

Survival analysis for interval censored data  
Part II: current status data

Marloes Maathuis

November 12, 2007

**Contents**

<b>1</b>	<b>Characterization and convex minorants</b>	<b>2</b>
1.1	Basic characterization . . . . .	3
1.2	The characterization in slightly different notation . . . . .	7
1.3	Exercises . . . . .	7
<b>2</b>	<b>Global and local consistency</b>	<b>8</b>
<b>3</b>	<b>Rate of convergence</b>	<b>9</b>
3.1	Global and local rate $n^{1/3}$ . . . . .	9
3.2	Why do we get a $n^{1/3}$ rate? . . . . .	12
3.3	Exercises . . . . .	15
<b>4</b>	<b>Local limiting distribution</b>	<b>22</b>
4.1	Localized characterization . . . . .	23
4.2	Subsequences . . . . .	24

We will now take a detailed look at the MLE for current status data. Current status censoring is the simplest form of interval censoring, but we will see that the MLE in this simple model has interesting and nonstandard limit behavior.

In Section 1, we start by characterizing the MLE in terms of necessary and sufficient conditions. This is a common starting point in the analysis of the MLE for censored data, because of the lack of a closed form for the MLE. In the special case of current status data, the characterization implies that the MLE is given by the slope of the convex minorant of a certain set of points. This convex minorant characterization leads to a fast algorithm for the computation of the MLE, and it also plays a role in the derivation of the limiting theory of the MLE. Moreover, it shows the connection of this problem with monotone regression problems. Note that such a simple convex minorant characterization does typically not exist for the MLE for more complicated forms of censored data, but convex minorants are often involved in some way.

In Sections 2 - 4, we discuss the large sample properties of the MLE for current status data, in three steps:

- Section 2 (Consistency). Does  $\hat{F}_n$  converge to the right thing? Answer: yes,  $\hat{F}_n$  converges to  $F_0$ , globally and locally.
- Section 3 (Rate of convergence). How fast does  $\hat{F}_n$  converge? Answer: at rate  $n^{1/3}$ , globally and locally. We also discuss the reason for this unusual rate of convergence.
- Section 4 (Limiting distribution). What does  $n^{1/3}(\hat{F}_n(t) - F_0(t))$  converge to? Answer: to the slope of the convex minorant of a Brownian motion process plus a parabola.

For consistency and the rate of convergence, we can take two perspectives: globally or locally. For the limiting distribution, we have to look locally in a neighborhood of a fixed point  $t_0$ , because  $\hat{F}_n(t_0) - F_0(t_0)$  is asymptotically independent of  $\hat{F}_n(t_1) - F_0(t_1)$  for  $t_0 \neq t_1$ . This is in contrast to for example the empirical distribution function. (Recall that for  $n$  independent uniform random variables, the re-centered and rescaled empirical distribution  $\sqrt{n}(F_n(t) - t)$  converges to a Brownian bridge process on  $[0, 1]$ ). In this sense, the MLE for current status data behaves more like a density estimation problem: we have slower rates of convergence and asymptotic independence at points  $t_0 \neq t_1$ .

We discuss the global consistency, local consistency and global rate results only briefly. We will look in detail at the characterization, the local rate of convergence, and the local limiting distribution.

## 1 Characterization and convex minorants

Recall the current status model described in Section 1.2 of Part I of the notes:

- Let  $X$  be a failure time with distribution  $F$
- Let  $T$  be an observation time with distribution  $G$
- Assume that the time of interest  $X$  is independent of the observation time  $T$
- We observe  $n$  i.i.d. observations of  $(T_1, \Delta_1), \dots, (T_n, \Delta_n)$  of  $(T, \Delta) = (T, 1\{X \leq T\})$
- Goal: estimate the distribution function  $F(x) = P(X \leq x)$

Let  $T_{(1)}, \dots, T_{(n)}$  be the order statistics of  $T_1, \dots, T_n$ , and let  $\Delta_{(1)}, \dots, \Delta_{(n)}$  be the corresponding  $\Delta$  values, i.e.,  $\Delta_{(i)} = \Delta_j$  if  $T_{(i)} = T_j$ .

Recall that the (relevant part of the) likelihood for  $F$  is given by

$$L_n(F) = \prod_{i=1}^n F(T_i)^{\Delta_i} (1 - F(T_i))^{1-\Delta_i},$$

and the corresponding log likelihood is

$$l_n(F) = \sum_{i=1}^n \Delta_i \log F(T_i) + (1 - \Delta_i) \log(1 - F(T_i)).$$

For convenience, we make the following assumptions throughout:

- (A)  $\widehat{F}_n$  is piecewise constant, with jumps only at the observation times. This is equivalent to assuming that all mass is assigned to the right endpoints of the maximal intersections. With this assumption,  $\widehat{F}_n$  is defined uniquely.
- (B) There are no ties between the observation times  $T_1, \dots, T_n$ , so that  $T_{(1)}, \dots, T_{(n)}$  are well-defined.
- (C)  $\Delta_{(1)} = 1$  and  $\Delta_{(n)} = 0$ . This assumption ensures that  $\widehat{F}_n(T_{(1)}) > 0$  and  $\widehat{F}_n(T_{(n)}) < 1$ , because otherwise the log likelihood equals  $-\infty$ .

None of these assumptions are needed to develop the theory, but they make the proofs a bit simpler.

## 1.1 Basic characterization

Let  $\mathcal{Y} = \{y \in \mathbb{R}^n : 0 < y_1 \leq \dots \leq y_n < 1\}$ , and define  $\widehat{y}_i \equiv \widehat{F}_n(T_{(i)})$ . Then the following proposition describes necessary and sufficient conditions for the MLE:

**Proposition 1.1** (*Groeneboom and Wellner (1992), Proposition 1.1, page 39*). *The vector  $\widehat{y} \in \mathcal{Y}$  is the MLE if and only if:*

$$\sum_{i \geq j} \left\{ \frac{\Delta_{(i)}}{\widehat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \widehat{y}_i} \right\} \leq 0 \quad \text{for all } j = 1, \dots, n \quad (1)$$

$$\sum_{i=1}^n \left\{ \frac{\Delta_{(i)}}{\widehat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \widehat{y}_i} \right\} \widehat{y}_i = 0. \quad (2)$$

**Proof.** This proposition can be proved using for example the Karush-Kuhn-Tucker theorem or the Fenchel theorem. However, we give a direct proof here, taken mostly from Groeneboom and Wellner (1992).

We first write the log likelihood in terms of  $y$ :

$$l_n(y) = \sum_{i=1}^n \Delta_{(i)} \log y_i + (1 - \Delta_{(i)}) \log(1 - y_i).$$

We now need to prove two things: (a) The MLE  $\widehat{y}$  satisfies conditions (1) and (2); and (b) any vector  $y \in \mathcal{Y}$  that satisfies conditions (1) and (2) is the MLE  $\widehat{y}$ .

We first prove (a). Suppose that  $\widehat{y}$  is the MLE, i.e.,  $\widehat{y} \in \mathcal{Y}$  maximizes  $l_n(y)$  over  $\mathcal{Y}$ . Take  $0 < \epsilon < 1 - \widehat{y}_n$ . Let  $\mathbf{1}_i$  be a vector of length  $n$ , for which elements  $i, \dots, n$  are 1, and the others

are 0. For example,  $\mathbf{1}_1 = (1, 1, 1, \dots, 1)$ ,  $\mathbf{1}_2 = (0, 1, 1, \dots, 1)$  and  $\mathbf{1}_3 = (0, 0, 1, \dots, 1)$ . Then for any  $j \in \{1, \dots, n\}$ , we have that  $\hat{y} + \epsilon \mathbf{1}_j \in \mathcal{Y}$ . Since  $\hat{y}$  maximizes  $l_n(y)$  over  $\mathcal{Y}$ , it follows that

$$\lim_{\epsilon \downarrow 0} \frac{l_n(\hat{y} + \epsilon \mathbf{1}_j) - l_n(\hat{y})}{\epsilon} \leq 0. \quad (3)$$

Note that we take a one-sided derivative ( $\epsilon \downarrow 0$ ), since for  $\epsilon < 0$ , we have no guarantee that  $\hat{y} + \epsilon \mathbf{1}_j \in \mathcal{Y}$ . Rewriting the expression in (3) (see blackboard) yields

$$\sum_{i \geq j} \left\{ \frac{\Delta(i)}{\hat{y}_i} - \frac{1 - \Delta(i)}{1 - \hat{y}_i} \right\} \leq 0.$$

This proves that  $\hat{y}$  satisfies condition (1). Next, note that  $\hat{y} + \epsilon \hat{y} \in \mathcal{Y}$  for  $|\epsilon|$  small enough (but  $\epsilon$  may be positive and negative now). It follows that

$$\lim_{\epsilon \rightarrow 0} \frac{l_n(\hat{y} + \epsilon \hat{y}) - l_n(\hat{y})}{\epsilon} = 0. \quad (4)$$

Rewriting this expression gives condition (2). This completes the proof of part (a).

Next, we prove part (b). Let  $\hat{y} \in \mathcal{Y}$  satisfy conditions (1) and (2). We will show that  $\hat{y}$  is the MLE, i.e., that  $l_n(\hat{y}) \geq l_n(x)$  for all  $x \in \mathcal{Y}$ . First, note that  $l_n(y)$  is concave in  $y$ . Hence, for any  $x, y \in \mathcal{Y}$ :

$$l_n(x) - l_n(y) \leq \langle \nabla l_n(y), x - y \rangle,$$

where  $\nabla l_n(y)$  is the vector of partial derivatives w.r.t.  $y_1, \dots, y_n$ :

$$\nabla l_n(y) = \left( \frac{\Delta(1)}{y_1} - \frac{1 - \Delta(1)}{1 - y_1}, \dots, \frac{\Delta(n)}{y_n} - \frac{1 - \Delta(n)}{1 - y_n} \right).$$

Hence,

$$\begin{aligned} l_n(x) - l_n(\hat{y}) &\leq \langle \nabla l_n(\hat{y}), x - \hat{y} \rangle \\ &= \sum_{i=1}^n \left\{ \frac{\Delta(i)}{\hat{y}_i} - \frac{1 - \Delta(i)}{1 - \hat{y}_i} \right\} (x_i - \hat{y}_i) \\ &= \sum_{i=1}^n \left\{ \frac{\Delta(i)}{\hat{y}_i} - \frac{1 - \Delta(i)}{1 - \hat{y}_i} \right\} x_i \quad (\text{since } \hat{y} \text{ satisfies (2)}). \end{aligned} \quad (5)$$

We now define  $\alpha_j = x_j - x_{j-1}$ ,  $j = 1, \dots, n$  (where  $x_0 = 0$ ), so that  $x_i = \sum_{j=1}^i \alpha_j$ . Plugging this into (4), and changing the order of summation, we get:

$$\begin{aligned} l_n(x) - l_n(\hat{y}) &\leq \sum_{i=1}^n \left\{ \frac{\Delta(i)}{\hat{y}_i} - \frac{1 - \Delta(i)}{1 - \hat{y}_i} \right\} \sum_{j=1}^i \alpha_j \\ &= \sum_{j=1}^n \alpha_j \sum_{i \geq j} \left\{ \frac{\Delta(i)}{\hat{y}_i} - \frac{1 - \Delta(i)}{1 - \hat{y}_i} \right\}. \end{aligned}$$

This expression is  $\leq 0$ , because of condition (1) and the fact that the all  $\alpha_j \geq 0$  for any  $x \in \mathcal{Y}$ . Hence, we showed that  $l_n(x) - l_n(\hat{y}) \leq 0$  for all  $x \in \mathcal{Y}$ .  $\square$

Rewrite Proposition 1.1 in a simpler form, leads to the following corollary:

**Corollary 1.2** *The vector  $\hat{y} \in \mathcal{Y}$  is the MLE if and only if*

$$\sum_{i < j} \{\Delta_{(i)} - \hat{y}_i\} \geq 0 \quad (6)$$

for all  $j = 1, \dots, n+1$ , and equality holds if  $\hat{y}_j > \hat{y}_{j-1}$  (with  $\hat{y}_0 = 0$  and  $\hat{y}_{n+1} = 1$ ).

