Empirical Likelihood Quantile Regression for Right-Censored Data

by

Shimeng Huang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Quantile estimation of time-to-event data plays a key role in many medical applications, especially conditional on covariates of interest. In such settings, bias due to model misspecification is an important concern. As such, Empirical Likelihood (EL) is a particularly attractive estimation approach, making minimal parametric modeling assumptions without unduly compromising statistical efficiency. However, observed survival times are typically subject to right-censoring, in which case most EL approaches cannot be applied directly. In this thesis, we revisit a widely-applicable Expectation-Maximization (EM) algorithm for right-censored EL. As the covariate-free EL function becomes discontinuous in the conditional setting, we propose a continuity correction for which the computational properties of EM are retained. Several approaches to obtaining confidence intervals are explored. We provide an implementation of our method and related algorithms in the R package `flexEL`. The source code is written in C++ for high computational performance, and a straightforward interface allows users to fit arbitrary EL models with little programming effort.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Martin Lysy, who made me understand that taking small and careful steps is crucial to achieving any goal. Working with Professor Martin Lysy has motivated me to always be ready and be curious to discover and tackle new challenges that come along the way, and to keep improving since there is always room to do so. I am extremely grateful for his guidance, patience, and support throughout my Master's program.

I would also like to thank all professors in our Statistics and Actuarial Science department. Thank you to Professor Pengfei Li and Professor Changbao Wu for their outstanding teaching as well as valuable advice on this thesis. Thank you to Professor Richard Cook for the conference and event opportunities.

Furthermore, I would like to thank Professor Henry Wolkowicz who first introduced me to academic research and guided me through my first research project when I was an undergraduate student, and has been a great mentor for me since then.

Moreover, I would like to thank our department coordinators Mary Lou Dufton and Lisa Baxter who constantly provide help to our graduate students. I would also like to thank Robyn Landers from MFCF who has answered me various questions regarding the computing services.

In addition, I would like to thank all of my friends and fellow students for their help and company, and because of them, my life as a graduate student has been a memorable and pleasant journey.

Last but not least, I would like to thank my father for his unconditioned and endless love, support and encouragement.

Dedication

To my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In many medical applications, the interest resides in the extreme quantiles of the survival times distributions rather than the expected survival times. In this case, quantile regression introduced by Basset and Koenker (1978) should be applied as opposed to the common mean regression.

The situation becomes more complicated when the values of the dependent variable are not fully observed. As an example, consider we observe the lifetimes of n patients $\{y_1, \cdots, y_n\}$, and for each of $i \in \{1, \cdots, n\}$, we also observe a set of measurements $\boldsymbol{x}_i \in \mathbb{R}^d$. Further, we know that not all y_i's are true lifetimes, some of the patients dropped out of the study and the corresponding y_i's are the times when they dropped out. These lifetimes are called right-censored, one of the length-bias issues that commonly occur in medical studies.

Some desired properties when conducting inference in this situation include minimal modeling assumptions, high statistical efficiency, the ease of obtaining confidence intervals, and the ability to work with right-censored data.

Cox regression (Cox, 1992) is one popular model which works for right-censoring data, however, the proportional hazards assumption may not hold in many situations. An alternative approach is the accelerated failure time (AFT) model, which has a more direct interpretation compared to the Cox regression model. Moreover, the Buckley-James estimator (Buckley and James, 1979) is a semi-parametric AFT model which gives promising results for right-censored data. Wang et al. (2015) propose a heteroscedastic AFT model for right-censored survival data, which is computationally simple but relies on parametric assumptions of the error distribution.

A more attractive approach is the empirical likelihood approach, which is a flexible framework suitable for different kinds of regression models and makes only moment assumptions but no distribution assumption, and has been shown to enjoy many statistical properties. The empirical likelihood approach is the focus of this thesis.

1.2 Empirical Likelihood

The empirical likelihood (EL) approach can be traced back to Thomas and Grunkemeier (1975). Its current framework is mainly developed by Owen (1988, 1990, 1991), where empirical likelihood ratio statistics is introduced, and the EL method is extended to linear regression models under fixed or random design. Kolaczyk (1994) further generalize the method to be used with generalized linear models. Qin and Lawless (1994) relate estimating equation and empirical likelihood and provide asymptotic properties of the estimator.

A Bayesian approach to EL considers the pseudo-posterior distribution $p_{\mathrm{EL}}(\boldsymbol{\theta}|Y) \propto \mathrm{EL}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ where $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, is usually straightforward to explore by Markov chain Monte Carlo (MCMC) algorithm. However, notice that since EL is not a true likelihood, neither is $p_{\mathrm{EL}}(\boldsymbol{\theta}|Y)$ a true posterior. The consequences of this have been investigated by e.g. Lazar (2003). Chaudhuri et al. (2017) considers using Hamiltonian Monte Carlo sampling for the Bayesian EL models.

EL approach generally requires a convex hall condition, which means that a solution may not exist if this condition is not satisfied. Chen et al. (2008) propose an adjustment to the constraints in the EL framework to ensure a solution always exist, and the theoretical properties are not affected.

An approach related to EL is the so-called exponentially tilting (ET) method (Efron, 1981). Schennach (2005, 2007) proposes the exponentially tilted empirical likelihood (ELET) approach, which enjoys the properties of both ET and EL methods. Newey and Smith (2004) also gives the theoretical results relating Generalized Method of Moment (GMM) and Generalized Empirical Likelihood (GEL), their higher order properties, as well as their bias-corrected forms in the absence of length-bias.

For length-biased data with EL, Zhou (2005) proposes an EM algorithm for censored an truncated data under mean type constraints without covariates. Zhou and Li (2008) combine the empirical likelihood with the Buckley-James estimator which works for regression models. Zhou et al. (2012) revisit the fixed and random design linear regression models but for right-censored data and show that the model works well even with heteroscedastic errors. Shen et al. (2016) develop a different EM algorithm under the EL framework for one- or two- sample doubly censored data.

The construction of confidence regions or intervals under the EL frameworks has been mainly discussed when there is no length-bias. In this case, an asymptotic χ^2 distribu-

tion of log EL is valid. When right-censoring is present, the asymptotic distribution is no longer a standard χ^2 distribution but subject to an unknown scaling factor. Some approaches have been proposed with modifications of the estimating equations under the EL framework. For example, He et al. (2016) consider using a special influence functions in the estimating equations to retain a standard χ^2 distribution. Li and Wang (2003) propose an adjusted EL for linear regression using synthetic data approach. They extend the EL method for inference on a linear combination of the coefficients and also incorporate auxiliary information on the covariates. Ning et al. (2013) consider length-biased right-censored data in a non-regression setting for the estimation of mean, quantile and survival function of the population as well as confidence intervals.

1.3 Quantile Regression

Quantile regression is originated by Basset and Koenker (1978). After the first paper based on a location model, the authors further consider a location-scale model and the consistency of the estimator is derived (Koenker and Bassett, 1982). Kocherginsky et al. (2005) propose a Markov chain marginal bootstrap approach for the confidence intervals of regression quantiles.

Yang and He (2012) introduce a Bayesian EL method which is able to estimate multiple quantile levels at the same time, and using prior on the parameters to leverage the bias and variance trade-off of estimating multiple quantile levels simultaneously. Lancaster and Jae Jun (2010) develop a Bayesian exponentially tilted empirical likelihood approach for quantile regressions. Noh and Lee (2016) propose a quantile regression location-scale model for heteroscedastic time series models.

For right-censored data, Reich and Smith (2013) considers a Bayesian quantile regres-

sion under a semi-parametric location-scale model, which is a linear combination of basis functions. The model jointly estimates multiple quantile levels, but may not be suitable for extreme quantiles and is computationally expensive. A kernel estimator under a similar setting is developed by Heuchenne and Van Keilegom (2010).

1.4 Contribution

For EL under right-censoring, we extend an existing EM algorithm to work with various regression problems. A primary challenge of estimation, in this case, is that the log EL is no longer continuous in most of the parameters, which causes difficulties in the optimization of log EL.

We propose a continuity correction to retain the smoothness of the objective function under right-censoring, so that direct optimization of log EL becomes possible. The same idea can also help with the non-smoothness introduced by the quantile regression constraint and together allows the problem to be solved computationally efficiently without losing statistical efficiency. We verify the correctness of the algorithm after the continuity correction.

We design a computationally efficient R package called `flexEL` with source code in C++, which is flexible enough for users to solve any type of regression problems with minimum programming effort. Other than our proposed methods, the package also includes various mean and quantile regressions for both right-censored and uncensored data, as well as other related algorithms to help with the computation of EL.

1.5 Outline

In the following chapters, we first introduce our location-scale model for mean and quantile regressions and derive the estimating equations for the parameters. Then we describe the EL framework for data without length-bias. After that, we discuss the case where right-censoring is present and an EM algorithm is extended. We then describe in detail our continuity correction for quantile regression and right-censored EL. We demonstrate our approaches and compare with other existing methods using simulated data. We conclude and discuss future works in the last chapter.

Chapter 2

Model and Estimating Equations

In this chapter, we first introduce our semi-parametric location-scale regression model, then we derive the estimating equations for both mean and quantile regressions.

2.1 Semi-Parametric Location-Scale Regression Model

Consider we observe survival times $y_1, \cdots, y_n$ of n individuals, where for each individual i, we also observe a d-dimensional vector of covariates $\boldsymbol{x}_i$. The general location-scale model has the form

$$y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\theta}) + \eta(\boldsymbol{x}_i; \boldsymbol{\theta}) \cdot \varepsilon_i, \tag{2.1}$$

where $\mu(\boldsymbol{x}_i; \boldsymbol{\theta})$ is the location function, $\eta(\boldsymbol{x}_i; \boldsymbol{\theta})$ is the scale function, and $\varepsilon_i \overset{\text{iid}}{\sim} F(\varepsilon)$ has mean 0 and variance 1, denoted as $\varepsilon_i \overset{\text{iid}}{\sim} (0, 1)$, and is independent of $\boldsymbol{x}_i$.

Unless mentioned otherwise, in this thesis, we consider

$$\mu(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\beta},$$
$$\eta(\boldsymbol{x};\boldsymbol{\theta}) = \sigma \cdot \exp(\boldsymbol{z}'\boldsymbol{\gamma}),$$

such that $\boldsymbol{x} = (\boldsymbol{x}, \boldsymbol{z})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma)$, where σ is a scale parameter assumed to be a positive scalar, $\sigma > 0$, and for identifiability purpose, $\boldsymbol{\gamma}$ should not contain an intercept term. Then a location-scale model is specified as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \exp(\boldsymbol{z}_i'\boldsymbol{\gamma}) \cdot \varepsilon_i, \quad i = 1, \cdots, n. \tag{2.2}$$

Notice that if we assume $E(\varepsilon_i) = 0$, an intercept should be included in $\boldsymbol{\beta}$, which means that $x_{i,1} = 1$ for all $i = 1, \cdots, n$. Also, the assumption $E(\varepsilon_i) = 0$ in the location-scale model means that the expectations of y_i's are linear in the covariates. If $E(\varepsilon_i) = \mu \neq 0$, then the conditional expectation of y_i is

$$\mathrm{E}[y_i|\boldsymbol{x}_i, \boldsymbol{z}_i] = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \exp(\boldsymbol{z}_i'\boldsymbol{\gamma}) \cdot \mu,$$

which is not linear in the covariates.

When $\mu(\boldsymbol{x};\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$ and $\sigma(\boldsymbol{z};\boldsymbol{\gamma}) = \sigma$, the model reduces to a linear regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \varepsilon_i, \quad i = 1, \cdots, n.$$

2.2 Estimating Equations for Mean Regression

For the general location-scale model (2.1), suppose for a moment that $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and consider the so-called quasi-likelihood

$$\mathrm{QL}(\boldsymbol{\theta}|y, \boldsymbol{x}) = \prod_{i=1}^{n} \left[\frac{1}{\eta(\boldsymbol{x}_i;\boldsymbol{\theta})} \cdot \exp\Big\{ -\frac{\mu^2(\boldsymbol{x}_i;\boldsymbol{\theta})}{\eta^2(\boldsymbol{x}_i;\boldsymbol{\theta})} \Big\} \right]. \tag{2.3}$$

If the true model were indeed $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then $\boldsymbol{\theta}$ would minimize the expected negative log QL, such that

$$\theta = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathrm{E}\left[\frac{\mu^2(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})}{2\eta^2(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})} + \log\Big\{\eta(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})\Big\}\right], \tag{2.4}$$

or equivalently,

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathrm{E}\left[\frac{\mu^2(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})}{2\eta^2(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})} + \log\Big\{\eta(\boldsymbol{x}_i;\tilde{\boldsymbol{\theta}})\Big\}\right] = 0. \tag{2.5}$$

Remarkably, one can verify that (2.5) holds not only for $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, but for any $\varepsilon_i \overset{\text{iid}}{\sim} (0,1)$. Therefore, we use (2.5) to define the moment conditions for our specific location-scale model (2.2), namely

$$\begin{aligned}
\mathrm{E}\Big[\frac{y-\boldsymbol{x}'\boldsymbol{\beta}}{\exp(2\boldsymbol{z}'\boldsymbol{\gamma})}\cdot\boldsymbol{x}\Big] &= 0 \\
\mathrm{E}\Big[(1-\frac{(y-\boldsymbol{x}'\boldsymbol{\beta})^2}{\sigma^2\cdot\exp(2\boldsymbol{z}'\boldsymbol{\gamma})})\cdot\boldsymbol{z}\Big] &= 0 \\
\mathrm{E}\Big[\frac{(y-\boldsymbol{x}'\boldsymbol{\beta})^2}{\sigma^2\cdot\exp(2\boldsymbol{z}'\boldsymbol{\gamma})}-1\Big] &= 0.
\end{aligned} \tag{2.6}$$

Note that the scale parameter σ is dropped in the first equation in (2.6) since it is a positive constant multiplier with respect to the expectation.

