Sliced Inverse Regression for Lifetimes and a Remark on High-Dimensional Graphical Models

THÈSE N^{O} 5816 (2013)

PRÉSENTÉE LE 19 JUILLET 2013

À LA FACULTÉ DES SCIENCES DE BASE
CHAIRE DE STATISTIQUE APPLIQUÉE
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Maya SHEVLYAKOVA

acceptée sur proposition du jury:

Prof. Th. Mountford, président du jury
Prof. S. Morgenthaler, directeur de thèse
Prof. K.-C. Li, rapporteur
Prof. F. Naef, rapporteur
Dr C. Posse, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

Life isn’t about finding yourself. Life is about creating yourself.
— George Bernard Shaw

Acknowledgements

Those who know me well might have guessed that I started thinking about this part of the thesis long before the first thesis template was even created. Simply because I have been lucky to have been surrounded by many people who made the everyday PhD routine so much brighter, even on the foggy day (and since we're in Lausanne, it comes as no surprise that there have been many of those).

First of all, I want to express my gratitude towards my advisor, Prof. Stephan Morgenthaler. For his patience, guidance and support throughout this journey. I came to EPFL wanting to study statistics, and now, 6 years later, my motivation and admiration for this field is higher than it's ever been, and I owe it to Stephan.

I would like to thank the members of my jury, Prof. Ker-Chau Li, Prof. Felix Naef and Dr. Christian Posse for their fruitful feedback and comments which helped improve this thesis, as well as Prof. Thomas Mountford for having presided my jury.

Six and a half years in Lausanne, the time truly flies. The list of people I want to thank is long, yet I will try. I'll start with all the "math" people: Stéphane, Julian, Anna, Jacques, Fanny, Oli, Laura, Jean-Benoît, Caroline, Gwenol, David and all the others that I obviously failed to name here, for all the fun, Sat coffees and math (de?)motivation. All the former statistics colleagues, especially Laurence and Sahar. And Nicolas, for the 7.30 coffee breaks. Yeay for the morning productivity! Our secretaries, Anne-Lise and Nadia, for the immense help with the paper work, Swiss bureaucracy and replacing my photocopy cards (yes, i am the trouble when it comes to magnetic stuff). Jean-Marie Helbling, for equilibrating the teaching load. Our IT crew, Léonard and Marc, for working the magic. Antonin, for everything; Oli and Xav, for the food, Apérolac, the mountains and style on the bike. Ok, non-math people now. All the ACIDE team, you guys rock! Pamela, Ruud, Sylvie, Willem-Jan, Brammert, Emile, Elena, Florent, Michela and Torsten! Whenever it's the dancefloor in Schneewittchen, cablecars in Saas-Fee or the tables in Great, we take it all. Special thanks to Tobi for being there in the moments of PhD despair and making me laugh about it. More EPFL crowd: Albrecht, for pure German awesomeness; Denisa and Maria, for constantly cheerful mood; Gizil, for the scepticism; Daniel, for Morf; German, for bringing the inspiration; Veronica, for being the inspiration. Finally, the fellows of the Swiss ring. Nathan, for bold fashion and music choices; Evelina, for being a girl on the bike; Toby, for the discussions; Sophie, for the straightforwardness; Steven, for the global nonconformism; Filip, for the concentration and motivation; Marius, for the open doors in all your houses; Simon, for never judging my dress style; Krem, for curiosity, support, tons of great vegan food and simply being my friend.

I've been a committee member of Xchange (ESN EPFL) for a number of years and I've had amazing time organizing and participating at ski and climb weekends, cultural trips and different soirées. I want to thank all the past and present committee members for that experience, especially Emanuele (for NP Bern and AGM in particular), Nora, Pierre, Cléo, Céline and Bernat P, for our triathlon trials and lots of fun together. Yann, Kim, Stefan, Matthieu, Julien, Fabian... the list is long. A special thank-you goes to Bernat M, for all the Vivapoly ideas, conneries espagnolles and hosting us at la Vallée. As Greg would say, "Mec, vive nous!".

LCVMM lab takes a lot of credit for the movie nights, ski outings and the lunch breaks (oh, and the coffee). A special thank-you goes to Marco for proof-reading a half of my thesis.

During last couple of years, I've taken part in a number of events organized by Polyathlon association. Many thanks go to both French and Swiss parties, especially René Bugnion, for promoting such a lovely non-competitive event. I keep the best memories of those trips and cherish all the friendships acquired through them. The Lebanese team, Polyliban, and Tina, George and Carole, also get the best references for introducing me to the land of Cedars. I will be back! The Luxembourgish team deserves a special reference for being great travel companions and good friends on and off the bike. Alain, Joël, Philippe and Christian, this one is for you.

I want to say a special thank you to Daiva Petkevičiūtė, who has been a loyal friend and a travel companion. Our trips together enriched my experience and the 4pm coffee breaks made the work afternoons so much more enjoyable.

Last but not least, I wish to thank my family for their support. Words cannot fully express my gratitude to my parents who have always been there for me and believed in me. My brother and my family in Boston also take the credit. Finally, I wish to mention my grandmother, Yevdakia who is the kindest and the wisest person I know and who never ceases to inspire me. This thesis is dedicated to her.

M.S., Lausanne, June 2013

Abstract

When analyzing multivariate data, one can appeal to the procedures of dimension reduction to describe its main features in the easiest way possible. In this thesis we work with one such methods, the sliced inverse regression (SIR), and propose a new adaptation to survival data.

A popular idea to account for censoring is to reweight the observed data points, often with the help of inverse probability weighting. We base our strategy on the estimation of the unobserved information.

Our idea is tested with different distributions for the two main survival data models, Accelerated Lifetime Model and Cox's proportional hazards model. In both cases and under different conditions of sparsity, sample size and dimension of parameters, this non-parametric approach evaluates the data structure successfully and can be viewed as a variable selector. We also compare our method with other existing techniques and find it to be competitive.

In the second part of the thesis, we concentrate on the problem of detection of a partial correlation. The ability to identify reliably a positive or negative partial correlation between the expression levels of two genes is determined by the number p of genes, the number n of analyzed samples, and the statistical properties of the measurements. Classical statistical theory teaches us that the product of the root sample size multiplied by the size of the partial correlation is the crucial quantity. But this has to be combined with some adjustment for multiplicity depending on p, which makes the classical analysis somewhat arbitrary. We investigate this problem through the lens of the Kullback-Leibler divergence, which is a measure of the average information for detecting an effect. As a results, it appears that commonly sized studies in genetical epidemiology are not able to reliably detect moderately strong links.

Keywords: survival data, sliced inverse regression, partial correlation, graphical models, Kullback-Leibler divergence.

Résumé

Lors de l'analyse des données multivariées, on peut faire appel aux procédures de réduction de la dimension afin de décrire les principales caractéristiques de la manière la plus simple possible. Dans cette thèse, nous travaillons avec une de ces méthodes, la régression inverse par tranches (SIR), et nous l'adaptons pour les données de survie.

Afin de tenir compte des données censurées il est courant de repondérer les points observés, souvent à l'aide de la probabilité inverse. Notre stratégie est basée sur l'estimation de l'information non observée.

On teste notre méthode avec des distributions différentes pour les deux modèles des données de survie principaux qui sont le modèle à temps accéléré et le modèle à risque proportionnel de Cox. Dans les deux cas et en changeant la distribution, la taille de l'échantillon et le nombre des paramètres, cette approche non paramétrique permet d'évaluer la structure de données avec succès et peut être considérée pour la sélection des variables. Nous comparons également notre méthode avec d'autres techniques existantes et la trouvons compétitive.

Dans la deuxième partie de la thèse, nous nous concentrons sur le problème de la détection d'une corrélation partielle. La capacité d'identifier de manière fiable une corrélation partielle positive ou négative entre les niveaux d'expression des deux gènes est déterminée par le nombre p de gènes, la taille d'échantillon n et les propriétés statistiques des mesures. La théorie statistique classique nous dit que la racine de la taille de l'échantillon multiplié par la taille de la corrélation partielle est une quantité importante. Mais il faut aussi considérer un certain ajustement pour la multiplicité en fonction de p, ce qui rend l'analyse classique quelque peu arbitraire. On étudie ce problème à l'aide de la divergence de Kullback-Leibler, qui est une mesure de la moyenne des informations de détection d'un effet. On conclut que les tailles des études utilisées habituellement en épidémiologie génétique ne semblent pas être en mesure de détecter de manière fiable les liens modérément forts.

Mots-clés: données de survie, régression inverse par tranches, corrélation partielle, modèles graphiques, divergence de Kullback-Leibler.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1. Motivation

Within the last ten years, the amount of data being collected has increased at high speed. One of the reasons for this development can be found in the automated data collection systems. The internet provides an obvious example. Web-based retailers such as Amazon suggest the products a customer might like based on a personal history of purchases. This is based on finding similarities with other clients.

In biomedical research, the introduction of microarrays at the end of the 90s and, most recently, the next-generation sequencing data as well as SNPs arrays changed the nature of data, linking thousands of variables to a single subject. While biologists might hope to gain a deeper understanding of processes on the molecular/-cellular level, the treatment of these huge databases represents a difficult challenge for statisticians and data analysts.

Even the primary analysis of high-dimensional (when the number of subjects n is smaller than the number of variables p) data is often not straightforward. A definition of distance (or similarity) between two individuals in not precise in this case, and its visualization is often problematic. Quite often some of the covariates are correlated which makes the interpretation unclear. The classical statistical theory, which relies on asymptotic approximation obtained by letting the size of the data n tend to infinity, can no longer be applied. Altogether, statistical inference under high-dimensional settings requires if not the creation of new methods, then at least an adaptation of existing ones. This thesis partially reflects on limitations of a few existing methods in different settings.

An intuitive thing, when dealing with the multivariate data, is to try to understand it in the best possible way, either by simplifying it to groups of new variables, or by finding a few original covariates which explain the most pertinent features of the data. If this can be achieved, then one can relate to the well-known techniques to get more accurate results.

Under low-dimensional settings ($n > p$), performing a statistical test requires computing the statistic and comparing it to the quantile of its distribution under the null hypothesis (on a chosen level) or working directly with its p-value. High-dimensionality brings in new questions to consider. Microarray data analysis, for example, usually aims to get a list of differentially expressed genes by fitting a linear model for two or multiple classes of arrays. When fitting the model, a test is performed for every single gene (which amounts to between 5000 and 10000). Such setup requires a correction for an error.

If a single test is employed to test a null hypothesis, using 0.05 as the significance level α and if the null hypothesis is actually true, the probability of reaching the right conclusion is $(1 - \alpha) = 0.95$. If two such hypotheses are tested independently, then the probability of reaching the right conclusion on both occasions would be $(1 - \alpha)^2 = 0.95 \cdot 0.95 = 0.90$. If m true hypotheses are tested independently, the probability of being right on all occasions would decrease substantially to $(1 - \alpha)^m = 0.95^m$. In other words, the probability of being wrong at least once (or getting a significant result erroneously) would become $(1 - (1 - \alpha)^m)$. This means that in the case of multiple hypotheses testing on a given data set there is an increasing probability of getting at least one false significant result.

There are a number of possible ways to correct for multiple testing. The Bonferroni correction, where the significance level α/m for each individual test is obtained after dividing the overall significance level α by the number of tests, ensured that the probability of at least one false rejection is bounded by α. Another popular choice is a False Discovery Rate technique, proposed by Benjamini and Hochberg (1995), where the expected proportion of errors among the rejected hypotheses is controlled.

What are the possible ways to treat multivariate data? One of the options is some kind of regression with regularizations to prevent the model from overfitting (being unnecessarily complex and noisy) and ensure the correct interpretation of results. The most popular choice for regularizations include the L_1-penalty (LASSO), which allows for sparse solutions, the L_2-penalty (ridge regression), their combination (elastic net) or different kinds of information criteria (AIC, BIC, etc.).

Another way to gain a better understanding of the data structure is clustering. It allows to regroup the subjects according to some similarity rule. It can be hierarchical (the objects nearby are most likely grouped together), based on the density, the distribution or centroid-based (when the clusters are defined by a central vector).

One can also consider the classification approach or assigning the observations to a set of classes. This family of procedures covers discriminant analysis, random forests, support vector machines, logistic regression and dimension reduction, each of them englobing a set of various methods. Dimension reduction aims to select a group of new variables which preserve the main features of the data. There are many working methods for dimension reduction, to name here a few: Principal Component Analysis (PCA), Principal Hessian Directions (PHD), Sliced Average Variance Estimation (SAVE), Sliced Inverse Regression (SIR). One can divide those into supervised and unsupervised algorithms. Unsupervised procedures (like PCA) are based on the response vector only, while the supervised ones take covariates into account. In this thesis we concentrate on the SIR method which is a form of the supervised dimension reduction technique. It reconsiders a p-variable multiple regression as a set of p univariate regressions, resulting in an effective dimension-reduction space, capturing the information about the response.

1.2. Thesis scope and outline

In this work we cover two topics. The major part considers an application of the sliced inverse regression (SIR) to survival data. First, the SIR performance is explained on a linear regression, then its adaptation to the censored data is discussed. Since it is not the first attempt to adapt SIR to survival data, some comparisons are made with existing methods. The second topic, presented in Chapter 4, deals with sizing studies when uncovering the structure of the graphical models. What is the power of the partial correlation test? How much information can we draw from the Kullback-Leibler divergence? What is the sample size needed in order to detect the edge in the graph? These are the questions we address. The main idea we wish to emphasize is that once the number of parameters is large enough, only strong dependencies can be uncovered, no matter what method is used.

This thesis has the following structure. Chapter 2 introduces background material, covering the main aspects of survival analysis, which are used throughout this manuscript. The main parameters, models and distributions are presented, as well as the generation of survival times under specific models. The second part of Chapter 2 presents the idea behind the SIR, illustrating it in a linear regression case. We review the main algorithm and the theoretical background of the method. Asymptotic properties of the SIR estimates are discussed.

While Chapter 2 is focused on linear regression, the SIR application to survival data is introduced in Chapter 3. The integration of censored observations into the algorithm is discussed in detail and is tested on simulations. Two models are considered, the one based on the proportional hazards and the one based on the accelerated lifetimes. In order to compare to other existing methods of SIR in the survival context, our approach is applied to some alternative models and to the DLBCL dataset. Variance estimation is considered, in terms of the maximum likelihood approach (for the accelerated lifetime model) and via bootstrap.

The material presented in Chapter 4 is a different topic concerned with detection of the structure of the Gaussian graphical models. The power of detecting is of interest, especially as a function of the sample size. A case of a single partial correlation is studied in detail, with the help of the local asymptotic power and the Kullback-Leibler divergence, allowing to view the problem from different perspectives. Conclusions about the feasibility of correlation detection are drawn.

CHAPTER

2

SURVIVAL AND REGRESSION

2.1. Survival data

One speaks of survival analysis if the variable one wants to understand is the time to an event. The event of interest is often a failure of some type, but it can represent many other occasions. Examples include time from operation to remission, time from diagnosis to death, time to retirement and so on. Originally, survival analysis was concerned with the time from treatment to death (hence the name), but it has proven to be a very practical tool in many areas other than mortality. In engineering, it is sometimes called reliability analysis or time to event analysis.

The specifics of the survival data are such that obtaining the observations takes time. If one wishes to analyze the level of cholesterol in a specific stratum of population, a simple survey involving the taking of blood samples would be sufficient. However, when dealing with the survival time for a rare disease, the sampling of patients would take place over a long period of time, sometimes decades.

Survival analysis originated from life-table analysis as used by life insurers, demographers and epidemiologists. It is one of the oldest statistical disciplines. First traces of actuarial science and demography go back to the 17-th century, with the first life-table presented in 1662 by John Graunt (Kreager, 1988). Until the 1950s, most approaches were actuarial. The modern development of survival methods started in 1958, when Kaplan and Meier (1958) proposed their estimator for a survival function. This paper has become one of the most cited in the history of statistics. Time recording in the lifetime tables was grouped by fixed intervals, often long (one-year to five-year) ones. The Kaplan-Meier approach took care of the observations coming up in clinical trials where the patients were monitored closely and the events were registered with much higher precision (Aalen et al., 2009).

The introduction of the Kaplan-Meier survival curve opened new areas for research, such as the comparison of the survival curves or hazard functions. The log-rank test of Mantel (1966) became a popular solution. The introduction of the proportional hazards model by Cox (1972) allowed for the adjustment for covariates, and due to its flexibility became a very popular model for survival data.

The fundamental theoretical advancements came later. Asymptotic properties and the theory behind the Cox's model were intensively studied. Among the main contributions one could list Breslow and Crowley (1974), Cox (1975) and Tsiatis (1981). The martingale theory proved to be a very helpful tool in survival theory. The notion of counting processes for survival data was first introduced in the Ph.D. thesis of Aalen (1975) and later developed in several papers and books (Aalen, 1978; Aalen et al., 2008; Andersen and Gill, 1982). Aalen et al. (2009) argue that the martingales, while not formally mentioned, could be intuitively uncovered in the logrank test and the partial likelihood paper of Cox (1975).

A survival dataset contains a response variable T, representing the survival time (failure time, lifetime), and a vector of covariates $x \in \mathbb{R}^p$. In this work, we mainly consider biological interpretations. There are many possible questions survival analysis may answer, the most common being the prediction of survival for a patient based on the covariates x, the identification of subgroups with low or high survival or a general understanding of survival as a function of x. Also, often the comparison between a treatment and a control group or more generally the comparison of groups, is of interest. As a new and flexible tool, we will investigate the use of sliced inverse regression of Li (1991) for survival data.

It can happen that information about some individuals is incomplete, in this case it is said to be censored. There are two main types of censoring. When the subject drops out of the study, the corresponding observation is said to be right-censored, the last available information being the fact that the subject is alive at a specific moment. This is a common situation in longitudinal studies since some participants lose their interest, relocate, or are lost for some other reason. Left-censoring is less common, a classical example being a limitation of the measurement technology which often happens in engineering and environmental research. In cancer studies, a time of metastasis can be left-censored if it occurs in between the patient follow-ups. It is not uncommon to have datasets with both types of censoring. The main focus in this thesis is on right-censored data.

Why keep censored observations? First of all, deleting them directly affects the power of the study, bringing the sample size down. In rare genetic studies (for example, hereditary diseases) one can simply not afford to lose information, the same goes for expensive medical tests. Another reason is that by deleting censored individuals from the study we introduce a selection bias.

The presence of censoring, which creates a mixture of complete and incomplete data, is an important characteristic of survival data and it requires the development of special methods for its analysis.

Example: Here is a simple example of a dataset with survival times. Suppose

$$T \; : \; 15,\ 18^*,\ 20,\ 20^*,\ 26,\ 28^*,\ 32,\ 32,\ 35,\ 37^*,$$

is a list of the survival times (in months) following a medical intervention. The observations marked with * are right-censored. To each observation may correspond a vector of covariates x_i, $i = 1, \ldots, 10$.

2.1.1. Main parameters

The main parameter of interest in survival analysis is the survival function, defined as

$$\bar{F}(t) = S(t) = \mathrm{P}(T \geq t)\,,$$

denoting the probability that the event does not happen before the time point t (the individual survives until t). Usually the condition $S(0) = 1$ is assumed. It can be easily seen that the survival function is a complement of the cumulative distribution function F,

$$S(t) = 1 - F(t-). \tag{2.1}$$

Survival functions can be estimated non-parametrically, for example, with a help of a Kaplan-Meier method (Kaplan and Meier, 1958), which is based on conditional probabilities. Their estimate can be written as

$$\hat{S}(t) = \prod_{j:\, t_j < t} \frac{r_j - d_j}{r_j}, \tag{2.2}$$

where $t_1, \ldots, t_m$ is the set of m distinct event times observed in the sample, d_j is the number of deaths at t_j, and r_j is the number of individuals "at risk" at time t_j-, right before the j-th event time. The number of censored observations between the time t_{j-1} and t_j, denoted as c_j, is taken into account when computing r_j, using the relationship

$$r_{j+1} = r_j - d_j - c_j.$$

Example: (continued) In our dataset, we have 6 events and 4 censored observations. That means that the value of the survival function changes 6 times, starting at 1 at the time point $T = 0$. However, two observations ($T = 32$) are tied, so we get 5 jumps of $\hat{S}(t)$. Figure 2.1 shows the Kaplan-Meier estimator. The fact that our estimated survival curve doesn't reach the value of 0 at the last event is explained by the presence of a censored observation at $T = 37$, which implies that $\hat{S}(35) = 0$ would not make sense.

Plotting estimated survival curves is an useful tool when comparing several groups, for example a treatment and a control group.

Another important aspect of survival times is the age-dependent mortality, which is called the hazard function and is for continuous survival times T defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathrm{P}(t \leq T < t + dt \mid T \geq t)}{dt}.$$

It follows that

$$\lambda(t) = \frac{f(t)}{S(t)},$$

where $f(t)$ is the density of the survival time.

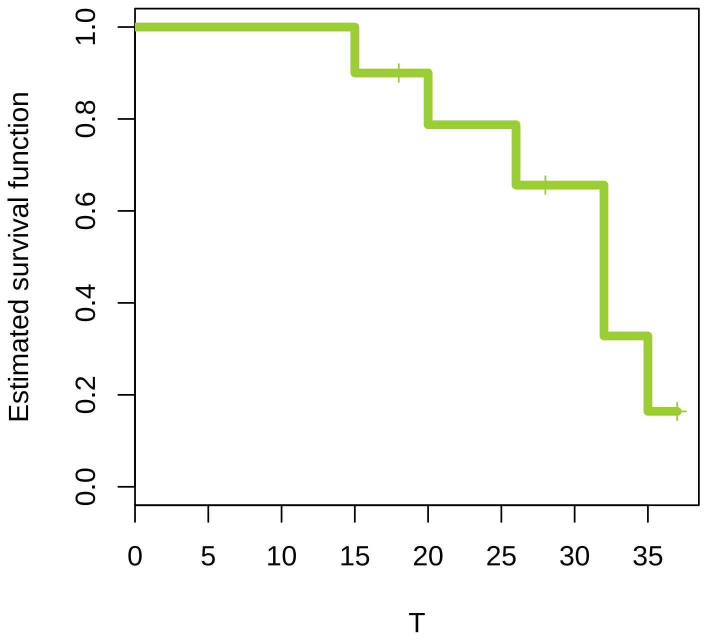

Figure 2.1.: The Kaplan-Meier estimate of the survival function for the Example 2.1.

Sometimes it can be helpful to consider the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u)du,$$

representing the risk accumulated by an individual at age t. The following relationship holds between the cumulative hazard function and the survival function:

$$\Lambda(t) = -\log S(t). \tag{2.3}$$

The Nelson-Aalen estimator (Aalen, 1975) allows for a non-parametric estimation of the cumulative hazard function. It is written as

$$\hat{\Lambda}(t) = \sum_{j:\, t_j < t} \frac{d_j}{r_j},$$

where r_j and d_j are the same as in (2.2). Using (2.3) we get an alternative estimator to (2.2).

Below we review the two most commonly used regression models for survival data. In a regression model, the effect of covariates on survival is described.

2.1.2. Cox's proportional hazards (PH) model

The most popular survival model is Cox's proportional hazards model (Cox, 1972), which has been extensively studied. It assumes the following relation for the hazard function for an individual with characteristics x:

$$\lambda_x(t) = \lambda_0(t) e^{\beta^T x}, \tag{2.4}$$

where $\lambda_0(t)$ is the baseline hazard function which is unknown and β contains the regression coefficients of interest.

The name for the model comes from the fact that the hazard ratio of two individuals with characteristics x and y satisfies

$$\frac{\lambda_x(t)}{\lambda_y(t)} = \frac{\lambda_0(t) e^{\beta^T x}}{\lambda_0(t) e^{\beta^T y}} = \frac{e^{\beta^T x}}{e^{\beta^T y}},$$

which is independent of the elapsed time t.

Cox's model is a semi-parametric model. While no assumptions about the form of $\lambda_0(t)$ are made, we assume a parametric form for the effect of the predictors on the hazard.

Parameter estimates are obtained by maximizing the partial likelihood function, which in the absence of tied survival times is

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{\beta^T x_i}}{\sum_{j \in R_i} e^{\beta^T x_j}}. \tag{2.5}$$

The product in (2.5) is computed over all the observed events (failures) and for each event we define the set R_i as the list of individuals at risk of failure time i. Censored data are naturally taken into account.

Cox originally called formula (2.5) a conditional likelihood since it is the product of conditional probabilities that a specific subject would fail at a given time. But the whole expression is not a conditional probability. In Cox (1975), the name of partial likelihood was suggested and a general justification was given for $L(\beta)$ to be the essential part of the full likelihood. Detailed justification of the partial likelihood in terms of counting processes can be found in Andersen et al. (1993).

2.1.3. Accelerated failure time model (ALT)

Another model for survival data which might be more reasonable in many medical studies is the accelerated failure time (lifetime) model. Cox's model assumes proportionality between hazards, while the accelerated lifetime model considers different behaviors for the hazards of different individuals. The survival time is expressed as

$$T = \exp(\beta^T x) T_0, \tag{2.6}$$

where T_0 denotes the base survival time and $\exp(\beta^T x)$ estimates the effect of the covariate values of the individual on his survival time. If $\beta^T x > 0$, the failure occurs later than for the base situation. Alternatively, when $\beta^T x < 0$, one speaks of accelerated lifetime since failure happens earlier.

Since $\exp(\beta^T x)$ acts as a scale parameter, the cumulative distribution function is

$$F_T(t) = F_0(\exp(-\beta^T x)t).$$

The main difference between Cox's model and the accelerated failure time is that the latter is often treated in a fully parametric manner. This fact makes this model less popular. However, in many genetic studies one would expect that certain conditions would shift the peak of the hazard curve compared to healthy individuals, which would rule out the proportionality of the hazards.

For example, an individual with a family history of breast cancer and with a mutated BRCA1 gene is known to have a higher risk of developing this disease compared to an individual with no such family history at all. In this case, the proportional hazards assumption is clearly wrong, while the accelerated risk may explain the action of this covariate. Another point in favor of the accelerated lifetime model is a simple interpretation of estimated coefficients of the parameters which is less intuitive in Cox's model.

2.1.4. Distributions in survival analysis

In this section we briefly review the most common distributions used in survival analysis. We list their densities, survival and hazard functions (where of interest).

Exponential distribution

The case of the exponential distribution of the survival time T corresponds to the simplest parametric model for survival data. It is denoted as $T \sim \mathcal{E}(\lambda)$.

- $f(t) = \lambda e^{-\lambda t}$, $t > 0$
- $S(t) = e^{-\lambda t}$
- $\lambda(t) = \lambda > 0$

The hazard function $\lambda(t)$ is constant.

Weibull distribution

The Weibull distribution for T is widely used in survival models. It is denoted as $T \sim Weibull(b, c)$, where $b > 0, c > 0$.

- $f(t) = c\frac{t^{c-1}}{b^c}e^{(-t/b)^c}$, $t > 0$
- $S(t) = e^{(-t/b)^c}$
- $\lambda(t) = c\frac{t^{c-1}}{b^c}$

When $c = 1$, the hazard is constant over time and we find the exponential distribution. When $c > 1$, the hazard monotonically increases with time, and with $c < 1$ the hazard monotonically decreases with time.

Log-normal distribution

A survival time T follows a log-normal distribution with parameters (μ, σ^2) if $\ln(T) \ \sim \mathcal{N}(\mu, \sigma^2)$.

- $f(t) = \dfrac{1}{\sigma t}\phi\left(\dfrac{\ln(t) - \mu}{\sigma}\right)$
- $S(t) = 1 - \Phi\left(\dfrac{\ln(t) - \mu}{\sigma}\right)$
- $\mathrm{E}(T) = e^{\mu + \sigma^2/2}$
- $\mathrm{Var}(T) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$

The functions Φ and ϕ in the formulas above denote the normal cumulative distribution function and its derivative, respectively.

Gompertz distribution

The Gompertz distribution is also used in modelling survival, although it is more popular to describe growth. For a $T \sim Gompertz(a, b)$ $(a > 0, b > 0)$, we get:

- $f(t) = ae^{bt}e^{-\frac{a}{b}(e^{bt}-1)}$
- $S(t) = e^{-\frac{a}{b}(e^{bt}-1)}$
- $\lambda(t) = ae^{bt}$

Gumbel distribution

The Gumbel distribution for $W = \log(T_0)$ is defined on the real line and can be used in the accelerated lifetime model to generate an exponential or a Weibull survival time (details can be found later on in this chapter). It has the following characteristics:

- $F(w) = 1 - e^{-e^{w}}$

- $f(w) = e^{w}e^{-e^{w}}$

To avoid confusion, we note that this form of the Gumbel distribution is the less common one, since this distribution is usually viewed as a part of the generalized extreme value distribution, with a cumulative distribution function $F(w) = e^{-e^{-w}}$. For our theory, however, it is more practical to use the other form, introduced above.

To illustrate some of the reviewed distributions, Figure 2.2 shows the survival and the hazard functions for the two following distributions: $\mathcal{E}(0.4)$, $Weibull(2, 1/3)$ and $Gompertz(0.05, 0.3)$.

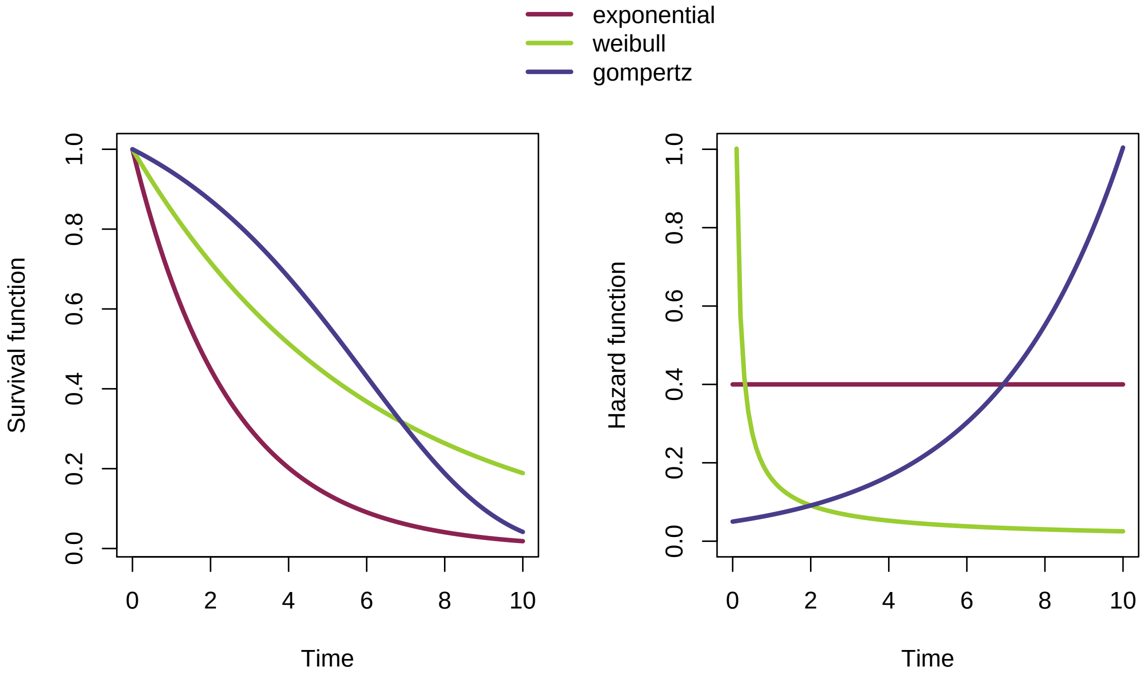

Figure 2.2.: Survival and hazard functions for different distributions.

Weibull distribution under different models

The proportional hazards and the accelerated lifetime specifications are identical only under the Weibull distribution (Cox and Oakes, 1984). We can illustrate this in a simple way. As seen earlier, the survival function for the Weibull distribution is

$$S(t) = e^{(-t/b)^c}.$$

This implies

$$-\ln S(t) = \left(\frac{t}{b}\right)^c$$

or

$$t = b(-\ln S(t))^{1/c}.$$

Now if we put $b = e^{\beta^T x}$, we get

$$t = (-\ln S(t))^{1/c} e^{\beta^T x},$$

which is the accelerated lifetime specification.

On the other hand, the hazard function for the Weibull model is $\lambda(t) = c\frac{t^{c-1}}{b^c}$, which can be re-written under the parametrization $b^{-c} = e^{\beta^T x}$ as

$$\lambda(t) = e^{\beta^T x} c t^{c-1}.$$

By taking ct^{c-1} as our baseline hazard in the formula above, we find the proportional hazards specification.

2.1.5. Simulating the survival times

In Chapter 3, we compare the sliced inverse regression technique for the ALT and PH models. Here we explain the generation of the survival times under those models.

