boldsymbol{\beta}^{i+1} \leftarrow (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$

end while

The algorithm of the ridge regression can be summarized as follows- see Algorithm 1.

3.3.4.2 The lasso penalty

The lasso [33] is a regularized estimation approach that constrains the $L1$ norm of coefficient vector. Instead of the log-likelihood we deal with its penalized version:

$$l_{pen}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| \tag{3.13}$$

where the first term is the log-likelihood which is concave and twice differentiable. Thee second term, the penalty, is concave and continuous but is not differentiable at points where $\boldsymbol{\beta}_i = 0$. To avoid that difficulty, we follow the idea of Goeman [13]. Consider directional vector $\mathbf{v} \in R^p$ and the corresponding derivative:

$$l'_{pen}(\boldsymbol{\beta}, \mathbf{v}) = \lim_{t \to 0} \left(l_{pen}(\boldsymbol{\beta} + t\mathbf{v}) - l_{pen}(\boldsymbol{\beta}) \right) \tag{3.14}$$

That gradient can be defined for every $\boldsymbol{\beta}$ as the scaled direction of the steepest ascent.

$$g_i(\boldsymbol{\beta}) = \begin{cases} \mathbf{G}_i(\boldsymbol{\beta}) - \lambda \text{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ \mathbf{G}_i(\boldsymbol{\beta}) - \lambda \text{sign}(\mathbf{G}_i(\boldsymbol{\beta})) & \text{if } \beta_i = 0 \text{ and } |\mathbf{G}_i(\boldsymbol{\beta})| > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

where $\mathbf{G}_i(\boldsymbol{\beta})$ is the i^{th} component of the gradient of the non-penalized log-likelihood given in equation (2.15). The gradient is ,therefore, discontinuous at every point where $\beta_i = 0$. The second derivative of the penalized log likelihood can also can be understood in terms of the quadratic form of non penalized log likelihood:

$$l''_{pen}(\boldsymbol{\beta}, \mathbf{v}) = \mathbf{v}'\mathbf{H}\mathbf{v} \tag{3.16}$$

where components of the matrix $\mathbf{H}$ are given by equation (2.16).

Consider second order Taylor approximation in the direction of the gradient (3.15) given by:

$$l_{pen}(\boldsymbol{\beta} + t \cdot \mathbf{g}(\boldsymbol{\beta})) \approx l_{pen}(\boldsymbol{\beta}) + t \cdot l'_{pen}(\boldsymbol{\beta}, \mathbf{g}(\boldsymbol{\beta})) + \frac{t^2}{2} l''(\boldsymbol{\beta}, \mathbf{g}(\boldsymbol{\beta})) \tag{3.17}$$

One needs to understand that the equation above has a meaning only within a subdomain of continuity of the gradient (3.15) i.e. for $0 < t < t_{edge}$, where:

$$t_{edge} = \min_i \left\{ -\frac{\beta_i}{g_i(\boldsymbol{\beta})} : \text{sign } \boldsymbol{\beta}_i = -\text{ sign } g_i(\boldsymbol{\beta}) \neq 0 \right\}$$

The optimal value inside the continuity domain according to equation(3.17) is at:

$$t_{opt} = -\frac{l'_{pen}(\boldsymbol{\beta}, g(\boldsymbol{\beta}))}{l''_{pen}(\boldsymbol{\beta}, g(\boldsymbol{\beta}))}$$

we only can accept t_{opt} if it is smaller than t_{edge} The update for $\boldsymbol{\beta}$ then proceed as follows:

$$\boldsymbol{\beta}^{i+1} = \boldsymbol{\beta}^i + \min(t_{opt}, t_{edge})\mathbf{g}(\boldsymbol{\beta}^i) \tag{3.18}$$

The algorithm described by equation(3.18)does not involve computational expensive operations such as matrix inversion, but requires a large number of iterations for convergence. For us it is undesirable because we need to calculate time-dependent likelihood which is computationally expensive process. Therefore the update in the work of Goeman [13] was beneficial for us. Goeman proposed to use Newton-Raphson algorithm, which is known for its fast convergence rate, and to switch to update described by equation(3.18) when Newton-Raphson step has failed.

Algorithm 2 L1 penalized Cox regression

$\boldsymbol{\beta} =$ initial $\boldsymbol{\beta}_0$ and $\epsilon = 0.001$

while $||\boldsymbol{\beta}^{i+1} - \boldsymbol{\beta}^i|| > \epsilon$ **do**

calculate l; $\quad \mathbf{G} = \frac{\partial l}{\partial \boldsymbol{\beta}}$; $\quad \mathbf{H} = \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$

calculate $g_i(\boldsymbol{\beta}) = \begin{cases} \mathbf{G}_i(\boldsymbol{\beta}) - \lambda \text{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ \mathbf{G}_i(\boldsymbol{\beta}) - \lambda \text{sign}(\mathbf{G}_i(\boldsymbol{\beta})) & \text{if } \beta_i = 0 \text{ and } |\mathbf{G}_i(\boldsymbol{\beta})| > \lambda \\ 0 & \text{otherwise} \end{cases}$

calculate $l'_{pen} = \text{Abs}(\mathbf{g})$ and $l''_{pen} = \mathbf{e}_\mathbf{g}^T \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \mathbf{e}_\mathbf{g}$

calculate $t_{opt} = -\frac{l'_{pen}(\boldsymbol{\beta}, g(\boldsymbol{\beta}))}{l''_{pen}(\boldsymbol{\beta}, g(\boldsymbol{\beta}))}$ and $t_{edge} = \min_i \left\{ -\frac{\beta_i}{g_i(\boldsymbol{\beta})} : \text{sign } \boldsymbol{\beta}_i = -\text{sign } g_i(\boldsymbol{\beta}) \neq 0 \right\}$

$\boldsymbol{\beta}^{i+1} \leftarrow \boldsymbol{\beta}^i + \min(t_{opt}, t_{edge})\mathbf{g}(\boldsymbol{\beta}^i)$

end while

The results are presented on the Figures 3.13-3.15. There are no informative patterns regarding advanced stages of cancer. But for the earlier stages we can see quite distinct maximum in the hazard rate for female patients comparing to male patients for all types of cancer presented. The absolute value therefore is not large. This is the explanation why it was to hard to show the

dependence of the hazard rate on age. The values at maximum of hazard rates for female patients vary from 0.01 for small cell carcinoma to 0.04 for Squamous cell carcinoma and up to to 0.1 for Adenocarcinoma. Whether the change in hazard rates is due to menopause or not is not possible to state based just on statistical data, but definitely there is exist some king of dependence and we provided evidence for deeper investigations of the phenomena.