2.3 Estimating Equations for Quantile Regression

For the location-scale model (2.2), the $\tau \times 100\%$ conditional quantile of y_i is

$$Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma\cdot\exp(\boldsymbol{z}_i'\boldsymbol{\gamma})\cdot\nu_\tau. \tag{2.7}$$

In this case, for parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and σ, we adopt the same estimating equation as in the mean regression case. For the quantile parameter ν_τ, we rely on the "check function" introduced by Basset and Koenker (1978), which is defined as

$$\rho_\tau(u) = u \cdot (\tau - \mathbb{1}\{u \leq 0\}), \tag{2.8}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

If the τ-th quantile value of $\varepsilon_i \overset{\text{iid}}{\sim} (0,1)$ is ν_τ, then $\varepsilon_i - \nu_\tau$ has τ-th quantile value 0. The estimator of ν_τ is then defined as

$$\hat{\nu}_\tau = \arg\min_{\tilde{\nu}_\tau} \mathrm{E}\left[\rho_\tau\Big(\frac{y - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma \cdot \exp(\boldsymbol{z}'\boldsymbol{\gamma})} - \tilde{\nu}_\tau\Big)\right]. \tag{2.9}$$

As before, we use the first order optimality condition of (2.9) to obtain the estimating equation for ν_τ. Therefore, we obtain all the moment conditions for quantile regression as follows

$$\begin{aligned}
\mathrm{E}\Big[\frac{y - \boldsymbol{x}'\boldsymbol{\beta}}{\exp(2\boldsymbol{z}'\boldsymbol{\gamma})} \cdot \boldsymbol{x}\Big] &= 0 \\
\mathrm{E}\Big[\big(1 - \frac{(y - \boldsymbol{x}'\boldsymbol{\beta})^2}{\sigma^2 \cdot \exp(2\boldsymbol{z}'\boldsymbol{\gamma})}\big) \cdot \boldsymbol{z}\Big] &= 0 \\
\mathrm{E}\Big[\frac{(y - \boldsymbol{x}'\boldsymbol{\beta})^2}{\sigma^2 \cdot \exp(2\boldsymbol{z}'\boldsymbol{\gamma})} - 1\Big] &= 0 \\
\mathrm{E}\Big[\rho'_\tau\big(\frac{y - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma \cdot \exp(\boldsymbol{z}'\boldsymbol{\gamma})} - \nu_\tau\big)\Big] &= 0.
\end{aligned} \tag{2.10}$$

Chapter 3

Empirical Likelihood with Fully-Observed Data

In this chapter, we describe the EL framework as well as the confidence interval construction in the absence of length-biases developed by previous works.

3.1 The Empirical Likelihood Framework

Let $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ where $\boldsymbol{y}_i \in \mathbb{R}^{d+1}$ be iid observations from an unknown distribution $F_0(\boldsymbol{y})$, about which a parameter of interest $\boldsymbol{\theta}$ is defined as satisfying an m-dimensional moment condition:

$$\mathrm{E}\big[\boldsymbol{g}(\boldsymbol{y};\boldsymbol{\theta})\big] = 0, \tag{3.1}$$

where $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{\theta}) = \big(g_1(\boldsymbol{y},\boldsymbol{\theta}), \ldots, g_m(\boldsymbol{y},\boldsymbol{\theta})\big)$.

The empirical likelihood $\mathrm{EL}(\boldsymbol{\theta})$ is defined as the profile likelihood over the distribution

function of $\boldsymbol{y}$:

$$\mathrm{EL}(\boldsymbol{\theta}) = \max_{F \in \mathcal{F}(\boldsymbol{\theta})} \prod_{i=1}^{n} \mathrm{d}F(\boldsymbol{y}_i), \tag{3.2}$$

where for any given $\boldsymbol{\theta}$, $\mathcal{F}(\boldsymbol{\theta})$ is the set of (valid) distribution functions satisfying (3.1).

It was shown (Owen, 1988) that for any $\boldsymbol{\theta}$, the maximum of (3.2) must be achieved by a PMF putting all mass on the support of the observed data $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$, such that the infinite-dimensional profile likelihood (3.2) reduces to a finite-dimensional one:

$$\mathrm{EL}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \hat{\omega}_i(\boldsymbol{\theta}), \tag{3.3}$$

where the n-dimensional vector of probability weights $\hat{\omega}(\boldsymbol{\theta})$ associated with the observations is the solution of an inner optimization problem which will be referred to as **EL inner optimization**

$$\begin{aligned} \max_{\omega} \quad & \sum_{i=1}^{n} \log(\omega_i) \\ \text{s.t.} \quad & \sum_{i=1}^{n} \omega_i \cdot \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{\theta}) = 0 \\ & \sum_{i=1}^{n} \omega_i = 1 \\ & \omega_i \geq 0, \quad i = 1, \cdots, n, \end{aligned} \tag{3.4}$$

The problem in (3.4) is a constrained convex optimization problem, and its optimal solution can be found by solving its dual problem derived through the Lagrangian function, as described by Owen (1990).

Specifically, provided that $\mathbf{0}$ is in the convex full of the points $\boldsymbol{g}(\boldsymbol{y}_1; \boldsymbol{\theta}), \cdots, \boldsymbol{g}(\boldsymbol{y}_n; \boldsymbol{\theta})$, a unique optimal weight vector exist and can be shown to be

$$\hat{\omega}_i(\boldsymbol{\theta}) = \frac{1}{n \cdot [1 - \hat{\boldsymbol{\lambda}}'(\boldsymbol{\theta}) \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{\theta})]}, \tag{3.5}$$

where the vector $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ solves the unconstrained optimization problem

$$\hat{\lambda}(\boldsymbol{\theta}) = \arg\max_{\lambda} \sum_{i=1}^{n} \log^{\star}\Big(1 - \boldsymbol{\lambda}' \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{\theta})\Big), \tag{3.6}$$

and where

$$\log^{\star}(x) = \begin{cases} \log(x) & x \geq \frac{1}{n} \\ -\frac{1}{2}n^2x^2 + 2nx - \frac{3}{2} - \log(n) & x < \frac{1}{n} \end{cases}. \tag{3.7}$$

Qin and Lawless (1994) has shown that $\boldsymbol{\lambda}(\boldsymbol{\theta})$ is a continuous differentiable function of $\boldsymbol{\theta}$ provided that convex hull condition is satisfied with $\boldsymbol{\theta}$ and $\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{\theta}) g'(\boldsymbol{y}_i; \boldsymbol{\theta})$ is positive definite. However, the support of $\boldsymbol{\theta}$ is not necessarily a convex set, as demonstrated by Chaudhuri et al. (2017).

3.2 Empirical Likelihood Confidence Intervals

Owen (1988) has shown that for sample mean, M-estimators and differentiable statistical functionals in 1-dimension, confidence intervals for the maximum empirical likelihood estimator (MELE) asymptotically follows a χ^2_1 distribution. Then the result is generalized to the construction of confidence region of statistics that depend smoothly on several means or linear estimating equations in multivariate case (Owen, 1990).

Qin and Lawless (1994) further links EL with estimating equations and provides a method to obtain a confidence region of the entire parameter vector as well as confidence intervals for any subset of parameters in $\boldsymbol{\theta}$, with the following theorem and corollary:

Theorem 1. *Suppose that x follows an unknown distribution F, and $\boldsymbol{\theta}$ is a d-dimensional parameter associated with F. Assume that $E[\boldsymbol{g}(x; \boldsymbol{\theta}_0)\boldsymbol{g}(x; \boldsymbol{\theta}_0)']$ is positive definite, $\frac{\partial \boldsymbol{g}(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and*

$\frac{\partial^2 g(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ *are continuous in a neighborhood of the true value* $\boldsymbol{\theta}_0$, $||\frac{\partial g(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}||$ *and* $||\frac{\partial g(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}||^3$ *are bounded by some integrable function in this neighborhood, and the rank of* $E[\frac{\partial g(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}]$ *is* d. *Let* $x_1, \cdots, x_n \overset{iid}{\sim} F$. *For* $0 < r < 1$, *let* $C_{r,n} = \{\boldsymbol{\theta} : W_E(\boldsymbol{\theta}) \geq r\}$, *then*

$$P(\boldsymbol{\theta}_0 \in C_{r,n}) \to P(\chi^2_d \leq -2\log r),$$

as $n \to \infty$, *where* $\boldsymbol{\theta}_0$ *is the true value of* $\boldsymbol{\theta}_0$, *and*

$$W_E(\boldsymbol{\theta}) = \frac{EL(\boldsymbol{\theta})}{EL(\hat{\boldsymbol{\theta}})},$$

where $\hat{\boldsymbol{\theta}}$ *is the MELE of* $\boldsymbol{\theta}$.

Proof. Theorem 2 by Qin and Lawless (1994). □

Corollary 1. *Let* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, *where* $\boldsymbol{\theta}_1 \in \mathbb{R}^q$ *and* $\boldsymbol{\theta}_2 \in \mathbb{R}^p$. *For* $0 < r < 1$, *let* $C_{r,n} = \{\boldsymbol{\theta}_1 : W_2(\boldsymbol{\theta}_1) \geq r\}$, *then*

$$P(\boldsymbol{\theta}_{1,0} \in C_{r,n}) \to P(\chi^2_q \leq -2\log r),$$

as $n \to \infty$, *where* $\boldsymbol{\theta}_{1,0}$ *is the true value of* $\boldsymbol{\theta}_1$, *and*

$$W_2(\boldsymbol{\theta}_1) = \frac{EL(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2^0)}{EL(\hat{\boldsymbol{\theta}})},$$

where $\hat{\boldsymbol{\theta}}_2^0$ *maximizes* $EL(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ *with respect to* $\boldsymbol{\theta}_2$ *with* $\boldsymbol{\theta}_1$ *being fixed, and* $\hat{\boldsymbol{\theta}}$ *is the MELE of* $\boldsymbol{\theta}$.

Proof. Corollary 5 by Qin and Lawless (1994). □

Example 1 illustrates how to use Corollary 1.

Example 1. *Recall the quantile regression location-scale model* (2.2)

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \exp(\boldsymbol{z}_i'\boldsymbol{\gamma}) \cdot \varepsilon_i,$$

so that for any individual i, its $\tau \times 100\%$ conditional quantile is given by (2.7)

$$Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \exp(\boldsymbol{z}_i'\boldsymbol{\gamma}) \cdot \nu_\tau.$$

In order to find the $\alpha \times 100\%$ CI of $Q_i = Q_\tau(y_i|\boldsymbol{x}_i)$ for individual i, we move the value of Q_i until the condition $W_2(Q_i) \geq r$ is violated, where $r = \exp(-\frac{1}{2}c_\alpha)$ and c_α is the $\alpha \times 100\%$ quantile of χ^2_1.

In particular, for any fixed Q_i, let

$$\hat{v}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma) = \frac{Q_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma \cdot \exp(\boldsymbol{z}_i'\boldsymbol{\gamma})}$$

and calculate

$$(\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\gamma}}^0, \hat{\sigma}^0) = \arg\max_{\boldsymbol{\beta},\boldsymbol{\gamma},\sigma} EL(\hat{v}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma), \boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}).$$

Then

$$W_2(Q_i) = \frac{EL(\hat{v}_i(\hat{\boldsymbol{\beta}}^0, \hat{\sigma}^0, \hat{\boldsymbol{\gamma}}^0), \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\gamma}}^0, \hat{\sigma}^0)}{EL(\hat{\boldsymbol{\theta}})}.$$

Notice that in Example 1 we have to go through this procedure for each one of the individuals if we want to obtain CI's for quantile values of all n individuals, which involves a large number of optimizations. Alternatively, we could use a bootstrap method, which only requires b times of optimizations, where b is the number of bootstraps, to obtain all the CI's at once, which could be much less computationally intensive when n is moderate or large.

Chapter 4

Empirical Likelihood under Right-Censoring

In this chapter, we will discuss two different approaches to work with right-censored data using EL: an EM algorithm extended from Zhou (2005) and the influence function approach by He et al. (2016).

4.1 Right-Censored EL for Regression Models

Consider the general location-scale model

$$y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\theta}) + \eta(\boldsymbol{x}_i; \boldsymbol{\theta}) \cdot \varepsilon_i, \tag{4.1}$$

with $\varepsilon_i \overset{\text{iid}}{\sim} (0, 1)$ and independent of $\boldsymbol{x}_i$, and m-dimensional conditional moment restrictions

$$\mathrm{E}\big[\boldsymbol{g}(\boldsymbol{x}, \varepsilon; \boldsymbol{\theta}) \,|\, \boldsymbol{x}\big] = 0. \tag{4.2}$$

When right-censoring is present, instead of observing y_i, we observe $u_i = \min(y_i, c_i)$ and $\delta_i = 1\{y_i \leq c_i\}$, where c_i is the censoring time. We assume that the censoring variable c_i is conditionally independent of y_i given $\boldsymbol{x}_i$.

The empirical likelihood with censored observations once again is defined by profiling over the unknown joint distribution function $F(\boldsymbol{x}, \varepsilon) = G(\boldsymbol{x}) \cdot H(\varepsilon)$, where G and H are the CDFs of $\boldsymbol{x}$ and ε:

$$\text{CEL}(\boldsymbol{\theta}) = \max_{F \in \mathcal{F}(\boldsymbol{\theta})} \prod_{i=1}^{n} dG(\boldsymbol{x}_i) \cdot dH(e_i)^{\delta_i} \cdot [1 - H(e_i)]^{1-\delta_i}, \tag{4.3}$$

where

$$e_i = e_i(\boldsymbol{\theta}) = \frac{u_i - \mu(\boldsymbol{x}_i; \boldsymbol{\theta})}{\eta(\boldsymbol{x}_i; \boldsymbol{\theta})},$$

and $\mathcal{F}(\boldsymbol{\theta})$ is the set of all valid distribution functions satisfying (4.2). It is not hard to show that for any choice of $H(\varepsilon)$, the maximum of (4.3) over $G(\boldsymbol{x})$ is attained as the empirical distribution $\hat{G}(\boldsymbol{x})$ which puts a point mass of $1/n$ on each covariate observation $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Restricting our attention to $G(\boldsymbol{x})$ uniform on the observed covariates, and considering only the weaker moment condition

$$\text{E}\big[\boldsymbol{g}(\boldsymbol{x}, \varepsilon; \boldsymbol{\theta})\big] = 0 \tag{4.4}$$

(which is true for any $G(\boldsymbol{x})$ if (4.2) holds), the CEL function reduces to

$$\text{CEL}(\boldsymbol{\theta}) = \max_{F \in \mathcal{F}^\star(\boldsymbol{\theta})} \prod_{i=1}^{n} dH(e_i)^{\delta_i} \cdot [1 - H(e_i)]^{1-\delta_i},$$

where $\mathcal{F}^\star(\boldsymbol{\theta})$ is the set of all valid distributions $F(\boldsymbol{x}, \varepsilon)$ satisfying (4.4).

