Chapter 2
Parameter Estimation for an i.i.d. Model

Оценивание параметров в модели с независимыми одинаково распределёнными наблюдениями

Кадры, овладевшие техникой, решают всё!

Personnels that became proficient in technique decide everything!

Joseph Stalin

Exercise 2.1 (Glivenko-Cantelli theorem). *Let F be the distribution function of a random variable X and let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from F. Define the edf as*

$$F_n(x) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \mathbf{1}(X_i \le x).$$

Prove that

$$\sup_x \left| F_n(x) - F(x) \right| \xrightarrow{a.s.} 0, \qquad n \to \infty$$

1. *If F is a continuous distribution function;*
2. *If F is a discrete distribution function.*

W.K. Härdle et al., *Basics of Modern Mathematical Statistics*, Springer Texts in Statistics,
DOI 10.1007/978-3-642-36850-9_2, © Springer-Verlag Berlin Heidelberg 2014

1. Consider first the case when the function F is continuous in y. Fix any integer N and define with $\varepsilon = 1/N$ the points $t_1 < t_2 < \ldots < t_N = +\infty$ such that

$$F(t_j) - F(t_{j-1}) = \varepsilon \text{ for } j = 2, \ldots, N. \tag{2.1}$$

For every j, by the law of large numbers: $F_n(t_j) \xrightarrow{a.s.} F(t_j)$. This implies that for some $n(\varepsilon)$, it holds for all $n \geq n(\varepsilon)$

$$\left|F_n(t_j) - F(t_j)\right| \leq \varepsilon, \qquad j = 1, \ldots, N. \tag{2.2}$$

$F(t)$ and $F_n(t)$ are nondecreasing functions. This implies that for every $t \in [t_{j-1}, t_j]$ it holds

$$F(t_{j-1}) \leq F(t) \leq F(t_j), \quad F_n(t_{j-1}) \leq F_n(t) \leq F_n(t_j). \tag{2.3}$$

Let us subtract the first inequality (2.3) from the second:

$$F_n(t_{j-1}) - F(t_j) \leq F_n(t) - F(t) \leq F_n(t_j) - F(t_{j-1}), \tag{2.4}$$

Let us continue with the right hand side using (2.1) and (2.2):

$$\begin{aligned} F_n(t) - F(t) &\leq F_n(t_j) - F(t_{j-1}) \\ &= \underbrace{\left\{F_n(t_j) - F(t_j)\right\}}_{\leq \varepsilon} + \underbrace{\left\{F(t_j) - F(t_{j-1})\right\}}_{=\varepsilon} \leq 2\varepsilon, \end{aligned}$$

In the same way (considering the left part of (2.4)), one can prove that

$$F_n(t) - F(t) \geq -2\varepsilon$$

So,

$$\left|F_n(t) - F(t)\right| \leq 2\varepsilon. \tag{2.5}$$

Thus for all $\varepsilon > 0$ there exists constant $n(\varepsilon) > 0$ such that for every $n > n(\varepsilon)$ the inequality (2.5) holds for all $t \in \mathbb{R}$.

2. By $T = \{t_m\}_{m=1}^{+\infty}$ we denote points of discontinuity of function $F(x)$. Of course, these points are also points of discontinuity of function $F_n(t)$ (for any n).

Let us fix some $\varepsilon > 0$ and let us construct some finite set $S(\varepsilon)$. We include in $S(\varepsilon)$ the following points:

(a) Points such that at least one inequality fulfills:

$$F(t_m) - F(t_{m-1}) > \varepsilon \quad \text{or} \quad F(t_{m+1}) - F(t_m) > \varepsilon$$

(b) Continuous set of points such that

$$F(t_m) - F(t_{m-1}) < \varepsilon$$

Denote amount of elements in $S(\varepsilon)$ by M.

We know that $F_n(t) \to F(t)$ almost sure. In particular

$$F_n(t_m) \xrightarrow{a.s.} F(t_m), \quad \forall\, m \in S(\varepsilon).$$

By definition

$$\exists n_m(\varepsilon) \in \mathbb{N}: \quad \forall n > n_m(\varepsilon) \quad |F_n(t_m) - F(t_m)| < \varepsilon$$

Define $n(\varepsilon) \stackrel{\text{def}}{=} \max\{n_1(\varepsilon), \ldots, n_M(\varepsilon)\}$. Then for all $t_m \in S(\varepsilon)$

$$\forall n > n(\varepsilon) \quad |F_n(t_m) - F(t_m)| < \varepsilon.$$

Let us prove that the inequality

$$\forall n > n(\varepsilon) \quad |F_n(t_m) - F(t_m)| < 2\varepsilon. \tag{2.6}$$

is also true for all points $t_m \notin S(\varepsilon)$. Fix some $t_m \notin S(\varepsilon)$ and find index s such that

$$F(t_{s-1}) \le F(t_m) \le F(t_s), \qquad F_n(t_{s-1}) \le F_n(t_m) \le F_n(t_s).$$

Consider

$$\begin{aligned} F_n(t_m) - F(t_m) &\le F_n(t_s) - F(t_{s-1}) \\ &= \underbrace{\{F_n(t_s) - F(t_s)\}}_{<\varepsilon} + \underbrace{\{F(t_s) - F(t_{s-1})\}}_{\le\varepsilon} \le 2\varepsilon, \end{aligned}$$

Similarly, one can prove that

$$F_n(t_m) - F(t_m) \ge -2\varepsilon$$

This means that

$$|F_n(t_m) - F(t_m)| \le 2\varepsilon$$

So, (2.6) is true for all $t_m \in T$.

For all t there exists some point $t_m \in T$ such that

$$F_n(t) = F_n(t_m) \quad \text{and} \quad F(t) = F(t_m).$$

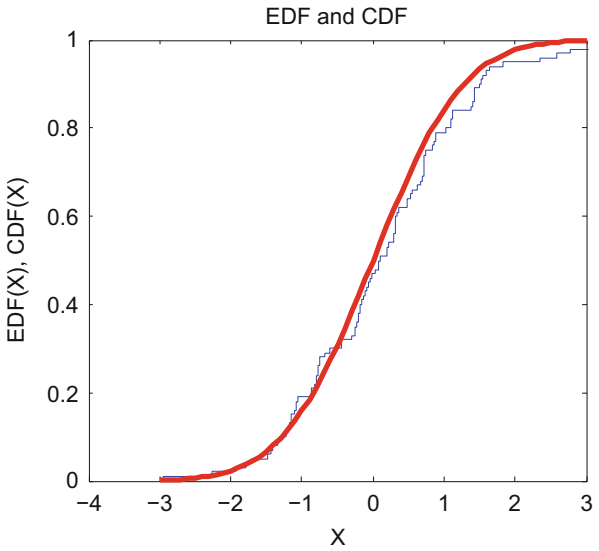

Fig. 2.1 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 100$. MSEedfnormal

Thus

$$\forall n > n(\varepsilon) \quad |F_n(t) - F(t)| < \varepsilon.$$

This observation completes the proof.

For an illustration of the asymptotic property, we draw $\{X_i\}_{i=1}^n$ i.i.d. samples from the standard normal distribution. Figure 2.1 shows the case of $n = 100$ and Fig. 2.2 shows the case of $n = 1{,}000$. The empirical cdf and theoretical cdf are close in the limit as n becomes larger.