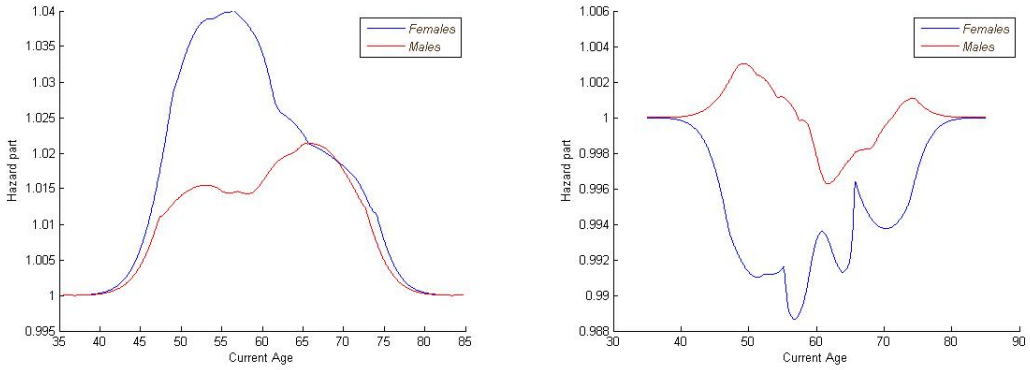


Figure 3.12: : Current age dependent part of hazard function for early (left) and advanced (right) stages of Squamous cell carcinoma

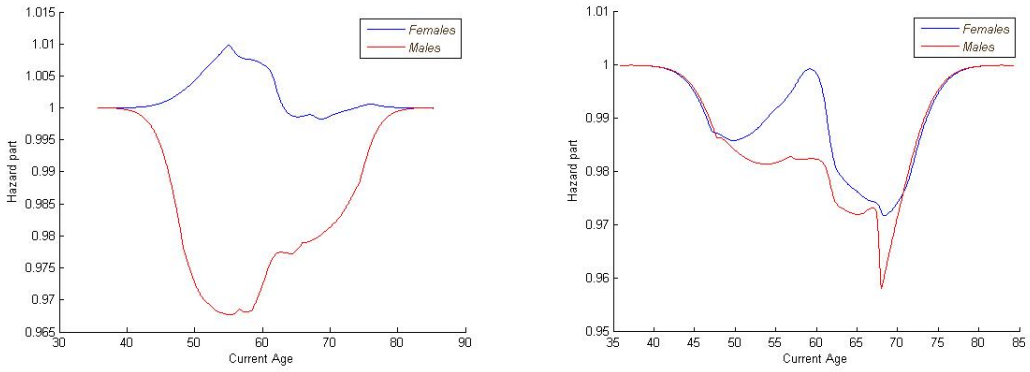


Figure 3.13: : Current age dependent part of hazard function for early (left) and advanced (right) stages of Neuroendocrine/small cell carcinoma

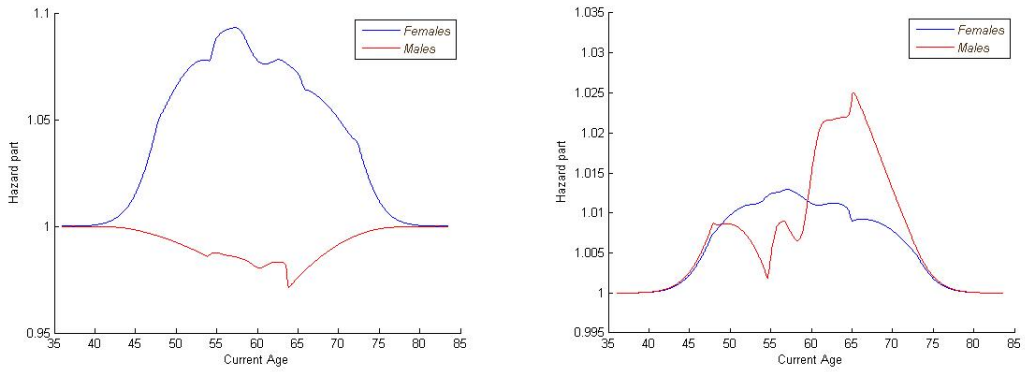


Figure 3.14: : Current age dependent part of hazard function for early (left) and advanced (right) stages of Adenocarcinoma

CHAPTER 4: ESTIMATION OF MATRIX-VARIATE FUNCTION.

4.1 Introduction

In vivo optical imaging exploits contrasts agents which interact with visible and near-infrared wavelengths in living tissues. Compared with another types of imaging like X-ray and magnetic resonance imaging (MRI), optical imaging has the major benefit of being able to exploit a rich palette of contrasts. Recently, considerable progress was achieved with Dynamic Contrast Enhanced (DCE) imaging. This method is used intensively in brain science and oncology. DCE has a great potential for cancer detection and characterization as well as for monitoring *in vivo* the effects of treatments.

The penetration of the contrast agent in living tissue is defined either by passive transport (diffusion) or active transport mechanisms. In each particular case the dynamic of contrast can shine light on deep internal cell processes and pathways such as receptor activity, antigen expression, endocytosis etc, see e.g. [35], [27], [22]. Intuitively it is clear that the change in the amount of contrast agent over time can be described by a continuous function, which opens opportunity for Functional Data Analysis methods. Consider the problem of estimation of a smooth matrix-variate function $f : [a, b] \to \mathbb{R}^{m_1 \times m_2}$ on the basis of its discrete noisy measurements

$$Y_i = f(t_i) + \epsilon\xi_i, \quad Y_i, f_i, \xi_i \in R^{m_1 \times m_2}, \ \ i = 1, \cdots, n. \tag{4.1}$$

Here, ξ_i are i.i.d. with matrix-variate normal distribution $\mathbb{N}(0, \mathbb{I}_{m_1} \times \mathbb{I}_{m_2})$ and function f is such that each of its components is a smooth function and, for each t_i, matrices $f(t_i)$ have low ranks. The objective is to recover the matrix-variate function f. Since observations Y_i, $i = 1, \cdots, n$, form a tensor, problem of estimating f is the problem of recovery of a sparse tensor from its noisy

observations. Denoting the 3-dimensional tensors of observations, values of function of interest and errors by Y, f and ξ, respectively, obtain the following model

$$Y = f + \epsilon\xi. \tag{4.2}$$

In the literature, there have been roughly three categories of solutions to establishing association between matrix/array covariates and clinical outcome [18]

1 Voxel-based methods, which take the image data at each voxel as responses and clinical variables such as age and gender as predictors, and then generate a statistical parametric map of test statistics or p-values across all voxels. Since the relationship between the correlation of voxels is not taken into consideration, this method we will not explore father.

2 FDA wa takes a one-dimensional function as predictor. Fitting such models commonly involves representing functions as a linear combination of basis functions which are either pre-specifie, or obtained from principal component decompositions.

3 Two step strategy: carrying out a dimension reduction followed by fitting a model.

Problems like (4.1) appear in image processing, medical imaging and computer vision. For example, in the case of Dynamic Contrast Enhanced medical imaging, one obtains a series of matrices where each of the matrices represents an image at a particular time instance and each of the pixel represents an amount of the contrast agent in a unit volume of tissue. Since the amount of contrast agent can only change gradually, each component of the matrix-variate function is itself a smooth function of time. Each of the images, however, does not necessarily form a smooth function since, for example, metastases of a tumor usually do not form a smooth continuous structure. Furthermore, since the amounts of contrast agent in a similar tissue types are proportional to each other,

one expects that many of the time-varying components of f are either equal or proportional to each other in value.