**Proof.** We will show that any  $\hat{y} \in \mathcal{Y}$  that satisfies the conditions of Proposition 1.1, also satisfies the conditions of Corollary 1.2. Let  $\hat{y} \in \mathcal{Y}$  and let  $\hat{\alpha}_j = \hat{y}_j - \hat{y}_{j-1}$ ,  $j = 1, \dots, n$  (with  $\hat{y}_0 = 0$ ). We rewrite condition (1) of Proposition 1.1 as follows:

$$\begin{aligned} 0 &= \sum_{i=1}^n \left\{ \frac{\Delta_{(i)}}{\hat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \hat{y}_i} \right\} \hat{y}_i \\ &= \sum_{i=1}^n \left\{ \frac{\Delta_{(i)}}{\hat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \hat{y}_i} \right\} \sum_{j=1}^i \hat{\alpha}_j \\ &= \sum_{j=1}^n \hat{\alpha}_j \sum_{i \geq j} \left\{ \frac{\Delta_{(i)}}{\hat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \hat{y}_i} \right\} \quad (\text{by changing the order of summation}). \end{aligned}$$

Since  $\hat{\alpha}_j \geq 0$  for all  $j$ , it is now clear that conditions (1) and (2) imply:

$$\sum_{i \geq j} \left\{ \frac{\Delta_{(i)}}{\hat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \hat{y}_i} \right\} \leq 0 \quad (7)$$

for all  $j = 1, \dots, n$ , and equality holds if  $\hat{y}_j > \hat{y}_{j-1}$  (with  $\hat{y}_0 = 0$ ). Next, let  $\sigma$  and  $\tau$  be the indices of two successive jump points of  $\hat{y}$ , in the sense that:

$$\hat{y}_{\sigma-1} < \hat{y}_\sigma = \hat{y}_{\sigma+1} = \dots = \hat{y}_{\tau-1} < \hat{y}_\tau.$$

Let  $s \in \{\sigma + 1, \dots, \tau\}$ . From (7), we know that:

$$\begin{aligned} \sum_{i \geq \sigma} \left\{ \frac{\Delta_{(i)}}{\hat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \hat{y}_i} \right\} &= 0, \quad \text{and} \\ \sum_{i \geq s} \left\{ \frac{\Delta_{(i)}}{\hat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \hat{y}_i} \right\} &\leq 0 \quad (\text{and equality holds if } s = \tau). \end{aligned}$$

Subtracting the second expression from the first gives:

$$\sum_{\sigma \leq i < s} \left\{ \frac{\Delta_{(i)}}{\hat{y}_i} - \frac{1 - \Delta_{(i)}}{1 - \hat{y}_i} \right\} \geq 0 \quad (\text{and equality holds if } s = \tau), \quad (8)$$

Since  $\hat{y}_i$  is constant for  $i \in \{\sigma, \dots, \tau - 1\}$ , we can multiply the expression by  $\hat{y}_i(1 - \hat{y}_i)$ , yielding:

$$0 \leq \sum_{\sigma \leq i < s} \{\Delta_{(i)}(1 - \hat{y}_i) - \hat{y}_i(1 - \Delta_{(i)})\} = \sum_{\sigma \leq i < s} \{\Delta_{(i)} - \hat{y}_i\}, \quad (9)$$

where equality holds if  $s = \tau$ . Note that  $\hat{y}_1$  is always a jump point (because of the assumption  $\Delta_{(1)} = 1$ ). Let  $\sigma_2$  be the index of the second jump point. Then (9) implies that for all  $s \in \{2, \dots, \sigma_2\}$ :

$$\sum_{i < s} \{\Delta_{(i)} - \hat{y}_i\} \geq 0,$$

and equality holds if  $s = \sigma_2$ . Let  $\sigma_3$  be the index of the next jump point of  $\hat{y}$ . Then for all  $s \in \{\sigma_2 + 1, \dots, \sigma_3\}$  we have

$$\begin{aligned} \sum_{i < s} \{\Delta_{(i)} - \hat{y}_i\} &= \sum_{i < \sigma_2} \{\Delta_{(i)} - \hat{y}_i\} + \sum_{\sigma_2 \leq i < s} \{\Delta_{(i)} - \hat{y}_i\} \\ &= 0 + \sum_{\sigma_2 \leq i < s} \{\Delta_{(i)} - \hat{y}_i\} \geq 0, \end{aligned}$$

and equality holds if  $s = \sigma_3$  (again using (9)). The proof is completed by continuing like this, and realizing that (9) also holds for  $\tau = n + 1$ .

Please check for yourself that the reversed implication also holds: Any  $\hat{y} \in \mathcal{Y}$  satisfying the condition of Corollary 1.2 also satisfies the conditions of Proposition 1.1.  $\square$

**Proposition 1.3** *Let  $\mathcal{P} = \{P_i = (i, \sum_{j \leq i} \Delta_{(j)}), i = 0, \dots, n\}$ . Let  $H$  be the greatest convex minorant of  $\mathcal{P}$ . Then  $\hat{y}$  is the MLE if and only if for all  $i = 1, \dots, n$ ,  $\hat{y}_i$  equals the left derivative of  $H$  at  $i$ .*

**Proof.** We will show that any  $\hat{y} \in \mathcal{Y}$  satisfying Proposition 1.1, also satisfies the convex minorant characterization. Let  $\sigma$  and  $\tau$  be the indices of two successive jump points of  $\hat{y}$ , and let  $s \in \{\sigma + 1, \dots, \tau\}$ . Then (9), together with the fact that  $\hat{y}_i = \hat{y}_\sigma$  for the entries in the sum, this implies:

$$\hat{y}_\sigma(s - \sigma) \leq \sum_{\sigma \leq i < s} \Delta_{(i)}$$

so that

$$\hat{y}_\sigma \leq \frac{\sum_{\sigma \leq i < s} \Delta_{(i)}}{s - \sigma},$$

where equality holds if  $s = \tau$ . Note that the right hand side of this expression is exactly the slope of the line connecting points  $P_{\sigma-1}$  and  $P_{s-1}$  in the cumulative sum diagram  $\mathcal{P}$ . Hence  $\hat{y}_\sigma$  must be  $\leq$  the slopes connecting points  $P_{\sigma-1}$  and  $P_{s-1}$  for  $s \in \{\sigma + 1, \dots, \tau\}$ , and it must be equal to the slope connecting the points  $P_{\sigma-1}$  and  $P_{\tau-1}$ . In other words,  $\hat{y}_i$  is the left derivative of the line segment connecting  $P_{\sigma-1}$  and  $P_{\tau-1}$ , for  $i = \{\sigma, \dots, \tau - 1\}$ . Since this holds for any two successive jump points, it follows that  $\hat{y}_i$  is the left derivative of  $H$  at  $i$ .

The other direction of can be proved similarly, and is left as an exercise.  $\square$

This characterization leads to a very fast and easy algorithm for the computation of the MLE for current status data, and we will later see that this characterization is also useful for deriving the limiting distribution of the MLE. Finally, note that the same convex minorant characterization also arise in monotone regression estimate (see, e.g., Barlow, Bartholomew, Bremner and Brunk (1972)). This shows that the MLE  $\hat{y}$  can also be viewed as a monotone regression estimate: it is the minimizer of the least squares criterion

$$\sum_{i=1}^n \{\Delta_{(i)} - y_i\}^2$$

over all  $y \in \mathcal{Y}$ . In other words,  $\hat{y}$  is the projection of  $\Delta_{(1)}, \dots, \Delta_{(n)}$  on  $\mathcal{Y}$ .

Another way to compute the MLE for current status data is via the Pool Adjacent Violators algorithm, see, e.g., Ayer, Brunk, Ewing, Reid and Silverman (1955) and Barlow, Bartholomew, Bremner and Brunk (1972, pages 13-15).

## 1.2 The characterization in slightly different notation

In order to derive asymptotic properties of the MLE, it is convenient to write the characterization in slightly different and more compact notation. Let  $G_n$  to denote the empirical distribution of  $T_1, \dots, T_n$ . Thus,  $G_n$  puts mass  $1/n$  at each value of  $T_1, \dots, T_n$ , or in other words:

$$G_n(t) = \frac{1}{n} \sum_{i=1}^n 1\{T_i \leq t\}.$$

Similarly, we let  $\mathbb{P}_n$  to denote the empirical distribution of  $(T_1, \Delta_1), \dots, (T_n, \Delta_n)$ . With this notation, we have

$$\int f(t, \delta) d\mathbb{P}_n(t, \delta) = \frac{1}{n} \sum_{i=1}^n f(T_i, \Delta_i).$$

We can then write the log likelihood (divided by  $n$ ) as

$$l_n(F) = \int \{\delta \log F(t) + (1 - \delta) \log(1 - F(t))\} d\mathbb{P}_n(t, \delta).$$

Using this notation, we get rewrite the characterization in Corollary 1.2 as follows:

**Proposition 1.4**  $\widehat{F}_n$  is an MLE if and only if

$$\int_{t < s} \{\delta - \widehat{F}_n(t)\} d\mathbb{P}_n(t, \delta) \geq 0, \quad \text{for all } s \tag{10}$$

and equality holds if  $s$  is a jump point of  $\widehat{F}_n$  and if  $s > T_{(n)}$ .