Unfortunately, with censored observations, it is no longer true that an optimal F is only on the support of the data points $D_i = (u_i, \boldsymbol{x}_i)$, as illustrated by Example 2.

Example 2. *Suppose we have* 3 *observations,* a_1, a_2 *and* a_3 *as shown in Figure 4.1, and* $\delta_1 = 1$, $\delta_2 = 0$ *and* $\delta_3 = 1$. *log CEL in this case is*

$$\log(\omega_1) + \log(\omega_2 + \omega_3) + \log(\omega_3) \tag{4.5}$$

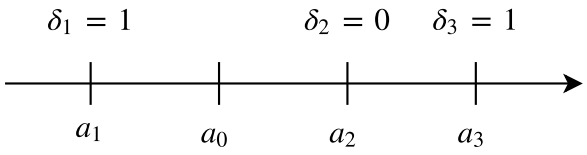


Figure 4.1: The optimal F may not have support only on the observations.

Consider a point a_0 between a_1 and a_2 might also take on a non-zero probability weight — the log CEL function does not change with this extra point. Suppose for some estimating function $g(\cdot)$ and a particular parameter $\boldsymbol{\theta}$, we have $g(a_1;\boldsymbol{\theta}) = 1$, $g(a_0;\boldsymbol{\theta}) = -5$, $g(a_2;\boldsymbol{\theta}) = -1$, and $g(a_3;\boldsymbol{\theta}) = 1$. Then the probability vectors $(\omega_1^{(1)}, \omega_0^{(1)}, \omega_2^{(1)}, \omega_3^{(1)}) = (0.4, 0.1, 0.2, 0.3)$ and $(\omega_1^{(2)}, \omega_0^{(2)}, \omega_2^{(2)}, \omega_3^{(2)}) = (0.1, 0, 0.5, 0.4)$ both satisfy $\sum_{i=0}^{3} \omega_i \cdot g(a_i;\boldsymbol{\theta}) = 0$. However, log CEL with $\omega^{(1)}$ is $\log(0.4 \times 0.5 \times 0.3) = \log(0.06)$ which is greater than log CEL with $\omega^{(2)}$ which is $\log(0.1 \times 0.9 \times 0.4) = \log(0.036)$.

However, if we restrict ourselves to this case, i.e., the support of F is $D_1, \ldots, D_n$, and if we also assume that $G(\boldsymbol{x})$ is the uniform distribution on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, then we arrive at the finite dimensional problem

$$\mathrm{CEL}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \Big[\hat{\omega}_i(\boldsymbol{\theta})^{\delta_i} \big(\sum_{j:e_j \geq e_i} \hat{\omega}_i(\boldsymbol{\theta}) \big)^{1-\delta_i} \Big]. \tag{4.6}$$

where similar as before, $\hat{\omega}(\boldsymbol{\theta})$ is the solution of an inner optimization problem

$$\begin{aligned} \max_{\omega} \quad & \sum_{i=1}^{n} \Big[\delta_i \log(\omega_i) + (1-\delta_i) \log\big(\sum_{j:e_j \geq e_i} \omega_i \big) \Big] \\ \text{s.t.} \quad & \sum_{i=1}^{n} \omega_i \cdot \boldsymbol{g}(\boldsymbol{x}_i, u_i; \boldsymbol{\theta}) = 0 \\ & \sum_{i=1}^{n} \omega_i = 1 \\ & \omega_i \geq 0, \quad i = 1 \cdots, n. \end{aligned} \tag{4.7}$$

Note that since we restrict the support of F to be only on the observations D_i, it is a PMF, so that the first constraint in (4.7) is indeed the moment condition itself rather than the sample version of it. Moreover, with the above formulation of CEL, a censored observation may also take on a positive probability weight.

Proposition 1. *The log CEL function defined above is concave in ω, so that if the convex hull condition is statisfied with $\boldsymbol{\theta}$, an optimal solution $\hat{\omega}(\boldsymbol{\theta})$ exists and it is global.*

Proof. See Appendix D. □

4.2 An EM Algorithm for EL Inner Optimization

Recall that when the survival times are fully observed, the EL inner optimization problem given a fixed parameter $\boldsymbol{\theta}$ can be converted to its dual problem through Lagrange multipliers. Under right-censoring, this conversion turns out to be difficult to achieve. However, notice that right-censoring essentially causes a missing data problem which can be handled by an EM algorithm. We consider an EM algorithm which is a generalization of Zhou (2005) to regression problems.

Let's denote the observed residuals given a specific $\boldsymbol{\theta}$ as e_i's (corresponding to u_i's), the complete residuals as ε_i' (corresponding to y_i's). Then if $\delta_i = 1$, we have $e_i = \varepsilon_i$, otherwise we do not observe ε_i. This means that the **unobserved (latent) variables** are the ε_i's such that $\delta_i = 0$, the **complete data likelihood** is

$$\ell(\omega, \varepsilon | e) = \sum_{i=1}^{n} \log(\prod_{j=1}^{n} \omega_j^{\mathbb{1}(\varepsilon_i = e_j)}) \tag{4.8}$$

The E-step of the EM algorithm takes the expectation of (4.8) with respect to ε (vector of all latent variables) conditioned on the observed values, the censoring indicator and

the current state of the parameters, and since for $\delta_i = 1$, we know that $e_i = \varepsilon_i$, we have

$$\begin{aligned}
\mathrm{E}_{\varepsilon|e,\delta,\omega_0}\big[\ell(\omega,\varepsilon|e)\big] &= \mathrm{E}_{\varepsilon|e,\delta,\omega_0}\Big[\sum_{i=1}^{n}\delta_i\log(\omega_i) + (1-\delta_i)\log(\prod_{j=1}^{n}\omega_j^{\mathbb{1}(\varepsilon_i=e_j)})\Big] \\
&= \sum_{i=1}^{n}\Big[\delta_i\log(\omega_i) + (1-\delta_i)\sum_{j=1}^{n}\mathrm{E}_{\varepsilon_i|e,\delta,\omega_0}[\mathbb{1}(\varepsilon_i=e_j)]\log(\omega_j)\Big] \qquad (4.9)\\
&= \sum_{i=1}^{n}\Big[\delta_i\log(\omega_i) + (1-\delta_i)\sum_{j=1}^{n}P_{\varepsilon_i|e,\delta,\omega_0}(\varepsilon_i=e_j)\log(\omega_j)\Big].
\end{aligned}$$

Notice that we can write

$$\log(\prod_{j=1}^{n}\omega_j^{\mathbb{1}(\varepsilon_i=e_j)}) = \sum_{i=1}^{n}\mathbb{1}(\varepsilon_i=e_j)\log(\omega_j),$$

because for any $i \in \{1,\cdots,n\}$, $\mathbb{1}(\varepsilon_i = e_j) = 1$ for one and only one $j \in \{1,\cdots,n\}$. Also, the latent $\varepsilon_i's$ are independent but not identically distributed, since each of them follows a different categorical distribution (multinomial distribution with one trial).

The conditional distribution in (4.10) is a categorical distribution conditioned on that the probability mass only allocates on the values in a subset of $\{e_1,\cdots,e_n\}$ such that $\mathbb{1}(e_j \geq e_i) = 1$ for $j = 1,\cdots,n$, which is still a multinomial distribution.

$$P_{\varepsilon_i|e,\delta,\omega_0}(\varepsilon_i = e_j) = \frac{\mathbb{1}(e_j \geq e_i)\cdot\omega_{0j}}{\sum_{k=1}^{n}\mathbb{1}(e_k \geq e_i)\cdot\omega_{0k}}. \qquad (4.10)$$

Therefore, the EM algorithm iterates between the following two steps:

- **E-step:** Given the observed values and the weights ω_0 from the previous iteration, the expectation of the log likelihood is

$$\begin{aligned}
\mathrm{E}_{\varepsilon|e,\delta,\omega_0}[\ell(\omega,\varepsilon|e)] &= \sum_{i=1}^{n}\Big[\delta_i\log\omega_i + (1-\delta_i)\sum_{j:e_j\geq e_i}\tilde{\omega}_{ij}\log\omega_j\Big] \\
&= \sum_{i=1}^{n}\Big[\delta_i + \sum_{k:e_k\leq e_i}(1-\delta_k)\cdot\tilde{\omega}_{ki}\Big]\cdot\log\omega_i,
\end{aligned} \qquad (4.11)$$

where

$$\tilde{\omega}_{ki} = \frac{\omega_{0i}}{\sum_{l:e_l \geq e_k} \omega_{0l}}, \quad k, i : e_k \leq e_i.$$

- **M-step:** Let $q_i = \delta_i + \sum_{k:e_k \leq e_i} (1 - \delta_k) \cdot \tilde{\omega}_{ki}$ for $i = 1, \cdots, n$, then the problem becomes

$$\begin{aligned} \max_{\omega} \quad & \sum_{i=1}^{n} q_i \log \omega_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} \omega_i \cdot \boldsymbol{g}(x_i, u_i; \boldsymbol{\theta}) = 0 \\ & \sum_{i=1}^{n} \omega_i = 1 \\ & \omega_i \geq 0, \quad i = 1 \cdots, n, \end{aligned} \tag{4.12}$$

  which is in the same form as in the case without right-censoring. It can be shown that (see Appendix A) the solution of (4.12) is

$$\hat{\omega}_i = \frac{q_i}{n + \hat{\boldsymbol{\lambda}}'(\boldsymbol{\theta}) \boldsymbol{g}(x_i, u_i; \boldsymbol{\theta})}, \tag{4.13}$$

  where $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ can be solved analogously which is

$$\hat{\lambda}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\lambda}} \sum_{i=1}^{n} q_i \cdot \log^{\sharp}\Big(n + \boldsymbol{\lambda}' \boldsymbol{g}(x_i, u_i; \boldsymbol{\theta})\Big), \tag{4.14}$$

  where

$$\log^{\sharp}(x_i; q_i) = \begin{cases} \log(x_i) & x_i \geq q_i \\ -\frac{1}{2q_i^2} x_i^2 + \frac{2}{q_i} x_i - \frac{3}{2} + \log(q_i) & x_i < q_i \end{cases}. \tag{4.15}$$

As before, the modification of the log function is to expand the domain from strictly positive numbers to the real line, while making sure that the optimal value remains the same.

An issue that may be concerned is that the log EL function may not be continuous with respect to the regression parameters. This issue will be addressed in the next chapter.

4.3 The Influence Function Approach

The influence function approach by He et al. (2016) is another promising method to deal with right-censored data with empirical likelihood. The paper mainly concerns about constructing confidence intervals for a 1-dimensional parameter θ, which is a functional of the lifetime distribution F. Here we discuss the method under regression models where the parameter $\boldsymbol{\theta}$ is a vector, as well as some issues therein.

Let $F(e) = \mathrm{P}(\varepsilon \leq e)$ and $G(s) = \mathrm{P}(c \leq s)$ be the distribution functions associated with the lifetime variable and the censoring variable, and F_n and G_n be the Kaplan-Meier estimators of them respectively. Denote $\overline{F} = 1 - F$ as the survival function of any distribution F.

Also, let $u = \min(\varepsilon, c)$ and denote $H(x) = \mathrm{P}(u \leq x)$ as the distribution function of u. Then given a sample of random pairs (u_i, δ_i) of (u, δ), denote their empirical CDF's as

$$\begin{aligned} H_n^1(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{u_i \leq x, \delta_i = 1\} \\ H_n^0(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{u_i \leq x, \delta_i = 0\}. \end{aligned} \tag{4.16}$$

EL in this case is defined the same as with fully-observed data,

$$\mathrm{EL}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \hat{\omega}_i(\boldsymbol{\theta}), \tag{4.17}$$

except that $\hat{\omega}(\boldsymbol{\theta})$ is the solution of the following optimization problem using influence

functions W_{ni}

$$\begin{aligned} \max_{\omega} \quad & \sum_{i=1}^{n} \log(\omega_i) \\ \text{s.t.} \quad & \sum_{i=1}^{n} \omega_i \cdot W_{ni}(\boldsymbol{\theta}) = 0 \\ & \sum_{i=1}^{n} \omega_i = 1 \\ & \omega_i \geq 0, \quad i = 1, \cdots, n. \end{aligned} \tag{4.18}$$

The influence functions are approximations of the iid random functions W_i defined by Akritas et al. (2000) or He and Huang (2003):

$$W_i = \frac{\boldsymbol{g}(u_i, \boldsymbol{\theta})\delta_i}{\overline{G}(u_i)} + \frac{\overline{\delta}_i}{\overline{H}(u_i)}\psi(u_i) - \int \psi(s)\frac{\mathbb{1}\{u_i \geq s\}}{\overline{H}^2(s)}dH^0(s), \tag{4.19}$$

where

$$\psi(s) = \int_{x \geq s} \boldsymbol{g}(x, \theta)dF(x), \tag{4.20}$$

and $\boldsymbol{g}(\cdot)$ is the estimating function.

It is shown that if $\boldsymbol{\theta}_0$ is the true parameter, then

$$\mathrm{E}[W_i(\boldsymbol{\theta}_0)] = \int \boldsymbol{g}(x, \boldsymbol{\theta}_0)dF(x) = 0. \tag{4.21}$$

The approximation is achieved by replacing all distribution functions by their EM estimators or empirical CDF's:

$$W_{ni} = \frac{\boldsymbol{g}(u_i, \boldsymbol{\theta})\delta_i}{\overline{G}_n(u_i)} + \frac{\overline{\delta}_i}{\overline{H}_n(u_i)}\psi_n(u_i) - \int \psi_n(s)\frac{\mathbb{1}\{u_i \geq s\}}{\overline{H}_n^2(s)}dH_n^0(s), \tag{4.22}$$

where

$$\psi_n(s) = \int_{x \geq s} \boldsymbol{g}(x, \boldsymbol{\theta})dF_n(x). \tag{4.23}$$

Therefore, the solution can be found the same way as in the fully-observed case. However, the log EL is also not continuous in $\boldsymbol{\theta}$ with regression models. Example 3 shows the discontinuity.