Due to these similarities, one can expect that if each of the functions is expanded over the same basis $\{\phi_j(t),\ j = 1, 2, \cdots\}$, the coefficients of the expansion should form low rank matrices with decreasing norms. For this reason, problem (4.2) is related to both the matrix regression and the linear regression problem.

4.2 Notations

Below, we provide a brief summary of the notation used throughout this chapter. Let $A \in \mathbb{R}^{m_1 \times m_2 \times M}$ be a 3-dimensional tensor and $B, B_1, B_2 \in \mathbb{R}^{m_1 \times m_2}$ be matrices. Denote the $k \times k$ identity matrix by $\mathbb{I}_k$. Denote elements of a matrix B by B^{j_1, j_2}.

- For any vector $\eta \in \mathbb{R}^p$, we denote the standard l_1 and l_2 vector norms by $\|\eta\|_1$ and $\|\eta\|_2$, respectively.
- For any numbers, a and b, denote $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.
- We define the *scalar product of matrices* $\langle B_1, B_2 \rangle = \text{Tr}(B_1^T B_2)$ where $\text{Tr}(\cdot)$ denotes the trace of a square matrix.
- For any matrix B denote by

$$\|B\|_1 = \sum_{j=1}^{m} \sigma_j(B) \quad \text{and} \quad \|B\|_2 = \left(\sum_{j=1}^{m} \sigma_j^2(B) \right)^{1/2} \quad \text{with} \quad m = m_1 \wedge m_2,$$

  respectively, the *trace*(*nuclear*), and *Frobenius* norms of matrix B. Here $\sigma_j(B)$ are the singular values of B in decresing order. Also, $\|B\| = \sigma_1(B)$ is the spectral norm of matrix

B.

- Denote the L^2 norm of tensor A by $\|A\|_2$:

$$\|A\|_2 = \left[\sum_{j_1=1}^{m_1}\sum_{j_2=1}^{m_2}\sum_{i=1}^{M}\left(A_i^{j_1,j_2}\right)^2\right]^{1/2}$$

- Denote Kronecker delta function by δ_{k_1,k_2}: $\delta_{k_1,k_2} = 1$ if $k_1 = k_2$, otherwise $\delta_{k_1,k_2} = 0$.

- For the tensor $\mathbf{T} \in \mathbb{R}^{l\times m\times k}$ and the matrix $\mathbf{M} \in \mathbb{R}^{l\times n}$ the tensor $\mathbf{O} \in \mathbb{R}^{n\times m\times k}$ is the following product:

$$\mathbf{O} = \mathbf{T} * \mathbf{M} = \sum_{i=1}^{l} T_i \otimes m_i$$

  where m_i is the i^{th} column of the matrix $\mathbf{M}$ and $T_i$is the i^{th} "slice" of the tensor $\mathbf{T}$ and the operation follows the following scheme:

  1 Unfold tensor $\mathbf{T}$ to matrix $\mathbf{MT} \in \mathbb{R}^{m\cdot k\times l}$

  2 Multiply $\mathbf{MT}$ and $\mathbf{M}$: $\mathbf{MO} = \mathbf{MT}\cdot\mathbf{M}$; $\quad \mathbf{MO} \in \mathbb{R}^{m\cdot k\times n}$

  3 Fold matrix $\mathbf{MO}$ into tensor $\mathbf{O}$

In what follows, we use the symbol C for a generic positive constant, which is independent of n, m_1, s and l, and may take different values at different places.

4.3 Estimation procedure and its risk

Consider a basis $\{\phi_j(t),\ j = 1, 2, \cdots\}$ on the interval $[a, b]$. Expand each of the components of f over this basis and approximate each slice $f(t_i)$ by

$$f_A(t_i) = \sum_{l=1}^{M} A_l \phi_l(t_i), \quad i = 1, \cdots, n, \tag{4.1}$$

where coefficients $A_l \in \mathbb{R}^{m_1 \times m_2}$, $l = 1, \cdots, M$, are matrices. Consider matrix $\Phi \in R^{M \times n}$

$$\Phi = \begin{pmatrix} \phi_1(t_1) & \cdots & \phi_1(t_n) \\ . & \cdots & . \\ \phi_M(t_1) & \cdots & \phi_M(t_n) \end{pmatrix}$$

and let $W = n^{-1}\ \Phi\Phi^T$ be matrix with elements

$$W^{j,k} = \frac{1}{n} \sum_{i=1}^{n} \phi_j(t_i)\phi_k(t_i)$$

. Let $f, f_A \in \mathbb{R}^{n \times m_1 \times m_2}$ and $A \in \mathbb{R}^{M \times m_1 \times m_2}$ be the 3-dimensional tensors of the values of functions $f(t)$ and $f_A(t)$ at points $t_i, i = 1, \cdots, n$, and of M matrix coefficients in the expansion (4.1). Then,

$$f_A = A * \Phi = \sum_{l=1}^{M} A_l \otimes \phi_l$$

where $\phi_l \in R^n$ is the l^{th} column of the matrix Φ. We obtain an estimator $\widehat{A}$ of A by minimizing the penalized empirical risk

$$\widehat{A} = \underset{A}{\operatorname{argmin}} \left(\frac{1}{n} \|Y - f_A\|_2^2 + \lambda \sum_{l=1}^{M} \rho_l \|A_l\|_1 \right), \tag{4.2}$$

where $\rho_l = W^{(l,l)}$. We subsequently estimate $f(t)$ by

$$f_{\widehat{A}}(t) = \sum_{l=1}^{M} \widehat{A}_l \phi_l(t), \quad t \in [a, b]. \tag{4.3}$$

and measure the quality of the estimator (4.3) by

$$R(f_{\widehat{A}} = n^{-1} \| f_{\widehat{A}} - f \|_2,$$

the L^2 norm of the difference of tensors $f_{\widehat{A}}$ and f at points $t_1, \cdots, t_n$.

In what follows, we assume that f is sparse, so that coefficients $A_l \in \mathbb{R}^{m_1 \times m_2}$ in (4.1) are the low rank matrices.

Let $\mathcal{L}$ be any subset of the set $\{1, 2, \cdots, M\}$. Then, the following oracle inequality holds.

