Technical Report on

"The Necessary And Sufficient Condition For Consistency Of The MLE"

By Qiqing Yu*

Department of Mathematical Sciences, SUNY, Binghamton, NY 13902, USA

email address: qyu@math.binghamton.edu

Current version: 9/20/2020

Short title: Consistency of the MLE

MSC 2010 subject classification: Primary 62 F 07; Secondary 62 F07.

Key Words: Maximum likelihood estimator, consistency, Kullback-Leibler Inequality, parametric family, sufficient condition.

Abstract: Suppose that the observations are i.i.d. from a density $f(\cdot;\theta)$, where θ is an identifiable parameter. One expects that the maximum likelihood estimator of θ is consistent. But its consistency proof is non-trivial and various sufficient conditions have been proposed (see, *e.g.*, the classical textbooks of Ferguson (1996), Lehmann and Casella (1998), Stuart and Arnold (1999), and Casella and Berger (2001), and more recently Rossi (2018) among others). All these sufficient conditions require $f(x;\theta)$ being somewhat upper semi-continuous (in θ), with various smoothness conditions or conditions needed for the dominated convergence theorem. We show that the sufficient and necessary condition is just that f is somewhat upper semi-continuous, without additional assumptions.

1. Introduction. In this paper, the sufficient and necessary condition is established for the strong consistency of the maximum likelihood estimator (MLE) under the assumptions (A1) $\mathbf{X}_1$, ..., $\mathbf{X}_n$ are i.i.d. from a random vector $\mathbf{X}$ with its density function $f(\cdot;\theta_o)$, $\theta_o \in \Theta$; (A2) the identifiability condition: $\int |f(\mathbf{x};\theta) - f(\mathbf{x};\theta_o)|d\mu(\mathbf{x}) = 0$ implies that $\theta = \theta_o$,

where μ is a measure, θ is a finite dimensional parameter in Θ (the parameter space). While (A1) is a common assumption of the MLE, (A2) is a necessary identifiability condition. If a parameter is not identifiable, we do not even know what is the true value of θ, let alone a consistent estimator of θ. **Hereafter, we assume that (A1) and (A2) hold.**

The new result is weaker than all the existing results in the literature (see, *e.g.*, Ferguson (1996, Part 4), Lehmann and Casella (1998, Section 6.3), Stuart and Arnold (1999, Chapter 18), and Casella and Berger (2001, p. 156)), among others), who only consider sufficient conditions, and is weaker than the results of Zhang (2017) and Rossi (2018) *etc.*, who study also necessary conditions under additional assumptions rather than (A1) and (A2).

For illustration, consider first a continuous random variable (r.v.) X, with its density function f. There are 4 typical cases as follows:

Case 1. $\int f(t)\ln f(t)dt$ is finite, *e.g.*, the uniform distribution $f(t) = \mathbf{1}(t \in (0,1))$.

Case 2. $\int f(t)\ln f(t)dt = -\infty$, *e.g.*, $f(t) = (r-1)\frac{\mathbf{1}(t>e)}{t(\ln t)^r}$, $r > 1$.

Case 3. $\int f(t)\ln f(t)dt = \infty$, *e.g.*, $f(t) = (r-1)\frac{\mathbf{1}(t\in(0,e^{-1}])}{t(|\ln t|)^r}$, $r > 1$.

Case 4. $\int f(t)\ln f(t)dt$ does not exist, *e.g.*, $f(t) = \frac{\mathbf{1}(t\in(0,e^{-1}]\cup[e,\infty))}{2t(\ln t)^2}$.

Each distribution in the examples of the four cases leads to a location or scale parameter family, among other possibilities. Thus we need to study the consistency of the MLE of the parameter in each case.

In many textbooks, people often say that the MLE is consistent under suitable conditions (see *e.g.*, Bickel and Doksum (1997, p. 139.)). On the other hand, in his classical textbook, Ferguson (1996, p.114) shows that the MLE of θ is consistent if the following

conditions hold:

(A3) $\overline{\lim}_{\theta' \to \theta} f(x;\theta') \leq f(x;\theta)\ \forall\ x$ (*i.e.*, $f(\cdot;\theta)$ is upper semi-continuous);

(A4) Θ is compact;

(A5) $\exists$ a function $K(x)$ such that $E_{\theta_o}(|K(X)|) < \infty$ and $\log \frac{f(x;\theta)}{f(x;\theta_o)} \leq K(x)\ \forall\ (x,\theta)$;

(A6) for all $\theta \in \Theta$, and sufficiently small $\delta > 0$, $\sup_{|\theta-\theta_o|<\delta} f(x;\theta)$ is measurable in x.

(A5) is needed in his proof so that the dominated convergence theorem is applicable, but in the examples of Cases 2, 3 and 4, (A3), (A4) and (A5) do not hold (see Remark 1).

Casella and Berger (2001, p.516) present a set of somewhat simpler sufficient conditions for the consistency of the MLE of θ in their popular textbook as follows:

(A7) The densities $f(x;\theta)$ have common support, and $f(x;\theta)$ is differentiable in θ.

(A8) The parameter space Θ contains an open set A and the true parameter $\theta_o \in A$.

Until recently (see *e.g.*, Rossi (2018)) the sufficient conditions for the MLE under the assumptions of (A1) and (A2) are still essentially combinations of (A3), ..., (A8). Notice that (A7) implies (A3), and (A7) and (A8) are much stronger smoothness assumptions on the density. (A8) weaken (A4), but it does not allow discrete Θ and we may not know the true A, just like that we do not know the true value of θ_o. One can verify that (A7) does not hold in the examples of Cases 2, 3 and 4 (see Remark 1).

It is actually more desirable to find the necessary and sufficient (NS) conditions for the consistency of the MLE. The NS conditions are studied for consistency of M-estimates in regression models with general errors (see Berlinet, Liese and Vajda (2000)) or other models, but not for the MLE until recently. Zhang (2017) establishes the NS condition for the weak consistency of the M-estimator under additional assumptions. In particular, Theorem 4 in Zhang (2017) can be stated for the MLE $\hat{\theta}_n$ as follows:

(S1) Suppose that X is a r.v.. $\rho_n(\theta)$ decreases with $\theta < \hat{\theta}_n$ and increases with $\theta > \hat{\theta}_n$, where

$\rho_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln f(X_i;\theta)$, $\theta \in \mathcal{R}$. Then $\hat{\theta}_n \xrightarrow{P} \theta_o$ iff

$$\inf_{||\theta-\theta_o||\geq\epsilon} \rho_n(\theta) - \rho_n(\theta_o) \xrightarrow{P} \delta(\epsilon) \ \forall\ \epsilon > 0 \text{ and some } \delta(\epsilon) > 0. \tag{1.1}$$

Zhang's result has several drawbacks: (1) condition (1.1) is not always easy to check, (2) it is not a strong consistency result; (3) X is not a random vector, (4) $\theta \in \mathcal{R}$, (5) convexity of $\mathcal{L}(\theta)$ in $\mathcal{R}$ may not be true (see Example 7).

Under (A1) and (A2), it is interesting to notice that almost all sufficient conditions in the literature imply that $f(\cdot;\theta)$ is upper semi-continuous in θ (see *e.g.*, (A3) or (A7) (as continuity implies upper semi-continuity)), or its weakened version:

(A9) $\overline{\lim_{n\to\infty}} f(x;\theta_n) \leq f(x;\theta_*)\ \forall\ x \in \mathcal{W}$, where $\int \mathbf{1}(x \notin \mathcal{W}) d\mu(x) = 0$ and $\lim_{n\to\infty} \theta_n = \theta_*$

(see van der Vaart (1998)). Under (A1) and (A2), in additional to upper semi-continuity (A3) or its weakened one (A9), all existing sufficient conditions in the literature need some additional regularity assumptions such as (A5), (A7), (A8) or (S1), among others. In this paper, it is shown that the sufficient condition for strong consistency of the MLE is only (A9) **alone** (see Theorems 1 and 2). The aforementioned parametric families all satisfy (A9) (see Section 4), but some of them do not satisfy (A3) and (A7) (see Remark 1). It is worth mentioning that since (A3) implies (A9), if (A3) holds then the MLE is consistent without (A4), (A5), (A7), (A8) and (S1), as imposed in the literature.

In Section 2, we present the sufficient condition for the strong consistency of the MLE when the observations are random variables (see Theorem 1). The extension of Theorem 1 to the case that the observations are $p \times 1$ random vectors is studied in Section 3 (see Theorem 2). In our proof of consistency, we need to modify the Kullback-Leibler (KL) (1951) inequality. In Section 4, we establish the consistency of the MLEs of the parameters related to the examples in Cases 2, 3 and 4. In Sections 5 and 6, we explain that Theorem 2 can be applied to multivariate regression analysis and survival analysis, respectively. In

Section 7, it is shown that (A9) is the necessary condition in some sense. Some proofs are relegated to Appendix for a better presentation.

2. Main results when observations are random variables. In order to motivate our main results, we shall first study some special cases. The examples in Cases 2, 3 and 4 can lead to parametric families that the existing sufficient conditions for the consistency of the MLE are not applicable. In some cases such as in Example 1, the MLE has an explicit solution. Their consistency proofs are relatively easy in such cases, but not so otherwise.

Example 1. Consider an example in Case 4 with the density function

$$f(x)=\begin{cases}\frac{0.5}{(1+\ln x)^2x} & \text{if } x\geq 1\\ \frac{0.5}{(1-\ln x)^2x} & \text{if } x\in(0,1)\end{cases} \text{ and } F(x)=P(X\leq x)=\begin{cases}1-\frac{0.5}{1+\ln x} & \text{if } x\geq 1\\ \frac{0.5}{1-\ln x} & \text{if } x\in(0,1).\end{cases}$$

It leads to a location parameter family. Suppose that X_1, ..., X_n are i.i.d. from $F(x-\alpha)$. The likelihood function is

$$\begin{aligned}\mathcal{L}(\alpha)=&0.5^n\prod_{i=1}^n[\mathbf{1}(\alpha<X_i)((1-\ln(X_i-\alpha))^2(X_i-\alpha))^{-\mathbf{1}(X_i-\alpha\in(0,1])}]\\ &\times\prod_{i=1}^n[\mathbf{1}(\alpha<X_i)((1+\ln(X_i-\alpha))^2(X_i-\alpha))^{-\mathbf{1}(X_i-\alpha>1)}]\begin{cases}=0 & \text{if } \alpha\geq X_{(1)}\\ =\infty & \text{if } \alpha=X_{(1)}-\\ <\infty & \text{if } \alpha<X_{(1)}.\end{cases}\end{aligned}$$

Thus the MLE under the location parameter family $F(x-\alpha)$ is $\hat\alpha=X_{(1)}$. It is well known that $P(\hat\alpha>t)=P(X_{(1)}>t)=(P(X_1>t))^n$, thus the consistency follows, as well as the distribution of $\hat\alpha$. It is proved in Appendix that $\hat\alpha$ is also strongly consistent.