Generating the survival times under the ALT model

To start, we take the logarithm of formula (2.6),

$$\log(T) = \beta^T x + \log(T_0) = \beta^T x + W, \tag{2.7}$$

which can be rewritten in the regression form:

$$Y = \alpha + \sigma u, \tag{2.8}$$

where $Y = \log(T)$, $\alpha = \beta^T x$, and $u = \frac{W}{\sigma}$ is a random error term. Distributional assumptions about u determine the survival distribution of T in the resulting ALT model, and it is straightforward to simulate Y.

- **Exponential distribution:** Suppose T has an exponential distribution with parameter λ, $T \sim \mathcal{E}(\lambda)$. It follows that $T = T_0/\lambda$, where $T_0 \sim \mathcal{E}(1)$. Then we rewrite equation (2.8) as

  $$Y = \log(T) = \log(1/\lambda) + \log(T_0) = \alpha + W,$$

  where $\alpha = -\log(\lambda)$, and W has a standard Gumbel (extreme value) distribution.

- **Weibull distribution:** If T follows a Weibull distribution with the distribution function

  $$F(t) = \mathrm{P}[T \leq t] = 1 - e^{(-t/b)^c},$$

  we find that

  $$\mathrm{P}[Y = \log(T) \leq w] = \mathrm{P}[T < e^w] = 1 - e^{(-e^w/b)^c} = 1 - e^{-e^{c(w-\log(b))}}.$$

  The variable $W = c(Y - \log(b))$ is called a reduced log-Weibull variable. More details on it can be found in (White, 1969). Its distribution is the Gumbel

  $$F_W(w) = \mathrm{P}[W \leq w] = \mathrm{P}[c(W - \log(b)) \leq z] = 1 - e^{-e^w}.$$

  If $T \sim Weibull(b, c)$, then

  $$Y = \log(T) = \alpha + \sigma W,$$

  where $\alpha = \log(b)$, $\sigma = 1/c$ and W has the Gumbel distribution.

- **Log-Normal distribution:** As mentioned earlier, T has a log-normal distribution if its logarithm follows a normal distribution. In this case, we have

$$Y = \log(T) = \alpha + \sigma W,$$

where $W \sim \mathcal{N}(0,1)$.

Generating the survival times under the PH model

The survival function of the Cox's model, for the covariate x can be written as

$$S(t) = e^{-\Lambda_0(t)e^{\beta^T x}},$$

where $\Lambda(t)$ is the cumulative hazard function. Thus, the distribution function is

$$F(t) = 1 - e^{-\Lambda_0(t)e^{\beta^T x}}.$$

If T is the survival time in Cox's model, then

$$U = e^{-\Lambda_0(T)e^{\beta^T x}} \sim U(0,1),$$

where U is a uniformly distributed variable. Bender et al. (2005) cover this topic nicely. If the baseline hazard $\lambda_0(t) > 0$ for all t, then we get

$$T = \Lambda_0^{-1}(-\log(U)e^{-\beta^T x}), \tag{2.9}$$

and it is straightforward to generate survival times using formula (2.9).

- **Exponential distribution:** We saw earlier that the hazard function for the exponential distribution with a scale parameter $\lambda > 0$ is a constant. The inverse cumulative hazard function is given by

  $$\Lambda_0^{-1}(t) = \frac{t}{\lambda_0}.$$

  By plugging this expression into formula (2.9), we get the following expression:

  $$T = -\frac{\log(U)}{\lambda_0 e^{\beta^T x}}. \tag{2.10}$$

- **Weibull distribution:** The inverse of the cumulative hazard function in this case is
$$\Lambda_0^{-1}(t) = b_0 t^{\frac{1}{c_0}},$$
which leads to the following survival time:
$$T = -b_0 \left(\frac{\log(U)}{e^{\beta^T x}} \right)^{\frac{1}{c_0}}. \tag{2.11}$$

- **Gompertz distribution:** For the Gompertz distribution, the inverse hazard function is of the following form:
$$\Lambda_0^{-1}(t) = \frac{1}{b_0} \log \left(1 + \frac{b_0}{a_0} t \right).$$
From equation (2.9) it follows that
$$T = \frac{1}{b_0} \log \left(1 - \frac{b_0}{a_0} \log(U) e^{-\beta^T x} \right) = \frac{1}{b_0} \log \left(1 - \frac{b_0 \log(U)}{a_0 e^{\beta^T x}} \right). \tag{2.12}$$

In this section, we gave an introduction on the purpose of survival analysis and covered its main parameters and models. We saw that while the proportional hazards and the accelerated lifetime models allow for the survival times from the different distributions, not all the combinations are possible. The Gompertz survival time can only be generated under the Cox's model, while the log-normal survival time exists only in terms of the ALT model. These distributions were selected on purpose, to get a better picture of the difference between these two models when evaluating the performance of our SIR algorithm on the survival data in Chapter 3.

2.2. Sliced Inverse Regression

Let us review the theory of the sliced inverse regression and illustrate this procedure on simple examples. We concentrate mostly on the univariate case, that is, the reduction to a single dimension, since it is the simplest to interpret graphically. We cover in detail the case of the linear regression, for which we show the asymptotic distribution to be equivalent to the one found by sliced inverse regression. The adaptation of this method to survival analysis is presented in the next chapter.

2.2.1. Dimension reduction

Dimension reduction is one of the fundamental principles for handling multivariate data. It aims to select a few new variables, usually linear combinations of the original ones, and which describe the most important features of the observed data. It is possible to construct predictive models for high-dimensional data directly (which is usually computationally intensive), but in practice it might be easier to reduce the dimension of the data beforehand. As in this example of searching for predictors, dimension reduction is often used as a preprocessing step. Among the classical examples are principal component analysis and factor analysis (see Muirhead (1982)), probably the most popular linear dimension reduction methods. Among the methods using non-linear reductions, neural networks and self-organizing maps are leading examples.

In this thesis, we concentrate on the problem of dimension reduction for the regression of a response variable Y on a p-dimensional predictor x. The reduction of the dimension of the regressors' space has become quite important in analyzing large datasets, the multivariate response regression analysis with a p-dimensional vector of regressors being a common example. Mathematically, a model in which dimension reduction makes sense can be written as

$$y = f(\beta_1^T x, \beta_2^T x, \dots, \beta_k^T x, \epsilon), \tag{2.13}$$

where the β's are unknown p-vectors, ϵ is a random variable independent of x, and f is an arbitrary unknown function on $\mathbb{R}^{k+1}$.

This model is reasonably general and is the starting point for the development of several methods. In 1991, Li introduced the sliced inverse regression (SIR) (Li, 1991). The same year, Cook and Weisberg (1991) published a discussion on SIR, suggesting a variance checking condition and calling this method sliced average variance estimation (SAVE). In 1992, Li (1992) provided another method, called principal Hessian directions (PHD), to find the inverse structure as well as a test for the dimension k. All three methods were further developed and were explored by many scientists, especially Cook (1994, 1998, 2000). All three of them are implemented in the *R* package **dr** for dimension reduction.

Li continued exploring the variations of SIR by considering second moment based variance methods (Li, 2000) and by comparing the slice covariance matrix with the mean slice covariance matrix, the method he called SIR-II.

Further developments in SIR include multivariate response vectors (Coudret et al., 2012; Li et al., 2003), recursive methods (Bercu et al., 2011) and other special cases. When it comes to the high dimensional predictors, one of the first articles to address this topic is a paper by Zhu et al. (2006), where the authors study the asymptotic behavior of the SIR estimate. A number of regularizations for SIR have been suggested, see papers by (Li and Yin, 2008) and (Scrucca, 2007).

In this thesis we propose a new SIR adaptation to survival data. We are mainly interested in the $k = 1$ case, that is, we look for a good linear combination of the covariates, $\beta^T x$, which can serve as a basis for predicting a variable Y. This case has been studied by Duan and Li (1991) and is reviewed in detail in this chapter.

2.2.2. The basic ideas behind SIR

The idea of the sliced inverse regression is to find a projection of a p-dimensional covariate x onto a k-dimensional linear subspace that contains most of the information about our response Y. If the subspace, rather than the precise basis, is of interest, any β such that $\beta^T x$ lies in the k-dimensional subspace is called an effective dimension reduction (e.d.r.) direction. We note here that the function f does not have to be linear in its components, the method is able to estimate the e.d.r. directions even if the link between Y and the subspace is of more complex form. We concentrate on estimating the e.d.r. directions only, not the form of the function f. We refer to the linear space generated by the β's as the e.d.r. space.

The SIR methodology avoids dealing directly with a possibly high-dimensional covariate vector by switching to the inverse problem. Instead of estimating Y as a function of x, we regress x against Y, which transforms a high-dimensional regression problem into a set of one-dimensional regression problems.

Why does this idea work? Before going into details, we first present an overview of the method. As stated in (Li, 1991), as Y varies, $\mathrm{E}(x \mid Y = y)$ draws a curve, called the inverse regression curve, in $\mathbb{R}^p$. Under the condition (2.13), this curve will oscillate around a k-dimensional affine subspace related to the linear space spanned by $\beta_1^T x, \dots, \beta_k^T x$. Under certain conditions, the inverse regression curve falls into the k-dimensional affine subspace determined by the e.d.r. directions. If the covariates x are standardized to have mean 0 and the identity covariance, then this subspace coincides with the e.d.r. space, allowing us the capture the main direction of the variation. We can rewrite formula (2.13) as

$$y = f(\eta_1^T z, \eta_2^T z, \dots, \eta_k^T z, \epsilon), \tag{2.14}$$

where $\eta_k = \Sigma^{1/2}\beta_k$, Σ being the covariance matrix of x.

The search starts with the standardization of x and proceeds with an estimate of the regression curve $\mathrm{E}(x \,|\, y)$. For that, we slice the sorted response vector y into several intervals and compute the slice means of x corresponding to each slice of y. The principal component analysis on the slice means of x defines the most important k-dimensional subspace for tracking the inverse regression curve $\mathrm{E}(x \,|\, y)$. The e.d.r. directions on the original scale are found by back-transformation.

A basic condition for sliced inverse regression is as follows:

Condition 2.2.1 *For any b in $\mathbb{R}^p$, the conditional expectation $E(b^T x \,|\, \beta_1^T x, \dots, \beta_k^T x)$ is linear in $\beta_1^T x, \dots, \beta_k^T x$; that is, for some constants $c_0, c_1, \dots, c_k$,*

$$E(b^T x \,|\, \beta_1^T x, \dots, \beta_k^T x) = c_0 + c_1\beta_1^T x + \dots + c_k\beta_k^T x. \tag{2.15}$$

This condition is satisfied when the distribution of x is elliptically symmetric. The normal distribution is a leading example.

The following theorem provides a foundation for the SIR method. This result was presented by Li (1991).

Theorem 2.2.2 *(Li, 1991) Under the conditions (2.13) and (2.15), the centered inverse regression curve $E(x \,|\, Y = y) - E(x)$ is contained in the linear subspace spanned by $\Sigma\beta_i$ $(i = 1, \dots, k)$, where Σ denotes the covariance matrix of x.*

Corollary 2.2.3 *Assume that x has been standardized to $z = \Sigma^{-1/2}x$. Then under (2.13) and (2.15), the standardized inverse regression curve $E(z \,|\, y)$ is contained in the linear space generated by the standardized e.d.r. directions $\eta_1, \dots \eta_k$.*

What follows from Corollary (2.2.3), is that the covariance matrix of $\mathrm{E}(z \mid y)$ is degenerate in any direction orthogonal to the linear subspace spanned by $\{\eta_1, \ldots, \eta_k\}$. Therefore, the eigenvectors η_i $(i = 1, \ldots, k)$, corresponding to the largest k eigenvalues of $\mathrm{Cov}(\mathrm{E}(z \mid y))$ are the standardized e.d.r. directions.

To illustrate these results we consider the case for $k = 1$ in more detail. The main condition for the SIR method is the elliptical distribution of covariates, as well as their normalization. Together they ensure the recovery of the direction of the main eigenvector.

Example: Consider $Z \sim \mathcal{N}(0, I)$, where I is the identity matrix, and let $Y \mid Z = z \sim \mathit{Weibull}(b = \exp(\beta^T z),\, c)$. Then the conditional density $f(Y \mid Z = z)$ is of form

$$f(y \mid Z = z) = c\frac{y^{c-1}}{b^c} \exp\left(-(\frac{y}{b})^c\right),$$

and the density of Z is

$$h(z) = (2\pi)^{-p/2} e^{-\frac{1}{2}z^T z}.$$

It follows that

$$\begin{aligned} f(z \mid Y = y) \;&\propto\; f(y \mid Z = z) h(z) \\ &\propto\; \exp\left(-\frac{1}{2}z^T z - c(\beta^T z - \log(y)) - e^{-c(\beta^T z - \log(y))}\right), \end{aligned}$$

which is a product of a spherically symmetric function with a positive function that depends only on $\beta^T z$. In any affine space orthogonal to β, $A_d = \{z \mid \beta^T z = d\} \subseteq R^{p-1}$, the inner product $\beta^T z = d$ is constant, which implies spherical symmetry of the conditional distribution of $Z \mid Y$ within any A_d. Because of this symmetry, the conditional expectation $\mathrm{E}(Z \mid Y)$ must be a multiple of β and $\mathrm{Cov}(\mathrm{E}(X \mid Y))$ is of rank one, with any v such that $\beta^T v = 0$ being an eigenvector with associated eigenvalue equal to zero.

2.2.3. One-component SIR

Theoretical justification

A good review on the one-dimensional (link-free regression) e.d.r. space can be found in (Duan and Li, 1991), where the authors also establish the asymptotic theory for the SIR estimate. We now give a summary of their results.

They assume that the true model is of the form

$$Y = f(\beta^T x, \epsilon), \quad \epsilon \,|\, x \sim F(\epsilon), \tag{2.16}$$

where f is a kind of link function, mapping $(\beta^T x,\, \epsilon)$ into $\mathbb{R}$, F is the error distribution, and x and β are in $\mathbb{R}^p$. We rewrite the condition (2.2.1) in a stronger version:

Condition 2.2.4 *The regressor variable x is sampled randomly from a nondegenerate elliptically symmetric distribution.*

Theorem 2.2.5 *Duan and Li (1991) Assume the general regression model (2.16) and the design condition (2.2.4). The inverse regression function falls along a line*

$$E(X \,|\, Y = y) = \mu + \Sigma\beta\kappa(y) \;\in \mathbb{R}^p, \tag{2.17}$$

where $\mu = E(X)$, $\Sigma = Cov(X)$ and $\kappa(y)$ is a scalar function of y, namely

$$\kappa(y) = \frac{E(\beta^T(X - \mu) \,|\, Y = y)}{\beta^T\Sigma\beta}.$$

Proof: The design condition (2.2.4) implies that

$$\mathrm{E}(X \,|\, \beta^T X) = \mu + \frac{\Sigma\beta\beta^T(X - \mu)}{\beta^T\Sigma\beta}. \tag{2.18}$$

By applying the law of iterated expectations, we get that

$$\mathrm{E}(X \,|\, Y = y) = \mathrm{E}(\mathrm{E}(X \,|\, \beta^T X) \,|\, Y = y) = \mu + \frac{\Sigma\beta}{\beta^T\Sigma\beta}\mathrm{E}(\beta^T(X - \mu) \,|\, Y = y). \quad \blacksquare$$

Discussion of this proof: While the second part of this proof is quite clear, the result (2.18) may not seem straightforward. We explore this statement in an example similar to the one earlier in this chapter.

Since X is elliptical we have $X = \Sigma^{1/2}U$, where U is spherically symmetric with $\mathrm{E}(UU^T) = I$.

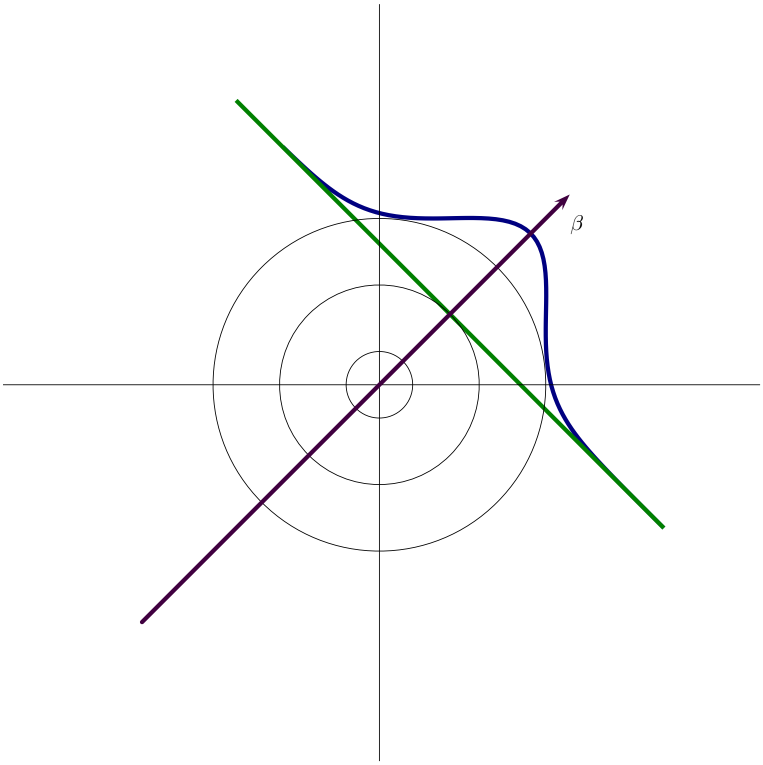

Figure 2.3.: An illustration to Theorem 2.2.5.

In Figure 2.3 we illustrate the spherically symmetric density (represented by the circles) and the conditional expectation $\mathrm{E}[U \mid \beta^T U = y]$ (in blue). Due to the symmetry, the mean of $U \mid \beta^T U = y$ lies in the linear subspace spanned by β. The scalar function $\kappa(y)$ is defined as

$$\mathrm{E}(U \mid \beta^T U = y) = \kappa(y)\beta, \tag{2.19}$$

that is, it determines how far from the origin the conditional mean lies.

Now we want to generalize the relationship to X:

$$\mathrm{E}(X = \Sigma^{1/2}U \mid \beta^T\Sigma^{1/2}U = y) = \Sigma^{1/2}\mathrm{E}(U \mid \tilde{\beta}^T U = y) = \Sigma^{1/2}\kappa(y)\tilde{\beta} = \Sigma\beta\kappa(y),$$

where $\tilde{\beta} = \beta^T\Sigma^{1/2}$ and we used formula (2.19).

Corollary 2.2.6 *Duan and Li (1991) Assume the same conditions as in Theorem 2.2.5. Let* $V = Cov(E(X \mid Y = y))$. *The slope vector* β *solves the following maximization problem:*

$$\max_{b \in \mathbb{R}^p} L(b), \quad \textit{where } L(b) = \frac{b^T V b}{b^T \Sigma b}. \tag{2.20}$$

The solution is unique (up to a multiplicative scalar) if and only if $\kappa(y) \neq 0$.

Proof: From (2.17) we get that

$$V = \mathrm{Var}(\kappa(Y)) \Sigma \beta \beta^T \Sigma$$

has rank one. This is most easily viewed graphically, as in Figure 2.3, the variation of $\mathrm{E}(X \mid Y)$ is in only one direction, namely β. The rest of the corollary follows from the Cauchy inequality. ▮.

Remark: Later on we compute the form of the matrix V and its eigenvalues for the normal covariates X and show mathematically that its rank in this case is 1.

Thus, β is the principal eigenvector for V, corresponding to the principal eigenvalue λ_1, and all the other eigenvalues are equal to zero. We can rewrite the function L in (2.20) as

$$L(b) = \frac{\mathrm{Var}(\mathrm{E}(b^T X \mid y))}{\mathrm{Var}(b^T X)}. \tag{2.21}$$

Since $L(b)$ in equation (2.21) compares the explained variance with the total variance of the data, it can be viewed as the R^2-coefficient for the nonparametric regression of $\beta^T X$ on Y, hence it measures how well we predict $\beta^T X$ from Y. The corollary states that among all linear combinations $b^T X$, Y predicts $\beta^T X$ the best.

Theorem 2.2.5 and Corollary 2.2.6 provide the theoretical foundation for the inverse regression. Their results can be applied in the sampling case. For a given sample $(y_i, x_i,\ i = 1, \ldots n)$ from the general model (2.16), the idea is to estimate the direction of the slope β. The estimate of the inverse regression curve $\mathrm{E}(X \mid Y = y)$ is necessary for the application of Theorem 2.2.5, and the idea behind sliced inverse regression is to use a step function estimate. For that, the whole range of y_i's is divided into S slices, $H_1, \ldots, H_S$. Within each slice, $\mathrm{E}(X \mid y \in H_s)$ can be estimated by the sample average of the corresponding x's. The estimated inverse regression curve can be written as

$$\hat{\mathrm{E}}(X \mid y \in H_s) = \hat{\xi}_s = \frac{\sum_{i=1}^n x_i I_{H_s}(y_i)}{\sum_{i=1}^n I_{H_s}(y_i)} \quad \text{if } y \in H_s,$$

where $I_{H_s}(y_i)$ is the indicator of the event that y_i in in the slice H_s.

The estimated inverse regression curve converges to the true one if we choose a suitable number of slices whose meshes decrease to zero as $n \to \infty$. Since $\mathrm{E}(X \mid Y = y)$ falls along a line, as shown earlier, a crude estimate for $\mathrm{E}(X \mid Y = y)$ is adequate for estimating its direction (Duan and Li, 1991).

We can rewrite Theorem 2.2.5 for the sampling case in the following way:

$$\xi_s = \mathrm{E}(\hat{\xi}_s) = \mu + \Sigma\beta k_s, \tag{2.22}$$

where

$$k_s = \mathrm{E}(\kappa(Y) \mid Y \in H_s) = \frac{\mathrm{E}((X-\mu)^T\beta \mid Y \in H_s)}{\beta^T\Sigma\beta}. \tag{2.23}$$

The idea is to combine the information from all the slices (where κ_s is nonzero) to estimate the direction of β. We introduce the following notation:

$$p_s = \mathrm{P}(y \in H_s), \quad k = (k_1, \dots, k_s)^T, \quad \xi = [\xi_1, \dots, \xi_s], \quad \hat{\xi} = [\hat{\xi}_1, \dots, \hat{\xi}_s].$$

We estimate V by

$$\hat{V} = \hat{\xi} W \hat{\xi}^T, \tag{2.24}$$

where W is an arbitrary symmetric nonnegative definite $s \times s$ matrix, chosen a priori, which satisfies $W\mathbb{1} = 0$ (Duan and Li, 1991). Equation (2.24) can be interpreted as a weighted covariance matrix for the $\hat{\xi}$ with a weight matrix W.

We can now rewrite the maximization problem (2.20) in a sample form for a given weight matrix W:

$$\max_{b \in \mathbb{R}^p} \hat{L}(b), \text{ where } \hat{L}(b) = \frac{b^T \hat{V} b}{b^T \hat{\Sigma} b}. \tag{2.25}$$

As stated earlier, the maximum

$$\hat{\lambda}_1 = \hat{L}(\hat{\beta}) \tag{2.26}$$

is the principal eigenvalue, and the slicing regression estimate $\hat{\beta}$ is the principal eigenvector.

There are many possible choices for the weight matrix W, but the main algorithm of SIR uses the *proportional to size* weight matrix. That means, each slice is weighted by the empirical probability of Y to fall inside the slice. This choice allows for a great simplification in the variance estimation.

Algorithm

The steps for computing the SIR estimate of $\hat{\beta}\ \ (k=1)$ are as follows (Haerdle and Simar, 2007). We operate on the data $(y_i, x_i,\ i=1,\dots,n)$, and $x_i \in \mathbb{R}^p$ is a vector containing the covariates for the i-th observation.

1. Standardize x to get
$$z_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x}),$$
where $\hat{\Sigma}^{-1/2}$ and $\bar{x}$ are the sample covariance matrix and sample mean of $x_1,\dots,x_n$, respectively.

2. Divide the range of $y_1,\dots,y_n$ into S nonoverlapping slices H_s, $s=1,\dots,S$. n_s denotes the number of observations within slice H_s, and I_{H_s} the indicator function for this slice:
$$n_s = \sum_{i=1}^{n} I_{H_s}(y_i).$$

3. Compute the sample mean of z_i over all slices, denoted by $\bar{z}_s$:
$$\bar{z}_s = \frac{1}{n_s}\sum_{i=1}^{n} z_i I_{H_s}(y_i).$$

4. Calculate the estimate for the weighted covariance matrix
$$\hat{V} = n^{-1}\sum_{s=1}^{S} n_s \bar{z}_s \bar{z}_s^T. \tag{2.27}$$

5. Identify the eigenvalues $\hat{\lambda}_i$ and eigenvectors $\hat{\eta}_i$ of $\hat{V}$.

6. Transform the standardized directions $\hat{\eta}_i$ back to the original scale.

$$\hat{\beta}_i = \hat{\Sigma}^{-1/2} \hat{\eta}_i.$$

In case of the one-component model ($k = 1$ or link-free regression by Duan and Li (1991)) we are only interested in the first eigenvector, while under the multicomponent model (2.13), the coefficients of interest are contained in the first k eigenvectors.

2.2.4. Asymptotic behavior

Determining the asymptotic distribution of the SIR estimator is a challenging problem. Li (1991) discusses how the choice of the number of slices may affect the asymptotic variance of $\hat{\beta}$. It is also stated that it is possible to establish the asymptotic normality of $\hat{\beta}$ and to calculate the asymptotic covariance matrices, which is treated in (Duan and Li, 1991). The formal proof of asymptotic normality can be found in (Saracco, 1997). Li also proposes a test for the principal eigenvector β_1, but states that such a test is not valid for the confidence interval (Li, 1991).

The asymptotic variance being rather complex, many papers studied the convergence of the SIR estimate. Hsing and Carroll (1992) list conditions under which $\hat{\beta}$ converges at the rate of $O_p(n^{-1/2})$. Zhu and Ng (1995) studied the consistency under a fixed number of observations per slice and when this number goes to infinity. The limiting behavior of the eigenvectors and eigenvalues of the matrix $\mathrm{E}(\mathrm{Cov}(X \mid Y = y))$ is also investigated in their paper.

The main results are established by Duan and Li (1991), but their formula does not allow an easy interpretation or even application. We present here the details and show their equivalency to the classical least squares formula in case of the simple linear regression. All the notations are the same as in Section 2.2.3.

Theorem 2.2.7 *Duan and Li (1991) Assume the general regression model (2.16), the design condition 2.2.4 (normal), and let the weight matrix W be symmetric and nonnegative definite with $W\mathbb{1} = 0$ and $k^T W k > 0$. The slicing regression estimate $\hat{\beta}$, which solves the maximization problem (2.25), is consistent for the direction of β. The estimated principal* *eigenvalue $\hat{\lambda}_1$ in (2.26) is a consistent estimate for the population principal eigenvalue*

$$\lambda_1 = k^T W k. \tag{2.28}$$

By the strong law of large numbers, $\hat{V}$ converges almost surely to

$$\xi W \xi^T = k^T W k \Sigma \beta \beta^T \Sigma,$$

which is proportional to V. Hence, both $\hat{L}(b)$ and $\tilde{L}(b)$ converge to a criterion function proportional to $L(b)$ in (2.20). The results follows from Corollary 2.2.6.

For simplicity, suppose that the design distribution is normal. General formulas can be found in Duan and Li (1991). We also assume that the β has been normalized to have unit length in the Σ - metric:

$$\beta^T \Sigma \beta = 1. \tag{2.29}$$

We also impose the normalization of the slicing regression estimate:

$$\hat{\beta}^T \hat{\Sigma} \hat{\beta} = 1. \tag{2.30}$$

We introduce a new parameter

$$u = Wk = (u_1, \ldots, u_H)^T,$$

where W is the weight matrix from equation (2.24) and $k = (k_1, \ldots, k_H)$ are defined by (2.23). The asymptotic results are presented in the following theorem:

Theorem 2.2.8 *Duan and Li (1991) Assume the general regression model (2.16), the design condition 2.2.4 (normal), the normalization (2.29) with a symmetric and nonnegative definite matrix W, which satisfies $W\mathbb{1} = 0$ and $k^T W k > 0$. The slicing regression estimate, which solves the maximization problem (2.25) and is normalized by (2.30), has the following normal approximation:*

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathcal{N}(0, \; A(\Sigma^{-1} - \beta\beta^T)),$$

where the scalar A is given by

$$A = \frac{\sum_{s=1}^{S} u_s^2 / p_s}{(u^T k)^2}.$$

This result is presented for the normal design distribution but can be simplified in case of the weight matrix proportional to size, which we use in our algorithm. The scalar A in Theorem 2.2.8 is equal to

$$A = \frac{1}{u^T k} = \frac{1}{\lambda_1}, \tag{2.31}$$

which can be estimated consistently by substituting $\hat{\lambda}_1$ for λ_1.

The proof of Theorem 2.2.8 is sketched in Duan and Li (1991).

Proof: Without loss of generality, assume that $u^T k = 1$. Throughout the proof, we leave out the terms of lower order. We approximate $\hat{V}$ by

$$\begin{aligned}\hat{V} = \hat{\xi} W \hat{\xi}^T &= (\xi + (\hat{\xi} - \xi)) W (\xi + (\hat{\xi} - \xi))^T \\ &\doteq \xi W \xi^T + (\hat{\xi} - \xi) W \xi^T + \xi W (\hat{\xi} - \xi)^T \\ &= \Sigma \beta k^T W k \beta^T \Sigma + (\hat{\xi} - \xi) W k \beta^T \Sigma + \Sigma \beta k^T W (\hat{\xi} - \xi)^T \\ &= (\Sigma \beta + (\hat{\xi} - \xi) u)(\Sigma \beta + (\hat{\xi} - \xi) u)^T. \end{aligned} \tag{2.32}$$

If we put $\Delta = (\hat{\xi} - \xi) u$, (2.32) becomes

$$\hat{V} \doteq (\Sigma \beta + \Delta)(\Sigma \beta + \Delta)^T.$$

Thus, the slicing regression estimate maximizes

$$\tilde{L}(b) = \frac{[b^T (\Sigma \beta + \Delta)]^2}{b^T \Sigma b}. \tag{2.33}$$

The problem in (2.33) can be viewed as

$$\arg\max\, (b^T a)^2$$

subject to the constraint

$$b^T \Sigma b = 1,$$

which is equivalent to

$$\arg\max\, [b^T a - \frac{\lambda}{2}(b^T \Sigma b - 1)],$$

where λ is a Lagrangian multiplier. From this formulation, it follows that $b \propto$

$\Sigma^{-1}a. = \Sigma^{-1}(\Sigma\beta + \Delta)$. In Section 2.2.3 we saw that $\hat{\beta}$ maximizes (2.33). We can write that

$$\hat{\beta} \propto \Sigma^{-1}(\Sigma\beta + \Delta) = \beta + \Sigma^{-1}\Delta.$$

The denominator in (2.33) is approximated

$$(\beta + \Sigma^{-1}\Delta)^T\Sigma(\beta + \Sigma^{-1}\Delta) \doteq \beta^T\Sigma\beta + 2\Delta^T\beta = 1 + 2\Delta^T\beta. \tag{2.34}$$

Again, we leave out the terms of the lower order. We apply the Taylor expansion to (2.34) and finish the computation of the normalized $\hat{\beta}$:

$$\hat{\beta} = \frac{\beta + \Sigma^{-1}\Delta}{\sqrt{1 + 2\Delta^T\beta}} = (\beta + \Sigma^{-1}\Delta)(1 - \Delta^T\beta) \doteq \beta + (\Sigma^{-1} - \beta\beta^T)\Delta. \tag{2.35}$$

The right-hand side of (2.35) is asymptotically normal with mean β. Its covariance equals

$$\mathrm{Cov}(\hat{\beta}) = (\Sigma^{-1} - \beta\beta^T)\mathrm{Cov}(\Delta)(\Sigma^{-1} - \beta\beta^T). \tag{2.36}$$

The last thing to compute is $\mathrm{Cov}(\Delta)$. The term $\Delta = (\hat{\xi} - \xi)u$ can be written as

$$\Delta = u_1(\hat{\xi}_1 - \xi_1) + u_2(\hat{\xi}_2 - \xi_2) + \ldots + u_s(\hat{\xi}_s - \xi_s),$$

and its covariance is of form

$$\mathrm{Cov}(\Delta) = \sum_{s=1}^{S} u_s^2 \mathrm{Cov}(X \mid Y \in H_s)\frac{1}{np_s}.$$

The design condition 2.2.4 implies that the covariance of $X \mid \beta^T X$ is of form

$$\mathrm{Cov}(X \mid \beta^T X) = \Sigma - \Sigma\beta\beta^T\Sigma, \tag{2.37}$$

and it follows that

$$\mathrm{Cov}(\Delta) = \frac{\sum_{s=1}^{S} u_s^2/p_s}{n}(\Sigma - \beta\beta^T\Sigma), \tag{2.38}$$

and plugging equation (2.38) into (2.36) we get

$$\mathrm{Cov}(\hat{\beta}) = (\Sigma^{-1} - \beta\beta^T)\frac{\sum_{s=1}^{S} u_s^2/p_s}{n}(\Sigma - \beta\beta^T\Sigma)(\Sigma^{-1} - \beta\beta^T)$$

$$= \frac{\sum_{s=1}^{S} u_s^2 / p_s}{n} (\Sigma^{-1} - \beta\beta^T), \tag{2.39}$$

which concludes the proof. ∎

Remark: The matrix $(\Sigma^{-1} - \beta\beta^T)$ in (2.39) is not of rank p, but of $p-1$. This is due to the chosen normalization, $\beta^T\Sigma\beta = 1$, which was used in the final step of the proof. This result implies, that the variation in the direction parallel to β is zero. This is most easily shown with the parametrization $(\Sigma^{-1} - \beta\beta^T)\Sigma\beta = \beta - \beta = 0$. Thus, the algorithm estimates the β precisely. This is shown in the next section for the linear regression, with the $\mathrm{Cov}(\mathrm{E}(X \mid Y))$ having only one non-zero eigenvalue.