Exercise 2.2 (Illustration of the Glivenko-Cantelli theorem). *Denote by F the cdf of*

1. *Standard normal law,*
2. *Exponential law with parameter $\lambda = 1$.*

Consider the sample $\{X_i\}_{i=1}^n$. Draw the plot of the empirical distribution function F_n and cumulative distribution function F. Find the index $i* \in \{1, \ldots, n\}$ such that

$$|F_n(X_{i*}) - F(X_{i*})| = \sup_i \big|F_n(X_i) - F(X_i)\big|.$$

The examples for the code can be found in the Quantnet. The readers are suggested to change the sample size n to compare the results (Figs. 2.3 and 2.4).

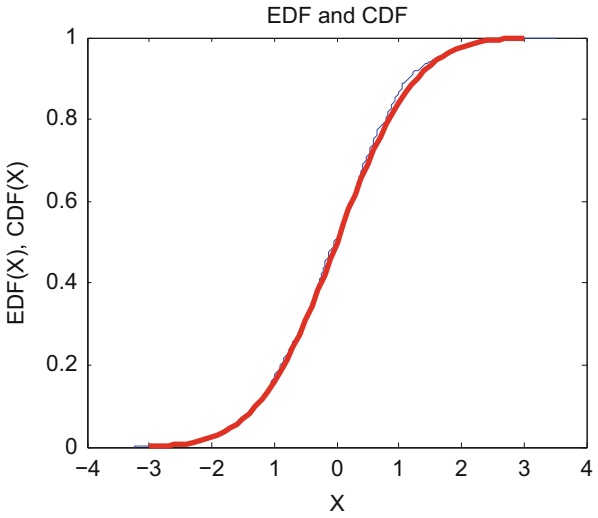

Fig. 2.2 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 1{,}000$. MSEedfnormal

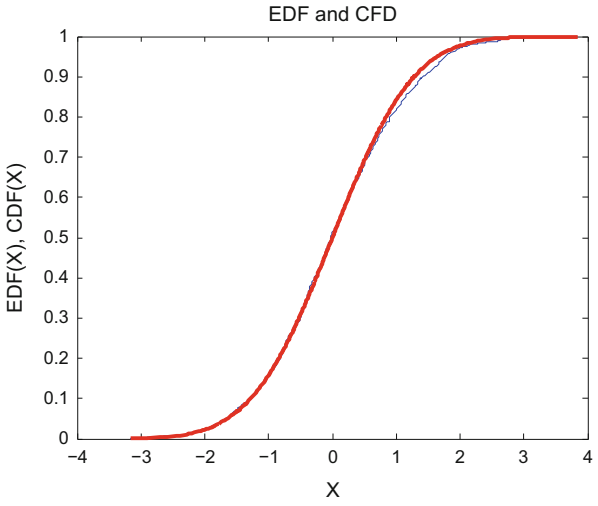

Fig. 2.3 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 1{,}000$. The maximal distance in this case occurs at $X_{i*} = 1.0646$ where $i* = 830$. MSEGCthmnorm

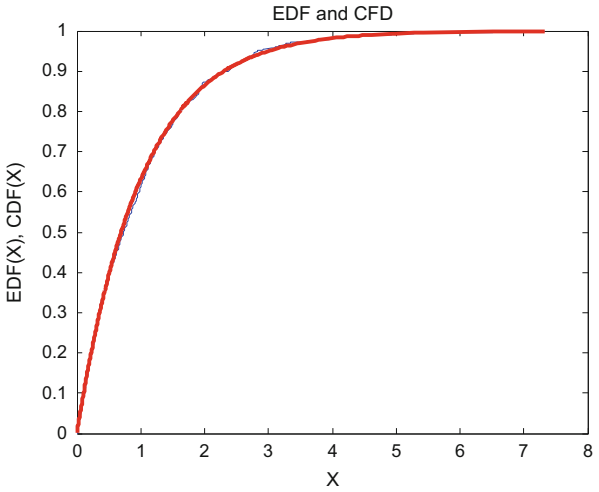

Fig. 2.4 The exponential ($\lambda = 1$) cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 1{,}000$. The maximal distance in this case occurs at $X_{i*} = 0.9184$ where $i* = 577$. MSEedfnormal

Exercise 2.3. *Compute the estimate of method of moments for the following parametric models:*

1. *Multinomial model:*

$$\mathbb{P}_\theta(X = k) = \binom{m}{k}\theta^k(1-\theta)^{m-k}, \qquad k = 0, \ldots, m.$$

2. *Exponential model*

$$\mathbb{P}_\theta(X > x) = e^{-x/\theta}.$$

In both cases one can follow the algorithm consisting of two steps:

- Calculate mathematical expectation $m(\theta) = \mathbb{E}_\theta X$;
- Solve the equation $m(\tilde{\theta}) = n^{-1}\sum_{i=1}^n X_i$; the solution is the required estimate.

Let us apply this:

1. Multinomial model, we first calculate expectation:

$$m(\theta) = n^{-1}\sum_{i=1}^n X_i = \sum_{k=0}^m k\binom{m}{k}\theta^k(1-\theta)^{m-k}$$
$$= m\sum_{k=1}^m \binom{m-1}{k-1}\theta^k(1-\theta)^{m-k} = m\theta.$$

Secondly we solve the equation

$$m(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

which gives the solution:

$$\tilde{\theta} = \frac{1}{nm}\sum_{i=1}^{n} X_i.$$

2. Exponential family. Both items are trivial: $m(\theta) = \frac{1}{\theta}$ and $\tilde{\theta} = n\left(\sum_{i=1}^{n} X_i\right)^{-1}$.

Exercise 2.4. *Let $\{X_i\}_{i=1}^{n}$ be an i.i.d. sample from a distribution with Lebesgue density*

$$f_\theta(x) = \frac{1}{2}(1+\theta x)\, I_{[-1,1]}(x)$$

1. *Find an estimator via the method of moments;*
2. *Find a consistent estimator.*

Let us begin with calculation of the mathematical expectation:

$$\mathbb{E}_\theta X_1 = \frac{1}{2}\int_{-1}^{1} (1+\theta x)\, x\, dx = \frac{1}{3}\theta$$

Both items of the exercise follow immediately:

1. The estimator of method of moments is a solution of the equality

$$\mathbb{E}_{\tilde{\theta}} X_1 = n^{-1}\sum_{i=1}^{n} X_i$$

So, $\tilde{\theta} = 3n^{-1}\sum_{i=1}^{n} X_i$

2. By the law of large numbers,

$$n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}X_i = \frac{1}{3}\theta, \quad n \to \infty.$$

This means that

$$3n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \theta, \quad n \to \infty,$$

hence the estimator $\hat{\theta} = 3n^{-1}\sum_{i=1}^{n} X_i$ is consistent.

Exercise 2.5. *Consider the model*

$$X_i = \theta^* + \varepsilon_i,$$

where θ^* is the parameter of interest and ε_i are independent normal errors $\mathcal{N}(0, \sigma_i^2)$.