The cumulative distribution function (cdf) $F(\cdot)$ in Example 1 also leads to a scale parameter family $\{F(\cdot/\theta):\ \theta>0\}$. The density is

$$\begin{aligned}f(x;\theta)=&\theta f(x/\theta)=0.5/\{[(1+\ln(x/\theta))^2x]^{\mathbf{1}(x>\theta)}[(1-\ln(x/\theta))^2x]^{\mathbf{1}(x\in(0,\theta])}\} && x>0\\ =&0.5/\{x(1+\ln(x/\theta))^{2\mathbf{1}(x>\theta>0)}(1-\ln(x/\theta))^{2\mathbf{1}(x\in(0,\theta])}\} && x>0;\\ \mathcal{L}(\theta)=&0.5^n/[\prod_{i=1}^n X_i\prod_{i=1}^n(1-\ln(X_i/\theta))^{2\mathbf{1}(X_i/\theta\in(0,1])}(1+\ln(X_i/\theta))^{2\mathbf{1}(X_i/\theta>1)}]\\ =&0.5^n/[\prod_{i=1}^n X_i\prod_{i=1}^m(1+\ln(X_{(i)}/\theta))^{2\mathbf{1}(X_{(i)}>\theta>0)}]\prod_{i=m+1}^n(1-\ln(X_i/\theta))^{2\mathbf{1}(0<X_{(i)}\leq\theta)}, \quad X_{(1)}>0\end{aligned}$$

$$g(\ln\theta) \stackrel{def}{=} (\ln\mathcal{L}(\theta) + \sum_i \ln X_i + n\ln 2)/2$$

$$= -\sum_{i>k}^{n} \ln(1 + \ln\frac{X_{(i)}}{\theta}) - \sum_{i=1}^{k} \ln(1 - \ln\frac{X_{(i)}}{\theta}) \text{ if } X_{(k)} \le \theta < X_{(k+1)},\ 0 \le k \le n$$

$$= \begin{cases} -\infty & \text{if } \theta = 0+ \\ -\sum_{i=1}^{n} \ln(1 + \ln\frac{X_{(i)}}{\theta}) - \sum_{i\in\emptyset} \ln(1 - \ln\frac{X_{(i)}}{\theta}) & \text{if } 0 < \theta < X_{(1)} \\ -\sum_{i>k}^{n} \ln(1 + \ln\frac{X_{(i)}}{\theta}) - \sum_{i=1}^{k} \ln(1 - \ln\frac{X_{(i)}}{\theta}) & \text{if } X_{(k)} \le \theta < X_{(k+1)},\ 1 \le k < n \\ -\sum_{i\in\emptyset} \ln(1 + \ln\frac{X_{(i)}}{\theta}) - \sum_{i=1}^{n} \ln(1 - \ln\frac{X_{(i)}}{\theta}) & \text{if } X_{(n)} \le \theta \\ -\infty & \text{if } \theta = \infty, \end{cases}$$

where $X_{(0)} = 0$. Let $h(x) = \ln(a \pm x)$, then $h' = \pm(a \pm x)^{-1}$ and $h''(x) = (a \pm x)^{-2}$. Thus

$$g'(t) = \sum_{i:\ X_{(i)} > e^t} (1 + \ln X_{(i)} - t)^{-1} - \sum_{i:\ X_{(i)} \le e^t} (1 - \ln X_{(i)} + t)^{-1}$$

$$= \sum_{i>k}^{n} (1 + \ln X_{(i)} - t)^{-1} - \sum_{i=1}^{k} (1 - \ln X_{(i)} + t)^{-1} \text{ if } t \ne \ln X_{(k)},\ k = 1, ..., n.$$

$$g''(t) = \sum_{i:\ X_{(i)} > e^t} (1 + \ln X_{(i)} - t)^{-2} + \sum_{i:\ X_{(i)} \le e^t} (1 - \ln X_{(i)} + t)^{-2} > 0$$

if $t \ne \ln X_{(k)}$, $k = 1, ..., n$. Thus $g'(\ln\theta)$ is a monotonely **increasing** function of $\ln\theta$ for $\theta \in (X_{(i)}, X_{(i+1)})$, $i \in \{0, 1, ..., n\}$. For example, consider a special case by letting $n = 2$, $X_1 = 1$ and $X_2 = e$. Then

$$g(t) = -\mathbf{1}(0 > t)\ln(1 - t) - \mathbf{1}(0 \le t)\ln(1 + t) - \mathbf{1}(1 > t)\ln(2 - t) - \mathbf{1}(1 \le t)\ln(t).$$

$$g(t) = \begin{cases} -\mathbf{1}(0 > t)\ln(1 - t) - \mathbf{1}(1 > t)\ln(2 - t) = -\infty & \text{if } t = -\infty \\ -\mathbf{1}(0 > t)\ln(1 - t) - \mathbf{1}(1 > t)\ln(2 - t) = -ln2 & \text{if } t = 0- \\ -\mathbf{1}(0 \le t)\ln(1 + t) - \mathbf{1}(1 > t)\ln(2 - t) = -\ln 2 & \text{if } t = 0+ \\ -\mathbf{1}(0 \le t)\ln(1 + t) - \mathbf{1}(1 > t)\ln(2 - t) = -\ln 2.25 & \text{if } t = 0.5 \\ -\mathbf{1}(0 \le t)\ln(1 + t) - \mathbf{1}(1 > t)\ln(2 - t) = -\ln 2 & \text{if } t = 1- \\ -\mathbf{1}(0 \le t)\ln(1 + t) - \mathbf{1}(1 \le t)\ln(t) = -\ln 2 & \text{if } t = 1+ \\ -\mathbf{1}(0 \le t)\ln(1 + t) - \mathbf{1}(1 \le t)\ln(t) = -\infty & \text{if } t = \infty \end{cases}$$

$$g'(t)=\begin{cases}-\frac{\mathbf{1}(0\le t)}{1+t}-\frac{\mathbf{1}(1\le t)}{t}=0- & \text{if } t=\infty\\ -\frac{\mathbf{1}(0\le t)}{1+t}-\frac{\mathbf{1}(1\le t)}{t}<0 & \text{if } t=1+\\ -\frac{\mathbf{1}(0\le t)}{1+t}+\frac{\mathbf{1}(1>t)}{2-t}>0 & \text{if } t=1-\\ -\frac{\mathbf{1}(0\le t)}{1+t}+\frac{\mathbf{1}(1>t)}{2-t}<0 & \text{if } t=0+\\ \frac{\mathbf{1}(0>t)}{1-t}+\frac{\mathbf{1}(1>t)}{2-t}>0 & \text{if } t=0-\\ \frac{\mathbf{1}(0>t)}{1-t}+\frac{\mathbf{1}(1>t)}{2-t}=0+ & \text{if } t=-\infty\end{cases}$$

$t:$	$-\infty$		$0-$	$0+$			$1-$	$1+$		∞
$g'(t):$	$0+$	$\nearrow$	1.5	-0.5	$\nearrow 0 \nearrow$		$+0.5$	-1.5	$\nearrow$	$0-$
$g(t):$	$-\infty$	$\nearrow$	$-\ln 2$	$-\ln 2$	$\searrow$	$\nearrow$	$-\ln 2$	$-\ln 2$	$\searrow$	$-\infty$

Notice that

(1) $\mathcal{L}(\theta)$ is a continuous function in θ;

(2) twice differentiable in the intervals $(X_{(i-1)}, X_{(i)})$, $i = 0, ..., n$;

(3) but not differentiable at $X_{(i)}$'s;

(4) $g''(t) > 0$ for $t \in (e^{X_{(i-1)}}, e^{X_{(i)}})$, $i = 0, ..., n$;

(5) thus $g'(t)$ is strictly increasing for $t \in (e^{X_{(i-1)}}, e^{X_{(i)}})$, $i = 0, ..., n$;

(6) thus $g'(t) = 0$ at the local minimum points.

The MLE $\hat{\theta}_n = \text{argmax}_{t\in\{X_1,...,X_n\}} g(\ln t)$ where

$g(\ln X_{(k)}) = \sum_{i>k}^{n}(1 + \ln(X_{(i)}/X_{(k)}))^{-1} - \sum_{i<k}(1 - \ln(X_{(i)}/X_{(k)}))^{-1}$, $k = 0, ..., n$.

The examples in Cases 2, 3 and 4 can also lead to certain location or scale parameter families. The existing sufficient conditions for the consistency of the MLE are not applicable (see Remark 1).

The example in Case 2. $f(t) = (r-1)\mathbf{1}(t > e)/[t(\ln t)^r]$, $r > 1$.

The location parameter: $f(t;\alpha) = (r-1)\mathbf{1}(t - e > \alpha)/[(t-\alpha)(\ln(t-\alpha))^r]$, $r > 1$. $f(t;\alpha) \uparrow$ in $\alpha \in (-\infty, t-e]$, as $\frac{1}{t-\alpha} \uparrow$ and $\frac{1}{(\ln(t-\alpha))^r} \uparrow$. Thus $\mathcal{L}(\alpha) \uparrow$ in $\alpha \in (-\infty, X_{(1)} - e]$, the MLE of α is $\hat{\alpha}_n = X_{(1)} - e$.

The scale parameter: $f(t;\theta) = (r-1)\mathbf{1}(t > e\theta)/[t(\ln(t/\theta))^r]$, $r > 1$. Thus the MLE of θ is $\hat{\theta}_n = X_{(1)}/e$.

The example in Case 3. $f(t) = (r-1)\mathbf{1}(t \in (0, 1/e])/[t(\ln t)^r]$, $r > 1$.

The location parameter: $f(t;\alpha) = (r-1)\mathbf{1}(t \in [\alpha, \alpha+1/e])/[(t-\alpha)(\ln(t-\alpha))^r]$, $r > 1$. $\mathcal{L}(\alpha) = \frac{\mathbf{1}(\alpha < X_{(1)} \le \cdots \le X_{(n)} < \alpha+1/e])}{\prod_{i=1}^n (r-1)[(X_i-\alpha)(\ln(X_i-\alpha))^r]} \uparrow$ in $\alpha \in [X_{(n)} - 1/e, X_{(1)}]$. Thus the MLE of α is $\hat{\alpha}_n = X_{(1)}$.

The scale parameter: $f(t;\theta) = (r-1)\mathbf{1}(t \in (0, \theta/e))/[t(\ln(t/\theta))^r]$, $r > 1$. Thus the MLE of θ is $\hat{\theta}_n = X_{(n)}e$.

An example in Case 4. Let $u, v > 1$, $p \in [0,1]$, $f(t) = p\frac{(v-1)\mathbf{1}(t\in(0,1])}{(1-\ln t)^v t} + q\frac{(u-1)\mathbf{1}(t>1)}{(1+\ln t)^u t}$, $q = 1-p$. Then $F(t) = \frac{p\mathbf{1}(t\in(0,1))}{(1-\ln t)^v} + q[1 - \frac{\mathbf{1}(t\ge 1)}{(1-\ln t)^u}]$.

The location parameter α: $f(t;\alpha) = p\frac{(v-1)\mathbf{1}(t-a\in(0,1])}{(1-\ln(t-a))^v(t-a)} + q\frac{(u-1)\mathbf{1}(t-\alpha>1)}{(1+\ln(t-a))^u(t-a)}$. Thus $\alpha < X_{(1)}$. Since $\mathcal{L}(\alpha)\begin{cases} = \infty & \text{if } \alpha = X_{(1)}- \\ < \infty & \text{otherwise} \end{cases}$, the MLE is $\hat{\alpha}_n = X_{(1)}$.