2.2.5. Linear regression

The asymptotic theory reviewed above is valid for the one-component SIR which assumes a general form of dependency between the response Y and the $\beta^T X$ and ϵ. The linear regression is a special case. In this section, we derive explicit results for the SIR estimator in the case of the linear regression.

Let $X \sim \mathcal{N}(0, \Sigma)$ be our explanatory variables, and consider the regression conditional on the observed x:

$$Y = \beta^T x + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \text{ independent of } x. \tag{2.40}$$

We shall start with the conditional distribution of $X \mid Y$, whose density is

$$f_{X \mid Y}(x \mid Y = y) = \frac{f_X(x) f_{Y \mid x}(y \mid x)}{f_Y(y)} \propto f_X(x) f_{Y \mid x}(y \mid x). \tag{2.41}$$

It follows that

$$f_{X \mid Y}(x \mid Y = y) \propto \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right) \exp\left(-\frac{1}{2\sigma^2}(y - \beta^T x)^2\right)$$

$$\propto \exp\left(-\frac{1}{2}(x^T \Sigma^{-1} x) + \frac{1}{\sigma^2}(\beta^T x)^2 - \frac{2}{\sigma^2}\beta^T x y\right) \tag{2.42}$$

$$\propto \exp\left(-\frac{1}{2}(x - \mu_1)^T V^{-1} (x - \mu_1)\right). \tag{2.43}$$

Thus, the conditional distribution of $X \,|\, Y = y$ is normal and we can easily deduce its parameters μ_1 and V. The conditional distribution is

$$X \,|\, Y \sim \mathcal{N}\left(\frac{y}{\sigma^2}\left(\Sigma^{-1} + \frac{1}{\sigma^2 \beta\beta^T}\right)^{-1}\beta,\; V = \left(\Sigma^{-1} + \frac{1}{\sigma^2 \beta\beta^T}\right)^{-1}\right). \tag{2.44}$$

In the context of SIR, formula (2.44) simplifies. Since we standardize the explanatory variable, Σ is an identity matrix. Moreover, our vector of interest β is an eigenvector for the covariance matrix V of the conditional distribution (2.44). We can show this by considering the inverse covariance matrix (since the eigenvectors stay the same).

$$V^{-1}\beta = \left(I + \frac{1}{\sigma^2}\beta\beta^T\right)\beta = \beta + \frac{1}{\sigma^2}\beta\beta^T\beta = \beta\left(1 + \frac{1}{\sigma^2}||\beta||^2\right) = \beta\left(1 + \frac{1}{\sigma^2}\right),$$

where we have used the fact that our eigenvectors are normalized to unit length. Thus, β is an eigenvector of the covariance matrix V^{-1} with eigenvalue $\frac{1+\sigma^2}{\sigma^2}$. We note that due to the orthogonality of eigenvectors $\beta_j, (j = 2, \ldots, p)$, the other eigenvalues of the matrix V^{-1} (and of V) are all equal to 1.

The expression for the covariance matrix V can be rewritten by considering its spectral decomposition.

$$\begin{aligned} V = \operatorname{Cov}(X \,|\, Y = y) &= \frac{\sigma^2}{1+\sigma^2}\beta\beta^T + \sum_{i=2}^{p} \beta_i\beta_i^T \cdot 1 + \beta\beta^T - \beta\beta^T \\ &= \frac{\sigma^2}{1+\sigma^2}\beta\beta^T + I - \beta\beta^T = I - \frac{1}{1+\sigma^2}\beta\beta^T. \end{aligned} \tag{2.45}$$

Duan and Li (1991) show that $\operatorname{Cov}(X \,|\, \beta^T X) = (I - \beta\beta^T)$, but in our model (2.40) we take ϵ into account directly, which brings us to (2.45).

From the law of the total variance we know that

$$\operatorname{Var}(X) = \operatorname{E}(\operatorname{Cov}(X \,|\, Y = y)) + \operatorname{Cov}(\operatorname{E}(X \,|\, Y = y)). \tag{2.46}$$

If we take X as our (standardized) covariates, equation (2.46) becomes

$$I = \operatorname{E}(V) + \tilde{V},$$

where the matrix V is defined by (2.45), and the matrix $\tilde{V}$ is the covariance matrix computed by the SIR algorithm (2.27). If we assume that the slices are small enough, then the variance within each slices could be considered constant. In this case, using formula (2.45) we find that

$$V = I - \tilde{V} = \frac{1}{1+\sigma^2}\beta\beta^T. \tag{2.47}$$

We see that for the linear regression, the matrix V indeed would only have one non-zero eigenvalue, which allows the accurate recovery of β.

In order to relate Theorem 2.2.8 to the classical regression, we have to consider the effect of normalization of the covariates. The ordinary least squares estimator based on a sample of n observations has the following distribution:

$$\hat{\beta}_{OLS} \sim \mathcal{N}\left(\beta, \frac{\sigma^2(X^TX)^{-1}}{n}\right),$$

where X is the design matrix. Since we standardize the covariate x, the matrix $(X^TX)^{-1}$ satisfies $(X^TX)^{-1} = I$. We now show that the variance of $\hat{\beta}_{OLS}/||\hat{\beta}_{OLS}||$ is equal to the variance of $\hat{\beta}_{SIR}$. To do that we shall apply the delta method with the transformation $g(\beta) = \frac{\beta}{\sqrt{\beta^T\beta}}$. We have

$$\frac{\partial}{\partial\beta}\frac{\beta}{\sqrt{\beta^T\beta}} = \frac{1}{||\beta||}\left(I - \beta\beta^T\frac{1}{||\beta||^2}\right) = A$$

and the asymptotic variance of the $g(\beta)$ becomes

$$\begin{aligned}\mathrm{Var}_{asy}(g(\hat{\beta})) = \sigma^2 AA^T\Big|_{\beta} &= \left(I - 2\frac{\beta\beta^T}{||\beta||^2} + \frac{\beta\beta^T}{||\beta||^2}\right)\frac{\sigma^2}{||\beta||^2} \\ &= \frac{\sigma^2}{||\beta||^2}\left(I - \frac{\beta\beta^T}{||\beta||^2}\right). \end{aligned} \tag{2.48}$$

Theorem 2.2.8 states that the asymptotic distribution of $\hat{\beta}_{SIR}$ satisfies

$$\sqrt{n}(\hat{\beta}_{SIR} - g(\beta)) \sim \mathcal{N}(0, \frac{1}{\lambda_1}(I - \frac{\beta\beta^T}{||\beta||^2})). \tag{2.49}$$

The formulas (2.48) and (2.49) are equivalent with $\lambda^{-1} = \sigma^2/||\beta||^2$.

Remark: Note that the model $X \sim \mathcal{N}(0, I)$ and $Y = \beta^T X + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is equivalent to

$$Y = g(\beta)^T X + \epsilon$$

with $X \sim \mathcal{N}(0, \frac{1}{||\beta||^2} I),\ \epsilon \sim \mathcal{N}(0, I)$, which explains the presence of the factor $\sigma^2/||\beta||^2$ in (2.48).

2.2.6. Discussions

The SIR procedure reduces the dimension of the predictor whevener a model of type (2.13) exists. The structure of the underlying function f is impossible to identify, only the inverse regression curve can be estimated. When identifying the β's in (2.13), we identify the e.d.r. space, not the $\beta_1, \dots, \beta_k$ individually. As presented later in this thesis, SIR can recover the main dependencies in both Cox's proportional hazards model and the accelerated lifetime model, despite their very different nature.

In this thesis, we do not consider the problem of convergence, but this has been largely studied and pertinent results can be found, for instance, in (Li, 1991; Zhu et al., 2006; Zhu and Ng, 1995). In general, the method provides root n consistent estimates for the e.d.r. directions.

How strong is the linearity condition (2.15)? First of all, it is imposed on true e.d.r. directions, the estimated directions may not fully satisfy this condition. Its possible relaxation was mentioned by Li (1989), and its robustness was studied in (Hall and Li, 1993), since low-dimensional projections often fall into the spherical space. Cook and Weisberg (1991) report that SIR is often not overly sensitive to the linear design condition. Li (1991) also states that by density estimation and reweighting, X can be forced into elliptical symmetry. Coudret et al. (2012) state that using the Bayesian argument from (Li, 1989) it can be inferred that the condition (2.15) approximately holds for many high-dimensional datasets.

When will the method fail to correctly identify the e.d.r. space? Li (1991) gives an example of the standardized inverse regression curve which falls within a proper subspace of the standardized e.d.r. space. In this case, the e.d.r. directions will not be completely recovered. Another limitation of SIR is in finding patterns symmetric about the vertical axes, a better method for such problem would be SAVE (Cook and Weisberg, 1991) or PHD (Cook, 1998; Li, 1992).

A good review and discussion of all the aforementioned features of SIR can be found in (Chen and Li, 1998), which draws comparisons between the SIR and the multiple regression estimates. Based on the theory and examples, they argue that SIR is a simple method which finds linear combinations of independent variables that maximize the correlation with the optimally transformed dependent variable. "It is powerful when it comes to interactive, multi-dimensional graphing techniques. It can be used whenever there is a need for visualization, which in turn can help functional approximation". Moreover, it can be used together with other methods which we shall discuss in the next chapter.

CHAPTER

3 SIR FOR SURVIVAL DATA

In the previous chapter, we reviewed in detail the sliced inverse regression (SIR) as a method for dimension reduction in regression problems. Here we shall discuss its adaptation to survival data. We start with an overview of literature on the topic, and then introduce a method for dealing with censored observations. We cover some aspects of the variance estimation and briefly discuss the high-dimensionality scenario.

3.1. Background

The first adaption of the SIR method in survival case was suggested by the author of the original method himself, Li et al. (1999). The paper distinguishes two censoring scenarios, the independence of the censoring distribution from the covariates and the true survival time, and the the independence of the censoring distribution from the true survival time, conditional on the covariates. In the first case, Li argues that the censoring does not introduce bias to the method, and that the SIR algorithm can be applied to the whole dataset without any modifications. When the independence between the censoring time and the true survival time holds but the covariates influence the censoring pattern, the idea is based on a weighting scheme to bypass the bias in estimating the slice means. The weight function is estimated via kernel smoothing techniques. However, they perform well only when the number of covariates is small ($p \leq 3$). When this is not the case, the estimation of the weights is preceded by location of the joint e.d.r. directions which allow for the dimension reduction. The joint e.d.r. directions are computed through the double slicing technique (slicing the observed and the censored cases separately and combining them at the third step of the algorithm, when computing the sliced mean). Under the certain conditions the estimates of this two-step procedure are root-n consistent. This method is implemented in the R package **censorSIR**.

Cook (2003) elaborated a version of SIR for bivariate responses, bivariate SIR, as an alternative procedure for survival data. His approach generalizes the double slicing of Li et al. (1999), which becomes a special case of the bivariate response. The proposed method can be found in the supplement to their software **Arc**, a program for linear regression analysis. Recently, in a paper about the model-free dimension reduction for bivariate regression, Wen and Cook (2009) introduced a new approach, called bivariate estimation across responses (BEAR), which allows the inclusion of a categorical response (an indicator function for censoring in the survival analysis case) and is based on the minimization of the quadratic discrepancy function, introduced by Cook and Ni (2005). Asymptotic theory for the BEAR estimator is presented as well.

An alternative to bivariate dimension reduction method is based on reweighting the censored observations. Li and Lu (2008) explored the sliced inverse regression with missing predictors with the help of the augmented inverse probability weighted estimator. The augmented version of the inverse probability weighted estimators is considered in order to obtain unbiased estimators, even if the model of the missingness indicator is misspecified. This approach is suitable under the different setups of missing data. Later on, Lu and Li (2011) expand their work from 2008 by focusing on the censored regression only. They employ the inverse censoring probability weighted estimator in order to handle censored responses. The authors obtain estimates of the e.d.r. directions by introducing the inverse of the survival function of the censored time as weight to the uncensored observations. Their weighted linear squares system has a closed-form solution. Among the ways to estimate the survival function of the censoring time, they call for the Kaplan-Meier method or adopt the semi-parametric approach by imposing the proportional hazards model and estimating the survival function based on its fit. They go further by considering variables selection via regularized sparse estimation, which is done by the lasso technique. A similar adaptation can be found in a paper by Nadkarni et al. (2011), introducing an inverse regression family for censored data, with SIR being a member of this family. To adjust for the censoring, inverse probability of censoring weighting is used. It is applied for the nonparametric estimation of the weighed Kaplan-Meier estimator for the censored time, the Kaplan-Meier estimator for the lifetime, and in the estimation of the sample estimators. For the basis estimation, the authors refer to the concept of inverse regression approach with a quadratic discrepancy function. Detail on inverse probability weighting can be found, for example, in Rotnitzky and Robins (2005).

To our knowledge, the first paper to mention high-dimensional covariates in SIR is Zhu et al. (2006). The main emphasis in put on estimating the dimension via the Bayes information criterion. More specific problems of linear dimension reduction methods under high-dimensionality are discussed by Li (2010). The author presents three possible ways to handle the $n < p$ problem. The first option is to use a two-way procedure, first reducing the dimension of predictors, and then applying SIR. Such an approach has been used in the microarray data analysis of Li and Li (2004), where the principal components analysis is used for dimension reduction, and the components serve as input data for the SIR algorithm later on. This algorithm is applied on a diffuse large B-cell lymphoma dataset and the results are compared with existing methods. Another example incorporating a two-step procedure is a paper by Wu et al. (2008). Instead of the principal component analysis, its authors preselect the genes using the liquid association measure (a way to characterize three-way interactions between genes) and the correlation with the Kaplan-Meier imputed survival probabilities.

The second way to deal with high-dimensionality is to use the partial least squares method, and the third one is to introduce some kind of regularization. Zhong et al. (2005) suggest the ridge regularization, by adding the identity matrix multiplied by a regularization parameter s to the $\mathrm{Cov}(\mathrm{E}(X \mid Y))$, but this is certainly not the only possibility. Lue et al. (2011) rely on an imputed spline approach to principal Hessian directions to reduce the dimension of covariates.

When it comes to a more general approach of dimension reduction in survival cases, there are many more papers addressing this topic. Some of them consider both the problems of high-dimensionality and censoring, others focus on censoring only. Witten and Tibshirani (2010) give a good review on the existing techniques, they not only discuss high-dimensional genomic data but introduce also the main aspects of survival analysis. Among the methods which are covered, one can count from stepwise selection and shrinkage methods to variance-based methods such as principal components and partial least squares. They treat SIR as in the paper by Li and Li (2004) and under the Cox's proportional hazards model. The most recent paper on dimension reduction for survival data is by Yan and Zhang (2012), where the authors study the estimation and variable selection via an iterative method which is a combination of L_1 penalty and the refined outer product of gradient method (OPG), which they call sparse hazard-function-based OPG algorithm.

All of the afore-mentioned method assume uninformative censoring, when the reasons for removal are unrelated to the event and does not bias the parameter estimation. When the independence condition is violated, information about the censoring mechanism is needed to adjust for the bias. Some of the papers on how to test for the informative censoring or to account for it, include (Koziol and Green, 1976; Lee and Wolfe, 1998; Scharfstein and Robins, 2002). The research on dimension reduction under the presence of informative censoring is ongoing.

3.2. Adaptation of SIR to censored data

What is the best way to handle the censored observations? One strategy, called the complete-case analysis, is to remove any missing datapoints. Such an approach can be judged appropriate when most of the data is complete, but it soon becomes inefficient, once the proportion of the censored cases increases. Moreover, this complete-case analysis is likely to create a bias in estimation.

Our idea is to propose a simple method of reintroducing the censored observations to the data slices. We concentrate on the low-dimensional covariates $(n > p)$, just as the classical algorithm of Li et al. (1999). Some suggestions on how to treat the high-dimensional setup can be found in the literature review in Section 3.1. We will explore another method of reweighting the censored cases, but more intuitively based and simpler than the inverse probability weighting.

We start with considering our response variable. Why does that matter? The classical method of SIR requires a relationship of some form between the response and the covariates. For the accelerated lifetime model, the logarithm of the survival time, $\log(T)$, depends on a linear function of the variables. For Cox's model, such a relationship exists for the hazards. While the reviewed papers consider both the survival time T and its logarithm, we opted to slice the survival time directly.

What does a right-censoring time imply? Given that the individual was censored at time t which falls in the slice i, the event for this individual could have taken place anytime after t. Based on this idea, and assuming that the slice sizes are small enough, we attribute this event with equal weights to all consequent slices. The total sum of the weights naturally equals one. This allows us to use the covariate information of the censored observations. Let us illustrate this procedure with a small example:

Suppose we have 7 observations, listed below:

$$10, 11^*, 13, 15^*, 17^*, 18, 20. \tag{3.1}$$

For this data we choose four slices: 10-12, 13-15, 16-18 and 19-20. Then we create a matrix of weights which shows in which slice each observation falls. The first censored observation 11^* is in the middle of the first slice, hence it is assigned to the second half of this slice and to the next 3 slices, giving a weight of $1/7$ to the first slice and $2/7$ to the slices 2-4. The last two censored observations 15^* and 17^* will be taken into account in the slices 3 and 4 but with the different weights. The observation 15^* is considered with the equal weights of $1/2$, while the 17^* will have the weight of $1/3$ in the slice 3 and $2/3$ in the slice 4. This matrix, listed in Table 3.1 is used at the third step of the SIR algorithm, when computing the slice mean for

obs	slice 1	slice 2	slice 3	slice 4
10	1	0	0	0
11*	0.14	0.29	0.29	0.29
13	0	1	0	0
15*	0	0	0.5	0.5
17*	0	0	0.33	0.66
18	0	0	1	0
20	0	0	0	1

Table 3.1.: An example of the weight matrix W for the dataset (3.1)

covariates. The slice ranges are computed by the R function from the package **dr**, which aims to put approximately the same number of observations in every slice, allowing for the asymptotic results to be valid.

3.3. Data Assumptions

In this section, we shall formalize the setup and the data assumptions. We adopt the notations from Li et al. (1999). Our main parameters are:

- Y^o = the true (unobservable) lifetime,

- C = the censoring time,

- δ = the censoring indicator; $\delta = 1$, if $Y^o \leq C$ and $\delta = 0$, otherwise,

- $T = \min\{Y^o, C\}$, the observed time.

We assume that Y^o follows the model

$$Y^o = f(\beta_1 x, \beta_2 x, \ldots, \beta_k x, \epsilon), \tag{3.2}$$

and that

$$C \perp\!\!\!\perp Y^o \mid x. \tag{3.3}$$

Condition (3.3) ensures identifiability under the random censoring scheme.

The data sample consists of n i.i.d. observations $(T_i, \delta_i, x_i,\ i = 1, \ldots, n.)$ The continuous random variables Y^o and C are not observed.

3.4. Simulation results

Simulation studies were performed to assess the performance of this approach under different conditions. We first present the results under different models and later on in this chapter we compare the estimation with other methods from Nadkarni et al. (2011) and Li and Lu (2008), as well as on the diffuse large B-cell lymphoma data.

3.4.1. Cox's proportional hazards model

The main idea was to generate the survival times following different distributions under the given models and assess the estimation of the β coefficient via our SIR approach. With the covariates generated from the normal distribution, we studied exponential, Weibull and Gompertz survival times in the PH case. While the Weibull and exponential distributions can both be put under the ALT and PH settings, the Gompertz case can only be interpreted in a PH format. We chose the following setup: $p = 5$, $x = (x_1, \dots, x_5) \overset{\text{iid}}{\sim} \mathcal{N}(0, 2)$, the regression coefficient $\beta = (0.5, -0.5, \frac{1}{\sqrt{2}}, 0, 0)$ and generated the survival times as described in Chapter 2. While different sample sizes and censoring patterns were considered, in the tables below we list the averaged results for $n = 50$ and $n = 500$, with censoring percentages of 25% and 50%, all estimated in 1000 runs. To allow for a better comparison, the estimates of $\hat{\beta}$ satisfy $||\hat{\beta}||_2 = 1$, which is the Euclidean norm of the true β. The censoring time C was computed as a random uniform variable from $Uniform(0, c_0)$, where the constant c_0 was selected to control the censoring proportion at the desired level.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Exponential, n=50	0.56 (0.04)	−0.41 (0.06)	0.67 (0.06)	−0.11 (0.05)	0.18 (0.04)
Weibull, n=50	0.43 (0.05)	−0.44 (0.05)	0.77 (0.05)	−0.09 (0.04)	0.13 (0.04)
Gompertz, n=50	0.45 (0.03)	−0.49 (0.04)	0.73 (0.02)	−0.07 (0.04)	0.15 (0.03)
Exponential, n=500	0.54 (0.01)	−0.52 (0.01)	0.66 (0.01)	−0.01 (0.01)	0.04 (0.01)
Weibull, n=500	0.54 (0.01)	−0.51 (0.01)	0.67 (0.01)	0.00 (0.01)	0.03 (0.01)
Gompertz, n=500	0.53 (0.01)	−0.48 (0.01)	0.71 (0.01)	0.00 (0.01)	0.02 (0.01)

Table 3.2.: SIR estimates and standard deviations of the coefficients of a PH model. 25% of the observations are right-censored.

From Table 3.2 we can see that the coefficients are pretty close to the true ones. Even on relatively small samples ($n = 50$), the method performs rather well. Having larger samples brings more accuracy, shrinking the fourth and the fifth coefficients more towards zero. The standard deviations of our estimates, as expected, get smaller with larger samples.

Table 3.3 contains the same results as Table 3.2, except that a larger proportion of the data was censored (50% instead of 25%). There is much more noise for the

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Exponential, n=50	0.56 (0.06)	−0.48 (0.07)	0.61 (0.07)	0.23 (0.08)	0.21 (0.08)
Weibull, n=50	0.47 (0.06)	−0.51 (0.08)	0.71 (0.09)	0.12 (0.07)	0.07 (0.07)
Gompertz, n=50	0.73 (0.03)	−0.43 (0.05)	0.52 (0.04)	0.04 (0.04)	0.12 (0.04)
Exponential, n=500	0.51 (0.01)	−0.47 (0.01)	0.72 (0.01)	−0.01 (0.01)	0.00 (0.01)
Weibull, n=500	0.52 (0.02)	−0.48 (0.02)	0.71 (0.01)	−0.01 (0.02)	0.01 (0.01)
Gompertz, n=500	0.48 (0.01)	−0.49 (0.01)	0.73 (0.01)	0.04 (0.01)	0.01 (0.01)

Table 3.3.: SIR estimates and standard deviations of the coefficients of a PH model. 50% of the observations are right-censored.

smaller sample size which makes the correct estimation of the non-zero variables quite challenging. While the estimates are not so close to the true values, such a procedure can be viewed as a variable selector, to distinguish the most important variables. The larger samples ($n = 500$) do not seem to be influenced much by the severe censoring.

In general, the underlying distribution does not seem to play a major role in successful recovery of the coefficients. One observes a slight underestimation of the second coefficient, which may be due to the bias caused by the equal reweighting. On larger sample this effect is less present.

3.4.2. Accelerated lifetime model

As a next step, we assess how our procedure performs under the ALT model. Here we also used the exponential and the Weibull distributions, replacing the Gompertz distribution with the log-normal one. The regression coefficients β remained the same as above, and the similar setups of sample sizes and censoring patterns were generated.

In general, the results listed in Tables 3.4 and 3.5 present similar properties to the accelerated lifetime model. One could again notice an underestimation of the second parameter in case of the mild censoring (25%) on smaller samples, however it is less pronounced than in the PH setup. The estimation under the severe censoring (50%) is of a slightly worse quality, as compared to the PH case.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Exponential, n=50	0.53 (0.07)	−0.40 (0.10)	0.74 (0.11)	−0.04 (0.08)	−0.09 (0.08)
Weibull, n=50	0.50 (0.11)	−0.54 (0.13)	0.66 (0.15)	−0.04 (0.12)	−0.16 (0.11)
Log-Normal, n=50	0.50 (0.06)	−0.37 (0.06)	0.77 (0.05)	−0.09 (0.08)	0.11 (0.07)
Exponential, n=500	0.54 (0.02)	−0.48 (0.02)	0.70 (0.02)	0.03 (0.03)	0.02 (0.03)
Weibull, n=500	0.50 (0.04)	−0.50 (0.04)	0.71 (0.03)	0.02 (0.03)	0.01 (0.03)
Log-Normal, n=500	0.50 (0.02)	−0.52 (0.02)	0.69 (0.02)	0.02 (0.02)	0.03 (0.02)

Table 3.4.: SIR estimates and standard deviations of the coefficients of an ALT model. 25% of the observations are right-censored.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Exponential, n=50	0.71 (0.07)	−0.46 (0.09)	0.52 (0.09)	0.01 (0.09)	0.10 (0.10)
Weibull, n=50	0.74 (0.11)	−0.20 (0.13)	0.63 (0.15)	0.09 (0.13)	0.11 (0.14)
Log-normal, n=50	0.71 (0.07)	−0.60 (0.08)	0.34 (0.08)	0.11 (0.09)	0.10 (0.09)
Exponential, n=500	0.52 (0.02)	−0.52 (0.02)	0.67 (0.02)	0.08 (0.02)	0.02 (0.02)
Weibull, n=500	0.54 (0.04)	−0.47 (0.04)	0.70 (0.03)	0.01 (0.04)	0.03 (0.04)
Log-normal, n=500	0.55 (0.02)	−0.51 (0.02)	0.67 (0.02)	0.01 (0.03)	−0.02 (0.03)

Table 3.5.: SIR estimates and standard deviations of the coefficients of an ALT model. 50% of the observations are right-censored.

Naturally, both the degree of censoring and the sample size influence the results. The larger the sample size, the better (and more accurate) estimates we get. The same pattern applies to the degree of censoring. But the sufficiently large sample size can compensate for the severely censored data. If we have a lot of data, we can get good results disregarding the fact that a major part of it has been censored.

3.5. Adjusting the weights for the censored observations

Our original idea to treat the censored observations was to redistribute them with equal weights to all the subsequent slices. This is justified because the event could have taken place any time after the censoring. The question is, how well does the equal weight correspond to reality? The individuals are not under the same risk in all the slices. Moreover, by not taking into the account the covariates, we run the risk of a strongly biased estimation. As an example, consider a dataset, where individuals with low, average and high values of $\beta^T x$ gave very different hazards. The equal distribution of the censored observation to the posterior slices creates a bias. Other options for the attribution of censored observations have to be considered.

As an example where the equal weights strategy does not perform well, let us consider a survival model investigated by Yan and Zhang (2012). Suppose $X = (X_1, \ldots, X_{10})^T \overset{\text{iid}}{\sim} Uniform(0,1)$. The true lifetime Y^0 depends on the covariates as follows

$$Y^0 = \exp(5 - 10(1 - \sqrt{2}\beta_0^T X)^2 + \epsilon), \tag{3.4}$$

where $\epsilon \sim \mathcal{N}(0,1)$ and is independent of X, and $\beta_0 = (1,0,0,0,1,0,0,0,0,0)^T/\sqrt{2}$. The right censoring time C is generated as

$$C = c\sqrt{2}\beta_1^T X, \tag{3.5}$$

where $\beta_1 = (0,0,0,1,1,0,0,0,0,0)^T/\sqrt{2}$ and c is a constant used to control the proportion of censoring.

Table 3.6 lists the SIR estimates (after 100 runs) under the equal weighting distribution of censored observations for the defined model with the true survival time (3.4) and the censoring time (3.5). The sample size of $n = 500$ was considered, under the two censoring percentages, 30% and 75%.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
n= 500	0.61 (0.26)	-0.21 (0.10)	-0.01 (0.05)	0.41 (0.19)	0.56 (0.15)
censoring 30%	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
	0.19 (0.09)	0.18 (0.08)	-0.12 (0.08)	0.05 (0.02)	0.02 (0.04)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
n = 500	0.41 (0.35)	-0.03 (0.08)	0.03 (0.05)	0.44 (0.22)	0.78 (0.61)
censoring 75%	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
	-0.01 (0.03)	0.03 (0.04)	-0.02 (0.04)	-0.06 (0.08)	0.10 (0.09)

Table 3.6.: SIR estimates and standard deviations of the coefficients of the model (3.4)-(3.5). 30% and 75% of the observations are right-censored.

We notice that apart from the 1st and the 5th components, corresponding to the true basis, the 4th component also stands out. Despite a relatively large sample size, the censoring pattern interferes with the true survival time, and the results are mixed up. Augmentation of the sample size does not solve this problem.

3.5.1. Cox's proportional hazards model

We shall base the reweighting on the direct computation of the probability of an event in the slice with the help of the survival function. One could write down the probability of a certain individual with covariate x experiencing the event (dying) in a slice i as

$$\begin{aligned} \mathrm{P}[\text{dying in slice } i,\, x] &= F(\text{upper},\, x) - F(\text{lower},\, x) \\ = S(\text{lower}, x) - S(\text{upper}, x) &= S_0(\text{lower})^{e^{\beta^T x}} - S_0(\text{upper})^{e^{\beta^T x}}, \end{aligned} \tag{3.6}$$

where $S_0(\text{upper})$ and $S_0(\text{lower})$ is the baseline survival function, estimated at the upper and lower bound of the slice i, respectively. This is depicted in Figure 3.1.

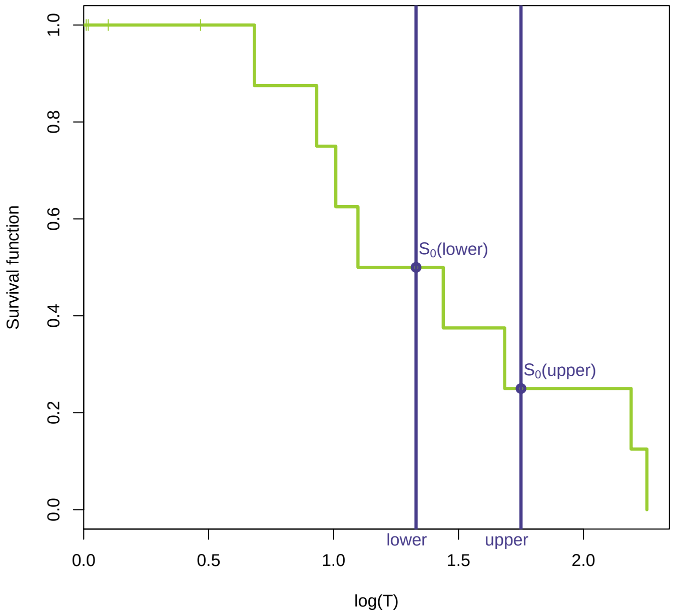

Figure 3.1.: Illustration of (3.6).

The application of this reweighting technique is quite straightforward. We start by computing $\hat{\beta}$, based on the equal distribution of the censored observations. We estimate S_0 with the Kaplan and Meier (1958) method. Then we compute (3.6) with $\hat{\beta}$ for every censored individual and for all posterior slices to get a probability of event in each of them. All the weights for any individual are then normalized to have their sum equal 1. In the end, we recompute the steps 3-5 of the algorithm from Section 2.2.3 to get the final $\hat{\beta}$.

Figure 3.2 shows the distribution of the weights among 11 slices for a certain censored individual, with equal weighting system, and after the adjustments for covariates. The weights values are in bold, while the grey lines mark the slice ranges, where the weights are different from zero. We see how drastically the weight assignment changes, with the individual in question having much higher probability to undergo an event later rather than sooner.

