Maximum likelihood estimator for double-truncation data under a special exponential family

Presenter: Ya-Hsuan Hu, Graduate Institute of Statistics, National Central University, Taiwan

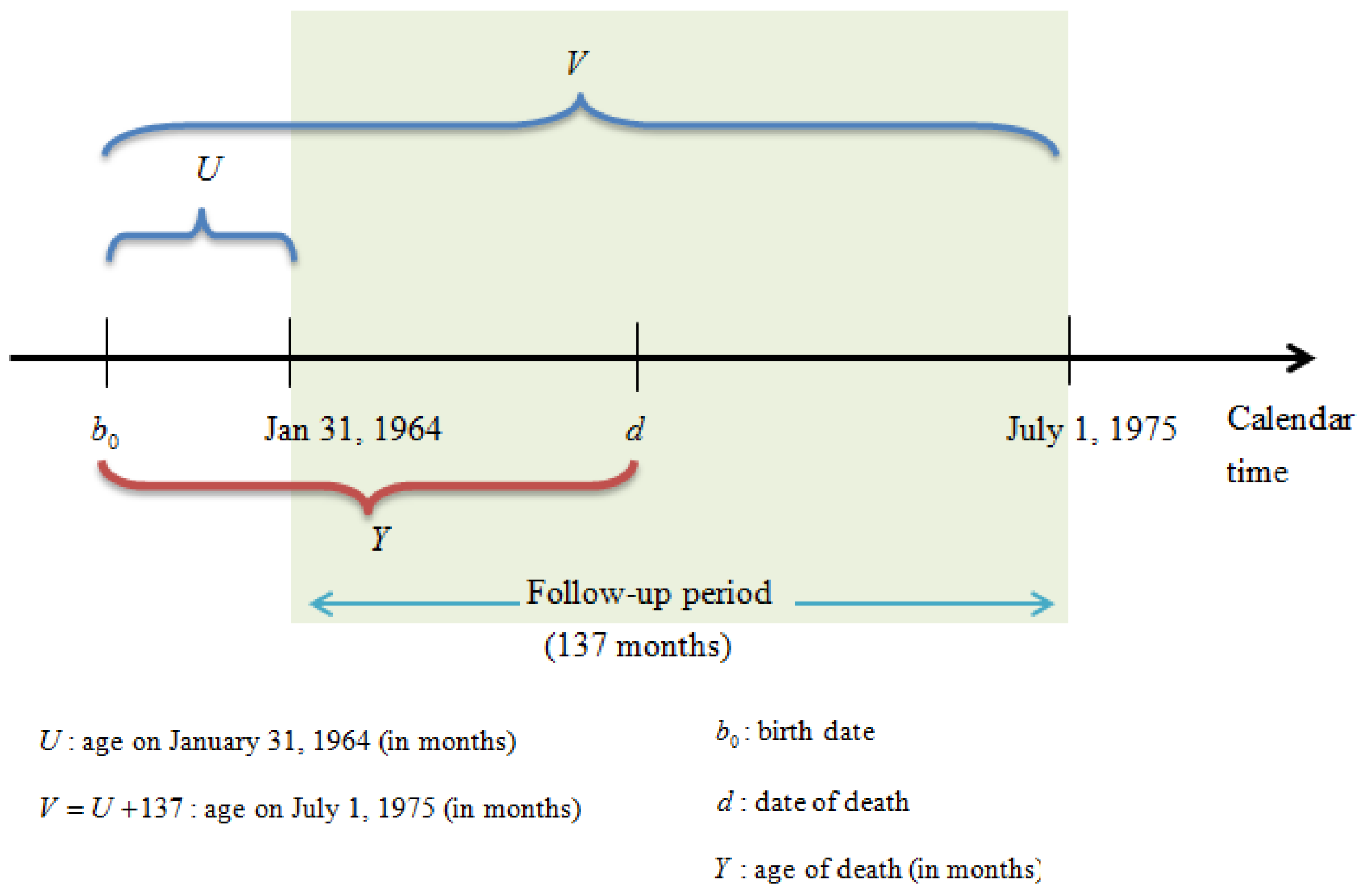
Advisor: Prof. Takeshi Emura, Graduate Institute of Statistics, National Central University, Taiwan

