NOVEL COMPUTATIONAL METHODS FOR CENSORED DATA AND REGRESSION

DISSERTATION

A dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of
Philosophy in the College of Arts and Sciences
at the University of Kentucky

By
Yifan Yang
Lexington, Kentucky
Director: Dr. Mai Zhou, Professor of Statistics
Lexington, Kentucky
2017

ABSTRACT OF DISSERTATION

NOVEL COMPUTATIONAL METHODS FOR CENSORED DATA AND REGRESSION

This dissertation can be divided into three topics. In the first topic, we derived a recursive algorithm for the constrained Kaplan-Meier estimator, which promotes the computation speed up to fifty times compared to the current method that uses EM algorithm. We also showed how this leads to the vast improvement of empirical likelihood analysis with right censored data. After a brief review of regularized regressions, we investigated the computational problems in the parametric/non-parametric hybrid accelerated failure time models and its regularization in a high dimensional setting. . We also illustrated that, when the number of pieces increases, the discussed models are close to a nonparametric one. In the last topic, we discussed a semi-parametric approach of hypothesis testing problem in the binary choice model. The major tools used are Buckley-James like algorithm and empirical likelihood. The essential idea, which is similar to the first topic, is iteratively computing linear constrained empirical likelihood using optimization algorithms including EM, and iterative convex minorant algorithm.

KEYWORDS: AFT model, Binary Choice Model, Empirical Likelihood

Author's signature: Yifan Yang

Date: January 2, 2017

NOVEL COMPUTATIONAL METHODS FOR CENSORED DATA AND REGRESSION

By
Yifan Yang

Director of Dissertation: Dr. Mai Zhou

Director of Graduate Studies: DGS name here

Date: January 2, 2017

To my friend Siyuan, and my parents.

ACKNOWLEDGMENTS

I would like to gratefully acknowledge several people who have been journeyed with me in the past five years.

First, I would like to thank my advisor, Dr Mai Zhou, who not only introduces me to the survival analysis world, but also provides and discusses with me enormous amount of deep insight of statistical and scientifically research.

Second, I would like to express my thanks to the committee members, Dr Arnold Stromberg, Dr Yanbing Zheng, Dr William Griffith, Dr Li Chen, and Dr Benjamin Braun, for their valuable suggestions and comments, and great patience.

In addition to the technical and instrumental assistance above, I received equally important assistance from family and friends. My parents Jianpin Yang and Aihua Zhang support me to chase the degree greatly. My friends, Dr Xiang Zhang, Dr Rui Tuo, Dr Lijie Wan, Dr Jin Hu, Dr Chen Chu, Dr Xiaoqin Pan, Dr Han Zhang, Dr Yue Ding, MD Yuan Zhou, and Siyuan Qian are always open to discuss and be with me since the beginning of this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

This thesis contains three relatively independent topics. All three are related to the novel computational methods for censored data and regression in survival analysis.

The second chapter develops a recursive algorithm to compute the Kaplan-Meier estimator fast with given mean constraints. Examples of such constraints are mean or median of the survival time, the cumulative probability at given time points, and so on. The direct Newton-Raphson optimization or the EM algorithm (Zhou, 2012) could also solve this problem, but either occupies too many memory spaces or converge slowly. We represented the Newton-Raphson approach and proposed a recursive algorithm to solve such problem fast, which could later be applied to solve different hypothesis testing problems on survival time or coefficients of survival regressions, in applications connects to empirical likelihood.

In Chapter 3, we considered high dimensional regression problem with right censored data. In particular, we investigated parametric accelerated failure time model with high-dimensional settings and illustrated the properties and performance of the proposed algorithm. It unifies the penalized regression method and the classical accelerated failure time model. There are some studies on non-parametric accelerated failure time model in high-dimensional setting now, but they are hard to use and lack rigorous proof. The parametric method and theory have a potential to contain more nuisance parameters and become more flexible.

In Chapter 4, we concentrated on the hypothesis testing problem for the so-called binary choice model. It uses a Buckley-James like method combined with EM algorithm. We maximized the empirical likelihood and derived the log-likelihood ratio statistics to solve the problem. Several algorithms and approaches could be replaced in each step of the proposed algorithm, which makes it flexible, and extendable.

The following flow chart covers the topics in each chapter and the relationship among them.

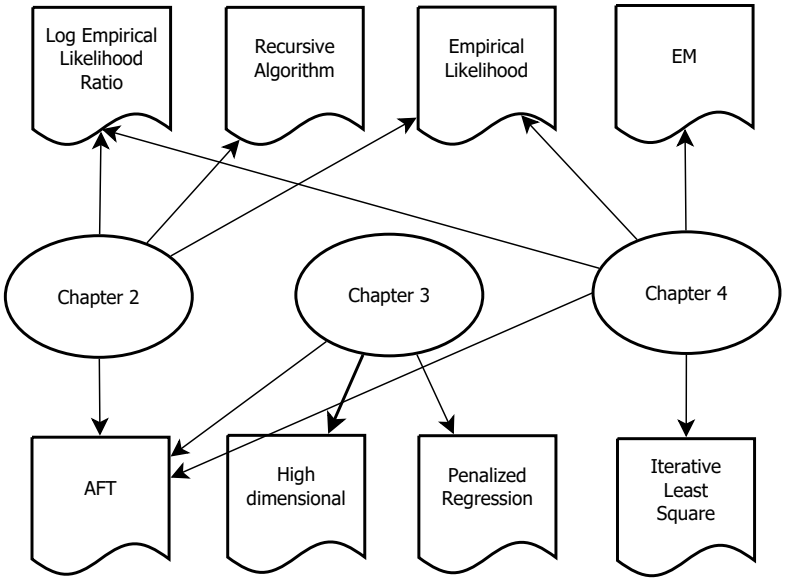


Figure 1.1: Topics in Chapter 2, Chapter 3 and Chapter 4.

Notation

$\arg\min_\beta f(\beta)$	the β that minimizes $f(.)$
$\arg\max_\beta f(\beta)$	the β that maximizes $f(.)$
$\mathbb{R}, \mathbb{C}, \mathbb{I}$	number fields and sets
$\|x\|_{\ell_p}$	$L-p$ norm of x if $p > 0$.
$\|x\|_{\ell_0}$	number of non-zero entries of x.
$\rightsquigarrow$, $\xrightarrow{d}$	convergence in distribution
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{a.s.}$	almost sure convergence
$N(\mu, \sigma^2)$, t_{df}, χ^2_{df}	normal, t and χ square distribution
Z_α	upper α quantile of normal distribution
$\triangleq$	define
$I_{\text{condition}}$	identification function
$\nabla f(x)$	first derivative of $f(x)$
$\Delta f(x)$	second derivative of $f(x)$, might be a matrix
$Q << P$	measure Q is absolutely continuous with respect to measure P
$o_p(1)$, $O_p(1)$	stochastic order symbols
$\mathbb{1}_p$	length p vector with all entries equal 1.
$\mathbb{P}_n$	empirical measure and process, e.g. $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$.

Chapter 2 A Recursive Formula for the Kaplan-Meier Estimator with Mean Constraints and Its Application to Empirical likelihood

2.1 Introduction

One essential problem in most survival analysis is to estimate the cumulative distribution function (CDF). Among all of the parametric, semi-parametric and non-parametric approaches, Kaplan-Meier estimator is the most famous one. The very first paper is (Bohmer 1912), but is not discovered and further studied by researchers until a major event of survival analysis came: the 1958 Kaplan, Meier paper was published.

In the June 1958 paper, Edward Kaplan and Paul Meier proposed a very important method to estimate and visualize incomplete survival observations: the Kaplan-Meier curve. The importance of this curve is highly appreciated not only in academia world but also in the medical researches and other fields. When Meier died, the news praised his achieve, said "that can affect the lives of millions", and "revolutionized medical trial." The original 1958 paper is the most cited statistical paper (ranks 11th) among all scientific fields of all time.

The plot of the Kaplan-Meier curve is now a standard approach to "depict time-to-events data for events like death or recurrence of disease", and it can show effects of treatment on major events of survival over time. Hence is now a must during clinical trial studies. Since it is a fundamental step to incomplete data study, most statistical software, either commercial (such SAS/MATLAB) or free (such as R/Python), provides several of methods to calculate and illustrate the Kaplan-Meier curve of data. But another side of the problem does not capture the same attention: hypothesis testing of Kaplan-Meier curve, which studies the propriety of random errors and further identifies/measures the uncertainty of the survival function estimation. Typically, there are two approaches to measure the uncertainty: local approach and the global approach.

For the first one, a so-called Greenwood formula will provide an estimator of the variance of the Kaplan-Meier estimator $\hat{S}(t)$ at any single fixed time point t.
Suppose that $X_1, \ldots, X_n$ are i.i.d. non-negative random variables denoting the life-times with a continuous distribution function F_0. Independent of the lifetimes there are censoring time $C_1, \ldots, C_n$ that are i.i.d. with a distribution G_0. Only the censored observations (T_i, δ_i)'s are available to us, where $T_i = min(T_i, C_i)$ and $\delta_i = I(X_i \leq C_i)$ for all i. Here $I(A)$ is the indicator function of A. Assume $0 = T_{t_1} < \ldots < T_{t_N}$, and for given time t, N_t be the number of observations that are still alive just before time t (denote as $t-$), and M_t be the number who survive from $t-$ to $t+$, i.e. $M_t = N_t - d_t$. Here d_t denotes the number of deaths that occur at time t. Then, the Kaplan-Meier estimator is :

$$\hat{S}(t) = \prod_{j:T_{t_j} \leq t} \frac{M_{t_j}}{N_{t_j}} \quad .$$

Besides, the Greenwood formula gives the variance at time point t:

$$\mathrm{Var}(\hat{S}(t)) \approx \hat{S}^2(t) \sum_{j:T_{t_j} \leq t} \frac{1 - \frac{M_{t_j}}{N_{t_j}}}{M_{t_j}} \quad .$$

More details could be found in any standard survival textbook, for example (Zhou, 2015). The Greenwood formula an also provide the Wald-type confidence interval:

$$(\hat{S}(t) - Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{S}(t))},\ \hat{S}(t) + Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{S}(t))}),$$

where $Z_{1-\alpha/2}$ is Normal distribution quantiles once the confidence level α is specified. Through the calculation, the drawback is clear:

1. Although Greenwood formula does a fine job in variance estimating, it has potential problem to give confidence interval when the distribution of $\hat{S}(t)$ is skewed;

2. It only works for a single point, in other words, it can not estimate the covariate

of Kaplan-Meier $\hat{S}(t)$ at any two distinguished time points.

There are papers on skewness correction (refs), which applies monotonic transformation function ($A(.)$) such as log (default transformation used in R) and log-log (default transformation used in SAS). It applies Delta-method again to calculate the confidence interval:

$$A^{-1}(A(\hat{S}(t)) - \alpha \text{ sd}(A(\hat{S}(t)))), A^{-1}(A(\hat{S}(t)) + \alpha \text{ sd}(A(\hat{S}(t)))).$$

But since each choice of $A(.)$ derive a different confidence interval, theoretically there are infinite many possible choices and no way to distinguish which is (are) the best. Besides, some authors also reported convergence speed of transformations may vary. For example, the Log-Log transformation outperforms the log transformation in this sense. Another comment is that the skewness of the Kaplan-Meier is also location dependent: $\hat{S}(t)$ may be quite skewed at t, but $\hat{S}(s)$ may have almost no skew at s. Therefore, we may need different transformation function for different locations s and t. So the asymptotical Wald-type confidence interval may have different forms at different time points.

The second problem is that the Greenwood formula does not provide the covariance of the Kaplan-Meier at two or more locations. Therefore, any quantity that depends on the Kaplan-Meier values at more than one places, the Greenwood falls short. A case in point is the mean value based on the Kaplan-Meier curve (trimmed mean or restricted mean are similar); see our examples in the method section for the restricted mean.

In the first section of this thesis will discuss the hypothesis testing problem of the Kaplan-Meier estimation. We advocate a new way of producing the confidence intervals that avoids the above two difficulties. This new way of producing confidence intervals is called empirical likelihood method and the theory was discussed in Owen and Zhou. The new method depends on the quick computation of *constrained* or *tilted* Kaplan-Meier Curve (Pan and Zhou, 1999).

2.2 Empirical Likelihood

The empirical likelihood (EL) of the censored data in terms of distribution F is defined as

$$\begin{aligned} EL(F) &= \prod_{i=1}^{n} [\Delta F(T_i))]^{\delta_i} [1 - F(T_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [\Delta F(T_i))]^{\delta_i} \{ \sum_{j:T_j > T_i} \Delta F(T_j) \}^{1-\delta_i} \end{aligned}$$

where $\Delta F(t) = F(t+) - F(t-)$ is the jump of F at t. See for example Kaplan and Meier (1958) and Owen (2001). The second line above assumes a discrete $F(\cdot)$. It is well known that the constrained or unconstrained maximum of the empirical likelihood are both obtained by discrete F (Zhou, 2005). Let $w_i = \Delta F(T_i)$ for $i = 1, 2, \ldots, n$. The likelihood at this F can be written in term of the jumps

$$EL = \prod_{i=1}^{n} [w_i]^{\delta_i} \{ \sum_{j=1}^{n} w_j I[T_j > T_i] \}^{1-\delta_i} ,$$

and the log likelihood is

$$\log EL = \sum_{i=1}^{n} \left\{ \delta_i \log w_i + (1 - \delta_i) \log \sum_{j=1}^{n} w_j I[T_j > T_i] \right\} . \tag{2.1}$$

If we maximize the log EL above without extra constraint (the probability constraints $w_i \geq 0$, and $\sum w_i = 1$ are always imposed), it is well known (Kaplan and Meier, 1958) that the Kaplan-Meier estimator $w_i = \Delta \hat{F}_{KM}(T_i)$ will achieve the maximum value of the log EL (Kaplan and Meier, 1958).

Definition 2.2.1. The empirical likelihood ratio statistics with uncensored data ((Thomas and Grunkemeier, 1975, Owen, 1988))

The empirical likelihood ratio statistics was proposed by Owen in a nonparametric

version of the well known Wilks theorem (1938). It is defined as:

$$\mathrm{ELR} = \frac{\mathrm{EL}_{H_0}}{\mathrm{EL}_{H_0 \cup H_1}} = \frac{\mathrm{EL}(\hat{F})}{\mathrm{EL}[\tilde{F}]}$$

Here $\hat{F}$ is the cumulative distribution function that maximizes the empirical likelihood under the null hypothesis H_0, and $\tilde{F}$ is the cumulative distribution function that maximizes the empirical likelihood under the hypothesis $H_0 \cup H_1$. (Owen, 1988) shows $-2 \log \mathrm{ELR}$ converges to χ^2 distribution under the linear type null hypothesis: $\int g(t) dF(t) = 0$ *when there is no censoring.*

Empirical likelihood ratio method was first proposed by Thomas and Grunkemeier (1975) in the context of a Kaplan-Meier estimator. This method has been studied by Owen (1988, 2001), Li (1995), Murphy and van der Vaart (1997) and Pan and Zhou (1999) among many others. When using the empirical likelihood with right censored data in testing a general hypothesis, Zhou (2005) gave an EM algorithm to compute the likelihood ratio. This paper also compared the EM algorithm with the sequential quadratic programming method (Chen and Zhou, 2007), and concluded that the EM was better. Though quite stable, the EM can be slow in certain data settings. See examples in the simulation section. We shall give a new recursive computation procedure for the constrained maximum of the log empirical likelihood above, which leads to a much faster computation algorithm of the empirical likelihood ratio for testing later.

The remain part of this chapter is organized as following: section 2.3 contains the derivation of the recursive algorithm, as well as its application to the empirical likelihood test; section 2.4 discusses the application of our novel algorithm on classical hypothesis testing problem in accelerated failure time model; section 2.5 reports the simulation results and a real data example. Finally we end the chapter with a discussion of further issues.

2.3 Method

In order to compute the empirical likelihood ratio, we need two empirical likelihoods: one with constraints, one without. The maximum of the empirical likelihood without constraint is achieved by F equals to the Kaplan-Meier estimator, as is well known. It remains to find the maximum of $\log EL$ under constraints. In this section, we first illustrate the recursive algorithm and then further discuss optimization and initial value problems.

The mean constrained Kaplan-Meier estimator

Using an argument similar to those in Owen (1988), we can show that we may restrict our attention in the EL analysis, i.e. search max under constrains, to those discrete CDF F that are dominated by the Kaplan-Meier: $F(t) \ll \hat{F}_{KM}(t)$. Owen (1988) restricted his attention to those distribution functions that are dominated by the empirical distribution.

The first step in our analysis is to find a discrete CDF that maximizes the log $EL(F)$ under the mean constraints, which are specified as follows:

$$\begin{aligned} \int_0^\infty g_1(t)\,\mathrm{d}F(t) &= \mu_1 \\ \int_0^\infty g_2(t)\,\mathrm{d}F(t) &= \mu_2 \\ \cdots \quad &\cdots \quad \cdots \\ \int_0^\infty g_p(t)\,\mathrm{d}F(t) &= \mu_p \end{aligned} \tag{2.2}$$

where $g_i(t)(i = 1, 2, \ldots, p)$ are given functions with finite second order moment, and μ_i $(i = 1, 2, \ldots, p)$ are given constants. Without loss of generality, we shall assume all $\mu_i = 0$. One examples of such constraints are shown in Figure 2.1, which present survival curve under certain hypothesis, i.e. $H_0 : E[X] = \mu_0$ or $H_0 : S(t_0) = s_0$. Examples of this "mean type" constraint are:

1 Sample mean $g(t) = t$;

2 Restricted mean $g(t) = tI(t \leq \tau)$;

3 Median $g(t) = I[t \leq m]$ and the constraint $\int [t \leq m] dF = 0.5$ defines implicitly the median m;

4 Survival probability at τ leads to $g(t) = I[t > \tau]$.

5 The difference or ratio of the above statistics. In this case, the two statistics can all be treated as two such linear functions after introduce one of them as the nuisance parameter.

6 Residual mean, and Residual median. In this scenario, we consider the CDF of the residual term in survival regression such as the accelerated failure time model. Then the residual mean and residual median can be considered as the "mean" type constraints,

The constraints (2.2) can be written as (for discrete CDFs with all $\mu_0 = 0$, and in terms of $w_i = \Delta F(T_i) = F(T_i) - F(T_i-)$)

$$\begin{aligned} \sum_{i=1}^{n} g_1(T_i) w_i &= 0 \\ \cdots \quad \cdots \quad \cdots & \\ \sum_{i=1}^{n} g_p(T_i) w_i &= 0 \ . \end{aligned} \tag{2.3}$$

We must find the maximum of the $\log EL(F)$ under these constraints. We shall use the Lagrange multiplier to find this constrained maximum.

Kaplan-Meier-Constraint algorithm

Since $F(t) \ll \hat{F}_{KM}(t)$, w_i is only positive at the uncensored observation T_i, except may be the last observation. Without loss of generality assume the observations are already ordered according to T and the smallest observation is an uncensored one ($\delta_1 = 1$). To see this, suppose the first observation is right censored and second one

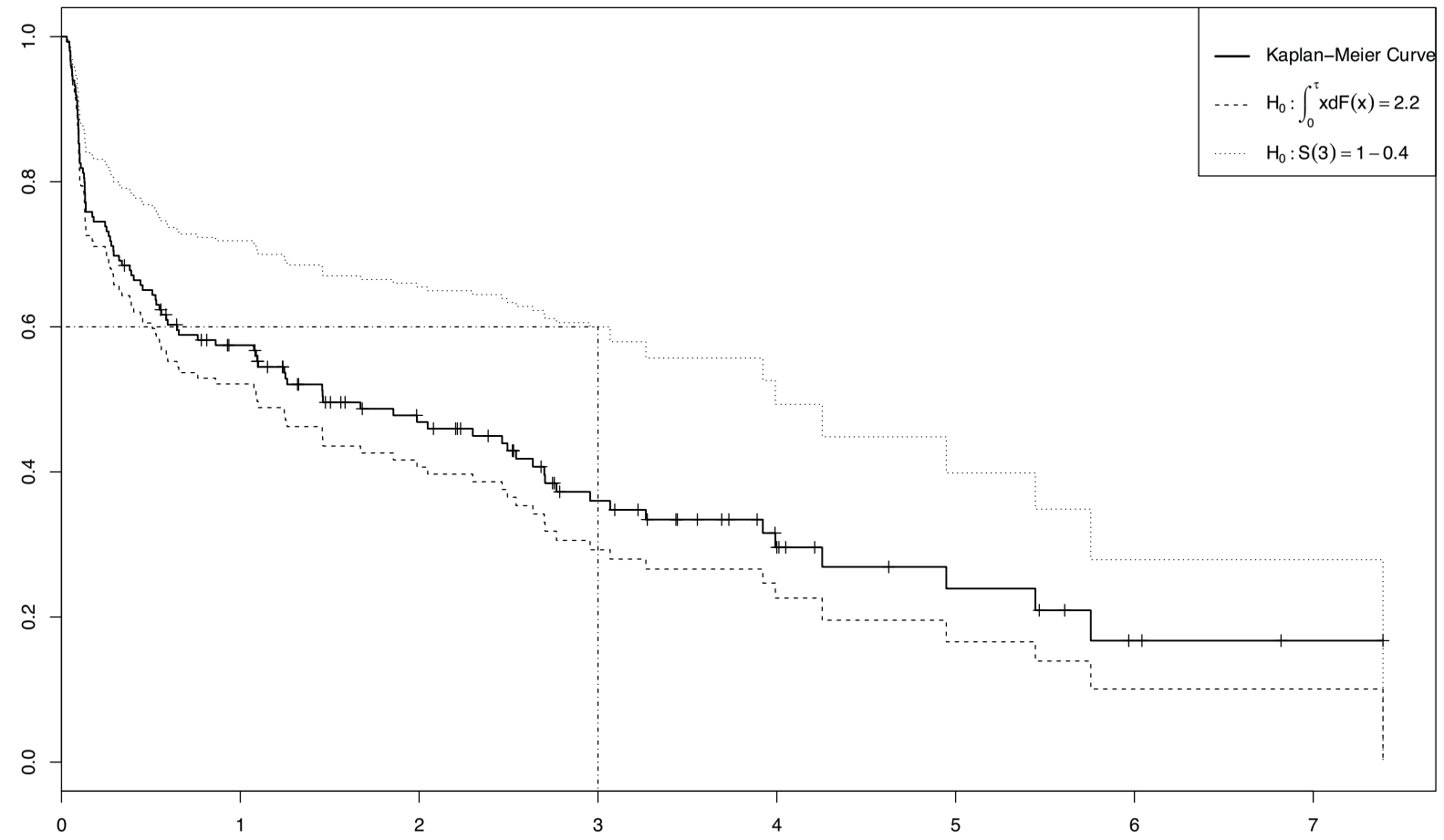


Figure 2.1: Kaplan-Meier plot with and without mean constraint for *Stanford 5* data discussed in Section **??**, time scaled by *500*. The two mean-type constraints are 1. - - - Mean survival time is 2.2, 2. ⋯ Survival probability at time 3 is 60%

is uncensored. In this case, $\delta_1 = 0$, and $w_1 = 0$. Hence

$$w_1 = 0 \quad , \quad \sum_{i=2}^{n} w_i = 1, \text{ and } \delta_1 \log w_1 \equiv 0 \ . \tag{2.4}$$

The i-th term in log empirical likelihood is

$$\delta_i \log w_i + (1 - \delta_i) \log \sum_{j=i+1}^{n} w_j \ .$$

This is true as observations are sorted according to T. Since $\delta_1 \log w_1 \equiv 0$, the log empirical likelihood only depends on $w_2, \ldots, w_n$. Additionally, the first observation with $\delta = 0$ has no contribution to the constraints. Therefore, we may focus on $w_2, , w_n$, with a positive w_2.

Assume $T_1 < \ldots < T_n$. Let $\mathcal{I} = \{i_1, \ldots, i_k\}$ be the index set of censored observations among the n's such that $T_{i_1} < \cdots < T_{i_k}$, k is the number of elements in $\mathcal{I}$. Thus we have only $n - k$ positive probability w_i's.