Theorem 1 *Let eigenvalues of matrix W be bounded above and below*

$$0 < w_{\min} = \sigma_{min}(W) \leq \sigma_{max}(W) = w_{\max} < \infty. \tag{4.4}$$

Let $\lambda \geq 2\frac{\epsilon}{\sqrt{n}} \sqrt{2\tau \log n + 2\log(m1 + m2) + \log M}$. Then, for any $\tau > 0$, with probability at least $1 - 2n^{-\tau}$ one has

$$\frac{1}{n} \| f_{\widehat{A}} - f \|_2 \leq \inf_{\tilde{A}} \left[\frac{3}{n} \| f_{\tilde{A}} - f \|_2^2 + \frac{16\lambda^2 w_{\max}^2}{w_{\min}^2} (m_1 \vee m_2) \sum_{l \in \mathcal{L}} \mathit{rank}(\tilde{A}_l) \right]. \tag{4.5}$$

Where $\| \sum_{l \in \mathcal{L}} A_l \phi_l - f \|_2^2$ is the bias, $\sum_{l \in \mathcal{L}} r_l$ is the number of "essential terms", $\frac{\epsilon^2}{n}$ is the error of estimating one value. $\frac{\sigma_{max}}{\sigma_{min}}$ is the conditional number of the basis matrix (if the grid is uniform, $W = \frac{1}{n}\Phi\Phi^T = I$ and $\sigma_{min} = \sigma_{max} = 1$). σ_{max} is magnification due to the basis matrix. $(2\tau \log n + 2\log r_0)$ is the log-factor (price for adaptability.

Proof of Theorem 1. Note that, for any $\tilde{A} \in \mathbb{R}^{m_1 \times m_2 \times M}$ one has

$$\frac{1}{n}\|Y - f_A\|_2^2 + \lambda \sum_{l=1}^{M} \rho_l \|\widehat{A}_l\|_1 \leq \frac{1}{n}\|Y - f_{\tilde{A}}\|_2^2 + \lambda \sum_{l=1}^{M} \rho_l \|\tilde{A}_l\|_1$$

Then, after some algebraic manipulations, for the true tensor f which generates data, obtain

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + \frac{2}{n}\left\langle Y - f, f_{\widehat{A}} - f_{\tilde{A}} \right\rangle + \lambda \sum_{l=1}^{M} \rho_l \left\{ \|\tilde{A}_l\|_1 - \|\widehat{A}_l\|_1 \right\}. \tag{4.6}$$

Since $Y - f = \epsilon\xi$ where $\xi \in \mathbb{R}^{m_1 \times m_2 \times M}$ is a tensor with i.i.d. $N(0,1)$ entries, obtain

$$\begin{aligned} \left\langle Y - f, f_{\widehat{A}} - f_{\tilde{A}} \right\rangle &= \epsilon \left\langle \xi, f_{\widehat{A}} - f_{\tilde{A}} \right\rangle = \epsilon \sum_{i=1}^{n} \left\langle \xi_i, \sum_{l=1}^{M} (\widehat{A}_l - \tilde{A}_l)\phi_l(t_i) \right\rangle \\ = \epsilon \sum_{i=1}^{n} \sum_{l=1}^{M} \phi_l(t_i) \left\langle \xi_i, \widehat{A}_l - \tilde{A}_l \right\rangle &= \epsilon \sum_{l=1}^{M} \text{Tr}\left[(\widehat{A}_l - \tilde{A}_l)^T \sum_{i=1}^{n} \phi_l(t_i)\xi_i \right] \end{aligned} \tag{4.7}$$

Consider random matrices

$$\Xi_l = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_l(t_i)\xi_i \quad l = 1, \cdots, M, \quad \Xi_l \in R^{m_1 \times m_2}$$

. Note that elements $\Xi_l^{j,k} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_l(t_i)\xi_i^{j,k}$ are normal variables with $E\left[\Xi_l^{j,k}\right] = 0$, and

$$\begin{aligned} Cov\left(\Xi_l^{j_1,k_1}, \Xi_l^{j_2,k_2}\right) &= \frac{1}{n} \sum_{i_1,i_2=1}^{n} \phi_l(t_{i_1})\phi_l(t_{i_2}) E\left[\xi_{i_1}^{j_1,k_1}\xi_{i_2}^{j_2,k_2}\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \phi_l^2(t_i)\delta_{j_1,j_2}\delta_{k_1,k_2} = \rho_l^2 \delta_{j_1,j_2}\delta_{k_1,k_2} \end{aligned}$$

Hence, $\Xi_l = \rho_l Z_l$ where Z_l is the $R^{m_1 \times m_2}$ matrix with i.i.d. $N(0,1)$ entries. Therefore, the

left-hand side of equation (4.7) can be bounded above as follows

$$\left|\frac{1}{n}\left\langle Y-f, f_{\widehat{A}}-f_{\tilde{A}}\right\rangle\right| \leq \frac{\epsilon}{\sqrt{n}}\sum_{l=1}^{M}\left|\mathrm{Tr}\left[(\widehat{A}_l-\tilde{A}_l)^T\Xi_l\right]\right| \leq \frac{\epsilon}{\sqrt{n}}\sum_{l=1}^{M}||\widehat{A}_l-\tilde{A}_l||_1\cdot\max_{1\leq l\leq M}||\Xi_l||$$

$$= \frac{\epsilon}{\sqrt{n}}\sum_{l=1}^{M}||\widehat{A}_l-\tilde{A}_l||_1\cdot\rho_l\max_{1\leq l\leq M}||Z_l||. \tag{4.8}$$

For any $\tau > 0$,

$$P(||Z_l|| > \tau\sqrt{m_1 \vee m_2}) \leq (m_1+m_2)e^{-\frac{\tau^2}{2}}$$

or,

$$P(\max_{1\leq l\leq M}||Z_l|| > t\sqrt{m_1 \vee m_2}) \leq (m_1+m_2)Me^{-\frac{t^2}{2}}$$

choosing $t = \sqrt{2\left[\log(m1+m2)+\log M+\tau\log n\right]}$ obtain that

$$P(\max_{1\leq l\leq M}||Z_l|| > t\sqrt{m_1 \vee m_2}) \leq n^{-\tau}$$

Hence, with probability at least $1-n^{-\tau}$ one has(see Tropp Ch. 4 [34]):

$$\sum_{l=1}^{M}\left|\mathrm{Tr}\left[(\widehat{A}_l-\tilde{A}_l)^T\Xi_l\right]\right| \leq \sum_{l=1}^{M}||\widehat{A}_l-\tilde{A}_l||_1\cdot\rho_l\sqrt{m_1 \vee m_2}\sqrt{2\left[\log(m1+m2)+\log M+\tau\log n\right]} \tag{4.9}$$

Let Ω_l be the set of points where inequality (4.9) is valid. Denote by $\Omega = \bigcap_{l=1}^{M}\Omega_l$ the set of points where (4.9) is true for every l. Then, by Morgan laws,

$$P(\Omega) \geq 1-n^{-\tau} \tag{4.10}$$

and, for any $\omega \in \Omega$, one has

$$2|\left\langle Y - f, f_{\widehat{A}} - f_{\tilde{A}} \right\rangle| \leq \lambda_0 \sqrt{m_1 \vee m_2} \sum_{l=1}^{M} \rho_l \|\widehat{A}_l - \tilde{A}_l\|_1, \tag{4.11}$$

where

$$\lambda_0 = 2\frac{\epsilon}{\sqrt{n}} \sqrt{2\tau \log n + 2\log(m1 + m2) + \log M}. \tag{4.12}$$