Compute the MLE $\tilde{\theta}$ of the parameter θ^* and prove that this estimate has the following properties:

(a) The estimate $\tilde{\theta}$ is unbiased: $\mathbb{E}_{\theta^*}\tilde{\theta} = \theta^*$.
(b) The quadratic risk of $\tilde{\theta}$ is equal to

$$\mathcal{R}(\tilde{\theta}, \theta^*) \stackrel{\text{def}}{=} \mathbb{E}_{\theta^*}|\tilde{\theta} - \theta^*|^2 = (\sum_{i=1}^{n} \sigma_i^2)^{-1}.$$

The corresponding log-likelihood reads

$$L(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\Big\{\log(2\pi\sigma_i^2) + \frac{(X_i - \theta)^2}{\sigma_i^2}\Big\}.$$

The first derivative is equal to

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^{n}\frac{X_i - \theta}{\sigma_i^2} = \sum_{i=1}^{n}\frac{X_i}{\sigma_i^2} - \theta\sum_{i=1}^{n}\frac{1}{\sigma_i^2}.$$

Then the MLE $\tilde{\theta}$ equals

$$\tilde{\theta} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta} L(\theta) = \frac{1}{N}\sum\frac{X_i}{\sigma_i^2},$$

where $N = \sum \sigma_i^{-2}$.
(a)

$$\mathbb{E}_{\theta^*}\tilde{\theta} = \frac{1}{N}\sum\frac{\mathbb{E}X_i}{\sigma_i^2} = \frac{1}{N}\sum\frac{\theta^*}{\sigma_i^2} = \frac{\theta^*}{N}\sum\frac{1}{\sigma_i^2} = \theta^*.$$

(b) The quadratic risk of $\tilde{\theta}$ is equal to the variance $\mathrm{Var}(\tilde{\theta})$:

$$\begin{aligned}
\mathcal{R}(\tilde{\theta}, \theta^*) \stackrel{\text{def}}{=} \mathbb{E}_{\theta^*}|\tilde{\theta}^\circ - \theta^*|^2 &= \mathbb{E}_{\theta^*}\left|\frac{1}{N}\sum\frac{X_i}{\sigma_i^2} - \theta^*\right|^2 \\
&= \mathbb{E}_{\theta^*}\left|\frac{1}{N}\sum\frac{X_i}{\sigma_i^2} - \theta^*\frac{1}{N}\sum\frac{1}{\sigma_i^2}\right|^2 \\
&= \mathbb{E}_{\theta^*}\left|\frac{1}{N}\sum\frac{X_i - \theta^*}{\sigma_i^2}\right|^2 = \frac{1}{N^2}\sum\mathbb{E}_{\theta^*}\left|\frac{X_i - \theta^*}{\sigma_i^2}\right|^2.
\end{aligned}$$

Note that random value $X_i - \theta^*$ has a normal distribution with zero mean and variance σ_i^2. Then $(X_i - \theta^*)/\sigma_i^2 \sim \mathcal{N}(0, \sigma_i^{-2})$ and

$$\mathcal{R}(\tilde{\theta}^\circ, \theta^*) = \frac{1}{N^2} \sum \sigma_i^{-2} = \frac{1}{N}.$$

Exercise 2.6. *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample with distribution that depends on some parameter θ. Let $\hat{\theta}_n$ be an estimate of parameter θ.*

Assume that this estimate is root-n normal, i.e. there exists a function $\sigma(\theta)$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma(\theta)^2\right), \qquad n \to \infty.$$

Prove that $\hat{\theta}_n$ is consistent,

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$$

This fact can be briefly formulated as "root-n normality implies consistency". We need Slutsky's Theorem:

1. Let a_n (sequence of real numbers) be convergent in probability,

$$a_n \xrightarrow{\mathbb{P}} a, \qquad n \to \infty$$

Let η_n (sequence of random variables) be convergent in distribution,

$$\eta_n \xrightarrow{\mathcal{L}} \mathrm{Law}(\eta), \qquad n \to \infty$$

Then

$$a_n \eta_n \xrightarrow{\mathcal{L}} \mathrm{Law}(a\eta), \qquad n \to \infty$$

2. Let ξ_n be a sequence of random variables that converges in law to the distribution that is degenerated in some point c (we denote this degenerated distribution by $\mathrm{Law}(c)$). Then ξ_n also tends to c in probability.

Let us apply these observations to our situation. We use the first part of Slutsky's Theorem with $a_n = \frac{1}{\sqrt{n}}$ and $\xi_n = \sqrt{n}(\hat{\theta}_n - \theta)$.

The sequence a_n tends to zero and the sequence ξ_n tends in probability to a normal distribution. So,

$$a_n \xi_n \xrightarrow{\mathcal{L}} \mathrm{Law}(0)$$

According to the second part, this sequence also tends to zero in probability. Thus,

$$a_n \xi_n = \hat{\theta}_n - \theta \xrightarrow{\mathbb{P}} 0.$$

Remark 2.1. In fact our proof is true for any estimate that has an asymptotic distribution (not necessarily normal).

Exercise 2.7. *Let F be the distribution function of a random variable X and let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from F. Let $g : \mathbb{R} \to \mathbb{R}$ be a function such that*

$$\sigma_g^2 \stackrel{\text{def}}{=} \operatorname{Var}\{g(X)\} < \infty$$

Denote

$$s \stackrel{\text{def}}{=} \mathbb{E} g(X), \quad S_n \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n g(X_i)$$

1. *Prove that*
   (a) $S_n \xrightarrow{\mathbb{P}} s, \qquad n \to \infty$
   (b) $\sqrt{n}(S_n - s) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_g^2), \qquad n \to \infty.$
2. *Let $h(z)$ be a twice continuously differentiable function on the real line such that $h'(s) \neq 0$ and $h''(s)$ is bounded in some neighborhood of s. Prove that*
   (a) $h(S_n) \xrightarrow{\mathbb{P}} h(s)$
   (b) $\sqrt{n}\{h(S_n) - h(s)\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_h^2), \qquad n \to \infty,$
   where $\sigma_h^2 \stackrel{\text{def}}{=} |h'(s)|^2 \sigma_g^2$.

1. (a) Note that $\{g(X_i)\}_{i=1}^n$ is a sample from the distribution with expectation equal to $\mathbb{E} g(X)$.
   One can apply the law of large numbers for the sequence $\{g(X_i)\}_{i=1}^n$:

$$n^{-1} \sum_{i=1}^n g(X_i) \xrightarrow{\mathbb{P}} \mathbb{E} g(X) \qquad n \to \infty.$$

   (b) This statement directly follows by the CLT for i.i.d. random variables:

$$\frac{n^{-1} \sum_{i=1}^n g(X_i) - \mathbb{E} g(X)}{\sqrt{\frac{1}{n} \operatorname{Var}\{g(X)\}}} \sim \mathcal{N}(0, 1)$$

In other words,

$$\sqrt{n}\left\{n^{-1}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X)\right\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_g^2), \qquad n \to \infty.$$

2. (a) We know:

$$S_n \xrightarrow{\mathbb{P}} s, \qquad n \to \infty$$

Then for any continuous function g:

$$g(S_n) \xrightarrow{\mathbb{P}} g(s), \quad n \to \infty$$

(b) One can find a neighborhood U of the point s such that

(i) S_n belongs with high probability to U;
(ii) $h''(s)$ is bounded in U.

Applying the Taylor expansion to h in this neighborhood U:

$$\sqrt{n}\left\{h(S_n) - h(s)\right\} = \sqrt{n}h'(s)\,(S_n - s) + \frac{\sqrt{n}}{2}h''(\tilde{s})\,(S_n - s)^2, \quad (2.7)$$

where $\tilde{s}$ is some point between s and S_n. The right hand side of (2.7) is a sum of two random variables. First random variable $\sqrt{n}h'(s)\,(S_n - s)$ tends to $\mathcal{N}\left(0, |h'(s)|^2\sigma_g^2\right)$ in distribution.