Introduce k new variables $\{\tilde{S}_1, \ldots, \tilde{S}_k\}$, one for each censored T observation, i.e. assume $j \in \mathcal{I}$, i.e. $\delta_{i_j} = 0$, and let:

$$\tilde{S}_j = \sum_{i:T_i > T_{i_j}} w_i = 1 - \sum_{i:T_i \leq T_{i_j}} w_i \ . \tag{2.5}$$

This adds k new constraints to the optimization problem. We write the vector of those k constraints as $\tilde{S}_j - \sum_{i:T_i > T_{i_j}} w_i = 0$. With these k new variables $\tilde{S}_i$, the log empirical likelihood in section 2.2 can be written simply as:

$$\log \mathrm{EL}\ (w, \tilde{S}) = \sum_{i=1, \delta_i = 1}^{n-k} \log w_i + \sum_{j=1, \delta_j = 0}^{k} \log \tilde{S}_j \ . \tag{2.6}$$

The Lagrangian function for constrained maximum is

$$G = \log EL(w, \tilde{S}) + \lambda^\top \left(\sum_{\delta_i = 1} \delta_i w_i g(T_i) \right) - \eta \left(\sum_{i=1}^{n} w_i - 1 \right) - \gamma^\top \left(\tilde{S} - W \right).$$

Here, $\lambda \in \mathbb{R}^p$, $g(T_i) = (g_1(T_i), \ldots, g_p(T_i))^\top$ is a vector corresponding to the constraints (2.3); $\eta \in \mathbb{R}$ is a scalar; $\gamma, \tilde{S}, W \in \mathbb{R}^k$, $\tilde{S}$ is a vector for those $\tilde{S}_j$'s defined in (2.5), and the j-th entry of W is $\sum_{i:T_i > T_{i_j}} w_i$.

Next we shall take the partial derivatives and set them to zero. We shall show that $\eta = n$.

First we compute

$$\frac{\partial G}{\partial \tilde{S}_j} = \frac{1-\delta_j}{\tilde{S}_j} - \gamma_j$$

Setting the derivative to zero, we have

$$\gamma_j = (1-\delta_j)/\tilde{S}_j \ . \tag{2.7}$$

Furthermore,

$$\frac{\partial G}{\partial w_l} = \frac{\delta_l}{w_l} + \delta_l \lambda^\top g(T_l) - \eta + \gamma^\top U^{(l)},$$

where $U^{(l)} \in \{0,1\}^k$ is a vector with the j-th entry to be an indicator $I[T_{i_j} < T_l] \times (1-\delta_{i_j}) = I[T_{i_j} < T_l]$. Then set the derivative to zero and write l as i:

$$\eta = \frac{\delta_i}{w_i} + \delta_i \lambda^\top g(T_i) + \gamma^\top U^{(i)}.$$

Multiply w_i on both sides and sum,

$$\sum_i w_i \eta = \sum_i \delta_i + \sum_i \delta_i w_i \lambda^\top g(T_i) + (\sum_i w_i \gamma^\top U^{(i)}) \ .$$

Make use the other constraints, this simplifies to

$$\eta = (n-k) + 0 + \sum_i w_i \gamma^\top U^{(i)} \ . \tag{2.8}$$

We now focus on the last term above. Plug in the γ_j expression we obtained in (2.7)

above and switch order of summation. It is not hard to see that

$$\begin{aligned} \sum_{i=1}^{n} w_i \gamma^\top U^{(i)} &= \sum_{i=1}^{n} w_i \left(\sum_{j=1}^{k} \gamma_j U_j^{(i)} \right) \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{w_i I[T_{i_j} < T_i] \times I[\delta_{i_j} = 0]}{\tilde{S}_j} = \sum_{j=1}^{k} 1 I[\delta_{i_j} = 0] \\ &= k \ . \end{aligned} \tag{2.9}$$

Therefore equation (2.8) becomes $\eta = (n-k) + 0 + k$. Therefore $\eta = n$, we have

$$w_i = \frac{\delta_i}{n - \lambda^\top \delta_i g(T_i) - \gamma^\top U^{(i)}} \ , \tag{2.10}$$

where we further note (plug in the γ, and $\delta_{i_j} = 0$):

$$\gamma^\top U^{(i)} = \sum_{j=1}^{k} \frac{(1 - \delta_{i_j})}{\tilde{S}_j} I[T_{i_j} < T_i, \delta_{i_j} = 0] = \sum_{j=1}^{k} \frac{I[T_{i_j} < T_i]}{\tilde{S}_j} \ .$$

This finally gives rise to

$$w_i = w_i(\lambda) = \frac{\delta_i}{n - \lambda^\top \delta_i g(T_i) - \sum_{j=1}^{k} \frac{I[T_{i_j} < T_i]}{\tilde{S}_j}} \tag{2.11}$$

which, together with (2.5), provides a recursive computation method for the probabilities w_i, provided λ is given:

1. Starting from the left most observation, and without loss of generality (as noted above) we can assume it is an uncensored data point: $\delta_1 = 1$. Thus

$$w_1 = \frac{1}{n - \lambda^\top g(T_1)} \ .$$

2. Once we have w_i for all $i \leq l$, we also have all $\tilde{S}_j$ where $T_{i_j} < T_{l+1}$ and $\delta_{i_j} = 0$, by using ($\tilde{S}_j = 1 - \sum_i I[T_i \leq T_{i_j}] w_i$), then we can compute

$$w_{l+1} = \frac{\delta_{l+1}}{n - \lambda^\top g(T_{l+1}) - \sum_{j=1}^{k} \frac{I[T_{i_j} < T_{l+1}]}{\tilde{S}_j}} \ . \tag{2.12}$$

So this recursive calculation will give us w_i and $\tilde{S}_j$ as a function of λ.

Lemma 2.3.1. *In the special case of no constraint of mean, then there is no* λ *(or* $\lambda = 0$*) and we have*

$$w_{l+1} = \frac{\delta_{l+1}}{n - \sum_{j=1}^{k} \frac{I[T_{i_j} < T_{l+1}]}{\tilde{S}_j}},$$

which is the jump of the Kaplan-Meier estimator.

Proof. Use the identity $(1 - \hat{F})(1 - \hat{G}) = 1 - \hat{H}$, we first get a formula for the jump of the Kaplan-Meier estimator: $w_i = 1/n \times 1/(1 - \hat{G})$. This works for $\hat{F}$ as well as for $\hat{G}$. We next show that

$$1 - 1/n \sum_{j=1}^{k} \frac{I[T_{i_j} < T_{k+1}]}{\tilde{S}_j} = (1 - \hat{G}(T_{k+1}))$$

since the left hand side is just equal to the summation of jumps of $\hat{G}$ before T_{k+1}. □

To compute the log empirical likelihood ratio statistics, we have to find the λ value that is determined from the constraint equation

$$0 = \sum_i \delta_i w_i(\lambda) g(T_i) = \sum_i \frac{\delta_i g(T_i)}{n - \lambda^\top \delta_i g(T_i) - \sum_{j=1}^{k} \frac{I[T_j < T_i]}{\tilde{S}_j}} \ . \tag{2.13}$$

So, the iteration goes like this:

(1) *Initialization*: Pick a λ value that is near zero but no equal to zero, as $\lambda = 0$ gives the Kaplan-Meier.

(2) *Updating* w *and* $\tilde{S}$: With this λ find all the w_i's and $\tilde{S}_j$'s by the recursive formula (2.12) and (2.5).

(3) *Updating* λ: Plug those w_i into the right hand side of equation (2.13) above and call it θ. The w_i's obtained in step (2) are actually the constrained Kaplan-Meier with the constraint being these θ instead of zero.

(4) *Checking* λ *and repeat*: Check if θ is zero. If not, change the λ value and repeat, until you find a λ which gives rise to w_i and $\tilde{S}_j$ that satisfy $\theta = 0$.

Notes

Two special cases of the above formula are worth some more discussion:

i we point out again when $\lambda = 0$, we get the Kaplan-Meier directly. The constraint on the Kaplan-Meier disappears and this recursive formula provide us with a way to calculate the jumps of the classical Kaplan-Meier;

ii when there is no censoring, i.e. when all $\delta_i = 1$, this formula becomes

$$w_{k+1} = \frac{1}{n - \lambda g(T_{k+1})}$$

This is non-recursive and is precisely what Owen obtained in his 1988 paperOwen (1988). A discrete distribution with the same support as the empirical distribution but with probability proportional to w_{k+1} is a 1-parameter family of distributions with parameter λ which has been called "hardest parametric submodel for estimating $\int gdF$" (Andersen et al., 2012); or "least favor" by Bickel et al. (1998) and (Pan and Zhou, 1999).

One interesting property of this parametric family of distributions given by (2.13) is that the parametric information for estimating $\int gdF$ is also the nonparametric information for estimating $\int gdF$. So, our recursive formula is just the "hardest parametric submodel for estimating $\int gdF$" in the random censorship data setting. The nonparametric information for estimating $\int gdF$ is discussed in the two books mentioned above, as well as (Zhou, 2015).

For one dimensional λ, solving (2.13) is going to be easily handled by any function that computes the root of a univariate function such as, for instance, the *uniroot* function in R. For multi dimensional λ this calls for each Newton type iteration.

The empirical likelihood ratio is then obtained as

$$-2\log ELR = -2\{\log EL(w_i, S_j) - \log EL(w_i = \Delta\hat{F}_{KM}(T_i))\} ;$$

for the first log EL inside the curly bracket above we use the expression (2.6) with the w_i, S_j computed from the recursive method in this section, and the second term is obtained by (2) with $w_i = \Delta \hat{F}_{KM}(T_i)$, the regular Kaplan-Meier estimator.

Observations (T_i, δ_i) are first ordered according to their T_i values. If there are several observations with identical T_i value, we then order according to their $-\delta_i$ value. That is, an uncensored T_i is considered as come before a censored T_j even when $T_i = T_j$. This is the usual convention in the calculation of the Kaplan-Meier estimators. When there are observations with identical T_i and δ_i value, we can either merge the tied observations and record the number of tie in another vector u_i; or we may just leave the tied observations as is, in the order of their input. When there are substantial(extensive) tied observations, it may save computational time to first merge the tied data record the number of tied in u_i before the recursive computation.

However, in most applications of survival analysis the Kapan-Meier estimator will be computed along with some covariates, as in regression analysis. So the actual data likely will look like (T_i, δ_i, x_i) where x_i are the covariates, e.g. treatment, gender, age, blood pressure, etc. of the i-th patient. In this case, even if two observations have identical T_i and δ_i should not be merged because they have different covariates. Therefore in the current implementation of *kmc* package, we choose not to merge any tied data.

Under the assumption that the variance of $\int g(t) d\hat{F}_{KM}(t)$ is finite (if $p = 1$), and variance-covariance matrix is nonsingular (if $p \geq 2$), we have a chi square limiting distribution for the above -2 log empirical likelihood ratio, under null hypothesis as stated in the Wilks' Theorem. Therefore we reject the null hypothesis if the computed -2 empirical likelihood ratio exceeds the chi square 95% percentile with p degrees of freedom. See Zhou (2010) for a proof of this theorem.

Root solving and initial values

For any optimization problem, there are always initial values/tuning parameter problems. Most of them relate to the properties of certain optimization algorithm. In our approach, the problem is not only about the optimization method we used, but, more

importantly, the nature of the algorithm itself.

If the Jacobian matrix of mean zero requirement (2.13) is not singular, Newton-Raphson method could be used to find the root(s) of (2.13):

$$0 = \sum_i \delta_i w_i(\lambda) g(T_i) = \sum_i \frac{\delta_i g(T_i)}{n - \lambda^\top \delta_i g(T_i) - \sum_{j=1}^k \frac{I[T_j < T_i]}{\tilde{S}_j}} \ .$$

We shall call the recursive computation for $w_i's$ plus the Newton iteration for λ as Kaplan-Meier-constrained (KMC) method, which is also the name of the R package *kmc* available on the comprehensive R archive network (CRAN).

Once $\tilde{S}$ is given, (2.13) has more that one roots of λ.

To simplify the proof, we assume there is only one constraint. Hence $dim(\lambda) = 1$. if we further define the i-th entry of $\lambda^\star$ as:

$$\lambda_i^\star = \frac{n - \sum_{j=1}^k \frac{I[T_j < T_i]}{\tilde{S}_j}}{g(T_i)}.$$

For $\tilde{S}_j = s_j$, denote $\psi(\lambda)$ as:

$$\psi_i(\lambda) = n - \lambda \delta_i g(T_i) - \sum_{j=1}^k \frac{I[T_j < T_i]}{\tilde{S}_j}$$

For those i's such that $\delta_i \neq 0$:

$$\begin{cases} \lim_{\lambda \searrow \lambda_i^\star} \psi_i(\lambda) \to 0^+ \\ \lim_{\lambda \nearrow \lambda_i^\star} \psi_i(\lambda) \to 0^- \end{cases}$$

Notice $\psi(\lambda)$ is the denominator term, hence the i-th term in (2.13) goes to infinity if $g(T_i) \neq 0$:

$$\begin{cases} \text{sign}(g(T_i)) \times \lim_{\lambda \searrow \lambda_i^\star} \delta_i w_i(\lambda) g(T_i) & \to +\infty \\ \text{sign}(g(T_i)) \times \lim_{\lambda \nearrow \lambda_i^\star} \delta_i w_i(\lambda) g(T_i) & \to -\infty \end{cases}$$

The combination of $w_i(\lambda)g(T_i)$'s that make non-zero sum gives the number of roots. This property of roots in the KMC computation is helpful to know the bound of

solution, or so called, feasible region. In each Newton iteration, when we try to find the root, it is obvious that those $\lambda' s$ such that $n - \delta_i \lambda^\top g(T_i) - \sum_{j=1}^{k} \frac{I[T_{i_j} < T_i]}{\tilde{S}_j} = 0$ will lead (2.13) to ∞. For one constraint problem, we could split the real line by $\lambda^\star$'s shown in Figure 2.3. For multi-constraints problems, we could just consider each dimension separately.

Meanwhile, as mentioned previously, the null parameter space that has no constraint corresponds to $\lambda = 0$. Then any i-th entry of the desired λ root for (2.13) must be in the region that contains 0, i.e. satisfies

$$\exists j \text{ such that } \lambda^\star_{i_j i} < 0, \lambda_i < \lambda^\star_{i_{j+1} i} \quad \forall i = 1, \ldots, p$$

where $\lambda^\star_{ij}$ is the j-th entry of vector $\lambda^\star_i$ such that $\lambda^\star_{i_1 i} < \ldots < \lambda^\star_{i_n i}$, and $\lambda^\star_{0,i} \stackrel{\Delta}{=} -\infty$, $\lambda^\star_{n+1,i} \stackrel{\Delta}{=} +\infty$. So, one suggested strategy is to start at 0 and try to stay within the feasible region at all times when carry out the Newton iterations, or only consider the λ in the feasible region that gives all $w(\lambda)$ that are non-negative.

We could also calculate the analytical derivatives used in the Newton iteration. Denote the right hand side of (2.13) as $f(\lambda)$, i.e.

$$f(\lambda) = \sum_i \delta_i w_i(\lambda) g(T_i) \ .$$

To compute $\frac{\partial}{\partial \lambda} f(\lambda)$, we only need to calculate $\frac{\partial}{\partial \lambda} w_i$ and $\frac{\partial}{\partial \lambda} \tilde{S}_j = \frac{\partial}{\partial \lambda} \left(1 - \sum_{k=1}^{j} w_k\right) = -\frac{\partial}{\partial \lambda} \sum_{k=1}^{j} w_k$. There are no closed forms of such derivatives, but it could again be derived recursively. The following lemma summarizes the calculation.

Lemma 2.3.2. *Recursive calculation of derivatives of* $w(\lambda)$

(1) Calculate $w_1(\lambda)$ *and* $\frac{\partial}{\partial \lambda} w_1(\lambda)$*:*

$$w_1 = \frac{1}{n - \lambda g(T_1)}, \ \textit{and} \ \frac{\partial}{\partial \lambda} w_1 = w_1^2 g(T_1)$$

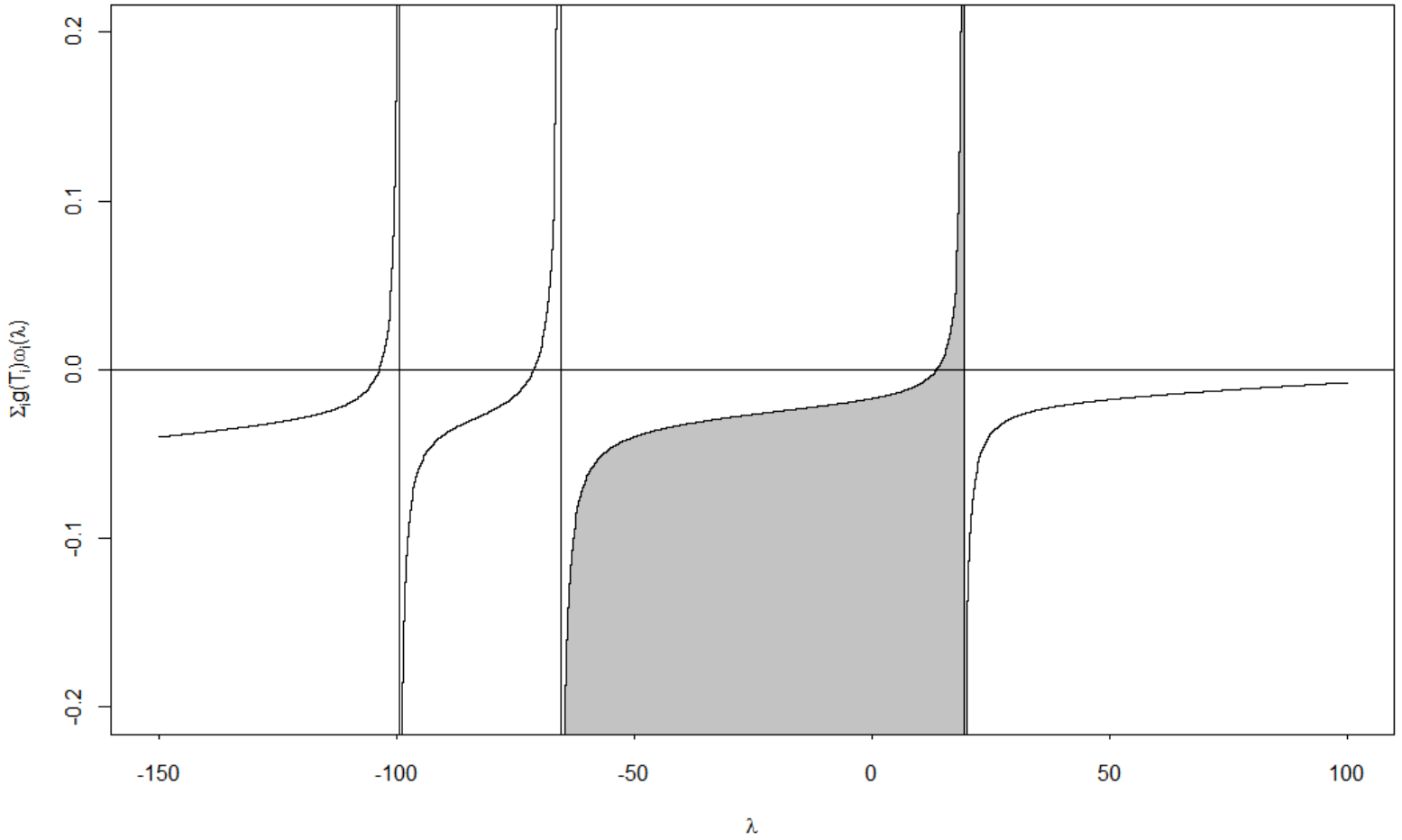


Figure 2.2: Example of λ v.s. $f(\lambda) = \sum_i \delta_i w_i(\lambda) g(T_i)$.

(2) Update $\frac{\partial}{\partial\lambda}w_{k+1},\ k \geq 1$*:*

$$\frac{\partial}{\partial\lambda}w_{k+1}(\lambda) = \delta_{k+1}(w_{k+1})^2\times \left(g(T_{k+1}) + \sum_{j=1}^{n}\left(I[T_{i_j} < T_{k+1}](\tilde{S}_j)^{-2}\frac{\partial}{\partial\lambda}\sum_{s=1}^{i_j}\frac{\partial}{\partial\lambda}w_s\right)\right).$$

Note: checking the constraints

We need to check whether the constraints are proper. For example, if the constraint is $\int_0^\infty xdF(x) = -1$, then there is no solution (as the left side is always positive). It is easy to check when there is only one constraint. For more than one constraints, we refer to (Dines, 1926).

Theorem 2.3.3 (Dines, Lloyd L)**.** *(Positive solutions of a system of linear equations) Consider the linear equations*

$$\sum_{s=1}^{n} a_{rs}x_s = 0, \quad r = 1,\ldots,m$$

with real coefficient a_{rs}*. There exists a solution* $(x_1,\ldots,x_n)$ *in which every component is positive if we can apply the mathematical induction algorithm till* $m = 1$*:*

Step 1*. if* $m = 1$*, at least one sign of* a*'s are different than others;*

Step 2*. if* $m > 1$*, then construct a new linear equation system with coefficients*

$$a_{r,ij}^{(new)} = a_{1i}a_{rj} - a_{1j}a_{ri}\forall r = 2,\ldots,m \quad .$$

Running Step 1 and Step 2 iteratively gives the final criterion.

For example, if $m = 3$, then we run step 2 to get a linear equation with $m = 2$, and run step 2 again to get a linear equation with $m = 1$. If all the coefficients have the same sign, then there is no positive solution for this problem.

This theorem presents a simple algorithm for determining whether there is a possible solution to (2.4). But we should notice that in step 2, it increase the number of the coefficients to be checked in a power trend. Considering the sample size n is

relative large, the checking procedure is not applicable for large number of constraints, i..e. p in (2.3) and m in Theorem 2.3.3 is large. The routine was implemented in *kmc::kmc.solve*.

2.4 Application: Hypothesis Testing Problem in Acceleration Failure Time Model

The accelerated failure time (AFT) model (Cox and Oakes, 1984, Kalbfleisch and Prentice, 2011) is an important alternative to the widely used proportional hazard model (Cox (1972)) in regression analyzing of censored failure time data, which not only focuses on properties of survival function instead of hazard function, but also provides a direct interpretation of linear relationship between logarithm of failure time and covariates. We will discuss more details of AFT model in later chapters.
In general, AFT model assumes

$$\log(T_i) = X_i^\top \beta + \epsilon_i, i = 1, \ldots, n, X_i \in \mathbb{R}^p \tag{2.14}$$

Here, T_i is the survival time of the i-th observation, X_i is the corresponding covariates p-dimension vector and the measurement error ϵ_i's are i.i.d. sampled from cumulative distribution function F_ϵ and are independent from X. For right censoring problem, we further assume the logarithm of the censoring time C_i is i.i.d distributed. Hence, for each $i = 1, \ldots, n$, we only observe a combination $(Z_i, \delta_i,\ X_i)$. Here $\delta_i = I_{Y_i \leq C_i}$ is the censoring indicator, $Z_i = \log \min(T_i,\ C_i)$.
We don't specify any distribution to the residual term ϵ. Otherwise, the likelihood of a parametric AFT model is easy to be calculated.
The hypothesis is to test coefficient of the AFT model, i.e.

$$H_0 : \beta = b_0$$

Here, β could be either a single value or a vector. In this section, we use KMC to do the hypothesis testing problem on coefficient β instead of classical EM method.