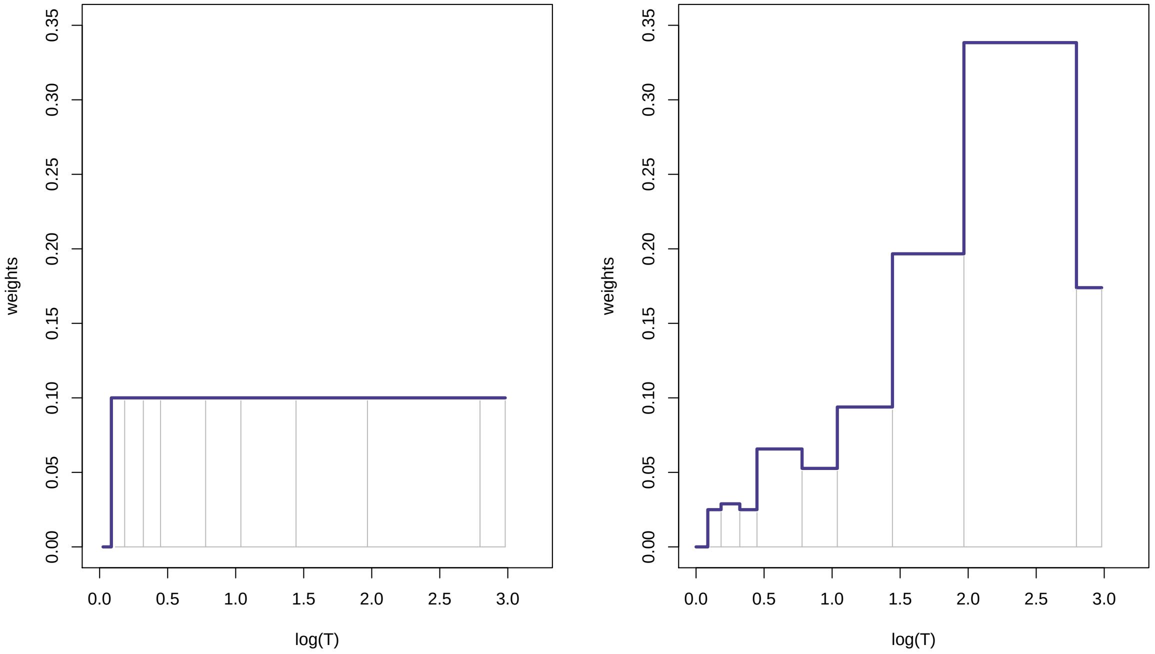

Figure 3.2.: Weights distribution (PH model), before and after the adjustment for the covariates.

Let us explore whether this adjustment actually affects the estimates for β. Below we list for comparison both averaged estimates, with equal weights and covariates effect. For convenience, p was taken to be 5, $x = (x_1, \dots, x_5) \overset{\text{iid}}{\sim} \mathcal{N}(0,2)$, and the true vector of the coefficients is normalized to be $\beta = (\frac{1}{\sqrt{5}}, -\frac{1}{\sqrt{5}}, 0, \frac{\sqrt{2}}{\sqrt{5}}, \frac{1}{\sqrt{5}})$. The censoring distribution was uniform, and the number of slices was fixed to be 10. This setup was run over 100 times on different distributions and sample sizes.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Weibull, n= 50	0.44 (0.07)	−0.51 (0.12)	0.06 (0.08)	0.60 (0.12)	0.32 (0.11)
	0.39 (0.12)	−0.45 (0.14)	0.04 (0.10)	0.54 (0.14)	0.24 (0.13)
Weibull, n = 500	0.44 (0.02)	−0.44 (0.02)	0.00 (0.01)	0.63 (0.02)	0.44 (0.02)
	0.44 (0.02)	−0.43 (0.02)	−0.01 (0.02)	0.64 (0.02)	0.44 (0.02)
Exponential, n = 50	0.51 (0.05)	−0.43 (0.05)	0.18 (0.05)	0.56 (0.05)	0.38 (0.05)
	0.45 (0.09)	−0.37 (0.06)	0.17 (0.05)	0.49 (0.07)	0.33 (0.08)
Exponential, n = 500	0.45 (0.02)	−0.44 (0.02)	0.00 (0.01)	0.64 (0.02)	0.44 (0.02)
	0.44 (0.04)	−0.44 (0.04)	0.00 (0.02)	0.64 (0.05)	0.44 (0.05)
Gompertz, n = 50	0.35 (0.06)	−0.47 (0.10)	−0.10 (0.07)	0.66 (0.10)	0.36 (0.08)
	0.33 (0.09)	−0.40 (0.12)	−0.10 (0.08)	0.57 (0.13)	0.30 (0.10)
Gompertz, n = 500	0.48 (0.02)	−0.47 (0.02)	−0.04 (0.02)	0.59 (0.02)	0.45 (0.02)
	0.48 (0.02)	−0.47 (0.02)	−0.04 (0.02)	0.59 (0.02)	0.45 (0.02)

Table 3.7.: SIR estimates and standard deviations of the coefficients of an PH model. 25% of the observations are right-censored. For each case, the estimates under two strategies are presented, equal weighting and after the adjustment for the covariates.

We run the SIR method for Cox's model under the same distributions as above: Weibull, exponential and Gompertz. The results for the 25% censoring are listed in Table 3.7. As earlier, two sample sizes, $n = 50$ and $n = 500$, were used. The true value of β is $(0.45, -0.45, 0, 0.63, 0.45)$. For each distribution there are two lines, the first one corresponding to the equal weighting strategy and the second one to the adjusted technique.

The estimates obtained after reweighting present a slightly higher variation which is due to the increased computation. These results were acquired on the $H = 10$ slices, considering a higher number of slices can reduce the variation.

Under a 50% censoring, as seen in Table 3.8, the situation is rather similar. For the smaller samples one could state that after reweighting the estimated components

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Weibull, n= 50	0.39 (0.09)	−0.40 (0.10)	0.04 (0.07)	0.66 (0.08)	0.28 (0.08)
	0.35 (0.11)	−0.33 (0.12)	0.01 (0.08)	0.55 (0.15)	0.28 (0.14)
Weibull, n = 500	0.46 (0.03)	−0.47 (0.02)	0.00 (0.02)	0.63 (0.02)	0.38 (0.03)
	0.46 (0.02)	−0.47 (0.03)	0.00 (0.02)	0.64 (0.02)	0.38 (0.02)
Exponential, n = 50	0.47 (0.06)	−0.41 (0.06)	0.06 (0.07)	0.60 (0.05)	0.35 (0.06)
	0.43 (0.08)	−0.38 (0.09)	0.05 (0.06)	0.57 (0.09)	0.33 (0.07)
Exponential, n = 500	0.47 (0.03)	−0.46 (0.03)	0.01 (0.03)	0.62 (0.02)	0.41 (0.02)
	0.47 (0.03)	−0.45 (0.03)	0.01 (0.03)	0.62 (0.03)	0.42 (0.03)
Gompertz, n = 50	0.33 (0.08)	−0.26 (0.11)	0.06 (0.07)	0.41 (0.17)	0.23 (0.11)
	0.28 (0.13)	−0.26 (0.16)	0.02 (0.05)	0.35 (0.20)	0.18 (0.14)
Gompertz, n = 500	0.44 (0.02)	−0.38 (0.02)	−0.02 (0.02)	0.66 (0.02)	0.46 (0.02)
	0.44 (0.02)	−0.39 (0.02)	−0.02 (0.02)	0.66 (0.02)	0.46 (0.02)

Table 3.8.: SIR estimates and standard deviations of the coefficients of an PH model. 50% of the observations are right-censored. For each case, the estimates under two strategies are presented, equal weighting and after the adjustment for the covariates.

of $\hat{\beta}$ are not so close to the true values. The nature of their relationship, however, mostly preserves (the equality between certain components, their ratio etc.)

3.5.2. Accelerated lifetime model

We write down the model as a case of the log-linear regression:

$$\log(T_i) = Y_i = \beta^T x_i + w_i, \quad i = 1, \ldots, n. \tag{3.7}$$

where $w_i = \log(T_{0,i})$ is a residual, following an unspecified distribution, $i = 1, \ldots, n$. We are going to compute the new weights for the censored data based on the distribution of w_i's.

We start by running the SIR algorithm with equal weights for get an estimate $\hat{\beta}$. Using all the uncensored observations, we create a histogram of residuals, $\hat{w}_i = Y_i - \hat{\beta}^T x_i$. For each censored observation T_j^*, the histogram is shifted by $T_j^* - \hat{\beta}^T x_j$ (where T_j^* is the censored value) and the density to the right of this point $T_j^* - \hat{\beta}^T x_j$

is reweighted to be equal to 1. All the weights are written in the weight matrix W in order to compute the covariance matrix $\hat{V}$ in the SIR algorithm.

Once computed, the histogram remains the same, only the zero level shifts for every censored observation. The detail to keep in mind during the implementation was the defined range of all the slices. Since the slice means are fixed by the first run of the procedure (performed to estimate $\hat{\beta}$), it is necessary to maintain all the slice limits. Therefore, the histogram had to be created with the breaks being the slice ranges.

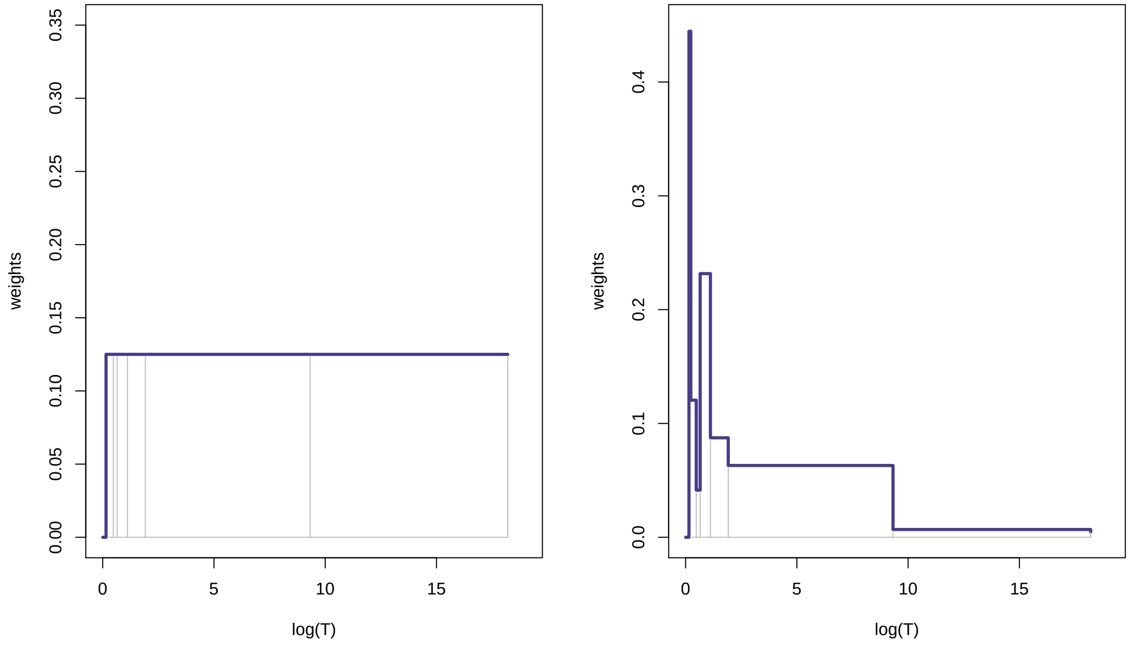

Figure 3.3.: Weights distribution (ALT model), before and after the adjustment for the covariates.

An example of reweighting is presented in Figure 3.3. For a specific observation, we plot the equal weights strategy and the adjustment for the covariates information. The slice intervals are defined by vertical lines. We note the uneven spread of slices among the $\log(T)$, this is due to the fact that the slices contain roughly the same number of observations.

There is another way to account for the information of the censored individuals.

We can view (3.7) as $T = e^{\beta^T x} e^{w} = e^{\beta^T x} T_0$, where T_0 is a baseline survival time. We write down the corresponding survival function as

$$S(t,x) = \mathrm{P}[T \geq t] = \mathrm{P}[e^{\beta^T x} T_0 \geq t] = \mathrm{P}[T_0 \geq te^{-\beta^T x}] = S_0(te^{-\beta^T x}), \tag{3.8}$$

where S_0 is the baseline survival function. Thus, we can compute the probability for a censored individual to experience an event in a posterior slice in the same way as in (3.6). We notice that in the PH model the baseline survival adjusted for covariates by taking it to the power of $e^{\beta^T x}$, while in the ALT model it is the time scale which gets shifted by $e^{-\beta^T x}$.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Weibull, n= 50	0.49 (0.09)	−0.47 (0.11)	0.15 (0.10)	0.65 (0.15)	0.28 (0.13)
	0.53 (0.17)	−0.48 (0.17)	0.09 (0.10)	0.64 (0.22)	0.27 (0.17)
	0.52 (0.15)	−0.48 (0.15)	0.16 (0.10)	0.62 (0.22)	0.29 (0.16)
Weibull, n = 500	0.46 (0.02)	−0.47 (0.02)	0.01 (0.02)	0.63 (0.02)	0.42 (0.02)
	0.47 (0.02)	−0.46 (0.02)	0.01 (0.02)	0.63 (0.02)	0.42 (0.02)
	0.46 (0.02)	−0.46 (0.02)	0.01 (0.02)	0.63 (0.02)	0.42 (0.02)
Exponential, n = 50	0.35 (0.06)	−0.50 (0.07)	−0.09(0.08)	0.66 (0.05)	0.43 (0.05)
	0.39 (0.11)	−0.50 (0.10)	−0.04 (0.08)	0.67 (0.09)	0.37 (0.10)
	0.36 (0.11)	−0.50 (0.10)	−0.07 (0.07)	0.65 (0.11)	0.44 (0.07)
Exponential, n = 500	0.43 (0.02)	−0.44 (0.02)	0.01 (0.02)	0.66 (0.02)	0.42 (0.02)
	0.43 (0.02)	−0.45 (0.02)	0.01 (0.02)	0.66 (0.02)	0.42 (0.02)
	0.43 (0.02)	−0.44 (0.02)	0.01 (0.02)	0.66 (0.02)	0.42 (0.02)
Log-Normal, n = 50	0.48 (0.05)	−0.52 (0.05)	0.03 (0.06)	0.63 (0.05)	0.31 (0.05)
	0.49 (0.07)	−0.52 (0.07)	0.02 (0.05)	0.63 (0.08)	0.31 (0.06)
	0.47 (0.07)	−0.53 (0.07)	0.02 (0.06)	0.63 (0.07)	0.31 (0.07)
Log-Normal, n = 500	0.44 (0.01)	−0.43 (0.02)	0.03 (0.01)	0.65 (0.01)	0.44 (0.01)
	0.44 (0.01)	−0.44 (0.01)	0.03 (0.01)	0.65 (0.01)	0.44 (0.01)
	0.44(0.02)	−0.43 (0.01)	0.03 (0.01)	0.64 (0.01)	0.44 (0.01)

Table 3.9.: SIR estimates and standard deviations of the coefficients of an ALT model. 25% of the observations are right-censored. For each case, the estimates under three strategies are presented, equal weighting, residuals' density and Kaplan-Meier reweighting.

Results in Tables 3.9 and 3.10 list the estimates for $\beta = (0.45, -0.45, 0, 0.63, 0.45)$ and the setup as in Section 3.5.1. However, for the ALT case we present three estimates for each distribution: equal weighting and two reweighting options. The first one is based on the density of the residuals, and the second one on the Kaplan-Meier estimate under the ALT scenario, as described above.

While the results on large samples in Table 3.9 do not differ much, on smaller samples ($n = 50$) one can observe a small improvement of the estimates after reweighting the observations. The residual approach (second line) seems to be more accurate in shrinking down the third component (which is truly zero) than other techniques. It is worth mentioning that the log-normal distribution yields very good estimates even on small samples, regardless of the weighting strategy.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Weibull, n= 50	0.78 (0.12)	−0.37 (0.17)	0.01 (0.08)	0.47 (0.20)	0.17 (0.17)
	0.74 (0.20)	−0.43 (0.21)	0.10 (0.13)	0.43 (0.26)	0.22 (0.20)
	0.81 (0.19)	−0.33 (0.22)	0.06 (0.12)	0.42 (0.24)	0.24 (0.20)
Weibull, n = 500	0.46 (0.03)	−0.41 (0.03)	0.02 (0.03)	0.64 (0.02)	0.47 (0.03)
	0.46 (0.03)	−0.40 (0.03)	0.02 (0.03)	0.64 (0.03)	0.48 (0.03)
	0.46 (0.03)	−0.40 (0.03)	0.02 (0.03)	0.64 (0.02)	0.47 (0.03)
Exponential, n = 50	0.48 (0.07)	−0.42 (0.12)	−0.02 (0.09)	0.70 (0.10)	0.33 (0.10)
	0.50 (0.16)	−0.48 (0.19)	−0.03 (0.12)	0.67 (0.16)	0.24 (0.15)
	0.52 (0.10)	−0.44 (0.17)	−0.04 (0.10)	0.66 (0.16)	0.32 (0.12)
Exponential, n = 500	0.45 (0.02)	−0.44 (0.02)	0.00 (0.02)	0.63 (0.02)	0.45 (0.02)
	0.45 (0.02)	−0.45 (0.02)	0.00 (0.02)	0.63 (0.02)	0.44 (0.02)
	0.45 (0.02)	−0.44 (0.01)	0.00 (0.02)	0.63 (0.01)	0.45 (0.02)
Log-Normal, n = 50	0.74 (0.06)	−0.21 (0.06)	0.11 (0.06)	0.58 (0.05)	0.22 (0.06)
	0.78 (0.12)	−0.23 (0.11)	0.12 (0.13)	0.56 (0.17)	0.13 (0.18)
	0.78 (0.11)	−0.22 (0.11)	0.11 (0.08)	0.56 (0.15)	0.16 (0.16)
Log-Normal, n = 500	0.45 (0.02)	−0.43 (0.02)	0.01 (0.02)	0.59 (0.02)	0.52 (0.02)
	0.44 (0.02)	−0.43 (0.02)	0.01 (0.02)	0.58 (0.02)	0.52 (0.02)
	0.45 (0.02)	−0.43 (0.02)	0.01 (0.02)	0.59 (0.01)	0.51 (0.02)

Table 3.10.: SIR estimates and standard deviations of the coefficients of an ALT model. 50% of the observations are right-censored. For each case, the estimates under three strategies are presented, equal weighting, residuals' density and Kaplan-Meier reweighting.

When comparing the results in the PH and the ALT case, one notices a higher variation for the Weibull distribution. This is an unexpected result and worth being explored. However, it is left for future work.

When half of the data are censored, as seen in Table 3.10, the correct identification of the coefficients is not really feasible. Surprisingly, the exponential distribution shows good identification, while other distributions yield distorted results. The best option in such a case is to consider these methods for variable selection (nomination of important covariates).

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
n= 500 censoring 30%	0.61 (0.26)	−0.21 (0.10)	−0.01 (0.05)	0.41 (0.19)	0.56 (0.15)
	0.58 (0.34)	−0.19 (0.13)	−0.06 (0.08)	0.32 (0.23)	0.63 (0.18)
	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
	0.19 (0.09)	0.18 (0.08)	−0.12 (0.08)	0.05 (0.02)	0.02 (0.04)
	0.18 (0.11)	0.22 (0.14)	−0.17 (0.11)	0.03 (0.04)	0.03 (0.08)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
n= 500 censoring 75%	0.41 (0.35)	−0.03 (0.08)	0.03 (0.05)	0.44 (0.22)	0.78 (0.61)
	0.67 (0.44)	−0.02 (0.09)	0.01 (0.06)	0.14 (0.25)	0.72 (0.66)
	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
	−0.01 (0.03)	0.03 (0.04)	−0.02 (0.04)	−0.06 (0.08)	0.10 (0.09)
	0.02 (0.08)	0.03 (0.08)	0.04 (0.07)	0.02 (0.06)	0.11 (0.13)

Table 3.11.: SIR estimates and standard deviations of the coefficients of the model (3.4)-(3.5). For each coefficient there are two values, the result of equal weighting (top line) and the results of Kaplan-Meier reweighting for ALT model (bottom line). 30% and 75% of the observations are right-censored.

Table 3.11 refers back to the model (with the true survival time (3.4 and the censoring distribution (3.5)) in the beginning of this section, presenting a case when the equal distribution of the censored observations resulted in a bias in estimation. We now take the same settings as before and apply the Kaplan-Meier reweighting (for ALT case) and present the results before and after this reweighting. We see that the effect of the fourth component (which is a part of the censoring distribution) is less present after reweighting the censored observations.

So far, we explored the performance of our SIR method for censored regressions with the uniform censoring pattern. This is quite simplistic, therefore we also wish to consider other approaches, especially when the censoring and the true survival time distributions depend on different covariates. This is reviewed in next section. Further on, the equal slice distribution of the censored observations is not considered.

It seems rather intuitive to re-iterate the process of adjusting the weights for the censored observations until reaching convergence. The stopping criterion was taken to be the maximal Euclidean distance between successive weight vectors divided by the number of censored observations. While one cannot exclude the possibility of oscillation between two solutions, in our simulations it led to improved results. For stability purposes, all the weights were rounded to four significant digits. All results we give from now on were computed with the iterated algorithm.

Another question of interest in the choice of slices in the model. There are few recommendations on the topic. Li (1991) states that even when the slice number is $n/2$ (resulting in two observations per slice), the resulting estimate will still be root n consistent. Later on, Chen and Li (1998) claim that the SIR algorithm is not too sensitive to the slice number H. These claims refer to the estimation of the dimension, but not the directions. Becker and Gather (2007) investigate the influence of the slice number on the number of directions and discuss that this parameter is of importance. It is concluded that too large H is to be avoided, as it presents the tendency to overestimate the e.d.r. space, and that a slice number of $H \approx 0.1n$ seems to be a reasonable choice. In the next section we briefly discuss this strategy.

3.5.3. Different models and comparison with other methods

As mentioned in the beginning of this chapter, Li et al. (1999) discussed two censoring patterns, when both the censoring C and the lifetime Y^o distributions are independent from the covariates, $(C, Y^o) \perp\!\!\!\perp x$, and their conditional independence given the covariates $C \perp\!\!\!\perp Y^o \mid x$. The censoring of the latter type creates the bias and is of interest to us. In the simulation studies we considered so far, the censoring pattern was generated from the uniform distribution. This setup, however, does not imply that the covariates do not affect the observed survival $T = \min(Y^o, C)$. Nevertheless, we shall review models with the censoring distribution depending on the covariates. In order to compare our approach with some of the methods mentioned in the literature review, we adopt the models from the papers in question and show our results.

We start with the paper by Li et al. (1999) and take the following example: *Model 1* assumes $x = (x_1, \dots, x_6) \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The true survival time Y^o and the censoring time C are generated from

$$Y^o = -\frac{\log(\epsilon_1)}{e^{x_1}}; \quad C = -\frac{\log(\epsilon_2)}{e^{x_2}},$$

where ϵ_1, ϵ_2 are independent uniform random variables from [0, 1]. Conditional on x, Y^o and C are seen to follow the exponential distributions with the parameters $\lambda_1 = e^{x_1}$ and $\lambda_2 = e^{x_2}$, respectively. We generate 300 independent observations of (T, δ), and have a censoring rate of 45%. To study the effect of the regressor dimension, we vary p from $6, 10, 15$ and 20. The setup has only one true e.d.r. lifetime direction, $\beta = (1, 0, \ldots, 0)^T$, and we compute an R-squared term for evaluating how close to the true e.d.r. direction the estimated direction is. In this case the R-squared term is the squared correlation coefficient between $\hat{\beta}^T X$ and $\beta^T X$. Table 3.12 lists the mean and the standard deviation for R^2 over 100 runs as the number of regressors increases. Different techniques of reweighting for the censored observations are listed, the density approach (for residuals) under the ALT model and the Kaplan-Meier adjustment under the ALT and the PH model. For comparison purposes, we present in the last column the results from the double slicing procedure DSIR, taken from the paper by Li et al. (1999).

Mean (standard deviation) for R^2				
p	Density approach	KM-ALT	KM-PH	DSIR
6	0.873 (0.062)	0.926 (0.039)	0.939 (0.034)	0.917 (0.059)
10	0.794 (0.068)	0.891 (0.041)	0.914 (0.042)	0.863 (0.063)
15	0.756 (0.071)	0.840 (0.555)	0.828 (0.071)	0.796 (0.089)
20	0.712 (0.080)	0.783 (0.062)	0.782 (0.073)	0.758 (0.082)

Table 3.12.: Performance of the proposed censored SIR under Model 1 with 100 runs. Reweighting of the censored observations is based on the densities of residuals and the Kaplan-Meier adjustment under the ALT and PH model. The last column contains the results from DSIR, the double slicing procedure by Li et al. (1999).

Since the true lifetime is exponential (which can be viewed both as PH or ALT setup), we test all of the three reweighting techniques. What we conclude from this example is that two out of the three methods we used for reweighting outperform the double slicing method. These are the Kaplan-Meier based methods. While the results of the density approach are satisfactory, in this model such an approach underperforms.

Next, we relate to the inverse regression by Nadkarni et al. (2011), where the authors estimate the conditional probability of survival with the help of a kernel conditional Kaplan-Meier estimator. In their study, both the accelerated lifetime and Cox's model are considered, and the basis for the central subspace is estimated by minimizing a quadratic discrepancy function. We test our procedure on two models from their article. In their paper, Nadkarni et al. (2011) compare the performace of their inverse regression method to the double slicing method by computing the mean angle between the basis vector and the eigenvector estimate. We did the same for our method and present below, in Figures 3.4 and 3.5, the results for comparison. The two following models were considered:

Model 2 has the censoring percentage of 45% under the following setup: p=6, $x = (x_1, \dots, x_6) \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. The true survival time Y^o is generated from

$$Y^o = \exp(x_1 + x_3)\epsilon_1,$$

where ϵ_1 follows the exponential distribution with parameter 1. The censoring time C is generated from

$$C = \exp(x_1 + x_2 + x_3)^4.$$

Model 3 has the same setup as Model 2, but the true survival time Y^o is generated from

$$Y^o = (-\log(\epsilon_2)/\exp(x_1 + x_3)),$$

where ϵ_2 follows the uniform distribution on [0,1]. The censoring time C is generated from

$$C = \exp(x_1 + x_2 + |x_3|)^2.$$

Figures 3.4 and 3.5 show the mean angles between the true basis and the estimates from three procedures from 100 simulation runs for Model 2 and Model 3, respectively. The true basis equals $\beta = (1,0,1,0,0,0)^T$ for $p = 6$. The considered procedures are listed as SIR (our method, we used the Kaplan-Meier reweighting), IR (the method by Nadkarni et al. (2011)) and DSIR (the double slicing). Each figure has two plots: part (a) presents the mean angles for a fixed number of pre-

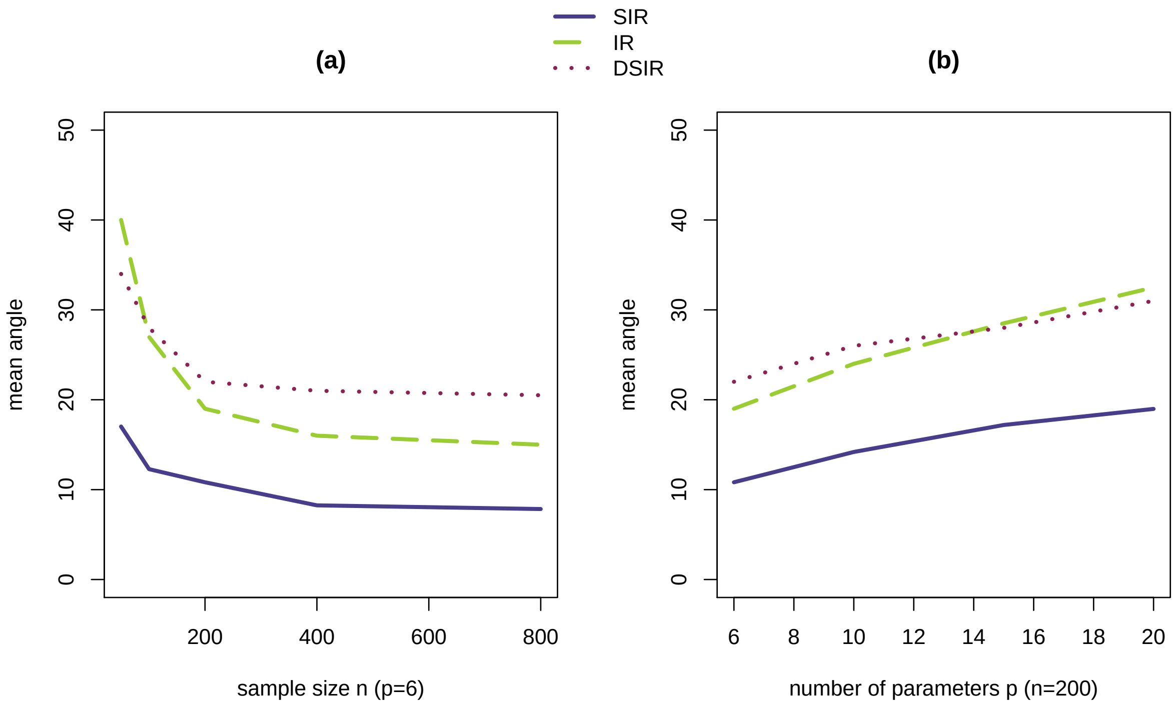

Figure 3.4.: Mean angles between the true basis and SIR (our), IR (alternative) and DSIR (double slicing) estimates in Model 2.

dictors ($p = 6$) as the sample size n grows, while part (b) is a function of p on the sample size of $n = 200$. We can see that our method in both cases largely outperforms the other two, while being computationally less challenging (both inverse regression and double slicing require solving an optimization problem and kernel estimation).

Table 3.13 lists the mean angles for Model 3 as a function of the sample size and the number of slices. We notice that on larger samples the choice of H idoes not have a big effect (aside from the very large values), while on small samples having a large number of slices (few observations per slice) leads to worse estimates of the basis. These results are in accordance with the conclusions of Becker and Gather (2007), where the authors suggest using $H \approx n/10$.

We also compared our method with a method described in Lu and Li (2011), which is also based on the inverse censoring probability weighted estimation, similar to the idea of Nadkarni et al. (2011). This method is more of a general case for dimension reduction (sliced inverse regression is just a special case), and the uncensored

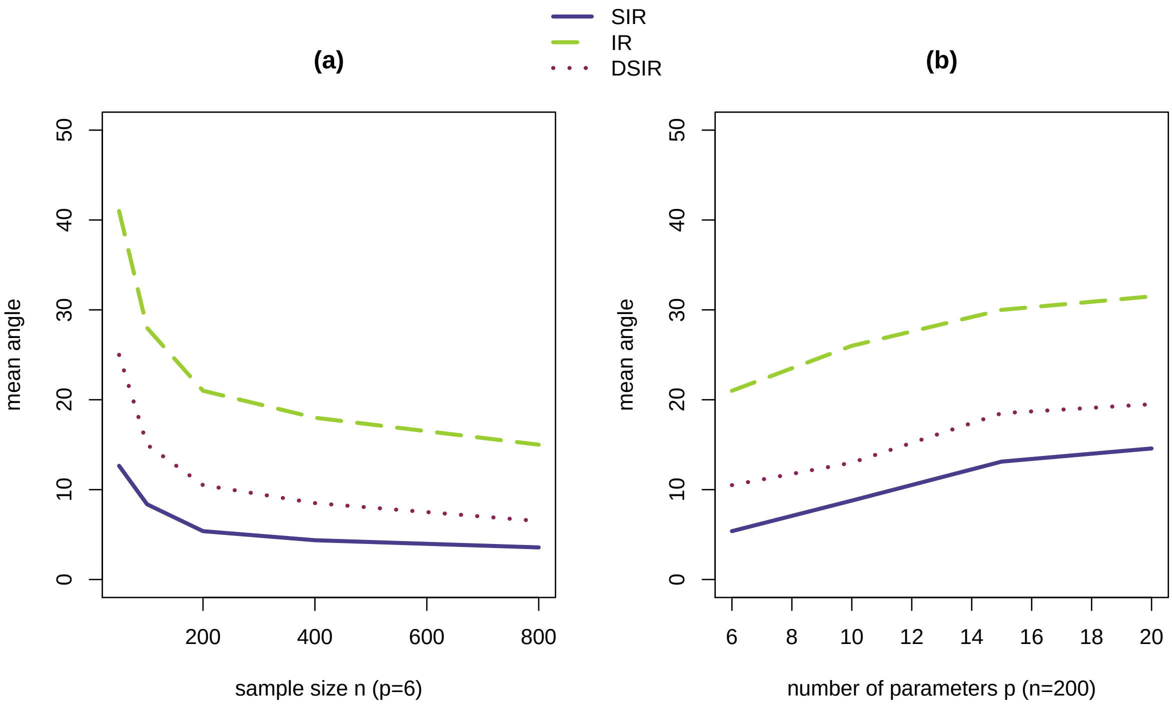

Figure 3.5.: Mean angles between the true basis and SIR (our), IR (alternative) and DSIR (double slicing) estimates in Model 3.

n=200		n=500		n=1000		n=3000	
H	mean angle	H	mean angle	H	mean angle	H	mean angle
10	4.5	10	3.11	10	5.15	10	2.96
20	5.47	20	5.06	20	4.19	20	3.10
50	13.29	50	6.77	50	3.97	50	4.15
100	14.45	100	3.10	100	4.03	100	4.09
		200	7.30	200	4.73	200	3.94
				500	6.02	500	8.54

Table 3.13.: Mean angle between the true and the estimated basis in Model 3 as a function of the sample size n and the number of slices H.

observations are weighted by the inverse of the survival function of the censored time. We adopt one of their models and test it on our procedure.