Let us show that the second component tends to zero in probability. Actually,

$$\left|\frac{\sqrt{n}}{2}h''(\tilde{s})\,(S_n - s)^2\right| \le \frac{U}{2}\frac{1}{\sqrt{n}}\left\{\sqrt{n}\,(S_n - s)\right\}^2,$$

where U is an upper bound for $h''(s)$ in the considering neighborhood. Expression in the right hand side is a product of the sequence $\frac{1}{\sqrt{n}}$, which tends to zero, and sequence $\left\{\sqrt{n}\,(S_n - s)\right\}^2$, which converges in distribution. Then

$$\left|\frac{\sqrt{n}}{2}h''(\tilde{s})\,(S_n - s)^2\right| \xrightarrow{\mathbb{P}} 0$$

Thus, the right hand side in (2.7) (and left hand side also) tends to $\mathcal{N}(0, |h'(s)|^2\sigma_g^2)$ in distribution.

Exercise 2.8. *(Analogue of the Exercise 2.7 for multi-dimensional case) Let* $\boldsymbol{g}(\cdot) = (g_1(\cdot), \dots, g_m(\cdot))^\top : \mathbb{R} \to \mathbb{R}^m$ *be a function such that*

$$\Sigma_{jk} \stackrel{\text{def}}{=} \mathbb{E}\left[g_j(X) g_k(X)\right] < \infty, \text{ for } j, k \le m.$$

Denote

$$\boldsymbol{s} = \mathbb{E}\boldsymbol{g}(X) = (\mathbb{E}g_1(X), \dots, \mathbb{E}g_m(X))^\top,$$
$$\boldsymbol{S}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{g}(X_i) = \left(\frac{1}{n}\sum_{i=1}^n g_1(X_i), \dots, \frac{1}{n}\sum_{i=1}^n g_m(X_i)\right)^\top.$$

1. *Prove that*
   (a) $\boldsymbol{S}_n \xrightarrow{\mathbb{P}} \boldsymbol{s}, \qquad n \to \infty$
   (b) $\sqrt{n}(\boldsymbol{S}_n - \boldsymbol{s}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma), \quad n \to \infty,$
   where $\Sigma = (\Sigma_{jk})_{j,k=1,\dots,m}$
2. *Let* $H(z) : \mathbb{R}^m \to \mathbb{R}$ *be a twice continuously differentiable function such that* $\nabla H(z)$ *and* $\left\|\nabla^2 H(z)\right\|$ *is bounded in some neighborhood of* $\boldsymbol{s}$. *Prove that*
   (a) $H(\boldsymbol{S}_n) \xrightarrow{\mathbb{P}} H(\boldsymbol{s})$
   (b) $\sqrt{n}\left\{H(\boldsymbol{S}_n) - H(\boldsymbol{s})\right\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_H^2), \qquad n \to \infty,$
   where $\sigma_H^2 \stackrel{\text{def}}{=} \nabla H(\boldsymbol{s})^\top \Sigma \nabla H(\boldsymbol{s})$.

First note that items 1a and 2a follow from items 1b and 2b correspondingly. Let us check items 1b and 2b.

Consider for every $\boldsymbol{v} = (v_1, \dots, v_m)^\top \in \mathbb{R}^m$ the scalar products $\boldsymbol{v}^\top \boldsymbol{g}(\cdot)$, $\boldsymbol{v}^\top \boldsymbol{s}$, $\boldsymbol{v}^\top \boldsymbol{S}_n$. For the statement 1b, it suffices to show that

$$\sqrt{n}\boldsymbol{v}^\top\left(\boldsymbol{S}_n - \boldsymbol{s}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{v}^\top \Sigma \boldsymbol{v}), \qquad n \to \infty.$$

Actually

$$\begin{aligned}
\sqrt{n}\boldsymbol{v}^\top\left(\boldsymbol{S}_n - \boldsymbol{s}\right) &= \sqrt{n}\sum_j v_j \left\{\frac{1}{n}\sum_i g_j(X_i) - \mathbb{E}g_j(X)\right\} \\
&= \sqrt{n}\left[\frac{1}{n}\sum_i \left\{\sum_j v_j g_j(X_i)\right\} - \mathbb{E}\left\{\sum_j v_j g_j(X)\right\}\right] \\
&= \sqrt{n}\left\{\frac{1}{n}\sum_i G(X_i) - \mathbb{E}G(X)\right\},
\end{aligned}$$

where $G(\cdot) = \sum_j v_j g_j(\cdot) = \boldsymbol{v}^\top \boldsymbol{g}(\cdot)$.

Now one can apply result of the Exercise 2.7 (item 1b) for the function $G(\cdot)$ and obtain the required statement.

For the statement 2b, consider the Taylor expansion

$$\sqrt{n}\,\{H(\boldsymbol{S}_n) - H(\boldsymbol{s})\} = \sqrt{n}\,\nabla H(s)^\top (\boldsymbol{S}_n - \boldsymbol{s}) + \frac{\sqrt{n}}{2} (\boldsymbol{S}_n - \boldsymbol{s})^\top \nabla^2 H(\tilde{s})\,(\boldsymbol{S}_n - \boldsymbol{s})\,.$$

This formula is an analogue of (2.7). One can continue the line of reasoning in the same way as in the proof of (2.7) (item 2b).

In fact,

$$\sqrt{n}\nabla H(s)^\top (\boldsymbol{S}_n - \boldsymbol{s}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla H(\boldsymbol{s})^\top \Sigma \nabla H(\boldsymbol{s})),$$

and

$$\left| \frac{\sqrt{n}}{2} (\boldsymbol{S}_n - \boldsymbol{s})^\top \nabla^2 H(\tilde{s})\,(\boldsymbol{S}_n - \boldsymbol{s}) \right| \;\le\; \frac{1}{2\sqrt{n}} \left\| \sqrt{n}\,(\boldsymbol{S}_n - \boldsymbol{s}) \right\|^2 \max_s \left\| \nabla^2 H(s) \right\| \xrightarrow{\mathbb{P}} 0$$

These two observations conclude the proof.

Exercise 2.9.

1. *Consider a sample* $\{X_i\}_{i=1}^n$ *from a distribution* $P_{\theta^*} \in (P_\theta, \theta \in \Theta \in \mathbb{R})$. *Let* $\tilde{\theta}$ *be an estimator of* θ *such that the bias*

$$b(\tilde{\theta}, \theta^*) \stackrel{\text{def}}{=} \mathbb{E}_{\theta^*} \tilde{\theta} - \theta^*$$

and the variance $\mathrm{Var}_{\theta^*}(\tilde{\theta})$ *tend to zero as* $n \to \infty$. *Prove that* $\tilde{\theta}$ *is consistent.*

2. *Let* $\{X_i\}_{i=1}^n$ *be a sample from the uniform distribution on* $[0, \theta]$. *Using the first item of this exercise, prove that the estimator*

$$\widetilde{\theta_1} = \max\{X_1, \ldots, X_n\}$$

is consistent.