Combining (4.11) and (4.6), obtain for any $\tilde{A}$ and $\omega \in \Omega$ that

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + \sum_{l=1}^{M} \lambda_0 \rho_l \sqrt{m_1 \vee m_2} \left[\|\widehat{A}_l - \tilde{A}_l\|_1 + \|\widehat{A}_l\|_1 - \|\tilde{A}_l\|_1\right].$$

Let $\mathcal{L} \subseteq \{1, 2, \cdots, M\}$ and $\mathcal{L}^c$ be its complement. Let $\tilde{A}_l = 0$ if $l \in \mathcal{L}^c$. Denote $r_l = \text{rank}(\tilde{A})$. Then, for $\omega \in \Omega$, one has

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + \Delta \tag{4.13}$$

where

$$\Delta = \sum_{l=1}^{M} \rho_l \left[\lambda_0 \sqrt{m_1 \vee m_2}\|\widehat{A}_l - \tilde{A}_l\|_1 + \lambda\sqrt{m_1 \vee m_2}\|\tilde{A}_l\| - \lambda\sqrt{m_1 \vee m_2}\|\widehat{A}_l\|\right].$$

Simple algebra yields

$$\Delta = \Delta_1 + (\lambda_0 - \lambda)\sqrt{m_1 \vee m_2} \sum_{l \in \mathcal{L}^c} \rho_l \|\widehat{A}_l\|_1 \tag{4.14}$$

where

$$\Delta_1 = \sum_{l \in \mathcal{L}} \rho_l \left[\lambda_0 \sqrt{m_1 \vee m_2}\|\widehat{A}_l - A_l\|_1 + \lambda\sqrt{m_1 \vee m_2}\|A_l\| - \lambda\sqrt{m_1 \vee m_2}\|\widehat{A}_l\|\right].$$

In order to obtain an upper bound for Δ_1, we apply Lemmas 6 and 7 of Rohde and Tsybakov [31]:

Lemma 1 *Let matrices A and B have the same size and if* $AB^T = 0; \quad A^T B = 0$ *then*

$$||A + B||_1 = ||A||_1 + ||B||_1 \tag{4.15}$$

The proof is from Recht, Fazel and Parrilo 2010.[2]

Proof of Lemma 1. Consider the singular value decomposition's of A and B keeping zero singular vectors:

$$A = [U_{A1}\ U_{A2}] \begin{pmatrix} \Sigma_{A1} & 0 \\ 0 & 0 \end{pmatrix} [V_{A1}\ V_{A2}]^T$$

and

$$B = [U_{B1}\ U_{B2}] \begin{pmatrix} \Sigma_{B1} & 0 \\ 0 & 0 \end{pmatrix} [V_{B1}\ V_{B2}]^T$$

The condition $AB^T = 0$ implies that $VA1^T V_{B1} = 0$, and similarly, $A^T B = 0$ implies that $U_{A1}^T U_{B1} = 0$. Hence, there exist matrices U_C and V_C such that $[U_{A1}\ U_{B1}\ U_C]$ and $[V_{A1}\ V_{B1}\ V_C]$ are orthogonal matrices. Thus, the following are valid singular value decomposition's for A and B:

$$A = [U_{A1}\ U_{A2}\ U_C] \begin{pmatrix} \Sigma_{A1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} [V_{A1}\ V_{A2}\ V_C]^T$$

and

$$B = [U_{B1}\ U_{B2}\ U_C] \begin{pmatrix} 0 & 0 & 0 \\ 0 & \Sigma_{B1} & 0 \\ 0 & 0 & 0 \end{pmatrix} [V_{B1}\ V_{B2}\ V_C]^T$$

In particular:

$$A + B = [U_{A1}\ U_{B1}] \begin{pmatrix} \Sigma_{A1} & 0 \\ 0 & \Sigma_{B1} \end{pmatrix} [V_{A1}\ V_{B1}]^T$$

This shows that the singular values of $A+B$ are equal to the union (with repetition) of the singular values of A and B. Hence,$||A + B||_* = ||A||_* + ||B||_*$ as desired.

Lemma 2 *Let $A \in \mathbb{R}^{m\times T}$ with rank$(A) = r$ and singular value decomposition $A = U\Lambda V_T$. Let $B \in \mathbb{R}^{m\times T}$ be arbitrary. Then there exists a decomposition $B = B_1 + B_2$ with the following properties:*

a. *rank*$(B_1) \leq 2$ *rank*$(A) \leq 2r$.

b. $AB_2^T = 0$; $A^T B_2 = 0$.

c. *Tr*$(B_1^T B_2) = 0$.

Proof of Lemma 2. SVD of A:

$$A = U \begin{pmatrix} \Sigma_{A1} & 0 \\ 0 & 0 \end{pmatrix} V^T$$

Let $\widehat{B} = U^T BV$, then partition $\widehat{B}$ as:

$$\widehat{B} = \begin{pmatrix} \widehat{B}_{11} & \widehat{B}_{12} \\ \widehat{B}_{21} & \widehat{B}_{22} \end{pmatrix}$$

Defining

$$B1 = U \begin{pmatrix} \widehat{B}_{11} & \widehat{B}_{12} \\ \widehat{B}_{21} & 0 \end{pmatrix} V^T \quad \text{and} \quad B2 = U \begin{pmatrix} 0 & 0 \\ 0 & \widehat{B}_{22} \end{pmatrix} V^T$$

it can be easily verified that B_1 and B_2 satisfy the conditions of the Lemma.

Therefore according Lemma2 the following is true:

1. There exists decomposition $\widehat{A}_l = \widehat{A}_l^{(1)} + \widehat{A}_l^{(2)}$ with the following properties:

   a. $\text{rank}(\widehat{A}_l^{(1)}) \leq 2\,\text{rank}(\tilde{A}_l) \leq 2r_l$.

   b. $\tilde{A}_l(\widehat{A}_l^{(2)})^T = 0$; $\tilde{A}_l^T(\widehat{A}_l^{(2)}) = 0$.

   c. $\text{Tr}\left[(\widehat{A}_l^{(1)} - \tilde{A}_l)^T \widehat{A}_l^{(2)}\right] = 0$.