Example 3. *We simulate data from a simple linear model with only one slope parameter* $\beta = 1$

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, \cdots, 200,$$

where $x_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$. *The log EL curve is shown in the left plot of Figure 4.2, and the right plot is the same plot but zoomed in to make the discontinuity more clearly.*

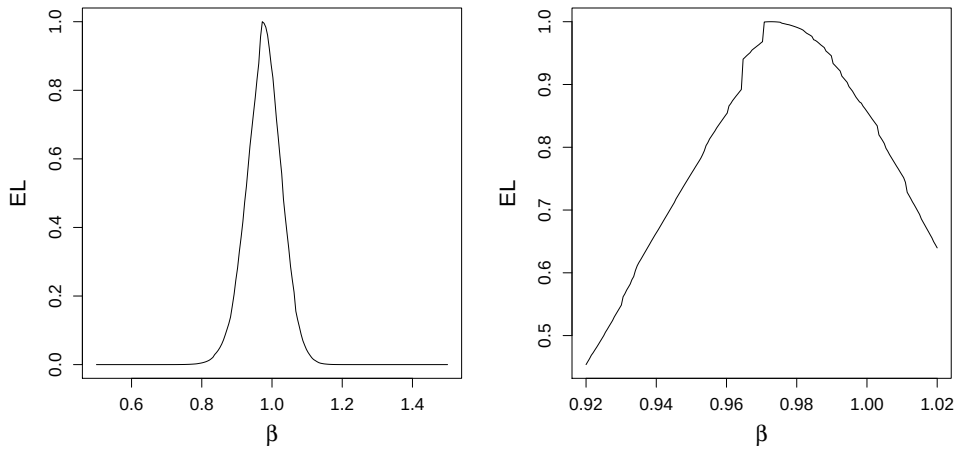


Figure 4.2: log EL curves for a simple linear regression with influence function.

Chapter 5

Smoothed Empirical Likelihood

There are two sources of discontinuity in the log EL function: the formulation of right-censored EL and the check function used in quantile regression. In this chapter, we discuss a single trick to deal with the two discontinuity issues.

5.1 Discontinuity due to Right-Censoring

As briefly mentioned in the last chapter, log CEL function is not a continuous function in $\boldsymbol{\theta}$, so that direct optimization of log CEL is difficult to achieve. Example 4 illustrates the problem.

Example 4. *Figure 5.1 shows two conditional log CEL curves with data generated from a simple linear model*

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \cdots, 200,$$

where $\beta_0 = 1$, $\beta_1 = 1.5$, $x_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$, $c_i \overset{iid}{\sim} \mathcal{N}(1.35,1)$, *and* $\epsilon_i = \min(\varepsilon_i, c_i)$. *Notice that*

conditionally, log CEL is continuous in β_0 for fixed β_1, but not the other way around. This is because moving β_0 continously while fixing β_1 does not change the ranking of the residuals.

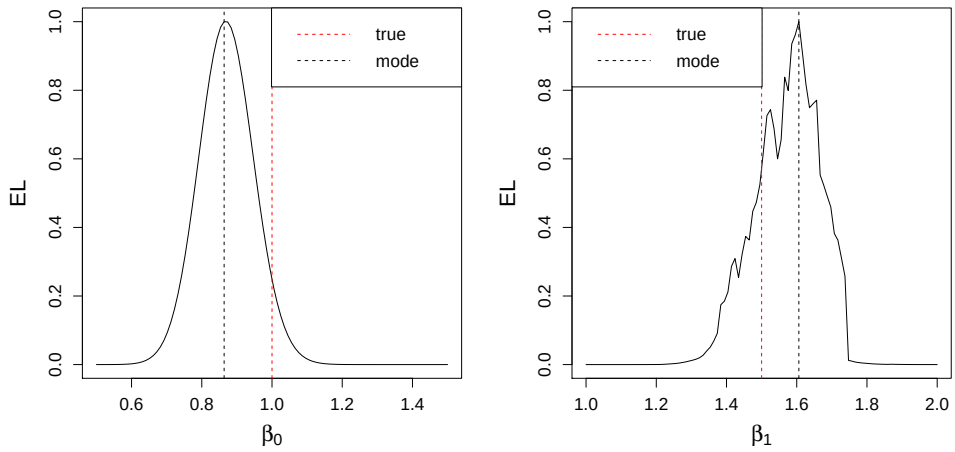


Figure 5.1: Conditional log CEL curves for a simple linear regression.

In order to obtain an estimate in this case, one could use Markov Chain Monte Carlo or global optimization algorithm such as simulated annealing, which are both time-consuming. Instead of relying on these methods, we consider a revision of the log CEL function.

The log CEL in Chapter 4 can be expanded as follows

$$\begin{aligned} \ell_{\mathrm{CEL}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\Big[\delta_i \log(\omega_i(\boldsymbol{\theta})) + (1-\delta_i)\log(\sum_{j:e_j \geq e_i} \omega_j(\boldsymbol{\theta}))\Big] \\ &= \sum_{i=1}^{n}\Big[\delta_i \log(\omega_i(\boldsymbol{\theta})) + (1-\delta_i)\log(\sum_{j=1}^{n} \mathbb{1}(e_j(\boldsymbol{\theta}) \geq e_i(\boldsymbol{\theta})) \cdot \omega_j(\boldsymbol{\theta}))\Big]. \end{aligned} \tag{5.1}$$

We can see that the discontinuity of $\ell_{\mathrm{CEL}}(\boldsymbol{\theta})$ in (5.1) is due to an indicator function. To smooth out this discontinuity, we replace the indicator function by a continuous approximation.

Let S be a transformed sigmoid function, i.e.,

$$S(x;s) = \frac{1}{1+\exp(s \cdot x)}, \tag{5.2}$$

where $s > 0$ is a smoothing parameter. A plot of the function is given in Figure 5.2. This function is radially symmetric around the point $(0, 0.5)$.

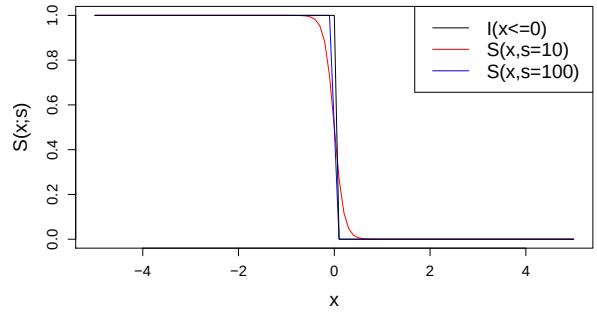


Figure 5.2: Smooth function for indicator function $\mathbb{1}(x \leq 0)$.

Specifically, we use

$$S_{ij}(\boldsymbol{\theta};s) := S(e_i(\boldsymbol{\theta}) - e_j(\boldsymbol{\theta});s) = \frac{1}{1+\exp(s \cdot (e_i(\boldsymbol{\theta}) - e_j(\boldsymbol{\theta})))}. \tag{5.3}$$

Notice that as long as $e_i(\boldsymbol{\theta})$ is a continous function of $\boldsymbol{\theta}$ for all $i = 1, \cdots, n$, then (5.3) is indeed a continuous function of $\boldsymbol{\theta}$.

The log smoothed censored EL (log SCEL) is then defined as

$$\ell_{\text{SCEL}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big[\delta_i \log(w_i(\boldsymbol{\theta})) + (1-\delta_i) \log(\sum_{j=1}^{n} S_{ij}(\boldsymbol{\theta};s) \cdot w_j(\boldsymbol{\theta})) \Big]. \tag{5.4}$$

If $\omega(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ (which depends on the support defined by the estimating equations), and each $S_{ij}(\boldsymbol{\theta};s)$ is continuous in $\boldsymbol{\theta}$, then since the sum of continuous

functions is a continuous, and the composition of continuous functions is continuous, then $\ell_{\text{SCEL}}(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$.

Proposition 2. *The log SCEL function defined by (5.4) is a concave function of ω, so that if the convex hull condition is satisfied with $\boldsymbol{\theta}$, an optimal solution $\hat{\omega}(\boldsymbol{\theta})$ exists and it is global.*

Proof. See appendix E. □

5.2 Discontinuity due to Quantile Regression Constraint

The log EL function is not continuous given the quantile regression constraints even with fully-observed data. An illustration is presented in Example 5.

Example 5. *Figure 5.3 shows two conditional log CEL curves with data generated from a simple linear model, where we aim to estimate the 75% quantile*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, 200,$$

where $x_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$, $\beta_0 = v_{\tau_{0.75}} = 0.6745$, and $\beta_1 = 1.5$.

Recall the check function in Section 2.3 is

$$\rho_\tau(u) = u \cdot (\tau - \mathbb{1}\{u \leq 0\}).$$

Observe that the discontinuity also comes from an indicator function $\mathbb{1}\{u \leq 0\}$. To make this function continuous, we again replace the indicator function by the continuous approximation in (5.2), so the smoothed check function is given by

$$\rho_{S,\tau}(u; s) = u \cdot (\tau - S(u; s)).$$

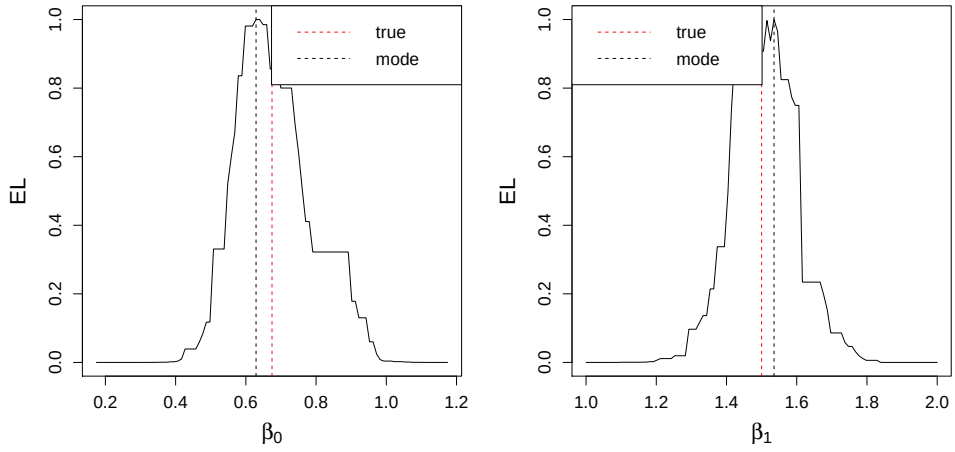


Figure 5.3: Conditional log EL curves for a simple quantile regression.

Although there are other approaches to smooth out the discontinuity in quantile regression, such as Chen (2007), splines or kernel methods, the trick above is particularly straightforward to apply.

As an example, we show the effect of the above continuity correction to both right-censoring and quantile regression constraint by 3D surface plots of log EL. We generate data from a simple linear model, and we aim to estimate the 75% quantile

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \cdots, 200,$$

where $\beta_0 = \nu_{0.75} = 0.6745$, $\beta_1 = 1.5$, $x_i, \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $c_i \overset{\text{iid}}{\sim} \mathcal{N}(1.35,1)$, and $\epsilon_i = \min(\varepsilon_i, c_i)$.

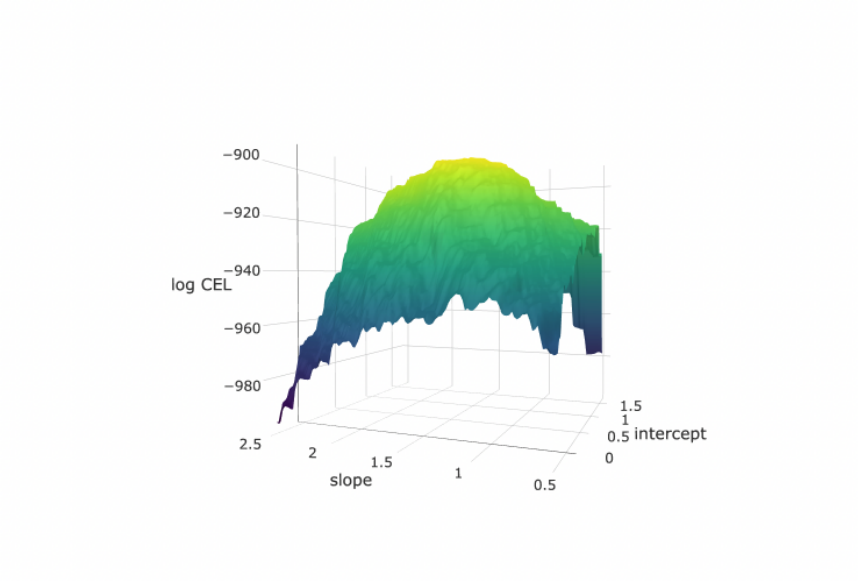
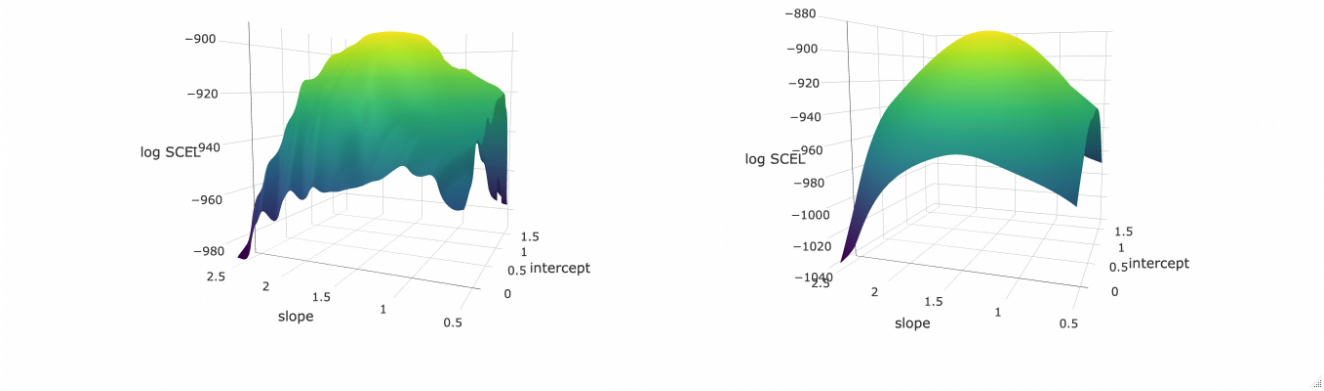


Figure 5.4: Original log CEL surface.