We chose the Buckley-James estimator and construct a log likelihood ratio statistics to solve the hypothesis testing problem. As we may later discuss in Chapter 3, the essential estimation equation for Buckley-James estimator is (3.7), i.e.

$$\sum_{i=1}^{n} \left\{ \delta_i e_i(\beta) + (1-\delta_i) \sum_{j:e_j>e_i} \frac{e_j(\beta) \Delta \hat{F}_\beta(e_j(\beta))}{1-\hat{F}_\beta(e_j(\beta))} \right\} X_i = 0$$

Here $e_i(\beta) = Z_i - X_i^\top \beta$, and $\hat{F}_\beta$ is the Kaplan-Meier estimator of $e_i(\beta)$'s once β is given.

Switch order of i and j we derive a linear constraint on $\Delta \hat{F}$ defined in the formula (5) in (Zhou and Li, 2008)'s paper. Then maximizing the empirical likelihood under such constraint leads to the NPMLE of residual under H_0.

By doing this, we could transform a regression coefficient hypothesis testing problem into a maximizing empirical likelihood under "mean" type constraint problem. Therefore, we could still use KMC to calculate the result fast once the problem could be represented into empirical likelihood with linear constraint problem. The real data example could be found in the next section. Besides, an R function *kmc::kmc.bjtest* was implemented in KMC to help researchers to test coefficients in AFT model.

2.5 Simulation

To evaluate the performance of this algorithm, a series of simulations had been done. We compared with standard EM algorithm (Zhou, 2005). Without further statement, all simulation works have been repeat 5,000 times and implemented in R language (R Core Team, 2014). R-2.15.3 is used on a Windows 7 (64-bits) computer with 2.4 GHz Intel(R) i7-3630QM CPU. The full parameter list of the R function in kmc package could be found in the help manual or Table 2.1.

Here is an example offered by a submitted paper of (Zhou and Yang, 2016) to illustrate the usage of both the KMC package and emplik package to solve the restricted mean survival time hypothesis testing problem. In cancer study, an often used measure of overall survival is the Restricted Mean Survival Time, especially when the

Table 2.1: Parameter List of kmc

Parameter	Function
x	Positive time
d	Status, 0: right censored; 1 uncensored
g	list of constraint functions. It should be a list of functions list(f1,f2,...)
em.boost	logical asking whether to use EM to get the initial value, default=TRUE. See 'Details' for EM control.
using.num	logical asking whether to use numeric derivative in iterations, default=TRUE.
using.Fortran	logical asking whether to use Fortran in root solving, default=F.
using.C	logical asking whether to use Rcpp in each iteration, default=T. This option will promote the performance of KMC algorithm. Development version works on one constraint only. Otherwise it will generate an Error information. It won't work on using.num=F.
tmp.tag	Development version needs it, keep it as TRUE.
rtol	Tolerance used in rootSolve(multiroot) package, see 'rootSolve::multiroot'.
control	nr.it controls max iterations allowed in N-R algorithm default=20, nr.c is the scaler used in N-R algorithm default=1,em.it is max iteration if use EM algorithm (em.boost) to get the initial value of lambda, default=3.
...	Unspecified yet.

proportional hazards assumption is in doubt and heavy censoring is present. See Royston and ParmarRoyston and Parmar (2013), also the R package `survRM2`Tian et al. (2014).

An expression of the restricted mean is

$$\mu(\tau) = \int_0^\tau 1 - \hat{F}_{KM}(s)ds \ ,$$

where $\hat{F}$ is the Kaplan-Meier and τ is the pre-specified restriction time. Another way to calculate the restricted mean survival time is

$$\mu(\tau) = \int_0^\infty \min(t, \tau) d\, \hat{F}_{KM}(t) \ .$$

We can construct a confidence interval of $\mu(\tau)$ by inverting the empirical likelihood

ratio test. The tests are computed via the tilted Kaplan-Meier with restricted mean survival time set at a value.

We used the dataset `ovarian` from the `survival` package and pre-select the time restriction $\tau = 700$.

```
library(survival)
library(kmc)
data(ovarian)
kf <- function(x){ pmin(x,700) - 532.6 }
kmc.solve( x= ovarian$futime, d = ovarian$fustat, g = list
    (kf))
```

This tests the hypothesis that the restricted mean survival is equal to 532.6: $H_0 : \mu(700) = 532.6$. You get same result but slower, by using the function `el.cen.EM2` from the package `emplik`.

```
el.cen.EM2(x = ovarian$futime, d = ovarian$fustat, fun =
    function(x){ pmin(x, 700) - 532.6 }, mu=0)
```

The previous code are used as template in the following simulations. More details could be found in `kmc`'s development page on GitHub.

Experiment 1 : Consider a right censored data with only one constraint:

$$\begin{cases} X \sim \text{Exp}(1) \\ C \sim \text{Exp}(\beta) \end{cases} \tag{2.15}$$

Censoring percentage of the data are determined by different $\beta's$. Three models are included in the experiments

(1) $\beta = 1.5$, then 40% data are *uncensored*

(2) $\beta = 0.7$, then 58.9% data are *uncensored*

(3) $\beta = 0.2$, then 83.3% data are *uncensored*

The common hypothesis is (2.3), where $g(x) = (1-x)1_{(0\le x\le 1)} - e^{-1}$. We could verify that the true expectation is zero: $\int g(x)\,\mathrm{d}F(x) = \int (1-x)1_{(0\le x\le 1)}e^{-x}dx - e^{-1}$=0. To compare the performances of KMC and EM algorithm, we use four different sample sizes, i.e. 200, 1,000, 2,000, and 5,000 in the experiments. To make fair comparisons, $\|f^{(t)} - f^{(t+1)}\|_{\ell_2} \le 10^{-9}$ is used as the convergence criterion for EM algorithm, and $\|f^{(t)}\|_{\ell_1} \le 10^{-9}$ is used for KMC. Average spending time is reported to compare the computation efficiency in Table 2.2. The no censored case is included for reference, this is equivalent to Newton solving λ without recursion. In all cases in our study,

Table 2.2: Average running time of EM/KMC (in second). "No Censor" column refers to time spend on solving empirical likelihood without censoring in R *el.test*(**emplik**). We use this as a comparison reference.

Censoring Rate	N	EM	KMC(nuDev)	KMC(An.Dev)	No Censor
60% $\beta = 1.5$	200	0.175	0.011	0.028	0.005
	1000	3.503	0.106	0.211	0.007
	2000	13.935	0.349	0.692	0.033
	5000	73.562	1.801	3.663	0.036
41% $\beta = 0.7$	200	0.064	0.010	0.029	0.000
	1000	1.058	0.115	0.268	0.010
	2000	4.104	0.385	0.836	0.020
	5000	22.878	2.367	4.693	0.037
17% $\beta = 0.2$	200	0.014	0.008	0.029	0.002
	1000	0.117	0.071	0.240	0.009
	2000	0.425	0.240	0.694	0.018
	5000	2.702	1.220	3.282	0.026

EM and KMC reported almost the same χ^2 test statistics and a quantile to quantile plot is shown in Figure 2.3. The plot shows good agreement to χ^2 distribution with $p - value = 0.5258$ using Kolmogrov-Smirnov test.

As shown in Table 2.2, we observed the following phenomenons:

(1) KMC always outperformed EM algorithm in speed at different simulation settings.

(2) Computation complexity of EM increased sharply with the percentage of censored data increasing. This is reasonable, since more censored data needs more E-step computation. But censored rate did not affect KMC much.

(3) Sample size is related to the computation complexity. We could see the running time of both EM and KMC increased along with the sample size.

(4) Another phenomenon is that, the computation of numeric derivative and analytic derivative of KMC is similar. But numerical derivative is slightly better. iteratively.

To summarize, when sample size is small and censored rate is low, the performance of EM and KMC is similar. But either in the large sample case or heavily censored case, KMC far outperformed EM algorithm with the same stopping criterion.

Experiment 2 : Consider a right censored data setting with two constraints. The i.i.d. right censored data are generated by:

$$\begin{cases} X \sim \text{Exp}(1) \\ C \sim \text{Exp}(.7) \end{cases} \tag{2.16}$$

with the following hypothesis:

$$H_0 : \sum_i g_j(T_i) w_i = 0; \quad j = 1, 2 \quad \text{where} \begin{cases} g_1(x) = (1-x)1_{(0 \le x \le 1)} - e^{-1} \\ g_2(x) = 1_{(0 \le x \le 1)} - 1 + e^{-1} \end{cases} \tag{2.17}$$

It is straightforward to verify that both g functions have expectation zero. In this

Table 2.3: Average running time of EM/KMC (in second)

Censoring Rate	N	EM	KMC(nuDev)
41%	200	3.055	0.033
41%	500	55.601	0.083

simulation study, we observed that EM spent great amount of time (3s $\sim$ 55s per case) to meet the converge criterion, while the average running time of KMC was considerable shorter (0.03s $\sim$ 0.08s). This dramatic result shows in multi-dimensional case, KMC runs much faster than EM algorithm. Only numerical derivatives were used in our simulations. One could implement the analytic ones using iteration shown previously. But in multi-dimensional case, the iterative type of derivatives do not have

advantage over numeric ones. We recommend using KMC with numeric derivative if one has more than one hypothesis even the sample size is small.

Experiment 3 : Other than exponential setting, considering a right censored data with one constraints:

$$\begin{cases} X \sim \Gamma(3,2) \\ C \sim U(0,\eta) \end{cases} \tag{2.18}$$

with hypothesis

$$H_0 : \sum_i g(T_i)w_i = 0, \text{ with } g(x) = x - 1.5$$

we carried out some experiments on censoring time C_i from uniform distribution. There were two models:

(1) $\eta = 5$, then 70.00% data are uncensored;

(2) $\eta = 3$, then 51.34% data are uncensored.

Table 2.4: Average running time of EM/KMC (in second) of one constraint and Uniform distributed censored time

Censoring Rate	**N**	**EM**	**KMC(nuDev)**	**KMC(An.Dev)**	**No Censor**
30.00%	200	0.124	0.018	0.044	0.005
$\eta = 5$	2000	11.237	0.725	1.197	0.112
48.66%	200	0.075	0.019	0.045	0.004
$\eta = 3$	2000	4.141	1.068	1.528	0.139

We found that the result shown in Table 2.4 is very similar to Table 2.2, which infers that different distribution of censored time will not affect the relative timings too much.

2.6 A Real Data Example

The speed advantage of KMC algorithm is more apparent in time consuming analysis such as drawing contour plot. In this real data example, we illustrate the proposed algorithm to analyze the Stanford Heart Transplants Data described in Miller and Halpern (1982) and considered regression with intercept and slope term of age. There

were 157 patients who received transplantation, among which 55 were still alive and 102 were deceased. We deleted cases that the survival times are less than 5 days, and used only 152 cases in the analysis, as suggested by Miller and Halpern. To draw such contour plot $51 \times 51 = 2601$ empirical likelihood ratios were calculated. In this example, we used KMC to calculate the empirical likelihood instead of EM described in (Zhou and Li, 2008).

Firstly, two hypothesizes on survival function are considered:

$$H_0 = \begin{cases} H_0^{(1)}: & \text{Mean} = \int_{x \geq 0} x dF(x) = \mu \\ H_0^{(2)}: & F(3) = \int_{x \geq 0} I(x \leq 3) dF(x) = \nu \end{cases} \tag{2.19}$$

In Figure 2.3, 30×30 combinations of (μ, ν) near NPMLE(0.5569,3.061), i.e. value plugged in with Kaplan Meier estimation, were used to construct a contour plot of the constrained log empirical likelihood. On the same computer, the program finished in 17 seconds. EM based method could also reproduce the same plot, but the time spend is not evaluated as some values fails to converge within 2 minutes.

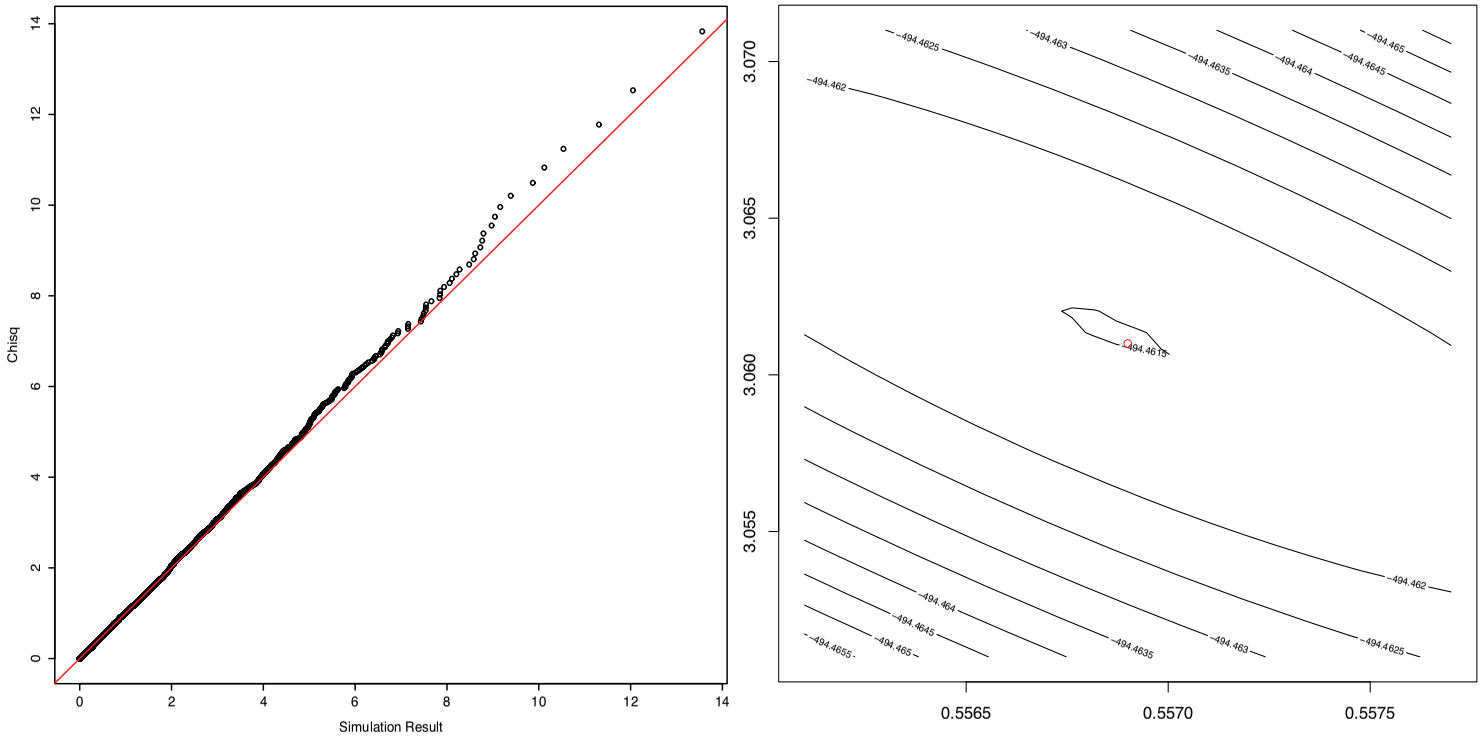


Figure 2.3: Left: Q-Q Plot for KMC with $\beta = .2$ N=1,000; Right: Contour plot of two hypothesizes.

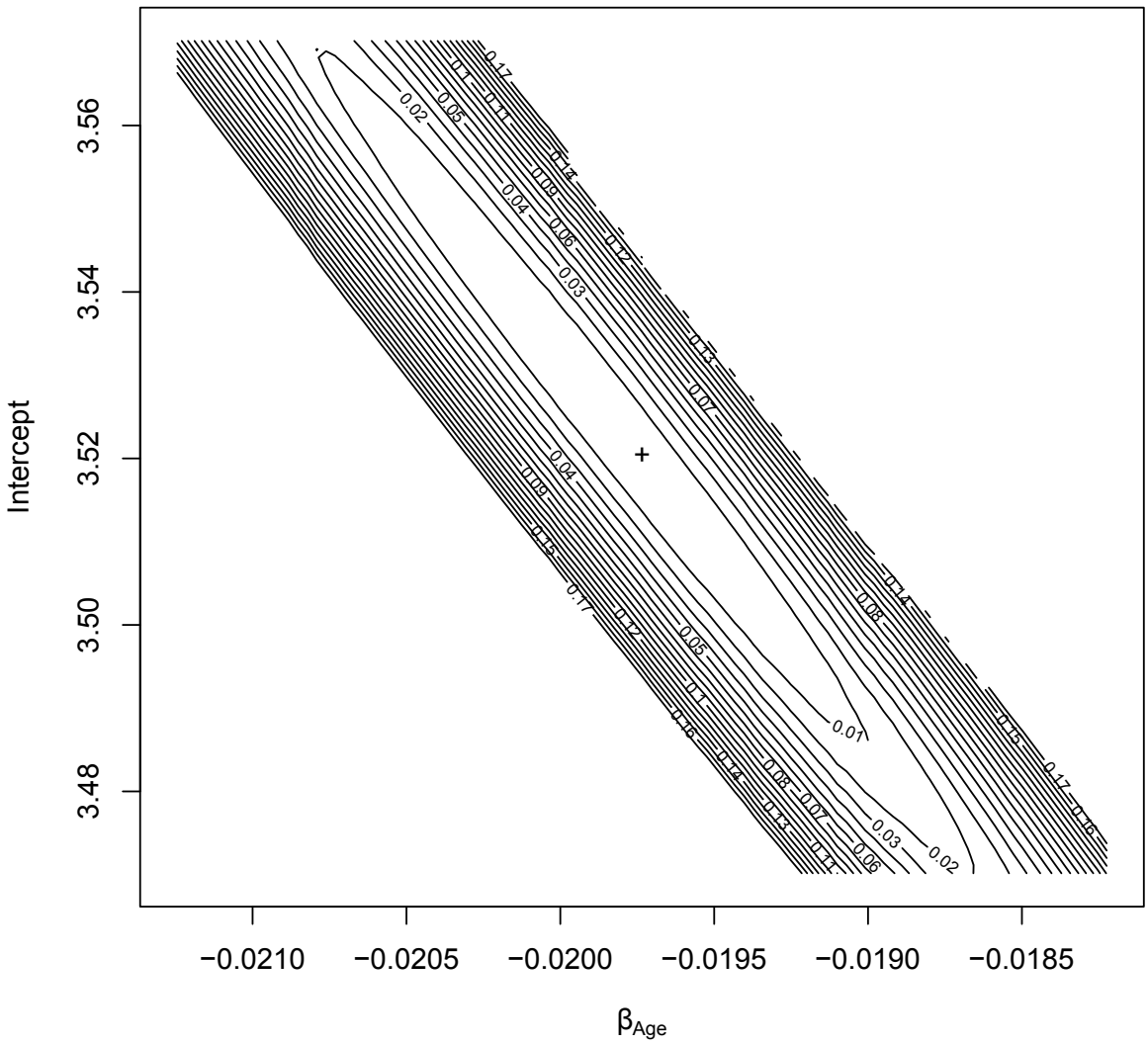


Figure 2.4: Contour plot of -2 log likelihood ratio corresponding to intercept and slope of age for the Stanford Heart Transplants Data.

In Figure 2.4, KMC could be able to derive the contour plot of -2 log likelihood ratio corresponding to intercept and slope of age very quickly too.

2.7 Discussion and Conclusion

In this chapter, we proposed a new recursive algorithm, KMC, to calculate mean constrained Kaplan-Meier estimator and log empirical likelihood ratio statistics of right censored data. Our algorithm used Lagrange multiplier method directly, and recursively computes the jumps of the constrained Kaplan Meier estimator.

Numerical simulations show this new method has an advantage over traditional EM algorithm in the sense of computational complexity. Our simulation work also shows that the performance of KMC does not depend on the censoring rate, and outperformed EM algorithm at every simulation setting. We recommend to use KMC in all cases but particular large gain are expected in the following cases:

(1) Sample size is large (e.g. > 1000 observations);

(2) Data are heavily censored (e.g. censored rate $> 40\%$);

(3) There are more than one constraints.

On the other hand, and somewhat surprisingly, the analytic derivative did not help speed up computation in our simulation study. Besides, since KMC with numeric derivative method could be extended to more than one constraints case, we highly recommend using numeric derivative in KMC rather than analytical one.

One of the issues of KMC is the initial value choosing, as is the case for most Newton algorithms. The performance of root solving relies on the precision of numerical derivative and Newton method. Our current strategy uses the M-step output of EM algorithm with only two iterations. Other better initial values are certainly possible. In addition, current KMC only works on right censored data, while EM algorithm works for right-, left- or doubly censored data; or even interval censored data. We were unable to find a proper way to derive such recursive computation algorithm in other censoring cases.

Software is available to download at `http://cran.r-project.org/web/packages/kmc` as a standard R package.

Chapter 3 High dimensional Accelerated Failure Time Model

3.1 Introduction

In recent years, high dimensional problem attracts many researchers' attention. One particular area is the variable selection problem in a regression setting:

$$Y = X\beta + \epsilon,\ \beta \in \mathbb{R}^p.$$

Here $X \in \mathbb{R}^{n\times p}$ is the explanatory variable, $Y \in \mathbb{R}^n$ is the response variable, and ϵ is i.i.d distributed error term. n is the sample size, the i-th row X represents the observed explanatory value of the i-th observation. We assume that the vector β contains several component that is/are zero. The model selection method aims to exclude those zero component from the model. If p is greater than n, or $p >> n$, such as $p = O(n^2)$, we call this the high dimensional problem.

Typically, there are several standard approaches to address this problem. Among them, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are very interesting. They both discuss likelihood function, which is much general than the regression problem. To see this, let us inspect the regression problem briefly. Assume the residual term ϵ is f_ϵ distributed, where f_ϵ is a given probability density function. Then we could estimate coefficient β through the maximum likelihood estimation (MLE):

$$\arg\max_{\beta} \prod_{i=1}^{n} f_\epsilon(Y_i - X_i^\top \beta),$$

where $X_i^\top$ is the i-th row of the matrix X.

For instance, if ϵ is Normally distributed, then the MLE equals the ordinary least squares estimation (OLSE). We now introduce the general idea which are applied to likelihood, such as AIC and BIC, and then move on to the linear regression problem in the later content.

AIC (Akaike, 1974, 1974) minimize the Kullback-Leibler (KL) divergence of the pre-

dicted model and the true model:

$$\mathrm{KL}(f, g(.|\theta)) = \int f(x) \log(\frac{f(x)}{g(x|\theta)})dx = \int f(x) \log f(x)dx - \int f(x) \log g(x|\theta)dx \tag{3.1}$$

Here

$f(.)$ is the truth in terms of the probability distribution function;

$g(.|\theta)$ is the approximation in terms of the probability distribution function.with θ as the parameter vector.

Notice

Take the regression problem for example. Here θ contains β and the parameters in f_ϵ. For example, if we assume the residual term ϵ is normally distributed with unknown standard deviation σ, then $\theta = (\beta^\top, \sigma)^\top$. Although in the problem we only focus on estimating β, the standard deviation σ can not be ignored. Other examples are extreme value distribution that can be found in later content.

The Akaike Information Criterion links the KL distance and maximized likelihood together, or equivalently, integrates the distance between two models and parametric estimation together:

$$\min_{g \in \mathcal{G}} E_y[\mathrm{KL}(f, g(.|\hat{\theta}_g(y)))]$$

Here

y is the random sample from the true density function $f(x)$;

$\mathcal{G}$ is the family of all "admissible" models;

$\hat{\theta}(.)$ is the MLE based on the model $g()$.

Plug in the definition of KL-distance, then we only need to solve the following optimization problem:

$$\max_{g\in\mathcal{G}} E_y E_x[\log g(x|\hat{\theta}_g(y))]$$

AIC is indeed an approximately unbiased estimator (Akaike, 1973) of $\max_{g\in\mathcal{G}} E_y E_x[\log g(x|\hat{\theta}_g(y))]$ for large sample and model "closed" to $f(.)$ in the sense of having small KL-distance with the following formula:

$$\log(L(\hat{\theta}|y)) - p$$

Here L is the likelihood function, $\hat{\theta}$ is the MLE of θ, p is the number of estimated parameters. There are several variations of AIC, for example: Takewchi's Information Criterion/TIC (Takewchi, 1976) is useful in cases where the model is not "closed" to the true model; AICc(Akaike, 1985) is useful in small-sample-size cases.