The scale parameter θ: $f(t;\theta) = p\frac{(v-1)\mathbf{1}(t/\theta\in(0,1])}{(1-\ln(t/\theta))^v t} + q\frac{(u-1)\mathbf{1}(t/\theta>1)}{(1+\ln(t/\theta))^u t}$. Since lnf=

$$-\ln t + \mathbf{1}(t \in (0,\theta))[\ln(p(v-1)) - v\ln(1-\ln t+\ln\theta)] + \mathbf{1}(t>\theta)[\ln(q(u-1)) - u\ln(1+\ln t-\ln\theta)]$$

$$\begin{aligned} g(\ln\theta) &\stackrel{def}{=} \ln\mathcal{L}(\theta) + \sum \ln X_i - \sum_i \mathbf{1}(X_i \in (0,\theta])\ln(p(v-1)) - \sum_i \mathbf{1}(X_i > \theta)\ln(q(u-1)) \\ &= -\sum_i \mathbf{1}(X_i \in (0,\theta])v\ln(1-\ln t+\ln\theta) - \sum_i \mathbf{1}(X_i>\theta))u\ln(1+\ln t-\ln\theta) \\ &= -u\sum_{i>k}^{n} \ln(1+\ln X_{(i)} - \ln\theta) - v\sum_{i\ge 1}^{k} \ln(1-\ln X_{(i)} + \ln\theta) \text{ where } X_{(k)} \le \theta < X_{(k+1)}, \end{aligned}$$

$k = 0, 1, ..., n$ and $X_{(0)} = 0$. Notice that $\frac{d}{d\theta}g(\ln\theta)$ does not exist at $\theta \in \{X_1, ..., X_n\}$ and $g''(t) = \sum_{i>k} \frac{u}{(1+\ln X_{(i)}-t)^2} + \sum_{i\ge 1}^k \frac{v}{(1-\ln X_{(i)}+t)^2} > 0$ if it exists.

Thus $\hat{\theta}_n = \text{argmax}_{\theta\in\{X_{(1)},...,X_{(n)}\}}\mathcal{L}(\theta)$.

It is worth mentioning that the Newton-Raphson algorithm does not work here.

One can also find the MLE of u. $\ln f \propto \ln(u-1) - u\ln(1+\ln t)$.

$\frac{d}{du}\sum_i \ln f(X_i;u) = \sum_{X_i>1}[\frac{1}{u-1} - \ln(1+\ln X_i)] = 0$ yields

$\sum_{X_i>1}\frac{1}{u-1} = \sum_{X_i>1}\ln(1+\ln X_i)$ => the MLE $\hat{u} = 1 + \frac{\sum_{X_i>1} 1}{\sum_{X_i>1}\ln(1+\ln X_i)}$,

as $\frac{d^2}{du^2}\sum_i \ln f(X_i;u) = -\sum_{X_i>1}\frac{1}{(u-1)^2} < 0$.

One can also find the MLE of p, u and v.

Remark 1. Condition (A7) in Casella and Berger (2001) and conditions (A3), (A4), (A5) in Ferguson (1996) are violated if $f(x;\theta,\alpha) = \frac{\mathbf{1}(x\in(\alpha,\alpha+\theta/e])}{(x-\alpha)(\ln((x-\alpha)/\theta))^2}$ (an example of Case 3).

(A7) fails: $f(x;\theta,\alpha)$ is not continuous in θ at $\theta=\theta_*$ for $x\in\{\alpha,\alpha+\theta_*/e\}$ and for each $\theta_*>0$, due to the factor $\mathbf{1}(x\in(\alpha,\alpha+\theta/e])$ in $f(x;\theta,\alpha)$.

The support of $f(x;\theta,\alpha)$ is $\begin{cases}[0,1/e] & \text{if } \theta=1 \text{ and } \alpha=0\\ [0,1/(2e)] & \text{if } \theta=0.5 \text{ and } \alpha=0\end{cases}$ (not the same).

(A3) fails: $\overline{\lim_{n\to\infty}} f(x;\theta_n,\alpha_n) = \infty$ ($\not\leq 0 = f(x;1,0)$) if $\theta_n = 1$ and $\alpha_n = (-1)^n/n + x$.

(A4) fails: Θ is not compact, as $\Theta = (0,\infty)\times(-\infty,\infty) \neq \overline{\Theta}$ (the closure of Θ).

(A5) fails: Letting $(\theta_o,\alpha_o) = (1,0)$ and $\theta\geq 1$, we have

$$\frac{f(x;\theta,\alpha)}{f(x;\theta_o,\alpha_o)} = \frac{\mathbf{1}(x-\alpha\in(0,\theta/e])x(\ln x)^2}{(x-\alpha)(\ln((x-\alpha)e/\theta))^2\mathbf{1}(x\in(0,1/e])} = \infty, \text{ if } x=\alpha+ > \alpha_o \text{ and} \quad (2.1)$$

$x<1/e$. If (A5) were true, then $\exists$ a function $K(x)$ such that

$E_{\theta_o,\alpha_o}(|K(X)|)<\infty$ and $K(x)\geq \ln\frac{f(x;\theta,\alpha)}{f(x;\theta_o,\alpha_o)}$ $\forall$ (x,θ,α),

then $K(x)\geq \ln\frac{f(x;\theta,\alpha)}{f(x;\theta_o,\alpha_o)} = \log\infty$ $\forall$ $x=\alpha+>\alpha_o=0$ and $x<1/e$ by (2.1),

i.e. $K(x)=\infty$ for all $x\in(0,1/e]$ and $\int_0^{1/e} f_o(x;\theta_o)dx = 1$.

Consequently $\infty > E_{\theta_o,\alpha_o}(|K(X)|) = E(\infty) = \infty$, a contradiction. ▫

$\rho_n(\theta) = -\sum_{i=1}^n \ln f(X_i;\theta)$

$= -\sum_{i=1}^n[\ln(0.5\theta/X_i) + 2\mathbf{1}(X_i>\theta>0)\ln(1+\ln\frac{X_i}{\theta}) + 2\mathbf{1}(0<X_i\leq\theta)\ln(1-\ln\frac{X_i}{\theta})]$

$\propto -n\ln\theta + 2\sum_{i=1}^n \ln\mathbf{1}(X_i>\theta>0)\ln(1+\ln\frac{X_i}{\theta}) + 2\sum_{i=1}^n \mathbf{1}(0<X_i\leq\theta)\ln(1-\ln\frac{X_i}{\theta})$

$$= \begin{cases}\infty & \text{if } \theta = 0+\\ -n\ln\theta + 2\sum_{i=1}^n \ln(1+\ln\frac{X_{(i)}}{\theta}) \ \ (\infty\downarrow) & \text{if } \theta < X_{(1)}\\ -n\ln\theta + 2\sum_{i=1}^m \ln(1+\ln\frac{X_{(i)}}{\theta}) + 2\sum_{i>m}^n \ln(1-\ln\frac{X_{(i)}}{\theta})] & \text{if } X_{(m)}\leq\theta<X_{(m+1)}\\ -n\ln\theta + 2\sum_{i=1}^n \ln(1-\ln\frac{X_{(i)}}{\theta}) \ \ (\downarrow-\infty) & \text{if } X_{(n)}\leq\theta\\ -\infty & \text{if } \theta=\infty.\end{cases}$$

Since the existing results on the sufficient conditions of the consistency of the MLE are not applicable to the families of distributions in Cases 2, 3 and 4, we propose a weaker sufficient condition for consistency of the MLE in Theorem 1. The consistency of the MLEs

in the examples in Cases 2, 3 and 4 can be proved by verifying the sufficient condition proposed in Theorem 1. This is done in Section 4.

Before we present the main theorem, we shall present some preliminary results. We shall make use of the following inequality:

KL inequality. $\int f_o(t)\ln(f_o/f)(t)d\mu(t) \geq 0$*; with equality iff* $\int |f(t) - f_o(t)|d\mu(t) = 0$.

Kullback and Leibler (1951) show that $\int f_o(t)\ln(f_o/f)(t)d\mu(t)$ exists, though it may be ∞.

The KL inequality requires that f and f_o are both densities w.r.t. the measure μ. That is, $\int f_o d\mu = \int f d\mu = 1$. However, we encounter the case $\int f d\mu < 1$ in our proof such as in Example 2.

Example 2. Let $f(x;\theta) = \mathbf{1}(x \in (0,\theta])/\theta$, $\theta > 0$. $F(x;0) = \lim_{\theta \downarrow 0} F(x;\theta) = \mathbf{1}(x \geq 0)\ \forall\ x$. $f(\cdot;0)$ is a point mass at 0 and $\int f(x;0)dx = 0 < 1 = \int f(x;\theta)dx$ if $\theta > 0$.

We thus modify the KL inequality as follows.

Proposition 1. *If* $f \geq 0$, $f_o \geq 0$, μ *is a measure,* $\int f_o(t)d\mu(t) = 1$ *and* $\int f(t)d\mu(t) \leq 1$, *then* $\int f_o(t)\ln\frac{f_o(t)}{f(t)}d\mu(t) \geq 0$, *with equality iff* $f = f_o$ *a.e. w.r.t.* μ.

Proof. In view of the KL inequality, it suffice to prove the inequality $\int f_o(t)\ln\frac{f_o(t)}{f(t)}d\mu(t) \geq 0$ under the additional assumption that $\int f(t)d\mu(t) < 1$, but $\int f(t)d(\mu(t) + \mu_2(t)) = 1$ and $\int f_o(t)\mu_2(t) = 0$. Since f_o and f are densities w.r.t. the measure $\nu = \mu + \mu_2$,

$$
\begin{aligned}
0 &\leq \int f_o(t)\ln\frac{f_o(t)}{f(t)}d\nu(t) && \text{(by the KL inequality)} \\
&= \int f_o(t)\ln\frac{f_o(t)}{f(t)}d\mu(t) + \int f_o(t)\ln\frac{f_o(t)}{f(t)}d\mu_2(t) \\
&= \int f_o(t)\ln\frac{f_o(t)}{f(t)}d\mu(t). && \square
\end{aligned}
$$

We would also make use of Fatou's Lemma with varying measures as follows.

Lemma 1 (Propositions 17 and 18 in Royden (1968, page 231)). *Let* $(\mathcal{S},\mathcal{B})$ *be a measurable space,* $\{\mu_n\}_{n\geq 1}$ *a sequence of measures which converge setwise to a measure* μ *(i.e.,*

$\lim_{n\to\infty} \mu_n(B) = \mu(B)$, $\forall\ B \in \mathcal{B}$), g_n *and* f_n *are non-negative measurable functions, and* $(f_n, g_n)(x)$ *converges pointwise to the vector of functions* $(f, g)(x)$. *Then*

(1) $\int f\, d\mu \le \underline{\lim}_{n\to\infty} \int f_n\, d\mu_n$;

(2) if $f_n \le g_n$ *and* $\lim_{n\to\infty} \int g_n d\mu_n = \int g d\mu$, *then* $\int f d\mu = \lim_{n\to\infty} \int f_n d\mu_n$.

Corollary 1. *Suppose that* μ_n *is a sequence of measures on the measurable space* $(\mathcal{S}, \mathcal{B})$ *such that* $\lim_{n\to\infty} \mu_n(B) \to \mu(B)$, $\forall\ B \in \mathcal{B}$, f *and* f_n *are integrable functions,* $n \ge 1$.