(a) log CEL surface with $s = 10$ (b) log CEL surface with $s = 1$

Figure 5.5: log CEL surfaces after continuity correction.

5.3 The EM Algorithm after Continuity Correction

Recall that in the EM algorithm, parameters are the ω_i's given a specific $\boldsymbol{\theta}$. In the E-step, we take the expectation of the log EL given the observations and the current value of the parameters.

With (5.1) where there is no smoothing, the expectation is in fact the expectation of a multinomial distribution conditional on the observed censoring indicators and indicator functions of e_i's. With smoothing as in (5.4), the expectation is taken conditional on the observed censoring indicators and smooth functions of e_i's.

Carrying similar steps as before, the E-step gives

$$q_i = \delta_i + \sum_{j=1}^{n} (1 - \delta_j) \cdot \tilde{\omega}_{ji} \cdot \log(\omega_i) \tag{5.5}$$

where $\tilde{\omega}_{ji} = \frac{S(e_j - e_i; s) \cdot \omega_{0i}}{\sum_{k=1}^{n} S(e_i - e_k) \cdot \omega_{0k}}$, and the S function is defined as in (5.2). Then the M-step solves a constrained weighted EL maximization problem as before.

With smoothing, we are modifying the meaning of censoring from "the value of ε_i being in a subset of $\{e_1, \cdots, e_n\}$" to "the value of ε_i getting a value of each $e_k \in \{e_1, \cdots, e_n\}$ with a certain probability which is the distance between e_i and e_k measured by S".

$$P_{\varepsilon_i | e, \delta, \omega_0}(\varepsilon_i = e_j) = \frac{S(e_j \geq e_i) \cdot \omega_{0j}}{\sum_{k=1}^{n} S(e_k \geq e_i) \cdot \omega_{0k}}. \tag{5.6}$$

The resulting distribution Eq (5.6) is still multinomial distribution but with support on all elements in $\{e_1, \cdots, e_n\}$.

In the M-step, we are maximizing Eq (4.9) to obtain the new ω. Therefore, the above algorithm is indeed an EM algorithm.

Chapter 6

Simulation Studies

In this chapter, we first give a brief description of the `flexEL` package, and then we present some simulation results using the methods proposed in this thesis.

6.1 Description of the `flexEL` Package

The package `flexEL` offers a flexible framework for users to solve EL regression problems with a fast computational speed. The package is written in C++ with an R interface providing functionalities including:

- EL inner optimization: a Newton-Raphson algorithm to solve the inner optimization problem of EL.
- log EL computation: calculate log EL of given parameters under a regression model.

- Mean regression: compute the estimating equations given a parameter value for mean regression under both location or location-scale models.
- Quantile regression: compute the estimating equations given a parameter value for quantile regression under both location or location-scale models.

The package has a clear structure to allow users to solve customized regression problems with minimum programming effort. Essentially, the only requirement is for one to design and implement the estimating equations of the regression problem, either in R for convenience or in C++ for computational speed. More details about this package will be provided in an upcoming GitHub release.

As an example for the computational speed, a comparison of the EL inner optimization function (fully-observed data case) implemented in C++ and R is provided in Figure 6.1. The computational times are the averages of 500 repetitions. Notice that to maximize EL with respect to $\boldsymbol{\theta}$, many times of this inner EL optimization need to be performed.

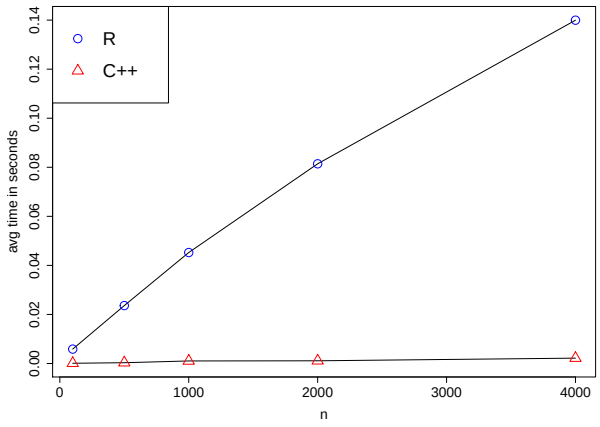


Figure 6.1: Computational time comparison for inner EL optimization.

6.2 Experiment Summary

In the following experiments, we consider the following location-scale quantile regression model

$$y_i = \beta_0 + \beta_1 \cdot x_i + \sigma \cdot \exp(\gamma \cdot z_i) \cdot \epsilon_i, \quad i = 1, \cdots, n, \tag{6.1}$$

where $\beta_0 = 0.5$, $\beta_1 = 1$, $\gamma = -0.5$, and $\sigma^2 = 1$. The covariates $x_i, z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, the censoring variable $c_i \overset{\text{iid}}{\sim} \mathcal{N}(1.35,1)$, and $\epsilon_i = \min(\varepsilon_i, c_i)$, where $\varepsilon_i \overset{\text{iid}}{\sim} F(\varepsilon)$, one of the error distributions summarized in Table 6.1. We consider estimating the 75% quantile value $\nu_{0.75}$ of the error distributions.

$F(\varepsilon)$	$\nu_{0.75}$	Censored (%)	Note
$\mathcal{N}(0,1)$	0.6745	17.0	Symmetric baseline
NCT$(-1,10)$	0.6530	16.6	Heavy left-tail
NCT$(1,10)$	0.5950	16.5	Heavy right-tail
NCT$(1,3)$	0.3219	13.7	Extremely heavy right-tail

Table 6.1: Summary of distributions used in simulations.

For all error distributions in Table 6.1, the information refers to the transformed pdf so that $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = 1$. "$\nu_{0.75}$" denotes the true 75% quantile value of the distribution, "Censored (%)" is the approximated percentage being censored with $c_i \overset{\text{iid}}{\sim} \mathcal{N}(1.35,1)$. NCT$(a,b)$ refers to a non-central t distribution with a as the non-central parameter and b as the degrees of freedom which control the tails of the distribution.

We compare the maximum smoothed censored EL (SCEL) estimator and the heteroscedastic accelerated failure time model for right-censored data by Wang et al. (2015) with Kaplan–Meier (HLM+KM).

Optimization of log SCEL is conducted through the non-linear minimization function

`nlm` in R. For the semi-parametric regression model (6.1), the HLM+KM approach proceeds in two steps:

1. Estimate the parameters using the QL with $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
2. Estimate $F(\varepsilon)$ by the Kaplan-Meier estimator of $e_i = \frac{u_i - \hat{\beta}_1 \cdot x_i - \hat{\beta}_0}{\sigma \cdot \exp(\gamma \cdot z_i)}$.

We will also present two methods to construct confidence intervals: bootstrap confidence intervals and mode-quadrature (normal approximation) confidence intervals. We must point out that, while the CEL estimator is asymptotically normal, to the best of our knowledge, convergence of the negative CEL Hessian to the true inverse variance matrix has not been formally established. However, the mode-quadrature approach is much faster than the bootstrap, hence our interest in the empirical comparison to follow.

6.3 Point Estimates of Parameters

In this section, we will compare the two methods for point estimates of the parameters using box plots. All the experiments are conducted through 200 times simulations.

1. $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$: See Figure 6.2. Since in this case the error distribution coincides with the imputation distribution by HLM+KM, for a large sample size the two methods perform very similarly.
2. $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(-1,10)$: See Figure 6.3. Notice that because the error distribution has a heavy left-tail and the right-censoring mechanism thus makes censored errors to have a smaller variance, the estimation of σ^2 tends to be smaller compared to HLM, although for larger sample size the bias does get smaller.

3. $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(1, 10)$: See Figure 6.4. Here the error distribution has a heavy right-tail, the SCEL estimator does perform better than HLM+KM especially on σ^2 and $\nu_{0.75}$.

4. $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(1, 3)$: See Figure 6.5. The error distribution has an extremely right-tail, although there is a clearer advantage of SCEL estimator, it needs a much larger sample size to reduce the bias, especially on the estimations of σ^2 and $\nu_{0.75}$.

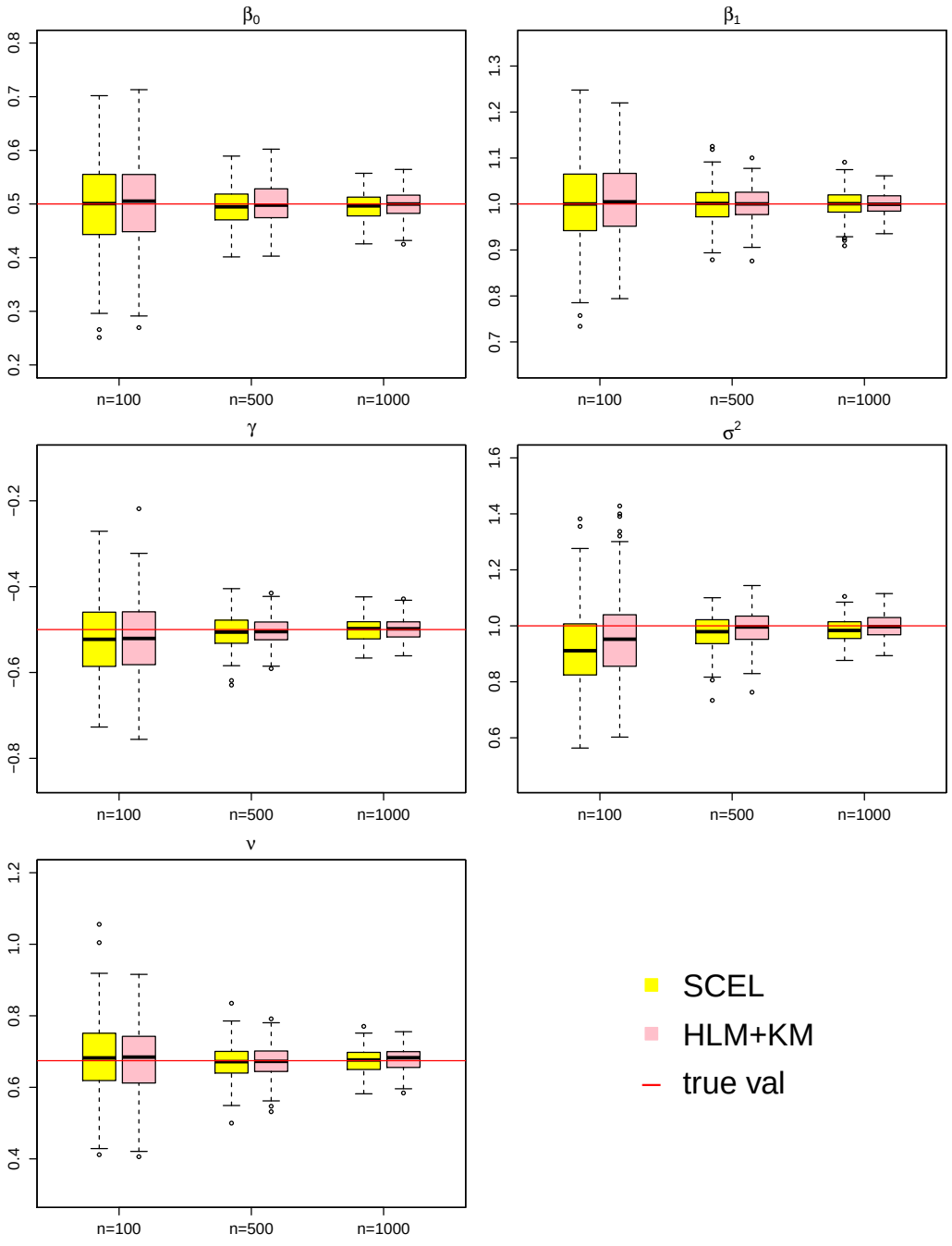


Figure 6.2: Estimates of $\boldsymbol{\theta}$ as n increases, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

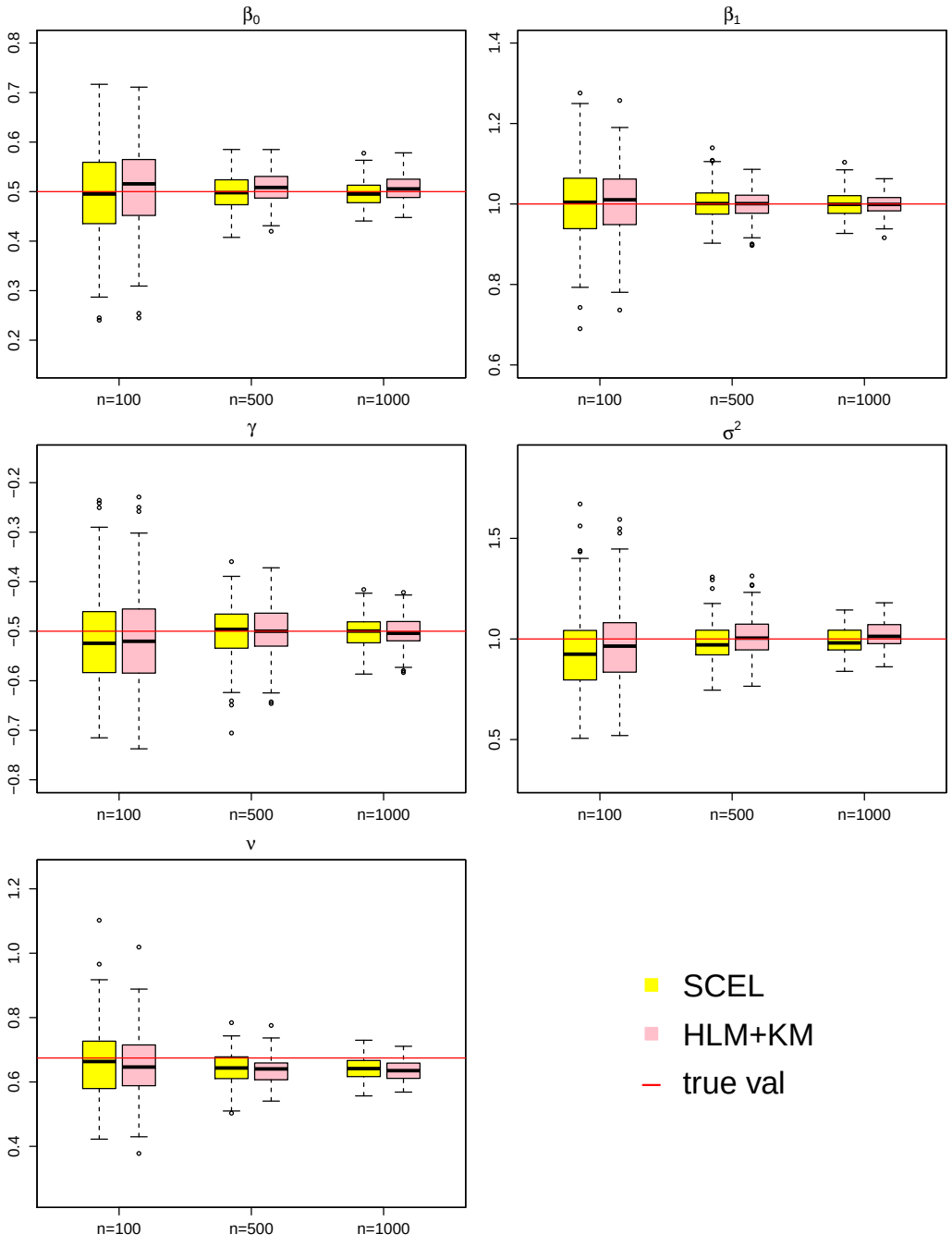


Figure 6.3: Estimates of $\boldsymbol{\theta}$ as n increases, $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(-1, 10)$.