In general, AIC considers to penalize the model complexity to determine the "best" model, which leads to researches on model complexity, and model degrees of freedom. Examples are (Friedman, et,.al 2001), and (Efron, 2004) among many others.

On the other hand, some Bayesian methods are also proposed. For example Bayesian Information Criterion/BIC (Schwarz, 1979) and reversible jump Monte-Carlo Markov chain (MCMC) by (Green, 1995). Especially, BIC has a similar formula as AIC:

$$\text{BIC} = -2\log \text{L}(\hat{\theta}) + p\log n$$

Here n is the sample size.

From the previous summary, we notice the most commonly used two criteria: AIC and BIC share the same form if we introduce a positive regulatory parameter λ:

$$\max_{\theta} \frac{1}{n}\ell(\theta) - \lambda\|\theta\|_{\ell_0} \ , \tag{3.2}$$

where ℓ is the log likelihood and $\|x\|_{\ell_0} = \#\text{supp}\{x\}$ is the number of non-zero elements in x. With given $\|x\|_{\ell_0} = k$, we could solve (3.2) by maximizing the likelihood constraint to all the subset with size k. As the true model is unknown, we need to

go over all possible k and all subsets (it is actually an NP-hard problem) to solve the problem. Hence classical approaches that solve (3.2), including AIC and BIC, are only applicable in low dimension setting (the number of parameter p is small, and $p < n$) and causes an impracticable computational complexity when p is large as a result of the curse of dimensionality(Friedman et al., 2001).

Now let us come back to the regression problem. Here we set $\theta = \beta$, if the residual term in normally distributed, then 3.2 is equivalent to minimize:

$$\|Y - X\beta\|_{\ell_2}^2 + \lambda\|\beta\|_{\ell_0} \tag{3.3}$$

This is an example of penalized regression, which are proposed and studied to solve the high dimensional problem. Examples are LASSO(Tibshirani et al., 1997), Adaptive LASSO(Zou, 2006), Elastic Network(Zou and Hastie, 2005), SCAD(Fan and Li, 2001), MCP/Mini-Max(Zhang et al., 2010), Dantzig(Candes and Tao, 2007), Compressive Sensing(Candes and Tao, 2007) and more. All these methods focus on the least squares problem with particular penalty term, i.e. penalized least squares (Fan and Lv, 2010), to derive "sparse" estimations and hence select variables automatically:

$$\left\{\frac{1}{2}\|Y - X\beta\|^2 + \sum_{j=1}^{p} p_\lambda(\beta_j)\right\},$$

where $\|Y - X^\top\beta\|^2$ is the ℓ^2-norm of the residual, $p_\lambda(.)$ is the penalty function other than $\|.\|_{\ell_0}$ used in (3.2). The penalty terms limit the parameter solution space of θ to a subset of one without penalty terms. Hence it has a potential to provide an estimation when $p > n$ or even $p >> n$, in which OLSE fails due to singularity of the matrix $X^\top X$. To be strict, (Fan and Li, 2001) proposed rules that "good" properties a penalty term should have, and name it as oracle properties (see Definition 3.2.1).

For (right) censored regression problem, there are several regression models that are produced and used widely. One approach is the proportional hazards model/Cox model, which studies the hazard rate and use partial likelihood(ref) as a powerful tool. Meanwhile, because Cox model has the partial likelihood, (Tibshrirani, 1997)

estimates the coefficients of Cox model in high dimensional settings though LASSO and partial likelihood directly:

$$\hat{\beta} = \arg\max \mathrm{PL}(\beta) + \lambda^{\star} \sum |\beta_j|,$$

or its dual form (Osborne et al., 2000):

$$\hat{\beta} = \arg\max \mathrm{PL}(\beta), \text{subject to } \sum |\beta_j| \leq \lambda.$$

Here, $PL(.)$ is partial likelihood proposed in (Cox, 1972). In this thesis, we focus on the accelerated failure time (AFT) model and shall not further study the proportional hazards model any more.

The AFT model is an important alternative to the widely used Cox model in regression analyzing of censored failure time data. It concentrates on and provides a direct interpretation of the linear relationship between the logarithm of failure time and explanatory variables. In general, for a random time-to-vent T, the accelerated failure time model is:

$$\log T_i = X_i^{\top}\beta + \epsilon_i,\ i = 1, \ldots, n,\ X_i \in \mathbb{R}^p. \tag{3.4}$$

Here, T_i is the survival time of the i-th observation, X_i is the corresponding covariates p-dimension vector and the measurement error ϵ_i's are i.i.d. sampled from cumulative distribution function F_{ϵ} and are independent from X. For right censoring problem, we further assume the logarithm of the censoring time C_i is i.i.d distributed. Hence, for each $i = 1, \ldots, n$, we only observe a combination $(Z_i, \delta_i,\ X_i)$. Here $\delta_i = I_{T_i \leq C_i}$ is the censoring indicator, $Z_i = \log \min(T_i,\ C_i)$.

In low-dimensional setting, at least two methods have become the standard ways to solve AFT model:

1. Rank based method;

2. Buckley-James method.

The rank-based method is motivated by the score function. But the computational complexity is too high in (Prentice 1978, Wei et al 1990, Ying 1993) to be applied in low/high dimensional setting until (Zhou 1992) and (Stute 1993). The latter two used inverse probability weighting (IPW) method and minimized a weighted least squares loss function. Some authors have used this IPW method and extended it into high-dimensional setting. But unfortunately, there is no mature result for the rank-based method in high dimensional settings.

The Buckley-James method is another choice. It uses Kaplan-Meier estimator to solve a particular estimation equation iteratively (Section 3.3). We will consider the iterative idea used in Buckley-James as a potential to solve the model selection problem in AFT model in the discussion section.

The structure of this chapter is as follows. In Section 3.2, we introduce existing penalized least squares models and the properties of different penalty function p_λ. The In Section 3.3, we propose our parametric accelerated failure time model in the high-dimensional setting and discuss the selection consistency of our model. This parametric AFT model with penalty is easy to use and has a potential to be extended to non-parametric setting. In addition, we also introduce an approach to tune p_λ and show the mechanism and performance of the tuning method. In Section 3.4, we illustrate the model performance in high-dimensional setting by repeating simulations. Section 3.5 summarized and concludes the parametric AFT model in high-dimensional setting with comments for some possible future work.

3.2 Penalized Least Squares Model for Linear Model

We now focus on the classical regression problem in the beginning of this chapter:

$$Y = X\beta + \epsilon$$

If ϵ is from a normal distribution $N(0, \sigma^2 I)$, then the penalized likelihood could be rewritten into a penalized least squares (PLS) form (Fan and Lv, 2010):

$$\min_{\beta \in \mathbb{R}} \{ \frac{1}{2n} \| Y - X\beta \|_{\ell_2}^2 + \sum_j p_\lambda(|\beta_j|) \} \tag{3.5}$$

Here $\| Y - X\beta \|_{\ell_2}^2 = \| \epsilon \|_{\ell_2}^2 = \epsilon^\top \epsilon$ is the L_2-norm of estimation of the residual term once β is specified. Ordinary least squares estimation uses the same formula, but its penalty terms are set to zero. Besides, from linear regression theories, the least squares loss function also provide consistent estimator when the error term is not normally distributed under the regularity conditions (Knight and Fu, 2000).
There are several important properties that a "good" penalized least squares estimation should have. They are defined in (Fan and Li, 2001)'s paper, and called "oracle" properties. We quote the definition directly:

Definition 3.2.1. Oracle properties
Sparsity: The resulting estimator automatically sets small estimated coefficients to zero to accomplish variable selection and reduce model complexity.
Unbiasedness: The resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, to reduce model bias.
Continuity: The resulting estimator is continuous in the data to reduce instability in model prediction(Breiman et al., 1996).

One example of such PLS estimators is smoothly clipped absolute deviation/SCAD proposed in the same paper (Fan and Li, 2001) and (Antoniadis and Fan, 2001).

Theorem 3.2.1. *PLS properties*
Assume $X^\top X = n I_p$, then (3.5) reduces to the minimization of

$$\frac{1}{2n} \| Y - X\hat{\beta} \|_{\ell_2}^2 + \| \beta - \hat{\beta} \|_{\ell_2}^2 + \sum_j p_\lambda(|\beta_j|) \ ,$$

where $\hat{\beta}$ is the ordinary least squares estimation $\hat{\beta} = (X^\top X)^{-1} X^\top Y = \frac{1}{n} X^\top Y$. This leads to consider the univariate PLS problem described in formula (2.4) of (Antoniadis

and Fan, 2001):

$$\hat{\theta}(z) = arg\min_{\theta\in\mathbb{R}}\{\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)\}.$$

Here $\theta = X\beta$ *and* $z = \frac{1}{n}X^\top Y$. *Then the PLS estimator* $\hat{\theta}(z)$ *holds the oracle properties:*

Sparsity: if $\min_{t\geq 0}\{t + p'_\lambda(t)\} > 0$;

Approximate unbiasedness: if $p'_\lambda(t) = 0$ *for large* t;

Continuity: if and only if $arg\min_{t\geq 0}\{t + p'_\lambda(t)\} = 0$.

In Table 3.1, we list some commonly used penalties terms. There are more that are produced every year, the full list could be very long.

3.3 Methods

Review of existing methods: The Accelerated Failure Time Model

The accelerated failure time (AFT) model (Cox and Oakes, 1984, Kalbfleisch and Prentice, 2011) is an important alternative to the widely used cox model (Cox, 1972) in regression analyzing of censored failure time data. It provides a direct interpretation of linear relationship between logarithm of failure time and covariates. As shown in the introduction section, it has a (log) linear form:

$$\log(T_i) = X_i^\top\beta + \epsilon_i, i = 1,\ldots,n, X_i \in \mathbf{R}^p$$

Assume the censoring time variable is C_i, then for each $i = 1,\ldots,n$, we only observe a combination $(Z_i, \delta_i,\ X_i)$. Here $\delta_i = I_{T_i\leq C_i}$ is the censoring indicator, $Z_i = \log\min(T_i,\ C_i)$.

As shown in (3.4), AFT model has a more direct way of interpreting coefficients comparing to Cox model in terms of quantification of (log transformation of) survival time instead of the relative risk and hazard rates.

There are several approaches that solve the AFT model. In (Prentice, 1978); (A. Tsiatis, 1990) among many others, one approach considers the weighted log-rank statistics to solve the problem. The rank based estimator is the solution to estima-

Table 3.1: Existing variable selection methods.

Model/Algorithms	**Authors**	**Penalty term**: $\sum_j p(\lvert\beta_j\rvert;\lambda)$
Ridge Rgression	(Frank and Friedman, 1993)	$\lambda\beta_j^2$
LASSO	(Tibshirani, 1996)	$\lambda\lvert\beta_j\rvert$
Elastic Net	(Zou and Hastie, 2005)	$\lambda_1\lvert\beta_j\rvert + \lambda_2\beta_j^2$
OSCAR	(Bondell and Reich, 2008)	$\sum\lvert\beta_j\rvert + \lambda\sum_{j<k}\max_{j<k}(\lvert\beta_j\rvert, \lvert\beta_k\rvert)$
Grouped LASSO	Friedman et al. (2010)	$\lambda\sum_{j<i}\lVert B_{i,j} - B_{j,i}\rVert_{\ell_2}$. Here the problem is $\hat{X}_{n\times p} = X_{n\times p}B_{p\times p}$.
Adaptive LASSO	(Zou, 2006)	$\lvert w_i\beta_j\rvert$
L2-Boosting	(Buehlmann, 2006)	Boosting method of LSE.
SCAD	(Fan and Li, 2001)	$\lambda^2 - (\lvert\beta_j\rvert - \lambda)^2 I(\lvert\beta_j\rvert < \lambda)$
Min-Max Convex (MCP)	(Zhang et al., 2010)	$\lambda\int_0^{\lvert\beta_j\rvert}(1 - v/(r\lambda))_+ dv$
Dantzig-Selector	(Candes and Tao, 2007)	solves min $\lVert\beta\rVert_{\ell_2}$ s.t. $\lVert X^T(y - X^T\beta)\rVert_{\ell\infty} \leq (1 + t^{-1})\sqrt{2\log p}\sigma$.
More		

tion equation of (Jin et al., 2003) shown in (3.6):

$$\sum_{i=1}^{n} \delta_i \phi(Z_i - X_i^T \beta)[X_i - \bar{X}_i(Z_i - X_i^T \beta)] = 0 \tag{3.6}$$

, where $\bar{X}_i(Z_i - X_i^T \beta)$ is the average of explanatory variable X_j, such that $Z_j - X_j^T \beta \geq Z_i - X_i^T \beta$. Different choice of $\phi(.)$ leads to different interpretations of the estimation equation.

Another approach, so called Buckley-James estimator, was discussed in (Buckley and James, 1979). With given $\beta = b$, let $e_i(b) = Z_i - X_i^\top b$ be the residual with respect to b. The Buckley-James estimator of β is the solution to the estimation equation:

$$\sum_{i=1}^{n} \Big\{ \delta_i e_i(\beta) + (1 - \delta_i) \sum_{j: e_j > e_i} \frac{e_j(\beta) \Delta \hat{F}_\beta(e_j(\beta))}{1 - \hat{F}_\beta(e_j(\beta))} \Big\} X_i = 0 \ , \tag{3.7}$$

where $\hat{F}_b$ is the non-parametric maximum likelihood estimator, or Kaplan-Meier computed from the residual and the censoring indicator: $(\delta_i, e(b))_{i=1}^n$ once b is given. Because it is for right censored data, $\hat{F}_b$ is the Kaplan-Meier estimator we discussed in Chapter 2.

It is worthy to pointing out that (3.7) is equivalent to

$$\sum_{i=1}^{n} \hat{\mathbb{E}}[e_i(\beta) | \beta, X_i, Z_i, \delta_i] X_i = 0$$

once we use a discrete distribution to estimate $\mathbb{E}[e_i(\beta) | \beta, X_i, Z_i, \delta_i]$.

Besides, (3.7) is also similar to the ordinary least squares estimator, which solves:

$$\sum_i (Y_i - X_i^\top \beta) X_i = 0.$$

To see this, we write (3.7) into the estimated conditional expectation of:

$$\hat{\mathbb{E}}[\sum_{i=1}^{n} (Z_i - X_i^\top \beta) X_i | \beta, X_i, Z_i, \delta_i X_i] = 0.$$

Hence the Buckley-James estimator is a conditional expectation version of the ordinary least squares estimator. Unfortunately, (3.7) is a implicit function about β, and there is no analytical formula to solve β.

Several authors (Miller, 1976, Buckley and James, 1979, Miller and Halpern, 1982) studied such least squares structure and they conclude that the Buckley-James estimator is more reliable (Miller and Halpern, 1982). (Ritov, 1990) and (Lai et al., 1991) modified Buckley James estimator by adding a particular smooth weight function and had developed a rigorous asymptotic theory including consistency and normality for their resulting estimator $\hat{\beta}$. They showed with such modification, " any consistent root of the BuckleyJames estimating function must be asymptotically normal and that the estimator is semi-parametrically efficient when the underlying error distribution is normal." Further, an empirical likelihood testing procedure for Buckley-James estimator was introduced by (Zhou and Li, 2008), which also extended the application of Buckley-James estimator and avoided its illusive variance estimation.

There is an EM algorithm to solve the Buckley-James estimator by updating the residual (calculating imputation step) and β (computing LSE step) sequentially. Other algorithm based on Buckley-James method has been implemented, such as hybrid methods of EM and the rank based method(ref).

On the other hands, other than the low dimensional setting, more and more researches have been concerning high dimensional problems in regression models, which is key to many research fields such as genome-wide associated study (GWAS), personalized medicine, data science and so on. One of the basic problem is to estimate the coefficients in linear model when the number of variables p is considerable large comparing to the number of observations n and leads to singularity in either rank based or Buckley-James estimator.

To solve this high-dimensional problem, there are a few studies. For example, (Huang et al., 2006), (Cai et al., 2009) and (Hu and Chai, 2013) applied LASSO and MCP to (3.6) directly; (Johnson, 2009) also discussed such rank-based estimation with ℓ_1-penalty term only with application to integrated analyses of clinical predictors and gene expression data; (Schmid and Hothorn, 2008) and (Liu et al., 2010) used kernel

based methods to ultra-high dimensional AFT model in the framework of boosting algorithm; (Li et al., 2014) introduced Dantzig Selector, which is different from the least squares setting. Due to the problem complexity, all the model are hard to use and lack rigorous proof. It is not clear if they have oracle property.

If we go deep into the topic, we could find further interesting phenomenon. Essentially, there is a contradiction between unbiasedness and sparseness in the Buckley-James estimator. On one side, in order to calculate the conditional expectation (3.7), an unbiased estimator $\hat{F}_\beta$ is needed. On the other side, PLS derived biased $\hat{\beta}$ in the high dimensional problem. The iteration steps used in the Buckley-James method indeed requires a proposed estimator $\hat{\beta}$ not only to be sparse but also unbiased. This contradiction is fundamentally rooted in each iteration of Buckley-estimation, and due to there is no closed form of the likelihood function. In consequence, a natural way is to use parametric AFT model instead.

In statistical teachings, people often contrast the parametric statistical methodology with non-parametric one, as if they are totally unrelated methods. Yet, if we study closely the development of the efficient estimation theory for nonparametric/semi-parametric models (Pfanzagl, 2012, Begun et al., 1983, Bickel et al., 1998). we will find the method proposed is based on the idea of parametric approaches, with obvious/necessary extension when needed. The tangent space and projection of score function are two such examples.

Besides, parametric models with a growing number of nuisance parameters may become (or getting very close to) a nonparametric model when sample size increases (Zhou, Chen). Therefore, understand thoroughly the mechanism of how parametric model estimation work, often lead to insight into the nonparametric methods.

We therefore shall study in this chapter the parametric AFT model with high dimensional covariates and the use of penalized MLE procedure. This is a worthy research topic of its own, since the nonparametric AFT model with high dimensional covariates is such a hard problem that we have not seen any established theory for its estimation, only a few proposed methods with various question marks and without theory can be found so far. Our hope is to be able to add more nuisance parameters to the

distribution of the error term ϵ in the future, so that this parametric approach will lead to clues for the fully nonparametric AFT models. At a minimal this parametric AFT model will be a competitive alternative, where theory is easier to obtain.
In this chapter, we stick on the discussion on the algorithms to solve high dimensional AFT in parametric settings and a theorem to prove the asymptotic properties of proposed estimator $\hat{\beta}$.

Parametric AFT in High-dimensional Settings

We see the closed form of likelihood plays a critical role in high-dimensional AFT model: 1. the essential problem of AFT model in the high-dimensional setting is that the likelihood is an implicit function about β; 2. on the other hand, the success of Cox model in high dimensional problem relies on the closed form of the partial likelihood. In this section, we propose a parametric approach that has a closed form of the likelihood function and can easily solve the high-dimensional AFT estimation problem through penalized least squares method.
Accordingly, we assume residual in (3.3) is from a particular family of distribution, i.e.

$$\epsilon_i \overset{i.i.d}{\sim} F \in \mathcal{F} \tag{3.8}$$

The optional choices could be exponential distribution family, including Normal distribution, Exponential distribution and many others, and the general extreme value distribution (GEV) family $\mathcal{F} = \text{GEV}\{\mu, \sigma, \xi\} \triangleq e^{-[1+\xi(\frac{x-\mu}{\sigma})]^{-\frac{1}{\xi}}}$ as the cumulative density function.
To simplify, we write the (3.4) into:

$$\log(T_i) = X_i^\top \beta + \sigma\epsilon_i, \quad \sigma > 0, \tag{3.9}$$

where $\epsilon \sim F(.)$ is some "standard" distribution, e.g. standard norm distribution $N(0,1)$ and exponential distribution $\text{Exp}(\lambda = 1)$. Without loss of generality, we assume for each i, $X_i^\top \mathbb{1}_p = 0$. This avoid involving intercept term in the regression formula. We could also set $\mu = 0$, otherwise β can not be uniquely estimated as the

estimation $\hat{\beta} = \hat{\beta}(\mu, \sigma)$. Once β is estimated, it helps to impute residuals ϵ_i's.
MLE could be easily derived through numeric method in low dimensional case. It has been implemented in standard statistical softwares, such as *lifereg* in SAS and *survreg* in R. Both of them use Newton-Raphson algorithm and have potential convergence problem if the number of variables is larger or equal to the number of observations and rely on initial values. To extend this into large p small n case, i.e. the number of explanatory variables p is larger than the number of observations n, some constraints on parameters must be considered. Otherwise the information matrix is singular, and the model is undetermined. Our method focuses on the following model with a penalty term:

$$\hat{\beta}_{n,\lambda_n} = \text{argmin}_\beta \, \mathbb{P}_n \rho_{\beta,\sigma} + \lambda_n^\star \mathcal{P}(\beta) \tag{3.10}$$

or equivalently, the constrained regression form of dual problem :

$$\begin{array}{lll} \hat{\beta}_{n,\lambda_n} & = & \text{argmin}_\beta \, \mathbb{P}_n \rho_{\beta,\sigma} \\ s.t. & & \mathcal{P}(\beta) \leq \lambda_n \end{array} , \tag{3.11}$$

where the loss function $\rho_n : \mathcal{L} \to \mathbb{R}$, given (x, δ)

$$\rho(., z) \;\; = \;\; -(1-\delta) \log \left(1 - F\left(\tfrac{z - x^T\beta}{\sigma}\right)\right) - \delta F'\left(\tfrac{z-x^T\beta}{\sigma}\right) \;\; .$$

The penalty terms $\mathcal{P}(.)$ in (3.11) can be LASSO(Tibshirani et al., 1997), SCAD (Fan and Li, 2001), Adaptive Lasso (Zou, 2006), and Mini-Max (Zhang et al., 2010). To perform the simulation, *eha*, *emplik*, *porcar* and *glmnet* packages in R (R Core Team, 2013) are used. The $\mathbf{P}_n \rho_{\beta,\sigma}$ part is actually the log-likelihood:

$$\mathbf{P}_n \rho_{\beta,\sigma} = \ell_n(\beta, \sigma) = \sum_{i=1}^{n} (1-\delta_i) \log(1 - F(\epsilon_i(\beta,\sigma))) + \delta_i \log F'(\epsilon_i(\beta,\sigma)),$$

where $\epsilon_i(\beta,\sigma) = \frac{\log(t_i)-x_i^T\beta}{\sigma}$. For random variable $\sigma\epsilon$, it is easy to get:

$$\begin{cases} \text{cdf}: & F_{\sigma\epsilon}(x) = P_\epsilon(\sigma\epsilon < x) = F_\epsilon\left(\frac{x}{\sigma}\right) \\ \text{pdf}: & f_{\sigma\epsilon}(x) = \frac{dF_{\sigma\epsilon}(x)}{dx} = \frac{1}{\sigma} f_\epsilon\left(\frac{x}{\sigma}\right) \\ j-\text{th deravitive}: & F^{(j)}_{\sigma\epsilon}(x) = \frac{1}{\sigma} F^{(j)}_\epsilon\left(\frac{x}{\sigma}\right) \quad j \geq 1 \end{cases}$$