(1) If f_n *are bounded below and* $f(x) = \lim_{n\to\infty} f_n(x)$, *then* $\int f\, d\mu \le \underline{\lim}_{n\to\infty} \int f_n\, d\mu_n$.

(2) If f_n *are bounded below then* $\int \underline{\lim}_{n\to\infty} f_n\, d\mu \le \underline{\lim}_{n\to\infty} \int f_n\, d\mu_n$.

Proof. (1) Let $k = \inf_n \inf_x f_n(x)$. If $k \ge 0$ then the corollary follows from Lemma 1. Otherwise, let $f_n^-(x) = 0 \wedge f_n(x)$, $f_n^+(x) = 0 \vee f_n(x)$, $f^-(x) = 0 \wedge f(x)$ and $f^+(x) = 0 \vee f(x)$. Then $f_n^+ \to f^+$ and $f_n^- \to f^-$ pointwisely, as $f_n \to f$ (assumed in statement (1)). Then

$$\underline{\lim}_{n\to\infty} \int f_n d\mu_n = \underline{\lim}_{n\to\infty} \int (f_n^+ + f_n^-) d\mu_n = \underline{\lim}_{n\to\infty} [\int f_n^+ d\mu_n + \int f_n^- d\mu_n]$$

$$\ge \underline{\lim}_{n\to\infty} \int f_n^+ d\mu_n + \underline{\lim}_{n\to\infty} \int f_n^- d\mu_n$$

$$= \underline{\lim}_{n\to\infty} \int f_n^+ d\mu_n + \int \lim_{n\to\infty} f_n^- d\mu \qquad \text{(by statement (2) of Lemma 1, as } |f_n^-(x)| \le k)$$

$$\ge \int \lim_{n\to\infty} f_n^+ d\mu + \int f^- d\mu \qquad \text{(by statement (1) of Lemma 1, as } f_n^+(x) \text{ is nonnegative)}$$

$$= \int f^+ d\mu + \int f^- d\mu = \int (f^+ + f^-) d\mu = \int f d\mu \ \ i.e., \text{ statement (1) holds.}$$

(2) Let $g_n(x) = \inf\{f_k(x) :\ k \ge n\}$, then $g_n(x) \to g(x) = \underline{\lim}_{n\to\infty} f_n(x)$. We have

$$\int \underline{\lim}_{n\to\infty} f_n\, d\mu = \int \lim_{n\to\infty} g_n\, d\mu \le \underline{\lim}_{n\to\infty} \int g_n\, d\mu_n \text{ (by statement (1), as } g_n \text{ is bounded below)}$$

$$= \underline{\lim}_{n\to\infty} \int \inf\{f_k :\ k \ge n\}\, d\mu_n \le \underline{\lim}_{n\to\infty} \int f_n\, d\mu_n. \ \square$$

Lemma 2. *Given each sequence of cdf's* $\{F(\cdot; \theta_n)\}_{n\ge1}$, $\exists$ *a pointwise convergence subsequence* $\{F(\cdot; \theta_{n_j}\}_{j\ge1}$ *with the limit function* $F(\cdot) \in \mathcal{F}$, *where* $\mathcal{F}$ *is a collection of all* F *satisfying that* $F(x)$ *is a nondecreasing function on* $[-\infty, \infty]$, $F(-\infty) = 0$ *and* $F(\infty) = 1$.

Here $\mathcal{F}$ includes F satisfying $\lim_{x\to\infty} F(x) < F(\infty) = 1$, or $\lim_{x\to-\infty} F(x) > F(-\infty) = 0$, or $\lim_{t\downarrow x} F(t) > F(x)$. Lemma 2 is a trivial special case of Helly's selection theorem (Rudin (1976) p.167), thus its proof is skipped. $F(x;\theta) = \int \mathbf{1}(t \le x) f(t;\theta) d\mu(t)$ if $\theta \in \Theta$. If $\theta_n \in \Theta$ and $\theta_n \to \theta_* \notin \Theta$, then there is a convergent sub-sequence of $\{F(\cdot;\theta_n)\}_{n\ge1}$ by Lemma 2, say $F(\cdot;\theta_{j_n}) \to F(\cdot;\theta_*)$. Then $f(x;\theta_*)$ needs to be defined. For technical reasons in the proof of consistency, define

$$f(x;\theta_*) = \overline{\lim_{n\to\infty}} f(x;\theta_{j_n}). \tag{2.2}$$

Remark 2. *It is worth mentioning that $f(\cdot;\theta_*)$ defined in Eq. (2.2) may not be a density function w.r.t. μ. If $\theta_* \notin \Theta$ then we may not have $F(x;\theta_*) = \int \mathbf{1}(t \le x) f(t;\theta_*) d\mu(t)$. For instance, let $f_n(x) = \mathbf{1}(x \in (0, 1/n))n$ as in Example 2, where $\theta_n = 1/n \to 0 = \theta_*$ and μ is the Lebesgue measure. Then $F_n(x) = \int \mathbf{1}(t \le x) f_n(t) dt \to \mathbf{1}(x \ge 0)$, denoted by $F(x;\theta_*)$. Eq. (2.2) yields $f(x;\theta_*) = \lim_{n\to\infty} f_n(x) = 0$, which is not a density function w.r.t. μ.*

Lemma 3. *If (A9) holds and if $\exists \{\theta_n\}_{n\ge1} \subset \Theta$ such that $\theta_n \to \theta_*$ and $\lim_{n\to\infty} F(x;\theta_n)$ exists $\forall\, x$, then by defining $f(\cdot;\theta_*) \overset{def}{=} \overline{\lim_{n\to\infty}} f(\cdot;\theta_n)$ if $\theta_* \notin \Theta$, we have*

$$\overline{\lim_{n\to\infty}} f(x;\theta_n) \le f(x;\theta_*) \text{ if } x \in \mathcal{W}, \text{ where } \int \mathbf{1}(x \notin \mathcal{W}) d\mu(x) = 0. \tag{2.3}$$

Proof. Assume (A9) holds. If $\theta_* \in \Theta$, (2.3) follows from (A9). Otherwise, (2.3) follows from $f(\cdot;\theta_*) \overset{def}{=} \overline{\lim_{n\to\infty}} f(\cdot;\theta_n)$ if $\theta_* \notin \Theta$. □

Remark 3. *(A9) is a weakened version of (A3) (the upper semi-continuity) in three senses:*

*(1) It does not require $\overline{\lim_{n\to\infty}} f(x;\theta_n) \le f(x;\theta_*)$ **for all** x;*

(2) It allows Θ be a nowhere-dense set;

(3) It yields Eq. (2.3), which does not require $\theta_* \in \Theta$, if Θ is not compact as in (A4).

Remark 4. *(A9) weakens (A7) too: (1) (A9) is implied by the existence of $\frac{\partial f(\cdot;\theta)}{\partial\theta}$ in (A7); (2) (A9) does not require that the densities $f(x;\theta)$ have common support as in (A7).*

Remark 5. *For technical reason, we define the MLE* $\hat{\theta} = \hat{\theta}_n = \text{argmax}_{\theta \in \overline{\Theta}} \prod_{i=1}^n f(X_i; \theta)$, *rather than* $\hat{\theta} = \text{argmax}_{\theta \in \Theta} \prod_{i=1}^n f(X_i; \theta)$. *Hereafter, let* $\hat{F}_n$ *be the empirical distribution function (edf) based on* X_i*'s and* Ω_o *the subset of the sample space* Ω *such that* $\hat{F}_n(x) \to F(x; \theta_o)\ \forall\ x$, *where* $F(\cdot; \theta_o)$ *is the true cdf of* X.

Theorem 1. *Suppose that (A1) and (A2) hold and* X_i*'s in (A1) are random variables. Then the MLE* $\hat{\theta} \overset{a.s.}{\to} \theta_o$ *and* $f(x; \hat{\theta}_n) \overset{a.s.}{\to} f(x; \theta_o)\ \forall\ x \in \mathcal{W}$, *if either (A9) or (A10) holds.*

(A10) $\overline{\lim_{n\to\infty}} f(x; \hat{\theta}_{j_n})(\omega) \le f(x; \theta_*(\omega))\ \forall\ x \in \mathcal{W}$, $\forall\ \omega \in \Omega_o$ *and* $\forall$ *convergent subsequence* $\{\hat{\theta}_{j_n}\}_{n\ge 1}$ *of the MLE* $\hat{\theta}_n$, *where* $\hat{\theta}_{j_n}(\omega) \to \theta_*(\omega)$ *and* $\int \mathbf{1}(x \notin \mathcal{W}) d\mu(x) = 0$.

Proof. Under assumptions (A1) and (A2), $P(\Omega_o) = 1$. For each $\omega \in \Omega_o$, let θ_* be a limiting point of $\hat{\theta}_n(\omega)$, where $\theta_* \in \overline{\Theta}$. Then the MLE $\hat{\theta}_n$ is consistent iff $\theta_* = \theta_o$, Thus the theorem is proved once we prove $\theta_* = \theta_o$, which is done next.

Hereafter, fixed $\omega \in \Omega_o$. Then there exists a convergent subsequence of $\{F(\cdot; \hat{\theta}_n)\}_{n\ge 1}$ by Lemma 2. By taking a convergence subsequences of $\{\hat{\theta}_n\}_{n\ge 1}$ and $\{F(\cdot; \hat{\theta}_n)\}_{n\ge 1}$, without loss of generality (WLOG), we can assume $\hat{\theta}_n \to \theta_*$ and $\hat{F}(\cdot, \hat{\theta}_n)$ converges to $F(\cdot; \theta_*)$ ($\in \mathcal{F}$ (see Lemma 2)) pointwisely. Thus (A9) and Lemma 3 yield Eq. (2.3), with $\theta_n = \hat{\theta}_n$, *i.e.*,

$$\overline{\lim_{n\to\infty}} f(x; \hat{\theta}_n) \le f(x; \theta_*),\ x \in \mathcal{W},\ \text{where} \int \mathbf{1}(x \notin \mathcal{W}) d\mu(x) = 0. \tag{2.4}$$

On the other hand, (2.4) follows from (A10) directly.