## 1.3 Exercises

**Exercise 1.** Why can we make assumption (C) on page 3 without loss of generality? And what do we do when this assumption does not hold?

**Exercise 2.** Show that rewriting expression (4) gives (2).

**Exercise 3.** Complete the proof of Corollary 1.2.

**Exercise 4.** Complete the proof of Proposition 1.3.

## 2 Global and local consistency

In many censored data problems, Hellinger consistency can be proved via empirical process theory. We will not discuss these techniques in this class, but you may look at the following references: Van der Vaart and Wellner (1996), Van der Vaart and Wellner (2000) and Van de Geer (2000). Schick and Yu (2000) also have a paper about consistency of the MLE for mixed case interval censored data, where they follow a more ‘bare-hands’ approach.

When we apply this theory to current status data, we obtain  $L_1(G)$  consistency (see, e.g., Wellner (2005), Example 1.1, page 85):

$$\int \left| \widehat{F}_n(t) - F_0(t) \right| dG(t) \rightarrow_{a.s.} 0. \quad (11)$$

Note that this is a ‘global’ consistency result, in the sense that we integrate the difference between  $\widehat{F}_n$  and  $F_0$  with respect to  $G$  over the entire real line. A very nice property of (11) is that we do not need any extra assumptions to derive it. Also, note the role of  $G$ ; the expression indicates that we cannot expect to get consistency on regions where  $G$  has no mass.

From the  $L_1(G)$  consistency we can also derive local consistency, under some additional regularity conditions:

**Proposition 2.1** *Let  $F_0$  be continuous at  $t_0$ . Moreover, let  $G$  be continuously differentiable at  $t_0$  with strictly positive derivative  $g(t_0)$ . Then we can choose  $r > 0$  such that*

$$\sup_{t \in [t_0 - r, t_0 + r]} \left| \widehat{F}_n(t) - F_0(t) \right| \rightarrow_{a.s.} 0.$$

**Proof.** Choose the constant  $r > 0$  such that  $g(t) > g(t_0)/2$  for all  $t \in [t_0 - 2r, t_0 + 2r]$ . Fix an  $\omega$  for which the  $L_1(G)$  consistency holds, and suppose there is an  $x_0 \in [t_0 - r, t_0 + r]$  for which  $\widehat{F}_n(x_0, \omega)$  does not converge to  $F_0(x_0)$ . Then there is an  $\epsilon > 0$  such that for all  $n_1 > 0$  there is an  $n > n_1$  such that  $|\widehat{F}_n(x_0, \omega) - F_0(x_0)| > \epsilon$ . Using the monotonicity of  $\widehat{F}_n$  and the continuity of  $F_0$ , this implies there is a  $\gamma > 0$  such that  $|\widehat{F}_n(t, \omega) - F_0(t)| > \epsilon/2$  for all  $t \in (x_0 - \gamma, x_0]$  or  $[x_0, x_0 + \gamma)$  and  $[x_0 - \gamma, x_0 + \gamma] \subset [t_0 - 2r, t_0 + 2r]$ . This yields that  $\int |\widehat{F}_n(t, \omega) - F_0(t)| dG(t) > \gamma \epsilon g(t_0)/4$ , which contradicts  $L_1(G)$  consistency. Uniform consistency follows since  $F_0$  is continuous.  $\square$

See Schick and Yu (2000) for more consistency results like this one, under a variety of conditions.

Finally, note that the MLE is very often consistent. But this is not *always* the case; in some models the MLE converges to something completely different than what you are trying to estimate. We will discuss one such example in class, namely the MLE for current status data with continuous marks.

### 3 Rate of convergence

#### 3.1 Global and local rate $n^{1/3}$

We can also use empirical process theory to prove a global rate of convergence. A useful theorem for this purpose is Van der Vaart and Wellner (1996, Theorem 3.4.1, page 322). For current status data, the global rate of convergence is  $n^{1/3}$ :

$$n^{1/3} \int \left| \widehat{F}_n(t) - F_0(t) \right| dG(t) = O_p(1).$$

Since we are ultimately interested in the local limiting distribution of the MLE, it is essential to also get a local rate of convergence. It will turn out that the local rate of convergence is also  $n^{1/3}$ , see Theorem 3.1 below. The local rate of convergence does typically not follow easily from the global rate of convergence, and there are no general techniques yet for local rate proofs in these type of problems. Therefore, proving the local rate of convergence is often an obstacle in proving the local limiting behavior of the MLE. The common theme in existing proofs is that they all rely heavily on the characterization of the MLE, and this is also the case in the proof of Theorem 3.1 below.

**Theorem 3.1** (Groeneboom and Wellner (1992, Lemma 5.4, page 95)) *Assume that  $0 < F_0(t_0) < 1$  and that  $F_0$  is continuously differentiable at  $t_0$  with positive derivative  $f_0(t_0)$ . Furthermore, assume that  $G$  is continuously differentiable at  $t_0$  with positive derivative  $g(t_0)$ . Let  $m > 0$ . Then*

$$n^{1/3} \sup_{t \in [-m, m]} \left| \widehat{F}_n(t_0 + n^{-1/3}t) - F_0(t_0) \right| = O_p(1).$$

As an introduction to the proof we first recall the definition of  $O_p(1)$ :

**Definition 3.2** *A sequence of random variables  $X_1, X_2, \dots$ , is said to be of order  $O_p(1)$  (or tight), if for every  $\epsilon > 0$  we can find  $c$  and  $N$  such that*

$$P(|X_n| > c) < \epsilon \quad \text{for all } n > N.$$

Thus, in order to prove Theorem 3.1, it is sufficient to show that for every  $\epsilon > 0$  we can find  $c$  and  $N$  so that

$$P \left( n^{1/3} \sup_{t \in [-m, m]} \left| \widehat{F}_n(t_0 + n^{-1/3}t) - F_0(t_0) \right| > 2f_0(t_0)c \right) < \epsilon \quad \text{for all } n > N.$$

Moreover, note that

$$\begin{aligned} & P \left\{ n^{1/3} \sup_{t \in [-m, m]} \left| \widehat{F}_n(t_0 + n^{-1/3}t) - F_0(t_0) \right| > 2f_0(t_0)c \right\} \\ &= P \left\{ \exists t \in [-m, m] : \left| \widehat{F}_n(t_0 + n^{-1/3}t) - F_0(t_0) \right| > 2f_0(t_0)cn^{-1/3} \right\} \\ &= P \left\{ \exists t \in [-m, m] : \widehat{F}_n(t_0 + n^{-1/3}t) \notin \left( F_0(t_0) - 2f_0(t_0)cn^{-1/3}, F_0(t_0) + 2f_0(t_0)cn^{-1/3} \right) \right\} \\ &\leq P \left\{ \exists t \in [-m, m] : \widehat{F}_n(t_0 + n^{-1/3}t) \notin \left( F_0(t_0 - cn^{-1/3}), F_0(t_0 + cn^{-1/3}) \right) \right\} \quad (12) \\ &\leq P \left\{ \widehat{F}_n(t_0 - mn^{-1/3}) < F_0(t_0 - cn^{-1/3}) \quad \text{or} \quad \widehat{F}_n(t_0 + mn^{-1/3}) > F_0(t_0 + cn^{-1/3}) \right\} \quad (13) \\ &\leq P \left\{ \widehat{F}_n(t_0 - mn^{-1/3}) < F_0(t_0 - cn^{-1/3}) \right\} + P \left\{ \widehat{F}_n(t_0 + mn^{-1/3}) > F_0(t_0 + cn^{-1/3}) \right\}. \end{aligned}$$

The only steps that may require explanation are (12) and (13). Line (12) follows from the continuous differentiability of  $F_0$  at  $t_0$ , so that

$$F_0(t_0) - 2f_0(t_0)cn^{-1/3} \leq F_0(t_0 - cn^{-1/3}) \leq F_0(t_0 + cn^{-1/3}) \leq F_0(t_0) + 2f_0(t_0)cn^{-1/3}$$

for all  $n$  sufficiently large. Line (13) follows from the monotonicity of  $\widehat{F}_n$ .

Hence, it is sufficient to show that we can find  $c$  and  $N$  so that for all  $n > N$ :

$$P \left\{ \widehat{F}_n(t_0 - mn^{-1/3}) < F_0(t_0 - cn^{-1/3}) \right\} < \epsilon/2 \quad \text{and} \quad (14)$$

$$P \left\{ \widehat{F}_n(t_0 + mn^{-1/3}) > F_0(t_0 + cn^{-1/3}) \right\} < \epsilon/2. \quad (15)$$

We will only show (15), since the proof of (14) is analogous.

**Proof of Theorem 3.1.** Let  $\epsilon > 0$  and  $m > 0$ . By the discussion above, it follows that it is sufficient to show that we can find  $c$  and  $N$  so that  $P(A_n) < \epsilon/2$  for all  $n > N$ , where

$$A_n = \left\{ \widehat{F}_n(t_0 + mn^{-1/3}) > F_0(t_{nc}) \right\} \quad \text{and} \quad t_{nc} = t_0 + cn^{-1/3}. \quad (16)$$

Let  $r > 0$ . Recall the uniform strong consistency of  $\widehat{F}_n$  on  $[t_0 - r, t_0]$  (Proposition 2.1). When we combine this with the assumptions  $F_0(t_0) > 0$  and  $f_0(t_0) > 0$ , it follows that there is an  $N$  so that with probability one  $\widehat{F}_n$  has at least one jump in the interval  $(t_0 - r, t_0]$  for all  $n > N$ , and hence also in the interval

$$I_{nm} = (t_0 - r, t_0 + mn^{-1/3}].$$

Let  $\tau_n$  be the largest jump point of  $\widehat{F}_n$  in  $I_{nm}$ .