1. Applying the so called bias-variance decomposition, which is true for any estimate $\tilde{\theta}$:

$$\mathbb{E}_{\boldsymbol{\theta}^*} \left(\tilde{\theta} - \theta^* \right)^2 = \mathrm{Var}_{\theta^*}(\tilde{\theta}) + b^2(\tilde{\theta}, \theta^*). \tag{2.8}$$

Let us prove (2.8):

$$\begin{aligned}\mathbb{E}_{\theta^*}\left(\tilde{\theta}-\theta^*\right)^2 &= \mathbb{E}_{\theta^*}\left\{\tilde{\theta}-\mathbb{E}(\tilde{\theta})+\mathbb{E}(\tilde{\theta})-\theta^*\right\}^2\\ &= \mathbb{E}_{\theta^*}\left\{\tilde{\theta}-\mathbb{E}(\tilde{\theta})+b(\tilde{\theta},\theta^*)\right\}^2\\ &= \mathrm{Var}_{\theta^*}(\tilde{\theta})+2b(\tilde{\theta},\theta^*)\mathbb{E}_{\theta^*}\left(\tilde{\theta}-\mathbb{E}\tilde{\theta}\right)+b^2(\tilde{\theta},\theta^*)\\ &= \mathrm{Var}_{\theta^*}(\tilde{\theta})+b^2(\tilde{\theta},\theta^*)\end{aligned}$$

If bias and variance tend to zero as $n\to\infty$, then

$$\mathbb{E}_{\theta^*}\left(\tilde{\theta}-\theta^*\right)^2\to 0,\quad n\to\infty$$

This means that $\tilde{\theta}$ tends to θ^* in L_2 sense. Then $\tilde{\theta}$ also tends to θ^* in probability, i.e. $\tilde{\theta}$ is a consistent estimator.

2. First of all, let us calculate the cdf of $\widetilde{\theta_1}$.

$$\begin{aligned}\mathbb{P}_{\theta^*}\left(\tilde{\theta}_1\le x\right) &= \mathbb{P}_{\theta^*}\left(X_1\le x,\dots,X_n\le x\right)\\ &= \left\{\mathbb{P}_{\theta^*}\left(X_1\le x\right)\right\}^n=\left(\frac{x}{\theta^*}\right)^n,\qquad x\in[0,\theta^*]\end{aligned}$$

Afterwards we can take the derivative and obtain the density function

$$p(x)=n(\theta^*)^{-n}x^{n-1}\mathbf{1}\left(0\le x\le\theta^*\right)$$

For applying the first item, one has to calculate expectation and variance of $\tilde{\theta}_1$:

$$\mathbb{E}\tilde{\theta}_1=\frac{n}{n+1}\theta^*,\quad \mathrm{Var}(\tilde{\theta}_1)=\frac{n}{(n+1)^2(n+2)}\theta^{*2}$$

Now we are ready for applying the first item:

$$b(\tilde{\theta}_1,\theta^*)=\frac{n}{n+1}\theta^*-\theta^*=-\frac{1}{n+1}\theta^*\to 0,\quad n\to\infty.$$

$$\mathrm{Var}_{\theta^*}(\tilde{\theta})=\frac{n}{(n+1)^2(n+2)}\theta^{*2}\to 0,\quad n\to\infty.$$

So, assumptions are fulfilled. This concludes the proof.

Exercise 2.10. *Check that the i.i.d. experiment from the uniform distribution on the interval* $[0,\theta]$ *with unknown* θ *is not regular.*

First condition from the definition of the regular family is the following one: the sets $A(\theta) \stackrel{\text{def}}{=} \{y : p(y, \theta) = 0\}$ are the same for all $\theta \in \Theta$.

The uniform distribution on the interval $[0, \theta]$ doesn't satisfy this condition,

$$A(\theta) = (-\infty, 0) \cup (\theta, +\infty)\,.$$

This exercise gives a local approximation of the Kullback-Leibler divergence.

Exercise 2.11. *Let (P_θ) be a regular family.*

1. *Show that the KL-divergence $\mathcal{K}(\theta, \theta')$ satisfies for any θ, θ':*

   (a)

   $$\mathcal{K}(\theta, \theta')\Big|_{\theta'=\theta} = 0;$$

   (b)

   $$\frac{d}{d\theta'}\mathcal{K}(\theta, \theta')\Big|_{\theta'=\theta} = 0;$$

   (c)

   $$\frac{d^2}{d\theta'^2}\mathcal{K}(\theta, \theta')\Big|_{\theta'=\theta} = I(\theta).$$

2. *Show that in a small neighborhood of θ, the KL-divergence can be approximated by*

   $$\mathcal{K}(\theta, \theta') \approx I(\theta)|\theta' - \theta|^2/2.$$

1. Note that

$$\mathcal{K}(\theta, \theta') = \mathbb{E}_\theta \log p(x, \theta) - \mathbb{E}_\theta \log p(x, \theta')$$

(a) First item is trivial.
(b)

$$\begin{aligned}
\frac{d}{d\theta'}\mathcal{K}(\theta, \theta') &= -\frac{d}{d\theta'}\mathbb{E}_\theta \log p(x, \theta') \\
&= -\frac{d}{d\theta'}\int \log p(x, \theta')\, p(x, \theta) dx \\
&= -\int \frac{p'_{\theta'}(x, \theta')}{p(x, \theta')}\, p(x, \theta) dx,
\end{aligned}$$

where $p'_{\theta'}(x,\theta') \stackrel{\text{def}}{=} \frac{d}{d\theta'} p(x,\theta')$. Substitution $\theta' = \theta$ gives

$$\begin{aligned}\frac{d}{d\theta'}\mathcal{K}(\theta,\theta')\Big|_{\theta'=\theta} &= -\int \frac{d}{d\theta'}\{p(x,\theta')\}dx\Big|_{\theta'=\theta} \\ &= -\frac{d}{d\theta'}\int p(x,\theta')dx\Big|_{\theta'=\theta} = 0.\end{aligned}$$

(c)

$$\begin{aligned}\frac{d^2}{d\theta'^2}\mathcal{K}(\theta,\theta') &= -\int \frac{d}{d\theta'}\left\{\frac{p'_{\theta'}(x,\theta')}{p(x,\theta')}\right\} p(x,\theta)dx \\ &= -\int\left[\frac{p''_{\theta'}(x,\theta')p(x,\theta') - \{p'_{\theta'}(x,\theta')\}^2}{\{p(x,\theta')\}^2}\right] p(x,\theta)dx.\end{aligned}$$

Substitution $\theta' = \theta$ yields

$$\frac{d^2}{d\theta'^2}\mathcal{K}(\theta,\theta')\Big|_{\theta'=\theta} = \underbrace{\int p''_{\theta'}(x,\theta')dx\Big|_{\theta'=\theta}}_{\frac{d^2}{d\theta'^2}\int p(x,\theta')dx\Big|_{\theta'=\theta}=0} + \underbrace{\int \frac{\{p'_\theta(x,\theta)\}^2}{p(x,\theta)}dx}_{=I(\theta)} = I(\theta).$$

2. The required representation directly follows from the Taylor expansion at the point $\theta' = \theta$.

The following exercise

1. Illustrates two methods for checking the R-efficiency;
2. Shows that the Fisher information can depend on the parameter (for some parametric families), but can be a constant (for other parametric families).