2. If matrices B_1 and B_2 are such that $B_1 B_2^T = 0$ and $B_1^T B_2 = 0$, then $\|B_1 + B_2\|_1 = \|B_1\|_1 + \|B_2\|_1$

Hence,

$$\begin{aligned} \Delta_1 &= \sum_{l \in \mathcal{L}} \rho_l \left[\lambda_0 \|\widehat{A}_l^{(1)} + \widehat{A}_l^{(2)} - \tilde{A}_l\|_1 + \lambda \|A_l\| - \lambda \|\widehat{A}_l\|\right] \\ &\leq \sum_{l \in \mathcal{L}} \rho_l \left[\lambda_0 \|\widehat{A}_l^{(1)} - \tilde{A}_l\|_1 + \lambda_0 \|\widehat{A}_l^{(2)}\| + \lambda \|\tilde{A}_l\| - \lambda \|\widehat{A}_l\|\right]. \end{aligned}$$

Note that, since

$$\|\widehat{A}_l\|_1 \geq \|\widehat{A}_l^{(2)} + \tilde{A}_l\|_1 - \|\widehat{A}_l^{(1)} - \tilde{A}_l\|_1 = \|\widehat{A}_l^{(2)}\|_1 + \tilde{A}_l\|_1 - \|\widehat{A}_l^{(1)} - \tilde{A}_l\|_1$$

one derives

$$\Delta_1 \leq \sum_{l \in \mathcal{L}} \rho_l \left[(\lambda_0 + \lambda)\|\widehat{A}_l^{(1)} - \tilde{A}_l\|_1 - (\lambda - \lambda_0)\|\widehat{A}_l^{(2)}\|_1\right] \tag{4.16}$$

Combining equations (4.14) and (4.16) and setting $\lambda \geq \lambda_0$, obtain from (4.13) that for any $\omega \in \Omega$ one has

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + 2\lambda \sum_{l\in\mathcal{L}} \rho_l \|\widehat{A}_l^{(1)} - A_l\|_1. \tag{4.17}$$

Note that, for any matrix B,

$$\|B\|_1 \leq \sqrt{\text{rank}(B)} \cdot \|B\|_2,$$

so that

$$\|\widehat{A}_l^{(1)} - A_l\|_1 \leq \sqrt{2r_l}\|\widehat{A}_l^{(1)} - A_l\|_2.$$

Since

$$\begin{aligned}\|\widehat{A}_l - \tilde{A}_l\|_2^2 &= \|\widehat{A}_l(1) - \tilde{A}_l\|_2^2 \\ &+\|\widehat{A}_l^{(2)}\| + 2\text{Tr}\left[(\widehat{A}_l^{(1)} - \tilde{A}_l)^T \widehat{A}_l^{(2)}\right] \\ &= \|\widehat{A}_l^{(1)} - \tilde{A}_l\|_2^2 + \|\widehat{A}_l^{(2)}\| \geq \|\widehat{A}_l^{(1)} - \tilde{A}_l\|_2^2,\end{aligned}$$

one derives

$$\|\widehat{A}_l^{(1)} - \tilde{A}_l\|_1 \leq \sqrt{2r_l}\|\widehat{A}_l - \tilde{A}_l\|_2. \tag{4.18}$$

Combination of formulae (4.17) and (4.18) yield that, for any $\omega \in \Omega$, one obtains

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + 2\lambda \sum_{l\in\mathcal{L}} \rho_l\sqrt{2r_l} \|\widehat{A}_l^{(1)} - A_l\|_2. \tag{4.19}$$

Note that $\rho_l = W^{ll}$, hence, $w_{\min} \leq \rho_l \leq w_{\max}$ and

$$\begin{aligned}\frac{1}{n}\|f_{\widehat{A}} - f_{\tilde{A}}\|_2^2 &= \frac{1}{n}\sum_{i=1}^n \|(\widehat{A}_l - \tilde{A}_l)\phi_l(t_i)\|_2^2 \\ = \sum_{l_1,l_2=1}^M \left\langle \widehat{A}_{l_1} - \tilde{A}_{l_1}, \widehat{A}_{l_2} - \tilde{A}_{l_2}\right\rangle W^{l_1,l_2} &\geq w_{\min} \sum_{l\in\mathcal{L}} \|\widehat{A}_l - \tilde{A}_l\|_2^2.\end{aligned}$$

Therefore, it follows from (4.19) that

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + 2\lambda w_{\max} \sum_{l\in\mathcal{L}} \sqrt{2r_l}\|\widehat{A}_l - A_l\|_2^2 \tag{4.20}$$

Since, for any $d > 0$, one has $2ab \leq d^{-1}\, a^2 + db^2$, obtain

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + 2d^{-1}\,\lambda^2 w_{\max}^2 \sum_{l\in\mathcal{L}} r_l + d\sum_{l\in\mathcal{L}} \|\widehat{A}_l - \tilde{A}_l\|_2^2.$$

Then, inequality (4.20) yields

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 + \frac{d}{w_{\min}}\frac{1}{n}\|f_{\widehat{A}} - f_{\tilde{A}}\|_2^2 + 2d^{-1}\lambda^2 w_{\max}^2 \sum_{l\in\mathcal{L}} r_l.$$

Since $\|f_{\widehat{A}} - f_{\tilde{A}}\|_2^2 \leq 2\|f_{\widehat{A}} - f\|_2^2 + 2\|f_{\tilde{A}} - f\|_2^2$, one obtains

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \left[1 - \frac{2d}{w_{\min}}\right] \leq \frac{1}{n}\|f_{\tilde{A}} - f\|_2^2 \left[1 + \frac{2d}{w_{\min}}\right] + \frac{2\lambda^2 w_{\max}^2}{d} \sum_{l\in\mathcal{L}} r_l.$$

Setting $d = w_{\min}/4$ and multiplying the last equation by 2, derive

$$\frac{1}{n}\|f_{\widehat{A}} - f\|_2^2 \leq \frac{3}{n}\|f_{\tilde{A}} - f\|_2^2 + 16\lambda^2 \frac{w_{\max}^2}{w_{\min}} \sum_{l\in\mathcal{L}} r_l,$$

where $r_l = \text{rank}(\tilde{A}_l)$. Since $\tilde{A}_l$ are arbitrary, this completes the proof.

4.4 Algorithm and Simulation results

Workin with a series of medical images i.e. with tensor data can be divided into two approaches: local and global. A local approach looks at neighboring pixels or voxels of a missing element and locally estimate the unknown values on basis of some difference measure between the adjacent

entries. In contrast, a global approach takes advantage of a global property of the data, and is the path that we use. For matrix-valued data, the rank of a matrix is a good notion of sparsity. As it is a non-convex function, matrix rank is difficult to minimize in general. Recently, the nuclear norm was advocated to be used as convex surrogate function for the rank function [3]. Generalizing this program, a convex surrogate for the tensor rank applied to the unfoldings of the unknown tensor. A related approach, which penalizes the unfoldings of the solution tensor to have low nuclear norm, was already presented in [25] for the special case of tensor completion. Gandi et al [12] have developed very convenient algorithm for solution problems like 4.2. Since the the optimization problem with low rank penalty is a NP problem, significant work were devoted to find a work around of this problem which manifested in the brilliant article of Cai, Candes, Shen [20] who introduced a novel algorithm to approximate the matrix with minimum nuclear norm among all matrices obeying a set of convex constraints. Thus the problem of a rank minimization may be understood as the convex relaxation problem.