Recall that the characterization in (10) must always hold for  $\widehat{F}_n$ . Hence, since  $\tau_n$  is a jump point of  $\widehat{F}_n$ ,

$$0 \leq \int_{[\tau_n, t_{nc})} \left\{ \delta - \widehat{F}_n(t) \right\} d\mathbb{P}_n(t, \delta).$$

Moreover, on the event  $A_n$ , we have  $\widehat{F}_n(t) \geq F_0(t_{nc})$  for  $t \geq \tau_n$ , since  $\tau_n$  is the last jump point before  $t_0 + mn^{-1/3}$ . Hence, on the event  $A_n$ , we have

$$0 \leq \int_{[\tau_n, t_{nc})} \left\{ \delta - \widehat{F}_n(t) \right\} d\mathbb{P}_n(t, \delta) \leq \int_{[\tau_n, t_{nc})} \left\{ \delta - F_0(t_{nc}) \right\} d\mathbb{P}_n(u, \delta).$$

This implies that for  $n > N$ ,

$$\begin{aligned} P(A_n) &= P(A_n \cap \{ \text{the characterization (10) holds} \}) \\ &\leq P \left( \int_{[\tau_n, t_{nc})} \left\{ \delta - F_0(t_{nc}) \right\} d\mathbb{P}_n(u, \delta) \geq 0 \right) \\ &\leq P \left( \exists t \in I_{nm} : \int_{[t, t_{nc})} \left\{ \delta - F_0(t_{nc}) \right\} d\mathbb{P}_n(u, \delta) \geq 0 \right), \end{aligned} \quad (17)$$

where the last inequality follows from the fact that with probability one  $\tau_n \in I_{nm}$  for  $n > N$ . Note that the probability in (17) does not involve the MLE  $\widehat{F}_n$  anymore. It is therefore much easier to analyze, and we can use Lemma 3.3 below to show that we can find  $N$  and  $c$  such that this probability is bounded by  $\epsilon/2$  for all  $n > N$ .  $\square$

**Lemma 3.3** *Let  $F_0$  be continuously differentiable at  $t_0$  with positive derivative  $f_0(t_0)$ , and suppose that  $G$  is continuously differentiable at  $t_0$  with positive derivative  $g(t_0)$ . Then for any  $\epsilon > 0$  and  $m > 0$  there exist  $c > m$ ,  $N > 0$ , and  $r > 0$  such that for all  $n > N$ ,*

$$P\left(\exists t \in I_{nm} : \int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} d\mathbb{P}_n \geq 0\right) < \frac{\epsilon}{2},$$

where  $I_{nm} = (t_0 - r, t_0 + mn^{-1/3}]$  and  $t_{nc} = t_0 + cn^{-1/3}$ .

Before we give the proof of this lemma, we try to give some intuition for it. We can write

$$\begin{aligned} & \int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} d\mathbb{P}_n(u, \delta) \\ &= \int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} d\{\mathbb{P}_n(u, \delta) - P(u, \delta)\} + \int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} dP(u, \delta) \\ &= I + II \end{aligned} \tag{18}$$

Note that  $I$  is a random term, and that  $II$  is deterministic. We will now look at the size of each of these terms.

We consider the first term  $I$ , for fixed  $t$ . This term is approximately normal with mean zero and variance

$$\begin{aligned} & n^{-1} \int_t^{t_{nc}} \{\delta - F_0(u)\}^2 dP(u, \delta) \\ &= n^{-1} \int_t^{t_{nc}} F_0(u)(1 - F_0(u)) dG(u) \\ &\approx n^{-1} g(t_0) F_0(t_0)(1 - F_0(t_0))(t_{nc} - t). \end{aligned}$$

So this term is of order  $O_p(n^{-1/2}\sqrt{t_{nc} - t})$ .

Next, we consider the deterministic term  $II$ . Note that the integrand of this term is always negative, since  $F_0$  is strictly increasing and  $u \leq t_{nc}$ . So this term will give a negative contribution:

$$\begin{aligned} & \int_t^{t_{nc}} \{F_0(u) - F_0(t_{nc})\} dP(u, \delta) \\ &= \int_t^{t_{nc}} \{F_0(u) - F_0(t_{nc})\} dG(u) \\ &= f_0(t_0)g(t_0) \int_t^{t_{nc}} (u - t_{nc}) du (1 + o(1)) \\ &= -\frac{1}{2} f_0(t_0)g(t_0)(t - t_{nc})^2 (1 + o(1)) \end{aligned}$$

for  $t \rightarrow t_0$  and  $t_{nc} \rightarrow t_0$ .

Now the  $n^{-1/3}$  rate arises because the random and deterministic terms

$$I = O_p\left(n^{-1/2}\sqrt{t_{nc} - t}\right) \quad \text{and} \quad II = -\frac{1}{2} f_0(t_0)g(t_0)(t - t_{nc})^2.$$

balance each other when  $t_{nc} - t = O_p(n^{-1/3})$ . If one of these terms were of different order, then we would get a different rate. For example, if the first term is of order  $O_p(n^{-1/2}(t_{nc} - t))$  and the deterministic term  $II$  is quadratic of order  $-c(t_{nc} - t)^2$ , then the terms balance out for  $t_{nc} - t = O_p(n^{-1/2})$ , giving the usual  $\sqrt{n}$  rate. Such a trade-off happens in many ‘regular’

problems, see the example of  $\bar{X}_n$  that we will discuss in class. Kim and Pollard (1990) provide a nice discussion of these issues on pages 193-194 of their paper.

Note that the argument above assumed a fixed value of  $t$  in order to determine the size of term  $I$ . To make the argument more precise we need a bound that is uniform in  $t$ . For this purpose we introduce the following lemma:

**Lemma 3.4** (Lemma 4.1 of Kim and Pollard (1990)). *Let  $r > 0$ ,  $s \in (t_0 - r, t_0 + r)$ . Then for every  $\gamma > 0$  there exist random variables  $M_n$  of order  $O_p(1)$  such that*

$$\int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} d\{\mathbb{P}_n(u, \delta) - P(u, \delta)\} \leq \gamma(t - s)^2 + n^{-2/3} M_n^2 \quad \text{for all } t \in (t_0 - r, s). \quad (19)$$

**Proof.** See Kim and Pollard (1990, Lemma 4.1), using the functions  $g_n(u, \delta) = \{\delta - F_0(t_{nc})\} 1_{[t, t_{nc})}(u)$ . Note that these functions depend on  $n$ , but that does not matter in the proof.  $\square$

**Proof of Lemma 3.3.** Let  $\epsilon > 0$  and  $m > 0$ . We use the decomposition of equation (18). The first term  $I$  can be bounded by Lemma 3.3: for every  $\gamma > 0$  there exist random variables  $M_n$  of order  $O_p(1)$  such that

$$\left| \int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} d\{\mathbb{P}_n(u, \delta) - P(u, \delta)\} \right| \leq \gamma(t_{nc} - t)^2 + n^{-2/3} M_n^2 \quad \text{for all } t \in (t_0 - r, t_{nc}).$$

Moreover, as we saw in the heuristic argument before Lemma 3.3, the deterministic term  $II$  can be bounded above by  $-\frac{1}{4}f_0(t_0)g(t_0)(t_{nc} - t)^2$  for  $r > 0$  sufficiently small and  $n$  sufficiently large.

By combining these bounds, we obtain that for all  $t \in I_{nm} = (t_0 - r, t_0 + mn^{-1/3}]$ ,

$$\begin{aligned} \int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} d\mathbb{P}_n(u, \delta) &\leq \gamma(t_{nc} - t)^2 + n^{-2/3} M_n^2 - \frac{1}{4}f_0(t_0)g(t_0)(t_{nc} - t)^2 \\ &= -\frac{1}{8}f_0(t_0)g(t_0)(t_{nc} - t)^2 + n^{-2/3} M_n^2 \\ &\leq \left( -\frac{1}{8}f_0(t_0)g(t_0)(c - m)^2 + M_n^2 \right) n^{-2/3}. \end{aligned} \quad (20)$$

The second line follows by choosing  $\gamma = f_0(t_0)g(t_0)/8$ , and the last line follows since  $t_{nc} - t \geq (c - m)n^{-1/3}$  for  $t \in I_{nm}$ . Since  $M_n$  is of order  $O_p(1)$ , we can choose  $N$  and  $c$  such that

$$\begin{aligned} P \left( \exists t \in I_{nm} : \int_{[t, t_{nc})} \{\delta - F_0(t_{nc})\} d\mathbb{P}_n(u, \delta) \geq 0 \right) \\ \leq P \left( M_n^2 \geq \frac{1}{8}f_0(t_0)g(t_0)(c - m)^2 \right) \leq \frac{\epsilon}{2} \quad \text{for all } n > N. \end{aligned}$$

$\square$

### 3.2 Why do we get a $n^{1/3}$ rate?

In order to understand why we get a  $n^{1/3}$  rate instead of the more common  $\sqrt{n}$ -rate, we look at two simpler problems that are given in Kim and Pollard (1990, page 193-194).

**Example 1. rate  $\sqrt{n}$** 

Suppose that we have an i.i.d. sample  $X_1, \dots, X_n$  from a population with mean  $\mu_0$ . We want to estimate  $\mu_0$ , using the value  $\hat{\mu}_n$  that minimizes  $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \mathbb{P}_n(x - \mu)^2$  over  $\mu$ . (Note that  $\hat{\mu}_n$  is simply the sample mean  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . However, we will not use that fact. Instead, we will work with the implicit definition of  $\hat{\mu}_n$  as minimizer of  $\mathbb{P}_n(x - \mu)^2$ , while in our censored data problems the MLE is also defined in such an implicit way.)

Since  $\hat{\mu}_n$  minimizes  $\mathbb{P}_n(x - \mu)^2$ , we have

$$\mathbb{P}_n(x - \hat{\mu}_n)^2 - \mathbb{P}_n(x - \mu_0)^2 \leq 0.$$

Next, note that for all  $\mu \in \mathbb{R}$

$$\begin{aligned} & \mathbb{P}_n(x - \mu)^2 - \mathbb{P}_n(x - \mu_0)^2 \\ &= [(\mathbb{P}_n - P)\{(x - \mu)^2 - (x - \mu_0)^2\}] + [P\{(x - \mu)^2 - (x - \mu_0)^2\}] \\ &= I + II. \end{aligned}$$

Note that  $I$  is a random term, and  $II$  is a deterministic term. We want to determine the size of these terms. By writing  $(x - \mu)^2 = ((x - \mu_0) + (\mu_0 - \mu))^2$ , it follows that

$$(x - \mu)^2 - (x - \mu_0)^2 = 2(x - \mu_0)(\mu_0 - \mu) + (\mu_0 - \mu)^2. \quad (21)$$

Hence,  $I = 2(\mu - \mu_0)(\mathbb{P}_n - P)(x - \mu_0)$ , and this is approximately normal with mean zero and variance  $4(\mu - \mu_0)^2 \text{Var}(X)/n$ . So term  $I$  is of order  $O_p(n^{-1/2}(\mu - \mu_0))$  (since for any sequence of variables  $Y_n$  with  $EY_n = 0$  and variance  $\text{Var}(Y_n) = O_p(\sigma_n^2)$ , we have that  $Y_n = O_p(\sigma_n)$ ; see the exercises). Equation (21) and the fact that  $P(x - \mu_0) = 0$  imply that  $II = (\mu - \mu_0)^2$ .