The normalized log-likelihood is $\sum_{i=1}^n \ln f(X_i; \theta)/n$. Let $H(t) = t\ln t$. We have

$$0 \ge \frac{1}{n}\sum_{i=1}^n \ln \frac{f(X_i; \theta_o)}{f(X_i; \hat{\theta}_n)} = \int \ln \frac{f(t; \theta_o)}{f(t; \hat{\theta}_n)} d\hat{F}_n(t) = \int H(\frac{f(t; \theta_o)}{f(t; \hat{\theta}_n)}) \frac{f(t; \hat{\theta}_n)}{f(t; \theta_o)} d\hat{F}_n(t). \tag{2.5}$$

Denote $A_k = \{t :\ \frac{f(t;\hat{\theta}_n)}{f(t;\theta_o)} \le k,\ \forall\ n\}$ and $B_k = A_k \setminus A_{k-1}$, $k \ge 1$. Notice that $\frac{f(t;\hat{\theta}_n)}{f(t;\theta_o)}$ is finite for each n, provided that $t \in \{x :\ |F(x+s; \theta_o) - F(x; \theta_o)| > 0\ \forall\ s \ne 0\}$. Then

$$\int \mathbf{1}(\cup_{k\ge 1} B_k) dF(t; \theta_o) = 1. \tag{2.6}$$

For each $k \geq 1$, let $a_k = \mathbf{1}(t \in B_k)\ln(\frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)})$ and define $\ln 0 = -\infty$. We have

$$
\begin{aligned}
&\varliminf_{n\to\infty} \int_{B_k} \ln \frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)} d\hat{F}_n(t) \\
&\geq \int_{B_k} \varliminf_{n\to\infty} \ln \frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)} dF(t;\theta_o) \qquad \text{(by (2) of Corollary 1 as } a_k \in [-\ln k, -\ln(k-1)],\ k \geq 1) \\
&= \int_{B_k} \ln \varliminf_{n\to\infty} \frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)} dF(t;\theta_o) \qquad \text{(as } \ln(x) \text{ is continuous)} \\
&= \int_{B_k} \ln \frac{f(t;\theta_o)}{\varlimsup_{n\to\infty} f(t;\hat{\theta}_n)} dF(t;\theta_o) \\
&\geq \int_{B_k} \ln \frac{f(t;\theta_o)}{f(t;\theta_*)} dF(t;\theta_o) \qquad \text{(by (2.4) and (A9))} \qquad (2.7) \\
&= \int_{B_k} H(\frac{f(t;\theta_o)}{f(t;\theta_*)}) \frac{f(t;\theta_*)}{f(t;\theta_o)} dF(t;\theta_o) \qquad \text{(see Eq. (2.5), as } H(t) = t\ln t) \\
&= \int_{B_k} H(\frac{f(t;\theta_o)}{f(t;\theta_*)}) f(t;\theta_*) d\mu(t) \qquad (as\ dF(t;\theta_o) = f(t;\theta_o)d\mu(t)) \\
&\geq \int_{B_k} (-1/e) f(t;\theta_*) d\mu(t) \qquad \text{(as } t\ln t \geq -1/e\ \forall\ t > 0) \\
&\geq -1/e \qquad \text{(as } \int_{\mathcal{B}_k} f(t;\theta_*) d\mu(t) \in [0,1]). \qquad (2.8)
\end{aligned}
$$

$$
\begin{aligned}
\text{Finally,} \quad 0 &\geq \varliminf_{n\to\infty} \int \ln \frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)} d\hat{F}_n(t) \qquad \text{(by Eq. (2.5))} \\
&= \varliminf_{n\to\infty} \sum_{k\geq 1} \int_{B_k} \frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)} d\hat{F}_n(t) \qquad \text{(by (2.6))} \\
&= \varliminf_{n\to\infty} \int_{k\geq 1} \int_{B_k} \ln \frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)} d\hat{F}_n(t) d\nu(k) \quad (d\nu \text{ is the counting measure}) \\
&\geq \int_{k\geq 1} \varliminf_{n\to\infty} \int_{B_k} \ln \frac{f(t;\theta_o)}{f(t;\hat{\theta}_n)} d\hat{F}_n(t) d\nu(k) \quad \text{(by (2) of Corollary 1 and (2.8))} \\
&\geq \int_{k\geq 1} \int_{B_k} \ln \frac{f(t;\theta_o)}{f(t;\theta_*)} dF(t;\theta_o) d\nu(k) \qquad \text{(by (2.7))} \\
&= \int \ln \frac{f(t;\theta_o)}{f(t;\theta_*)} dF(t;\theta_o) \\
&= \int \ln \frac{f(t;\theta_o)}{f(t;\theta_*)} f(t;\theta_o) d\mu(t) \geq 0 \qquad \text{(by Proposition 1).}
\end{aligned}
$$

That is, $\int \ln \frac{f(t;\theta_o)}{f(t;\theta_*)} f(t;\theta_o) d\mu(t) = 0$. It follows that $\int |f(x;\theta_*) - f(x;\theta_o)| d\mu(x) = 0$ by the second statement of Proposition 1. Consequently $\theta_* = \theta_o$ by (A2). Since $P(\Omega_o) = 1$, the MLE $\hat{\theta} \overset{a.s.}{\to} \theta_o$ and $f(x;\hat{\theta}_n) \overset{a.s.}{\to} f(x;\theta_o)$ $\forall$ $x \in \mathcal{W}$, □

Remark 6. *In view of Remarks 3 and 4, under assumptions (A1) and (A2), Theorem 1 presents a simple sufficient condition, namely (A9) or (A10) alone, which is much weaker than all similar results in the literature. Notice that in the literature, even though some people show that the sufficient conditions include (A9), it is not known that the sufficient condition can be (A9) alone. We shall show in Theorem 3 that (A10) alone is the NS condition for the MLE being strong consistent. But (A9) is easier to verify than (A10).*

3. Extension of Theorem 1 to Random Vectors. In this section, assume that $\mathbf{X}$ is a $p \times 1$ random vector. Notice that (A1) says that $\mathbf{X}_1$, ..., $\mathbf{X}_n$ are i.i.d. observations from $f(\cdot;\theta)$, $\theta \in \Theta$. Here $\mathbf{X}_i$ can be a random variable or a random vector. (A1), (A2) and (A9), as well as Eq. (2.2), Lemmas 1 and 3 do not need to be revised except replacing x by $\mathbf{x}$ $(= (x_1, ..., x_p))$, *etc.*.

For a better presentation, we shall first extend Theorem 1 to the case $p = 2$. Hereafter, write $\mathbf{x} = (x_1, x_2)$, *etc.*, denote $\mathbf{x} \geq \mathbf{y}$ if $x_i \geq y_i$, $i \in \{1, 2\}$; denote $\mathbf{x} > \mathbf{y}$ if $x_1 \geq y_1$ and $x_2 \geq y_2$ with at least one strict inequality. Lemma 2 is a key in the proof of Theorem 1, and it is extended as follows.

Lemma 4. *Given each sequence of bivariate cdf's* $\{F(\cdot;\theta_n)\}_{n\geq 1}$*, there exists a convergence subsequence such that* $\lim_{j\to\infty} F(\mathbf{x};\theta_{n_j}) = F(\mathbf{x};\theta_*)$ $\forall$ $\mathbf{x}$*, where* $F(\mathbf{x};\theta_*) \in \mathcal{F}_2$*, a collection of all* F *such that* $F(\mathbf{x}) \geq F(\mathbf{y})$ *whenever* $\mathbf{x} \geq \mathbf{y}$*,* $F(-\infty, -\infty) = 0$ *and* $F(\infty, \infty) = 1$.

Proof. Given a sequence of cdf's $F(\cdot;\theta_n)$, $\{F(x, \infty;\theta)\}_{n\geq 1}$ is a bounded nondecreasing sequence in x. Helly's selection theorem ensures that there exists a convergent subsequence. WLOG, we can assume $\{F(x, \infty;\theta_n)\}_{n\geq 1}$ converges. Likewise, $F(\infty, x;\theta_n)$ is a bounded nondecreasing sequence in x. Helly's selection theorem ensures that there exists a convergent

subsequence. WLOG, we can assume $F(\infty, x; \theta_n)$ converges too. Moreover, $\{F(x, x; \theta_n)\}_{n\geq 1}$ is also a bounded nondecreasing sequence, thus Helly's selection theorem ensures that there exists a convergent subsequence. WLOG, we can further assume $F(x, x; \theta_n)$ converges again. Denote the limiting functions of $F(x, \infty; \theta_n)$, $F(\infty, x; \theta_n)$ and $F(x, x; \theta_n)$, by F_1, F_2 and F_3, respectively. Let Q_k be the collection of $i/2^k$th quantiles of F_j's for $i \in \{1, 2, ..., 2^k - 1\}$ and $j \in \{1, 2, 3\}$. Notice that Q_k is a finite set with at most $3(2^k - 1)$ elements. For $k = 1$, there is a convergent subsequence of $\{F(\mathbf{x}, \theta_n)\}_{n\geq 1}$ for $\mathbf{x} \in Q_1$, denoted by $\{F(\mathbf{x}, \theta_{1,n})\}_{n\geq 1}$. Inductively, for $k \geq 2$, there is a convergent subsequence of $\{F(\mathbf{x}, \theta_{k-1,n})\}_{n\geq 1}$ for $\mathbf{x} \in \cup_{j=1}^{k} Q_j$, denoted by $\{F(\mathbf{x}, \theta_{k,n})\}_{n\geq 1}$. Then the subsequence $\{F(\mathbf{x}, \theta_{n,n})\}_{n\geq 1}$ converges $\forall$ $\mathbf{x} \in \cup_{k\geq 1} Q_k$. WLOG, we can assume $F(\mathbf{x}, \theta_n)$ converges for $\mathbf{x} \in \cup_{k\geq 1} Q_k$.

We now show that $F(\mathbf{x}; \theta_n)$ converges pointwisely. By the previous construction, it suffices to show that $F(\mathbf{x}; \theta_n)$ converges if $\mathbf{x} \notin \cup_{j\geq 1} Q_j$. In the latter case, given $\epsilon > 0$, $\exists$ $k > 0$ and $\exists$ $\mathbf{z}$, $\mathbf{y} \in \cup_j Q_j$ satisfying $\mathbf{z} < \mathbf{x} < \mathbf{y}$ such that

$$F(\mathbf{y}; \theta_n) - F(\mathbf{z}; \theta_n) \leq \epsilon, \ |F(\mathbf{y}; \theta_m) - F(\mathbf{y}; \theta_n)| \leq \epsilon, \text{ and } |F(\mathbf{z}; \theta_m) - F(\mathbf{z}; \theta_n)| \leq \epsilon \quad (3.1)$$

whenever $n, m \geq k$. It follows that

$$\begin{aligned}
&|F(\mathbf{x}; \theta_n) - F(\mathbf{x}; \theta_m)| \\
\leq &|F(\mathbf{y}; \theta_m) - F(\mathbf{y}; \theta_m) + F(\mathbf{y}; \theta_n) - F(\mathbf{y}; \theta_n) + F(\mathbf{x}; \theta_n) - F(\mathbf{x}; \theta_m) + F(\mathbf{z}; \theta_m) - F(\mathbf{z}; \theta_m)| \\
= &|F(\mathbf{y}; \theta_m) - F(\mathbf{z}; \theta_m) + F(\mathbf{y}; \theta_n) - F(\mathbf{y}; \theta_m) + F(\mathbf{x}; \theta_n) - F(\mathbf{y}; \theta_n) + F(\mathbf{z}; \theta_m) - F(\mathbf{x}; \theta_m)| \\
\leq &|F(\mathbf{y}; \theta_m) - F(\mathbf{z}; \theta_m)| + |F(\mathbf{y}; \theta_n) - F(\mathbf{y}; \theta_m)| \\
&+ |F(\mathbf{x}; \theta_n) - F(\mathbf{y}; \theta_n)| + |F(\mathbf{z}; \theta_m) - F(\mathbf{x}; \theta_m)| \\
\leq &|F(\mathbf{y}; \theta_m) - F(\mathbf{z}; \theta_m)| + |F(\mathbf{y}; \theta_n) - F(\mathbf{y}; \theta_m)| + 2|F(\mathbf{z}; \theta_m) - F(\mathbf{y}; \theta_m)| \qquad (\text{as } \mathbf{z} < \mathbf{x} < \mathbf{y}) \\
\leq &4\epsilon, \text{ whenever } n, m \geq k \text{ by (3.1).}
\end{aligned}$$

Consequently, $F(\mathbf{x}; \theta_k)$ converges pointwisely. This completes the proof of the lemma. □

Theorem 2. *Suppose that (A1) and (A2) hold and* $\mathbf{X}_i$*'s in (A1) are* $2-$*dimensional random vector. Then the MLE* $\hat{\theta}$ *is consistent if (A9) or (A10) holds.*

Proof. Let $\hat{F}_n$ be the empirical distribution function (edf) based on $\mathbf{X}_i$'s and Ω_o the subset of the sample space Ω such that $\hat{F}_n(\cdot) \to F(\cdot;\theta_o)$, where $F(\cdot;\theta_o)$ is the true cdf of $\mathbf{X}$. Under assumptions (A1) and (A2), $P(\Omega_o) = 1$. For each $\omega \in \Omega_o$, let θ_* be a limiting point of $\hat{\theta}_n(\omega)$, where $\theta_* \in \overline{\Theta}$. Then the MLE $\hat{\theta}_n$ is consistent iff $\theta_* = \theta_o$, Thus the theorem is proved once we prove $\theta_* = \theta_o$, which is done next.