Abstract: Truncation often occurs in lifetime data analysis, where samples are collected under certain time constraints. We consider parametric inference when random samples are subject to double-truncation, i.e., both left- and right-truncations. In particular, we consider the proposal of Efron and Petrosian (1999)[2], which is based on special exponential family (SEF). We develop computational algorithms for Newton-Raphson and fixed point iteration techniques to obtain maximum likelihood estimator (MLE) of the parameters, and then compare the performance of these two methods by simulations. We observe that the Newton-Raphson method has faster rate of convergence than the fixed point iteration for moderate sample sizes. Also, we study the asymptotic properties of the MLE based on the asymptotic theory of independent but not identical random variables. Real data are used for illustration.

1. Introduction

Double-truncation of survival data occurs when only those individuals whose event time lies within a certain follow-up period are observed. For our example of life expectancy data of Hyde (1980)[3], the lifetime of the Channing House residents suffer from doubly truncated. (Figure below)

Statistical inference for doubly truncated data have been popular research with a variety of applications.[2,5,6]


V
U
b_0
Jan 31, 1964
d
July 1, 1975
Calendar time
Y
Follow-up period
(137 months)
U : age on January 31, 1964 (in months)
$V=U+137$: age on July 1, 1975 (in months)
b_0 : birth date
d : date of death
Y : age of death (in months)

If we ignore double-truncation, it will introduces a systematic bias in estimation.

2. Specially exponential family

Efron and Tibshirani (1996)[1] first introduced the special exponential family (SEF). The idea of the SEF combined two methods. In order to estimates the probability density, we can use maximum likelihood fitting within some parametric family or nonparametric methods. Combining these two methods are the specially exponential family.