Figures 1, 2 and 3 show these processes for random samples of size  $n = 100$ ,  $n = 1000$  and  $n = 10000$ . Note that  $I$  is a line with a random slope of order  $O_p(n^{-1/2})$ , and that  $II$  is a parabola. The minimizer  $\hat{\mu}_n$  of  $\mathbb{P}_n(x - \mu)^2$  can only occur in regions where the terms  $I$  and  $II$  are of the same order of magnitude. So

$$O_p\left(n^{-1/2}(\hat{\mu}_n - \mu_0)\right) = (\hat{\mu}_n - \mu_0)^2,$$

which implies that  $\hat{\mu}_n - \mu_0 = O_p(n^{-1/2})$ . This gives the familiar  $\sqrt{n}$  rate of convergence for the sample mean.

**Example 2. rate  $n^{1/3}$** 

Suppose that we have an i.i.d. sample  $X_1, \dots, X_n$  from some distribution  $F$  with density  $f$ . We want to estimate the midpoint  $\mu_0$  of the interval of length 2 for which  $P1\{\mu - 1 \leq x \leq \mu + 1\}$  is maximal. As our estimator, we take the value  $\tilde{\mu}_n$  that maximizes the proportion of observations in the interval  $[\mu - 1, \mu + 1]$  over  $\mu$ . In other words,  $\tilde{\mu}_n$  is the argmax of  $\frac{1}{n} \sum_{i=1}^n 1\{\mu - 1 \leq X_i \leq \mu + 1\} = \mathbb{P}_n 1\{\mu - 1 \leq x \leq \mu + 1\}$ .

Note that

$$\mathbb{P}_n 1\{\tilde{\mu}_n - 1 \leq x \leq \tilde{\mu}_n + 1\} - \mathbb{P}_n 1\{\mu_0 - 1 \leq x \leq \mu_0 + 1\} \geq 0,$$

since  $\tilde{\mu}_n$  maximizes  $\mathbb{P}_n 1\{\mu - 1 \leq x \leq \mu + 1\}$ . Next, note that

$$\begin{aligned} & \mathbb{P}_n 1\{\mu - 1 \leq x \leq \mu + 1\} - \mathbb{P}_n 1\{\mu_0 - 1 \leq x \leq \mu_0 + 1\} \\ &= [(\mathbb{P}_n - P)(1\{\mu - 1 \leq x \leq \mu + 1\} - 1\{\mu_0 - 1 \leq x \leq \mu_0 + 1\})] \\ &\quad + [P(1\{\mu - 1 \leq x \leq \mu + 1\} - 1\{\mu_0 - 1 \leq x \leq \mu_0 + 1\})] \\ &= I + II. \end{aligned}$$

Suppose that  $X$  has a smooth density  $f(x)$ . Then, for fixed  $\mu$ ,  $I$  is approximately normal with mean zero and variance (assuming without loss of generality that  $\mu_0 < \mu$  and  $|\mu - \mu_0| < 1$ ):

$$\begin{aligned}
& \frac{1}{n} \text{Var}(1\{\mu - 1 \leq X \leq \mu + 1\} - 1\{\mu_0 - 1 \leq X \leq \mu_0 + 1\}) \\
&= \frac{1}{n} \text{Var}(1\{\mu_0 - 1 \leq X \leq \mu - 1\} - 1\{\mu_0 + 1 \leq X \leq \mu + 1\}) \\
&\approx \frac{1}{n} P[1\{\mu_0 - 1 \leq X \leq \mu - 1\} + 1\{\mu_0 + 1 \leq X \leq \mu + 1\}] \\
&= \frac{1}{n} \int_{\mu_0-1}^{\mu-1} f(x) dx + \frac{1}{n} \int_{\mu_0+1}^{\mu+1} f(x) dx \\
&\approx \frac{1}{n} (\mu - \mu_0) \{f(\mu_0 - 1) + f(\mu_0 + 1)\}.
\end{aligned}$$

So we conclude that

$$I = O_p\left(n^{-1/2} \sqrt{|\mu - \mu_0|}\right).$$

Note that this is different from the order we got for the sample mean, which was  $O_p(n^{-1/2}(\mu - \mu_0))$ .

Next, we analyze term  $II$ , again for a fixed value of  $\mu$ :

$$II = \int_{\mu_0+1}^{\mu+1} f(x) dx - \int_{\mu_0-1}^{\mu-1} f(x) dx.$$

This looks similar to what we got for the variance of term I, but we now take the difference of these two expectations instead of their sum. Because of that, and because of the fact that  $f(\mu_0 - 1) \approx f(\mu_0 + 1)$  (since  $\mu_0$  maximizes  $P\{1\{\mu - 1 \leq x \leq \mu + 1\}\}$ ), the first order terms cancel. We have

$$\begin{aligned}
\int_{\mu_0-1}^{\mu-1} f(x) dx &\approx \int_{\mu_0-1}^{\mu-1} \{f(\mu_0 - 1) + (x - (\mu_0 - 1))f'(\mu_0 - 1)\} dx \\
&= f(\mu_0 - 1)(\mu - \mu_0) + \frac{1}{2} f'(\mu_0 - 1)(\mu - \mu_0)^2,
\end{aligned}$$

and similarly

$$\int_{\mu_0+1}^{\mu+1} f(x) dx \approx f(\mu_0 + 1)(\mu - \mu_0) + \frac{1}{2} f'(\mu_0 + 1)(\mu - \mu_0)^2.$$

Hence, using  $f(\mu_0 - 1) \approx f(\mu_0 + 1)$ , we have that

$$II \approx \frac{1}{2} (f'(\mu_0 + 1) - f'(\mu_0 - 1)) (\mu - \mu_0)^2.$$

Note that this is a parabola with a negative coefficient, since  $f'(\mu_0 + 1) < 0$  and  $f'(\mu_0 - 1) > 0$ . I simulated data from a normal distribution with mean 1 and standard deviation 1. In that case,  $\mu_0 = 1$  and  $f(x) = (2\pi)^{-1/2} \exp(-(x - \mu_0)^2/2)$ , so that

$$\begin{aligned}
I &\approx n^{-1/2} \sqrt{2(2\pi)^{-1/2} \exp(-1/2) |\mu - 1|}, \\
II &\approx -(2\pi)^{-1/2} \exp(-1/2) (\mu - 1)^2
\end{aligned}$$

Figures 4, 5 and 6 show these processes in pictures for  $n = 100$ ,  $n = 1000$  and  $n = 10000$ . Note that  $I$  is a now a random process, and that  $II$  is still a parabola. The maximizer  $\tilde{\mu}_n$  of

$\mathbb{P}_n 1\{\mu - 1 \leq x \leq \mu + 1\}$  can only occur in regions where the terms  $I$  and  $II$  are of the same order of magnitude. So

$$O_p\left(n^{-1/2}\sqrt{|\tilde{\mu}_n - \mu_0|}\right) = (\tilde{\mu}_n - \mu_0)^2,$$

which implies that  $\tilde{\mu}_n - \mu_0 = O_p(n^{-1/3})$ . This gives a  $n^{1/3}$  rate of convergence.

This argument shows roughly how the  $n^{1/3}$  rate arises. To make it more precise, one needs to establish the bounds of terms  $I$  and  $II$  not only for fixed values of  $\mu$ , but uniformly in  $\mu$ , see Kim and Pollard (1990).

One can study the random and deterministic component in more complicated estimation problems as well, and the order of magnitudes of these two terms determine the rate of convergence. In the case of current status data, we saw that we get a deterministic and a random term that are of the same order as in Example 2. This explains where the  $n^{1/3}$  rate of convergence comes from.

### 3.3 Exercises

**Exercise 1.** Let  $Y_n$ ,  $n = 1, 2, \dots$  be a sequence of variables with  $EY_n = 0$  and variance  $\text{Var}(Y_n) = O_p(\sigma_n^2)$ . Show that  $Y_n$  is of order  $O_p(\sigma_n)$ .