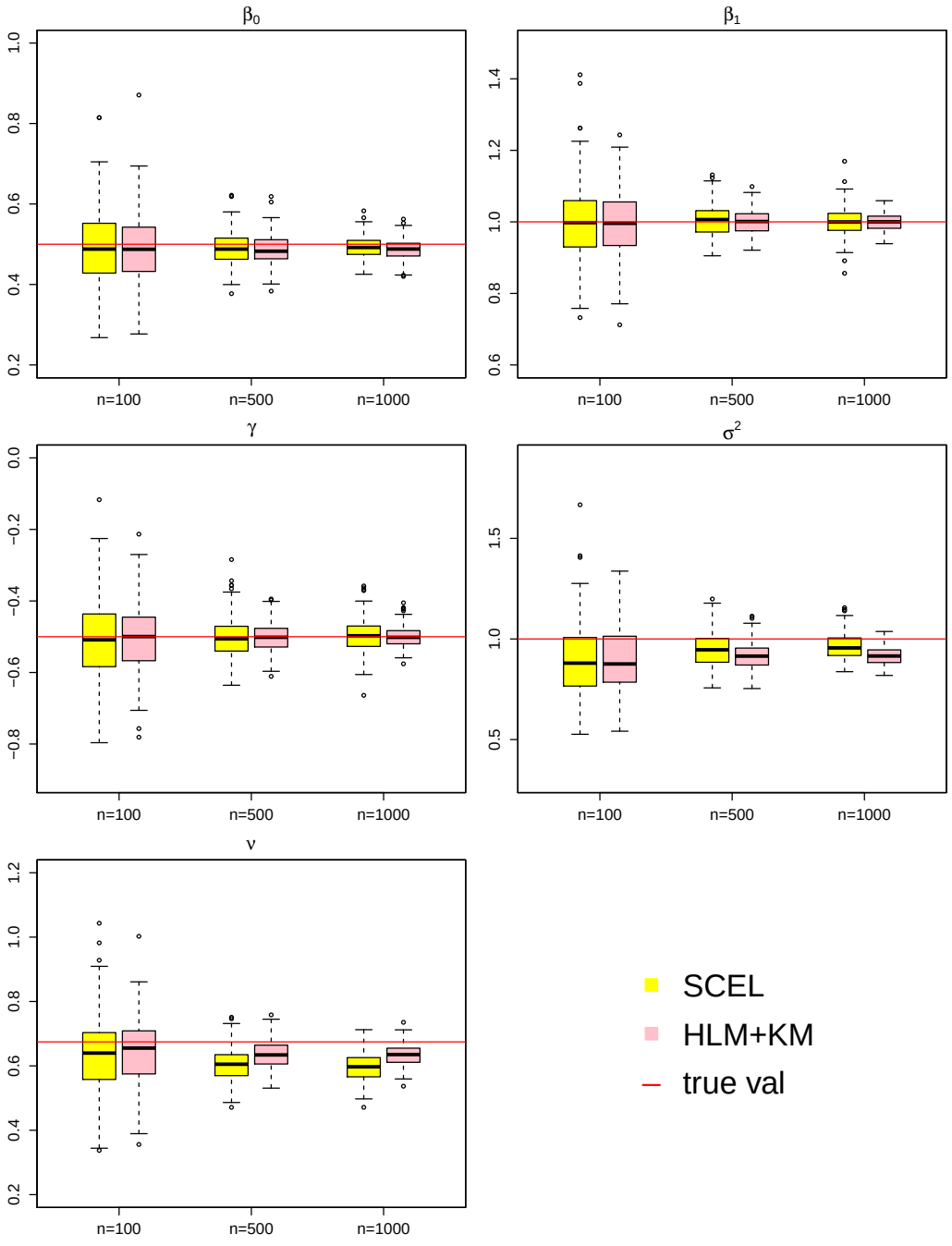


Figure 6.4: Estimates of $\boldsymbol{\theta}$ as n increases, $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(1, 10)$.

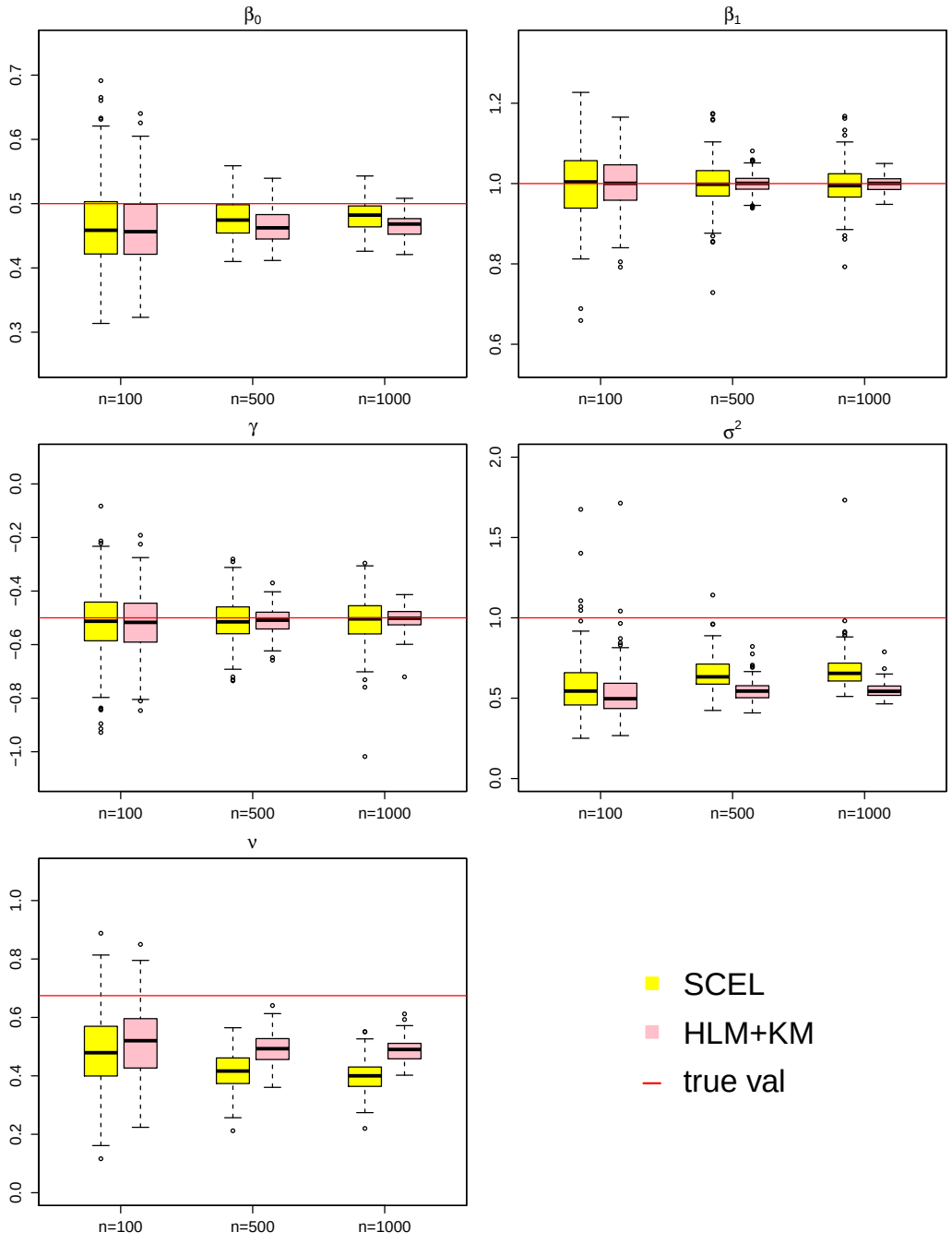


Figure 6.5: Estimates of $\boldsymbol{\theta}$ as n increases, $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(1,3)$.

6.4 Construction of Confidence Intervals

Here we compare bootstrap confidence intervals and mode-quadrature approximated confidence intervals for the above error distributions. The confidence level is chosen to be 95%. The bootstrap CIs are obtained from 100 bootstrap samples using the bootstrap percentile method, and the mode-quadrature approximation is obtained by using a normal distribution with the estimate as the mean, and the diagonal elements of the hessian matrix returned from `nlm` as the variances. Due to computational time, the bootstrap method is done up to $n = 1000$. The results are given in Table 6.2 to Table 6.5.

From the tables we can see that the mode-quadrature approximation performs well for $\mathcal{N}(0,1)$ error, while for other skewed error distributions, it needs a larger sample size to achieve the right coverage. Moreover, the coverage probabilities for $\text{NCT}(1,3)$ is very poor especially for σ^2 and $\nu_{0.75}$. From the previous section, we can see that this is due to the bias which is still obvious even sample size is 1000, although the bias is decreasing as n gets larger.

	Bootstrap			Mode-Quad			
n	100	500	1000	100	500	1000	1500
β_0	91.5	92.0	92.5	85.0	90.0	92.5	95.0
β_1	90.0	91.5	92.5	76.5	91.0	89.5	94.0
γ	92.5	90.0	92.5	77.5	85.0	89.5	91.5
σ^2	81.5	90.5	91.0	79.5	93.0	93.5	95.0
$\nu_{0.75}$	96.0	93.5	92.5	88.0	90.0	92.0	93.5

Table 6.2: Coverage probabilities, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

	Bootstrap			Mode-Quad			
n	100	500	1000	100	500	1000	1500
β_0	92.0	95.5	93.5	85.0	93.0	92.5	91.0
β_1	89.0	92.0	95.5	79.5	87.0	92.0	92.0
γ	91.0	92.0	94.5	77.5	82.5	90.0	90.0
σ^2	79.5	91.5	92.0	76.0	90.5	91.5	91.0
$\nu_{0.75}$	96.5	94.0	92.0	86.0	90.5	87.0	90.5

Table 6.3: Coverage probabilities, $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(-1, 10)$.

	Bootstrap			Mode-Quad			
n	100	500	1000	100	500	1000	1500
β_0	94.0	91.0	91.0	81.0	89.0	89.0	89.0
β_1	96.5	93.0	92.0	80.5	86.0	85.5	85.5
γ	93.0	91.0	95.0	72.0	78.5	82.0	81.0
σ^2	69.5	80.5	82.5	66.5	77.0	78.5	76.0
$\nu_{0.75}$	92.5	94.0	93.5	83.0	91.0	90.0	94.5

Table 6.4: Coverage probabilities, $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(1, 10)$.

	Bootstrap			Mode-Quad			
n	100	500	1000	100	500	1000	1500
β_0	89.5	83.5	81.5	75.0	73.5	73.5	72.5
β_1	92.5	92.5	91.5	73.5	77.5	75.0	72.0
γ	94.0	93.5	90.0	65.5	65.5	62.0	61.5
σ^2	17.0	10.5	9.5	12.5	5.5	3.0	0.5
$\nu_{0.75}$	58.5	49.5	39.0	60.5	50.5	38.5	31.5

Table 6.5: Coverage probabilities, $\varepsilon_i \overset{\text{iid}}{\sim} \text{NCT}(1, 3)$.

Chapter 7

Conclusion and Future Works

Quantile estimation conditional on covariates is discussed in the case of right-censored data using empirical likelihood. For the discontinuity in the EL due to the quantile regression constraint as well as right-censoring, a continuity-correction is proposed, and the maximum SCEL estimator is compared through simulations with another method which essentially imputes censored values by a normal distribution.

Simulations show that SCEL method works better for right-skewed error distributions compared to left-skewed ones because of right-censoring, although for a very skewed error distribution, large sample size is required for bias to diminish empirically. Moreover, for slightly skewed error distributions, bootstrap confidence intervals does provide the correct coverage, while for an extremely skewed error distribution, the coverage could be very poor due to the bias of the estimators, especially with comparatively small sample sizes.

There are several directions of future works: Firstly, the current optimization of log SCEL using `nlm` in R relies on numerical approximation of the gradient. If an analytic

gradient can be derived for the log SCEl in a similar fashion as in Chaudhuri et al. (2017), the optimization can be achieved much faster and more accurate, especially combined with the support correction by Chen et al. (2008).

Secondly, the discontinuity of the log EL using the influence function might be corrected too. We expect the performance of the corresponding estimator to be promising, and the confidence intervals are also easier to construct under this setting.

Thirdly, a better way of obtaining the confidence intervals using SCEL should be investigated. One direction is to use a better bootstrap method in the case of censored data. Also, since bias is the main reason for the under-coverage of the confidence intervals when sample size is small, finding a way to approximate and correct this bias would help correct the coverage probabilities.

Lastly, a more thorough comparison of SCEL with a few other methods such as the EL with Buckley-James estimator by Zhou and Li (2008), as well as the influence function approach by He et al. (2016) under the location-scale quantile regression model should be conducted. The selection of smoothing parameter s should also be investigated.

References

Akritas, M. G. et al. (2000), 'The central limit theorem under censoring', *Bernoulli* **6**(6), 1109–1120.

Basset, G. and Koenker, R. (1978), 'Regression quantiles', *Econometrica* **46**(1), 33–50.