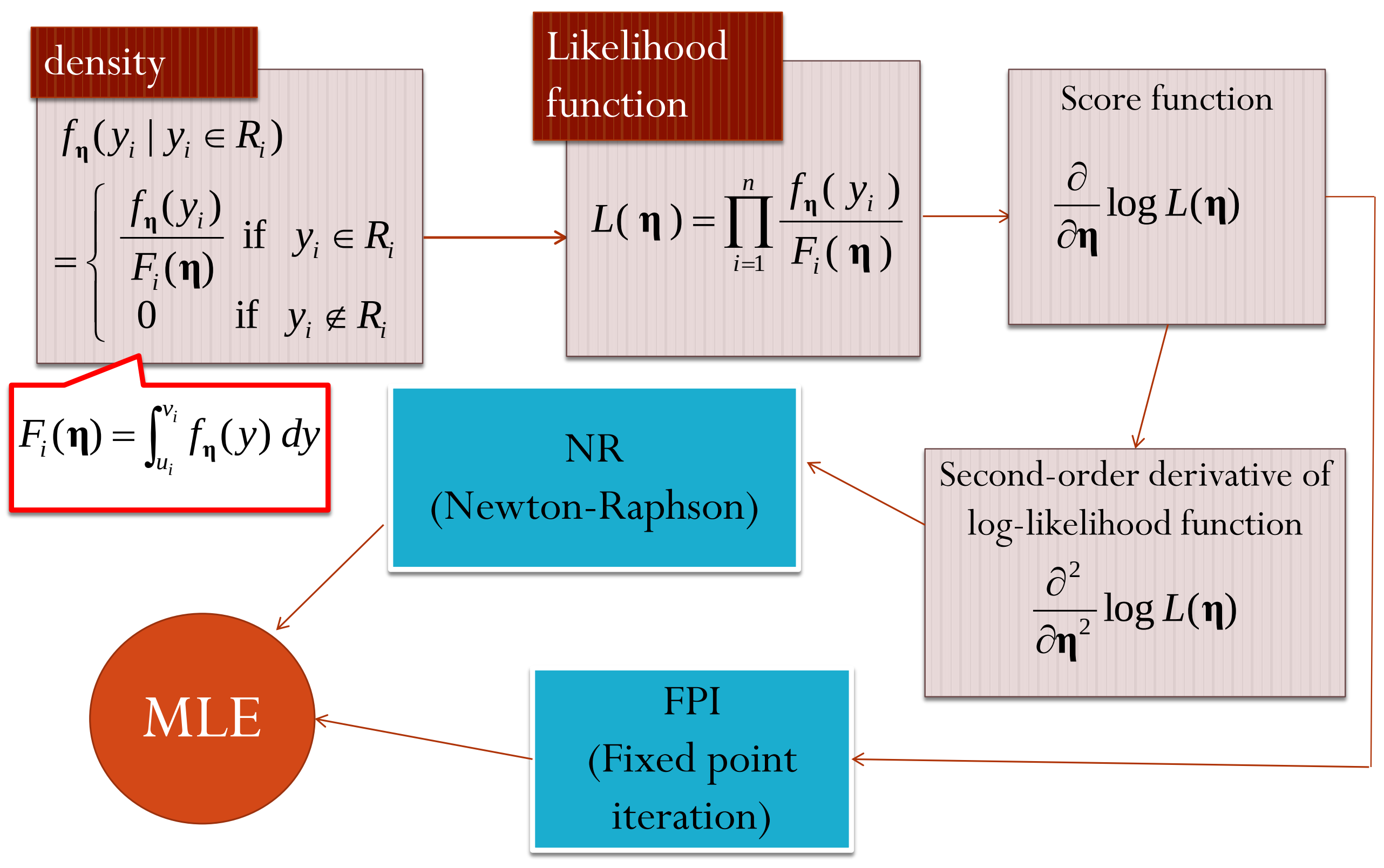
The density of SEF is

$$f_{\boldsymbol\eta}(y)=\exp\{\boldsymbol\eta^{\mathrm T}\mathbf t(y)-\phi(\boldsymbol\eta)\},\quad y\in\mathcal Y,$$

where $\mathcal Y\subset\Re$ is the support of Y , $\mathbf t(y)=(y,y^2,\cdots,y^k)^{\mathrm T}$, and $\boldsymbol\eta=(\eta_1,\eta_2,\cdots,\eta_k)^{\mathrm T}\subset\Theta\subset\Re^k$. $\phi(\boldsymbol\eta)$ is chosen to make $\int_{\mathcal Y} f_{\boldsymbol\eta}(y)\,dy=1$, that is, $\phi(\boldsymbol\eta)=\log[\int_{\mathcal Y}\exp\{\boldsymbol\eta^{\mathrm T}\mathbf t(y)\}\,dy]$, provided $\int_{\mathcal Y}\exp\{\boldsymbol\eta^{\mathrm T}\mathbf t(y)\}\,dy<\infty.$

3. Method

We introduce the likelihood function with the doubly truncated data proposed by Efron and Petrosian (1999)[2] . Let $R_i=[u_i,v_i]$ be the interval, where u_i is the left-truncation limit and v_i is the right-truncation limit. We consider estimation of density $f(y)$ when the random samples $y_1,y_2,\cdots,y_n$ are subject to the constraints $y_i\in R_i$. Then the following chart is how to obtain the maximum likelihood estimators.



density

$$f_{\boldsymbol\eta}(y_i\mid y_i\in R_i)=\begin{cases}\dfrac{f_{\boldsymbol\eta}(y_i)}{F_i(\boldsymbol\eta)} & \text{if } y_i\in R_i\\ 0 & \text{if } y_i\notin R_i\end{cases}$$

$$F_i(\boldsymbol\eta)=\int_{u_i}^{v_i} f_{\boldsymbol\eta}(y)\,dy$$

Likelihood function

$$L(\boldsymbol\eta)=\prod_{i=1}^{n}\frac{f_{\boldsymbol\eta}(y_i)}{F_i(\boldsymbol\eta)}$$