To simplify the computation, we could use multi-variate Taylors' expansion as an approximation. The partial deravitives are:

$$\begin{cases} \frac{\partial\epsilon_i(\beta,\sigma)}{\partial\beta_j} &= -\frac{x_{ij}}{\sigma} \\ \frac{\partial\epsilon_i(\beta,\sigma)}{\partial\sigma} &= -\epsilon_i(\beta,\sigma)/\sigma = -\frac{\log(t_i)-x_i^T\beta}{\sigma^2} \\ \frac{\partial\ell_n(\beta,\sigma)}{\partial\beta_j} &= \frac{1}{\sigma}\sum_{i=1}^n \left\{(1-\delta_i)\frac{f(\epsilon_i(\beta,\sigma))}{1-F(\epsilon_i(\beta,\sigma))} - \delta_i \frac{f'(\epsilon_i(\beta,\sigma))}{f(\epsilon_i(\beta,\sigma))}\right\} x_{ij} \\ \frac{\partial^2\ell_n(\beta,\sigma)}{\partial\beta_j\partial\beta_k} &= \frac{1}{\sigma}\sum_{i=1}^n \frac{\partial}{\partial\beta_k}\left\{(1-\delta_i)\frac{f(\epsilon_i(\beta,\sigma))}{1-F(\epsilon_i(\beta,\sigma))} - \delta_i \frac{f'(\epsilon_i(\beta,\sigma))}{f(\epsilon_i(\beta,\sigma))}\right\} x_{ij}x_{ik} \\ &= \frac{1}{\sigma^2}\sum_{i=1} x_{ij} \times \left\{-(1-\delta_i)\frac{f^2+(1-F)f'}{(1-F)^2} - \delta_i\frac{(f')^2-f''f}{f^2}\right\} \times x_{ik} \\ &= \frac{1}{\sigma^2}\sum_{i:\delta_i=0} x_{ij}\times a_i \times x_{ik} + \frac{1}{\sigma^2}\sum_{i:\delta_i=1} x_{ij}\times b_i \times x_{ik} \\ \cdots \end{cases}$$

Here:

$$\begin{cases} a_i = & -\frac{f^2(\epsilon_i(\beta,\sigma))+(1-F(\epsilon_i(\beta,\sigma)))f'(\epsilon_i(\beta,\sigma))}{(1-F(\epsilon_i(\beta,\sigma)))^2} \\ b_i = & -\delta_i\frac{(f'(\epsilon_i(\beta,\sigma)))^2-f''(\epsilon_i(\beta,\sigma))f(\epsilon_i(\beta,\sigma))}{f^2(\epsilon_i(\beta,\sigma))} \end{cases}$$

To simplify the notation, we could use a matrix form to present the Hessian matrix D corresponding to β:

$$D = \Delta\ell \triangleq \frac{\partial^2\ell}{\partial\beta^\top\partial\beta} = \frac{1}{\sigma^2}X^\top \begin{pmatrix} \ddots & & \\ & (1-\delta_i)a_i + \delta_i b_i & \\ & & \ddots \end{pmatrix} X \tag{3.12}$$

Given σ,by using Taylor's expansion at initial β_0, ℓ_n could be approximated in the

following way:

$$\ell_n(\beta,\sigma) = \ell_n(\beta_0,\sigma) + \nabla\ell_n^\top|_{\beta_0,\sigma}(\beta-\beta_0) + \frac{1}{2}(\beta-\beta_0)^\top \Delta\ell_n^\top|_{\beta_0,\sigma}(\beta-\beta_0) + R(\|\beta-\beta_0\|_{\ell_2}^2)\ , \tag{3.13}$$

where $R(\|\beta-\beta_0\|_{\ell_2}^2)$ is the remainder term. In the low dimensional case, i.e. $p < n$, $\text{rank}(\Delta\ell_n) = p$, we could solve (3.13) very easily through standard Newton-Raphson algorithm or iteratively reweighted least squares widely used in generalized linear regression if the ϵ is from a exponential family. But this is not true in large p small n problem. To see this, (3.12) shows $\text{rank}(\Delta\ell_n) \leq \text{rank}(X) \leq n < p$. The Hessian matrix is singular, extra penalty is needed to estimate β. Similar to (3.11), the problem we concentrate could be summarized into a constrained regression form:

$$\begin{array}{lll} \hat{\beta}_{n,\lambda_n,\sigma} & = & \text{argmax}_\beta\, \ell_n(\beta_0,\sigma) + \nabla\ell_n^\top|_{\beta_0,\sigma}(\beta-\beta_0) + \frac{1}{2}(\beta-\beta_0)^T \Delta\ell_n^\top|_{\beta_0,\sigma}(\beta-\beta_0) \\ s.t. & & \mathcal{P}(\beta) \leq \lambda_n \end{array}\ , \tag{3.14}$$

Possible penalty functions could be found in Table (3.1). In this thesis, we use elastic net and LASSO.

Under the following regular conditions, we could derive some useful properties of the parametric AFT model in high dimensional settings.

C1 $\|\beta^0\|_{\ell_0} = K < n$;

C2 With regards to $S_0 = supp\{\beta^0\}$: $\frac{1}{\sqrt{n}}\nabla^T\ell_{S_0,\beta^0} \rightsquigarrow W \sim N(0,\Sigma)$, and $\Sigma > 0$;

C3 $F \in C^3$ and $\frac{1}{n}\left(\frac{1}{\sigma^2}X^T DX\right)_{ij} \overset{a.s}{\to} C_{ij} < C_0 < \infty$, the Hessian matrix D is described in 3.12;

C4 $\lim_n \lambda_n/n^{1/2} = \lambda_0 > 0$

Lemma 3.3.1 proves the existence of $\frac{1}{n}(\Delta\ell_{\beta,\sigma})$.

Lemma 3.3.1. *Under Conditions 1-4, for any j-k-th entry of the Hessian matrix satisfies:* $\exists c_{jk} \in \mathbb{R} s.t.\ \frac{1}{n}(\Delta\ell_{\beta,\sigma}) \overset{a.s.}{\to} c_{jk}^{\beta,\sigma}$.

Lemma 3.3.2 further derives the asymptotic properties of $\frac{1}{n}(\Delta\ell_{\beta,\sigma})$.

Lemma 3.3.2. *Given* σ, $\|\beta_0\|_{\ell_0} = K < n$, *then* $\frac{1}{\sqrt{n}}\nabla\ell_{\beta_0} \rightsquigarrow N(0, \Sigma)$. Σ *is the inverse of Fisher's information matrix with regard to given index set* $\text{supp}\{\beta_0\}$.

These two lemmas could be summarized into the following theorem.

Theorem 3.3.3. *Asymptotic Properties of the parametric model*
Given σ,

$$\sqrt{n}(\hat{\beta} - \beta^0) \xrightarrow{d} \text{argmin}\, T(\eta) \ , \tag{3.15}$$

where $T(\eta) = W^{*T}\eta + \eta^T C\eta + \lambda_0 \sum_j \{\eta_j \,\text{sign}(\beta_j^0) 1_{\{\beta_j^0 \neq 0\}} + |\eta_j| 1_{\{\beta_j^0 = 0\}}\}$. *In* $T(.)$, *the* $i-j-th$ *entry of* C *is* C_{ij} *and* $W^{\star}_{\text{supp}\{\beta_0\},\text{supp}\{\beta_0\}} = W$, other entries are 0.

The proof could be found in the Appendix section.
Then we could update σ using $\hat{\beta}$, i.e. $\hat{\sigma}$ is the root to $\frac{\partial \ell_n(\beta,\sigma)}{\partial\sigma}|_{\beta=\hat{\beta}} = 0$. In most case, Newton-Raphson algorithm and other methods could be used to estimate $\hat{\sigma}$ numerically and effectively. Iteratively, we could update $\hat{\beta}$ and then $\hat{\sigma}$ repetitively until convergence criterion is met.

Tuning Parameters

One important issue of the proposed method is tuning the regulatory parameter λ. In a standard LASSO problem:

$$\text{argmin}_\beta \, \|Y_n - X_{n\times p}\beta_p\|_{\ell_2}, \quad s.t. \quad \|\beta\| \leq \lambda \ , \tag{3.16}$$

k-fold cross-validation Kohavi et al. (1995) is commonly used to determine the tuning parameter λ. It randomly splits the data into k folds with equal sample size, and repeated using one fold as the training data set to generate model and measuring the performances on the remain $k-1$ folds. The average performance is reported to evaluate the model/chose proper tuning parameters. In this part, we describe another linear searching algorithm to perform the same task.

Notations

1. q: number of independent *dummy* variables;

2. $Z_i \in \mathbb{R}^n, i \in 1, \ldots, q$: i-th dummy variable;

3. $Z = [Z_1, \ldots, Z_q] \in \mathbb{R}^{n \times q}$;

4. $\xi = (\xi_1, \ldots, \xi_q)^\top \in \mathbb{R}^q$: dummy coefficient corresponding to Z.

5. $\gamma = [\beta^\top, \xi^\top]^\top$

The standard LASSO problem is then transformed to

$$\text{argmin}_\gamma \|Y_n - [X, Z]\gamma\|_{\ell_2}, \quad s.t. \quad \|\gamma\| \leq \lambda. \tag{3.17}$$

Here we assume:

1. $X \perp Z$.

2. $\epsilon \perp Z$, which is the same as linear regression.

3. $Z \sim F_Z()$, *i.e.* dummy variables are from $F_Z()$.

Given λ, solving (3.17) could be done by standard algorithm such as L2-boosting (Friedman et al., 2000) and LARS (Efron et al., 2004) among many others. Denote the corresponding estimation of γ as $\hat{\gamma}$, or $\hat{\beta}, \hat{\xi}$ respectively. We introduce the failure ratio to help determine λ:

$$r_\lambda = \frac{\|\hat{\xi}_\lambda\|_{\ell_0}}{q}, \tag{3.18}$$

where the ℓ_0 counts the number of non zero elements. Given a series of λ, i.e. $\lambda_1, \ldots, \lambda_m$, the optimal one is:

$$\lambda^{(\text{optim})} = \text{argmin}_\lambda |r_\lambda - r_0|.$$

Here r_0 should be a prefixed ratio ranged $(0, 1)$. Different levels of r_0 lead to various performances.

There are several parts that could be tuned in this algorithm:

1. $q = q(p, n)$

2. $F_Z(.)$

3. r_0

They offer a different aspect other than tune λ alone. We conduct a simple simulation to illustrate the usage of our new tuning parameter method. We focus on the LASSO problem with sample size N = 100, number of parameter p =80, and tuning parameter q = 80 or 40, and $r_0 = .05/.1/.2$, $F_Z \sim N(0,1)$, 90% of β's are randomly set to 0. The model performance on test set (N=100) is shown in Figure 3.1, PMSE is the mean square estimation. The lower value PMSE is, the better performance the model has. The behavior of 5-fold cross-validate was shown in the model, typically it has a "U" shape pattern (Friedman et al., 2001) like we get in Figure 3.1. The red text is the value of r_0, hence there is a one to one mapping to r_0 and λ_0. The full algorithm is implemented in R on GitHub: yfyang86/optimise2. Further study are need to provide rigorous details and proof.

3.4 Simulation Study

In this section, we conduct simulation to illustrate the performance of the proposed parametric accelerated failure time model in both low and high dimensional settings. The penalized term we considered is LASSO or elastic net:

$$\lambda_1|\beta_j| + \lambda_2\beta_j^2.$$

When $\lambda_2 = 0$ then it is LASSO, and $\lambda_1 = 0$ reduces to ridge regression. Here is some notation we use to define the parameters used in the simulation.

X $X \in \mathbb{R}^{n\times p}$ is the explanatory variable, where n is the sample size, and p is the number of parameters;

β $\beta \in \mathbb{R}^p$ is the coefficient;

ϵ $\epsilon \sim F_\epsilon(.)$ is the residual term;

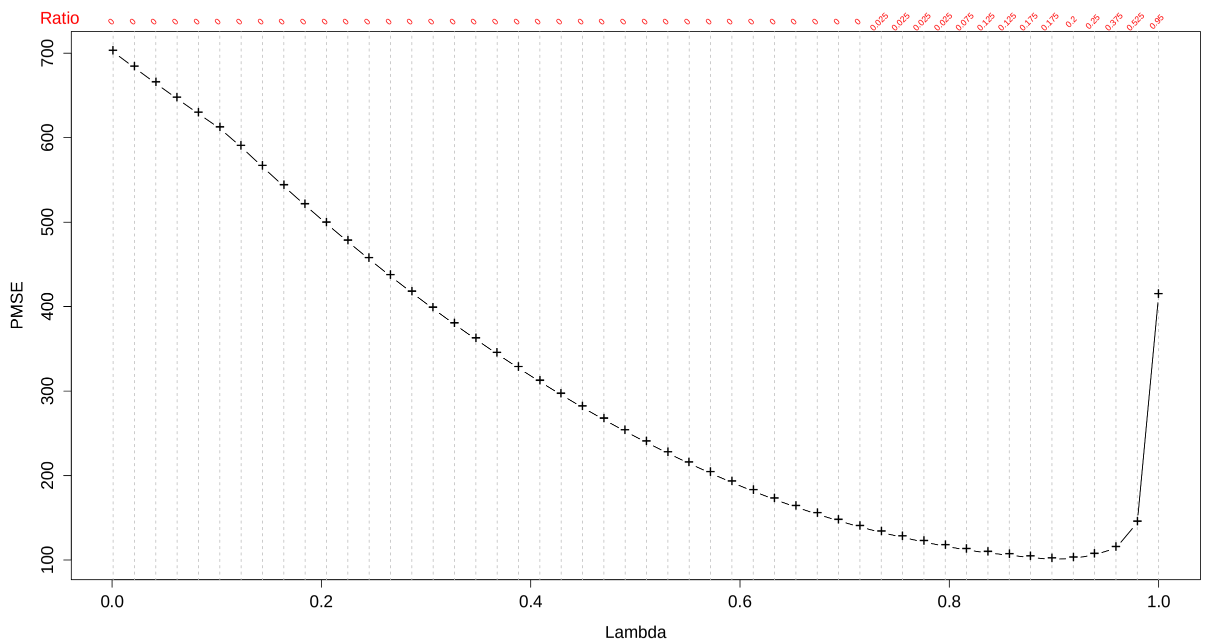


Figure 3.1: Performance of tuning parameter method on test set. The top X-axis Ratio is the value of r_0 corresponding to lambda.

C C is the censoring time, which is i.i.d. sampled from corresponding cumulative distribution function $F_C(.)$;

I_β I_β is the non zero parameters' indexes of β, i.e. $I_\beta = \#\text{supp}\{\beta\}$

Here, the model discussed is a parametric AFT model with the i-th observation (Z_i, δ_i, X_i):

$$Y_i = X_i^\top \beta + \epsilon_i \ , \ Z_i = \log \min(Y_i, C_i) \ , \ \delta_i = I(Y_i \leq C_i).$$

The simulation code is attached in the appendix. Notice, without loss of generality, we transform each column of X as $X_i^{(new)} = X_i - \frac{1}{n}\sum_i X_{ij}$. In this way, the model does not include the intercept term. i.e. in R:

```
# set X, beta and eps first
f_Xst <- function(x) t(t(x)-apply(X,2,mean))
X= f_Xst(X)
Y= X%*%beta + eps
```

Simulation case 1

In the first simulation , we assume:

1. Sample size $n = 300$, number of parameter $p = 300$;
2. $X_{ij} \sim N(0, \frac{1}{16})$;
3. $\epsilon \sim N(0,1)$ and in another case, $\log \epsilon \sim \text{Exp}(0,1)$;
4. $\beta_{I_\beta} \sim \text{Unif}(-1,1)$;
5. $\#I_\beta = 14$, and I_β is uniformly sampled from $\{1, 2, \ldots, p\}$;
6. C is sample from a combination of $Exp(1)$ and absolute value of $t_{\text{df}=10}$ distribution.

In all simulation, we assume the parametric model to be Normal, and check the performance of this normal assumption parameter AFT with different ϵ settings. The stopping rule used in our simulation is $\|\beta^{(t+1)} - \beta^{(t)}\|_{\ell^2} \leq 10^{-5}$ and number of iterations is lager than or equal to 10. 4-fold cross validation is used to tune λ_1 and

λ_2 with parallel computing. The performance of the proposed is illustrated in Figure 3.2. The figure compares the Kaplan-Meier estimation of $\{Y - X\hat{\beta}\}_{i=1}^n$ and with $\{Y - X\beta\}_{i=1}^n$. Besides, it also draws the Q-Q plot to check whether the two samples are from the same distribution. We could see for the normal assumption simulation, the results fits in a straight line which suggest the two distribution (residual and estimated residual) are very similar to each other. On the other side, if the real residual is exponential-logarithm distributed, the Q-Q plot is nearly a straight line, but the tail part has a U-shape pattern. Hence, our parametric AFT model performs robustly in this scenario.

Simulation case 2: large sample size

In the second simulation , we examine a simulation with large sample size:

1. Sample size $n = 800$, number of parameter $p = 800$;
2. $X_{ij} \sim N(0, \frac{1}{16})$;
3. $\beta_{I_\beta} \sim \text{Unif}(-1, 1)$;
4. $\#I_\beta = 19$, and I_β is uniformly sampled from $\{1, 2, \ldots, p\}$;
5. C is sampled from a combination of $Exp(1)$ and absolute value of t_{10} distribution.

In this simulation, we assume the parametric model to be Normal. The stopping rule used in our simulation is $\|\beta^{(t+1)} - \beta^{(t)}\|_{\ell^2} \leq 10^{-5}$ and number of iterations is lager than or equal to 500. The Figure 3.3 shows the performance. The red cross is the value real β, while the dash line represents the estimated $\hat{\beta}$. It is shown that the large value of β is captured by our algorithm. But since we use elastic net as the penalty term, the estimator contains about 100 non-zero $\hat{\beta}_i$.

The drawback of this method is the computation speed is slow, on a 4-cores computer platform (2.20 GHz per thread), it converges in more that one minute. To solve the elastic net problem, we used the glmnet in the package with the same name. Besides, in the appendix, we also offers approaches in LARS and L2 boosting. The package is under development and is planned to transform into C++ language.

Figure 3.2: Performance of Parametric AFT model.

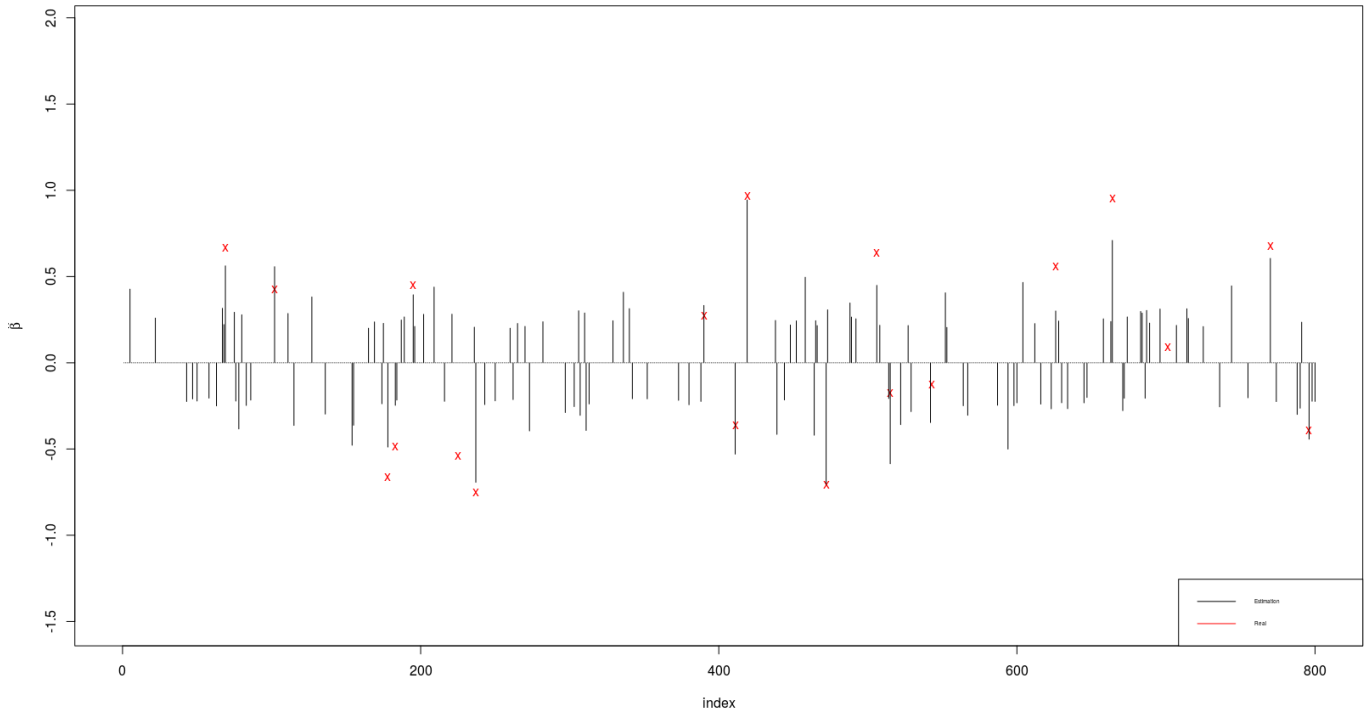


Figure 3.3: β vs. $\hat{\beta}$ of Parametric AFT model.

3.5 Conclusion, Discussion and Future Work

Conclusion and Discussion

In this chapter, we illustrate a straightforward parametric accelerated failure time algorithm to address the coefficient estimation problem of AFT model in high dimensional setting. The proposed method use the likelihood with penalty term directly:

$$\begin{aligned} \hat{\beta}_{n,\lambda_n} &= \text{argmin}_\beta \, \mathbf{P}_n \rho_{\beta,\sigma} \\ s.t. \quad & \mathcal{P}(\beta) \leq \lambda_n \end{aligned} \tag{3.19}$$

Besides, under certain assumptions, we provide a rigorous proof to show the asymptotic properties of the proposed algorithm. Further, we illustrate a new way to tune parameters in proposed model, which is later shown to be equivalent to tune λ directly. It examines the ratios that the model falsely chooses the added independent variables. By set up proper value, it has a similar performance to widely used cross-validation method.

All the algorithm is implemented in R and plan to be released as standard R packages.

Future work: More Flexible Semiparametric Model

In our model, a parametric AFT model is used and requires a pre-determined distribution of the error term. This assumption is generally considered too strong:

1. It assumes that we know the distribution of the error term. More essentially, we could write the closed form of the likelihood distribution;

2. Although we could put any possible distribution in practice, only a few of them could be easily accessed in standard software;

3. It would be hard to compare the performances of the estimations based on different distributions.

We can use distribution with several nuisance parameters instead, e.g. piecewise exponential distribution or piecewise Weibull distribution, to replace the parametric

approach. When the number of pieces involved in the model is exactly the same as the number of event points, this piecewise distribution is indeed the empirical distribution, i.e. a non-parametric approach. Hence, we name this a semi-parametric model, which greatly differs from those use spline to approximate the functions of interests. The difficulty is there is no computational efficient way to split $\mathbb{R}^p$ into disjoint regions and maximize the likelihood function simultaneously. One possible way is to mimic the coordinate descent algorithm, which update the $\hat{\beta}$ and $\hat{F}_\beta$ estimator dimension by dimension using segmented model(Davies, 1987, Muggeo, 2008). But since in each Buckley-James iteration, we solve an segmented model with penalty terms. The computational complexity is high and the model performance is unknown. Additionally, the gap between parametric model and non-parametric model is not as huge as it appears to be. There are several references [ref Royston, Parmar] [ref Carstensen] about a parametric proportional hazards regression model, with the baseline been either piece-wise exponential or some spline function. They all reach the conclusion that when the number of pieces in the piece-wise exponential, or the spline become more flexible, the parametric model will either become a Bona Fide Cox model or getting so close to a Cox model that they are practically the same.
We end up this chapter with the interesting comment from Sir Davide Cox, he himself stated that he actually preferred the parametric model: [1]

[1]Quote from Sir David Cox (Reid 1994 [10])

Reid "What do you think of the cottage industry thats grown up around [the Cox model]?"

Cox "In the light of further results one knows since, I think I would normally want to tackle the problem parametrically....Im not keen on non-parametric formulations normally."

Reid "So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasnt quite right."

Cox "Thats right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution. And if you want to do things like predict the outcome for a particular patient, its much more convenient to do that parametrically."