**Exercise 2.** Work out the other details in Examples 1 and 2 of Section 3.2.

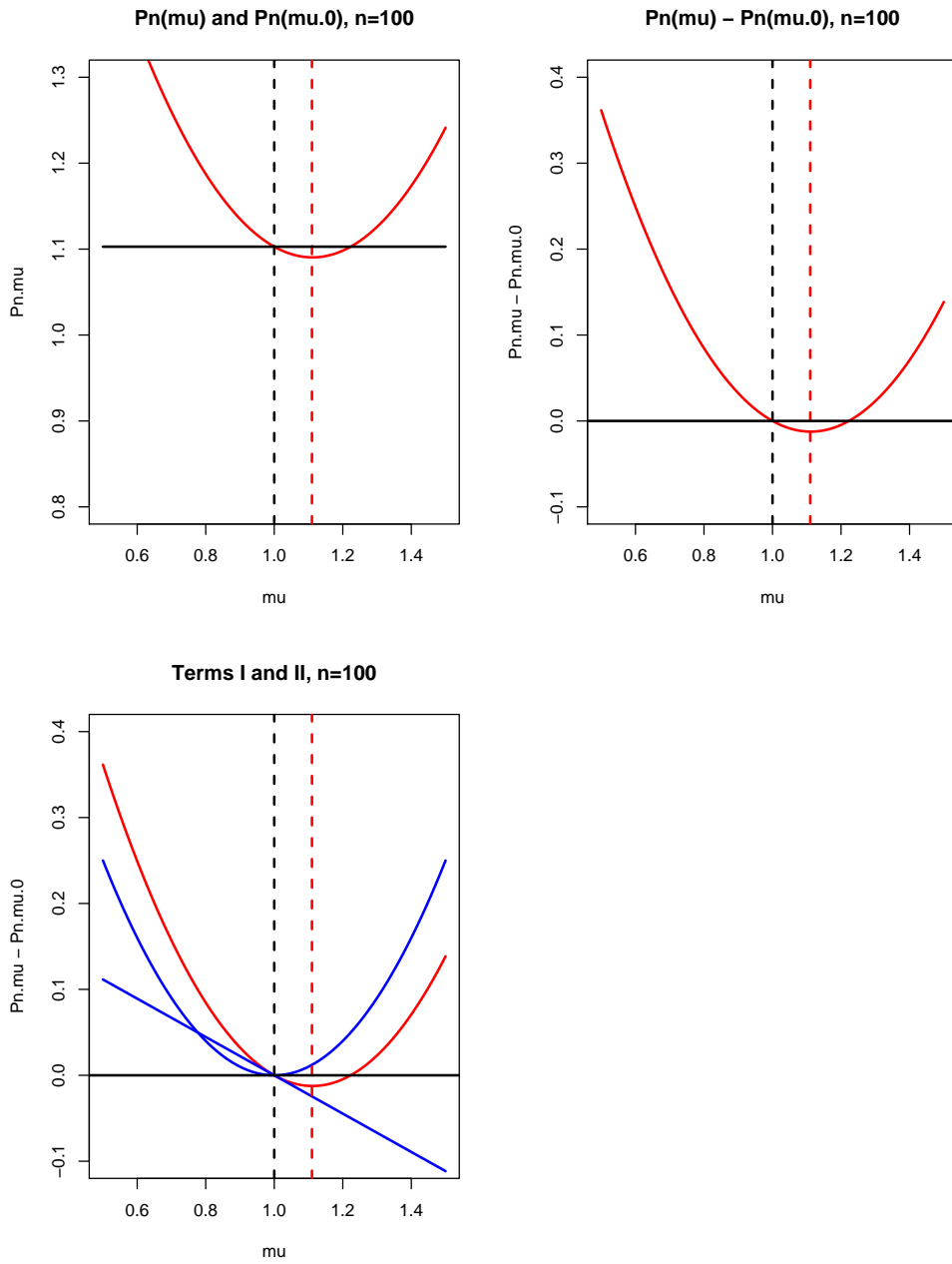


Figure 1: The top left panel shows  $\mathbb{P}_n(x - \mu)^2$  (red) and  $\mathbb{P}_n(x - \mu_0)^2$  (black). The values  $\hat{\mu}_n$  (red) and  $\mu_0$  (black) are depicted by vertical dashed lines. The top right panel shows  $\mathbb{P}_n(x - \mu_0)^2 - \mathbb{P}_n(x - \mu)^2$ . Note that this process is nonpositive at  $\hat{\mu}_n$ . The lower left panel separates this process in the random term *I* (blue line) and the deterministic term *II* (blue parabola). The figures are based on a random sample of size 100 from a Normal(1,1) distribution.

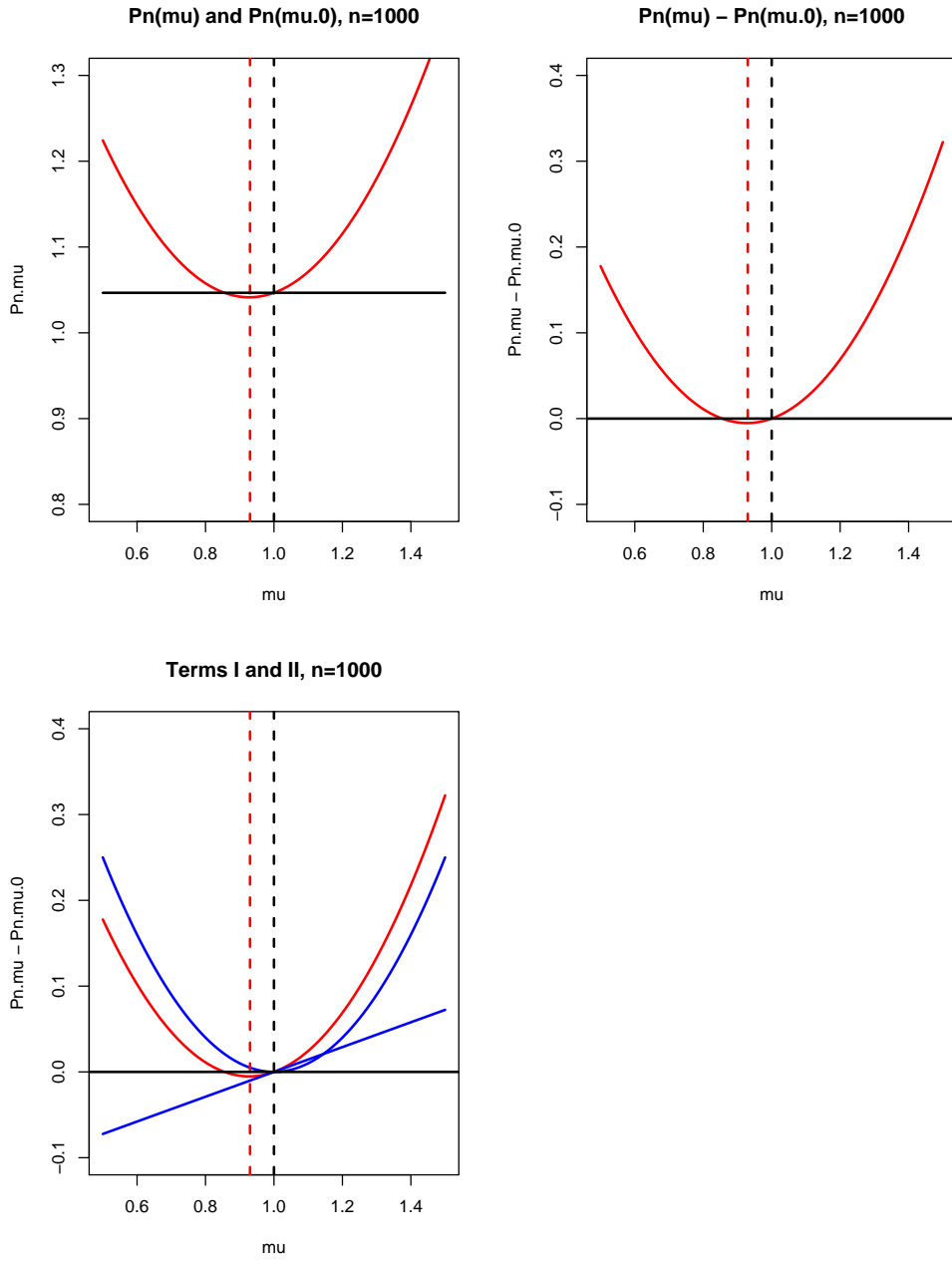


Figure 2: Same as Figure 1, but for a sample of size 1000.

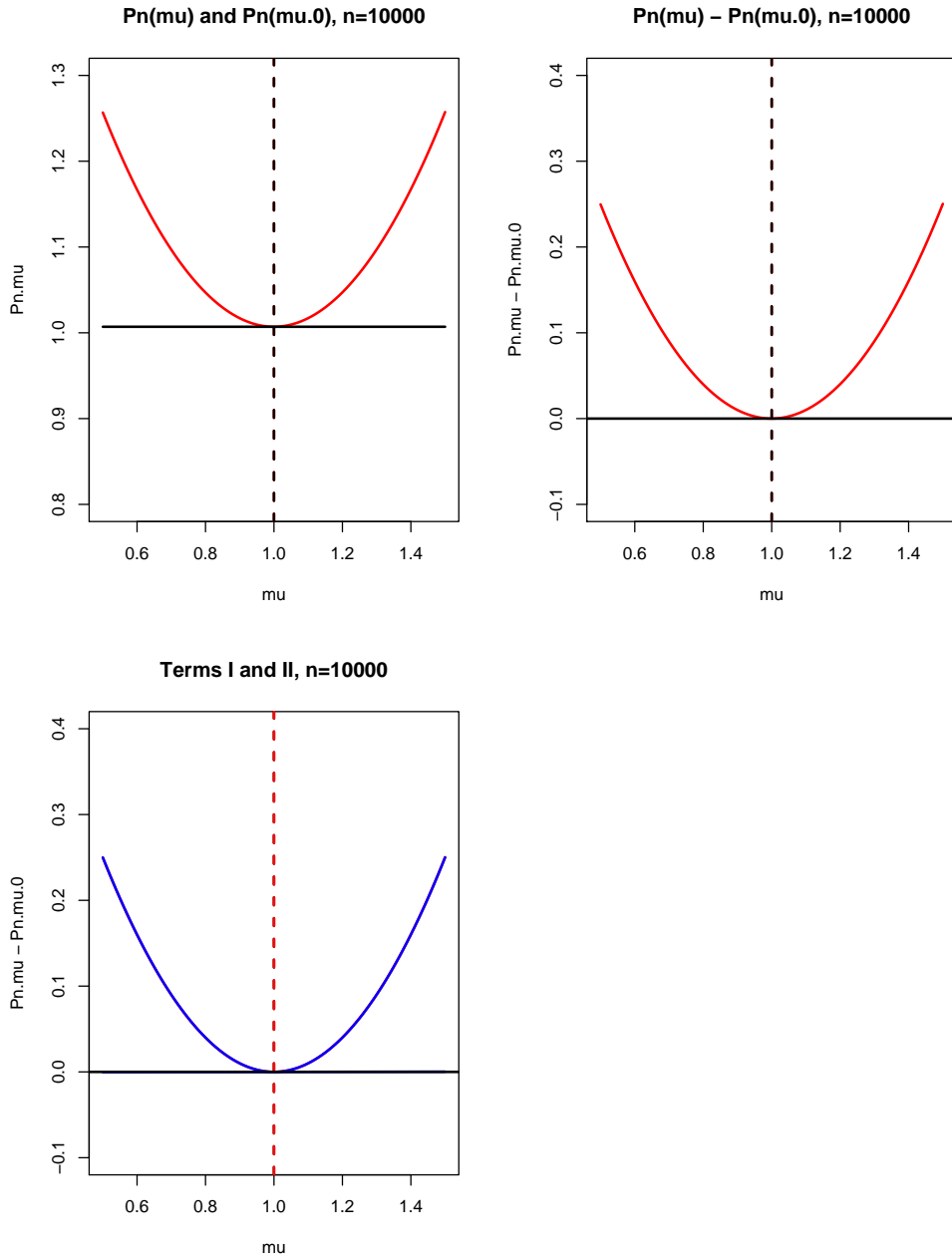


Figure 3: Same as Figure 1, but for a sample of size 10000.