Model 4 has the the censoring percentage of 40% and takes p=10. Covariates $x = (x_1, \ldots, x_{10}) \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. The true survival time is generated from

$$Y^o = \exp(-2.5 + \sin(0.1\pi\beta^T x) + 0.1(\beta^T x + 2)^2 + 0.25\epsilon_3, \qquad (3.9)$$

where ϵ_3 follows the extreme value distribution (to have a PH setup), and $\beta = (1,1,1,0,0,0,0,0,0,0)^T$. The censoring time is generated as

$$C = \exp(c + \beta_c^T X + \epsilon), \qquad (3.10)$$

with $\beta_c = (-1,0,0,1,0,0,0,0,0,0)^T$, and c being a constant that controls the censoring percentage.

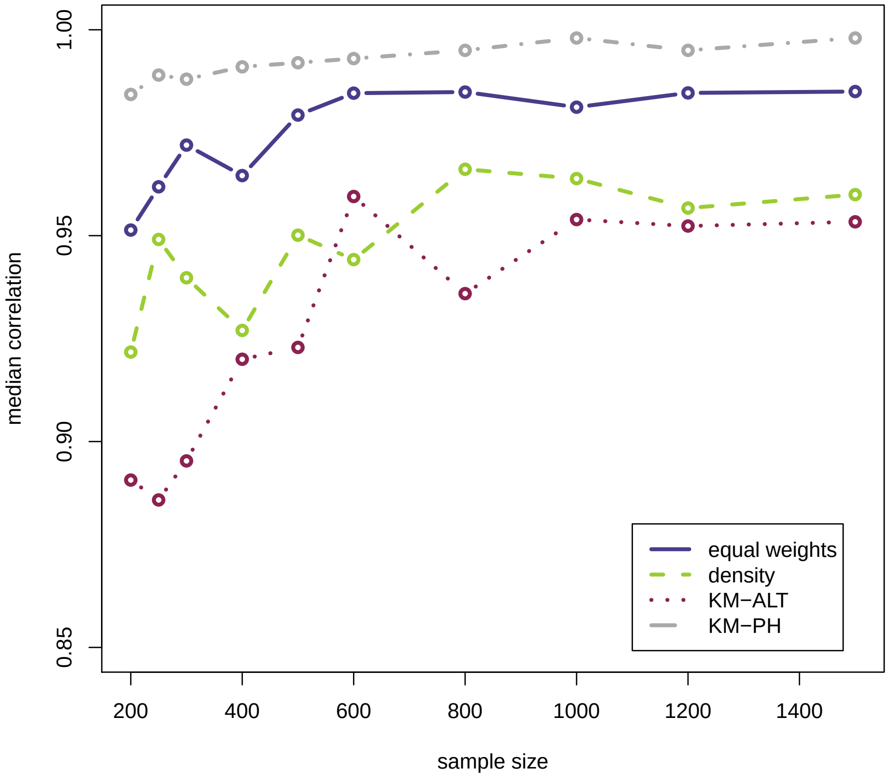

Figure 3.6.: Correlation coefficient between the true and estimated basis for Model 4. Presented results include all three reweighting techniques, as well as equal distribution of the censored observations.

For performance evaluations, the authors rely on the vector correlation coefficient between the true and estimated basis. The results can be seen in Figure 3.6, where the median correlation coefficient is plotted versus the sample size. In their paper, Lu and Li (2011) were comparing different methods of estimating the inverse of the survival function of the censored time, and their resulting curves vary between the 0.93 and the 0.98 range for the correlation coefficient, which is comparable with our results. We highlight the fact that our Kaplan-Meier reweighting under the PH model performs better than all the methods in Lu and Li (2011). They also list the results for the double slicing for this example, which fluctuate between 0.80 and 0.85 for the correlation coefficient.

We also performed a sensitivity analysis, inspired by Lu and Li (2011). They suggested to evaluate the effect of misspecification of the censoring time distribution. For this purpose, different variations of the censoring distribution (3.10) are studied. The hazard function of ϵ_c in (3.10) is given by

$$\lambda_c(t,r) = \frac{\exp(t)}{1 + r\exp(t)}, \tag{3.11}$$

where the constant r controls the level of deviation from a PH model. When $r = 0$, it is a PH model, but as r increases, it deviates from the PH specification, and when $r = 1$, it corresponds to a proportional odds (PO) or ordered logistic regression model. For the sensitivity study, we take $r = 0.25, 0.5, 0.75$, and 1. In addition, a log-normal censoring distribution is considered by generating ϵ_c from a standard normal distribution (denoted by log-normal in Table 3.14), and a misspecified PH model $(r = 0)$ with an interaction term of covariates, that is, $C = \exp(c - X_1 + X_4 + 0.5X_1X_2 + \epsilon_c)$ (denoted by misspec. PH). The true survival time is still generated by (3.9). Aside from the proportional hazards model for the lifetime, the proportional odds model was taken into account, with ϵ following a logistic distribution in (3.9). The same deviations for the censoring time were considered.

Table 3.14 presents the median correlation based on 100 simulation runs, with 40% censoring. Since it is a study of deviation from a PH model, we used the Kaplan-Meier reweighting for PH models. For comparison, the results from Lu and Li (2011) are also reported, under ICPW (inverse censoring probability weighted) estimation. For both methods, the performance degrades a bit as the model deviates

		n=400		n=800		n=1200	
ϵ	ϵ_c	SIR	ICPW	SIR	ICPW	SIR	ICPW
PH	r = 0.25	0.976	0.969	0.985	0.976	0.997	0.980
	r = 0.5	0.972	0.968	0.984	0.969	0.998	0.971
	r = 0.75	0.971	0.964	0.986	0.969	0.997	0.965
	r = 1	0.971	0.964	0.983	0.957	0.996	0.958
	log-normal	0.980	0.962	0.991	0.967	0.995	0.964
	misspec. PH	0.979	0.969	0.984	0.976	0.994	0.981
PO	r = 0.25	0.973	0.968	0.989	0.977	0.994	0.981
	r = 0.5	0.969	0.967	0.989	0.972	0.994	0.972
	r = 0.75	0.971	0.960	0.985	0.967	0.992	0.964
	r = 1	0.970	0.959	0.986	0.958	0.991	0.958
	log-normal	0.983	0.958	0.991	0.966	0.993	0.967
	misspec. PH	0.965	0.967	0.978	0.975	0.987	0.980

Table 3.14.: Sensitivity analysis: median vector correlation for various censoring time distributions different than the proportional hazards model. For each scenario, the model for the lifetime distribution (PH or PO) and the censoring distribution (3.11) is specified.

increasingly from a PH setting, but our method yields slightly higher correlation coefficients indicating the more accurate estimation of the basis. It is also less sensitive to a deviation towards a log-normal or a misspecified model. One cannot report a big difference in performance between the PH and PO models for the true lifetime, for both listed methods. While the results for the double slicing are not listed in this table, its estimation is far less precise.

In this section, we evaluated the performance of our method on different models and applied it to several examples from different papers. Since each of the papers in question used different ways to judge the quality of the estimation, we tried to be consistent with their methods and did the same. Overall comparison results are quite satisfactory. In fact, our procedure is competitive with the others, and even yields a better estimation in most of the cases. While we designed our approach keeping in mind the nature of the model in question (proportional hazards or accelerated lifetime), it seems that often the reweighing techniques are not overly sensitive to minor deviations from the model.

3.5.4. DLBCL data

An an example of a an application to real data, we chose the diffuse large B-cell lymphoma (DLBCL) data, which was first analyzed by Rosenwald et al. (2002). This dataset consists of 240 patients with DLBCL and there are 138 patient deaths during the followup. As covariate information, we chose the aggregated information (the microarray data have 7399 gene expression levels for each patient), from Appendix to the original paper. Out of 138 recorded events, 5 cases had a survival time of zero, and therefore have been excluded. The rest of the pre-selection was done according to Nadkarni et al. (2011), the IPI subgroup has been removed because of multiple missing entries, and the categorical variable of gene expression subgroup was replaced by dummy variables, ABC and GCB groups. The other variables included gene expression signatures (germinal center B-cell signature, major-histocompatibility-complex (MHC) class II signature, lymph node signature, and the proliferation signature), value for the BMP6 gene, and the outcome predictor score. In total, there are eight covariates.

Basis estimate (our approach)	Covariate	Basis estimate (Nadkarni et al.)
0.284 (0.172)	ABC	0.020 (0.537)
0.055 (0.563)	GCB	0.029 (0.640)
−0.205 (0.698)	B-cell sig.	−0.251 (0.161)
−0.245 (0.452)	Lymph sig.	−0.212 (0.152)
0.096 (0.305)	Prolif. sig	0.201 (0.267)
0.207 (0.168)	BMP6	0.267 (0.216)
−0.154 (0.242)	MHC sig.	−0.266 (0.187)
0.317 (0.211)	Out. pred. score	−0.842 (0.248)

Table 3.15.: Estimates and their bootstrap standard errors of the basis for the DLBCL data. In the left column, our results are listed, in the right column, the results from Nadkarni et al. (2011).

Since Nadkarni et al. (2011) did not provide the details on introducing the dummy variables for the ABC and the GCB groups, the comparison of results does not make much sense. Our basis estimates are listed in Table 3.15, together with the results of Nadkarni et al. (2011). The most important variable selected is the outcome predictor score, however, in our results it is not significant.

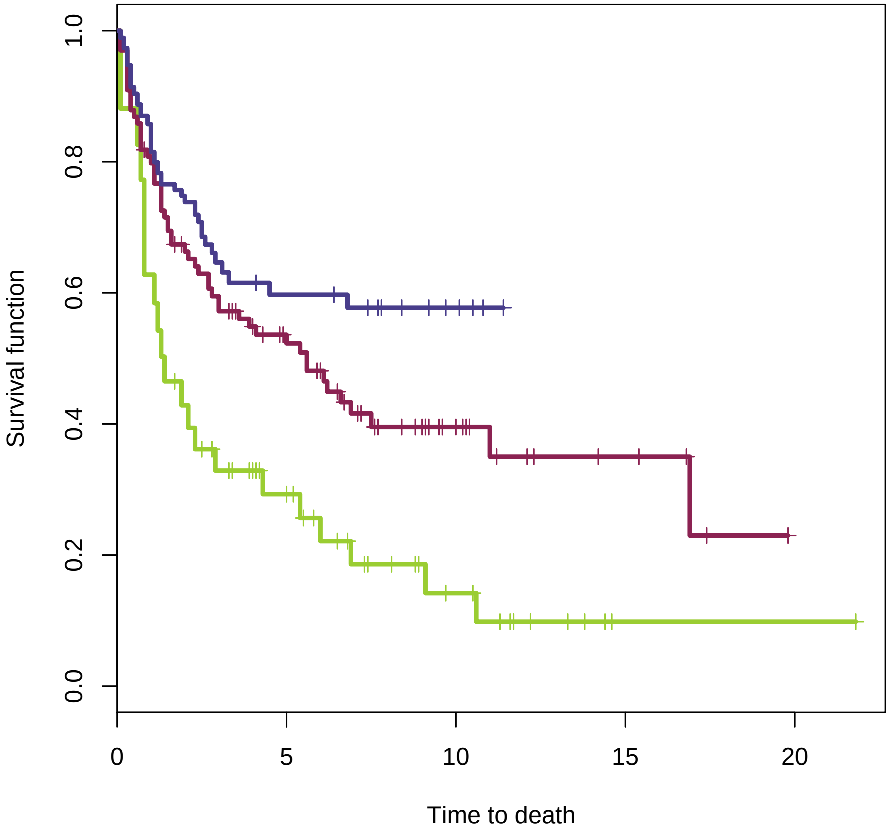

Figure 3.7.: Kaplan-Meier estimates of survival curves for three risk groups of patients defined by the fitted PH model based on SIR basis component for the DLBCL data.

Inspired by a paper by Li (2006), where the double slicing in SIR was applied to the DLBCL data, we fitted a Cox model based on the extracted component from Table 3.15. Based on the fit, we separated the individuals into three groups, corresponding to the low, intermediate, and high risk respectively. The separation was determined by the 33% and 66% quantiles of the estimated score. The resulting Kaplan-Meier survival curves can be seen in Figure 3.7, where the three groups are easily recognized. While we did not consider a further examination of this dataset, this example is a nice illustration of the SIR application on real cases.

So far, we used the correlation coefficient, the mean angle between the basis vectors and the R-squared coefficient to judge the quality of the estimation. The aspects of the variance estimation are discussed in the next section.

3.6. Variance estimation

Now, let us address the variance estimation for the SIR in survival models. We start with the likelihood-based approach (for the ALT model) and further on review the difficulties of the asymptotic estimation.

3.6.1. Maximum likelihood approach

To study the likelihood application on censored regressions, we consider an ALT Weibull model of the following kind:

$$\log(T_i) = Y_i = \alpha + \beta^T x_i + \sigma w_i, \quad i = 1, \dots, n. \tag{3.12}$$

The survival time T follows the Weibull distribution, and the random variable w follows the Gumbel distribution with the density (2.1.4). The model (3.12) is of a same type as the model (3.7), where the coefficients α and σ were absorbed as the shape and the scale parameters of w's Gumbel distribution. Here, for convenience w has a standard Gumbel distribution.

The likelihood of Y can be written in terms of $w = (Y - \beta^T x)/\sigma$ (Kalbfleisch and Prentice, 1980). The density function for Y is of the form

$$f(y) = \frac{1}{\sigma} f(w) = \frac{1}{\sigma} e^w e^{-e^w},$$

where $f(\cdot)$ is the standard Gumbel density.

We start with the likelihood for Y:

$$L(\alpha, \beta, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sigma} f(w_i) \right)^{\delta_i} \bar{F}(w_i)^{1-\delta_i}, \tag{3.13}$$

where $\delta_i = 1$, if Y_i is not censored, and $\delta_i = 0$, otherwise. $\bar{F}(w_i)$ denotes the survival function of the observation Y_i^*.

The log-likelihood takes the following form:

$$\log L(\alpha, \beta, \sigma) = \sum_{i=1}^{n} \delta_i[-\log(\sigma) + \log(f(w_i))] + (1-\delta_i)\log(\bar{F}(w_i)), \tag{3.14}$$

and its partial derivatives are:

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n}\left(\delta_i \frac{d\log f(w_i)}{dw_i}\frac{dw_i}{d\alpha} + (1-\delta_i)\frac{d\log \bar{F}(w_i)}{dw_i}\frac{dw_i}{d\alpha}\right)$$

$$= \frac{1}{\sigma}\sum_{i=1}^{n}[-\delta_i \frac{d\log f(w_i)}{dw_i} + (1-\delta_i)\lambda(w_i)]$$

$$= \frac{1}{\sigma}\sum_{i=1}^{n}[-\delta_i(1-e^{w_i}) + (1-\delta_i)e^{w_i}] = \frac{1}{\sigma}\sum_{i=1}^{n} x_{ji}(e^{w_i} - \delta_i). \tag{3.15}$$

In (3.15) we used the fact that the hazard function for the Gumbel distribution is $\lambda(w_i) = e^{w_i}$.

We can get the partial derivative with respect to α simply by replacing x_{ji} in (3.15) by 1.

$$\frac{\partial \log L}{\partial \alpha} = \frac{1}{\sigma}\sum_{i=1}^{n}(e^{w_i} - \delta_i). \tag{3.16}$$

And the last partial derivate equals

$$\frac{\partial \log L}{\partial \sigma} = \sum_{i=1}^{n}\left[\delta_i\left(-\frac{1}{\sigma} + \frac{d\log f(w_i)}{dw_i}\frac{dw_i}{d\sigma}\right) + (1-\delta_i)\frac{d\log \bar{F}(w_i)}{dw_i}\frac{dw_i}{d\sigma}\right]$$

$$= \frac{1}{\sigma}\sum_{i=1}^{n}(w_i(e^{w_i} - \delta_i) - \delta_i). \tag{3.17}$$

In order to compute Fisher's information, we also need the second-degree partial derivatives. We have the following:

$$-\frac{\partial^2 \log L}{\partial \beta_k \partial \beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ji}x_{ki}e^{w_i}.$$

$$-\frac{\partial^2 \log L}{\partial\alpha\partial\beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^n x_{ji}e^{w_i}.$$

$$-\frac{\partial^2 \log L}{\partial\alpha^2} = \frac{1}{\sigma^2}\sum_{i=1}^n e^{w_i}.$$

$$\begin{aligned}-\frac{\partial^2 \log L}{\partial\beta_j\partial\sigma} &= \frac{1}{\sigma^2}\sum_{i=1}^n x_{ji}w_ie^{w_i} + \frac{1}{\sigma^2}\sum_{i=1}^n (e^{w_i}-\delta_i)x_{ji} \qquad (3.18)\\ &= \frac{1}{\sigma^2}\sum_{i=1}^n x_{ji}(w_ie^{w_i}+e^{w_i}-\delta_i).\end{aligned}$$

$$-\frac{\partial^2 \log L}{\partial\alpha\partial\sigma} = \frac{1}{\sigma^2}\sum_{i=1}^n (w_ie^{w_i}+e^{w_i}-\delta_i).$$

$$-\frac{\partial^2 \log L}{\partial\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (w_i^2e^{w_i}+\delta_i) + \frac{2}{\sigma^2}\sum_{i=1}^n (w_i(e^{w_i}-\delta_i)-\delta_i).$$

By taking the expected value of this matrix of second-degree partial derivatives, we get Fisher's information.

We wish to compare the maximum likelihood (ML) estimates and the SIR estimates. In order to compute the ML estimates, we need to solve the system of equation (3.15), (3.16) and (3.17). We use the Fisher scoring algorithm to do that. For an initial value of $\theta_0 = (\alpha_0, \beta_0, \sigma_0)$, the score statistic at $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$ can be written as

$$\nabla L(\hat{\theta}) \approx \nabla L(\theta_0) - I(\theta_0)(\hat{\theta} - \theta_0),$$

where $I(\theta_0)$ is Fisher's information matrix, evaluated at θ_0. Given that $\nabla L(\hat{\theta}) = 0$, we get

$$\hat{\theta} = \theta_0 + I(\theta_0)^{-1}\nabla L(\theta_0). \qquad (3.19)$$

Equation (3.19) gives us an iteration procedure for the ML estimates. We start with θ_0 and at each step update the current $\hat{\theta}$ by the term $I(\hat{\theta})^{-1}\nabla L(\hat{\theta})$. Once the algorithm has converged, $\nabla L(\hat{\theta}) = 0$, and the iterations stop. The choice of θ_0 is important for convergence.

We did a comparison between the SIR and ML estimates for $p = 2$ case, that is, a model of the following kind:

$$\log(T_i) = Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \sigma w_i, \quad i = 1, \ldots, n.$$

Besides the estimates themselves, we are also interested in their respective variances.

All the results presented below were achieved with 100-repetitions with $n = 1000$. For the SIR method, we used a reweighting technique (density-based) on 20 slices. Independent censoring situations of 16%, 48% and 64% were considered, as well as the value of $\sigma = 0.5, 2, 4$. The true β was normalized (in order to allow for a comparison between the two methods) and was equal to $\beta = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$. The starting value for the ML procedure was β. The main results are aggregated in Figures 3.8-3.10.

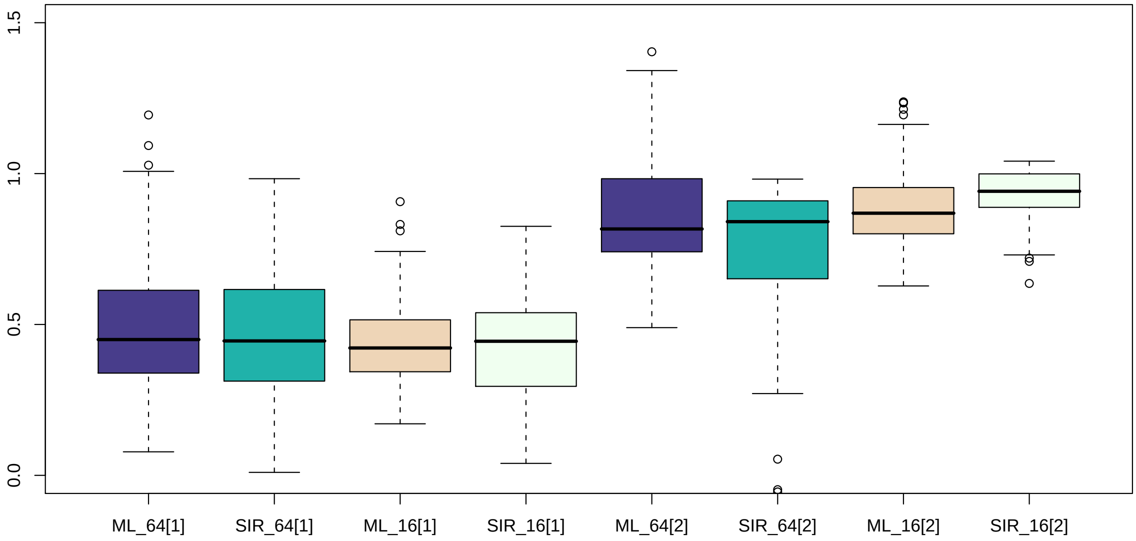

Figure 3.8.: ML and SIR estimates with $\sigma = 4$. Censoring is 16% and 64%.

Figure 3.8 depicts the boxplots for the 2 components of the ML and SIR estimates when $\sigma = 4$. The censoring percentage is either 16% or 64% and is indicated in the label of each boxplot. The estimation (the recovery of β) is quite good. In general, SIR tends to have a bit larger variance which is not surprising, since the ML estimation is fully efficient. Another expected result is that the variation decreases with lower censoring.

Figure 3.9 shows the variation under the fixed censoring of 48%, but for the models with different σ's, namely $\sigma = 0.5$ or $\sigma = 4$ (as in previous figure, the respective

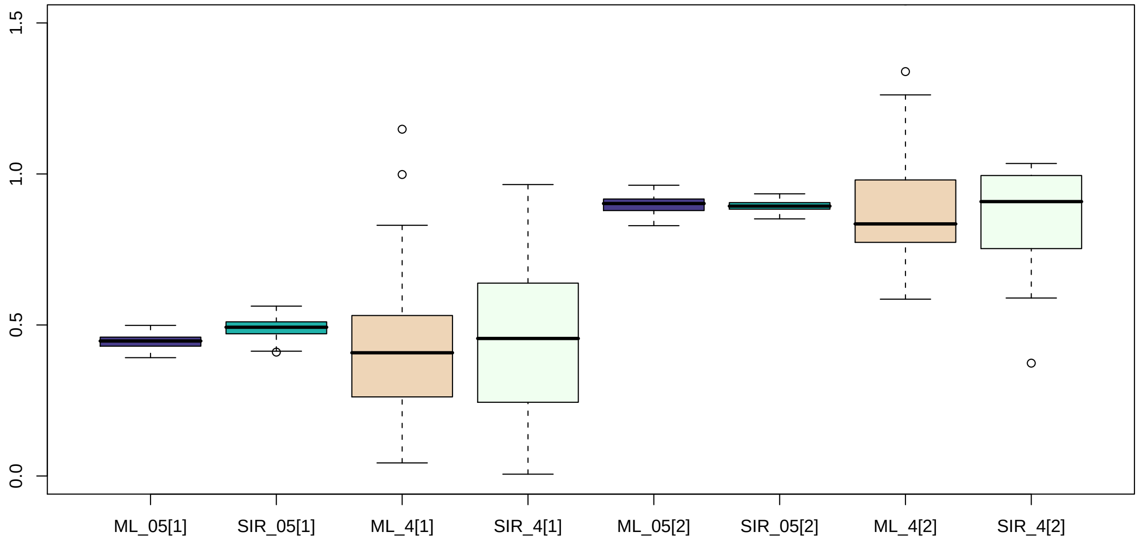

Figure 3.9.: ML and SIR estimates under the 48% of censoring, $\sigma = 0.5$ or 4.

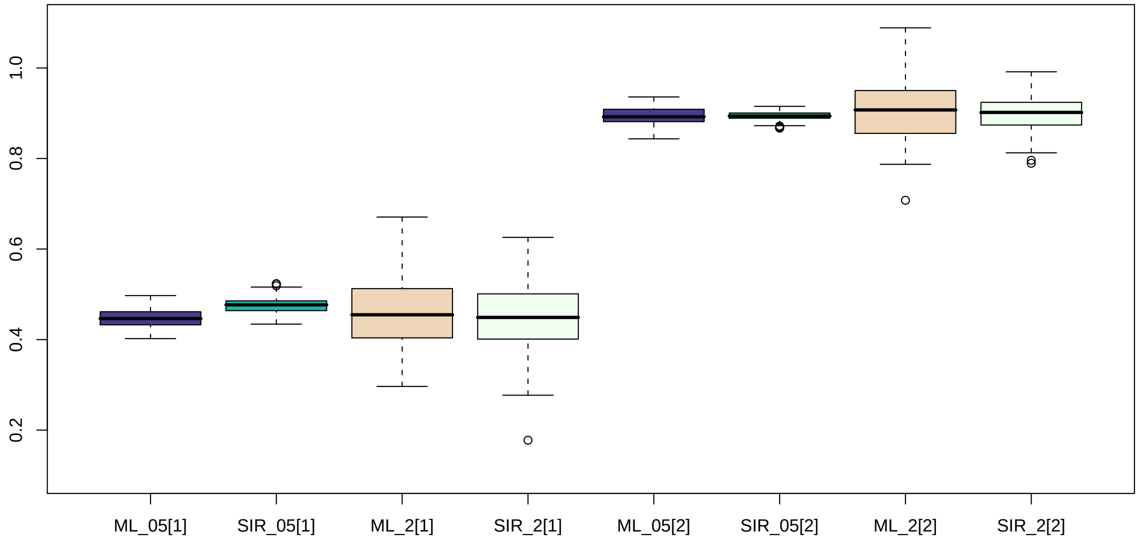

Figure 3.10.: ML and SIR estimates under the 16% of censoring, $\sigma = 0.5$ or 2.

values of σ are indicated in the label of each boxplot). What we observe here, is the effect of σ in (3.12). The variation depends on it and increases with the value of σ.

An interesting remark could be made on the basis of Figure 3.10. It seems that under small fractions of censoring, SIR almost outperforms the ML approach, or at least, performs equally well.

Let us study the variance of the ML method for the Weibull model (3.12) with no censoring. We shall investigate the structure of the inverse Fisher information matrix. In order to do that, we need to compute the expected value of the second-order partial derivatives of the log-likelihood, presented in equations (3.18).

The censoring indicator δ_i in our case is 1, since we assume the no-censoring scenario. Most terms in equations (3.18) are constants, we only need to compute a few expected values.

$$\mathrm{E}(e^w) = 1, \tag{3.20}$$

since w follows the Gumbel distribution with the cumulative distribution function $F(w) = 1 - e^{-e^w}$. By replacing e^w by y, we clearly recognize the exponential distribution $\mathcal{E}(1)$, hence $\mathrm{E}(Y) = 1$.

The next integral,

$$\mathrm{E}(w) = \int_{\infty}^{\infty} w e^w e^{-e^w} dw = \int_0^{\infty} e^{-y} \log y \, dy = -\gamma, \tag{3.21}$$

turns out to be the negative of Euler's constant, $\gamma \approx 0.5772$. Here we substituted $y = e^w$.

The two remaining integrals can be computed by parts, with the help of (3.21):

$$\mathrm{E}(we^w) = \int_{\infty}^{\infty} w e^{2w} e^{-e^w} dw = \int_0^{\infty} y e^{-y} \log y \, dy = 1 - \gamma;$$

$$\mathrm{E}(w^2 e^w) = \int_{\infty}^{\infty} w^2 e^{2w} e^{-e^w} dw = \int_0^{\infty} y e^{-y} \log^2 y \, dy = \frac{\pi^2}{6} + \gamma^2 - 2\gamma. \tag{3.22}$$

In last integral (3.22), we used the fact that $\mathrm{Var}(w) = \pi^2/6$.

The information matrix is:

$$-\mathrm{E}\left(\frac{\partial^2 \log L}{\partial\beta_k \partial\beta_j}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ji}x_{ki}.$$

$$-\mathrm{E}\left(\frac{\partial^2 \log L}{\partial\alpha \partial\beta_j}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ji}.$$

$$-\mathrm{E}\left(\frac{\partial^2 \log L}{\partial\alpha^2}\right) = \frac{n}{\sigma^2}.$$

$$-E\left(\frac{\partial^2 \log L}{\partial\beta_j \partial\sigma}\right) = \frac{1-\gamma}{\sigma^2}\sum_{i=1}^{n} x_{ji}. \tag{3.23}$$

$$-\mathrm{E}\left(\frac{\partial^2 \log L}{\partial\alpha \partial\sigma}\right) = \frac{n}{\sigma^2}(1-\gamma).$$

$$-\mathrm{E}\left(\frac{\partial^2 \log L}{\partial\sigma^2}\right) = \frac{n}{\sigma^2}\left(\frac{\pi^2}{6} + \gamma^2 - 2\gamma + 1\right).$$

In the context of the SIR, the covariates x come from a elliptically symmetric distribution and satisfy $\sum_{i=1}^{n} x_{ji} = 0$, for a given j. That results in $\mathrm{Cov}(\beta,\alpha) = \mathrm{Cov}(\beta,\sigma) = 0$. Finally, we write down the Fisher's information matrix. It is block-diagonal, and its structure is specified for σ, α and β. We do not write the element $*$ part but its formula can be found above in (3.23). We note that all the elements except the (β,β) part are constants.

$$I(\alpha,\beta,\sigma) = \frac{1}{\sigma^2}\cdot \begin{array}{c|c|c|c} & \sigma & \alpha & \beta \\ \sigma & * & n(1-\gamma) & 0 \\ \hline \alpha & n(1-\gamma) & n & 0 \\ \hline \beta & 0 & 0 & X^T X \end{array}$$

What interest us in the matrix $I(\alpha,\beta,\sigma)$, is the variance of the parameters of interest, namely β. Since the Weibull distribution with fixed shape parameter is an exponential family, the asymptotic variance of β is the inverse of the Fisher's information matrix. Its block-diagonal structure allows for a simple inverse:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}. \tag{3.24}$$

In formula (3.24) we easily recognize the variance of the linear regression estimates, derived from the ML estimation. Thus, we conclude that the Weibull regression without censoring is fully efficient (in terms of maximum likelihood).

3.6.2. Asymptotics

What about the asymptotic variance? We saw in Chapter 2 the asymptotic theory for a SIR estimate. Can the effect of the censoring and reweighting be somehow integrated into Theorem 2.2.8? In principe, yes, but not in a convenient and easy to use way. Buckley and James (1979) suggested an estimator based on a modification of the ordinary least squares method. An alternative way to address this problem is based on linear rank tests, derived by Prentice (1978) using a score test for the marginal likelihood of generalized ranks. Wei et al. (1990) proposed a method to make inference on a subset of the regression coefficients, while Tsiatis (1990) introduced a method to consistenly estimate the variances for the linear rank estimates, using the counting process approach. Ritov (1990) showed the asymptotic equivalence between the Buckley-James estimator and the linear rank test estimator of Tsiatis. All his results are based on counting process martingale theory.

Based on these facts, it is difficult to integrate the martingale theory of censored regressions into the asymptotic theory of SIR. Even the formula itself for the asymptotic variance of the regression coefficients by Tsiatis (1990) is difficult to apply in practice. Therefore, a good strategy to estimate the variance of the adapted SIR is to apply the bootstrap.

The bootstrap of estimators based on censored data were first studied by Efron (1981). He showed that for the observed data of $(T_i, \delta_i, x_i,\ i = 1, \ldots, n)$, a bootstrap sample of $(T_1^*, \delta_1^*, x_1^*)$, $(T_2^*, \delta_2^*, x_2^*)$, $\ldots$, $(T_n^*, \delta_n^*, x_n^*)$, drawn by independent sampling n times with replacement, yields an appropriate variance estimation. The "obvious" method, demanding the independent sampling for the lifetime and the censoring time points, keeping the minimal value between the two, and defining the indicator accordingly, is equivalent to the triplets sampling.

To check the validity of the bootstrap variance estimation in our case, we performed a number of simulations. Figures 3.11 and 3.12 show the boxplots of the bootstrap estimates under the censoring rates of $20\%, 33\%, 50\%, 60\%$ and 75% for the PH and

ALT models respectively, on two sample sizes, $n = 300$ and $n = 1000$. The top row of each figure lists the boxplots for a non-zero component, while the bottom line corresponds to a zero component. The models were generated in a similar way as in Section 3.5, for $p = 10$ variables, and the number of bootstrap replications was set to 100.