Exercise 2.12. *Consider two families:*
(a) the Gaussian shift (b) the Poisson family

1. *Compute the Fisher Information for these families.*
2. *Check that the Cramér-Rao inequality for the empirical mean estimate* $\tilde{\theta} = n^{-1}\sum_{i=1}^n X_i$ *is in fact an equality, i.e.*

$$\mathrm{Var}_\theta(\tilde{\theta}) = n^{-1}I^{-1}(\theta).$$

3. *Check R-efficiency of* $\tilde{\theta}$
   (i) Using only the definition;
   (ii) Using the Theorem 2.6.3. of Spokoiny and Dickhaus (2014)

1. (a) Recall

$$p(x,\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\}.$$

Then

$$\begin{aligned} I(\theta) = \mathbb{E}_\theta \left| \frac{\partial \log p(X,\theta)}{\partial \theta} \right|^2 &= \mathbb{E}_\theta \left| \frac{\partial}{\partial \theta} \left\{-\frac{(X-\theta)^2}{2}\right\} \right|^2 \\ &= \mathbb{E}_\theta \, |X-\theta|^2 \\ &= \mathbb{E}_\theta \, |X - \mathbb{E}_\theta X|^2 \\ &= \mathrm{Var}(X) = 1. \end{aligned}$$

Therefore, the Fisher information is equal to 1 for any values of the parameter θ.

(b)

$$p(x,\theta) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 1, 2, \ldots$$

$$\begin{aligned} I(\theta) = \mathbb{E}_\theta \left| \frac{\partial \log p(X,\theta)}{\partial \theta} \right|^2 &= \mathbb{E}_\theta \left| \frac{\partial}{\partial \theta} (X \log \theta - \log X! - \theta) \right|^2 \\ &= \mathbb{E}_\theta \left| \frac{X}{\theta} - 1 \right|^2 = \frac{1}{\theta^2} \mathbb{E}_\theta \, |X-\theta|^2 \\ &= \frac{1}{\theta^2} \mathbb{E}_\theta \, |X - \mathbb{E}_\theta X|^2 = \frac{1}{\theta^2} \mathrm{Var}_\theta(X) = \frac{1}{\theta}. \end{aligned}$$

So, in the case of the Poisson family, the Fisher information depends on θ.

2. Estimator $\tilde{\theta}$ is unbiased for both cases. Then the Cramér-Rao inequality stands that

$$\mathrm{Var}_\theta(\tilde{\theta}) = \mathrm{Var}_\theta \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} \mathrm{Var}_\theta(X_1) \geq n^{-1} I^{-1}(\theta).$$

So, the aim is to check that

$$\mathrm{Var}_\theta(X_1) \, I(\theta) = 1. \tag{2.9}$$

(a) For the Gaussian shift $\mathrm{Var}_\theta(X_1) = 1$ and $I(\theta) = 1$. Hence, (2.9) is fulfilled.
(b) For the Poisson family, $\mathrm{Var}_\theta(X_1) = \theta$ and $I(\theta) = 1/\theta$. Hence, (2.9) is also fulfilled.

3. (i) The definition says that R-efficient estimators are exactly the estimators that give the equality in the Cramér-Rao inequality. So, this item is already proved.
   (ii) The estimate $\tilde{\theta}$ can be represented as

$$\tilde{\theta} = n^{-1} \sum U(Y_i)$$

with $U(x) = x$. The aim is to show that the log-density $\ell(y, \theta)$ of P_θ can be represented as

$$\ell(x, \theta) = C(\theta)x - B(\theta) + \ell(x), \tag{2.10}$$

for some functions $C(\cdot)$ and $B(\cdot)$ on Θ and a function $\ell(\cdot)$ on $\mathbb{R}$.
(a)

$$\ell(x, \theta) = \theta x - \theta^2/2 + \left(-\frac{x^2}{2} + \log \frac{1}{\sqrt{2\pi}} \right),$$

and (2.10) follows with $C(\theta) = \theta$, $B(\theta) = \theta^2/2$, and $\ell(x) = -x^2/2 + \log 1/\sqrt{2\pi}$.
(b)

$$\ell(x, \theta) = \log(\theta)x - \theta + \log(x!), \tag{2.11}$$

and (2.11) follows with $C(\theta) = \log \theta$, $B(\theta) = \theta$, and $\ell(x) = \log(x!)$.

Exercise 2.13. *Let $\boldsymbol{X}$ be a random variable with a distribution from $(P_\theta, \theta \in \Theta \subset \mathbb{R})$. Let also a function $\psi^\circ : \mathbb{R} \times \Theta \to \mathbb{R}$ be such that*

$$\psi^\circ(x, \theta) = a\,(x - \theta)^2 + b\,(x - \theta) + c,$$

where $a, b, c \in \mathbb{R}$.

1. *Find a condition on the constants a, b, c and the family (P_θ) such that the function $\psi^\circ(x, \theta)$ is a contrast.*
2. *Find a condition on the constants a, b, c such that such that the function $\psi^\circ(x, \theta)$ is a contrast for the model of the Gaussian shift $\mathcal{N}(\theta, 1)$.*

1. By definition, the function ψ° is a contrast if and only if

$$\operatorname*{argmin}_{\theta'} \mathbb{E}_\theta\, \psi^\circ(\boldsymbol{X}, \theta') = \theta, \quad \forall \theta.$$

Introduce a function

$$\begin{aligned} f(\theta, \theta') &\stackrel{\text{def}}{=} \mathbb{E}_\theta \left[\psi^\circ(\boldsymbol{X}, \theta') \right] \\ &= \left(a\, \mathbb{E}_\theta \boldsymbol{X}^2 + b\, \mathbb{E}_\theta \boldsymbol{X} + c \right) - (2a \mathbb{E}_\theta \boldsymbol{X} + b)\, \theta' + a\theta'^2. \end{aligned} \tag{2.12}$$

The aim is to find a condition on the constants a, b, c and the family (P_θ) such that

$$\operatorname*{argmin}_{\theta'} f(\theta, \theta') = \theta, \quad \forall \theta. \tag{2.13}$$

Take the derivative of the function $f(\theta, \theta')$ with respect to θ' and solve the equation $\partial f(\theta, \theta')/\partial \theta' = 0$:

$$\frac{\partial f(\theta, \theta')}{\partial \theta'} = -(2a\mathbb{E}_\theta \boldsymbol{X} + b) + 2a\theta' = 0$$

This means that

$$\operatorname*{argmin}_{\theta'} f(\theta, \theta') = \mathbb{E}_\theta \boldsymbol{X} + \frac{b}{2a}.$$

Together with (2.13), this yields the required condition on the constants a, b and the family (P_θ):

$$\theta = \mathbb{E}_\theta \boldsymbol{X} + \frac{b}{2a}, \quad \forall \theta. \tag{2.14}$$

Constant c can be chosen arbitrary.

2. For the model of the Gaussian shift $\mathcal{N}(\theta, 1)$,

$$\mathbb{E}_\theta \boldsymbol{X} = \theta, \quad \forall \theta.$$

Condition (2.14) in this case yields $b = 0$. This means, that any function $\psi^\circ(x, \theta)$ with $b = 0$ and any constants a and c is a contrast for the Gaussian shift.

Exercise 2.14. *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution $P_{\theta^*} \in (P_\theta, \theta \in \Theta \subset \mathbb{R})$.*

1. *Let also $g(x)$ satisfy $\int g(x) dP_{\theta^*}(x) = \theta^*$, leading to the moment estimate*

$$\tilde{\theta} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n g(X_i).$$

Show that this estimate can be obtained as the M-estimate for a properly selected function $\psi(\cdot)$.