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

Chaudhuri, S., Mondal, D. and Yin, T. (2017), 'Hamiltonian monte carlo sampling in bayesian empirical likelihood computation', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(1), 293–320.

Chen, C. (2007), 'A finite smoothing algorithm for quantile regression', *Journal of Computational and Graphical Statistics* **16**(1), 136–164.

Chen, J., Variyath, A. M. and Abraham, B. (2008), 'Adjusted empirical likelihood and its properties', *Journal of Computational and Graphical Statistics* **17**(2), 426–443.

Chen, S. X. and Van Keilegom, I. (2009), 'A review on empirical likelihood methods for regression', *Test* **18**(3), 415–447.

Cox, D. R. (1992), Regression models and life-tables, *in* 'Breakthroughs in statistics', Springer, pp. 527–541.

Efron, B. (1981), 'Nonparametric standard errors and confidence intervals', *canadian Journal of Statistics* **9**(2), 139–158.

He, S. and Huang, X. (2003), 'Central limit theorem of linear regression model under right censorship', *Science in China Series A: Mathematics* **46**(5), 600–610.

He, S., Liang, W., Shen, J. and Yang, G. (2016), 'Empirical likelihood for right censored lifetime data', *Journal of the American Statistical Association* **111**(514), 646–655.

Heuchenne, C. and Van Keilegom, I. (2010), 'Estimation in nonparametric location-scale regression models with censored data', *Annals of the Institute of Statistical Mathematics* **62**(3), 439–463.

Kocherginsky, M., He, X. and Mu, Y. (2005), 'Practical confidence intervals for regression quantiles', *Journal of Computational and Graphical Statistics* **14**(1), 41–55.

Koenker, R. and Bassett, G. (1982), 'Robust tests for heteroscedasticity based on regression quantiles', *Econometrica: Journal of the Econometric Society* pp. 43–61.

Kolaczyk, E. D. (1994), 'Empirical likelihood for generalized linear models', *Statistica Sinica* pp. 199–218.

Lancaster, T. and Jae Jun, S. (2010), 'Bayesian quantile regression methods', *Journal of Applied Econometrics* **25**(2), 287–307.

Lazar, N. A. (2003), 'Bayesian empirical likelihood', *Biometrika* **90**(2), 319–326.

Li, G. and Wang, Q.-H. (2003), 'Empirical likelihood regression analysis for right censored data', *Statistica Sinica* pp. 51–68.

Newey, W. K. and Smith, R. J. (2004), 'Higher order properties of gmm and generalized empirical likelihood estimators', *Econometrica* **72**(1), 219–255.

Ning, J., Qin, J., Asgharian, M. and Shen, Y. (2013), 'Empirical likelihood-based confidence intervals for length-biased data', *Statistics in medicine* **32**(13), 2278–2291.

Noh, J. and Lee, S. (2016), 'Quantile regression for location-scale time series models with conditional heteroscedasticity', *Scandinavian Journal of Statistics* **43**(3), 700–720.

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

Owen, A. (1990), 'Empirical likelihood ratio confidence regions', *The Annals of Statistics* pp. 90–120.

Owen, A. (1991), 'Empirical likelihood for linear models', *The Annals of Statistics* pp. 1725–1747.

Qin, J. and Lawless, J. (1994), 'Empirical likelihood and general estimating equations', *The Annals of Statistics* pp. 300–325.

Reich, B. J. and Smith, L. B. (2013), 'Bayesian quantile regression for censored data', *Biometrics* **69**(3), 651–660.

Schennach, S. M. (2005), 'Bayesian exponentially tilted empirical likelihood', *Biometrika* **92**(1), 31–46.

Schennach, S. M. (2007), 'Point estimation with exponentially tilted empirical likelihood', *The Annals of Statistics* pp. 634–672.

Shen, J., Yuen, K. C. and Liu, C. (2016), 'Empirical likelihood confidence regions for

one-or two-samples with doubly censored data', *Computational Statistics & Data Analysis* **93**, 285–293.

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.

Turnbull, B. W. (1976), 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 290–295.

Wang, Y., You, T. and Lysy, M. (2015), 'A heteroscedastic accelerated failure time model for survival data', *arXiv preprint arXiv:1508.05137* .

Yang, Y. and He, X. (2012), 'Bayesian empirical likelihood for quantile regression', *The Annals of Statistics* **40**(2), 1102–1131.

Zhou, M. (2005), 'Empirical likelihood ratio with arbitrarily censored/truncated data by em algorithm', *Journal of Computational and Graphical Statistics* **14**(3), 643–656.

Zhou, M., Kim, M.-O. and Bathke, A. C. (2012), 'Empirical likelihood analysis for the heteroscedastic accelerated failure time model', *Statistica Sinica* pp. 295–316.

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

Appendices

Appendix A

Derivation of M-step in the EM Algorithm

For simplicity of notation, for now let's denote $\boldsymbol{g}_i = \boldsymbol{g}(\boldsymbol{x}_i, u_i, \boldsymbol{\theta})$. The Lagrangian function for the optimization problem (4.12) (after converting max to min)

$$\begin{aligned} \min_{\omega} \quad & -\sum_{i=1}^{n} q_i \log \omega_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i \cdot \boldsymbol{g}_i = 0 \\ & \sum_{i=1}^{n} \omega_i = 1 \\ & \omega_i \geq 0, \quad i = 1 \cdots, n, \end{aligned} \tag{A.1}$$

is given by

$$L(\omega) = -\sum_{i=1}^{n} q_i \log \omega_i + \gamma(\sum_{i=1}^{n} \omega_i - 1) + \boldsymbol{\lambda}' \sum_{i=1}^{n} \omega_i \cdot \boldsymbol{g}_i \tag{A.2}$$

Then by first order condition of optimality, $\frac{\partial \mathbf{L}(\boldsymbol{w})}{\partial w_i} = 0$ for all $i = 1, \cdots, n$, we have

$$\sum_{i=1}^{n} w_i \frac{\partial \mathbf{L}(\boldsymbol{w})}{\partial w_i} = -\sum_{i=1}^{n} q_i + \sum_{i=1}^{n} w_i \cdot \gamma = 0$$

which gives $\gamma = \sum_{i=1}^{n} q_i$. In fact, we can check that

$$\begin{aligned}
\sum_{i=1}^{n} q_i &= \sum_{i=1}^{n} \Big[\delta_i + \sum_{k:e_k \le e_i} (1-\delta_k) \cdot \tilde{\omega_{ki}}\Big] \\
&= \sum_{i=1}^{n} \Big[\delta_i + \sum_{k:e_k \le e_i} (1-\delta_k) \cdot \frac{\omega_{0i}}{\sum_{l:e_l \ge e_k} \omega_{0l}}\Big] \\
&= \sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} \sum_{k:e_i \ge e_k} (1-\delta_k) \cdot \frac{\omega_{0i}}{\sum_{l:e_l \ge e_k} \omega_{0l}} \\
&= \sum_{i=1}^{n} \delta_i + \sum_{k=1}^{n} (1-\delta_k) \cdot \frac{\sum_{i:e_i \ge e_k} \omega_{0i}}{\sum_{l:e_l \ge e_k} \omega_{0l}} \\
&= \sum_{i=1}^{n} \delta_i + \sum_{k=1}^{n} (1-\delta_k) \\
&= n.
\end{aligned} \tag{A.3}$$

Substituting $\gamma = n$ back to equation (A.2) and with $\frac{\partial \mathbf{L}(\omega)}{\partial \omega_i} = 0$, we have

$$w_i = \frac{q_i}{n + \boldsymbol{\lambda}' \boldsymbol{g}_i} \tag{A.4}$$

Also, because of the constraints $\sum_{i=1}^{n} \omega_i \cdot \boldsymbol{g}_i = 0$, we have

$$\sum_{i=1}^{n} \frac{q_i \cdot \boldsymbol{g}_i}{n + \boldsymbol{\lambda}' \boldsymbol{g}_i} = 0 \tag{A.5}$$

Instead of sovling equation (A.5) directly, we can transform it into a minimization problem of a convex function

$$f(\boldsymbol{\lambda}) = -\sum_{i=1}^{n} q_i \log(n + \boldsymbol{\lambda}' \boldsymbol{g}_i) \tag{A.6}$$

Notice that from the constraints of the optimization problem (A.1), we must have $\omega_i \leq 1$, and with (A.4), we should only consider λ for which

$$n + \boldsymbol{\lambda}' \boldsymbol{g}_i \geq q_i$$

By a similar argument as in Owen (1990), we can define a quadratic function $\log^\sharp(x_i)$ which matches $\log(x_i)$ and its first two derivatives at q_i (through Taylor expansion). That is, at $x_i = q_i$

$$\begin{aligned} \log(x_i) &= \log q_i + \frac{1}{q_i}(x_i - q_i) - \frac{1}{q_i^2}(x_i - q_i)^2 + o(x_i^2) \\ &= -\frac{1}{2q_i^2}x_i^2 + \frac{2}{q_i}x_i - \frac{3}{2} + \log q_i + o(x_i^2) \end{aligned} \tag{A.7}$$

which gives the expression (4.15). So that we can combine (A.6) with the support restriction as

$$f^\sharp(\boldsymbol{\lambda}) = -\sum_{i=1}^{n} q_i \log^\sharp(n + \boldsymbol{\lambda}' \boldsymbol{g}_i) \tag{A.8}$$

to obtain $\hat{\lambda}$ as in (4.14).

□

Appendix B

Equivalence between the EM Algorithm and log CEL Maximization

Does the EM algorithm produce equivalent results as directly maximizing (5.1), or with smoothing, (5.4) subject to the same constraints? We know that EM algorithm is equivalent to maximizing the marginal likelihood of the observed data, but it does not seem obvious that (5.1) or (5.4) are the corresponding marginal distributions or not. However, we can follow the method by Turnbull (1976) and Zhou (2005) and show that this is indeed the case.

Before continuity correction, recall that in the EM algorithm, we have

$$\hat{\omega}_i(\boldsymbol{\theta}) = \frac{q_i}{n + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i(\boldsymbol{\theta})}, \tag{B.1}$$

where

$$
\begin{aligned}
q_i &= \delta_i + \sum_{k=1}^{n} (1-\delta_k) \mathbb{1}(e_k \le e_i) \cdot \tilde{\omega}_{ki}, \\
\tilde{\omega}_{ki} &= \frac{\omega_{0i}}{\sum_{l=1}^{n} \mathbb{1}(e_k \le e_l) \cdot \omega_{0l}}, \\
\hat{\boldsymbol{\lambda}} &= \arg\max_{\boldsymbol{\lambda}} \sum_{i=1}^{n} q_i \cdot \log^{\sharp}\Big(n + \boldsymbol{\lambda}' \boldsymbol{g}(x_i, y_i; \boldsymbol{\theta})\Big).
\end{aligned}
\tag{B.2}
$$

To maximize the log CEL in (5.1) with respect to ω subject to the constraints for a given $\boldsymbol{\theta}$, we look at the (negated) Lagrangian function which is shown to be concex. $\boldsymbol{\theta}$ is omitted since it is fixed here:

$$
\begin{aligned}
L(\omega, \gamma, \boldsymbol{\lambda}) = &- \sum_{i=1}^{n} \Big[\delta_i \log \omega_i + (1-\delta_i) \log\big(\sum_{j=1}^{n} \mathbb{1}(e_j \ge e_i) \cdot \omega_j\big) \Big] \\
&+ \gamma\big(\sum_{i=1}^{n} \omega_i - 1\big) + \boldsymbol{\lambda}' \sum_{i=1}^{n} \omega_i \cdot \boldsymbol{g}_i.
\end{aligned}
\tag{B.3}
$$

At optimality, the first derivatives of the Lagrangian function with respect to w, γ and λ must be 0 respectively. Denote the first derivative of L w.r.t w_k as $d_k(w)$, which is

$$
d_k(\omega) = -\left[\frac{\delta_k}{\omega_k} + \sum_{i=1}^{n} \frac{(1-\delta_i)\mathbb{1}(e_k \ge e_i)}{\sum_{j=1}^{n} \mathbb{1}(e_j \ge e_k) \cdot \omega_k} \right] + \gamma + \boldsymbol{\lambda}' \boldsymbol{g}_i. \tag{B.4}
$$

Since $\frac{\partial L}{\partial \omega_k} = 0$ for all $k = 1, \cdots, n$ at optimality, we have

$$
\begin{aligned}
\sum_{k=1}^{n} d_k(\omega) \cdot \omega_k &= -\sum_{k=1}^{n} \left[\frac{\delta_k}{\omega_k} + \frac{(1-\delta_k)\sum_{j=1}^{n} \mathbb{1}(e_j \ge e_k)}{\sum_{j=1}^{n} \mathbb{1}(e_j \ge e_k) \cdot \omega_k} \right] \cdot \omega_k + \gamma \sum_{k=1}^{n} \omega_k \\
&= -n + \gamma = 0.
\end{aligned}
\tag{B.5}
$$

This gives $\gamma = n$, so that we can replace γ by n in the Lagrangian function.

With (B.4) and $\sum_{i=1}^{n} q_i = n = \gamma$, we can write Eq (B.1) as

$$\begin{aligned}\hat{\omega}_i &= \frac{\omega_{0i}}{n + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i} \cdot \frac{q_i}{\omega_{0i}} \\ &= \frac{\omega_{0i}}{n + \boldsymbol{\lambda}' \boldsymbol{g}_i} \cdot \left(-d_i(w_0) + n + \boldsymbol{\lambda}' g_i\right) \\ &= \left(1 - \frac{d_i(\omega_0)}{n + \hat{\boldsymbol{\lambda}}' g_i}\right) \cdot \omega_{0i},\end{aligned} \tag{B.6}$$

where subscript "0" indicates the value at the previous iteration during the EM algorithm. Therefore, if ω_0 gives the optimal value for the original log EL, since $d_i(\omega_0) = 0$, Eq (B.6) implies $\hat{\omega}_i = \omega_{0i}$, so that EM converges; conversely, when EM converges, we have $\hat{\omega}_i = \omega_{0i}$, so that $d_i(\omega_0) = 0$, and thus the original log EL reaches optimality at ω_0.

After continuity correction, the Lagrangian function and the EM algorithm only have the indicator functions replaced by S. Since the function S does not have ω as its argument, everything follows as above and Eq (B.6) still holds. Therefore, the smoothed EM algorithm is indeed equivalent to maximizing the smoothed version of the original log EL under the constraints.