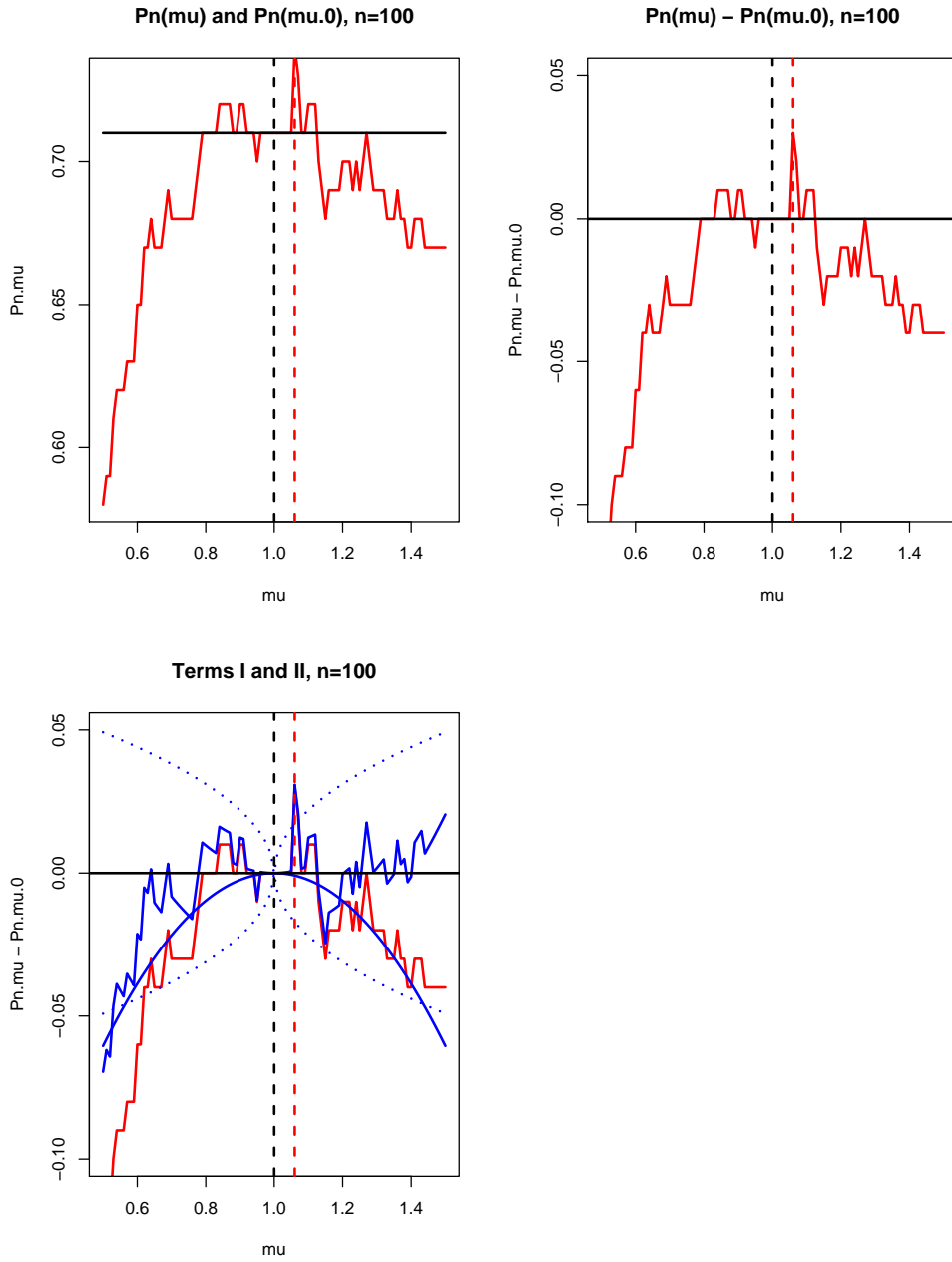


Figure 4: The top left panel shows  $\mathbb{P}_n 1\{\mu - 1 \leq x \leq \mu + 1\}$  (red) and  $\mathbb{P}_n 1\{\mu_0 - 1 \leq x \leq \mu_0 + 1\}$  (black). The values  $\tilde{\mu}_n$  (red) and  $\mu_0$  (black) are depicted by vertical dashed lines. The top right panel shows  $\mathbb{P}_n 1\{\mu - 1 \leq x \leq \mu + 1\} - \mathbb{P}_n 1\{\mu_0 - 1 \leq x \leq \mu_0 + 1\}$ . Note that this process is nonnegative at  $\tilde{\mu}_n$ . The lower left panel separates this process in the random term  $I$  (blue jagged line) and the deterministic term  $II$  (blue parabola). The blue dotted line indicate the size of term  $I$ . The figures are based on a random sample of size 1000 from a Normal(1,1) distribution.

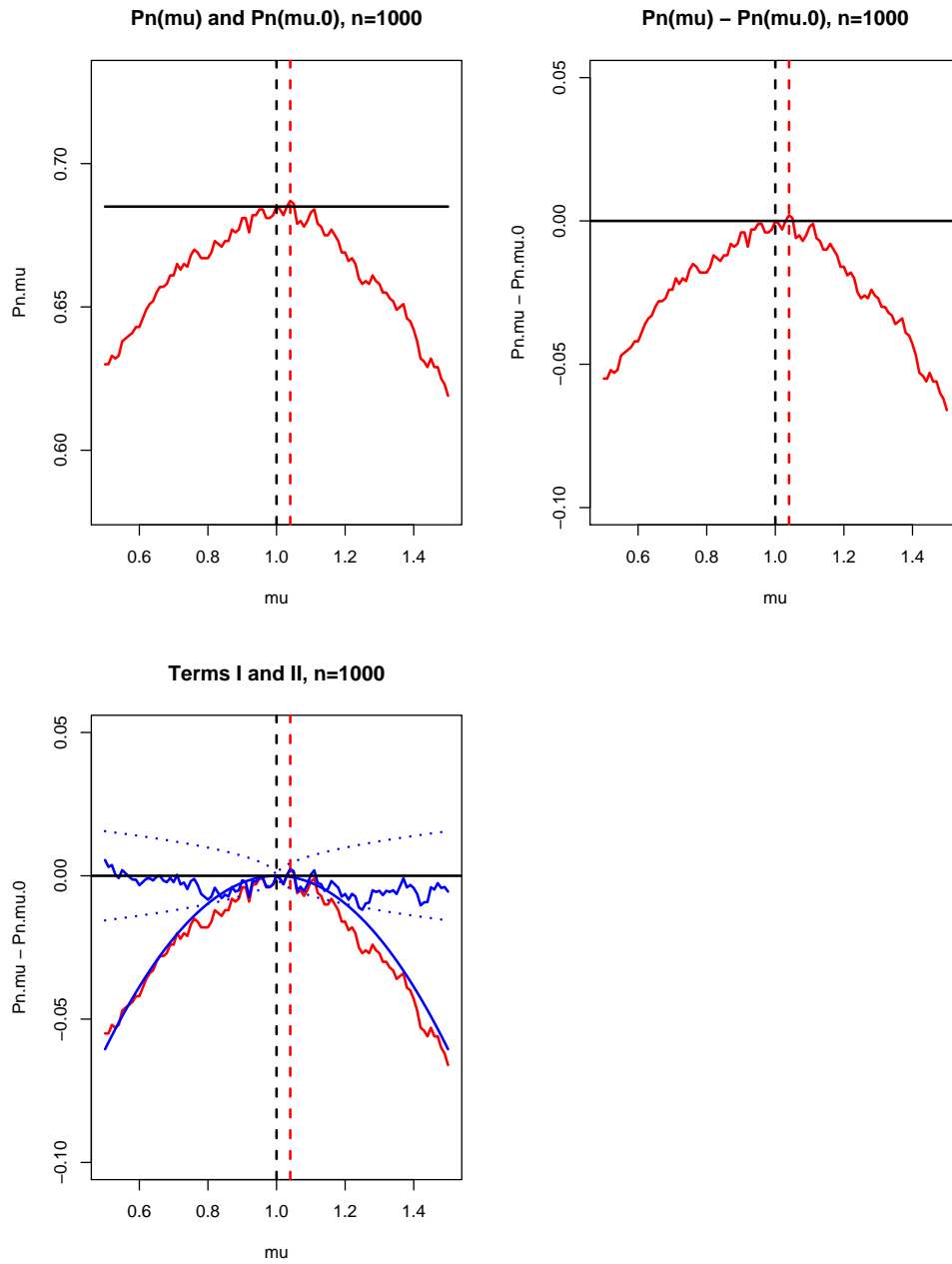


Figure 5: Same as Figure 4, but for sample size  $n = 1000$ .

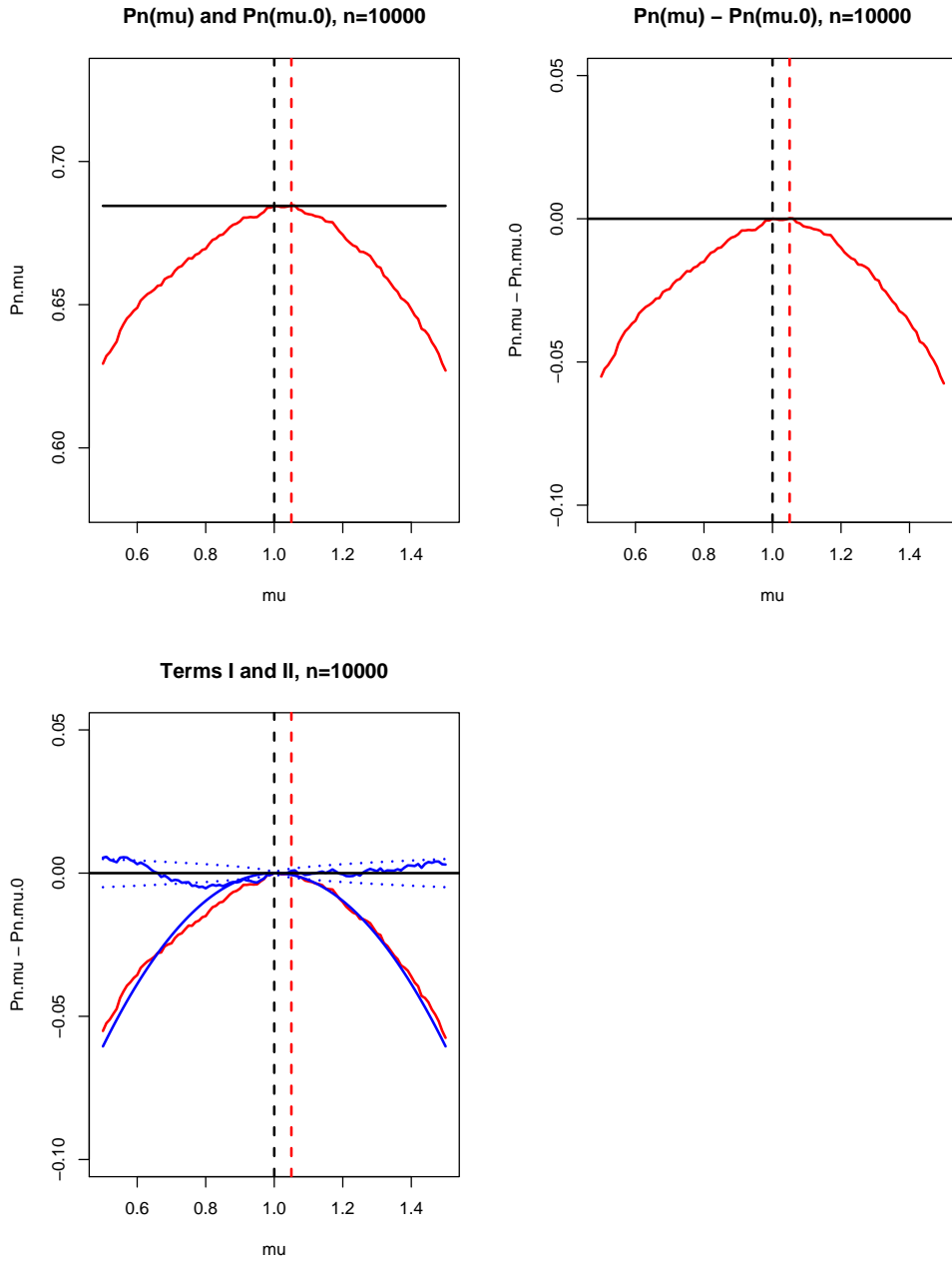


Figure 6: Same as Figure 4, but for sample size  $n = 10000$ .