4.4.1 Douglas-Rachford technique

Douglas-Rachford algorithm [9] is known for more than 50 years, it addresses the minimization of the sum of two functions $(f+g)(x)$. Where f and g are from a class of all lower semicontinuous convex functions from a real Hilbert space. To proceed lets consider the following sequence:

$$x_{n+1} = x_n + t_n \left\{ \underset{\gamma,f}{\text{prox}}\,[\,2 \underset{\gamma,g}{\text{prox}}\,[\,x_n] - x_n] - \underset{\gamma,g}{\text{prox}}\,[\,x_n] \right\} \tag{4.1}$$

where $t_n \in [0,2]$ and $\sum_n t_n(2-t_n) = \infty$, and the proximal map, of index $\gamma \in (0,\infty)$ of function f is given by the following equation:

$$y = \underset{\gamma,f}{\text{prox}}\,[\,x] \Leftrightarrow y = \underset{y}{\text{argmin}} \left[f(y) + \frac{1}{2\gamma}||y-x||^2 \right] \tag{4.2}$$

Under certain conditions see e.g [5], [6], the sequence (4.1) converges weakly to the minimizer of $(f+g)(x)$. To recast our minimization problem given by equation (4.2 to the unconstrained minimization problem. Do the following:

$$\underset{\mathbb{A}}{\text{minimize}}\,[\,f(\mathbb{A})+g(\mathbb{A})];$$

where $\mathbb{A}=(A_0,A_1)$ and $D=\{\mathbb{A}|A_0=A_1\}$ and

$$\begin{cases} f(\mathbb{A})=\sum_{i=0}^{1} f_i(A_i)=\frac{\lambda}{2}||A_0*\Phi-Y||_2^2+||A_1||_* \\ g(\mathbb{A})=i(\mathbb{A})_D=\begin{cases} 0, & \text{if } \mathbb{A}\in D \\ +\infty, & otherwise \end{cases} \end{cases} \tag{4.3}$$

This formalizm is equivalent to equation (4.2) , we only need to establish proximity maps of f and g to apply algorithm(4.1). The proximal map of f is given by:

$$\underset{\gamma,f}{\text{prox}}\,[\,\mathbf{X}]=\underset{\mathbf{Y}}{\text{argmin}}\left\{\sum_{i=0}^{1} f(X_i)+\frac{1}{2\gamma}||\mathbf{Y}-\mathbf{X}||_F^2\right\} \tag{4.4}$$

For $i=1$ the proximal map of f_1 is essentially the shrinkage operator as described in [20]:

$$\underset{\tau,||\cdot||_*}{\text{prox}}\,[\,X]=\underset{Z}{\text{argmin}}\left[||Z||_*+\frac{1}{2\tau}||Z-X||_2^2\right]=\text{shrink}(X,\tau) \tag{4.5}$$

The operator shrink(X,τ) acts by applying the soft thresholding to the singular values of X. Consider singular value decomposition of the matrix X:

$$X=U_X\Sigma_X V_X^T$$

where $\Sigma_X = diag(\sigma_1(X), \cdots, \sigma_r(X))$, hence action of "shrink" operator can be described as:

$$\text{shrink}(X, \tau) = U_X \text{diag}\left[\max\left\{\sigma_1 - \tau\right\}, \cdots, \max\left\{\sigma_r - \tau\right\}\right] V_X^T$$

To apply shrinking operator to a tensor we need to unfold the tensor to matrix first and then upon shrinkage fold it back. As for proximal map for $i = 0$, we have:

$$\begin{aligned}\underset{\tau, f_0}{\text{prox}}\left[\, X\right] &= \underset{A_0}{\text{argmin}} \left\{ \frac{\lambda}{2} ||B_0 * \Phi - Y||_2^2 + \frac{1}{2\tau} ||A_0 - B_0||_2^2 \right\} \\ &= \left(\lambda \Phi^T \Phi + \frac{1}{\tau}\mathbb{I}\right)^{-1} \left(\lambda \Phi^T Y + \frac{1}{\tau} A\right)\end{aligned}$$

Finally, the proximal map of the indicator function i_D is the orthogonal projection onto the set D and is given by:

$$i_D = \left[I_{A_0}, I_{A_1}\right] \frac{A_0 + A_1}{2} \tag{4.6}$$

Now we can develop splitting Douglas-Rachford algorithm for our problem (4.2)

Algorithm 3 splitting Douglas-Rachford algorithm

input: tensor Y of observed data; functional matrix Φ

input: t_i-divergent sequence

input: λ τ γ -parameters for Lagrange multipliers and for soft thresholding

input: ϵ - to check convergence

Init tensor $A_0^{(0)}$ and $A_1^{(0)}$ with zeros

while $||A^{i+1} - A^i|| > \epsilon$ **do**

$\widehat{A} = \text{mean}(A_0^{(i)} + A_1^{(i)})$

calculate $A_0^{(i+1)} = A_0^{(i)} + t_i \left(\text{prox}\left[_{\tau, f_0} 2\widehat{A} - A_0^{(i)}\right] - \widehat{A}\right)$

calculate $A_1^{(i+1)} = A_1^{(i)} + t_i \left(\text{refold}(\text{shrink}(\text{unfold}(2\widehat{A} - A_1^{(i)}))) - \widehat{A}\right)$

end while

4.4.2 Simulation results

We generated some functional data using elementary functions, contaminated them with Gaussian noise and then recovered the data using algorithm (4.4.1) the functions were evaluated over integer values of dependent variable $x \in \{1, \cdots, 30\}$. The following functions were tabulated:

$$x;\ \frac{x^2}{30};\ x(30-x);\ \frac{x^3}{900};\ \tan x;\ x\log(1+x)$$

We created tensor of zeros $T \in \mathbb{R}^{30\times30\times30}$ and randomly inserted functional fibers in it. Then a random Gaussian noise was added. The results are presented on the Figure [9] below. The nuclear norm of the "denoised" tensor is even smaller that the one for the original tensor before addition of the noise. From another point of view the bias are still significant and for some functions it is larger than for anothers.