Hereafter, fixed $\omega \in \Omega_o$. Then there exists a convergent subsequence of $\{F(\cdot;\hat{\theta}_n)\}_{n\geq 1}$ by Lemma 4. By taking a convergence subsequences of $\{\hat{\theta}_n\}_{n\geq 1}$ and $\{F(\cdot;\hat{\theta}_n)\}_{n\geq 1}$, WLOG, we can assume $\hat{\theta}_n \to \theta_*$ and $\hat{F}(\cdot,\hat{\theta}_n)$ converges to $F(\cdot;\theta_*)$ pointwisely, where $F(\cdot;\theta_*) \in \mathcal{F}_2$. Thus either (A10) yields (2.4), or both (A9) and Lemma 3 yield (2.3) with $\theta_n = \hat{\theta}_n$, that is,

$$\overline{\lim_{n\to\infty}} f(\mathbf{x};\hat{\theta}_n) \leq f(\mathbf{x};\theta_*),\ \mathbf{x} \in \mathcal{W},\ \text{where } \int \mathbf{1}(\mathbf{x} \notin \mathcal{W}) d\mu(\mathbf{x}) = 0 \qquad (see\ (2.4)).$$

The normalized log-likelihood is $\sum_{i=1}^n \ln f(\mathbf{X}_i;\theta)/n$. Let $H(t) = t\ln t$. We have

$$0 \geq \frac{1}{n}\sum_{i=1}^n \ln\frac{f(\mathbf{X}_i;\theta_o)}{f(\mathbf{X}_i;\hat{\theta}_n)} = \int \ln\frac{f(\mathbf{t};\theta_o)}{f(\mathbf{t};\hat{\theta}_n)} d\hat{F}_n(\mathbf{t}) = \int H(\frac{f(\mathbf{t};\theta_o)}{f(\mathbf{t};\hat{\theta}_n)})\frac{f(\mathbf{t};\hat{\theta}_n)}{f(\mathbf{t};\theta_o)} d\hat{F}_n(\mathbf{t}) \quad (see\ (2.5)).$$

The rest of the proof is skipped, as it is identical to the proof of Theorem 1 after Eq. (2.5), provided that x is replaced by $\mathbf{x}$. □

Remark 7. The extension of Theorem 1 to the case $p > 1$ can be done through a mathematical induction on p. The proof of Theorem 1 can be viewed as the step $p = 1$ in the mathematical induction. The proof of Theorem 2 can be viewed as the simple version of step $p + 1$. For simplicity, we ignore the details.

4. Direct Applications. We first consider the case of a random vector specified in the next example.

Example 3. Let $f_{\mathbf{X}}(\mathbf{x};\theta,\alpha,r) = \prod_{i=1}^3 f_i(\frac{x_i-\alpha}{\theta})$, be the density of $\mathbf{X} = (X_1, X_2, X_3)$, where $f_1(x) = \frac{(r-1)\mathbf{1}(x>e)}{(\ln x)^r x}$, $f_2(x) = \frac{(r-1)\mathbf{1}(x\in(0,1/e))}{(\ln x)^r x}$ and $f_3(x) = \frac{\mathbf{1}(x>e)+\mathbf{1}(x\in(0,1/e))}{2(\ln x)^r x}$, $r > 1$,

$\theta > 0$. Notice that f_1, f_2 and f_3 are the examples corresponding to Cases 2, 3 and 4 (mentioned in Section 1), respectively. Then the MLE of the parameter (θ, α, r) based on i.i.d. observations from $f_{\mathbf{X}}$ is consistent. It suffices to show that (A2) and (A9) holds.

$\vdash$: (A2) holds. Or equivalently, $(\theta, \alpha, r) = (\theta_o, \alpha_o, r_o)$ if $F(\mathbf{x}; \theta, \alpha, r) = F(\mathbf{x}; \theta_o, \alpha_o, r_o)\ \forall\ \mathbf{x}$.

By taking derivative, it is easy to check that $\int \frac{r-1}{x(\ln x)^r} dx = -\text{sign}(\ln x)|\ln x|^{1-r} + c$, where c is a constant. Then

$$F(\mathbf{x}; \theta, \alpha, r) = \mathbf{1}(x_1 > \alpha + e\theta)[1 - (\ln\frac{x_1 - \alpha}{\theta})^{1-r}]\mathbf{1}(x_2 \in (\alpha, \alpha + \theta/e))|\ln\frac{x_2 - \alpha}{\theta}|^{1-r} \quad (4.1)$$
$$\times \{\mathbf{1}(x_3 \in (\alpha, \alpha + \theta/e))/|\ln\frac{x_3 - \alpha}{\theta}| + \mathbf{1}(x_3 > \alpha + e\theta)[1 - 1/(\ln\frac{x_3 - \alpha}{\theta})]\}/2.$$

If $(\theta, \alpha, r) \neq (\theta_o, \alpha_o, r_o)$, $\theta, \theta_o > 0$, and r, $r_o > 1$, due to symmetry, it suffices to consider these 4 cases: (1) $\alpha < \alpha_o$, (2) $\alpha + \theta/e < \alpha_o + \theta_o/e$, (3) $\alpha + \theta e < \alpha_o + \theta_o e$, (4) $r < r_o$. We shall show that if $F(\mathbf{x}; \theta, \alpha, r) = F(\mathbf{x}; \theta_o, \alpha_o, r_o)\ \forall\ \mathbf{x}$ then none of the 4 cases is possible.

In Case (1), $\exists\ x_2 \in (\alpha, \alpha_o \wedge (\alpha + \theta/e))$ and $x_1 = x_3 \approx \infty$ such that $F(\mathbf{x}; \theta, \alpha, r) = \mathbf{1}(x_2 \in (\alpha, \alpha + \theta/3))|\ln\frac{x_2 - \alpha}{\theta}|^{1-r} > 0 = F(\mathbf{x}; \theta_o, \alpha_o, r_o)$, by Eq. (4.1). Thus Case (1) is impossible.

In Case (2), $\exists\ x_2 \in ((\alpha + \theta/e) \vee \alpha_o, \alpha_o + \theta_o/e)$ and $x_1 = x_3 \approx \infty$ such that $F(\mathbf{x}; \theta, \alpha, r) = 0 < \mathbf{1}(x_2 \in (\alpha_o, \alpha_o + \theta_o/e))|\ln\frac{x_2 - \alpha_o}{\theta_o}|^{1-r_o} = F(\mathbf{x}; \theta_o, \alpha_o, r_o)$ by Eq. (4.1). Thus Case (2) is impossible.

In Case (3), $\exists\ x_1 \in (\alpha + \theta e, \alpha_o + \theta_o e)$ and $x_2 = x_3 \approx \infty$ such that $F(\mathbf{x}; \theta, \alpha, r) = \mathbf{1}(x_1 > \alpha + \theta e))[1 - (\ln\frac{x_1 - \alpha}{\theta})^{1-r}] > 0 = F(\mathbf{x}; \theta_o, \alpha_o, r_o)$. Thus Case (3) is impossible.

The previous discussion implies that if $F(\mathbf{x}; \theta, \alpha, r) = F(\mathbf{x}; \theta_o, \alpha_o, r_o)\ \forall\ \mathbf{x}$ then $(\theta, \alpha) = (\theta_o, \alpha_o)$ and thus $F(\mathbf{x}; \theta, \alpha, r) = F(\mathbf{x}; \theta, \alpha, r_o)\ \forall\ \mathbf{x}$. The latter equality together with Eq. (4.1) and $x_1 = x_3 = \infty$ further implies

$\mathbf{1}(x_2 \in (\alpha, \alpha + \theta/e))|\ln\frac{x_2 - \alpha}{\theta}|^{1-r} = \mathbf{1}(x_2 \in (\alpha, \alpha + \theta/e))|\ln\frac{x_2 - \alpha}{\theta}|^{1-r_o}$ for all x_2.

$=> 1 = |\ln\frac{x_2 - \alpha}{\theta}|^{r_o - r}$ if $x_2 \in (\alpha, \alpha + \theta/e))$.

Letting $x_2 \downarrow \alpha$, the last equation yields $1 \to \infty$ which is a contradiction. It implies that $r = r_o$. Thus (A2) holds.

Moreover, $f_{\mathbf{X}}(\mathbf{x}; \theta, \alpha, r)$ is continuous in (θ, α, r) for all $\mathbf{x} \notin B$, where $\mathbf{x} \in B$ implies that either $x_1 = \alpha + \theta e$ or $x_2 \in \{\alpha, \alpha + \theta/e\}$ or $x_3 \in \{\alpha, \alpha + \theta/e, \alpha + \theta e\}$. Furthermore, $\int \mathbf{1}(\mathbf{x} \in B) d\mathbf{x} = 0$. Thus (A9) holds. □

5. Applications to the multivariate regression analysis. It seems that Theorem 2 is just for non-regression data, however it can also be applied to the regression data. For instance, the common regression model is the linear regression model, which can be specified by $\mathbf{Y} = \beta' \mathbf{Z} + \mathbf{W}$, where $\mathbf{Y}$ is a k-dimensional response vector, $\mathbf{Z}$ is a p-dimensional covariate vector which may take value zero, and $E(W)$ may not exist. The conditional density function of $\mathbf{Y}$, given $\mathbf{Z} = \mathbf{z}$, is

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) = f_{\mathbf{W}}(\mathbf{y} - \beta' \mathbf{z}; \gamma), \text{ where } \beta \text{ is a } k \times p \text{ dimensional matrix,}$$

β and γ are parameters. The marginal density function of $\mathbf{Z}$ is $f_{\mathbf{Z}}(\mathbf{z})$, which does not depend on (β, γ). Then the joint density function of $\mathbf{X} = (\mathbf{Y}', \mathbf{Z}')'$ becomes

$$f_{\mathbf{X}}(\mathbf{x}; \theta) = f_{\mathbf{W}}(\mathbf{y} - \beta' \mathbf{z}; \gamma) f_{\mathbf{Z}}(\mathbf{z}), \text{ where } \theta = (\beta, \gamma) \text{ and } \mathbf{x}' = (\mathbf{y}', \mathbf{z}'). \tag{5.1}$$

For a random sample from the density $f_{\mathbf{X}}(\cdot; \theta)$, one can apply Theorem 2 to prove the consistency of the MLE, such as the next example.