Score function

$$\frac{\partial}{\partial\boldsymbol\eta}\log L(\boldsymbol\eta)$$

Second-order derivative of log-likelihood function

$$\frac{\partial^2}{\partial\boldsymbol\eta^2}\log L(\boldsymbol\eta)$$

NR (Newton-Raphson)

FPI (Fixed point iteration)

MLE

Finding the maximum likelihood estimator, we can directly calculated. But, sometimes even for common densities, it is difficult to find maximum likelihood estimator. Here, we do it by using numerical method. One is Newton-Raphson method. Another is fixed-point iteration.

Newton-Raphson method

Step 1: Choose the initial value $\boldsymbol\eta^{(0)}$.

Step 2: Consider the recursive process

$$\boldsymbol\eta^{(k+1)}-\boldsymbol\eta^{(k)}=\left[-\frac{\partial^2}{\partial\boldsymbol\eta^2}\log L(\boldsymbol\eta^{(k)})\right]^{-1}\frac{\partial}{\partial\boldsymbol\eta}\log L(\boldsymbol\eta^{(k)})$$

Step 3: The iteration procedure then continuous until convergence, i.e., until $|\boldsymbol\eta^{(k+1)}-\boldsymbol\eta^{(k)}|<10^{-4}$.

Fixed-point iteration method

To solve the score function $S(\boldsymbol\eta)=\partial\log L(\boldsymbol\eta)/\partial\boldsymbol\eta=0$, we rewrite $S(\boldsymbol\eta)=0$ as $\boldsymbol\eta=g(\boldsymbol\eta)$.

Step 1: Choose the initial value $\boldsymbol\eta^{(0)}$.

Step 2: Consider the recursive process

$$\boldsymbol\eta^{(k+1)}=g(\boldsymbol\eta^{(k)})$$

Step 3: The iteration procedure then continuous until convergence, i.e., until $|\boldsymbol\eta^{(k+1)}-\boldsymbol\eta^{(k)}|<10^{-4}$.

The advantage of the fixed-point iteration is that it does not require the second derivatives of the log-likelihood which are often complicated. The disadvantage is that the choice of g is not unique.

4. Theory

Our model is not iid case, only independent but not identical, we need to consider more assumptions to prove consistency.

Assumption (A) There exists an open subset ω of Θ containing the true parameter point $\boldsymbol\eta^0=(\eta_1^0,\eta_2^0,\eta_3^0)$.

Assumption (B) There exists a 3×3 positive definite matrix

$$I(\boldsymbol\eta)=\{I_{jk}(\boldsymbol\eta)\}_{j,k=1,2,3}$$

such that $\sum_{i=1}^{n}I_{ijk}(\boldsymbol\eta)/n\to I_{jk}(\boldsymbol\eta)$ for $j,k=1,2,3$ and all $\boldsymbol\eta\in\omega$ as $n\to\infty$.

Assumption (C) Suppose that there exists a measurable function M_{jkl} such that for all $i=1,2,\cdots,n$ and $\boldsymbol\eta\in\omega$

$$\left|\frac{\partial^3}{\partial\eta_j\partial\eta_k\partial\eta_l}\log f_i(y_i\mid\boldsymbol\eta)\right|\le M_{jkl}(y_i),$$

where

$$m_{ijkl}=E_{\boldsymbol\eta^0}\{M_{jkl}(Y_i)\}<\infty \text{ for all } j,k,l \text{ and } i=1,2,\cdots,n.$$

And assume that $\sum_{i=1}^{n}m_{ijkl}^2/n\to m_{jkl}^2$ and $\sum_{i=1}^{n}m_{ijkl}/n\to m_{jkl}$ as $n\to\infty$.