The results from our simulations indicate that the bootstrap estimation of our SIR parameters is consistent. We only present the results for the two components for each model (a true zero and non-zero one), but they reflect the general behavior. Figures 3.11 and 3.12 indicate that the variation in estimation decreases as the sample size grows, allowing to extrapolate for the asymptotic case. For the PH model, the variances of the true zero coefficients decrease faster than those of the non-zero coefficients. The ALT model presents more variability in estimation, which is linked to the censoring pattern and its sensitivity to it. Even on $n = 1000$, we observe many outliers in our estimation. However, starting from $n = 5000$ (not pictured), the results are more stable and accurate. We note here that the bootstrap estimation does not reach the true value of the non-zero coefficient (which is 0.45), unlike in Section 3.5, but captures the overall structure of the regression coefficient vector (signs, ratios) flawlessly.

A comparison between the bootstrap and the simulated variances reveals that the bootstrap variances give good estimates of the simulated variances.

3.7. High-dimensional case

The main problem why the original SIR method cannot be applied lies in the fact that when $n \leq p$, the covariance matrix Σ is ill-defined, making its inversion a problem (we perform this at the very first step of the algorithm). There are many possible techniques of regularization, and some papers on that topic were mentioned in the beginning of this chapter. The most popular technique seems to be ridge regularization. The choice of methods and respective regularization parameters is out of the scope of this thesis.

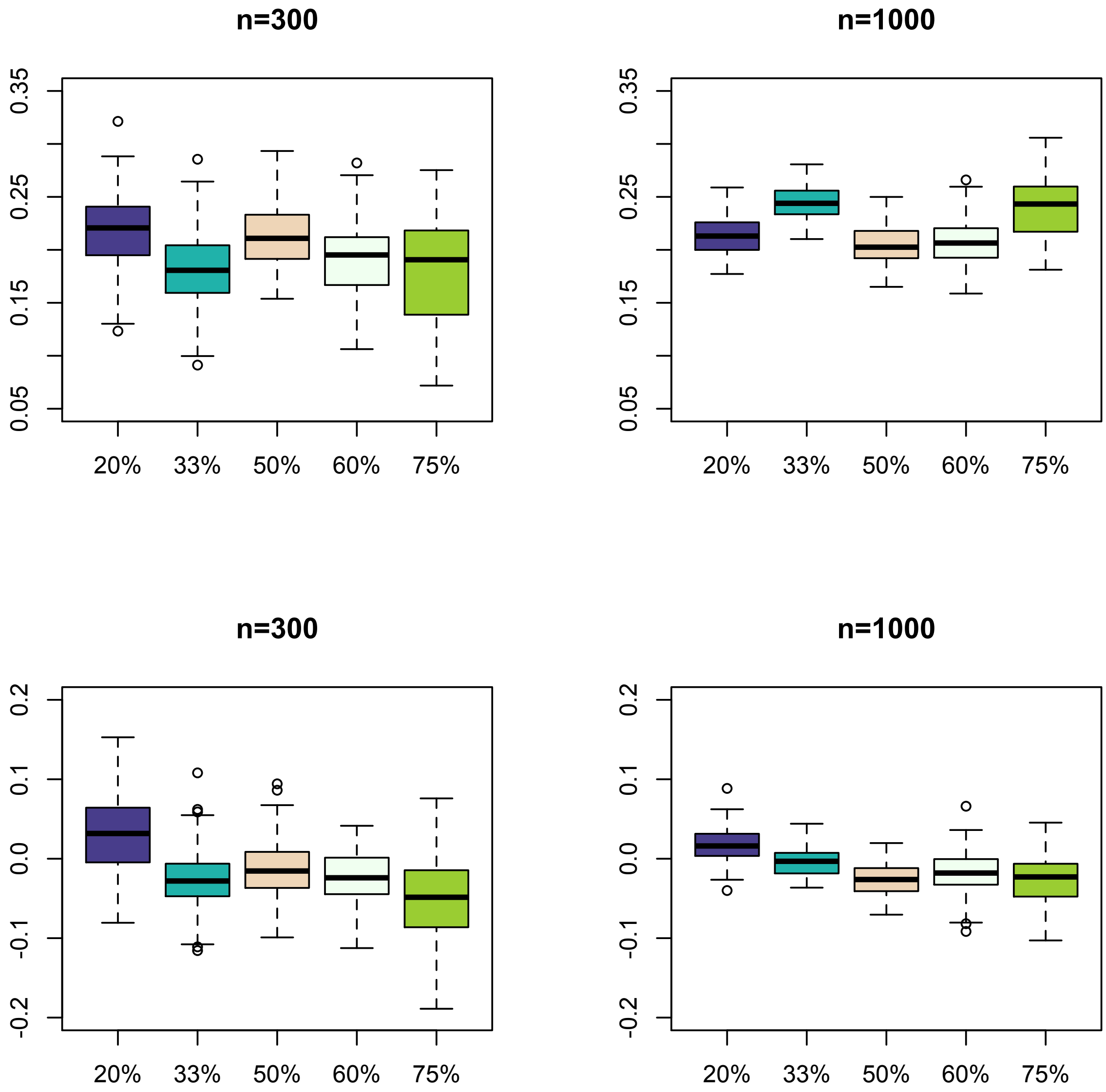

Figure 3.11.: Boxplots of the bootstrap estimates (PH model) under 5 different censoring rates for a non-zero (top line) and a zero (bottom line) component. Results for two samples sizes are listed, $n = 300$ and $n = 1000$.

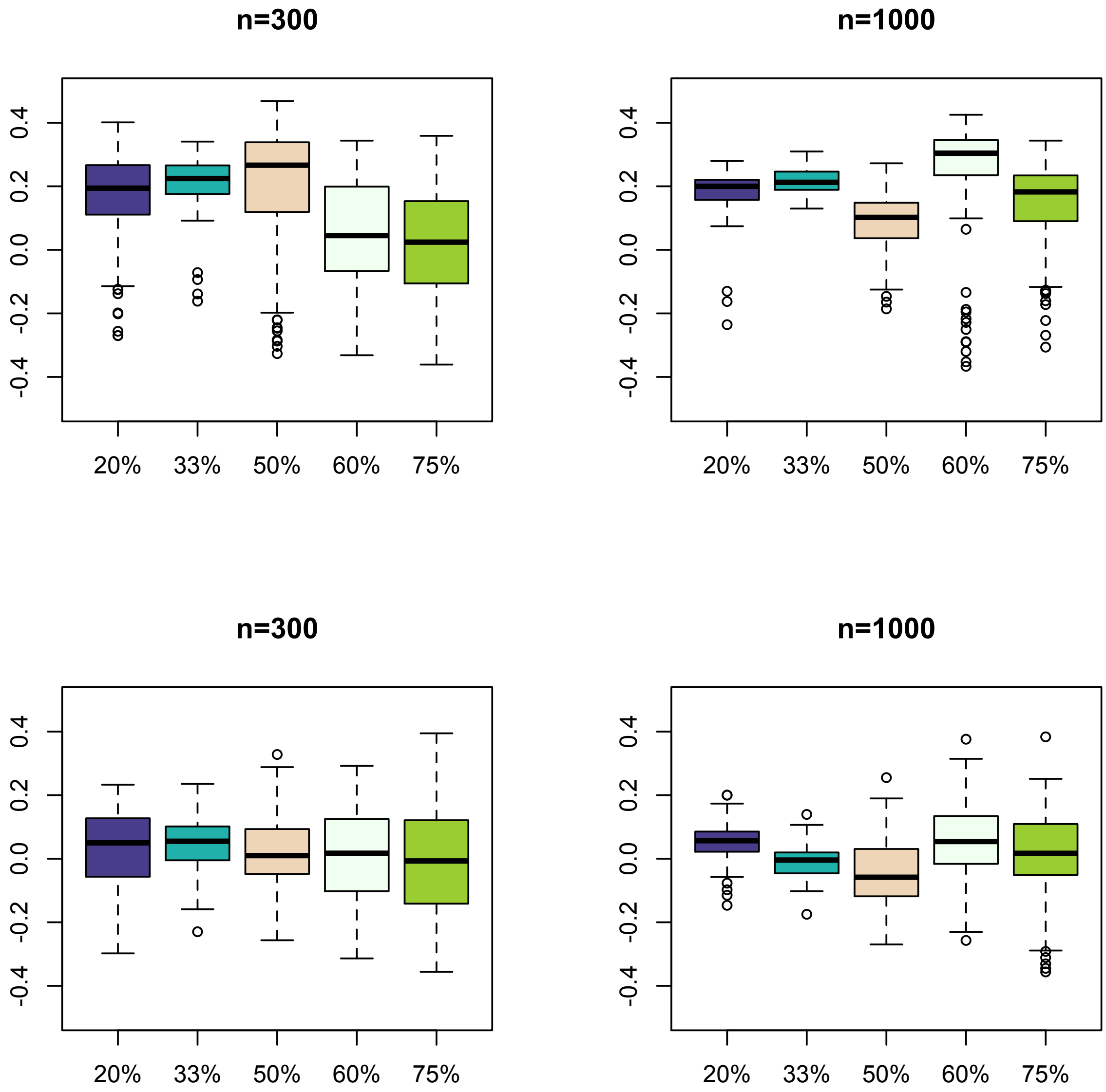

Figure 3.12.: Boxplots of the bootstrap estimates (ALT model) under 5 different censoring rates for a non-zero (top line) and a zero (bottom line) component. Results for two samples sizes are listed, $n = 300$ and $n = 1000$.

3.8. Defining the true zero coefficients and confidence intervals

While visualizing all the coefficients of the principal eigenvector as we have done so far in Tables above, may be informative, it is not the best way to judge the performance of the algorithm, especially in the case of a larger number of variables. In practice one is interested in finding the regression coefficients which differ significantly from zero. Confidence intervals seem to be a natural choice for such a task.

From the simulations listed above we see that in most cases the true zero coefficients are random and of order of $O_p(1/\sqrt{n})$. Since we estimate the variance of the SIR regressors via bootstrap, the classical confidence interval for a coefficient β_i, a component of a vector β, would be:

$$[\hat{\beta}_i - c \cdot \hat{\sigma}_{boot}, \quad \hat{\beta}_i - c \cdot \hat{\sigma}_{boot}], \tag{3.25}$$

where $\hat{\beta}_i$ is an estimated value for β_i, $\hat{\sigma}_{boot}$ is its estimated standard deviation, and c is a Student's quantile.

In our method, we constructed the confidence intervals based on the percentile method, first introduced by Efron (1981) and later studied by Efron and Tibshirani (1993). For a given α, we construct a $(1 - 2\alpha)$ confidence interval by using the percentiles of the bootstrap distribution β^* in a following way:

$$[\beta^*_\alpha, \quad \beta^*_{1-\alpha}], \tag{3.26}$$

where $\beta^*_\alpha = CDF^{-1}(\alpha)$ and $\beta^*_{1-\alpha} = CDF^{-1}(1-\alpha)$ are the corresponding percentiles of the sample distribution of β^*.

We used Formula (3.26) to test the null hypothesis $H_0 : \ \beta_i = 0$ in our models, are the results were very satisfactory.

In this chapter, we covered the aspects of applying the sliced inverse regression to the survival datasets. Despite SIR being a non-parametric method, we studied its performance on two types of the models, the accelerated lifetime and the proportional hazards, claiming that the consideration of the censored observations, based on the suggested model, improves the estimation. Different adaptations of the method are often not too sensitive to the model specification. Various simulations prove our approach to be effective in selecting the important variables.

CHAPTER

4

SIZING STUDIES FOR DETECTING GRAPHICAL MODELS

This chapter studies the factors that influence the power of the partial correlation test in detecting the structure of the Gaussian graphical models. We are particularly interested in how the sample size n affects this power. We concentrate on the case of a single partial correlation, where a local asymptotic power approach and a Kullback-Leibler divergence are considered. The Kullback-Leibler approach allows us to get a better understanding of the complexity of the edge detection with regard to a value of the partial correlation.

4.1. Background

Probabilistic graphical models are graphs in which nodes represent random variables X_u and the edges represent conditional dependence. Any two nodes or variables that are not connected are independent, conditional on the values of all the other random variables (in this work we only consider undirected graphs).

Such models provide a compact representation of a joint probability distribution. The theoretical aspects of graphical models are nicely covered by Lauritzen (1996) and Whittaker (1990). As an application to genetical epidemiology, the variables X_u can be viewed as gene expressions measured from tissue samples of n patients. The graphical model is used to describe the association between genes. We write $X_1 \perp\!\!\!\perp X_2 \mid X_3 \ldots X_p$ to indicate that X_1 and X_2 are conditionally independent, given $X_3 \ldots X_p$. It turns out that for the multivariate normal distribution, conditional independence is equivalent to zero entries in the inverse covariance matrix Σ^{-1} (also called a concentration or precision matrix). Thus, if $X \sim \mathcal{N}_p(\mu, \Sigma)$ is a p-dimensional normal random vector with regular Σ, then for $1 \leq u, v \leq p$ with $u \neq v$

$$X_u \perp\!\!\!\perp X_v \mid X_{rest} \Longleftrightarrow \sigma^{uv} = 0\,, \tag{4.1}$$

where $\Sigma^{-1} = (\sigma^{uv})_{u,v=1}^{p}$. This fact follows from the following theorem:

Theorem 4.1.1 *If* $\begin{pmatrix} X \\ Y \end{pmatrix}$ *is normally distributed with the expected value* $\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$ *and covariance matrix* $\begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$ *then* $(X \mid Y)$ *is also normally distributed with the expected value* $\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y)$ *and covariance matrix* $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$*.*

Thus the conditional independence in the case of the multivariate normal distribution is expressed through zero elements in the concentration matrix. We can identify the matrix

$$\Sigma^{uv} = \begin{pmatrix} \sigma^{uu} & \sigma^{uv} \\ \sigma^{vu} & \sigma^{vv} \end{pmatrix}$$

as the concentration matrix of the conditional distribution of $(X_u,\, X_v) \mid X_{rest}$. The covariance matrix of this conditional distribution is therefore equal to

$$\Sigma_{uv|rest} = \frac{1}{|\Sigma^{uv}|} \begin{pmatrix} \sigma^{vv} & -\sigma^{uv} \\ -\sigma^{vu} & \sigma^{uu} \end{pmatrix}. \tag{4.2}$$

From (4.2) it follows that

$$\mathrm{Cov}(X_u, X_v \mid X_{rest}) = \frac{-\sigma^{uv}}{\sigma^{uu}\sigma^{vv} - (\sigma^{uv})^2},$$

which proves (4.1).

Estimating the structure of the concentration matrix (or a graphical model) from data can be solved with a variety of statistical procedures. They can be classified as score-based methods (for instance, a penalized likelihood), Bayesian methods (identifying posterior distributions over graphs) and constraint-based methods (testing for conditional independencies and identifying compatible independence structures). Penalized likelihoods have been extensively studied over the last couple of years, allowing for a sparse solution in high-dimensional scenario (see Cai et al. (2011); Loh and Wainwright (2012); Meinshausen and Buehlmann (2006). In this chapter, we consider testing the inclusion of every edge separately, edge by edge (which would be a part of the constraint-based approach). Thus, we have to test $H_0^{uv} : \rho_{uv\cdot\text{rest}} = 0$ for all $\binom{p}{2}$ choices of u and v, where rest refers to the variables with indices in $\{1, 2, \ldots, p\} \setminus \{u, v\}$.

4.1.1. Testing for Edges

Based on a sample $x_1 \ldots x_n \in \mathbb{R}^p$, we estimate the covariance matrix Σ by $S_n = (s_{uv})_{u,v=1}^{p}$, whose elements in the case $n > p$ can be taken as $s_{uv} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{iu} - \bar{x}_u)(x_{iv} - \bar{x}_v)$, whereas in high-dimensional cases, some regularization needs to be applied. In the following we assume that $n > p$. Without loss of generality, let $u = 1$, $v = 2$ and consider

$$H_0^{12} : \rho_{12\cdot\text{rest}} = 0 \quad \text{against} \quad H_A^{12} : \rho_{12\cdot\text{rest}} \neq 0.$$

The standard estimate of $\rho_{12\cdot\text{rest}}$ is $s_{12\cdot\text{rest}} = -\frac{s^{12}}{\sqrt{s^{11} s^{22}}}$. Under multivariate Gaussianity, its sampling distribution is equal to the sampling distribution of the ordinary product-moment correlation with the sample size n replaced by $n - (p - 2)$. Assuming the null hypothesis H_0^{12} is true, this implies that

$$s_{12\cdot\text{rest}} \sqrt{\frac{n-p}{1 - s_{12\cdot\text{rest}}^2}} \sim t_{n-p}, \tag{4.3}$$

a Student's t - distribution with $n - p$ degrees of freedom. In order to achieve level α with a two-sided test, we reject H_0^{12} if the absolute value of the test statistic exceeds the $(1-\alpha/2)$ quantile of the t_{n-p} - distribution, which we denote by $t_{n-p,\alpha/2}$. We thus reject if

$$\left| s_{12\cdot\text{rest}} \sqrt{\frac{n-p}{1-s^2_{12\cdot\text{rest}}}} \right| > t_{n-p,\alpha/2} \text{ or } |s_{12\cdot\text{rest}}| > \frac{t_{n-p,\alpha/2}}{\sqrt{n-p+(t_{n-p,\alpha/2})^2}}\,. \tag{4.4}$$

This rule depends only very weakly on p, but this is misleading, because we have to test $m = p(p-1)/2$ null hypotheses. Without some correction, the probability of getting a falsely significant edge simply by chance increases with p. To avoid this, we can use the Bonferroni correction or choose to control the False Discovery Rate (FDR).

Other approaches based on test statistics are possible. In Edwards (2000), for example, the backward stepwise model selection on the basis of the χ^2 - distribution is suggested. Drton and Perlman (2004), on the other hand, discuss the estimation of the confidence interval for the partial correlation coefficient.

In many modern applications one has $n < p$, that is, there are more variables than observations. In this case, the above method is clearly no longer available. In fact, it is not possible to estimate Σ^{-1} by maximum likelihood and only a regularized procedure can assure a positive definite estimate. A variety of solutions have been proposed, see for example Kraemer et al. (2009).

4.1.2. The Kullback-Leibler Divergence

The investigation of the feasibility of edge-detection can also be based on the Kullback-Leibler divergence. The Kullback-Leibler divergence (KLD) measures the difference between two probability distributions F_1 and F_2 with densities f_1 and f_2 (see Kullback (1997)). It is defined as

$$D(f_1 \,|\, f_2) = \int f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx\,. \tag{4.5}$$

The divergence equals the expected value of the log-likelihood-ratio for a single observation $X \sim F_1$ when testing the model F_1 vs the model F_2. This interpretation shows that the KLD is in fact a universal information number and is not tied to the particular model being considered. It is easy to show that this divergence is always positive unless $F_2 = F_1$. Furthermore, the bigger the KLD, the easier it is to distinguish F_1 from F_2 by likelihood tests and the more powerful a test distinguishing between the two hypothesis F_1 and F_2 is. If we dispose of n independent observations, the KLD is multiplied by n. If we test the absence of partial correlations vs. the presence of partial correlations and assume multivariate Gaussianity, the KLD is a useful tool to determine the average amount of information in the data. Because it is based on likelihoods rather than estimates, the KLD can be computed for any two models, without reference to additional conditions such as $n > p$. This is an advantage of this approach.

Further on in this chapter, we will examine how information accumulates when trying to fit a graphical model. When testing for edges, we will be interested in the power of the test, while in the KLD approach, we can directly compute the relevant amount of information.

4.2. Evaluation of the partial correlation test

We investigated the performance of this testing procedure both by simulation and via asymptotic power calculations.

4.2.1. Asymptotic power

When testing the null hypothesis $H_0^{12} : \rho_{12 \cdot \text{rest}} = 0$ against one-sided alternative $\rho_{12 \cdot \text{rest}} > 0$ based on the estimator $s_{12 \cdot \text{rest}}$ and the Bonferroni correction for the number of tests $m = p(p-1)/2$, the power function for large sample sizes is approximately equal to

$$1 - \Phi\left(z_{1-\alpha/m} - c\sqrt{n}\rho_{12 \cdot \text{rest}}\right) = \Phi\left(c\sqrt{n}\rho_{12 \cdot \text{rest}} - z_{1-\alpha/m}\right) , \tag{4.6}$$

where $z_{1-\alpha/m}$ denotes the $(1-\alpha/m)$- quantile of the standard normal distribution and c is a constant. This approximation of the power comes from the Pitman asymptotic relative efficiency theory and is based on the asymptotic normality of the estimator of the partial correlation and on the consideration of alternatives close to the null hypothesis (details can be found, for example, in Chapter 10 of Serfling

(1980), or in Chapter 14 of Van der Vaart (1998)). The constant c is called the Pitman efficacy or the slope of the test and is defined in our case as

$$c = \frac{\mu^{'}(0)}{\sigma(0)}, \tag{4.7}$$

where $\mu(0)$ and $\sigma(0)$ are the expected value and the standard deviation of the partial correlation test evaluated under the null hypothesis. To compute the Pitman efficacy, we get the needed results from Muirhead (1982). The expected value of the partial correlation for an estimate based on a normal sample is

$$2/f(\Gamma[(f+1)/2]/\Gamma[f/2])^2 \rho_{12\cdot\text{rest}} \, {}_2F_1\left[\tfrac{1}{2}, \tfrac{1}{2}; (f+2)/2, \rho^2_{12\cdot\text{rest}}\right], \tag{4.8}$$

where ${}_2F_1[\cdot]$ is a hypergeometric function and $f = n + 1 - p$. It follows that its derivative with respect to $\rho_{12\cdot\text{rest}}$ and evaluated at $\rho_{12\cdot\text{rest}} = 0$ equals

$$2/f(\Gamma[(f+1)/2]/\Gamma[f/2])^2 \, {}_2F_1\left[\tfrac{1}{2}, \tfrac{1}{2}; (f+2)/2, \rho^2_{12\cdot\text{rest}}\right]\Bigg|_{\rho^2_{12\cdot\text{rest}}=0}$$

$$= 2/f(\Gamma[(f+1)/2]/\Gamma[f/2])^2.$$

Using Stirling's formula one can approximate the values of the Gamma function and finds the following value:

$$2/f(\Gamma[(f+1)/2]/\Gamma[f/2])^2 \approx \frac{(1+1/f)^f}{e} \xrightarrow[f\to\infty]{} \frac{e}{e} = 1. \tag{4.9}$$

Because the asymptotic variance of the partial correlation statistic assuming that the null hypothesis is true is 1, the slope of the test or its Pitman efficacy is equal to 1.

Using the asymptotic approximation for the quantile $z_{1-\alpha/m} \sim \sqrt{2\log(m/\alpha)}$, and the fact that $c = 1$, we can rewrite (4.6) as

$$\Phi\left(\sqrt{n}\rho_{12\cdot\text{rest}} - \sqrt{2\log(p(p-1)/(2\alpha))}\right) \tag{4.10}$$

for an actual partial correlation of size $\rho_{12\cdot\text{rest}} > 0$. For a two-sided test one can replace α by $\alpha/2$ to obtain an approximate power value. The approximation of the quantile $z_{1-\alpha/m}$ is quite crude and gives values that are typically too large so that the power might be underestimated. The formula shows that the asymptotic power increases with increasing strength of the partial correlation, where the strength is measured by

$$\frac{\sqrt{n}\sigma^{12}}{\sqrt{\sigma^{11}\sigma^{22}}} = \sqrt{n}\,\rho_{12\cdot\text{rest}}\,. \tag{4.11}$$

The approximation (4.10) is valid if $n - p$ is large. Of course, in high-dimensional situations this is typically not the case.

We can use the approximation (4.10) to compute n as a function of the complexity of the graph p and the size of the partial correlation. The formula shows that the sample sizes required to reach a certain power (for example 0.8), are inversely proportional to the square of the partial correlation we wish to detect. The required sample size also increases roughly in parallel with the square root of $\log(p)$.

4.2.2. Simulated power

We simulated the power function by simple Monte Carlo according to the following schema:

1. Set up the concentration matrix Σ^{-1} with diagonal elements equal to 1 and nonzero off-diagonal elements equal to ρ and of a given sparseness of 10%; 20%; 40%.
2. Repeatedly simulate data $X_1 \ldots X_n \overset{\text{iid}}{\sim} \mathcal{N}_p(0, \Sigma)$ with the chosen values of n and p.
3. Compute the sample covariance matrix $S(X_1 \ldots X_n)$ and its inverse S^{-1}.
4. Using (4.4), test whether the elements of S^{-1} are significantly different from zero.
5. Compare the original Σ^{-1} and the graph derived from S.

To compute the finite sample power of the test, we started with a sparseness of $1/\binom{p}{2}$, that is, a single non-zero off-diagonal element, which we positioned at $(1, 2)$. All simulations were performed for matrices of dimension $p = 40$, $p = 100$

and a sample size of $n = 4000$ and $n = 300$. The power of detection is estimated as

$$\text{estimated Power} = \frac{\text{No. times } H_0^{12} \text{ is rejected}}{\text{No. trials}}.$$

It turns out that the power does not depend strongly on the dimension p. Moreover, due to the extreme sparseness, the performances of the two corrections are almost equal. Figure 4.1 below shows the power curves from simulations and asymptotic formula (4.10) for two cases of parameters ($p = 40$ and $p = 100$) on a sample set of 300. The asymptotic curve describes well the power obtained via simulations even on small samples, and it only gets better for larger samples (not pictured).

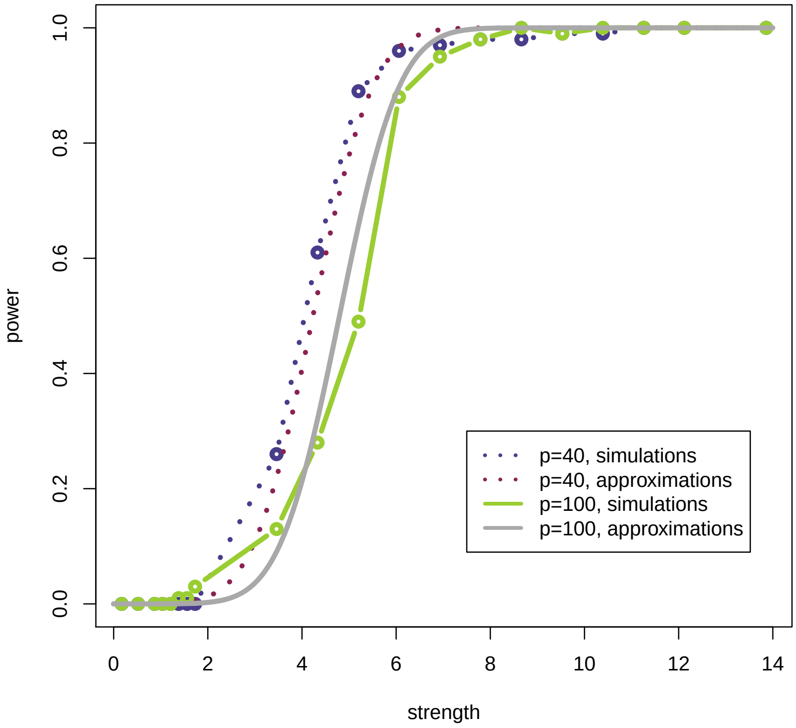

Figure 4.1.: Estimated and local asymptotic power as a function of edge strength (4.11) for a sample size $n = 300$. The values of p are 40 (dotted) and 100 (solid).

We also checked how well the test detects multiple connections. A set of simulations was performed on a 20%-sparse concentration matrix for different sample sizes. We tested whether the procedure could detect all the existing edges. As expected, only very strong connections were detected on small samples.

4.3. The Kullback-Leibler Divergence

Suppose we have two p-variate normal populations $\mathcal{N}_p(\mu_i, \Sigma_i)$ (i=1, 2), with $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})$ two vectors of mean values, and $\Sigma_i \in \mathbb{R}^{p \times p}$ two covariance matrices. Their respective population densities are

$$f_i(x) = \frac{1}{|2\pi\Sigma_i|^{1/2}} \exp(-\tfrac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))\,. \tag{4.12}$$

It follows that

$$\log\frac{f_1(x)}{f_2(x)} = \tfrac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \tfrac{1}{2}\mathrm{tr}\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T + \tfrac{1}{2}\mathrm{tr}\Sigma_2^{-1}(x-\mu_2)(x-\mu_2)^T\,.$$

Taking the expectation of the above, we can rewrite (4.5) in the case of these two multivariate normal populations as

$$D(f_1\,|\,f_2) = \tfrac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + \tfrac{1}{2}\mathrm{tr}\Sigma_1(\Sigma_2^{-1} - \Sigma_1^{-1}) + \tfrac{1}{2}\mathrm{tr}\Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T. \tag{4.13}$$

We will next evaluate this formula with normal populations with equal means but unequal covariance matrices, namely Σ_2 equal to the identity matrix and Σ_1 with an inverse which is nearly equal to the identity matrix. In the context of partial correlations this describes a situation where the p variables have equal variance and only a very small proportion of all partial correlations is non-zero.

4.3.1. Divergence for a single non-zero partial correlation with known placement

Let f_{ij} be the density of the multivariate normal $\mathcal{N}_p(0, \Sigma_{ij})$ and let f be the density of $\mathcal{N}_p(0, I)$, where Σ_{ij}^{-1} has diagonal elements equal to 1 and has a value of $\rho > 0$ in positions (i,j) and (j,i), that is, exactly one partial correlation is non-zero. The divergence then takes the following form:

$$D(f_{ij}\,|\,f) = \tfrac{1}{2}\log\frac{1}{|\Sigma_{ij}|} + \tfrac{1}{2}\mathrm{tr}\Sigma_{ij}(I - \Sigma_{ij}^{-1}) = \tfrac{1}{2}(\log|\Sigma_{ij}^{-1}| + \mathrm{tr}\Sigma_{ij} - p). \qquad (4.14)$$

Because of our assumption about Σ_{ij}^{-1}, we have

$$\mathbf{\Sigma_{ij}^{-1}} = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & \dots & \rho & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \rho & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix} = I_p + U_{ij}\,,$$

where the off-diagonal elements are in positions (i, j) and (j, i). Without loss of generality we take $i < j$. The expansion of the determinant gives

$$|\Sigma_{ij}^{-1}| = 1 + (-1)^{i+j}\rho(-1)^{j-1+i}\rho = 1 + \rho^2(-1)^{2i+2j-1} = 1 - \rho^2.$$

Thus, the determinant does not depend on the position (i, j), nor does it depend on the dimension p. In order to compute the trace of Σ_{ij}, we only need its diagonal elements, which are straightforward to compute. They are equal to 1 except in positions (i, i) and (j, j), where they are $\frac{1}{1-\rho^2}$. Thus, the trace we are looking for is equal to $p - 2 + \frac{2}{1-\rho^2}$. From the previous expression (4.14) we then find

$$D(f_{ij}\,|\,f) = \frac{\log(1-\rho^2)}{2} + \frac{\rho^2}{1-\rho^2}\,. \qquad (4.15)$$

4.3.2. Divergence when the correlation pair is unknown

In our formulation of the density f_{ij}, we make use of the knowledge of the placement of the positive partial correlation. Because of this, Equation (4.15) is only useful in understanding the test of H_0^{ij} if no correction for p is made.

How does the divergence change if we do not know the pair of correlated variables?

To answer this question, we consider the mixture of normal distributions

$$g(x) = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \frac{1}{\binom{p}{2}} f_{ij}(x) \,. \tag{4.16}$$

We showed earlier that the determinant of the identity matrix with two off-diagonal elements of value ρ added equals $1-\rho^2$. The quadratic form in the exponential can also be explicitly evaluated as $x_1^2 + \cdots + x_p^2 + 2\rho x_i x_j$. It follows that the elements of the mixture density g are:

$$f_{ij}(x) = \frac{\sqrt{1-\rho^2}}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(x_1^2 + \cdots + x_p^2 + 2\rho x_i x_j)\right).$$

The ratio of the two densities becomes

$$\begin{aligned}\frac{g(x)}{f(x)} &= \frac{\frac{\sqrt{1-\rho^2}}{\binom{p}{2}} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \exp\left(-\frac{1}{2}(x_1^2 + \cdots + x_p^2 + 2\rho x_i x_j)\right)}{\exp(-\frac{1}{2}(x_1^2 + \cdots + x_p^2))} \\ &= \frac{\sqrt{1-\rho^2}}{\binom{p}{2}} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \exp(-\rho x_i x_j) \,.\end{aligned}$$

It follows that

$$\begin{aligned} D(g \,|\, f) &= \int_{\mathcal{R}^p} g(x) \log\left(\frac{g(x)}{f(x)}\right) dx \\ &= \int_{\mathcal{R}^p} g(x) \log\left(\frac{\sqrt{1-\rho^2}}{\binom{p}{2}} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \exp(-\rho x_i x_j)\right) dx \\ &= \int_{\mathcal{R}^p} f_{12}(x) \log\left(\frac{\sqrt{1-\rho^2}}{\binom{p}{2}} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \exp(-\rho x_i x_j)\right) dx, \end{aligned} \tag{4.17}$$

where the last line follows from the fact that the ratio of the densities is invariant with respect to the permutations of the x_i.