Chapter 4 Hypothesis Testing for Binary Choice Model

4.1 Introduction

The binary/discrete choice model (BCM) is also known as the current status model, and case I interval censored data. It describes the behave of a dichotomous output that is closely related to the commonly used logistic regression model, or the generalized linear model, which is usually defined as a binomial outcome (dependent variable) and assumes a link function (logit/probit and so on). There is an equivalent latent variable definition of the logistic regression model that matches the essential idea of BCM, where it has a usual linear model with a residual term of logistic distribution, and the observed binary outcome is whether larger than a particular value or not (Rodríguez, 2007). So, using this definition, in BCM, we are interested in a model that do not assume a distribution of the error in this thesis. We use (Wang and Zhou, 1995)'s notation in this thesis, i.e. each study subject is observed only once and the observed information is that the observation Y takes one of the two possible values (without loss of generality 0 and 1). We may introduce a latent, continuous variable $Y\star$, and define:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^\star = \beta^\top X_i + \epsilon_i > 0 \\ 0 & \end{cases} \tag{4.1}$$

Here $Y_i^\star$ is the latent response variable, $X_i \in \mathbb{R}^p$ is a p-dimensional real vector of explanatory variable. In this thesis, we assume ϵ_i's are independent and identically distributed with cumulative distribution function (CDF) F_ϵ. Here,we may think of the variable $Y_i^\star$ being either left censored or right censored. In order to make the model identifiable, we further assume F_ϵ a distribution with zero mean and finite variance, and the intercept term in the regression is always 1.

To be specific, the binary choice model is interested in "whether a decision has been made or an action carried out"(Pagan and Ullah, 1999), and can be commonly seen in

medical and biological statistics, and social science such as economics and marketing (Amemiya and Powell, 1981). Examples of binary choice model are studying the effects of drug dosage and control variables upon a unit, or whether the seats of a game is over-sold, a government bonds is issued and so on.

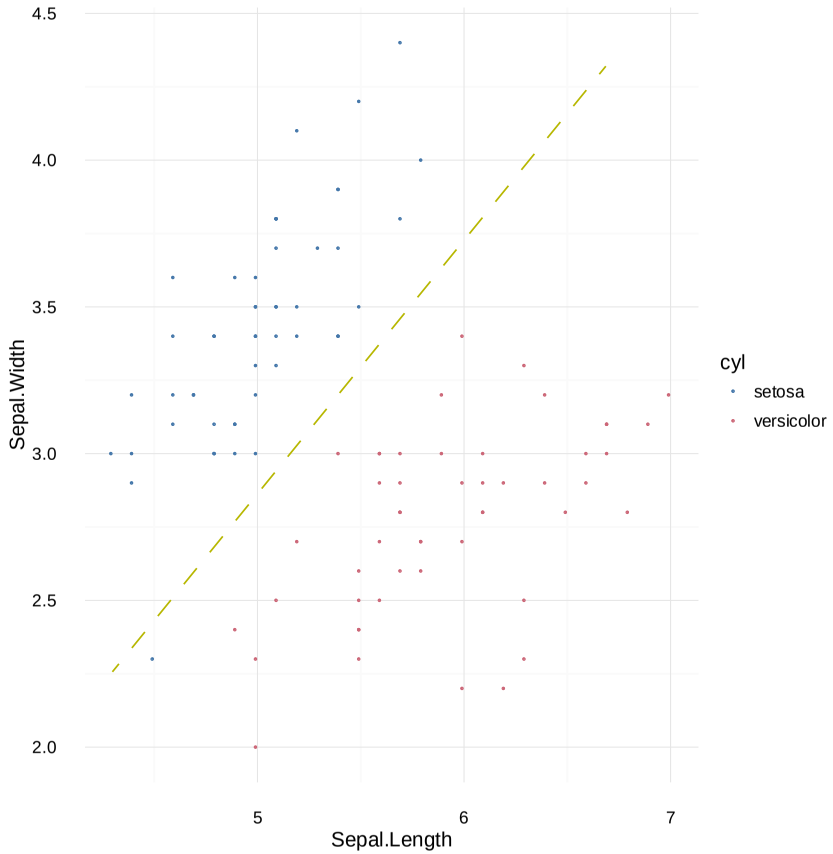
As shown in (4.1), the binary choice model is a regression model with qualitative and dichotomous response variable, i.e. taking value zero or one, rather than quantitative. We should point out that although binary choice model contains only dichotomous statuses: "a decision has been made" and "a decision has not been made", it is different from classification problem widely discussed in machine learning such as support vector machine, decision tree, neuron-network and many others, which focus on the prediction like the left side of Figure 4.1. Apparently, binary choice model reflects positive or negative decisions and contains censored status, which suggests there is a latent procedure of data generation like the right side of Figure 4.1. This differs from a simple binary outcome 0/1 that only identifies the categories.

There are a large number of published works on the properties of cumulative distribution estimator (NPMLE) of the case I interval censored data and coefficient estimation of binary choice model. Various sources are available such as (Huang and Wellner, 1997)'s review paper, and (Sun, 2007)'s book. Traditionally, there are two ways to estimate the coefficients:

1 Specify a proper link function and distribution that could be later used to estimate and test hypothesis using generalized linear model (Chambers and Cox, 1967, Han, 1987, Ichimura, 1993) or heteroscedastic non-linear model;

2 Based on empirical likelihood that holds no pre-given distributions (Wang and Zhou, 1995, Horowitz, 2009).

Either approach could be derived through maximizing likelihood function, or in another word, they are M-estimators. But recent studies have shown that popular parametric methods, such as the Probit or Logit, "can be highly misleading if the error distribution is misspecified" (Horowitz, 1992). One example is: the first derivatives of coefficients may be no longer zero and leads the estimation non consistent. To



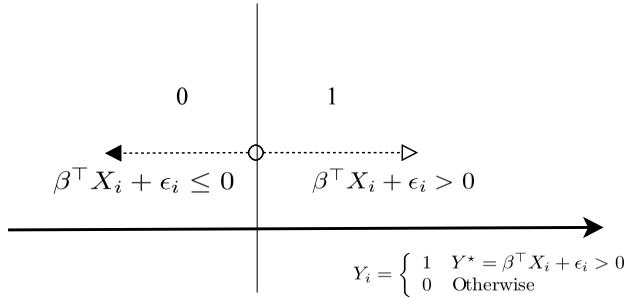


Figure 4.1: Difference between classification and binary choice model. Left: Fisher's or Anderson's iris data. Right: hidden $y^\star$ generation mechanism of binary choice model

avoid the link function specification and distributional assumptions, extensive studies have been done on semi-parametric estimation of binary choice models. For example, (Manski, 1975, 1985), (Han, 1987), (Horowitz, 1992, 2009), (Ichimura, 1993), (Sherman, 1993),(Klein and Spady, 1993), (Wang and Zhou, 1995),(Li and Racine, 2007), (Dominitz and Sherman, 2005), (Rothe, 2009) among others.

In this chapter, we proposed a semi-parametric approach to solve the hypothesis testing problems on coefficients parameters, which calculates a log-likelihood ratio statistics based on the empirical likelihood that could be fast computed through Buckle-James (B-J) like algorithms under expectation-maximization (EM)principle. The estimation is cited as the semiparametric least squares(SLS), and had been well studied in (Tanaka, 2008).

We started from (Wang and Zhou, 1995)'s iterative least squares (ILS) version, which is later proved by (Hisatoshi Tanaka, 2011) as $\sqrt{n}$-consistent. Under the null hypothesis, we used conditional least squares method similar to B-J algorithm to derive linear type constraints for survival probabilities and further applying EM algorithm or iterative convex minorant algorithm (ICM) (Pan, 1999) to maximize the empirical likelihood and get the maximum likelihood under the null hypothesis. Then, a semi-parametric approach described in (Tanaka, 2008, Hisatoshi Tanaka, 2011) was used to compute the coefficients using maximum likelihood estimations without such constraints. These steps directly construct a log-likelihood ratio statistics for the hypothesis testing problem.

The advantages of the proposed method are:

1 Flexibility. The proposed method is a semi-parametric method that does not assume the error term is from any given probability distribution.

2 Extendability. Because the proposed method uses mature techniques such as (iterative) least square estimation, empirical likelihood, EM algorithm. It has a potential to be extended to hypothesis testing problem for: a) multiple choices model(MCM) b) monotone single index models (MSIM). The former extends binary status to multiple statuses. The latter includes models such as linear re-

gression, accelerated failure time (AFT) model, transformation model, duration model and binary choice model.

3 Modularity. The proposed algorithm contains two main steps. Each step could be approached by various methods: there are different algorithms to be chosen to estimate coefficient under $H_0 \cup H_1$; there are at least two algorithms to calculate NPMLE for cumulative distribution function of the case I interval censored data.

The structure of this chapter is listed as below: section 4.2 introduces the general description of binary choice model and its NPMLE's; discusses the empirical likelihood configuration used in binary choice model and the ILS approach to derive the log-likelihood; section 4.3 illustrates simulation results; section 4.4 summarizes and discusses the semi-parametric approach of hypothesis testing and future work.

4.2 Method

The main purpose of this chapter is to study the hypothesis testing problem:

$$H_0 : \beta = b \leftrightarrow \beta \neq b$$

The method we used in this chapter is a combination of chapter 2 and 3, i.e. combination of the log empirical likelihood test and Buckley-James method. Hence, we will first introduce the existing theories and algorithms of properties of the residuals (case I interval censored data distribution) and coefficient estimation, then construct the log empirical likelihood ratio statistics with an EM algorithm similar to Buckley-James method (see Section 3.3 for details) used in AFT model.
In Buckley-James method, each iteration updates $\hat{\beta}$ with the NPMLE of the cumulative distribution function. Therefore, we focus on th basic properties algorithms of case I interval censored data first.[1] (Sun, 2007) and (Huang and Wellner) provided more general discussion on this topic.

[1]However, in our treatment we are only testing the hypothesis of $\beta = b$. So, under null hypothesis we do not need to update the beta since it is assumed to take the null value.

Basic Property of case I interval censored data

Shown in the introduction section, we notice the hypothesis testing problem is related to the case I interval censored data distribution once the coefficient is given, i.e. $\beta = b$. Now we only observe:

$$(e_i(b) = -b^\top X_i, \delta_i),$$

where $\delta_i = 1 - Y_i = I(e_i(b) + b^\top X_i \leq 0)$ indicates the residual is either left or right censored or larger than $-\beta^\top X_i$ or not. To distinguish the symbol, we use $e_i(b)$ instead of ϵ_i. By this observation, let $S_\epsilon(t) = 1 - F_\epsilon(t)$ denote the survival function of ϵ_i at time t, then the likelihood function is:

$$L = \prod_{i=1}^{n} F_\epsilon^{\delta_i}(e_i(b)) S_\epsilon^{1-\delta_i}(e_i(b)). \tag{4.2}$$

Without loss of generality, we assume $e_i = e_i(b)$, $e_1 < e_2 < \ldots < e_n$ and Y_i is ordered according to $-b^\top X_i$'s. Similar to Kaplan-Meier constraint (KMC) chapter, we can use empirical likelihood (EL) to represent the likelihood into:

$$\mathrm{EL}(p) = \prod_{i=1}^{n} \left(1 - \sum_{j:e_j \leqslant e_i} p_j\right)^{1-\delta_i} \left(\sum_{j:e_j \leqslant e_i} p_j\right)^{\delta_i}. \tag{4.3}$$

Here $p_i = F_\epsilon(e_i) - F_\epsilon(e_i-)$ is the jump at time e_i. (Robertson and Robertson, 1988) provided an isotonic regression (Ayer et al., 1955) approach to maximize (4.3):

$$\begin{aligned} &\hat{F} = \arg\min_F \sum(\delta_i - F_i)^2 \\ &s.t. \\ &\begin{cases} F_i = \sum_{j \leqslant i} p_i \\ F_1 \leq F_2 \leq \ldots \leq F_n \\ F_i \in [0,1] \end{cases} \end{aligned}$$

There are several algorithms that can solve the isotonic regression. For example (Robertson, et al., 1988) derived a closed form of the NPMLE using a max-min

formula:

$$\hat{F}_\epsilon(e_j) = \max_{u \leq j} \min_{v \geq j} \frac{\sum_{l=u}^{v} \delta_l}{\sum_u^v 1}. \tag{4.4}$$

Also, (Barlow et al, 1972) suggested adjacent violators algorithm that searches convex maximal minorant hull iteratively. (Jongbloed, 1998) proved iterative convex minorant algorithm (see Algorithm 1) also works to solve the isotonic regression problem. All these methods are implemented in the Appendix.

Algorithm 1: Pool Adjacent Violators Algorithm

Data: INPUT : $Y_i = (e_i, \delta_i)$
Step 1 Order the examination times: $e_1, \ldots, e_n$ and relabel δ_i accordingly to obtain $\delta_1, \ldots, \delta_n$
Step 2 Loops **while** $i = 1, \ldots, n$ **do**
plot(i ,$\sum_{j=1}^{i} \delta_j$)
end
;
Step 3 Form the greatest convex minorant (GCM) G* of the points in **3**.
Result: OUTPUT: $\hat{F}_n(s_i)$=left derivative of G* at i, $i = 1, \ldots, n$.

Notes

1. We should point out again, Y_i is an indicator of sign of $-b^\top X_i$ rather than the actual response variable in regression. Hence we define $e_i(b) = -b^\top X_i$, not $e_i(b) = Y_i - b^\top X_i$.

2. Similar to discussion in KMC, if there are ties, we could assume m-distinguished points $\{s_1, \ldots, s_m\}$ such that $F_\epsilon(s_1) < \ldots < F_\epsilon(s_m)$. Apparently, $m \leq n$ and the likelihood is:

$$L = \prod_{i=1}^{m} F_\epsilon^{\sum_j (1-\delta_j) I(e_j = s_i)}(e_i) S_\epsilon^{\sum_j (1-\delta_j) I(e_j = s_i)}(e_i). \tag{4.5}$$

Hence the empirical likelihood is:

$$\text{EL} = \prod_{i=1}^{m} (1 - \sum_{j: s_j \leq s_i} p_j)^{\sum_j (1-\delta_j) \times I(e_j = s_i)} (\sum_{j: s_j \leq s_i} p_j)^{\sum_j \delta_j \times I(e_j = s_i)}. \tag{4.6}$$

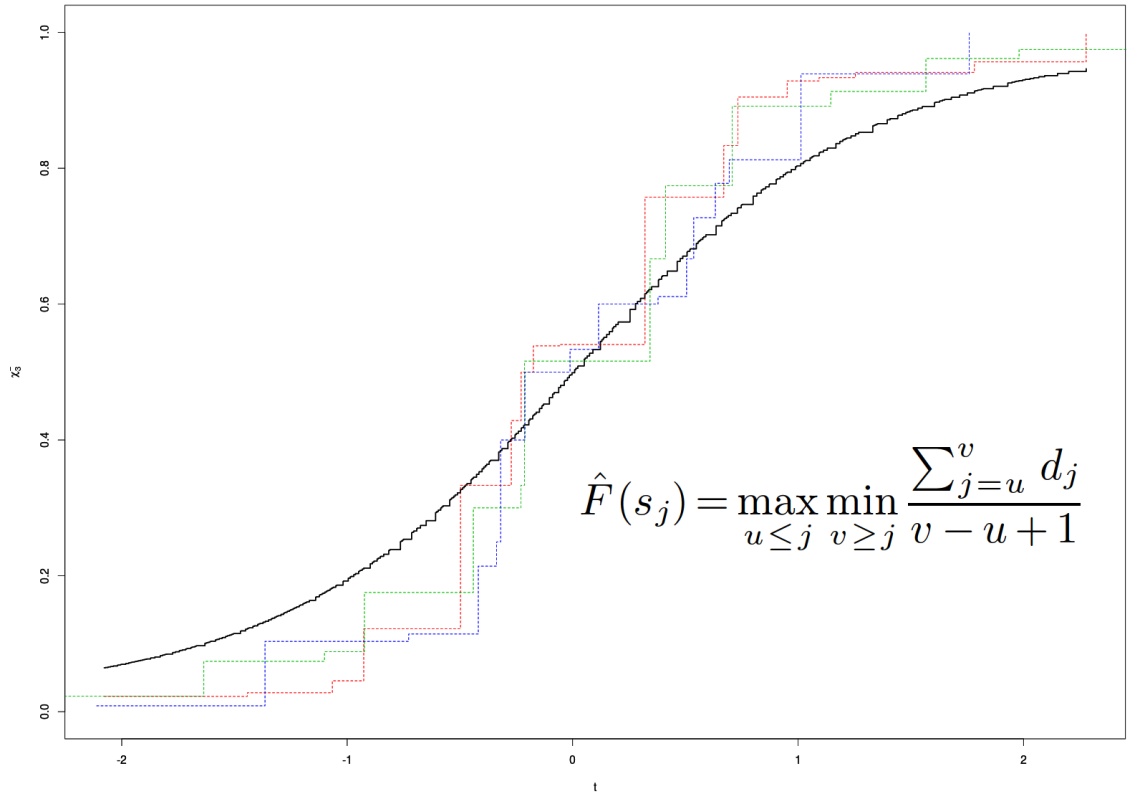


Figure 4.2: An example of NPMLE estimator. The number of jumps is $O\left(N^{\frac{1}{3}}\right)$.

In this scenario, the isotonic regression still exists, but needs to be corrected with some weight. (Sun, 2007) provides more details.

The NPMLE is much more complicated than Kaplan-Meier estimator. (Huang and Wellner, 1995) shows it is not $\sqrt{n}$-consistent, but Theorem 5.1 of the same paper proves the expectation with regards to the NPMLE $int g(t) d\hat{F}_\epsilon(t)$ for smooth function $g(t)$ is $\sqrt{n}$-consistent under some regularity conditions.

The Binary choice model and the empirical likelihood

As described in the beginning of this section, the binary choice model used in this thesis has the following linear form:

$$\begin{cases} y_i &= I(y^\star > 0) \\ y^\star &= \beta^\top x_i + \epsilon_i \end{cases}$$

Only y_i, and x_i are observed. Besides, the latent variable $y^\star$ and corresponding variable x are assumed to be independent.

To test the hypothesis, empirical likelihood ratio statistics was derived similar to the one used in Chapter 2:

$$\begin{aligned} \text{ELLR} &= -2\log\frac{\max_{p\in H_o}\text{EL}(p)}{\max_{p\in H_o\bigcup H_A}\text{EL}(p)} \\ &= -2[\log\text{EL}(\hat{p}) - \log\text{EL}(\tilde{p})] \\ &= 2[\log\text{EL}(\tilde{p}) - \log\text{EL}(\hat{p})] \end{aligned}$$

Here:

$$\tilde{p} = \arg\max_p EL_{H_o\bigcup H_A}(p)$$

is the estimation under $H_o \bigcup H_A$, and

$$\hat{p} = \arg\max_p EL_{H_o}(p),$$

is the estimation under H_o. We will discuss more details in the following two subsections on the estimation problem.

Notes

1. In order to calculate $\hat{p}$, we could use the conditional expectation as the linear constraint and EM algorithm. It is similar to (Zhou and Li, 2005)'s AFT hypothesis paper, but our problem is binary choice model with case I interval censored data.
2. ILS uses EM algorithm that involves calculating NPMLE of F_ϵ, hence we could derive $\tilde{p}$ by ILS directly.

Solve $\max_{p \in H_0 \bigcup H_A}$ EL(p): the $\tilde{p}$ estimation

One important step is to estimate the coefficient without H_0, $\beta = b$ constriant. (Wang and Zhou, 1995) published the iterative least square (ILS) estimation to solve the problem. In paper, the $\tilde{p}$ could be derived simultaneously.

The ILS method is a standard Expected and Maximization algorithm (EM), and also a special case of the semiparametric Least Squares (SLS). More discuss could be found in (Ichimura, 1993). In this section, we briefly summarize the main steps used in ILS and relate to the current study.

In (4.1), only the dichotomous output indicator Y_i is observed. Hence we could approximate the latent response variable dichotomous by a givenβ, and then update β by the least square estimator. Using the conditional expectation $E[Y^\star|X, \beta]$ as the approximation or proxy of $Y^\star$, the proposed ILS method is indeed standard EM algorithm described in Wang and Zhou (1995).

Solve $\max_{p \in H_0}$ EL(p): the $\hat{p}$ estimation

When $\beta = b \in \mathbb{R}^p$, the interval censored residual ϵ is:

$$e_i(b) = -b^\top x_i \ ,$$

with case I interval censored indicator δ_i:

$$\begin{cases} \delta_i = 1: \ Y_i = 0 \\ \delta_i = 0: \ Y_i = 1 \end{cases}$$

We further assume $e_i(b)$ is monotone increasing without tie, i.e.:

$$e_1(b) < e_2(b) < \ldots < e_n(b).$$

Otherwise, we could sort $\{e_i(b)\}$ and order $\{x_i\}$ accordingly.