Assumption (D) Suppose that there exists a measurable function W_{jk} such that for all $i=1,2,\cdots,n$ and $\boldsymbol\eta\in\omega$

$$\left|\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log f_i(y_i\mid\boldsymbol\eta)\right|\le W_{jk}(y_i),$$

where

$$w_{ijk}=E_{\boldsymbol\eta^0}\{W_{jk}(Y_i)\}<\infty \text{ for all } j,k \text{ and } i=1,2,\cdots,n.$$

And assume that $\sum_{i=1}^{n}w_{ijk}/n\to w_{jk}$ as $n\to\infty$.

Theorem: If Assumptions (A)-(D) holds, then $\hat\eta_{jn}$ is consistent for estimating η_j .

5. Simulations

We compare the performance of the Newton-Raphson method and Fixed-point iteration via simulation. Data are generated from the special exponential family. Also, we consider that all the simulations are conducted under $P(U\le Y\le V)\approx0.5$.

We report the MLE results in terms of the average of 100 Monte Carlo replications. The notation of the table defined as following :

① $E(\hat{\boldsymbol\eta})=\sum_{i=1}^{n}\hat{\boldsymbol\eta}^{(i)}/n,$ and $MSE(\hat{\boldsymbol\eta})=E(\hat{\boldsymbol\eta}-\boldsymbol\eta)^2.$

② AI=Average number of iteration.

Simulation results under one-parameter special exponential family with parameter $\eta>0$

	Initial value	method	$E(\hat\eta)$	$MSE(\hat\eta)$	AI
$\eta=3$	$\eta^{(0)}=3$	FPI	3.093683	0.2314271	12.73
$n=100$		NR	3.093673	0.2314795	4.49
	$\eta^{(0)}=\frac{1}{y_{(n)}-\bar y}$	FPI	3.093723	0.2314482	12.28
		NR	3.093673	0.2314795	4.29
$\eta=3$	$\eta^{(0)}=3$	FPI	3.038339	0.09521812	12.06
$n=200$		NR	3.038342	0.09525377	4.24
	$\eta^{(0)}=\frac{1}{y_{(n)}-\bar y}$	FPI	3.038394	0.09523499	11.9
		NR	3.038342	0.09525377	4.19

Simulation results under one-parameter special exponential family with parameter $\eta<0$

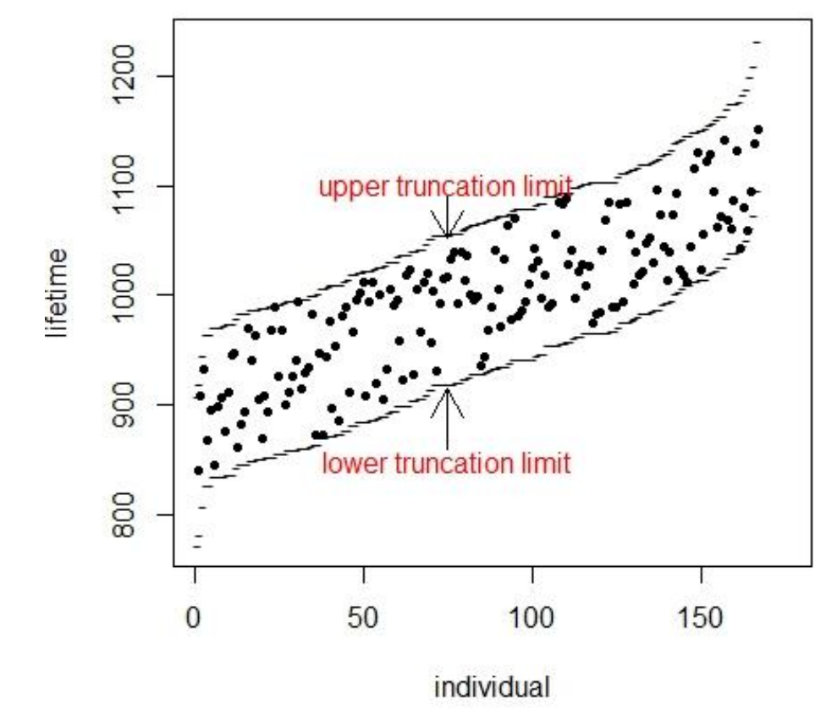
	Initial value	method	$E(\hat\eta)$	$MSE(\hat\eta)$	AI
$\eta=-3$	$\eta^{(0)}=-3$	FPI	-3.036629	029635	13.07
$n=100$		NR	-3.036605	0.29641	4.51
	$\eta^{(0)}=\frac{1}{y_{(1)}-\bar y}$	FPI	-3.036634	0.29635	12.16
		NR	-3.036605	0.29641	4.34
$\eta=-3$	$\eta^{(0)}=-3$	FPI	-3.002451	0.10555	12.23
$n=200$		NR	-3.002436	0.10559	4.25
	$\eta^{(0)}=\frac{1}{y_{(1)}-\bar y}$	FPI	-3.002467	0.10556	11.63
		NR	-3.002436	0.10559	4.18