The integral (4.17) can be approximated as follows for large p. Let $(X_1, \ldots, X_p)$ be a random vector with density f_{12}. The integrand in (4.17) is

$$\begin{aligned}
&\log\left(\frac{\sqrt{1-\rho^2}}{\binom{p}{2}}\left(e^{-\rho X_1X_2}+\sum_{j=3}^{p}\left(e^{-\rho X_1X_j}+e^{-\rho X_2X_j}\right)+\sum_{i=3}^{p-1}\sum_{j=i+1}^{p}e^{-\rho X_iX_j}\right)\right)\\
&\quad=\log\left(\sqrt{1-\rho^2}\left(O_P(p^{-2})+\frac{2}{p^2-p}\left(\sum_{j=3}^{p}\left(e^{-\rho X_1X_j}+e^{-\rho X_2X_j}\right)\right)\right.\right.\\
&\qquad\left.\left.+\frac{2}{p^2-p}\left(\sum_{i=3}^{p-1}\sum_{j=i+1}^{p}e^{-\rho X_iX_j}\right)\right)\right). \qquad (4.18)
\end{aligned}$$

For large p, the two last terms should be well-behaved and can be approximated by their asymptotic limits. The theory of a product of two independent normal variables is well described by Aroian (1947). We can easily show that if Z_1, Z_2 are independent and $\mathcal{N}(0,1)$, then

$$\begin{aligned}
\mathrm{E}(\exp(-\rho Z_1Z_2)) &= 1/\sqrt{1-\rho^2},\\
\mathrm{Var}(\exp(-\rho Z_1Z_2)) &= 1/\sqrt{1-4\rho^2}-1/(1-\rho^2).
\end{aligned}$$

Thus, the variance is only finite for $\rho < 0.5$ and a transition in the KLD value happens as one passes to $\rho > 0.5$. These equations are computed in a straightforward way:

$$\begin{aligned}
\mathrm{E}(\exp(-\rho Z_1Z_2)) &= \iint \exp(-\rho Z_1Z_2)\phi(z_1)\phi(z_2)dz_1dz_2\\
&= \iint \frac{1}{2\pi}\exp\left(-\frac{z_1^2}{2}-\frac{z_2^2}{2}-\rho z_1z_2\right)dz_1dz_2\\
&= \int \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z_2^2}{2}+\frac{\rho^2z_2^2}{2}\right)\left(\int\frac{1}{\sqrt{2\pi}}\exp(-\frac{(z_1+\rho z_2)^2}{2})dz_1\right)dz_2\\
&= \frac{1}{\sqrt{1-\rho^2}}\int\frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}}\exp\left(-\frac{z_2^2}{2}(1-\rho^2)\right)dz_2=\frac{1}{\sqrt{1-\rho^2}},
\end{aligned}$$

where the ϕ function represents the density of the standard normal distribution. The result for $\mathrm{Var}(\exp(-\rho Z_1Z_2))$ is computed analogously.

Thus, if $\rho < 0.5$, we can appeal to the central limit theorem to deduce that

$$\left(\sum_{i=3}^{p-1}\sum_{j=i+1}^{p} e^{-\rho X_i X_j}\right)/[(p-2)(p-3)/2] \overset{\text{asy}}{\sim} \mathcal{N}\left(\frac{1}{\sqrt{1-\rho^2}},\ O(p^{-2})\right).$$

The variables $X_1, \ldots, X_n$ have the density f_{12}, that is, their inverse covariance matrix is diagonal from $i = 3, \ldots, n$ and has ρ in the positions (1, 2) and (2, 1). Hence, the marginal variance of X_1 and X_2 is

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \frac{1}{1-\rho^2}.$$

The second summand in (4.18) is based on the variables X_1 and X_2, and the expectation of the typical term $\exp(-\rho X_1 Z_i)$ can be computed in the following way:

$$\begin{aligned}\mathrm{E}(\exp(-\rho X_1 Z_i)) &= \iint \exp(-\rho x_1 z_i)\frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}}\exp\left(-\frac{x_1^2}{2}(1-\rho^2)\right)\phi(z_i)dz_i dx_1 \\ &= \sqrt{1-\rho^2}\int \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(z_i+\rho x_1)^2}{2}\right)dz_i \int \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_1^2}{2}(1-2\rho^2)\right)dx_1 \\ &= \frac{\sqrt{1-\rho^2}}{\sqrt{1-2\rho^2}}. \qquad (4.19)\end{aligned}$$

The variance of this term $\exp(-\rho X_1 Z_i)$ is also $O(p^{-2})$. In this computation, we replaced X_i by Z_i to highlight the fact that X_i, $(i = 3, \ldots, p) \sim \mathcal{N}(0,1)$.

The last thing to account for is a number of terms in each of the summands in (4.18). We get that

$$\begin{aligned}\mathrm{E}\left(\frac{2}{p^2-p}\sum_{j=3}^{p}\left(e^{-\rho X_1 X_j}+e^{-\rho X_2 X_j}\right)\right) &= \frac{4(p-2)}{p(p-1)}\frac{\sqrt{1-\rho^2}}{\sqrt{1-2\rho^2}} \\ &\approx \frac{4}{p}\frac{\sqrt{1-\rho^2}}{\sqrt{1-2\rho^2}}; \\ \mathrm{E}\left(\frac{2}{p^2-p}\sum_{i=3}^{p-1}\sum_{j=i+1}^{p} e^{-\rho X_i X_j}\right) &= \frac{(p-2)(p-3)}{p^2-p}\frac{1}{\sqrt{1-\rho^2}}\end{aligned}$$

$$= (1 - \frac{4}{p} + O(p^{-2}))\frac{1}{\sqrt{1-\rho^2}}.$$

Combining everything together we get that the KLD is approximately equal to the expectation of $\log(\sqrt{1-\rho^2}(Y + O_p(p^{-2})))$, where $\mathrm{Var}(Y) = O(p^{-2})$ and its mean is $(1-4/p)/\sqrt{1-\rho^2} + (4/p)\sqrt{1-\rho^2}/\sqrt{1-2\rho^2}$. Multiplying by $\sqrt{1-\rho^2}$ and expanding the logarithm leads to an approximate KLD value of

$$\begin{aligned} D(g\,|\,f) &= \log\left(1 - \frac{4}{p} + \frac{4}{p}\frac{(1-\rho^2)}{\sqrt{1-2\rho^2}} + O(p^{-2})\right) \\ &= \frac{4}{p}\left(\frac{1-\rho^2}{\sqrt{1-2\rho^2}} - 1\right) + O(p^{-2})\,. \end{aligned} \quad (4.20)$$

A comparison of (4.15) and (4.20) reveals that the effect of an unknown placement of the partial correlation is a much slower increase with ρ of the information content. Table 4.1 shows values of the KLD, calculated by Monte Carlo integration.

ρ	$p = 10$	$p = 100$	$p = 1000$
0.1	4×10^{-4}	2×10^{-5}	6×10^{-6}
0.5	2.026×10^{-2}	1.77×10^{-3}	8×10^{-5}
0.9	2.05	1.1	5.9×10^{-1}

Table 4.1.: The entries of this table show the information about a single randomly placed partial correlation of size ρ provided by one subject. The information is shown as a function of the number of genes p. For studies involving n subjects, the information can be multiplied by n.

Figure 4.2 compares the numerical true values with the approximation. Our approximation predicts that the information content decreases for small values of the partial correlation ($\rho < 0.5$) very rapidly with the dimension p, much more rapidly than the decrease predicted by the local power, where the dimension p entered via $\log(p)$. The analysis via the KLD leads to the conclusion that only very large studies will be able to confirm a weakish connection in a graphical model. The approximated and the numerical values show a good agreement, even for $p = 10$. The approximate divergence decreases as $1/p$, which translates into a slope of -1 in the plot.

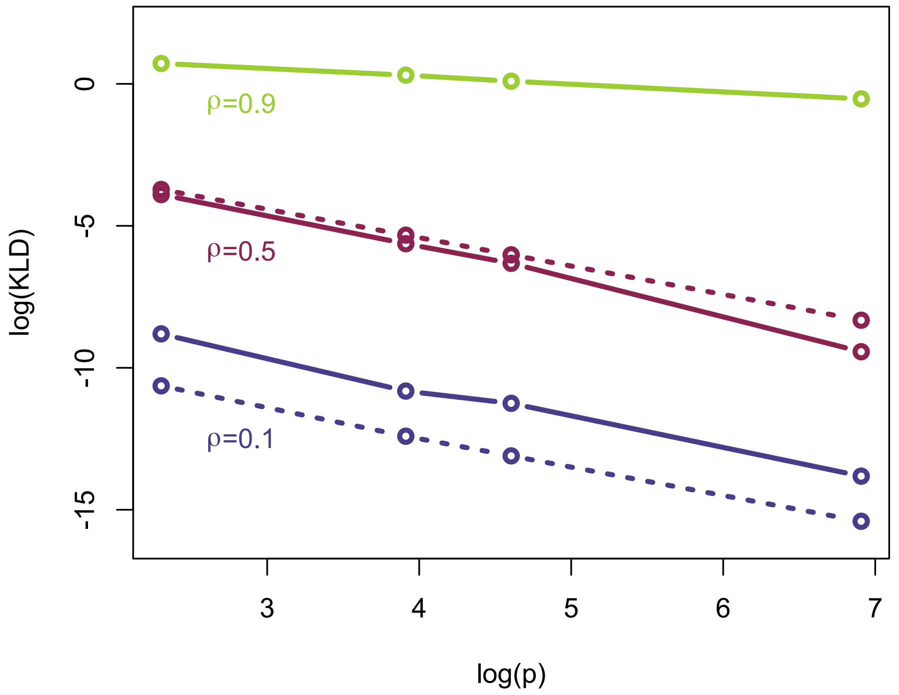

Figure 4.2.: The comparison of the numerical (solid) and the approximate (dotted) divergence as a function of a number of parameters p.

The case $\rho \geq 0.5$ requires more refined methods and is mainly left for future work. Figure 4.2 includes also the numerical results for $\rho = 0.9$. The plot makes it clear that in this case the information content decreases much less rapidly with increasing dimension p. The linearity in the plot of log(KLD) as a function of log(p) remains, but the slope passes from -1 to -0.25. In the end of this chapter we quickly reflect on the ratio between the sample size and the number of parameters, allowing for detection of the partial correlation.

The divergence value will grow to infinity for any fixed value of p as the sample size $n \to \infty$. When does this method undergo difficulties in detecting a partial correlation, i.e. when does the divergence remain small despite the large sample size n? A Table 4.2 below shows some simulation on that subject.

From the simulations it is clear that under when ρ is of order $O(n^{-1/2})$, there are

	$\rho = \frac{1}{\sqrt{n}}$		$\rho = \frac{1}{n}$	
n	$p = 10$	$p = 100$	$p = 10$	$p = 100$
100	0.002	0.0005	0.003	7.73e-06
500	0.044	0.002	0.0007	5.85e-05
1000	0.08	0.011	0.0003	4.53e-05
5000	0.18	0.016	0.0012	0.0002

Table 4.2.: Kullback-Leibler divergence when the partial correlation $\rho = \rho_n$ is a function of the sample size n.

almost no chances of discovering the partial correlation. The only exception might be the case with an extremely large sample and very few variables. This problem is even more evident when ρ is of order $O(n^{-1})$.

Table 4.3 gives a very rough estimate (based on simulations) for the necessary sample size in order to detect a single partial correlation of value ρ among p variables. We note that the presented ratio gives a lower bound for the ratio upon which the partial correlation remains undetected. The behavior of the method around this ratio strongly depends on the value of ρ. A high correlation of $\rho = 0.9$ is easily detected even with small samples, and only a large number of variables can mask it. For the average partial correlation ($\rho \approx 0.5$), this table suggests it can not be detected in high-dimensional settings.

	ρ = 0.9	ρ=0.5
$\frac{p}{n} \leq$	100	1

Table 4.3.: Critical ratio between p and n enabling detection of a single partial correlation.

4.3.3. Exact value for two and more non-zero partial correlations

The exact divergence between $g = \mathcal{N}_p(0, \Sigma)$ and $f = \mathcal{N}_p(0, I)$, when Σ^{-1} contains more than two non-zero off-diagonal elements, becomes more problematic to evaluate. As seen in the Formula (4.14), the determinant of Σ^{-1} needs to be computed. However, even if we assume all the non-zero entries to be of the same value ρ, the determinant of Σ^{-1}, and hence, the Kullback-Leibler divergence takes multiple values. The different values of the determinant depend on the positions of the non-zero entries in the matrix.

In the case of two partial correlations of equal value, there are two possible cases:

$$|\Sigma^{-1}| = \left[\begin{array}{l} 1-2\rho^2, \quad \text{if the partial correlations are in the same row or column,} \\ (1-\rho^2)^2, \quad \text{if the partial correlations are in different rows and columns.} \end{array}\right.$$

This transforms into the following values for the divergence:

$$D(g\,|\,f) = \left[\begin{array}{l} \frac{1}{2}\left(\log(1-2\rho^2) - \frac{4\rho^2}{2\rho^2-1}\right), \\ \frac{1}{2}\left(\log((1-\rho^2)^2) - \frac{4\rho^2}{\rho^2-1}\right). \end{array}\right.$$

For three partial correlations of equal size we have three solutions:

$$|\Sigma^{-1}| = \left[\begin{array}{l} 1-3\rho^2, \quad \text{if all three partial correlations are in the same row or column,} \\ 1-3\rho^2+2\rho^3, \quad \text{if two out of three correlations are in the same row or column,} \\ 1-3\rho^2+\rho^4, \quad \text{if all three of them are in different rows and columns.} \end{array}\right.$$

$$D(g\,,f) = \left[\begin{array}{l} \frac{1}{2}\left(\log(1-3\rho^2) - \frac{6\rho^2}{3\rho^2-1}\right), \\ \frac{1}{2}\left(\log(1-3\rho^2+2\rho^3) - \frac{6\rho^2}{2\rho^2-\rho-1}\right), \\ \frac{1}{2}\left(\log(1-3\rho^2+\rho^4) - \frac{4\rho^4-6\rho^2}{\rho^4-3\rho^2-1}\right). \end{array}\right.$$

These divergence values do not depend on the dimension p but they illustrate how quickly the difficulty of the problem increases. Distinguishing between a case with no dependencies and a case with a few of them is not straightforward as the number of the correlations grows. Nevertheless, we shall address the case of unknown placement of correlation in the case of two partial correlations of equal value.

4.3.4. Divergence in the case of two equal correlations with unknown placement

To write down the mixture density for two partial correlations (of the same value), we need to know the exact number of cases when both correlations are in the same line or column and otherwise. The total number of placements (in the upper off-diagonal) for two elements is $N = \binom{\binom{p}{2}}{2} = (p^4 - 2p^3 - p^2 + 2p)/8$. To determine the number of cases when two partial correlations are placed in the same row or column, we first explore the case of $p = 7$.

$$\mathbf{\Sigma^{-1}} = \begin{pmatrix} 1 & \rho & \rho & 0 & 0 & 0 & 0 \\ * & 1 & 0 & 0 & 0 & 0 & 0 \\ * & * & 1 & 0 & 0 & 0 & 0 \\ * & * & * & 1 & 0 & 0 & 0 \\ * & * & * & * & 1 & 0 & 0 \\ * & * & * & * & * & 1 & 0 \\ * & * & * & * & * & * & 1 \end{pmatrix}.$$

Here, for illustration purposes, we have placed the elements in the positions $(1,2)$ and $(1,3)$. What are the other options? If we fix the first element to be (1,2), then the possible cases within the same row for the second element are $(1,3), (1,4), (1,5), (1,6)$ and $(1,7)$. For all the cases when one of the correlations is in the first line (columns 2-7), there are 5 options to place the second one to be on the same row (or the column) with the first one. For the second row, there are 5 possible columns, each of them yielding 4 potential placements for the second correlation, and so on. In total, we get $6 \cdot 5 + 5 \cdot 4 + 4 \cdot 3 + 3 \cdot 2 + 2 \cdot 1 = 70$ possible placements. In the general case, for a dimension of p, this can be written as

$$N_1 = \#\{\text{same row/column}\} = \sum_{i=1}^{p-2} (p-i)(p-i-1) = \frac{1}{3}p^3 - p^2 + \frac{2}{3}p. \tag{4.21}$$

Equation (4.21) also gives us the number of placements of two correlations in different rows and columns:

$$N_2 = \#\{\text{different rows/columns}\} = N - N_1 = \frac{1}{8}p^4 - \frac{7}{12}p^3 + \frac{7}{8}p^2 - \frac{5}{12}p. \tag{4.22}$$

Since for two partial correlations there are two distinct cases of $|\Sigma^{-1}|$, our mixture density (4.16) becomes

$$g(x) = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \frac{1}{\binom{\binom{p}{2}}{2}} \left(\sum_{\substack{k=i+1 \\ k \neq j}}^{p} f_{ijk}(x) + \sum_{\substack{k=1 \\ k \neq i}}^{p-1} \sum_{\substack{l=k+1 \\ l \neq j}}^{p} f_{ijkl}(x) \right), \tag{4.23}$$

where f_{ijk} and f_{ijkl} are the normal densities with Σ^{-1} having two non-zero upper-diagonal elements in positions (i,j), (i,k) for f_{ijk}, and (i,j), (k,l) for f_{ijkl}. These densities have the following form:

$$f_{ijk} = \frac{\sqrt{1-2\rho^2}}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(x_1^2 + \ldots + x_p^2 + 2\rho x_i x_j + 2\rho x_i x_k)\right),$$

$$f_{ijkl} = \frac{1-\rho^2}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(x_1^2 + \ldots + x_p^2 + 2\rho x_i x_j + 2\rho x_k x_l)\right),$$

where the multipliers $\sqrt{1-2\rho^2}$ and $1-\rho^2$ are the square roots of the corresponding determinants of Σ^{-1}.

For a divergence $D(g \,|\, f)$, where f is a density $\mathcal{N}_p(0, I)$, as before, the ratio of the two densities becomes:

$$\frac{g(x)}{f(x)} = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \frac{1}{\binom{\binom{p}{2}}{2}} \left(\sqrt{1-2\rho^2} \sum_{\substack{k=i+1 \\ k \neq j}}^{p} \exp(-\rho x_i x_j - \rho x_i x_k) \right.$$

$$\left. + (1-\rho^2) \sum_{\substack{k=1 \\ k \neq i}}^{p-1} \sum_{\substack{l=k+1 \\ l \neq j}}^{p} \exp(-\rho x_i x_j - \rho x_k x_l) \right) = A + B. \tag{4.24}$$

It follows that

$$D(g \,|\, f) = \int_{\mathcal{R}^p} g(x) \log\left(\frac{g(x)}{f(x)}\right) dx = \int_{\mathcal{R}^p} f_{123}(x) \log\left(\frac{g(x)}{f(x)}\right) dx, \tag{4.25}$$

where without loss of generality we choose the density f_{123}, with the partial correlations in the positions (1,2), (1,3) or (1,2), (2,3), depending on the summand in (4.24).

We wish to employ the same strategy as in the case of a single partial correlation by approximating the integrand in (4.25) and using the central limit theorem to get the expected values of the summands in (4.24). To do this, we need the consider in more detail the terms arising in the sum and their behavior, as in (4.18) .

We start with the first summand in (4.24). Let $(X_1, \dots, X_p)$ be a random vector with density f_{123}. We write down the corresponding ratio of densities:

$$A = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \frac{1}{\binom{\binom{p}{2}}{2}} \left(\sqrt{1-2\rho^2} \sum_{\substack{k=i+1 \\ k \neq j}}^{p} e^{-\rho X_i X_j} e^{-\rho X_i X_k} \right) \tag{4.26}$$

$$= \frac{\sqrt{1-2\rho^2}}{\binom{\binom{p}{2}}{2}} \left(e^{-\rho X_1 X_2} e^{-\rho X_1 X_3} + e^{-\rho X_1 X_3} e^{-\rho X_2 X_3} + \sum_{i=\{2,3\}} \sum_{j=4}^{p} e^{-\rho X_1 X_i} e^{-\rho X_1 Z_j} \right.$$
$$+ \sum_{i=\{1,2\}} \sum_{\substack{j=\{2,3\} \\ j \neq i}} \sum_{k=4}^{p} e^{-\rho X_i Z_k} e^{-\rho X_j Z_k} + \sum_{i=4}^{p} e^{-\rho X_2 X_3} e^{-\rho X_2 Z_i} + \sum_{i=1}^{3} \sum_{j=4}^{p} \sum_{\substack{k=4 \\ k \neq j}}^{p} e^{-\rho X_i Z_j} e^{-\rho X_i Z_k}$$
$$\left. + \sum_{i=1}^{3} \sum_{j=4}^{p} \sum_{\substack{k=4 \\ k \neq j}}^{p} e^{-\rho X_i Z_j} e^{-\rho Z_k Z_j} + \sum_{i=4}^{p-1} \sum_{j=i+1}^{p} \sum_{\substack{k=i+1 \\ k \neq j}}^{p} e^{-\rho Z_i Z_j} e^{-\rho Z_i Z_k} + \sum_{i=4}^{p-1} \sum_{j=i+1}^{p} \sum_{\substack{k=4 \\ k \neq i}}^{p-1} e^{-\rho Z_i Z_j} e^{-\rho Z_k Z_j} \right).$$

In Equation (4.26) we explicitly write down all the arising pairs of correlations and we change the notations for X_i to Z_i since $X_i, (i = 4, \dots, p) \sim \mathcal{N}(0, 1)$.

For large p, the sums in (4.26) should be well-behaved and be approximated by their asymptotic limits. Before computing the expected values of these terms, we shall first estimate the number of terms in each of these summands. The last two terms in (4.26) correspond to the number of placements of two elements in the same row or column in the upper off-diagonal matrix of dimension $p-3$. Hence, we get $n_{13} = \sum_{i=1}^{p-5}(p-3-i)(p-4-i) = p^3/3 - 4p^2 + 47p/3 - 20$. We can show that

$$\mathrm{E}(\exp(-\rho Z_1 Z_2 - \rho Z_1 Z_3)) = \iiint \exp(-\rho Z_1 Z_2 - \rho Z_1 Z_3) \phi(z_1) \phi(z_2) \phi(z_3) dz_1 dz_2 dz_3$$

$$= \iint \frac{1}{(2\pi)^{3/2}} \exp\left(-\frac{z_1^2}{2} - \frac{z_2^2}{2} - \frac{z_3^2}{2} - \rho z_1 z_2 - \rho z_1 z_3\right) dz_1 dz_2 dz_3$$

$$= \left(\int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_1 + \rho(z_2 + z_3))\right)^2 dz_1\right)$$

$$\iint \frac{1}{2\pi} \exp\left(-\frac{1}{2}(z_2^2(1-\rho^2) + z_3^2(1-\rho^2) - 2\rho^2 z_2 z_3)\right) dz_2 dz_3$$

$$= \int \frac{1}{2\pi} \exp\left(-\frac{(1-\rho^2)}{2}\left(z_2 - \frac{\rho^2}{(1-\rho^2)} z_3\right)^2\right) dz_2$$

$$\int \frac{1}{2\pi} \exp\left(-\frac{1}{2}\frac{(1-2\rho^2)}{(1-\rho^2)} z_3^2\right) dz_3 = \frac{1}{\sqrt{1-\rho^2}} \frac{\sqrt{1-\rho^2}}{\sqrt{1-2\rho^2}} = \frac{1}{\sqrt{1-2\rho^2}}. \qquad (4.27)$$

The variance of this term is computed analogously and equals

$$\mathrm{Var}(\exp(-\rho Z_1 Z_2 - \rho Z_1 Z_3)) = \frac{1}{\sqrt{1-8\rho^2}},$$

which states that the transition of KLD in this case happens as $\rho > 1/\sqrt{8}$.

The terms in the sums $\sum_{i=1}^{3} \sum_{j=4}^{p} \sum_{\substack{k=4 \\ k\neq j}}^{p} e^{-\rho X_i Z_j} e^{-\rho X_i Z_k}$ and $\sum_{i=1}^{3} \sum_{j=4}^{p} \sum_{\substack{k=4 \\ k\neq j}}^{p} e^{-\rho X_i Z_j} e^{-\rho Z_k Z_j}$ need to be separated. The possible placements are $((X_1, Z_i), (X_1, Z_j))$, $((X_2, Z_i), (X_2, Z_j))$, $((X_3, Z_i), (X_3, Z_j))$, $((X_1, Z_i), (Z_k, Z_i))$, $((X_2, Z_i), (Z_k, Z_i))$ and $((X_3, Z_i), (Z_k, Z_i))$. Each of these combinations takes place in $n_{12} = \sum_{i=4}^{p-1}(p-i) = p^2/2 - 7p/2 + 6$ cases. As an example in $p = 7$ for the placement of the kind $((X_1, Z_i), (X_1, Z_j))$ we get $\{((1,4),(1,5))\ ((1,4),(1,6))\ ((1,4),(1,7))\ ((1,5),(1,6))\ ((1,5),(1,7)), ((1,6),(1,7))\}$, i.e. $3 + 2 + 1$ cases. The rest can be shown in a similar way.

The remaining random terms, when separated, present the following combinations: $\{(X_1, X_2), (X_1, Z_i))$, $((X_1, X_3), (X_1, Z_i))$, $((X_1, Z_i), (X_2, Z_i))$, $((X_1, Z_i), (X_3, Z_i))$, $((X_2, X_3), (X_2, Z_i))$, $((X_2, Z_i), (X_3, Z_i))\}$. Each of them appears in $n_{11} = (p-3)$ cases, for example $\{((1,2),(1,4))$, $((1,2),(1,5))$, $((1,2),(1,6))$, $((1,2),(1,7))\}$ for $(X_1, X_2)\,(X_1, Z_i)$ in a $p = 7$ case.

These coefficient sum up to N_1, the total number of placements in case of same row or column, in a following way:

$$N_1 = n_{11} + n_{12} + n_{13} + 2,$$

where 2 comes from the terms $e^{-\rho X_1 X_2} e^{-\rho X_1 X_3}$ and $e^{-\rho X_1 X_3} e^{-\rho X_2 X_3}$. In our sum (4.26), however, we divide them by the total number of placements, $N = (p^4 - 2p^3 - p^2 + 2p)/8$. Since $n_{11} = O(p^{-3})$ and $n_{12} = O(p^{-2})$, we only keep the

$$\frac{n_{13}}{N} = \frac{p^3/3 - 4p^2 + 47p/3 - 20}{(p^4 - 2p^3 - p^2 + 2p)/8} = \frac{8}{3p} + O(p^{-2}).$$

Finally, we appeal to the central limit theorem and get that

$$\mathrm{E}\left(\sum_{i=4}^{p-1} \sum_{j=i+1}^{p} \sum_{\substack{k=i+1 \\ k \neq j}}^{p} e^{-\rho Z_i Z_j} e^{-\rho Z_i Z_k} + \sum_{i=4}^{p-1} \sum_{j=i+1}^{p} \sum_{\substack{k=4 \\ k \neq i}}^{p-1} e^{-\rho Z_i Z_j} e^{-\rho Z_k Z_j} \right) = \frac{8}{3p} \frac{1}{\sqrt{1 - 2\rho^2}}.$$

And our first term A is approximated for large p by

$$A = \sqrt{1 - 2\rho^2} \frac{8}{3p} \frac{1}{\sqrt{1 - 2\rho^2}} = \frac{8}{3p}. \tag{4.28}$$

Now we consider the case when the partial correlations are placed in different rows and columns. There are $N_2 = N - N_1 = p^4/8 - 7p^3/12 + 7p^2/8 - 5p/12$ possibilities and the detailed breakdown of scenarios is more complicated than before. Let $(X_1, \ldots, X_p)$ be a random vector with density f_{123}, where Σ^{-1} has ρ in positions $(1,2)$ and $(2,3)$. The corresponding ratio of densities is

$$B = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \frac{1}{\binom{\binom{p}{2}}{2}} \left((1 - \rho^2) \sum_{\substack{k=1 \\ k \neq i}}^{p-1} \sum_{\substack{l=k+1 \\ l \neq j}}^{p} \exp(-\rho x_i x_j - \rho x_k x_l) \right). \tag{4.29}$$

The possible placements and the number of cases they present are listed in Table 4.4 below. While the coefficients n_{21}, n_{22} and n_{23} are rather easy to deduce, the coefficients n_{24} and n_{25} are computed as the total number of placements in respective scenarios minus the number of placements in the same row or column. We conclude that we are keeping only the coefficients n_{24} and n_{25} since after division by N they are of order $O(p^{-1})$ or higher. More precisely,

$$\frac{n_{24}}{N} = \frac{p^3/2 - 11p^2/2 + 20p - 24}{(p^4 - 2p^3 - p^2 + 2p)/8} = \frac{4}{p} + O(p^{-2}),$$

$$\frac{n_{25}}{N} = \frac{p^4/8 - 25p^3/12 + 39p^2/8 - 43p/12 + 35}{(p^4 - 2p^3 - p^2 + 2p)/8} = 1 + O(p^{-2}).$$

Possible combinations	Number of cases
$((X_1, X_2), (X_2, Z_i))$ $((X_1, X_2), (X_3, Z_i))$ $((X_1, X_3), (X_2, Z_i))$ $((X_1, X_3), (X_3, Z_i))$ $((X_1, Z_i), (X_2, X_3))$ $((X_2, X_3), (X_3, Z_i))$	$n_{21} = (p-3) = n_{11}$
$((X_1, X_2), (Z_i, Z_j))$ $((X_1, X_3), (Z_i, Z_j))$ $((X_2, X_3), (Z_i, Z_j))$	$n_{22} = \sum_{i=4}^{p}(p-i) = n_{12}$
$((X_1, Z_i), (X_2, Z_j))$ $((X_1, Z_i), (X_3, Z_j))$ $((X_2, Z_i), (X_3, Z_j))$	$n_{23} = (p-4)(p-3)$
$((X_1, Z_i), (Z_i, Z_k))$ $((X_2, Z_i), (Z_i, Z_k))$ $((X_3, Z_i), (Z_i, Z_k))$	$n_{24} = \frac{(p-4)(p-3)^2}{2} - n_{12} = p^3/2 - 11p^2/2 + 20p - 24$
$((Z_i, Z_j), (Z_k, Z_l))$	$n_{25} = \frac{(p-4)(p-3)}{4}\left(\frac{(p-4)(p-3)}{2} - 1\right) - n_{13}$

Table 4.4.: Possible scenarios and their counts for placing two partial correlations in upper off-diagonal matrix in different rows and columns.

The last thing to compute is the expected values of terms which enter Equation (4.29).

$$\mathrm{E}(e^{-\rho Z_i Z_j} e^{-\rho Z_k Z_l}) = \mathrm{E}(e^{-\rho Z_i Z_j})\mathrm{E}(e^{-\rho Z_k Z_l}) = \frac{1}{1-\rho^2}.$$

To compute $\mathrm{E}(e^{-\rho X_i Z_j} e^{-\rho Z_i Z_l})$, we need the marginal variance for X_1, X_2 and X_3.

Since their Σ^{-1} has ρ in positions $(1,2)$ and $(1,3)$, we get that

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_3) = \frac{1-\rho^2}{1-2\rho^2},$$

$$\mathrm{Var}(X_2) = \frac{1}{1-2\rho^2}.$$

The computation of the necessary expected values is straightforward and resembles (4.19). The results are:

$$\mathrm{E}(e^{-\rho X_1 Z_j}) = \mathrm{E}(e^{-\rho X_3 Z_j}) = \frac{\sqrt{1-2\rho^2}}{\sqrt{1-3\rho^2+\rho^4}},$$

$$\mathrm{E}(e^{-\rho X_3 Z_j}) = \frac{\sqrt{1-2\rho^2}}{\sqrt{1-3\rho^2}}.$$

By multiplying each of them by $\mathrm{E}(e^{-\rho Z_j Z_k}) = 1/\sqrt{1-\rho^2}$, we get the desired expected values.