In Appendix C, we provide an alternative argument for the validity of the algorithm, that the value of the objective function is indeed non-decreasing during the iterations.

Appendix C

Another Proof of Validity for the EM Algorithm

Here we discuss an alternative way of showing the algorithm indeed produces monotone increasing sequence of log EL's. The following argument is similar to the one by Turnbull (1976).

We have shown that log CEL is a concave function and it is differentiable, and we know that for a differentiable concave function, the difference in the function values of two points is bounded by its first-order Taylor approximation. That is, for two valid and

consecutive steps $\omega^{(n)}$ and $\omega^{(n+1)}$, and using (B.6)

$$\begin{aligned}
\ell(\omega^{(n+1)}) - \ell(\omega^{(n)}) &\geq \sum_{i=1}^{n} \frac{\partial \ell}{\partial \omega_j} \cdot (\omega_i^{(n+1)} - \omega_i^{(n)}) \\
&= \sum_{i=1}^{n} \frac{\partial \ell}{\partial w_j} \cdot \frac{-d_i(\omega^{(n)}) \cdot \omega_i^{(n)}}{n + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i} \\
&= \sum_{i=1}^{n} (-d_i(\omega^{(n)}) + n + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i) \cdot \frac{-d_i(\omega^{(n)}) \cdot \omega_i^{(n)}}{n + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i} \\
&= \sum_{i=1}^{n} \frac{d_i^2(\omega^{(n)}) \cdot \omega_i^{(n)}}{n + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i} \geq 0
\end{aligned} \tag{C.1}$$

as long as $n + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i \geq 0$. Also, the last step uses the fact that the constraints are satisfied at ω and thus $\sum_{i=1}^{n} d_i(\omega) \cdot \omega_i = 0$.

Notice that the function $\log^\sharp$ in (4.14) always gives $\hat{\boldsymbol{\lambda}}$ such that $n + \boldsymbol{\lambda}' \boldsymbol{g}_i \geq 0$, because it is a modification of the log function such that for any $\boldsymbol{\lambda}$ such that $n + \boldsymbol{\lambda}' \boldsymbol{g}_i < 0$, the function value is smaller than any other $\tilde{\boldsymbol{\lambda}}$ such that $n + \tilde{\boldsymbol{\lambda}}' \boldsymbol{g}_i \geq 0$, and thus contradicts that $\hat{\boldsymbol{\lambda}}$ give the maximum value.

This argument is also not affected by the smoothing function, so it also applies to the smoothed version of EM.

□

Appendix D

Proof of Proposition 1

We first show that the negation of the objective function in (4.7) is convex (Part I), then we will show that the Lagrangian function is convex, and it is strictly convex if and only if only if no observation other than the one with the largest survival time can be a censored observation (Part II).

Part I:

Maximizing the objective function in (4.7) is equivalent to

$$\min -\sum_{i=1}^{n}\Big[\delta_i \log \omega_i + (1-\delta_i)\log(\sum_{j:e_j\geq e_i}\omega_j)\Big]. \tag{D.1}$$

Let $A(\omega) = -\sum_{i=1}^{n}\delta_i \log\omega_i$ and $B(\omega) = -\sum_{i=1}^{n}(1-\delta_i)\log(\sum_{j:e_j\geq e_i}\omega_j)$, the first and the section terms of (D.1) respectively. Since the sum of convex functions is still convex, we only need to show that $A(\omega)$ and $B(\omega)$ are both convex. We show this by finding the Hessian matrices of them.

The second derivative of $A(\omega)$ w.r.t any w_i is

$$\tau_i := \frac{\partial^2 A(\omega)}{\partial w_i} = \frac{\delta_i}{w_i^2} \geq 0.$$

The Hessian matrix is a diagonal matrix with the above entries on the diagonal, therefore, the Hessian matrix for $A(\omega)$ is indeed positive semidefinite (PSD). Thus $A(\omega)$ is a convex function.

Assumed w.l.o.g. that the residual e_i's are ordered, and the weights and indicators are ordered accordingly. We have

$$B(\omega) = -\Big[(1-\delta_1)\log(\sum_{j\geq i}\omega_j) + \cdots + (1-\delta_n)\log(\sum_{j\geq i}\omega_j)\Big].$$

First derivative of $B(\omega)$ w.r.t ω_1 is

$$\frac{\partial B(\omega)}{\partial \omega_1} = -\frac{1-\delta_1}{\sum_{i=1}^n \omega_i},$$

so that second derivative of $B(\omega)$ first w.r.t w_1 then w.r.t w_i for any $i \geq 1$ is

$$\sigma_1 := \frac{\partial^2 B(\omega)}{\partial \omega_i \partial \omega_1} = \frac{1-\delta_1}{(\sum_{i=1}^n \omega_i)^2} \geq 0.$$

First derivative of $B(\omega)$ w.r.t ω_2 is

$$\frac{\partial B(\omega)}{\partial \omega_2} = -\frac{1-\delta_1}{\sum_{i=1}^n \omega_i} - \frac{1-\delta_2}{\sum_{i=2}^n \omega_i},$$

so that second derivative of $B(\omega)$ first w.r.t ω_2 then w.r.t ω_1 is

$$\frac{\partial^2 B(\omega)}{\partial \omega_1 \partial \omega_2} = \frac{1-\delta_1}{(\sum_{i=1}^n \omega_i)^2} = \sigma_1,$$

and the second derivative of $B(\omega)$ first w.r.t ω_2 then w.r.t. ω_i for any $i \geq 2$ is

$$\sigma_2 := \frac{\partial^2 B(\omega)}{\partial \omega_i \partial \omega_2} = \frac{1-\delta_1}{(\sum_{i=1}^n \omega_i)^2} + \frac{1-\delta_2}{(\sum_{i=2}^n \omega_i)^2}.$$

Then by induction, we obtain the $n \times n$ Hessian matrix in the form

$$H = \begin{bmatrix} \sigma_1 & \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma_2 & \sigma_2 & \cdots & \sigma_2 \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_n \end{bmatrix}.$$

We can do row operations on the matrix — subtracting i-th row from $i+1$-th row for $i = 2, \cdots, n$ — without changing its determinant, we get

$$H = \begin{bmatrix} \sigma_1 & \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ 0 & \sigma_2 - \sigma_1 & \sigma_2 - \sigma_1 & \cdots & \sigma_2 - \sigma_1 \\ 0 & 0 & \sigma_3 - \sigma_2 & \cdots & \sigma_3 - \sigma_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n - \sigma_{n-1} \end{bmatrix}.$$

Since it is an upper triangular matrix, its determinant is the product of the diagonal elements. These diagonal elements are all nonnegative since $\sigma_{i+1} \geq \sigma_i$ for all $i = 1, \cdots, n-1$ by the calculation above.

Therefore, the objective function in (4.7) is indeed a convex function.

Part II:

Let $f(\omega)$ denote the objective function in (D.1), $m(\omega) = \sum_{i=1}^{n} \omega_i - 1$, and $h(\omega) = \sum_{i=1}^{n} \omega_i \cdot g_i$. Then the Lagrangian function of the optimization problem (4.7) is given by

$$L(\omega) = f(\omega) + \gamma \cdot m(\omega) + \boldsymbol{\lambda}' h(\omega). \tag{D.2}$$

To determine the convexity of the Lagrangian function, we take the second derivative so we get the Hessian matrix H^* as

$$\begin{aligned} H^* &= \frac{\partial^2 L(\omega, \gamma, \lambda)}{\partial \omega^2} \\ &= \frac{\partial^2 L(\omega)}{\partial \omega^2} \\ &= \frac{\partial^2 A(\omega)}{\partial \omega^2} + \frac{\partial^2 B(\omega)}{\partial \omega^2} = H, \end{aligned}$$

since the second derivatives of m and h are both $\mathbf{0}$.

From Part I,

$$\frac{\partial^2 A(\omega)}{\partial \omega^2} = \begin{bmatrix} \tau_1 & 0 & \cdots & 0 \\ \vdots & \tau_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \tau_n \end{bmatrix}$$

and

$$\frac{\partial^2 B(\omega)}{\partial \omega^2} = \begin{bmatrix} \sigma_1 & \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma_2 & \sigma_2 & \cdots & \sigma_2 \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_n \end{bmatrix}.$$

As in Part I, we can perform row operations on the matrix H without changing its determinant. Here, let P be the permutation matrix such that PM subtracts ith row by

$(i+1)$th row in matrix $M \in \mathbb{R}^{n\times n}$ for $i = 1, \cdots, n-1$. Then we have

$$PH^* = \begin{bmatrix} \tau_1 & -\tau_2 + \sigma_1 - \sigma_2 & \sigma_1 - \sigma_2 & \cdots & \sigma_1 - \sigma_2 \\ 0 & \tau_2 & -\tau_3 + \sigma_2 - \sigma_3 & \cdots & \sigma_2 - \sigma_3 \\ 0 & 0 & \tau_3 & \cdots & \sigma_4 - \sigma_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \tau_n + \sigma_n \end{bmatrix}.$$

As what we have shown in Part I, the above matrix is indeed PSD. Here we would like to find the sufficient and necessary conditions such that the above matrix is (strictly) positive definite (PD).

Notice that for any observation, it is either censored or uncensored, so we have $\tau_n + \sigma_n > 0$. Let H_i be the upper $i \times i$ matrix of H, for $i = 1, \cdots, n-1$. Then for PH^* to be PD, the sufficient condition is that the determinant of H_i

$$|H_i| = \prod_{i=1}^{i} \tau_i$$

are all positive for $i = 1, \cdots, n-1$.

This means that the sufficient condition is that $\tau_i > 0, i = 1, \cdots, n-1$. If $\tau_i > 0, i = 1, \cdots, n-1$, then $\sigma_i = 0, i = 1, \cdots, n-1$. So we obtained a upper triangular matrix and all the diagonal elements are strictly positive. Therefore, the necessory condition is that $\tau_i > 0, i = 1, \cdots, n-1$ as well.

Hence the negated Lagrangian function for the optimization problem (4.7) is convex, and it is strictly convex if and only if no observation other than the one with the largest survival time can be a censored observation.

□

Appendix E

Proof of Proposition 2

We will show that the negated log SCEL in (5.4) is convex with respect to ω.

Denote $A(\omega) = -\sum_{i=1}^{n} \delta_i \log \omega_i$ and $B(\omega) = -\sum_{i=1}^{n}(1-\delta_i)\log\big(\sum_{j=1}^{n} S(e_i - e_j)\cdot\omega_j\big)$, the negated first and the section terms of (5.4) respectively, for a fixed $\boldsymbol{\theta}$. We will show that both of the two terms are convex functions of ω.

The second derivative of $A(\omega)$ w.r.t any ω_i is

$$\frac{\partial^2 A(\omega)}{\partial \omega_i} = \frac{\delta_i}{\omega_i^2} \geq 0.$$

The Hessian matrix is a diagonal matrix with the above entries on the diagonal, therefore, the Hessian matrix for $A(\omega)$ is indeed PSD. Thus $A(\omega)$ is a convex function.

The first derivative of $B(\omega)$ w.r.t ω_1 is

$$\frac{\partial B(\omega)}{\partial \omega_1} = -\sum_{i=1}^{n} \frac{(1-\delta_i)\cdot S'(e_i - e_1)}{\sum_{j=1}^{n} S(e_i - e_j)\cdot\omega_j}. \tag{E.1}$$

Denote the second derivative of $B(\omega)$ w.r.t w_1 as

$$\sigma_{11} := \frac{\partial^2 B(\omega)}{\partial \omega_1^2} = \sum_{i=1}^{n} \frac{(1-\delta_i) \cdot S'(e_i - e_1)^2}{\left(\sum_{j=1}^{n} S(e_i - e_j) \cdot \omega_j\right)^2}.$$

The second derivative of $B(\omega)$ w.r.t. ω_1 and then ω_2 as

$$\sigma_{12} := \frac{\partial^2 B(\omega)}{\partial \omega_1 \partial \omega_2} = \sum_{i=1}^{n} \frac{(1-\delta_i) \cdot S'(e_i - e_1) \cdot S'(e_i - e_2)}{\left(\sum_{j=1}^{n} S(e_i - e_j) \cdot \omega_j\right)^2}.$$

Continue this way the hessian matrix of $B(w)$ can be written as

$$H = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2n} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \cdots & \sigma_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \sigma_{3n} & \cdots & \sigma_{nn} \end{bmatrix}.$$

We can write $H = \sum_{i=1}^{n} H_i$, where

$$\begin{aligned} H_i &= \frac{1-\delta_i}{\left(\sum_{j=1}^{n} (e_i - e_j) \cdot \omega_j\right)^2} \cdot \begin{bmatrix} \eta_{i11} & \eta_{i12} & \eta_{i13} & \cdots & \eta_{i1n} \\ \eta_{i12} & \eta_{i22} & \eta_{i23} & \cdots & \eta_{i2n} \\ \eta_{i13} & \eta_{i23} & \eta_{i33} & \cdots & \eta_{i3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta_{i1n} & \eta_{i2n} & \eta_{i3n} & \cdots & \eta_{inn} \end{bmatrix} \\ &= \frac{1-\delta_i}{\left(\sum_{j=1}^{n} (e_i - e_j) \cdot \omega_j\right)^2} \cdot \bar{H}_i, \end{aligned} \tag{E.2}$$

and where

$$\eta_{ijk} = S'(e_i - e_j) \cdot S'(e_i - e_k)$$

If each $\bar{H}_i$ such that $\delta_i = 0$ is PSD, then H is PSD. In fact, each $\bar{H}_i$ can be written as the outer product of two vectors of the same value

$$\bar{H}_i = \begin{bmatrix} S'(e_i - e_1) \\ S'(e_i - e_2) \\ -s \\ \vdots \\ S'(e_i - e_n) \end{bmatrix} \begin{bmatrix} S'(e_i - e_1) \\ S'(e_i - e_2) \\ -s \\ \vdots \\ S'(e_i - e_n) \end{bmatrix}',$$

where $-s$ appears at the i-th entry. This means that $\bar{H}_i$ is a rank 1 matrix with one and only one positive eigenvalue.[1]

□

[1] $A = vv'$, then $Av = vv'v = v||v||^2 =: v\lambda$, which means the eigenvalue $\lambda = ||v||^2 > 0$, as long as $v \neq \mathbf{0}$.