Example 5. Suppose that the density function of $\mathbf{X}$ is as in Eq. (5.1), where $W = (W_1, W_2)'$, W_1 and W_2 are independent, $f_{W_i}(t; \gamma) = \frac{\gamma_i - 1}{t(\ln t)^{\gamma_i}} \mathbf{1}(t > e)$ $(\gamma_i > 1)$. By Eq. (5.1),

$$f_{\mathbf{X}}(\mathbf{x}; \theta) = f_W(x_1 - \sum_{j=1}^{p} \beta_{1j} z_j, \gamma_1) f_W(x_2 - \sum_{j=1}^{p} \beta_{2j} z_j, \gamma_2) f_{\mathbf{Z}}(\mathbf{z}),$$

where $\theta = (\gamma_1, \gamma_2, \beta)$, and β is a $2 \times p$ matrix. The MLE of θ can be derived by a numerical method, but the consistency can be proved easily by Theorem 2. In order to prove its

consistency, it is suffices to prove that

$$\overline{\lim_{n\to\infty}} f_{\mathbf{X}}(\mathbf{x};\theta_n) \le f_{\mathbf{X}}(\mathbf{x};\theta_*), \text{ whenever } \theta_n,\ \theta_* \in \Theta, \text{ and } \theta_n \to \theta_*.$$

Notice that

$$f_{W_i}(t_i - \sum_{j=1}^{p} \beta_{ij} z_j; \gamma_i) = \frac{(\gamma_i - 1)\mathbf{1}((t_i - \sum_{j=1}^{p} \beta_{ij} z_j) > e)}{(t_i - \sum_{j=1}^{p} \beta_{ij} z_j)(\ln(t_i - \sum_{j=1}^{p} \beta_{ij} z_j))^{\gamma_i}}$$

is continuous in $\theta \in \Theta$ a.e. in $\mathbf{x} \notin B_\theta$, where $\mathcal{B}_\theta = \{(t_1, t_2, z_1, ..., z_p) :\ t_i - \sum_{j=1}^{p} \beta_{ij} z_j = e\}$ and $\int_{B_\theta} 1 d\mathbf{x} = 0$. Thus $f_{\mathbf{X}}(\mathbf{x};\theta)$ is continuous a.e. in $\mathbf{x}$. Consequently, (A9) holds and thus the MLE is consistent.

Example 4. Suppose that the survival time $Y \sim U(0,\theta)$, $\theta \in \Theta = \mathcal{R}$. The censoring time $C \sim Exp(1)$. Y and C are independent. The observable random vector is $\mathbf{X} = (M, \delta)$, where $M = Y \wedge C$ and $\delta = \mathbf{1}(Y \le C$. Let (M_1, δ_1), ..., (M_n, δ_n) be i.i.d. observations. Then $f(m, \delta; \theta) = \frac{\mathbf{1}_{(m\in[0,\theta])}}{\theta}(\theta - m)^{1-\delta}$. It is easy to verify that (A4), (A5) and (A6) do not hold. However, it is easy to show that (A9) holds with $\mathcal{W} = \{0, y\} \times \{0, 1\}\ \forall\ y > 0$. Thus the MLE of θ is strong consistent by Theorem 2.

6. Applications to censored data. It also seems that Theorem 2 is only applicable to complete data. However, Theorem 2 is applicable to various censored data too, such as the right-censored data, current status data, case 2 interval-censored data and the double censored data. In particular, in this section, we consider the controlled experiment with k groups under the right censorship model and the Cox regression model. Then our observations are i.i.d. from $(M_1, \delta_1, ..., M_k, \delta_k, \mathbf{Z})$, where $M_i = Y_i \wedge C$, $i \in \{1, ..., k\}$, Y_1, ..., Y_k are independent random variables, C is a random censoring variable common to the Y_1, ..., Y_k, $\delta_i = \mathbf{1}(Y_i \le C)$, where $\mathbf{1}(A)$ is the indicator function of an event A, and $\mathbf{Z}$ is a random covariate vector. Assume that Y_i is a continuous random variable. Then the Cox model can be specified by

$$S_{Y_i|\mathbf{Z}}(y|\mathbf{z}) = (S_o(t;\gamma_i))^{\exp(\beta'\mathbf{Z})},\ i = 1, ..., k, \tag{6.1}$$

where $S_o(t;\gamma_i) = S_{Y|\mathbf{Z}}(t|\mathbf{0})$ is the baseline survival function and $\gamma = (\gamma_1, ..., \gamma_k)$ is a parameter. One can derive $f_{Y_i|\mathbf{Z}}(t|\mathbf{z})$ from $S_{Y_i|\mathbf{Z}}(t|\mathbf{z})$. That is,

$$f_{Y_i|\mathbf{Z}}(t|\mathbf{z}) = e^{\beta'\mathbf{Z}}(S_o(t;\gamma_i))^{\exp(\beta'\mathbf{Z})-1} f_o(t;\gamma_i) = e^{\beta'\mathbf{Z}}(S_o(t;\gamma_i))^{\exp(\beta'\mathbf{Z})} h_o(t;\gamma_i). \tag{6.2}$$

Let $\tau_C = \sup\{t : \ S_C(t) > 0\}$. Under the right censorship model, Y_i is not observable if $Y_i > \tau_C$. Write $\mathbf{X} = (M_1, \delta_1, ..., M_k, \delta_k, \mathbf{Z}')'$. Under the right censorship model and assuming that the densities $f_{\mathbf{Z}}$ and f_C do not depend on (β, γ), the density of $\mathbf{X}$ becomes

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\theta) =& c\prod_{i=1}^{k}(f_{Y_i|\mathbf{Z}}(t_i|\mathbf{z}))^{\delta_i}(S_{Y_i|\mathbf{Z}}(t_i|\mathbf{z}))^{1-\delta_i} \\ =& c\prod_{i=1}^{k}(e^{\beta'\mathbf{Z}}h_o(t_i;\gamma_i))^{\delta_i}(S_o(t_i;\gamma_i))^{\exp(\beta'\mathbf{Z})} \end{aligned} \tag{6.3}$$

for $\max_i t_i \le \tau_C$, where $\theta = (\gamma, \beta)$, $c = f_{\mathbf{Z}}(\mathbf{z})\prod_{i=1}^{k}(S_C(t_i))^{\delta_i}(f_C(t_i))^{1-\delta_i}$, $t_i = t_j$ if $\delta_i = \delta_j = 0$ and $\mathbf{x} = (t_1, \delta_1, ..., t_k, \delta_k, \mathbf{z}')$. In view of Eqs. (6.2) and (6.3), it is easy to prove the next lemma.

Lemma 5. *Under the assumptions leading to Eq. (6.3), if* $\exists\{\theta_*\} \cup \{\theta_n\}_{n\ge1} \subset \Theta$ *such that* $\lim_{n\to\infty} \theta_n = \theta_*$ *and* $\lim_{n\to\infty} F_{\mathbf{X}}(\cdot;\theta_n)$ *exists, and if*

$$\overline{\lim_{n\to\infty}} h_o(t;\gamma_{in}) \le f_o(t;\gamma_{i*}) \ \forall \ t \in \mathcal{W}, \ \textit{where} \int \mathbf{1}(t \notin \mathcal{W})dt = 0, \ \gamma_n = (\gamma_{1n}, ..., \gamma_{kn})$$

and $\gamma_* = (\gamma_{1*}, ..., \gamma_{k*})$, *then (R9) holds for* $f_{\mathbf{X}}(\mathbf{x};\theta)$.

Example 6. Suppose that a random sample of right censored regression data is from the distribution specified by Eq. (6.3) with $k = 2$ and the baseline density functions are $f_o(t;\gamma_i) = \frac{\gamma_i - 1}{t(\ln t)^{\gamma_i}}\mathbf{1}(t > e)$, where $\gamma_i > 1$, $i = 1, 2$. The density of the random vector $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}) = c\prod_{i=1}^{2}(\exp(\beta/\mathbf{z})\frac{\gamma_i - 1}{t_i \ln t_i}\mathbf{1}(t_i > e))^{\delta_i}(\ln t_i)^{\exp(\beta'\mathbf{Z})(1-\gamma_i)}, \ \mathbf{x} = (t_1, \delta_1, t_2, \delta_2, \mathbf{z}')'.$$

The parameter is $\theta = (\gamma, \beta)$, $\gamma = (\gamma_1, \gamma_2)$, $\gamma_i > 1$ and $\beta \in \mathcal{R}^p$, the p-dimensional Euclidean space. Since the baseline density function is $f_o(t;\gamma_i) = \frac{\gamma_i - 1}{t(\ln t)^{\gamma_i}}\mathbf{1}(t > e)$, we have $S_o(t;\gamma_i) =$

$(\ln t)^{(1-\gamma_i)}\mathbf{1}_{(t>e)}$ and $h_o(t;\gamma_i) = \frac{\gamma_i - 1}{t \ln t}\mathbf{1}(t > e)$. In order to show that the MLE of θ based on the random sample is consistent, by the previous lemma, it suffices to show that $f_o(t;\gamma_i)$ is continuous in γ_i a.e in t, which is indeed true. In fact, $f_o(t;\gamma_i)$ is continuous in γ_i except at $t = e$, provided that $\gamma_i > 1$.

7. Concluding Remark. In this paper we assume (A1) and (A2). We show that (A9) is the only sufficient condition needed for the MLE $\hat{\theta}$ being consistent. In almost all natural parametric distribution families, (A9) is valid and is easy to verified, thus Theorems 1 and 2 confirm our belief that the MLE is consistent in almost all practical cases. One may wonder whether it is the necessary condition. The answer is "No", as shown in Example 7 below.

Example 7. Let $f(x;\theta) = \begin{cases} \mathbf{1}(x \in (0,2])/2 & \text{if } \theta = 0 \\ \frac{k+1}{k}\exp(-x(k+1)/k) & \text{if } \theta = 1/k,\ k \in \mathcal{K} \end{cases}$, $\mathcal{K} = \{1,2,3,...\}$, and $\Theta = \{0\} \cup \{1/k :\ k \in \mathcal{K}\}$. Then $\hat{\theta}_n$ is consistent, (A10) holds but not (A9) and (1.1).