2. *Let $\int g(x) dP_{\theta^*}(x) = m(\theta^*)$ for the given functions $g(\cdot)$ and strictly monotonic and continuously differentiable $m(\cdot)$. Show that the moment estimate $\tilde{\theta} = m^{-1}\{\sum g(X_i)/n\}$ can be obtained as the M-estimate for a properly selected function $\psi(\cdot)$.*

1. It is the worth mentioning that

$$\tilde{\theta} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} g(X_i) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \{g(X_i) - \theta\}^2 . \tag{2.15}$$

This observation helps us to find an appropriate function ψ. Fix

$$\psi(x, \theta) \stackrel{\text{def}}{=} \{g(x) - \theta\}^2$$

and prove that

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\theta^*} \psi(\boldsymbol{X}, \theta), \tag{2.16}$$

where $\boldsymbol{X}$ is a variable that has the distribution P_{θ^*}.

The proof of (2.16) is straightforward:

$$\mathbb{E}_{\theta^*} \psi(\boldsymbol{X}, \theta) = \mathbb{E}_{\theta^*} \{g(\boldsymbol{X}) - \theta\}^2 = \mathbb{E}_{\theta^*} g^2(\boldsymbol{X}) - 2\theta \mathbb{E}_{\theta^*} g(\boldsymbol{X}) + \theta^2$$

Minimizing the right hand side expression by θ yields

$$\theta_{min} = \mathbb{E}_{\theta^*} g(\boldsymbol{X}) = \theta^*.$$

This concludes the proof.

2. The proof follows the same lines as the proof of the first statement. Note that

$$\tilde{\theta} \stackrel{\text{def}}{=} m^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(X_i) \right\} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \{g(X_i) - m(\theta)\}^2 .$$

Function

$$\psi(x, \theta) \stackrel{\text{def}}{=} \{g(x) - m(\theta)\}^2$$

is appropriate because of

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\theta^*} \{g(\boldsymbol{X}) - m(\theta)\}^2 \tag{2.17}$$

In fact, fix some $\theta \in \Theta$ and find a minimum value of the function

$$f(\theta) = \mathbb{E}_{\theta^*} \{g(\boldsymbol{X}) - m(\theta)\}^2$$

In order to minimize this function, solve the equation $f'(\theta) = 0$:

$$\frac{df(\theta)}{d\theta} = \frac{df(\theta)}{dm(\theta)} \frac{dm(\theta)}{d\theta} = 2\mathbb{E}_\theta^* \{g(\boldsymbol{X}) - m(\theta)\} \frac{dm(\theta)}{d\theta} = 0$$

The first derivative of the function $m(\theta)$ doesn't change the sign because of monotonicity. This means that the minimum value of function f satisfies the following equation

$$m(\theta_{min}) = \mathbb{E}_{\theta *} g(\boldsymbol{X}).$$

Then (2.17) fulfills. This completes the proof.

Exercise 2.15. *Let $\{X_i\}_{i=1}^{n_1}$ be a sample from the distribution with the pdf*

$$p(x, \theta) = \frac{2x}{\theta^2}, \quad x \in [0, \theta].$$

Find the MLE of the median of the distribution.

First let us find a relation between θ and the median m. By the definition of the median,

$$\int_{-\infty}^{m} p(x, \theta) dx = \int_0^m \frac{2x}{\theta^2} dx = 1/2,$$

i.e. $\theta = \sqrt{2}m$. Then the likelihood function

$$L(m) = \prod_{i=1}^{n} p(X_i, \sqrt{2}m) = \prod_{i=1}^{n} \frac{X_i}{m^2} \mathbf{1}\left(X_i \in [0, \sqrt{2}m]\right)$$

has a maximum at the point $\hat{m} = \max_i X_i / \sqrt{2}$.

Exercise 2.16. *Let $\left\{X_i^{(1)}\right\}_{i=1}^{n_1}$ and $\left\{X_i^{(2)}\right\}_{i=1}^{n_2}$ be two independent samples from the Poisson distributions with unknown parameters μ_1 and $\mu_2 = \mu_1 + \mu$ correspondingly. Find the maximum likelihood estimator for the parameter μ.*
Hint: Is it possible to find separately $\hat{\mu}_1$ (the MLE for μ_1) from the first sample, $\hat{\mu}_2$ (the MLE for μ_2) from the second sample, and then obtain the MLE estimator for μ as the difference $\hat{\mu} = \hat{\mu}_1 - \hat{\mu}_2$?

Denote by $L_1(\mu_1)$ and $L_2(\mu_2)$ the log-likelihood functions for the first and the second samples correspondingly.

The MLE estimate for the parameter μ is determined as

$$(\hat{\mu}_1, \hat{\mu}) = \underset{\mu_1, \mu}{\operatorname{argmax}} \{L_1(\mu_1) + L_2(\mu_1 + \mu)\}.$$

Hence, $\hat{\mu} = \operatorname{argmax}_{\mu} L_2(\hat{\mu}_1+\mu)$. The maximal value of the function L_2 is achieved at the point $\hat{\mu}_2$. This yields

$$\max_{\mu} L_2(\hat{\mu}_1 + \mu) = L_2(\hat{\mu}_2) = L_2\{\hat{\mu}_1 + (\hat{\mu}_2 - \hat{\mu}_1)\}.$$

So, $\hat{\mu} = \hat{\mu}_2 - \hat{\mu}_1$.

In the case of the Poisson distribution,

$$L_1(\mu_1) = \sum_{i=1}^{n_1} \log\left(e^{-\mu_1}\frac{\mu_1^{X_i^{(1)}}}{X_i^{(1)}!}\right) \quad \text{and} \quad L_2(\mu_2) = \sum_{i=1}^{n_2} \log\left(e^{-\mu_2}\frac{\mu_2^{X_i^{(2)}}}{X_i^{(2)}!}\right),$$

and the MLE of the parameter is the mean value, i.e. $\hat{\mu}_j = n_j^{-1}\sum_{i=1}^{n_j} X_i^{(j)} \stackrel{\text{def}}{=} \overline{X}^{(j)}$, $j = 1, 2$. Thus, we conclude that

$$\hat{\mu} = \overline{X}^{(2)} - \overline{X}^{(1)}.$$

Exercise 2.17. *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution with the Lebesque density*

$$p(x, \boldsymbol{\theta}) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} I_{(0,\beta)}(x),$$

where $\alpha, \beta > 0$ and $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\alpha, \beta)$. Find estimators for the multivariate parameter $\boldsymbol{\theta}$ using the following approaches:

1. *Maximum likelihood approach;*
2. *Method of moments.*

1. The likelihood function in this case

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(X_i, \boldsymbol{\theta}) = \frac{\alpha^n}{\beta^{\alpha n}} \prod_{i=1}^{n} X_i^{\alpha-1} I_{(0,\beta)}(X_i)$$

$$= \frac{\alpha^n}{\beta^{\alpha n}} I_{(0,\beta)}(X_{(n)}) \prod_{i=1}^{n} X_i^{\alpha-1}$$

is equal to zero if $\beta < X_{(n)}$ and decreases for $\beta \geq X_{(n)}$. Therefore the maximum likelihood estimator for the parameter β is $\tilde{\beta} = X_{(n)}$. In order to find MLE for the parameter α, one should maximize the function

$$f(\alpha) = C_1 \alpha^n C_2^{\alpha-1},$$

where $C_1 = I_{(0,\tilde{\beta})}(X_{(n)})\tilde{\beta}^{-n}$, $C_2 = \prod_{i=1}^n X_i/\tilde{\beta} = \prod_{i=1}^n X_i/X_{(n)}$. The equation $f'(\alpha) = 0$ gives the MLE of the parameter α:

$$\tilde{\alpha} = \frac{n}{-\log C_2} = \frac{n}{\sum_{i=1}^n \log \frac{X_{(n)}}{X_i}}.$$