Hence, the approximated term B becomes

$$\begin{aligned} B &= (1-\rho^2)\left(\frac{4}{p\sqrt{1-\rho^2}}\left(\frac{2\sqrt{1-2\rho^2}}{\sqrt{1-3\rho^2+\rho^4}} + \frac{\sqrt{1-2\rho^2}}{\sqrt{1-3\rho^2}}\right) + \frac{1}{1-\rho^2}\right) \\ &= \frac{4}{p}\sqrt{1-2\rho^2}\sqrt{1-\rho^2}\left(\frac{2}{\sqrt{1-3\rho^2+\rho^4}} + \frac{1}{\sqrt{1-3\rho^2}}\right) + 1. \end{aligned} \tag{4.30}$$

Finally, we can write down the approximate KLD value in case of two partial correlations of the same value by taking the logarithm of $(A+B)$

$$\begin{aligned} D(g\,|\,f) &= \log\left(1 + \frac{8}{3p} + \frac{4}{p}\sqrt{1-2\rho^2}\sqrt{1-\rho^2}\left(\frac{2}{\sqrt{1-3\rho^2+\rho^4}} + \frac{1}{\sqrt{1-3\rho^2}}\right)\right) \\ &\approx \frac{4}{p}\left(\frac{2}{3} + \sqrt{1-2\rho^2}\sqrt{1-\rho^2}\left(\frac{2}{\sqrt{1-3\rho^2+\rho^4}} + \frac{1}{\sqrt{1-3\rho^2}}\right)\right). \end{aligned} \tag{4.31}$$

Comparing (4.31) and (4.20), the KLDs for one and two partial correlations, we see that both values are of order $1/p$, making the detection rather difficult. Formula (4.31) is only valid for $\rho < 1/\sqrt{8} \approx 0.35$, while the KLD for a single partial correlation is valid for $\rho < 0.5$. While no theoretical justification can be made for the behavior on higher values of ρ, numerical simulation suggest that detection gets easier, as expected. In this case, this translates as two-correlation case will be detected easier (since the "bad" behavior stops at $\rho \approx 0.35$) than the single correlation. When comparing their values, the KLD for two correlations is much higher then the one for a single one, for example, 0.1 versus $2e-06$ for $\rho = 0.1$ and $p = 100$.

4.4. Comparison of local power and KLD

The analysis using the KLD is related, but different, from the asymptotic power computed in Section 4.2.1. When using the KLD, there is no correction for multiplicity involved, no constraints of the type $n > p$ are needed and no limits towards infinite study sizes are taken. The KLD thus provides a more solid foundation for judgeing the sample sizes needed in order to reliably detect effects.

A rough comparison can be based on the fact that in order to reach a power of about 0.5 at level α, the KLD of an experiment must exceed $z^2_{1-\alpha}$. From this, one can derive a formula for the needed size of a study,

$$n = pz^2_{1-\alpha}/[4((1-\rho^2)/\sqrt{1-2\rho^2}-1)].$$

The equivalent value of n from the asymptotic power on the other hand predicts that

$$n = (z_{1-\alpha/m}/\rho)^2,$$

where $m = p(p-1)/2$ is the number of tests. For values of $\rho < 0.5$, the KLD-based formula gives much higher values of the study size n. For example, around $n = 20,000$ subjects would be required to detect a partial correlation in a single pair of $p = 1000$ genes. The asymptotic power wrongly suggests that $n = 135$ subjects would be sufficient. Generally speaking, when $\rho < 0.5$, the situation is hopeless, unless the number of candidate genes that are tested can be reduced below $p = 100$. Figure 4.2 also gives an indication of what will happen for a strong effect, $\rho = 0.9$. The value of KLD decreases by about a factor of 0.24 for each increase of p by a factor of 10. If we extrapolate to $p = 10^6$, we have a KLD value of about 8×10^{-3}. We thus would need a study involving at least $n = 330$ subjects, which is doable. The qualitatively different behavior of the local power and the KLD confirms the observed overestimation of power for large p from Figure 4.1.

The divergence value will grow to infinity for any fixed value of p as the sample size $n \to \infty$. This means that a sufficiently big study will always detect a partial correlation of fixed size. In order to approximate high-dimensional cases ($n < p$), it is more interesting to study limits where ρ is fixed, but n and p both tend to ∞. For $\rho < 0.5$ we found that KLD $= O(p^{-1})$ from which it follows that the KLD will grow to infinity as long as $p/n \to 0$. When $\rho = 0.9$, our numerical values suggest that $p/n^{1.6} \to 0$ in order for the effect to be detectable. Thus, with $n = 1000$, we can hope to sift through a few tens of thousands of genes and detect strong linkage.

To sum up, it is easier to detect the structure for sparse graphs while the detection of multiple edges requires stronger linkages. We studied in detail the case of a single true alternative in dimension p, where $m = p(p-1)/2$ tests have to be performed. The local asymptotic power and the Kullback-Leibler divergence were used to assess the sample size needed for detection.

The KLD increases linearly in the number of samples n and we showed that it decreases inverse proportionally with p when the linkage is weak. In high-dimensional smoothing problems, it is usually found that p enters logarithmically, which is much more favorable. A transition phenomenon happens as the partial correlation increases beyond 0.5. We were, however, not able to give a rigorous description of the “strong effect” situation, although numerical integration suggests that a power law in p remains in effect.

We conclude that weak partial correlations require very large samples in order for a study to be able to detect them reliably. Our results are consistent with the findings of Arias-Castro et al. (2012), where the detection of correlation has been studied as a part of a structure of a high-dimensional vector. The authors show that, under certain conditions (value of the correlation relative to n and p), the correlations would not be detected.

CHAPTER

5 CONCLUSIONS

This thesis covers two topics. The first and the main result consists in proposing a new approach on how to integrate the censored observations arising in survival data into the sliced inverse regression procedure. This is done via a two-step approach. At first, all the censored observations are distributed with equal weights to consecutive slices of the data. Since the individual is censored, the event could have taken place anytime after the censoring appeared (we consider the slices to be small enough). When computing the slice means, we appeal to the weights of the censored points and apply them to the covariate information. This allows us to get the first SIR estimate (we only keep the principal direction). For better precision, we separate the cases of the accelerated lifetime and the proportional hazards model and suggest a method to recompute the weights for the covariates based on the supposed model. Our results prove to be competitive and in some cases better than other methods used for incorporating the censored data into the SIR.

When it comes to strategies to account for incomplete information (and censoring is a special case of such a situation), a popular idea is to reweight the observed data points, often with the help of the inverse probability weighting. This is the strategy of the other methods of SIR we have compared our results to. Our approach is inspired by the EM algorithm, by trying to estimate the unobserved events.

Since the asymptotic variance for our estimate is difficult to derive, we base the variance estimation of the bootstrap technique. When studying the likelihood approach for the accelerated lifetime models, we get the nice result that the Weibull regression is fully efficient (in terms of the maximum likelihood) in the absence of censoring.

In our algorithm, we devoted our attention to the case of the principal direction. The generalization to the several directions can be easily integrated in our application, while the testing for the dimensions is left for future work.

In the second part of the thesis (Chapter 4) we studied the factors that influence the power of the partial correlation test in detecting the structure of the Gaussian graphical models. The purpose of the study was to determine the limits between the sample size and the number of parameters one includes in the graph allowing for the detection. This was done based on two approaches, by deriving the local asymptotic power of the the partial correlation test and by considering the Kullback-Leibler divergence.

The main case for the study was the problem of detection of a single edge in the graph, which transforms for the partial correlation $\rho_{ij\cdot\text{rest}}$ in a single true alternative for the hypothesis $H_0 : \rho_{ij\cdot\text{rest}} = 0$.

While we do not elaborate this topic in detail, we found that it is easier to detect the structure for sparse graphs while the detection of multiple edges requires stronger linkages. This is not a new result, many penalized approaches these days aim for a sparse solution. Considering a single partial correlation in the concentration matrix allowed us to establish the local asymptotic power which we found to be overoptimistic in terms of the sample size value.

The Kullback-Leibler divergence (KLD) allows us to overcome the constraint $n > p$, set by the test of the partial correlation. While establishing its value, we uncovered the change in its behavior once the value of the correlation ρ passes the threshold of 0.5. This transition phenomenon allows for the asymptotic approximation only for the values of $\rho < 0.5$. In this case we uncovered that while the KLD increases linearly in the number of samples n, it also decreases inverse proportionally with p. That is, the ability to detect is proportional to n/p^{α}, when the linkage is weak. In high-dimensional smoothing problems, it is usually found that p enters logarithmically, in the $n/\log(p)$ kind, which is much more favorable for detection. While we were not able to give a rigorous description of the "strong effect" situation $\rho > 0.5$, our numerical results suggest that a power law in p remains in effect. The asymptotic approximation of the divergence for two partial correlations of the same value showed the same behavior, but the transition appears when $\rho \geq 1\sqrt{8} \approx 0.35$, expanding the region of an "easier" detection.

Our findings suggest that weak partial correlations require very large (larger, than usually anticipated) samples in order for a study to be able to detect them reliably. This results in big challenge for getting the necessary amount of data in genetical epidemiology and other biomedical studies, where samples rarely exceed a few hundred individuals.

APPENDIX

A

R FUNCTIONS

Here we list the main functions for our algorithm in R.

```
library(survival)
library(dr)

#-------------------------------------------------------------

GetWeightHazard<-function(i, j, a, survival_hat, effect)
{ # weight for the W matrix, i-th censored obs, j-th slice
w<-numeric(length = a$nslices)
if(j ==a$nslices){ #it’s in the last slice
   w[a$nslices]<-1
}else{
   for(k in (j+1):a$nslices){
      w[k]<-survival_hat[k]^effect - survival_hat[k+1]^effect
   }
   if(sum(w)==0){w[a$nslices]<-1}
}
return(w)
}
```

```
#---------------------------------------------------------------

GetPosition<-function(data_sorted, i, a)
{ # place the censored observation in the W matrix
   if (i < nrow(data_sorted)){   # next position isn't censored
      j<-1
      while(data_sorted[i+j, ncol(data_sorted)-1]==0 && (i+j) <
          nrow(data_sorted)){
         j<-j+1
      }
      position<-data_sorted[i+j, ncol(data_sorted)]
      if(position == 0){
         position<-a$nslices
      }
   }else{position<-a$nslices}
return(position)
}

#---------------------------------------------------------------

SIR_slicing<-function(NUM_Patients, NUM_Variables, a, x_mod, W,
    Sigma_sqrt_inv)
{ # steps 3-5 of the algorithm
p<-numeric(length = a$nslices)
for (i in 1:a$nslices){
   p[i]<-sum(W[,i])/NUM_Patients
}

m<-numeric(0)
for(i in 1:a$nslices)
{
   line<-numeric(length = NUM_Variables)
   for(j in 1:NUM_Patients)
   {
      line<-line+x_mod[j,]*W[j,i]
   }
   line<-line/(p[i]*NUM_Patients)
   m<-rbind(m, line)
   var(x_mod*W[,i])
}
m<-t(m)

#--- step 4: building a weighted covariance matrix
V<-matrix(nrow=NUM_Variables, ncol=NUM_Variables, data=0)
for (i in 1:a$nslices)
{
   V<-V+p[i]*m[,i]%*%t(m[,i])
}

#--- step 5: Eigenvalues of V
EV<-eigen(V)
beta<-EV$vectors%*%Sigma_sqrt_inv
return(beta[,1])
}

#--------------------------------------------------------------

VectorNorm<-function(data){ return(data/sum(data)) }

#--------------------------------------------------------------

BuildW<-function(NUM_Patients, NUM_Variables, a, data_sorted,
   censored_indices)
{ # assembling a weight matrix at step 2
W<-matrix(nrow=NUM_Patients, ncol=a$nslices, data=0)
indices<-which(data_sorted[, NUM_Variables+2]==1)
j<-1
for(i in indices){
   data_sorted[i, NUM_Variables+3]<-a$slice.indicator[j]
   j<-j+1
   W[i, data_sorted[i, ncol(data_sorted)]]<-1
}

for(i in censored_indices){
   position<-GetPosition(data_sorted, i, a)
   data_sorted[i, NUM_Variables+3]<-position
   for(j in position:a$nslices){
      W[i,j]<-1/(1+(a$nslices-position))
   }
}
return(W)
}

#-------------------------------------------------------------

SpherizeX<-function(NUM_Patients, NUM_Variables, X, Sigma_sqrt_
   inv)
{ # data normalization
X_mod<-matrix(nrow = nrow(X), ncol=ncol(X), dat=0)
X_mean<-apply(X, 2, mean)
for (i in 1:NUM_Variables){
   X_mod[,i]<-X[,i] - X_mean[i]
}
for(i in 1:(NUM_Patients)){
   X_mod[i,]<-Sigma_sqrt_inv%*%X_mod[i,]
}
return(X_mod)
}

#-------------------------------------------------------------

SortData<-function(NUM_Patients, NUM_Variables, X, T, event)
{ # preprocessing
data<-cbind(X,T, event)
data_sorted<-matrix(nrow = NUM_Patients, ncol = NUM_Variables
   +2, data=0)
index<-order(data[,NUM_Variables+1])
for (i in 1:NUM_Patients){
   data_sorted[i,]<-data[index[i],]
}
return(data_sorted)
}

#-------------------------------------------------------------
```

```
SingleRun_PH<-function(NUM_Patients, NUM_Variables, H, X, T,
   event, beta)
{ # PH version of SIR

#--- preprocess the data
data_sorted<-SortData(NUM_Patients, NUM_Variables, X, T, event)
y<-data_sorted[,NUM_Variables +1]
x<-data_sorted[,1:NUM_Variables]

censored_indices<-which(data_sorted[,NUM_Variables+2] == 0)
x_uncensored<-x
y_uncensored<-y
for (i in NUM_Patients:1){
   if (event[i] ==0){  #if censored
      x_uncensored<-x_uncensored[-i,]
      y_uncensored<-y_uncensored[-i]
   }
}

#--- step 1: standardization
Sigma<-cov(x)
s<-svd(Sigma)
d<-diag(sqrt(s$d))
Sigma_sqrt_inv<-s$v%*%solve(d)%*%t(s$u)
x_mod<-SpherizeX(NUM_Patients, NUM_Variables,x, Sigma_sqrt_inv)

#--- step 2: slicing
a<-dr.slices(y_uncensored, H)

#--- building the vector of slices indices in the data
slices<-numeric(length=NUM_Patients)
data_sorted<-cbind(data_sorted, slices)
W<-BuildW(NUM_Patients, NUM_Variables, a, data_sorted, censored
   _indices)
W<-round(W, 4)
W_old<-W

#--- SIR estimate with equal weights
```

```
beta_hat<-SIR_slicing(NUM_Patients, NUM_Variables, a, x_mod, W,
    Sigma_sqrt_inv)

eps<-1
while(eps > 0.01){

#--- start reweighting with Kaplan-Meier
if (length(y_uncensored)< NUM_Patients) {
slice_breaks<-numeric(0)
slice_breaks[1]<-min(y)
for (i in 2:a$nslices){
   slice_breaks[i]<-min(y_uncensored[a$slice.indicator==(i)])
}
slice_breaks[a$nslices+1]<-max(y)+0.01

temp<-summary(survfit(Surv(data_sorted[,NUM_Variables+1],
1-data_sorted[,NUM_Variables+2]) ~ 1, type="kaplan-meier"))
survival<-temp$surv

#--- adjusting the K-M curve to the slice ranges
survival_hat<-numeric(length=length(slice_breaks))
survival_hat[1:length(survival_hat)]<-1
survival_hat[length(survival_hat)]<-survival[length(survival)]
for(k in 2:(length(survival_hat)-1)){
   for(l in 1:(length(temp$time)-1)){
      if (slice_breaks[k] > temp$time[l] && slice_breaks[k] <=
         temp$time[l+1]){
         survival_hat[k]<-survival[l]
      }
      if(slice_breaks[k] > max(temp$time)){
         survival_hat[k]<-survival_hat[length(survival_hat)]
        }
    }
}

W[censored_indices,]<-0
for(i in censored_indices){
   effect<-as.double(exp(beta_hat%*%x_mod[i,]))
   for (j in 1:a$nslices){
      if(y[i] >= slice_breaks[j] && y[i] < slice_breaks[j+1]){
            W[i,]<-GetWeightHazard(i, j, a, survival_hat,
               effect)
   }
   }
   W[i,]<-VectorNorm(W[i,])   #rescale W[i,]
}
W[NUM_Patients,]<-0
W[NUM_Patients,a$nslices]<-1
W<-round(W, 4)
beta_coeffs<-SIR_slicing(NUM_Patients, NUM_Variables, a, x_mod,
    W, Sigma_sqrt_inv)
}
beta_hat<-beta_coeffs
dist.max<-0

for(u in censored_indices){
dist<-sqrt(sum((W[u,]-W_old[u,])^2))/length(censored_indices)
if (dist > dist.max){ dist.max<-dist}
}
W_old<-W
eps<-dist.max
#cat("maximal distance", round(eps, 6), "\n")
}
return(beta_coeffs)
}

#-------------------------------------------------------------

SingleRun_ALT<-function(NUM_Patients, NUM_Variables, H, X, T,
   event)
{# ALT version of SIR

data_sorted<-SortData(NUM_Patients, NUM_Variables, X, T, event)
y<-data_sorted[,NUM_Variables +1]
x<-data_sorted[,1:NUM_Variables]

censored_indices<-which(data_sorted[,NUM_Variables+2] == 0)
x_uncensored<-x
```

```
y_uncensored<-y
for (i in NUM_Patients:1){
   if (event[i] ==0){  #if censored
      x_uncensored<-x_uncensored[-i,]
      y_uncensored<-y_uncensored[-i]
   }
}

#### step 1: standardization
Sigma<-cov(x)
s<-svd(Sigma)
d<-sqrt(s$d)
d<-diag(d)
Sigma_sqrt_inv<-s$v%*%solve(d)%*%t(s$u)
x_mod<-SpherizeX(NUM_Patients, NUM_Variables, x, Sigma_sqrt_inv
   )

#### step 2: slicing
a<-dr.slices(y_uncensored, H)
#building the vector of slices indices in the data
slices<-numeric(length=NUM_Patients)
data_sorted<-cbind(data_sorted, slices)
W<-BuildW(NUM_Patients, NUM_Variables, a, data_sorted, censored
   _indices)
W<-round(W, 4)
W_old<-W
#--- equal weights
beta_hat<-SIR_slicing(NUM_Patients, NUM_Variables, a, x_mod, W,
    Sigma_sqrt_inv)

##### start convergence
eps<-1
while (eps > 0.01){

#--- reweighting
if (length(y_uncensored)< NUM_Patients) {
w<-numeric(0)
for (i in 1:length(y_uncensored)){
   w[i]<-y_uncensored[i] - beta_hat%*%x_uncensored[i,]
```

```
}

minT<-0
maxT<-0
for (i in censored_indices){
   if (min(w+ beta_hat%*%x_mod[i,]) <=minT){
      minT<-min(w+ beta_hat%*%x_mod[i,])
   }
   if (max(w+ beta_hat%*%x_mod[i,]) >=maxT){
      maxT<-max(w+ beta_hat%*%x_mod[i,])
   }
}

slice_breaks<-0
slice_breaks[1]<-min(y,minT)
for (i in 2:a$nslices){
   slice_breaks[i]<-min(y_uncensored[a$slice.indicator==(i)])
}
slice_breaks[a$nslices+1]<-max(y, maxT)+0.01

W[censored_indices, ]<-0

for(i in censored_indices){
   for (j in 1:(a$nslices+1)){
         if(y[i] == max(slice_breaks)){
            W[NUM_Patients,]<-0
            W[NUM_Patients,a$nslices]<-1
         }else{
         if (y[i] >= slice_breaks[j] && y[i]< slice_breaks[j
            +1]){  #-- Y_i* is in the j-th slice
            if (j==a$nslices){
               W[i,a$nslices]<-1
            }else{
               for(k in (j+1):a$nslices){# taking the weights
                  from the histogram density
                  W[i, k]<-hist(w+beta_hat%*%x_mod[i,], breaks
                     = slice_breaks)$density[k]
               }
               if (sum(W[i])==0){
                  W[i,j+1]<-1 #-- T_i* falls beyond the
                     histogram
               }
            }
         }
      }
   }
   W[i,]<-VectorNorm(W[i,])   #rescale W[i,]
}

beta_density<-SIR_slicing(NUM_Patients, NUM_Variables, a, x_mod
   , W, Sigma_sqrt_inv)

#---Kaplan-Meier
#-- transforming the timeline
for(i in 1:NUM_Patients){
data_sorted[i, NUM_Variables+1]<-data_sorted[i, NUM_Variables
   +1]*exp(-beta_hat%*%x_mod[i,])
}
temp<-summary(survfit(Surv(data_sorted[,NUM_Variables+1], 1-
   data_sorted[,NUM_Variables+2]) ~ 1, type="kaplan-meier"))
survival<-temp$surv

#--- adjusting the K-M curve to the slice ranges
survival_hat<-numeric(length=length(slice_breaks))
survival_hat[1:length(survival_hat)]<-1
survival_hat[length(survival_hat)]<-survival[length(survival)]
for(k in 2:(length(survival_hat)-1)){
   for(l in 1:(length(temp$time)-1)){
      if (slice_breaks[k] > temp$time[l] && slice_breaks[k] <=
         temp$time[l+1]){
         survival_hat[k]<-survival[l]
      }
   }
}

W[censored_indices,]<-0
for(i in censored_indices){
   for (j in 1:a$nslices){
      if(y[i] == max(slice_breaks)){ #the last observation is
         censored
         W[NUM_Patients,]<-0
         W[NUM_Patients,a$nslices]<-1
      }else{
      if(y[i] >= slice_breaks[j] && y[i] < slice_breaks[j+1]){
            W[i,]<-GetWeightHazard(i, j, a, survival_hat, 1)
       }
       }
   }
   W[i,]<-VectorNorm(W[i,])   #rescale W[i,]
}
W[NUM_Patients,]<-0
W[NUM_Patients,a$nslices]<-1
W<-round(W, 4)
beta_km<-SIR_slicing(NUM_Patients, NUM_Variables, a, x_mod, W,
   Sigma_sqrt_inv)
}

beta_hat<-beta_km
#beta_hat<-beta_density
dist.max<-0

for(u in censored_indices){
dist<-sqrt(sum((W[u,]-W_old[u,])^2))/length(censored_indices)
if (dist > dist.max){ dist.max<-dist}
}
W_old<-W
eps<-dist.max
#cat("maximal distance", round(eps, 6), "\n")
}
return(c(beta_density, beta_km))
}

#-------------------------------------------------------------

SIR_bootstrap<-function(T, X, event, H, method=c("alt", "cox"))
{ # the main function, return the coefficients and their
   standard deviations
```

```
NUM_Patients<-length(T)
NUM_Variables<-ncol(X)
if(length(T)!= length(event) || length(event)!= nrow(X))
   stop("dimensions do not match");
 if(NUM_Patients < NUM_Variables) stop("high-dimensional
    case!");
NUM_bootstrap<-100 # repetitions for the bootstrap

if (method == "alt"){
   beta<-SingleRun_ALT(NUM_Patients, NUM_Variables, H, X, T,
       event)
   beta_matrix<-matrix(nrow = NUM_bootstrap, ncol = NUM_
      Variables*2, data = 0)
   for(i in 1:NUM_bootstrap){
      indices<-sample(NUM_Patients, replace=TRUE)
      T_new<-T[indices]
      event_new<-event[indices]
      X_new<-X[indices,]
      beta_matrix[i,]<-SingleRun_ALT(NUM_Patients, NUM_
         Variables, H, X_new, T_new, event_new)
      if (sign(beta_matrix[i,1]) == -1){
            beta_matrix[i,]<-beta_matrix[i,]*(-1)
         }
      }
      beta_sd<-apply(beta_matrix, 2, sd)
      beta_coefficients<-cbind(beta[1:NUM_Variables], beta_
         sd[1:NUM_Variables], beta[(NUM_Variables+1):(2*NUM_
         Variables)], beta_sd[(NUM_Variables+1):(2*NUM_
         Variables)])
      colnames(beta_coefficients)<-c("beta_density", "sd_
         beta_density", "beta_km", "sd_beta_km")
}else{ #cox
   beta<-numeric(length = NUM_Variables)
   beta<-SingleRun_PH(NUM_Patients, NUM_Variables, H, X, T,
      event)
   beta_matrix<-matrix(nrow = NUM_bootstrap, ncol = NUM_
      Variables, data = 0)
   for(i in 1:NUM_bootstrap){
      indices<-sample(NUM_Patients, replace=TRUE)
```

```
         T_new<-T[indices]
         event_new<-event[indices]
         X_new<-X[indices,]
         beta_matrix[i,]<-SingleRun_PH(NUM_Patients, NUM_
            Variables, H, X_new, T_new, event_new)
         if (sign(beta_matrix[i,1]) == -1){
               beta_matrix[i,]<-beta_matrix[i,]*(-1)
            }
         }
      beta_sd<-apply(beta_matrix, 2, sd)
      beta_coefficients<-cbind(beta, beta_sd)
      colnames(beta_coefficients)<-c("beta_km", "sd_beta_km")
      }
return(beta_coefficients)
}

#------------------------------------------------------------
# Example: Cox-Weibull regression, 10 variables, 30 samples,
   25% censoring
beta<-c(1, -1, 0, sqrt(2), 1,0,0,0,0,0)/sqrt(5)
X<-rnorm(3000, 0, 2)
X<-matrix(X, ncol = 10)
T<-numeric(length = 300) #observed time
Y<-numeric(length = 300) #real time
C<-numeric(length = 300) #censored time
event<-numeric(length = 300) #  1 if observed, 0 if censored
lambda<-numeric(length = 300)
U<-runif(300)

for (i in 1:300){ # generating times

   lambda[i]<-X[i,]%*%beta
   Y[i]<-(-log(U[i])/exp(lambda[i]))^{1/4}*10 #Cox-Weibull
   C[i]<-runif(1,0,45) # 25% censoring

   if(C[i] > Y[i]){
      T[i]<-Y[i]
      event[i]<-1
   }else
```

```
      {
      T[i]<-C[i]
      event[i]<-0
   }
}

beta_hat<-SIR_bootstrap(T, X, event, 10, "cox") # results
```

BIBLIOGRAPHY

Aalen, O. (1975). *Statistical inference for a family of counting processes*. PhD thesis, Univ. of California, Berkeley.

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and Event History Analysis. A process point of view*. Springer.

Aalen, O. O., Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2009). History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, 5(1).

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Arias-Castro, E., Bubeck, S., and Lugosi, G. (2012). Detection of correlations. *The Annals of Statistics*, 40(1):412–435.

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, 18:265–271.

Becker, C. and Gather, U. (2007). A note on the choice of the number of slices in sliced inverse regression. Technical Report 11, Universität Dortmund.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards model. *Statistics in Medicine*, 24:1713–1723.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society*, 57:289–300.

Bercu, B., T.M.N.Nguyen, and Saracco, J. (2011). A new approach of recursive and non recursive sir methods. *Journal of the Korean Statistical Society*, 41:17–36.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66(3):429–436.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607.

Chen, C.-H. and Li, K.-C. (1998). Can sir be as popular as multiple linear regression? *Statistica Sinica*, 8:289–316.

Cook, D. R. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189.

Cook, D. R. (1998). Principal hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93(441):84–94.

Cook, D. R. (2000). Save: a method for dimension reduction and graphics in regression. *Communications in Statistics*, 29:2109–2121.

Cook, D. R. (2003). Dimension reduction and graphical exploration in regression including survival analysis. *Statistics in Medicine*, 2(9):1399–1413.

Cook, D. R. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *Journal of the American Statistical Association*, 100:410–428.

Cook, D. R. and Weisberg, S. (1991). Discussion of "sliced inverse regression" by k.c. li. *Journal of the American Statistical Association*, 86:328–332.

Coudret, R., Girard, S., and Saracco, J. (2012). A new sliced inverse regression method for multivariate response regression. Technical report, HAL.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of Royal*

Statistical Society, 34:187–200.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall.

Drton, M. and Perlman, M. D. (2004). Model selection for gausssian concentration graphs. *Biometrika*, 91(3):591–602.

Duan, N. and Li, K.-C. (1991). Slicing regression: A link-free regression method. *The Annals of Statistics*, 19(2):505–530.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer.

Efron, B. (1981). Censored data and bootstrap. *Journal of the American Statistical Association*, 76(374):312–319.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Haerdle, W. and Simar, L. (2007). *Applied Multivariate Statistical Analysis*. Springer.

Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21(2):867–889.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20(2):1040–1061.

Kalbfleisch, J. and Prentice, R. L. (1980). *The statistical analysis of failure time data*. Wiley.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Koziol, J. and Green, S. (1976). A cramér-von mises statistic for randomly censored data. *Biometrika*, 63(3):465–474.

Kraemer, N., Schaefer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(384).

Kreager, P. (1988). New light on graunt. *Population studies*, 42.

Kullback, S. (1997). *Information Theory and Statistics*. Dover Publications.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lee, S. and Wolfe, R. (1998). A simple test for independent censoring under the proportional hazards model. *Biometrics*, 54:1176–1182.

Li, K.-C. (1989). Data visualization with sir: a transformation based projection pursuit method. *UCLA statistical series*, 24.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87:1025–1039.

Li, K.-C. (2000). High dimensional data analysis via the sir/phd approach.

Li, K.-C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, 98(461):99–109.

Li, K.-C., Wang, J.-L., and Chen, C.-H. (1999). Dimension reduction for censored regression data. *The Annals of Statistics*, 27(1):1–23.

Li, L. (2006). Survival prediction of diffuse large-b-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, 22(4):466–471.

Li, L. (2010). Dimension reduction for high-dimensional data. In *Statistical Methods in Molecular Biology*, volume 620, pages 417–434. Springer Protocols.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, 20(18):3406–3412.

Li, L. and Lu, W. (2008). Sufficient dimension reduction with missing predictors. *Journal of the American Statistical Association*, 103(482):822–831.

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64(1):124–131.

Loh, P.-L. and Wainwright, M. (2012). Structure estimation for discrete graphical models: generalized covariance matrices and their inverses. *ArXiv e-prints*.

Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regressions. *Biometrics*, 67:513–523.

Lue, H.-H., Chen, C.-H., and Chang, W.-H. (2011). Dimension reduction in survival regression with censored data via an imputed spline approach. *Biometrical Journal*, 59(3):426–443.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170.

Meinshausen, N. and Buehlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, New

York.

Nadkarni, N. V., Zhao, Y., and Kosorok, M. (2011). Inverse regression estimation for censored data. *Journal of the American Statistical Association*, 106(493):178–190.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, 65:167–179.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics*, 18(1):303–328.

Rosenwald, A., Wright, G., Chan, W., Connors, J., and Campo, E. (2002). The use of molecualr profiling to predict survival after chemotherapy for diffuse large b-cell lymphoma. *New England Journal of Medicine*, 346:1937–1947.

Rotnitzky, A. and Robins, J. (2005). Inverse probability weighted estimation in survival analysis.

Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Communications in Statistics*, 26(9):2141–2171.

Scharfstein, D. and Robins, J. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634.

Scrucca, L. (2007). Class prediction and gene selection for dna microarrays using regularized sliced inverse regression. *Computational Statistics & Data Analysis*, 52(1):438–451.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

Tsiatis, A. (1981). A large sample study of cox's regression model. *The Annals of Statistics*, 9(1):93–108.

Tsiatis, A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18(1):354–372.

Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, 77(4):845–851.

Wen, X. and Cook, D. R. (2009). New approaches to model-free dimension reduction for bivariate regression. *Journal of Statistical Planning and Inference*, 139(3):734–748.

White, J. S. (1969). The moments of log-weibull order statistics. *Technometrics*, 11(2):373–386.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. WILEY, New Haven.

Witten, D. M. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19:29–51.

Wu, T., Sun, W., Yuan, S., Chen, C.-H., and Li, K.-C. (2008). A method for analyzing censored survival phenotype with gene expression data. *BMC Bioinformatics*, 9(417).

Yan, C. and Zhang, D. (2012). Sparse dimension reduction for censored data. *Computational Statistics*, pages 1–18.

Zhong, W., Zeng, P., Ma, P., Liu, J., and Zhu, Y. (2005). Rsir: regularized sliced inverse regression for motif discovery. *Bioinformatic*, 21:4169–4175.

Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):630–643.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, 5:727–736.

Curriculum Vitae

Maya Shevlyakova was born in Grodno, Belarus in September of 1983. In 2000 she graduated from the Classical Gymnasium 610 of Saint-Petersburg, Russia and went on with the Bachelor and Master studies in applied mathematics and informatics at the Saint-Petersburg Polytechnical University. Wanting to pursue her studies in biostatistics, she moved to Switzerland in 2006 where she completed Master programme in applied mathematics at Ecole Polytechnique Fédérale de Lausanne (EPFL). Between the fall of 2008 and the spring of 2013, she worked on her PhD thesis in biostatistics in EPFL.

Before EPFL and its teaching activities, Maya has volunteered for a number of years at the NGO for the HIV-positive people. Between her Master and PhD programme, she took up an internship at the Global Fund to fight AIDS, tuberculosis and malaria in Geneva.

Beside her mother tongue Russian, Maya is fluent in English and French and is rather proficient in German. She also has a basic understanding of Italian.

In her free time, Maya can be found cycling around Switzerland and not so-nearby countries. She also appreciates good food, mountain getaways and is a triathlon amateur. During her EPFL studies, she took part in organizing ski and climbing weekends for PhD and Erasmus students.

Maya can be reached at maya.shevlyakova@gmail.com.