Step (1). $\vdash$: **The MLE $\hat{\theta}_n$ is consistent**. The reason is as follows.

$$\mathcal{L}(\theta) = \begin{cases} 2^{-n}\mathbf{1}(X_{(n)} < 2) & \text{if } \theta = 0, \\ (\frac{k+1}{k})^n \exp(-\sum_{i=1}^n X_i(k+1)/k) & \text{if } \theta = 1/k,\ k \in \mathcal{K}, \end{cases}$$

$$\approx \begin{cases} \begin{cases} 2^{-n} & \text{if } \theta = 0, \\ (\frac{k+1}{k})^n \exp(-n(k+1)/k) & \text{if } \theta = 1/k,\ k \in \mathcal{K}, \end{cases} & \text{if } \theta_o = 0, \\ \begin{cases} 0 & \text{if } \theta = 0, \\ (\frac{m+1}{m})^n \exp(-n) = (\frac{m+1}{me})^n & \text{if } \theta = 1/m, \\ (\frac{k+1}{k})^n \exp(-n\frac{m}{m+1}\frac{k+1}{k}) & \text{if } \theta = 1/k,\ k \in \mathcal{K}, \end{cases} & \text{if } \theta_o = 1/m,\ m \in \mathcal{K}, \end{cases} \quad \text{if } n \text{ is large.}$$

$$\hat{\theta}_n = \begin{cases} 0 & \text{if } 0.5^n \geq \max_k (\frac{k+1}{k})^n \exp(-\sum_{i=1}^n X_i(k+1)/k) \\ 1/m & \text{if } 0.5^n < \max_k (\frac{k+1}{k})^n \exp(-\sum_{i=1}^n X_i \frac{k+1}{k}) = (\frac{m+1}{m})^n \exp(-\sum_{i=1}^n X_i \frac{m+1}{m}). \end{cases}$$

Notice that $\hat{\theta}_n = 0$ if $\theta_o = 0$ and n is large, as

$$\mathcal{L}(0) = 2^{-n} > 2^n e^{-2n} = (2/e^2)^n = \mathcal{L}(1) \geq \mathcal{L}(1/m) = (\frac{m+1}{m\exp(\frac{m+1}{m})})^n,\ m \in \mathcal{K}. \qquad (7.1)$$

Similarly, if $\theta_o = 1/m$, then $\hat{\theta}_n = 1/m$ if n is large, as $\mathcal{L}(0) = 0$ and it is essentially the exponential distribution $Exp(\rho)$, with the MLE $\hat{\rho} \approx \rho_o$ $(= \frac{m+1}{m}$ if $\theta_o = 1/m)$.

Step (2). $\vdash$: **(A9) and (1.1) do not hold**. Eq. (1.1) fails as $\mathcal{L}(0) > \mathcal{L}(1) > \mathcal{L}(1/2)$ by (7.1). On the other hand, (A9) fails by letting $\theta_k = 1/k \to 0 = \theta_* \in \Theta$, then

$$\overline{\lim_{n\to\infty}} f(x;\theta_k) = \lim_{n\to\infty} f(x;\theta_k) = e^{-x} > \mathbf{1}(x \in (0,2])/2 = f(x;\theta_*) \text{ if } x \in (0, -\ln 0.5) \cup (2,\infty).$$

Step (3) $\vdash$: **(A10) holds**. We have just proved in Step (2) that $\hat{\theta}_n = \theta_o$ if $\theta = \theta_o$ and n is large enough. Thus $\hat{\theta}_{j_n}(\omega) \to \theta_*(\omega)$, $\theta_* = \theta_o$, and $\overline{\lim_{n\to\infty}} f(x;\hat{\theta}_{j_n})(\omega) = f(x;\theta_*(\omega))\ \forall\ x \in \mathcal{W}$, $\forall\ \omega \in \Omega_o$, where $\int \mathbf{1}(x \notin \mathcal{W}) d\mu(x) = 0$. So (A10) holds naturally. □

Theorem 3. *Suppose that* $\mathbf{X}$ *is a random vector, (A1) and (A2) hold,* $\theta_o \in \Theta$ *and* $\hat{\theta}_n$ *is the MLE of* θ_o*. Then statement (A10) holds iff statement (A11) holds.*

(A11) $\hat{\theta}_n(\omega) \to \theta_o$ *and* $f(x;\hat{\theta}_n(\omega)) \to f(x;\theta_o)\ \forall\ x \in \mathcal{W}$ *and* $\omega \in \Omega_o$*, where*

$\int \mathbf{1}(x \notin \mathcal{W}) d\mu(x) = 0$ *and* Ω_o *is a subset of the sample space satisfying* $P(\Omega_o) = 1$*.*

Proof. By Theorems 1 and 2, (A10) implies (A11). On the other hand, letting $\theta_* = \theta_o$ in (A11), it is trivially true that (A11) yields (A10). □

Remark 8. Theorem 3 actually presents the NS condition for the MLE $\hat{\theta}_n$ being strongly consistent, though the consistency is referred to both $\hat{\theta}_n$ and $f(\cdot;\hat{\theta}_n)$.

References:

* Berlinet, A., Liese, F. and Vajda, I. (2000). Necessary and sufficient conditions for consistency of M-estimates in regression models with general errors *Journal of Statistical Planning and Inference* 89(1-2):243-267.
* Royden, H.L. (1968). Real analysis. Macmillan. N.Y.
* Bickel, P.J. and Doksum, K.A. (1997) Mathematical Statistics. Holden-Day. Oakland.
* Casella, G. and Berger, R. (2001) Statistical inference, 2nd Ed. Duxbury. N.Y.
* Ferguson, T.S. (1996). A course in large sample theory. Chapman & Hall. N.Y.
* Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.*, 22, 7986.
* Lehmann, E.L. and Casella, G. (1998). Theory of Point estimation. 2nd edition. Springer-Verlag. NY.
* Rossi, R. J. (2018). Mathematical Statistics : An Introduction to Likelihood Based Inference. 227. John Wiley & Sons. NY.

* Rudin, W. (1976). Principles of mathematical analysis. McGraw-Hill. NY.

* Shannon C.E. (1948). The mathematical theory of communication. *Bell System Tech. J.*. 27, 379−423, 623−656.

* Stuart, A. Ord, J.K., and Arnold, S. (1999). Advanced Theory of Statistics, Vol. 2A: Classical Inference and the Linear MOdel, 6th edition. London: Oxford University Press.

* Van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge University Press.

* Zhang, J. (2017). Consistency of MLE, LSE and M-estimation under mild conditions. *Statistical Papers* 61, 189-199.

Appendix For the convenience of readers, we shall give the proofs of some statements in Sections 1 and 2 here. The Appendix can be deleted later on.

Proof of Example 1. $\vdash$: The MLE $\hat{\alpha}$ is strongly consistent.

Let $X_{(1),n} = \min_{i\in\{1,...,n\}} X_i$, where n is the sample size. Since $X_{(1),n}(\omega) \geq X_{(1),n+1}(\omega)\ \forall$ n and $\forall\ \omega$, $\hat{\alpha} = X_{(1),n}$ converges pointwisely as $n \to \infty$. $X_{(1),n+1}(\omega) > \epsilon => X_{(1),n}(\omega) > \epsilon$. Thus $\{\omega \in \Omega :\ X_{(1),n+1}(\omega) > \epsilon\} \subset \{\omega :\ X_{(1),n} > \epsilon\}$. Consequently,

$\{\omega \in \Omega :\ \lim_{n\to\infty} X_{(1),n}(\omega) > \epsilon\} = \cap_{n\geq 1}\{\omega :\ X_{(1),n} > \epsilon\} = \lim_{n\to\infty}\{\omega :\ X_{(1),n} > \epsilon\}$, and

$$P(\{\omega \in \Omega :\ \lim_{n\to\infty} X_{(1),n}(\omega) > \epsilon\}) = \lim_{n\to\infty} P(\{\omega :\ X_{(1),n} > \epsilon\}) = \lim_{n\to\infty} (P(X_1 > \epsilon))^n = 0.$$

Thus $\hat{\alpha}$ is strongly consistent. □

Lemma 6. If $r > 1$, then (1) $\int \frac{1}{x(|\ln x|)^r} dx = \text{sign}(\ln x)(1-r)^{-1}(|\ln x|)^{-r+1} + c$, and

(2) $\int \frac{1}{x(\ln x)^2} \ln \frac{1}{x(|\ln x|)^r} dx = -\ln|\ln x| + \frac{r\ln|\ln x|}{\ln x} + \frac{r}{\ln x} + c$, where c is a constant and $\int f(x)dx$ is the indefinite integral.

Proof of Lemma 6. If $x \in (0, 1)$ then

$\int \frac{1}{x(|\ln x|)^r} dx = \int \frac{1}{x(-\ln x)^r} dx = -\int (-\ln x)^{-r} d(-\ln x) = \frac{-(-\ln x)^{-r+1}}{-r+1} + c.$

If $x > 1$ then $\int \frac{1}{x(|\ln x|)^r} dx = \int (\ln x)^{-r} d\ln x = \frac{1}{-r+1}(\ln x)^{-r+1} + c$. Thus statement (1) holds.

$\int \frac{1}{x(\ln x)^2}\ln\frac{1}{x(\ln x)^r}dx \qquad$ (if $x>1$)

$= -\int \frac{1}{x(|\ln x|)^2}[\ln x + \ln(|\ln x|)^r]dx = -\int \frac{1}{x(\ln x)}dx - \int \frac{r\ln\ln x}{x(\ln x)^2}dx = -\ln\ln x - \int r\ln\ln x d(-\frac{1}{\ln x})$

$= -\ln\ln x - r\ln\ln x(-\frac{1}{\ln x}) + r\int(-\frac{1}{\ln x})d\ln\ln x = -\ln\ln x + \frac{r\ln\ln x}{\ln x} - r\int \frac{1}{(\ln x)^2}d\ln x$

$= -\ln\ln x + \frac{r\ln\ln x}{\ln x} + \frac{r}{\ln x} + c.$

$\int \frac{1}{x(|\ln x|)^2}\ln\frac{1}{x(|\ln x|)^r}dx = \int \frac{1}{x(-\ln x)^2}\ln\frac{1}{x(-\ln x)^r}dx \qquad$ (if $x\in(0,1)$)

$= -\int \frac{1}{x(-\ln x)^2}[\ln x + \ln(-\ln x)^r]dx = -\int \frac{1}{x(\ln x)}dx - \int \frac{r\ln|\ln x|}{x(\ln x)^2}dx$

$= -\ln|\ln x| - \int r\ln|\ln x| d(-\frac{1}{\ln x})$

$= -\ln|\ln x| - r\ln|\ln x|(-\frac{1}{\ln x}) + r\int(-\frac{1}{\ln x})d\ln|\ln x| = (-\ln|\ln x| + r\ln|\ln x|(\frac{1}{\ln x})) - r\int \frac{1}{(\ln x)^2}d\ln x$

$= -\ln|\ln x| + \frac{r\ln|\ln x|}{\ln x} + \frac{r}{\ln x} + c.$ □

Remark 9. $\frac{1}{x(\ln x)^r} = \infty$ if $x = 1$ and $\int \frac{1}{x(\ln x)^r}dx = (1-r)^{-1}(\ln x)^{-r+1} = \infty$ if $x = 1$.

Proof of the Example in Case 3 (mentioned in Section 1). Consider a family of distributions:

$f(x;r) = \frac{r-1}{x(|\ln x|)^r}$, $x\in(0,e^{-1})$ and $r>1$, where r is a parameter.

Notice that $\int_0^{e^{-1}} f(x;r)dx = (|\ln x|)^{-r+1}\Big|_0^{e^{-1}} = 1$ by Lemma 6.

$\vdash$: $\int f(t;2)\ln f(t;r)dt = \infty$, where $r>1$.

$\int_0^{e^{-1}} f(x;2)\ln f(x;r)dx = \int_0^{e^{-1}} \frac{\ln(r-1)-\ln(x(|\ln x|)^r)}{x(\ln x)^2}dx$

$= \ln(r-1) + [(-\ln|\ln x| + \frac{r\ln|\ln x|}{\ln x} + \frac{r}{\ln x})\Big|_0^{e^{-1}}] = \infty\ \forall$ by Lemma 6. □