Simulation results under normal distribution with parameter (μ,σ^2) which correspond to $(\eta_1,\eta_2)=(\mu/\sigma^2,-1/2\sigma^2)$

True (η_1,η_2)	Initial value	method	$E(\hat\eta_1)$	$E(\hat\eta_2)$	$MSE(\hat\eta_1)$	$MSE(\hat\eta_2)$	AI
(30,-0.5)	(30,-0.5)	FPI	30.61147	-0.51085	61.09232	0.01697	36.14
$n=100$		NR	30.61134	-0.51085	61.09615	0.01695	5.23
	$(\frac{\bar y}{s^2},\frac{-1}{2s^2})$	FPI	30.61163	-0.51086	61.09445	0.016974	38.89
		NR	30.61134	-0.51085	61.09615	0.01695	6.21
(30,-0.5)	(30,-0.5)	FPI	31.3292	-0.52253	31.86291	0.00885	30.99
$n=200$		NR	31.32923	-0.52253	31.86499	0.00885	4.93
	$(\frac{\bar y}{s^2},\frac{-1}{2s^2})$	FPI	31.32947	-0.52253	31.86498	0.00885	35.12
		NR	31.32923	-0.52253	31.86499	0.00885	6.07

6. Data Analysis

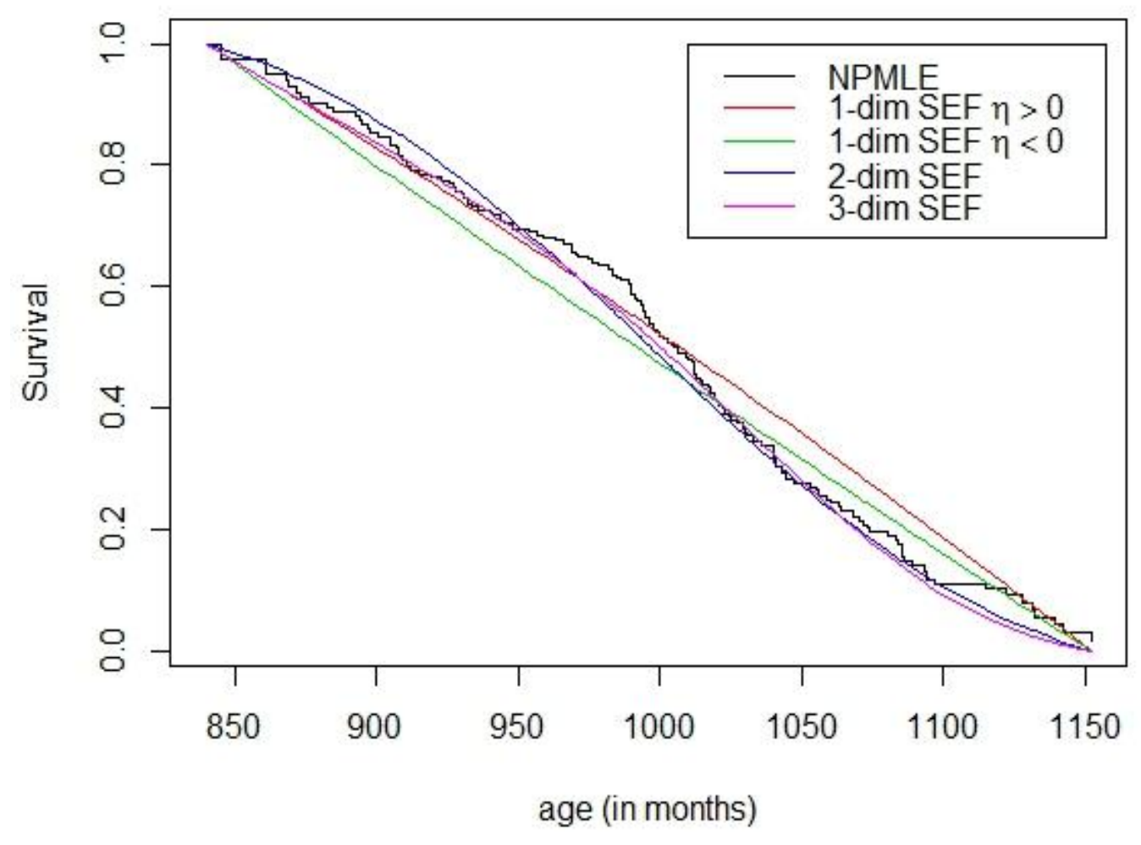
We revisit the Channing House data from Hyde (1980)[3].This dataset contains 462 elderly residents form the Channing House retirement center in Palo Alto, California. The data are collected from 31 January 1964 to 1 July 1975 and hence, the length of the observed period is 11 years and 5 months ($11\times12+5=137$ months). For individual i, let y_i be the age of death and let u_i be the age of the individual at January 31, 1964. Thus, the observed data satisfy $u_i\le y_i\le u_i+137$, for $i=1,2,\cdots,462$. Hence, the lifetime Y is doubly truncation by (U,V) where $V=U+137$. We concentrate on the subset of 167 cases who died during the study period by ignoring individuals who survived at the study end.