Note:

1. *Since* $\{e_i(b)\}$ *us a function of* b*, we should reorder* $\{e_i, x_i, \delta_i\}$ *every time* b *is changed. Assume* $H_0 : \beta = b$*, this order will not change.*

2. In order to make the model identifiable, we need to normalize the model. Use the assumption in (Wang and Zhou, 1995), we assume $E[\epsilon] = 0$ and put an intercept "1" in the model, i.e.:

$$y_i^\star \;=\; 1 + \beta x_i + \epsilon_i.$$

Let $\hat{F}_{BCM}(t)$ be the NPMLE of F_ϵ based on the observation $(e_i(b), \delta_i)$, i.e.

$$\hat{F}_{BCM}(e_i(b), b) = \arg\max_F \prod_{i=1}^{n} F^{\delta_j}(e_i(b), b) S^{1-\delta_j}(e_i(b), b) \quad .$$

Computation Details

Similar to (Wang and Zhou, 1995), the estimation equation is (4.7) using conditional expectation:

$$\frac{1}{n}\sum_{i=1}^{n}(E[y_i^\star | b, x_i] - 1 - bx_i)x_i \;=\; 0,$$

which could be represented to:

$$\begin{aligned} 0 \;=\; & \textstyle\sum_{i=1}^{n} \delta_i x_i \sum_{j:j\le i} e_j(b) \frac{\Delta \hat{F}_{BCM}(e_j)}{\hat{F}_{BCM}(e_i)} \\ & + (1-\delta_i) x_i \textstyle\sum_{j:j>i} e_j(b) \frac{\Delta \hat{F}_{BCM}(e_j(b))}{1-\hat{F}_{BCM}(e_i(b))} \end{aligned} \tag{4.7}$$

Furthermore, there is a matrix form to simplify the notation. We can form a weight matrix $m \in \mathbb{R}^{p\times p}$, with positive entries:

$$m_{ij} = \begin{cases} \frac{\hat{F}_{BCM}(e_j(b))}{1-\hat{F}_{BCM}(e_i(b))} & \delta_i = 0; j > i \\ \frac{\hat{F}_{BCM}(e_j(b))}{\hat{F}_{BCM}(e_i(b))} & \delta_i = 1; j \leq i \\ 0 & \text{Otherwise} \end{cases}$$

In this way, the estimation equation could denote as:

$$0 = \sum_{i=1}^{n} \left((1-\delta_i) \sum_{j:j>i} e_j(b) m_{ij} + \delta_i \sum_{j:j\leq i} e_j(b) m_{ij} \right) x_i \tag{4.8}$$

We can write the Binary Choice Model estimation (4.7) according to $e_i(b)$ by doing the following calculation:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \delta_i x_i \sum_{j:j\leq i} e_j(b) \frac{\Delta \hat{F}_{BCM}(e_j)}{\hat{F}_{BCM}(e_i)} \\ &\quad + (1-\delta_i) x_i \sum_{j:j>i} e_j(b) \frac{\Delta \hat{F}_{BCM}(e_j(b))}{1-\hat{F}_{BCM}(e_i(b))} \\ \Leftrightarrow \quad 0 &= \sum_{i=1}^{n} x_i \frac{\delta_i (1-\hat{F}_{\mathrm{BCM}}(e_i)) A_i + (1-\delta_i)\hat{F}_{\mathrm{BCM}}(e_i) B_i}{\hat{F}_{\mathrm{BCM}}(e_i)(1-\hat{F}_{\mathrm{BCM}}(e_i))} \end{aligned}$$

Here $A_i = \sum_{j:j>i} e_j(b) \Delta \hat{F}_{bcm}(e_j)$ and $B_i = \sum_{j:j\leq i} e_j(b) \Delta \hat{F}_{bcm}(e_j)$.
Notice the mean of residual is 0, then

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} t \, \mathrm{d}\hat{F}_{bcm}(t) \\ \Leftrightarrow \quad 0 &= \int_{-\infty}^{e_i(b)} + \int_{e_i(b)}^{\infty} t \, \mathrm{d}\hat{F}_{bcm}(t) \quad \forall i \\ \Leftrightarrow \quad 0 &= \sum_{j:j>i} e_j(b) \Delta \hat{F}_{bcm}(e_j) + \sum_{j:j\leq i} e_j(b) \Delta \hat{F}_{bcm}(e_j) \\ 0 &= A_i + B_i \end{aligned}$$

Then

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i \frac{\delta_i (1-\hat{F}_{\mathrm{BCM}}(e_i)) A_i + (1-\delta_i)\hat{F}_{\mathrm{BCM}}(e_i) B_i}{\hat{F}_{\mathrm{BCM}}(e_i)(1-\hat{F}_{\mathrm{BCM}}(e_i))} \\ \Leftrightarrow \quad 0 &= \sum_{i=1}^{n} x_i \frac{\delta_i A_i + \hat{F}_{\mathrm{BCM}}(e_i) B_i - \delta_i \hat{F}_{\mathrm{BCM}}(e_i) A_i - \delta_i \hat{F}_{\mathrm{BCM}}(e_i) B_i}{\hat{F}_{\mathrm{BCM}}(e_i)(1-\hat{F}_{\mathrm{BCM}}(e_i))} \\ \Leftrightarrow \quad 0 &= \sum_{i=1}^{n} x_i \frac{\delta_i A_i + \hat{F}_{\mathrm{BCM}}(e_i) B_i - \delta_i \hat{F}_{\mathrm{BCM}}(e_i)(A_i + B_i)}{\hat{F}_{\mathrm{BCM}}(e_i)(1-\hat{F}_{\mathrm{BCM}}(e_i))} \\ \Leftrightarrow \quad 0 &= \sum_{i=1}^{n} x_i \frac{\delta_i A_i + \hat{F}_{\mathrm{BCM}}(e_i) B_i}{\hat{F}_{\mathrm{BCM}}(e_i)(1-\hat{F}_{\mathrm{BCM}}(e_i))} \end{aligned}$$

For cases of $\delta = 1$ or $\delta = 0$, plug in $A_i = \sum_{j:j>i} e_j(b)\Delta\hat{F}_{bcm}(e_j)$ and $B_i = \sum_{j:j\leq i} e_j(b)$, and change the summation order of i and j, we derived the Binary Choice Model estimation (4.7) according to $e_i(b)$ by doing the following calculation:

$$\sum_j e_j \left(\sum_{i:1\leq i<j,\ \delta_i=1} m_{ij}x_i + \sum_{i:n\geq i>j,\ \delta_i=0} m_{ij}x_i \right) = 0 \tag{4.9}$$

Similar to (Li and Zhou, 2002), we could derive the linear constraint according to (4.9):

$$\sum_j e_j \frac{\sum_{i:1\leq i<j,\ \delta_i=1} m_{ij}x_i + \sum_{i:n\geq i>j,\ \delta_i=0} m_{ij}x_i}{\Delta\hat{F}_{BCM}(e_j(b))} p_j = 0 \tag{4.10}$$

Note

1. In (Wang and Zhou, 1995), the authors derived another form of estimation equation, which focuses on the parameter estimation:

$$\sum_{i=1}^{n} x_i \frac{\delta_i A_i + \hat{F}_{\mathrm{BCM}}(e_i)(-A_i)}{\hat{F}_{\mathrm{BCM}}(e_i)(1-\hat{F}_{\mathrm{BCM}}(e_i))} = 0\ .$$

2. The estimation equation (4.9) is very similar to formula (5) in (Li and Zhou, 2002). Hence, the Buckley-James alike algorithm described in (Li and Zhou) has a potential to solve more hypothesis problems than AFT model. Besides, since ILS uses the same estimation equation, if $b = \hat{\beta}_{ILS}$, then $\hat{p}_i = \Delta\hat{F}_{BCM}(t, \hat{\beta}_{ILS})$ satisfies the constraint (4.10) and maximize the empirical likelihood.

The empirical likelihood for the $e_i(b)$ is defined in (4.3) as:

$$\mathrm{EL} = \prod_{i=1}^{n} \left(1 - \sum_{j:e_j\leq e_i} p_j\right)^{1-\delta_i} \left(\sum_{j:e_j\leq e_i} p_j\right)^{\delta_i}\ .$$

Then we are to find p_i's such that it :

i maximize the empirical likelihood;

ii satisfies the linear constraint (4.10).

To summarize, $\hat{p}$ could be solve by the following empirical likelihood optimization with linear constraint problem:

$$\begin{aligned} &\arg\max_p \mathrm{EL}(p) \\ &\text{s. t. :} \\ &\begin{cases} \text{formula (4.10)} \\ \sum p_i = 1 \\ p_i \geq 0 \end{cases} \end{aligned} \tag{4.11}$$

Compute the constraint EL

The right censored data empirical likelihood optimization with linear constraint problem (4.11) is discussed in our first chapter, but unfortunately, the proposed KMC algorithm does not hold for case I interval censored data, hence could be applied to this chapter. But there is a modified EM algorithm proposed in (Zhou, 2012) could solve such problem:

Algorithm 2: Empirical likelihood ratio with arbitrarily censored data by EM algorithm

Data: Initial : $Y_i = (C_i, \delta_i)$ and assume the "uncensored" time points are $X_1, \ldots, X_m$;

E-Step Given F, compute the weight w_j at location t_j

$$w_j = \sum_i E[I_{X_i = t_j} | X_i, \delta_i]$$

M-Step Maximize $\sum_i w_i \log P_i$ with proper linear constraints.

Run E-step and M-step until converge.

More details could be found in (Zhou, 2016). Notice in our empirical likelihood, there is no "uncensored" time point, otherwise it is the same as the empirical likelihood in (Zhou, 2012).

Numeric problem

In this thesis, the R function el.test.wt2 is used to solve the follow M-step:

$$\text{argmax}_{p_i} \sum_i \omega_i \log p_i \quad s.t. \begin{cases} \sum_i p_i x_i = \mu \\ \sum_i p_i = 1 \\ \forall i : \ p_i \geq 0 \end{cases}$$

It has the same numeric problem as we discussed in chapter one. We need to check if there is a positive solution for the following constraint condition before we do each iteration:

$$\begin{cases} \sum_i p_i x_i = \mu \\ \sum_i p_i = 1 \end{cases}$$

4.3 Simulation

In this section, two simulation were reported to illustrate the χ^2 approximation of proposed log empirical likelihood ratio test statistics for hypothesis problem in binary choice model. ILS and EM approach is used in both simulation. Due to there is only $n^{\frac{1}{3}}$ jumps, the sample size is chosen as 3,000 and each simulation has been repeated in 1,000 times.

The isotonic regression, without nay with tie, is solved by the max-min formula in C++ and wrapped as a R function (see Appendix), E-M algorithm is an extension of el.cen.EM function in **emplik** package as the constraint used in this thesis is linear type. A standalone R package is on working and will be release with KMC as a future work.

Hypothesis for single explanatory variable

The first example is a simple binary choice model with only one explanatory variable. with the following setting:

$$Y = \begin{cases} 1 & \text{if } 1 + \beta_i X_1 + \epsilon > 0 \\ 0 & \text{if } 1 + \beta_i X_1 + \epsilon \leq 0 \end{cases}$$

Here, we use a random design:

i $X_1 \sim N(1,1)$;

ii $\epsilon \sim \frac{1}{\sqrt{3}} T_{df=3}$

iii $\beta_i = 2$

Notice that the intercept is set as 1 to avoid identification issue, and the error term is $t_{df=3}$ distributed, which means its mean is 0 and variance is finite.
There is $E[I(3 + 2 \times X + \epsilon) > 0] \approx 91.54\%$ of Y is 1. Here $X \sim N(0,1)$, and $\epsilon \sim t_3$. The simulation result is shown in a quantile-to-quantile (QQ) plot comparing with χ^2 distribution with degree of freedom 1. Because ILS algorithm (Wang and Zhou 1995) was used to calculate max $\mathrm{EL}_{H_0 \cup H_1}$ in this approach, the same starting value of β and stopping criterion is applied, i.e.

i Use logistic regression (with logit link function) to give a initial value of β;

ii Since there is only one explanatory variable, $|\hat{\beta}_{\text{step t+1}} - \hat{\beta}_{\text{step t}}| \leq 10^{-4}$ is used as the stopping criterion.

The QQ plot is illustrated in Figure (4.3), which draws the quantiles of the simulated test statistics (X axis)against the quantiles of the $\chi^2_{\text{df=1}}$ distribution (Y axis). In the figure, a 45-degree reference line is also plotted and the dots fall approximately along this line, which suggest they come from a population with the same distribution.

Hypothesis for multiple explanatory variables

The second example extend the first one into binary choice model with multiple explanatory variable. The parameters setting are as follows:

$$Y = 1 + \beta_1 \times X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 \times X_4 + \epsilon,$$

and

1. $X_1 \sim N(1,1)$, $X_2 \sim N(1,1)$, $X_3 \sim N(0,1.2)$, $X_4 \sim N(0,.8)$

2. $\epsilon \sim \frac{1}{\sqrt{3}} T_{df=3}$

3. $\beta_1 = 2$, $\beta_2 = 1$, $\beta_3 = 1$, $\beta_4 = 0.6$.

Choices of the distribution of X and ϵ could vary, but they show similar result. For multiple explanatory variables, we use ℓ_2-norm instead of absolute value to set up the stopping criterion:

$$\sqrt{\sum_i \hat{\beta}_{i,\text{step t+1}} - \hat{\beta}_{i,\text{step t}}} \leq 10^{-4}$$

In this setting, around 63.96% of Y's are 1.

Similar to the first simulation, the QQ plot is illustrated in Figure (4.3). In this QQ plot, the Y axis is the quantiles of the $\chi^2_{\text{df}=4}$ distribution. Again, the dots fall approximately along the 45-degree reference line, which suggest they come from a population with the same distribution.

Notes

In this simulation, we use both max-min formula and pool adjacent algorithm to calculate the NPMLE for cumulative distribution function of the case I interval censored data. But there is no difference among the two. In the thesis, the result of max-min formula one is presented with implement in corresponding section in the Appendix chapter.

4.4 Conclusion, discussion and future work

Summary

In this chapter, we discussed the hypothesis testing problem in binary choice model. Log empirical likelihood ratio statistics was used. The proposed method contains two steps:

1 Compute NPMLE of cumulative distribution function under $H_0 \cup H_1$ and plug it into the empirical likelihood;

2 Maximize the empirical likelihood under H_0.

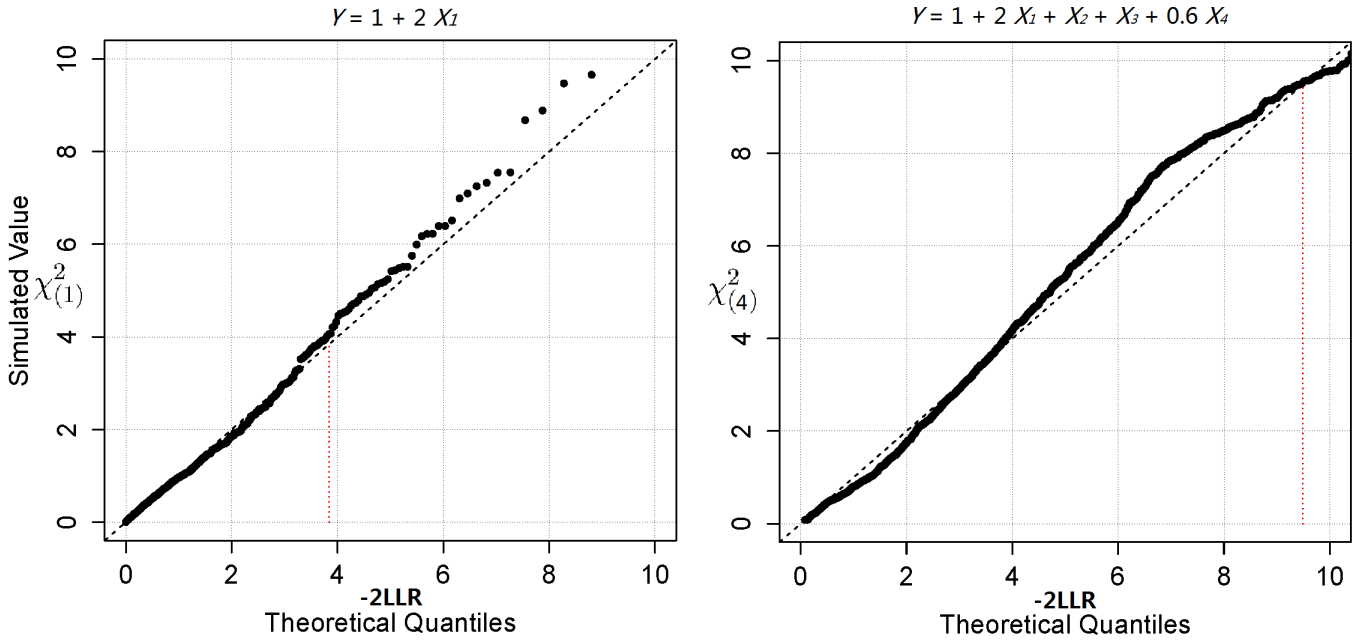


Figure 4.3: Simulation result: Quatile-Qunatile plot.

(Wang and Zhou)'s iterative least square estimation for binary choice model is used to solve the first step above. A Buckley-James alike algorithm was proposed to compute the maximum empirical likelihood under $\boldsymbol{H_0}$. Here are some highlights of this algorithm:

1. using least square estimation and conditional expectations to derive the estimation equation;

2. represent the estimation equation as a linear constraint on jumps of cumulative distribution function;

3. use EM algorithm to maximize the empirical likelihood function, which simplified the problem into iteratively solving a weighted log sum with linear constraint.

Advantages of the proposed algorithm could summarized into a few key words: assembly, flexibility, and extendability :

1. Modularity. The proposed algorithm contains two main steps, each step could be approached by various methods: there are different algorithms to be chosen to estimate coefficient under $\boldsymbol{H_0} \cup \boldsymbol{H_1}$, and at least two to calculate NPMLE for cumulative distribution function of the case I interval censored data.

2. Flexibility. The proposed method is a semi-parametric method that does not assume the error term is from any given probability distribution.

3. Extendability. Because the proposed method used mature techniques such as (iterative) least square estimation, empirical likelihood, EM algorithm. It has a potential could be extend to hypothesis testing problem a) multiple choices model b) monotone single index models.

The drawbacks of the proposed method are 1. computation speed is slow 2. need large value of observations.

Future works

In the thesis, we illustrated the proposed log empirical likelihood ratio test statistics is approximately χ^2 distributed through the Q-Q plots. Hence one direction is to provide a rigorous proof of the asymptotic properties of the log empirical likelihood ratio test statistics. Unlike AFT model or other survival analysis problem, the cumulative distribution function is only $n^{\frac{1}{3}}$-consistent, but the mean estimate, i.e. $\int g(d)\mathrm{d}F(t)$, using this $\hat{p}_{\mathrm{NPMLE}}$ is still $\sqrt{n}$-consistent, which could give us direction to finish the proof.

Meanwhile, since the proposed method use coefficient estimation and empirical likelihood separately, we see possibility that we could extend it into multi choice model. But the formula of empirical likelihood would become very complex in this case.

Last but not least, further research are encouraged to study the properties of the log empirical likelihood ratio and extend it to cover more scenarios. (Zhou, 2005) and (Zhou, 2015) discussed the EM algorithm very carefully, and we could see possibilities the EM algorithm that maximizes the empirical likelihood, and log empirical likelihood ratio statistics to become a standard procedure to handle many kinds of survival regression estimation/hypothesis testing problems.

1. AFT model;
2. Right/Left/doubly censored data;
3. Extension of Cox model such as Yang-Prentice model (Yang and Prentice, 2010) using empirical likelihood approach;
4. Binary/multiple choice model;
5. More.

Appendix

KMC package

The KMC algorithm described in this manuscript is available on `https://github.com/yfyang86/kmc` and `cran.r-project.org/web/packages/kmc/`. The standard CRAN version could be installed in R directly but lacks features than the GitHub one. All source code are following GPL-3 license.

R code: KMC Real data example

The speed advantage of KMC algorithm could be used in time consuming analysis such as drawing contour plot. In this real data example, we illustrate the proposed algorithm to analyze the Stanford heart transplants program described in (Miller 1982). There were 157 patients who received transplants collected in the data, among which 55 were still alive and 102 were deceased. Besides, the survival time were scaled by 365.25. We could draw a contour plot of intercept and slope for a AFT model.

```
LL= 50
beta0 <- 3.52016
beta1 <- -0.01973458 #-0.0185
beta.grid <- function(x0,range,n0,type="sq",u=5){
n0 = as.double(n0)
if (type=="sq"){
o1 <- c(
-range*(u*(n0:1)^2)/(u*n0^2),0,
range*(u*(1:n0)^2)/(u*n0^2)
)
}else{
if (type=='sqrt'){
o1 <- c(
```

```
-range*(u*sqrt(n0:1))/(u*sqrt(n0)),0,
range*(u*sqrt(1:n0))/(u*sqrt(n0)))
}else{
o1=c(
-range*(n0:1)/n0,
0,
range*(1:n0)/n0
)
}
}
return(
x0+o1
);
}

beta.0 <- beta.grid(beta0,0.05,LL,"l")
beta.1 <- beta.grid(beta1,.00151,LL,"l")#0.00051

set.seed(1234)
y=log10(stanford5$time)+runif(152)/1000

d <- stanford5$status

oy = order(y,-d)
d=d[oy]
y=y[oy]
x=cbind(1,stanford5$age)[oy,]

ZZ=matrix(0,2*LL+1,2*LL+1)

```

```
44 library(kmc)
45 tic=0
46 for(jj in 1:(2*LL+1)){
47 for(ii in 1:(2*LL+1)){
48 beta=c(beta.0[ii],beta.1[jj])
49 ZZ[jj,ii]=kmc.bjtest(y,d,x=x,beta=beta,init.st="naive")$"
       -2LLR"
50 }
51 }
52 ZZ2<-ZZ
53 ZZ[ZZ<0]=NA ## when KMC.BJTEST fails to converge, it'll
       return a negative value.
54
55 range(ZZ,finite=T) -> zlim
56 floor.d<-function(x,n=4){floor(x*10^n)/(10^n)}
57
58 postscript("C:/Temp/Fig2_1.eps",width=7,height=7)
59 contour(
60 y=beta.0,
61 x=beta.1,
62 ZZ,
63 zlim=c(0,.17),
64 levels=unique(floor.d(
65 beta.grid(x0=mean(zlim),range=diff(zlim)/2,n0=15,type="
       sqrt",u=10),
66 4)),
67 ylab="Intercept",
68 xlab=expression(beta[Age])
69 )
```

Proof to Lemma 3.3.1

Proof. The j-k-th entry is $\frac{1}{\sigma^2}\sum_{i:\delta_i=0} x_{ij}\times a_i \times x_{ik} + \frac{1}{\sigma^2}\sum_{i:\delta_i=1} x_{ij}\times b_i \times x_{ik}$, then

$$\begin{aligned}
\text{LHS} &= \frac{1}{n}\frac{1}{\sigma^2}\sum_{i=1} X_{ij}\times\{(1-\delta_i)a_i+\delta_i b_i\}\times X_{ik} \\
&= \frac{1}{\sigma^2}\times\frac{n_1}{n}\times\frac{1}{n_1}\sum_{\delta_i=1} b_i X_{ij}X_{ik} \\
&\quad +\frac{1}{\sigma^2}\times\frac{n_0}{n}\times\frac{1}{n_0}\sum_{\delta_i=0} a_i X_{ij}X_{ik} \\
&\overset{n\to\infty}{\to} \frac{1}{\sigma^2}\times\alpha\times\frac{1}{n_1}\sum_{\delta_i=1} b_i X_{ij}X_{ik} \\
&\quad +\frac{1}{\sigma^2}\times(1-\alpha)\times\frac{1}{n_0}\sum_{\delta_i=0} a_i X_{ij}X_{ik}
\end{aligned}$$

, where $n_1 = \sum\delta_i$, and $n_0 = n - n_1$.