## 4 Local limiting distribution

To get an intuitive idea of the limiting distribution of the MLE, please first read Section 4.1.

To make the reasoning given there more rigorous, we need some extra notation. For any interval  $I$ , let  $D(I)$  denote the collection of cadlag functions on  $I$  (right continuous with left limits), let  $D^-(I)$  denote the collection of caglad functions on  $I$  (left continuous with right limits), and let  $C(I)$  denote the collection of continuous functions on  $I$ . Moreover, we use the following conventions:

$$\begin{aligned} 1_{[t_0,t)}(u) &= -1_{[t,t_0)}(u), \\ \int_{[t_0,t)} f(u)dA(u) &= - \int_{[t,t_0)} dA(u), \\ \int_{t_0}^t f(u)dW(u) &= - \int_t^{t_0} f(u)dW(u), \end{aligned}$$

for any Lebesgue-Stieltjes measure  $dA$  and Brownian motion process  $W$ .

**Definition 4.1** *Let  $W$  be a two sided Brownian motion process with mean zero and variance*

$$E(W(s)W(t)) = (|s| \wedge |t|)1\{st > 0\}F_0(t_0)\{1 - F_0(t_0)\}/g(t_0).$$

*Let  $V(t) = W(t) + \frac{1}{2}f_0(t_0)t^2$ . Let  $H$  be the greatest convex minorant of  $V$ , i.e.,  $H$  is a convex function that satisfies:*

$$H(t) \leq V(t) \quad \text{for all } t,$$

*and equality holds if  $H$  has a change of slope.*

We will then prove the following theorem:

**Theorem 4.2** *Let  $0 < F_0(t_0) < 1$  and let  $F_0$  and  $G$  be continuously differentiable at  $t_0$  with continuous derivatives  $f_0(t_0)$  and  $g(t_0)$ . Then*

$$n^{1/3}(\widehat{F}_n(t_0 + n^{-1/3}t) - F_0(t_0)) \rightarrow_d H'(t) \quad \text{in the Skorohod topology on } D(\mathbb{R}).$$

The proof will contain the three steps given below. These steps can also be followed in more complicated problems:

1. Establish that the limiting process  $H$  exists and is unique.
2. Localize the characterization.
3. Take limits of subsequences of the localized processes (using tightness).

Step 1 is almost automatic: just take the pointwise maximum of all convex functions  $f$  for which  $f(t) \leq V(t)$  for all  $t$ . Step 2 is discussed in Section 4.1, and the main ideas for Step 3 are discussed in Section 4.2.

## 4.1 Localized characterization

The local rate of convergence (Theorem 3.1) leads to the following corollary which we will give without proof.

**Corollary 4.3** *Let  $\tau_n$  be the last jump point of  $\widehat{F}_n$  before  $t_0$ . Then  $t_0 - \tau_n = O_p(n^{-1/3})$ .*

**Definition 4.4** *We define the following localized processes:*

$$\begin{aligned}\widehat{F}_n^{loc}(t) &= n^{1/3}(\widehat{F}_n(t_0 + n^{-1/3}t) - F_0(t_0)) \\ V_n^{loc}(t) &= \frac{n^{2/3}}{g(t_0)} \int_{(t_0, t_0 + n^{-1/3}t]} \{\delta - F_0(t_0)\} d\mathbb{P}_n(u, \delta) \\ \widehat{H}_n^{loc}(t) &= \frac{n^{2/3}}{g(t_0)} \int_{t_0}^{t_0 + n^{-1/3}t} \{\widehat{F}_n(u) - F_0(t_0)\} dG(u) + \frac{n^{2/3}}{g(t_0)} c_n, \\ R_n^{loc}(t) &= \frac{n^{2/3}}{g(t_0)} \int_{[\tau_n, t_0 + n^{-1/3}t)} \{\widehat{F}_n(u) - F_0(t_0)\} d(G(u) - G_n(u)),\end{aligned}$$

where

$$c_n = \int_{\tau_n}^{t_0} \{\widehat{F}_n(u) - F_0(t_0)\} dG(u) - \int_{[\tau_n, t_0)} \{\delta - F_0(t_0)\} d\mathbb{P}_n(u, \delta) \quad (22)$$

Note that  $(\widehat{H}_n^{loc})'(t) = \widehat{F}_n^{loc}(t) + o(1)$ .

Let  $\tau_n$  be a jump point of  $\widehat{F}_n$ . Then the convex minorant characterization tells us that the MLE satisfies

$$\int_{[\tau_n, s)} \widehat{F}_n(u) dG_n(u) \leq \int_{[\tau_n, s)} \delta d\mathbb{P}_n(u, \delta), \quad \text{for all } s,$$

and equality must hold if  $s$  is also a jump point of  $\widehat{F}_n$ . We now ‘localize’ this expression by subtracting  $F_0(t_0)$  from the integrands on both sides:

$$\int_{[\tau_n, s)} \{\widehat{F}_n(u) - F_0(t_0)\} dG_n(u) \leq \int_{[\tau_n, s)} \{\delta - F_0(t_0)\} d\mathbb{P}_n(u, \delta).$$

Next, we replace  $G_n$  by  $G$ :

$$\int_{\tau_n}^s \{\widehat{F}_n(u) - F_0(t_0)\} dG(u) \leq \int_{[\tau_n, s)} \{\delta - F_0(t_0)\} d\mathbb{P}_n(u, \delta) + R_n(\tau_n, s),$$

where  $R_n(\tau_n, s) = \int_{[\tau_n, s)} \{\widehat{F}_n(u) - F_0(t_0)\} d(G(u) - G_n(u))$ . Subsequently, we replace the lower bound of the integrals by  $t_0$ :

$$\int_{t_0}^s \{\widehat{F}_n(u) - F_0(t_0)\} dG(u) + c_n \leq \int_{[t_0, s)} \{\delta - F_0(t_0)\} d\mathbb{P}_n(u, \delta) + R_n(\tau_n, s),$$

and we replace the upper bound of the integrals by  $t_0 + n^{-1/3}t$ :

$$\begin{aligned}\int_{t_0}^{t_0 + n^{-1/3}t} \{\widehat{F}_n(u) - F_0(t_0)\} dG(u) + c_n \\ \leq \int_{[t_0, t_0 + n^{-1/3}t)} \{\delta - F_0(t_0)\} d\mathbb{P}_n(u, \delta) + R_n(\tau_n, t_0 + n^{-1/3}t),\end{aligned}$$

with  $c_n$  defined in (22). Finally, by multiplying both sides by  $n^{2/3}/g(t_0)$ , this can be rewritten as

$$\widehat{H}_n^{loc}(t) \leq V_n^{loc}(t) + R_n^{loc}(t).$$

This is our localized characterization. One can show that  $R_n^{loc}(t) = o_p(1)$  for  $t \in [-m, m]$ . Hence,  $R_n^{loc}(t)$  will disappear in the limit. Furthermore,  $V_n^{loc}$  will become the drifted Brownian motion process  $V$ , and  $\widehat{H}_n^{loc}$  will become its greatest convex minorant  $H$ . Since  $\widehat{F}_n^{loc}(t) \approx (\widehat{H}_n^{loc})'(t)$ , we can see that the limiting distribution of the MLE will be given by the slope of the convex minorant of  $V$  at the point  $t$ .

## 4.2 Subsequences

Define  $\widehat{U}_n = (R_n^{loc}, V_n^{loc}, \widehat{H}_n^{loc}, \widehat{F}_n^{loc})$ . Let  $\widehat{U}_n|[-m, m]$  denote the restriction of  $\widehat{U}_n$  on  $[-m, m]$ . Note that  $R_n^{loc}$  is left continuous,  $V_n^{loc}$  is right continuous,  $\widehat{H}_n^{loc}$  is continuous, and  $\widehat{F}_n^{loc}$  is right continuous. Hence, an appropriate space for  $\widehat{U}_n|[-m, m]$  is

$$E[-m, m] = (D^-[-m, m]) \times D[-m, m] \times C[-m, m] \times D[-m, m] = I \times II \times III \times IV,$$

endowed with the uniform topology on  $I \times II \times III$  and the Skorohod topology on  $IV$ .

Take a subsequence of  $\widehat{U}_n|[-m, m]$ . Using the local rate result, we can prove that  $\widehat{U}_n|[-m, m]$  is tight in  $E[-m, m]$  for each fixed  $m$ . Hence, for each  $m$  there is a further subsequence that converges to some limit. By a diagonal argument, it follows that there is a limit on  $\mathbb{R}$ . By the continuous mapping theorem, the limit must satisfy the characterization on  $[-m, m]$  for each  $m$ . Letting  $m \rightarrow \infty$ , the limit must satisfy the characterization on  $\mathbb{R}$ . By uniqueness of the limiting process, all subsequences must converge to the same limit, that is characterized in Definition 4.1.

Using scaling properties of Brownian motion, one can rewrite the limiting distribution as a constant (depending on  $F_0(t_0)$ ,  $f_0(t_0)$  and  $g(t_0)$ ) times Chernoff's distribution. Chernoff's distribution is computed and tabulated in Groeneboom and Wellner (2001).

## References

- AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics* **26** 641–647.
- BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression*. John Wiley & Sons, New York.
- GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser Verlag, Basel.
- GROENEBOOM, P. and WELLNER, J. A. (2001). Computing Chernoff’s distribution **10** 388–400.
- KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *The Annals of Statistics* **18** 191–219.
- SCHICK, A. and YU, Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics* **27** 45–55.
- VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York.
- VAN DER VAART, A. W. and WELLNER, J. A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High Dimensional Probability II*. Birkhäuser, Boston, 115–133.
- WELLNER, J. A. (2005). Empirical processes: Theory and applications. Lecture notes for special topics course at Delft University of Technology, the Netherlands. Available at <http://www.stat.washington.edu/jaw/RESEARCH/TALKS/Delft/emp-proc-delft-big.pdf>.