We select the suitable model by the criterion AIC. The preferred model is the one with minimum AIC value. $AIC=-2\log L+2k$, where k is the number of unknown parameters in the model and L is the maximized value of the likelihood function. We also choose the goodness of fit test for continuous data, Kolmogorov-Smirnov (K-S) test statistic.

	$\hat\eta_1$	$\hat\eta_2$	$\hat\eta_3$	$\log L$	AIC	K-S statistic
Model 1	0.00090671	0	0	-817.8168	1637.634	0.1022534
Model 2	-0.0003247	0	0	-819.7299	1641.46	0.1000025
Model 3	0.09459514	-4.7424×10^{-5}	0	-817.1781	1638.356	0.07243756
Model 4	-0.8972449	9.436624×10^{-4}	-3.288192×10^{-7}	-814.5028	1635.006	0.06077817

Model 1 is one-parameter special exponential family ($\eta_1>0$), Model 2 is one-parameter special exponential family ($\eta_1<0$), Model 3 is two-parameter special exponential family, Model 4 is cubic special exponential family. We also give the survival function for each model.



7. Conclusion

(1) We propose use Newton-Raphson method to obtain MLE.

(2) For estimating the lifetime data, we propose to use cubic special exponential family. This is similar to the skew-normal distribution mentioned in Robertson and Allison (2011)[4].

Reference

1 Efron B, Tibshirani R. The Annals of Statistics 1996; **24**: 2431-2461.
2 Efron B, Petrosian R. Jornal of the American Statistical Association 1999; **94**: 824-834.
3 Rupert G, John Hyde et. al. Biostatistics casebook 1980. pp. 31-46.
4 Robertson HT, Allison DB. PLOS one 2011; **7**: e37025.
5 Stovring H, Wang MC. BMC Medical Research Methodology 2007; **7**: 53.
6 Zhu H, Wang MC, Biometrika 2012; **99**: 345-361.