Assume random variable $\xi_{jk}^{(0)} = X_j X_k \frac{f^2\left(\frac{y-X^T\beta}{\sigma}\right)+\left(1-F\left(\frac{y-X^T\beta}{\sigma}\right)\right)f'\left(\frac{y-X^T\beta}{\sigma}\right)}{\left(1-F\left(\frac{y-X^T\beta}{\sigma}\right)\right)^2}$, and $\xi_{jk}^{(1)} = \frac{\left(f'\left(\frac{y-X^T\beta}{\sigma}\right)\right)^2 - f''\left(\frac{y-X^T\beta}{\sigma}\right)f\left(\frac{y-X^T\beta}{\sigma}\right)}{f^2\left(\frac{y-X^T\beta}{\sigma}\right)}$, where X_i is the i-th element of random vector X.then $\max\{E[|\xi_{jk}^{(0)}|], E[|\xi_{jk}^{(1)}|]\} < M_2M_4$.

$$\begin{aligned}
\lim_n \text{LHS} &= \lim_n \alpha\mathbb{P}_n(\xi_{jk}^{(0)}) + (1-\alpha)\lim_n \mathbb{P}_n(\xi_{jk}^{(1)}) \\
&= \alpha E[\xi_{jk}^{(0)}] + (1-\alpha)E[\xi_{jk}^{(1)}] \\
&\overset{\Delta}{=} c_{jk}
\end{aligned}$$

□

Proof to Lemma 3.3.2

Proof. Assume we use LASSO-type penalty, then the minimizing problem is

$$\hat{\beta}_{n,\lambda_n,\sigma,\beta_0} = \text{argmin}_\beta \frac{1}{n}\|(\Delta\ell\beta - \nabla\ell)\Big|_{\beta=\beta_0} - \Delta\ell|_{\beta_0}\beta\|_{\ell_2}^2 + \frac{\lambda_n}{n}\|\beta|_{\ell_1}$$

Let

$$S_n(\beta) = \ell_n(\beta_0,\sigma) + \nabla\ell_n^T|_{\beta_0,\sigma}(\beta-\beta_0) + \frac{1}{2}(\beta-\beta_0)^\top\Delta\ell_n^\top|_{\beta_0,\sigma}(\beta-\beta_0)$$

, and $T_n(\eta) = n\left[S_n\left(\beta^0 + \frac{\eta}{n^{1/2}}\right) - S_n(\beta^0)\right]$, then

$$\begin{aligned} T_n(\eta) &= \left(S_n\left(\beta^0 + \tfrac{\eta}{n^{-1/2}}\right) - S_n(\beta^0)\right) + \lambda_n\left(\|\beta^0 + \tfrac{\eta}{n^{1/2}}\|_{\ell_1} - \|\beta^0\|_{\ell_1}\right) \\ &= \tfrac{1}{\sqrt{n}}\nabla\ell_{0,n}^\top\eta + \tfrac{1}{2n}\eta^\top\Delta\ell_{0,n}\eta + \lambda_n\left(\|\beta^0 + \tfrac{\eta}{n^{1/2}}\|_{\ell_1} - \|\beta^0\|_{\ell_1}\right) \end{aligned}$$

Under certain assumptions, $F \in C^3 \Rightarrow \frac{1}{\sqrt{n}}[\nabla\ell_{\beta_0}]_{\text{supp}\{\beta_0\}} \rightsquigarrow W \sim N_K(0, \Sigma)$.
If $\lim_n \lambda_n/n^{1/2} = \lambda_0 > 0 \Rightarrow$

$$\begin{aligned} & \lim_n \lambda_n\left(\|\beta^0 + \tfrac{\eta}{n^{1/2}}\|_{\ell_1} - \|\beta^0\|_{\ell_1}\right) \\ = \; & \lim_n \|\lambda_n\beta^0 + \lambda_0\eta\|_{\ell_1} - \|\lambda_n\beta^0\|_{\ell_1} \\ = \; & \lambda_0\textstyle\sum_j \eta_j \operatorname{sign}(\beta_j^0)1_{\{\beta_j^0\neq 0\}} + |\eta_j|1_{\{\beta_j^0=0\}} \quad \because \lambda_n \to \infty \end{aligned}$$

Hence by argmax version continuous mapping theory:

$$\sqrt{n}(\hat{\beta} - \beta^0) \xrightarrow{d} \operatorname{argmin} T(\eta)$$

, where $T(\eta) = W^{\star\top}\eta + \eta^\top C\eta + \lambda_0\sum_j\{\eta_j \operatorname{sign}(\beta_j^0)1_{\{\beta_j^0\neq 0\}} + |\eta_j|1_{\{\beta_j^0=0\}}\}$, and $W^\star_{\text{supp}\{\beta_0\},\text{supp}\{\beta_0\}} = W$, other entries are 0. Hence we derive the following asymptotic properties: □

```
1 ##########################
2 #### PARAMETRIC AFT ####
3 ##########################
4 library(survival)
5 library(emplik)
6 library(parcor)
7 require(foreach)
8 require(doParallel)
9 library(Runuran)
10 # paftoga and optimise2 could be find on my GitHub page:
```

```
# yfyang86
# The code is very long, I won't paste here.

library(paftoga)
library(optimise2)


rgumbel<-function(n){
distr <- udgumbel()
gen <- pinvd.new(distr)
x <- ur(gen,n)
x
}

set.seed(1234)
p = 100 # number of parameters
n = 200 # sample size
p.zero= p-ceiling(log(n)^1.5)# 600-32 # N(0)
ST=1; # simulation settings

using.glmnet=T
using.oga=F
using.inteceptstragtegy=F

cores<- 8
cl <- makeCluster(cores-1, methods=FALSE)
registerDoParallel(cl)
scaless=rep(0,ST)
for (bigsimu in 1:ST){
X=matrix(rnorm(p*n)/4,ncol=p);
```

```
41 beta=4*runif(p)-2
42 beta=(beta+0.1*sign(beta))
43 zero.loc=sample(1:p,p.zero);
44 #beta[zero.loc]=runif(p.zero,-.002,.002)
45 beta[zero.loc]=0
46 sigma=2
47 #eps= - sigma*log(rexp(n)) # exp: std Gumbel * sigma
48 eps= - sigma*rnorm(n) # exp: std Gumbel * sigma
49 f_Xst <- function(x) t(t(x)-apply(X,2,mean))
50 X= f_Xst(X)
51 Y= X%*%beta + eps
52
53 cen=rexp(n,rate=1)*2+abs(rt(n=n,df=10))*2
54 YC=apply(cbind(cen,Y),1,min)
55 delta=as.double( YC == Y);
56
57 sigma0=1;
58 scale=2;
59 Y=Y/scale
60 YC=YC/scale
61 beta0=rep(.2,p);
62 beta.real<-beta/scale
63
64 #
65 locate.nonzero<-function(x){(1:length(x))[abs(x)>1e-10]}
66 plotbeta.paftoga <- function(beta,beta.real){
67 plot((1:p)[abs(beta.real)>0],beta.real[abs(beta.real)>0],
      col=2,xlim=c(0,p),pch='x',ylim=2*range(beta.real),
68 xlab="index",
69 ylab=expression(hat(beta))
```

```
70 )
71 points(beta,type='h')
72 legend('bottomright',col=1:2,legend =c('Estimation','Real'
       ),lty=1,cex=.4)
73 }
74 #plotbeta.paftoga(beta=beta.real,beta.real=beta.real)
75
76 summarybeta.paftoga<-function(a,b){
77 a.ind=which(abs(a)>1e-8);
78 b.ind=which(abs(b)>1e-8);
79 list(joint=sum(a.ind%in%b.ind),betahat=length(a.ind),beta0
       =length(b.ind))
80 }
81
82 observartion.index = 1:n
83 ob.index   = observartion.index[delta==1]
84 cen.index = observartion.index[delta==0]
85
86 beta.current =beta0
87 sigma.current=sigma0
88 X.inner=X
89 f.eb   <-function(b){as.double(YC-X.inner%*%b)}
90 # F(x), F'(x) ,... Standard form!
91 f.S <-function(eb,sigma0){x=eb/sigma;
92 1-pnorm(q=x)}
93 f.f <-function(eb,sigma0,log.t=F){x=eb/sigma0;
94 if (!log.t) {
95 return(dnorm(x=x));
96 }else{
97 return(dnorm(x=x,log=T));
```

```
98 }
99 }
100 f.df<-function(eb,sigma0){x=eb/sigma0;
101 -x*dnorm(x=x)
102 }
103 f.ddf<-function(eb,sigma0){x=eb/sigma0;
104 dnorm(x=x)*(x*x-1)
105 }
106
107 ###ITERATION
108
109 f.ell.devs <- function(b,sigma){# return dev and Hessian
110 var.eb <- f.eb(b=b);
111 var.f.ddf <- f.ddf(var.eb,sigma);
112 var.f.df <- f.df(var.eb,sigma);
113 var.f.f <- f.f(var.eb,sigma);
114 var.f.S <- f.S(var.eb,sigma);
115 # vecterized
116 #ell.firstdev = rep(0,p)
117 ell = sum(
118 log(var.f.S[cen.index]) ) + sum(llog(var.f.f[ob.index],1e
    -10))
119 ell.firstdev = as.double(matrix(var.f.f[cen.index]/var.f.
    S[cen.index],nrow=1)%*%X.inner[cen.index,])/sigma
120 ell.firstdev = ell.firstdev - as.double(matrix(var.f.df[
    ob.index]/var.f.f[ob.index],nrow=1)%*%X.inner[ob.index
    ,])/sigma
121 tmp.vec = rep(0,n)
122 tmp.vec[cen.index] = -as.double(
123 (var.f.df[cen.index]*var.f.S[cen.index]+
```

```
124 var.f.f[cen.index]^2)/(var.f.S[cen.index])^2
125 )/sigma^2
126 tmp.vec[ob.index]          =  as.double(
127 (var.f.ddf[ob.index]*var.f.f[ob.index]-
128 var.f.df[ob.index]^2)/(var.f.f[ob.index])^2
129 )/sigma^2
130 ell.Hessian                =  sign(tmp.vec[1])*tcrossprod(t(X
        .inner)%*%diag(sqrt(abs(tmp.vec))))
131 #   t(X)%*%diag(tmp.vec)%*%X
132 # NOTICE:  t(X*tmp.vec)%*%X= t(X)%*%diag(tmp.vec)%*%X,
        which is slower in the latter case?
133 # crossprod()???
134 return(
135 list(
136 lik=ell,
137 dev=ell.firstdev,
138 Hessian=ell.Hessian
139 )
140 );
141 }
142
143 f.ell.sigma.solve <- function (sigma0,b){
144 var.eb = as.double(YC-X.inner%*%b)
145 fsigma.f0 <- Vectorize(function(sigma){
146 var.f.f.loged <-  f.f(eb=var.eb,sigma0=sigma,log.t=T);
147 var.f.S <-  f.S(eb=var.eb,sigma0=sigma);
148 # vecterized
149 ell           =  sum(log(var.f.S[cen.index])) + sum(var.f.f
        .loged[ob.index])-log(sigma)*length(ob.index)
150 ell
```

```
151 })
152 optimize(interval=c(0.01,80),f=fsigma.f0,maximum = T)[[1]]
153 }
154
155 comm.lasso<-function (X, y, k = 4,use.Gram=TRUE,normalize=
        TRUE,offset=NULL)
156 {
157 n<-length(y)
158 all.folds <- split(sample(1:n),rep(1:k,length=n))
159
160 if (use.Gram==TRUE){
161 type="covariance"
162 }
163 if (use.Gram==FALSE){
164 type="naive"
165 }
166
167 globalfit<-glmnet(X,y,offset=offset,family="gaussian",
        alpha=1,standardize=normalize,type.gaussian=type)
168 lambda<-c(1,globalfit$lambda)
169
170 re<-cv.glmnet(X,y,offset=offset,lambda=lambda,family='
        gaussian',
171 #type.measure="mae",
172 parallel=T,alpha=1,
173 nfolds=k)
174
175 lambda.opt<- .1*re$lambda.1se + .9*re$lambda.min
176 cat("\t LAMmin:\t ",re$lambda.min)
```

```
177 coefficients=predict(globalfit,type="coefficients",s=
        lambda.opt)
178
179 intercept=coefficients[1]
180 coefficients=coefficients[-1]
181 names(coefficients)=1:ncol(X)
182 object <- list(lambda=lambda,lambda.opt=lambda.opt,cv.
        lasso=re,intercept.lasso=intercept,coefficients.lasso=
        coefficients)
183 return(object);
184 }
185
186 ITERATIONS=300;
187 simu.report=rep(0,ITERATIONS);
188 beta.update=rep(0,p)
189 #beta.current=c(1,beta+runif(p)/20)
190 sigma.current=1
191 testing.lars=F
192 beta.current[1] = mean(YC)
193 if (p < 1) {
194 beta.current = coefficients(survreg(Surv(exp(YC),delta)~X)
        );
195 beta0          = beta.current
196 }
197 #####begin_of_iteration#####
198 ii.flag=0;
199 iter=1
200 while (iter <ITERATIONS){
201 flag=T;
202 while(flag){ # GUESS A SOLUTION
```

```
203 tmp.reg =  f.ell.devs(beta.current,sigma.current)
204 if ((sum(is.na(tmp.reg$dev))==0) && (sum(is.na(tmp.reg$
    Hessian))==0) ) break
205 ii.flag=ii.flag+1
206 if (ii.flag>20) stop('not coverge!')
207 beta.current = beta0+runif(n=p+1,min=-1,max=1)/20;
208 beta.current[1] = 0
209 cat('Tring initial:\t',ii.flag,'Reseting\n
    ----------\n')
210 iter=1
211 }
212
213 if(p>(n/10)){
214 #update beta
215 if(using.glmnet){
216 if (!using.inteceptstragtegy){
217 b.update.re =  comm.lasso(X=tmp.reg$Hessian,y=tmp.reg$
    Hessian%*%beta.current-tmp.reg$dev,k=cores);
218 #b.glmnet.re = glmnet::cv.glmnet(x=tmp.reg$Hessian,y=tmp.
    reg$Hessian%*%beta.current-tmp.reg$dev,parallel=T,
    offset=rep(F,p+1))
219 beta.update  =b.update.re$coefficients.lasso
220 }else{
221 XXXX=tmp.reg$Hessian[,-1]
222 XXXX=t(t(XXXX)-colMeans(XXXX))
223 XXXX.scaler = sqrt(colMeans(XXXX*XXXX))
224 YYYY=tmp.reg$Hessian%*%beta.current-tmp.reg$dev
225 YYYY= YYYY-beta.update[1]
226 b.update.re =  comm.lasso(X=XXXX,y=YYYY,k=cores);
227 beta.update = b.update.re$coefficients.lasso
```

```
#* (XXXX.scaler^2)*4
}

}

if (using.oga){
oga(tmp.reg$Hessian%*%beta.current-tmp.reg$dev,tmp.reg$
    Hessian,k=20) -> rere
beta.update=rere$beta
beta.update[1]=rere$alpha

}

if(testing.lars){b.update.re.lars = lars::lars(x=tmp.reg$
    Hessian,
y=tmp.reg$Hessian%*%beta.current-tmp.reg$dev,
type="stepwise"
)}
# or use PGA-OGA
# uniroot upate sigma


}else{ # p<<n
beta.update=as.double(solve(tmp.reg$Hessian,tmp.reg$
    Hessian%*%beta.current-tmp.reg$dev))
}
sigma.update=f.ell.sigma.solve(sigma.current,beta.update)
#if (iter> 10) sigma.update=1.25
simu.report[iter]=mean((beta.update-beta.current)^2)+(
    sigma.current-sigma.update)^2
```

```
254
255 beta.current  <- beta.update
256 sigma.current <- sigma.update
257 cat('\nITER\t',iter,'ERROR:\t',simu.report[iter],'\tSigma:
      ',sigma.update,'\talpha:',beta.current[1],'\tK:\t',sum(
    abs(beta.current)>0));
258 if (sum(abs(beta.current)>0)<5 & iter==(ITERATIONS-1))
      iter=iter-1
259 iter=iter+1
260 if(simu.report[iter]<1e-4 & iter>10) break;
261 }
262 if (iter==ITERATIONS) cat('\nMay not converge! Hit max
      iteration!')
263 #####end_of_iteration#####
264 #  plotbeta.paftoga(beta.current,c(1/scale,beta.real))
265 sigma.current -> scaless[bigsimu]
266 par(mfrow=c(1,2))
267 plot(survfit(Surv(exp(as.double(YC-X.inner%*%beta.current)
    ),delta)~1))
268 lines(survfit(Surv(exp(as.double(YC-X.inner%*%beta.real)),
    delta)~1),col=2)
269 legend("topright",legend=c('PAFT','real'),col=1:2,lty=1,
    lwd=1.5)
270 plot(sort(as.double(YC-X.inner%*%beta.current)),sort(as.
    double(YC-X.inner%*%beta.real)),
271 xlab="Estimated residual",ylab="Real residual")
272 abline(0,1,col=2,lty=2)
273 par(mfrow=c(1,1))
274 plotbeta.paftoga(beta.current,c(beta.real))
275 print(summarybeta.paftoga(beta.current,c(beta.real)))
```

```
}
stopCluster(cl)
```

Proof to Proposition ??

Isotonic regression: max-min formula

This version offers a way to solve the weighted isotonic regression, i.e. case I interval censored data CDF's NPMLE with ties. R has a standard function isoreg to solve the problem without tie.

```
//File: isot.cpp
//Author: Yifan Yang
//Time: 2015-01
//License: GPL-2
// include R.h
#include <R.h>


extern "C" {
  void isot_C(
  int *y,
  int *x,
  int *L,
  double *ps,
  int *J,
  int *ms
){
int n=0[L];
int j;
int starting=0;

```

```
22 //int *ms = new int[n];
23 double *ys = new double[n];
24 double *ns =  new double[n];
25
26 ms[0]=1;
27 ys[0]=y[0];
28 ns[0]=x[0];
29
30 j=1;
31
32 //Start iteration
33 for (int i=1;i<n;i++){
34   j++;//TODO length of ys in Line 1
35   ys[j-1] = y[i];
36   ns[j-1] = x[i];
37   ms[j-1] = 1;
38   //Line 1
39   for (int jj = 0;jj<j;jj++){
40     ps[jj] = (double)ys[jj]/(double)ns[jj];
41   }
42
43
44   while ( (j > 0) && (ps[j-2]>ps[j-1])){
45     ys[j-2] += ys[j-1] ;
46     ns[j-2] += ns[j-1] ;
47     ms[j-2] += ms[j-1] ;
48     for (int jj = 0;jj<j;jj++){ps[jj] = (double)ys[jj]/(
    double)ns[jj];}
49     j--;
50   }
```

```
51
52    }
53    0[J]=j++;
54    delete[] ys,ns;
55    }
56 }
```

```
1 dyn.load("isot.so")
2
3 isoreg_yifan <- function(n,d,tie=T){
4    if(tie){
5    L=length(n)
6    re=numeric(L)
7    J=integer(1)
8    n=as.integer(n)
9    d=as.integer(d)
10   ms= integer(L)
11   re3=.C("isot_C",d,n,L,re,J,ms);
12   LL=re3[[5]]
13   rep(re3[[4]][1:LL],re3[[6]][1:LL]) -> Fdist
14   Fdist[Fdist< 10* .Machine$double.eps] = 0;
15   return(Fdist);
16   }else{
17   return(isoreg(d)$yf);
18   }
19 }
```

Bibliography

Amemiya, T. and Powell, J. L. (1981). A comparison of the box-cox maximum likelihood estimator and the non-linear two-stage least squares estimator. *Journal of Econometrics*, 17(3):351–381.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W., Silverman, E., et al. (1955). An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, 26(4):641–647.

Begun, J. M., Hall, W., Huang, W.-M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, pages 432–452.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66(3):429–436.

Buehlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, pages 559–583.

Cai, T., Huang, J., and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics*, 65(2):394–404.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

Chambers, E. A. and Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika*, 54(3-4):573–578.

Chen, K. and Zhou, M. (2007). Computation of the empirical likelihood ratio from censored data. *Journal of Statistical Computation and Simulation*, 77(12):1033–1042.

Cox, D. R. (1972). Regression models and life tables. *JR stat soc B*, 34(2):187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, volume 21. CRC Press.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1):33–43.

Dines, L. L. (1926). On positive solutions of a system of linear equations. *Annals of Mathematics*, pages 386–392.

Dominitz, J. and Sherman, R. P. (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory*, 21(04):838–863.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Technical report, Stanford University.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316.

Hisatoshi Tanaka, T. H. (2011). Consistency of the iterative least squares estimator for the binary choice model with varying coefficients.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica: journal of the Econometric Society*, pages 505–531.

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer.

Hu, J. W. and Chai, H. (2013). Adjusted regularized estimation in the accelerated failure time model with high dimensional covariates. *Journal of Multivariate Analysis*, 122:96–114.

Huang, J., Ma, S. G., and Xie, H. L. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62(3):813–820.

Huang, J. and Wellner, J. A. (1995). Asymptotic normality of the npmle of linear functionals for interval censored data, case 1. *Statistica Neerlandica*, 49(2):153–163.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 123–169. Springer.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120.

Jin, Z., Lin, D., Wei, L., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353.

Johnson, B. A. (2009). Rank-based estimation in the l(1)-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics*, 10(4):659–666.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7(3):310–321.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145.

Lai, T. L., Ying, Z., et al. (1991). Large sample theory of a modified buckley-james estimator for regression analysis with censored data. *The Annals of Statistics*, 19(3):1370–1402.

Li, G. (1995). On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statistics & Probability Letters*, 25(2):95–104.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Li, Y., Dicker, L., and Zhao, S. D. (2014). The dantzig selector for censored linear regression models. *Statistica Sinica*, 24(1):251.

Liu, Z., Chen, D., Tan, M., Jiang, F., and Gartenhaus, R. B. (2010). Kernel based methods for accelerated failure time model with ultra-high dimensional data. *BMC bioinformatics*, 11(1):606.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, 27(3):313–333.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika*, 69(3):521–531.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, 63(3):449–464.

Muggeo, V. M. (2008). Segmented: an r package to fit regression models with broken-line relationships. *R news*, 8(1):20–25.

Murphy, S. A. and van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *The Annals of Statistics*, pages 1471–1509.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. B. (2001). *Empirical likelihood*. CRC press.

Pagan, A. and Ullah, A. (1999). *Nonparametric econometrics*. Cambridge university press.

Pan, W. (1999). Extending the iterative convex minorant algorithm to the cox model for interval-censored data. *Journal of Computational and Graphical Statistics*, 8(1):109–120.

Pan, X.-R. and Zhou, M. (1999). Using one-parameter sub-family of distributions in empirical likelihood ratio with censored data. *Journal of statistical planning and inference*, 75(2):379–392.

Pfanzagl, J. (2012). *Contributions to a general asymptotic statistical theory*, volume 13. Springer Science & Business Media.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, 65(1):167–179.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics*, pages 303–328.

Robertson, T. and Robertson, T. (1988). Order restricted statistical inference. Technical report.

Rodríguez, G. (2007). Lecture notes on generalized linear models.

Rothe, C. (2009). Semiparametric estimation of binary response models with endogenous regressors. *Journal of Econometrics*, 153(1):51–64.

Royston, P. and Parmar, M. K. (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology*, 13(1):152.

Schmid, M. and Hothorn, T. (2008). Flexible boosting of accelerated failure time models. *BMC bioinformatics*, 9(1):269.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, pages 123–137.

Sun, J. (2007). *The statistical analysis of interval-censored failure time data*. Springer Science & Business Media.

Tanaka, H. (2008). Semiparametric least squares estimation of monotone single index models and its application to the iterative least squares estimation of binary choice models. Technical report, Citeseer.

Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70(352):865–871.

Tian, L., Zhao, L., and Wei, L. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15(2):222–233.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. et al. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Wang, W. and Zhou, M. (1995). Iterative least squares estimator of binary choice models: a semi-parametric approach. *On line*.

Yang, S. and Prentice, R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, 66(1):30–38.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhou, M. (2005). Empirical likelihood ratio with arbitrarily censored/truncated data by em algorithm. *Journal of Computational and Graphical Statistics*, 14(3).

Zhou, M. (2010). A wilks theorem for the censored empirical likelihood of means.

Zhou, M. (2012). Empirical likelihood ratio with arbitrarily censored/truncated data by em algorithm. *Journal of Computational and Graphical Statistics*.

Zhou, M. (2015). *Empirical likelihood method in survival analysis*, volume 79. CRC Press.

Zhou, M. and Li, G. (2008). Empirical likelihood analysis of the buckley–james estimator. *Journal of multivariate analysis*, 99(4):649–664.

Zhou, M. and Yang, Y. (2016). Constrained kaplan-meier curve and empirical likelihood. *submitted*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Vita

Education

PhD student in Statistics: University of Kentucky,**2011 to Present**

MS in Statistics: University of Science and Technology of China, **2008 to 2011**

BS in Statistics: University of Science and Technology of China, **2004 to 2008**

Professional publications in University of Kentucky

1. Zhou, M., **Yang, Y.** (2015). A Recursive Formula for the Kaplan-Meier Estimator with Mean Constraints. *Computational Statistics*, 30.4 (2015): 1097-1109.
2. Zhu, S., Yang, Y., Zhou, M. (2015). A Note on Empirical Likelihood Inference on the Hazards Ratio with Non-proportional Hazards. *Biometrics* 71.3 (2015): 859-563.
3. Zhang,Y., Yang, Y. , Xu, B., Zang, Q., and et al, & Shi, Q (2016). IsomiR Bank: A research resource for tracking isomiRs, *Bioinformatics* (accepted)
4. Hua, J., Xu, B., Yang, Y., Ban, R., Iqbal, F., Cooke, H. J., ... & Shi, Q. (2015). Follicle Online: an integrated database of follicle assembly, development and ovulation. *Database*, 2015, bav036
5. Zhang, Y., Zhong, L., Xu, B., Yang, Y., Ban, R., Zhu, J., et al. & Shi, Q. (2013). SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. *Nucleic acids research*,41(D1), D1055-D1062.
6. Zhang, Y., Xu, B., **Yang, Y.**, Ban, R., Zhang, H., Jiang, X., et al. & Shi, Q. (2012). CPSS: a computational platform for the analysis of small RNA deep sequencing data. *Bioinformatics*, 28(14), 1925-1927.
7. Yuanwei Zhang, Qiguang Zang, Huan Zhang, **Yifan Yang**, Rongjun Ban, Furhan Iqbal, Ao Li, and Qinghua Shi (05-May-2016), Nucleic Acids Research, DeAnnIso: a tool for online detection and annotation of isomiRs from small RNA sequencing data
8. Zhang H, Wheeler W, Hyland PL, Yang Y, Shi J, Chatterjee N, Yu K. PLoS Genet. 2016 Jun 30;12(6):e1006122. doi: 10.1371/journal.pgen.1006122. eCollection 2016. A Powerful Procedure for Pathway-Based Meta-analysis Using Summary Statistics Identifies 43 Pathways Associated with Type II Diabetes in European Populations
9. Han Zhang, Colin O. Wu, Yifan Yang, Sonja I. Berndt, Stephen J. Chanock,Kai Yu Statistical Methods in Medical Research, 2016-07-11, A multi-locus genetic association test for a dichotomous trait and its secondary phenotype