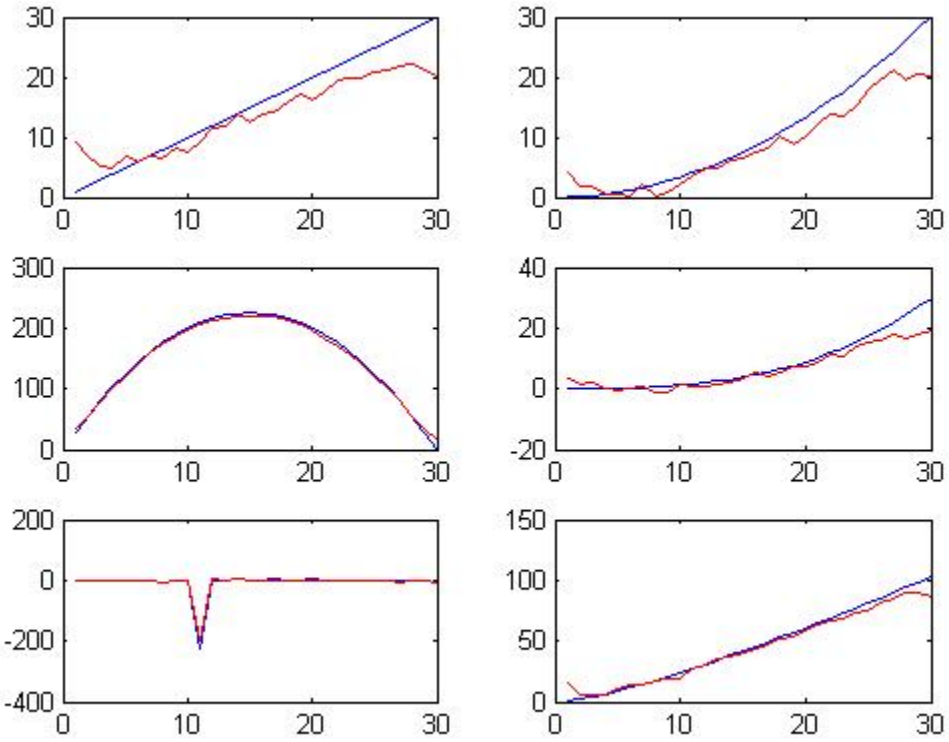


Figure 4.1: Simulation data for Douglas-Rachford algorithm. The true value are in blue. Denoised are in red.

CHAPTER 5: DISCUSSION

5.1 Conclusion

- Current Age dependent hazard model We found some evidence that in the case of the Adenocarcinoma and squamous cell carcinoma the hazard increasing for women of ages $52 - 56$ is comparison with men
- Matrix-variate denoising We developed oracle in equality for the matrix-variate data least squares fit. We also developed algorithm which carries out denoising and delivers optimal convergence rates.

5.2 Future work

We are planning to continue our investigation in two different direction. One venue is a survival analysis combined with genetic and biochemical data. There is also possibility of applying renewal mathematical models together with cancer growth mathematical models toward understanding carcinogenics and cancer mortality.

Another direction is father developing algorithm for functional data for Dynamic Contrast Enchanced imaging. The results obtained can be improved by taking into considerations different penalty functions including the ones based on metabolic kinetics from one side and on advanced mathematical conceptions such as fractal dimension for example from another side. There is a room for improvement: undeniably, however, this is a matter of future investigation.

LIST OF REFERENCES

[1] American Lung Cancer Assosiation. Lung cancer fact sheet, 2014.

[2] B.Recht, M. Fazel, and P.A. Parrilo. Guaranteed minium-rank solution of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501, 2010.

[3] E. J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundation of Computational Mathematics*, 9:717–772, 2009.

[4] A. Cichocki, R. Zdunek, A. Huy Phan, and S. Amari. *Nonegative matrix anf tensor factorization*. Wiley, 2009.

[5] P.L. Combettes and J.C. Pesquet. A douglas-rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Selected Topics in Signal Processing*, 1:564–574, 2007.

[6] P.L. Combettes and J.C. Pesquet. A proximal decomposition method for solving convex variational inverse problems. *Inverse Problems,*, 25:065014, 2008.

[7] Alex Cook. Censoring and truncation, 2008.

[8] R. A. DeVore and B.Y J. Lucier. Wavelets. *Acta Numerica*, 1:1–56, 1992.

[9] J. Douglas and H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Trans. of the American Mathematical Society.*, 82:421439, 1956.

[10] J.B. Fu, Y. Kau, R.K. Severson, and G.P. Kalemkerian. Lung cancer in women: analysis of the national surveillance, epidemiology, and end results database. *Chest*, 127(3), 2005.

[11] W.J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 3:397–416, 1998.

[12] S. Gandy, B. Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27, 2011.

[13] J.J. Goeman. l_1 penalized estimation in the cox proportional hazards model. *2010*, 52:70–84.

[14] Jürgen Groß. *Linear Regression*. Springer, 2003.

[15] W. Hardle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical Application*. Springer, 1998.

[16] C.I. Henschke and O.S. Miettinen. Women's susceptibility to tobacco carcinogens. *Lung Cancer*, 43(1):1–5, 2004.

[17] A.E. Hoerl and R.W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[18] Zhou Hua, Li Lexin, and Zhu Hongtu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108:540–552, 2013.

[19] J. Huang and D. Harrington. Penalized partial likelihood regression for right censored data with bootstrap selection of the penalty parameter. *Biometrics*, 58:781–791, 2002.

[20] JF.Cai, E. J. Cands, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20:1956–1982, 2008.

[21] J.O.Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, 2005.

[22] Li Jun, Yu Yanming, and Zhang Yibao. A clinically feasible method to estimate pharmacokinetic parameters in breast cancer. *Medical Physics*, 36:3786–3794, 2009.

[23] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assn.*, 53:457–481, 1956.

[24] D. G. Kleinbaum and M. Klein. *Survival Analysis a Self-Learning Text*. (Springer, 2005.

[25] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on pattern analysis and machine intelligence*, 35:208–220, 2013.

[26] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 2008.

[27] F. Montemurro, F. Russo, and L. Martincich. Dynamic contrast enhanced magnetic resonance imaging in monitoring bone metastases in breast cancer patients receiving bisphosphonates and endocrine therapy. *Acta Radiologica*, 45:71–74, 2004.

[28] M.Tableman and J.S. Kim. *Survival Analysis Using S*. Chapman&Hall/CRC, 2004.

[29] G.V. Pendse. A tutorial on the lasso and the ”shooting algorithm”., 2011.

[30] P. Reiss and R.Ogden. Functional generalized linear models with images as predictors. *Biometrics*, 66:61–69, 2010.

[31] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39:887930, 2011.

[32] Jean-Luc Starck, Fionn Murtagh, and Jalal M. Fadili. *Sparse Image and Signal Processing*. Cambridge University Press, 2010.

[33] R. Tibshirani. The lasso method for variable selection in the cox model. *Stat. Med.*, 16:385–395, 1997.

[34] J. A. Tropp. User-friendly tools for random matrices: An introduction, 2012.

[35] A. Wismuller, O. Lange, D.R. Dersch, G.L. Leinsinger, K. Hahn, B. Putz, and D. Auer. Cluster analysis of biomedical image time-series. *International Journal of computer vision*, 2:103–128, 2002.

[36] X. Yan and X.G. Su. *Linear Regression Analysis*. World Scientific, 2009.

[37] Yi Yang and Hui Zou. A coctail algorithm for solving the elastic net penalized cox regression in high dimensions. *Statistics and Its Interface*, 6:167–173, 2013.

[38] Hao Helen Zhag and Wenbin Lu. Adaptive lasso for cox proportional hazards model. *Biometrika*, 94:691–703, 2007.

[39] H. Zhou, L. Li, and H.Zhu. Tensor regression with applications in neuroimaging data analysis. *arXiv preprint*, (1203. 3209), 2012.