So, the MLE is

$$\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\beta}) = \left(\frac{n}{\sum_{i=1}^n \log \frac{X_{(n)}}{X_i}},\ X_{(n)} \right).$$

2. Firstly we compute the first and the second moments:

$$m_1(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} X_1 = \int x p(x, \boldsymbol{\theta}) dx = \int_0^\beta \frac{\alpha}{\beta^\alpha} x^\alpha dx = \frac{\alpha\beta}{\alpha + 1}$$

$$m_2(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} X_1^2 = \int x^2 p(x, \boldsymbol{\theta}) dx = \int_0^\beta \frac{\alpha}{\beta^\alpha} x^{\alpha+1} dx = \frac{\alpha\beta^2}{\alpha + 2}$$

The empirical counterparts are

$$M_1 = \frac{1}{n}\sum_{i=1}^n X_i, \qquad M_2 = \frac{1}{n}\sum_{i=1}^n X_i^2. \tag{2.18}$$

The required estimators are the solutions of the system of equations

$$\begin{cases} M_1 = \alpha\beta/(\alpha + 1) \\ M_2 = \alpha\beta^2/(\alpha + 2) \end{cases} \tag{2.19}$$

Raise both parts of the first equation to the second power and divide it to the second equation:

$$\frac{M_1}{M_2} = \frac{\alpha(\alpha + 2)}{(\alpha + 1)^2}.$$

This yields the following quadratic equation w.r.t α:

$$\alpha^2 + 2\alpha + \frac{M_1}{M_1 - M_2} = 0. \tag{2.20}$$

If $\frac{M_1}{M_1 - M_2} < 0$ (or equivalently $M_1 < M_2$) then (2.20) has one positive solution

$$\hat{\alpha} = -1 + \sqrt{1 - \frac{M_1}{M_1 - M_2}}.$$

The first equation of system (2.19) gives

$$\hat{\beta} = \frac{\hat{\alpha}+1}{\hat{\alpha}} M_1.$$

So, the estimate by the method of moments is

$$(\hat{\alpha}, \hat{\beta}) = \left(-1 + \sqrt{1 - \frac{M_1}{M_1 - M_2}}, \quad \frac{\sqrt{1 - \frac{M_1}{M_1 - M_2}}}{-1 + \sqrt{1 - \frac{M_1}{M_1 - M_2}}} M_1 \right),$$

where M_1 and M_2 are given by (2.18).

Exercise 2.18. *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution with the Lebesgue density that depends on the parameter $\theta \in \mathbb{R}$ (σ is a fixed positive number):*

$$p(x, \theta) = (2\sigma)^{-1} e^{-|x-\theta|/\sigma}.$$

Compute the maximum likelihood estimate for the parameter θ.

This model is known as a shift of a Laplace law.
The maximum likelihood approach leads to maximizing the sum

$$L(\theta) = -n \log(2\sigma) - \sum_{i=1}^{n} |X_i - \theta|/\sigma,$$

or equivalently to minimizing the sum $\sum_{i=1}^n |X_i - \theta|$:

$$\tilde{\theta} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} |X_i - \theta|.$$

Order the observations $X_{(1)} \le X_{(2)} \le \cdots X_{(n)}$ and consider two cases.

1. Suppose that n is even. Denote $k = n/2 \in \mathbb{N}$. It is worth mentioning that

$$|X_{(1)} - \theta| + |X_{(n)} - \theta| \ge |X_{(n)} - X_{(1)}|, \tag{2.21}$$

where equality takes place if and only if $\theta \in \left[X_{(1)}, X_{(n)}\right]$ Analogously,

$$|X_{(2)} - \theta| + |X_{(n-1)} - \theta| \ge |X_{(n-1)} - X_{(2)}| \tag{2.22}$$

$$\cdots$$

$$|X_{(k)} - \theta| + |X_{(k+1)} - \theta| \ge |X_{(k+1)} - X_{(k)}| \tag{2.23}$$

This yields that

$$\sum_{i=1}^{n} |X_i - \theta| = \sum_{i=1}^{n} |X_{(i)} - \theta| = \sum_{j=1}^{k} \left(|X_{(j)} - \theta| + |X_{(n-j+1)} - \theta| \right)$$

$$\geq \sum_{j=1}^{k} |X_{(n-j+1)} - X_{(j)}|. \quad (2.24)$$

Equality in (2.24) takes place if and only if all the inequalities (2.21)–(2.23) are in fact equalities. This means that $\operatorname{argmin} \sum |X_i - \theta|$ is minimized by any $\theta \in \left[X_{(k)}, X_{(k+1)} \right]$, in particular by

$$\tilde{\theta} = \operatorname{med} X_i = \frac{X_{(k)} + X_{(k+1)}}{2}.$$

2. Suppose that n is odd. Denote $k = (n-1)/2 \in \mathbb{N}$. Equalities (2.21)–(2.23) are still true. This yields the analogue for (2.24):

$$\sum_{i=1}^{n} |X_i - \theta| = \sum_{i=1}^{n} |X_{(i)} - \theta| = \underbrace{|X_{(k+1)} - \theta|}_{\geq 0} + \sum_{j=1}^{k} \underbrace{|X_{(j)} - \theta| + |X_{(n-j+1)} - \theta|}_{\geq |X_{(n-j+1)} - X_{(j)}|}$$

$$\geq \sum_{j=1}^{k} |X_{(n-j+1)} - X_{(j)}|. \quad (2.25)$$

Note that the following two equalities take place only in the case of $\tilde{\theta} = \operatorname{med} X_i = X_{(k+1)}$:

$$|X_{(k+1)} - \theta| = 0$$

$$\sum_{j=1}^{k} \left[|X_{(j)} - \theta| + |X_{(n-j+1)} - \theta| \right] = \sum_{j=1}^{k} |X_{(n-j+1)} - X_{(j)}|$$

This completes the proof.

Exercise 2.19. *Consider the volatility model with parameter θ:*

$$Y = \xi^2, \quad \xi \sim \mathcal{N}(0, \theta).$$

1. *Prove that θ is a natural parameter.*
2. *Find a canonical parameter for this model.*
3. *Compute the Fisher information for this model with canonical parameter.*

1. The proof is straightforward:

$$\mathbb{E}Y = \mathbb{E}\xi^2 = \underbrace{\operatorname{Var}\xi}_{=\theta} + \underbrace{(\mathbb{E}\xi)}_{=0}^{\,2} = \theta.$$

2. Denote by $p_\xi(x)$ the pdf of ξ:

$$p_\xi(x) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right).$$

The density function of Y can be derived from $p_\xi(x)$:

$$p_Y(y,\theta) = \frac{1}{2\sqrt{y}} p_\xi(\sqrt{y}) = \frac{1}{2\sqrt{2\pi\theta y}} \exp\left(-\frac{y}{2\theta}\right)$$
$$= \frac{1}{2\sqrt{2\pi y}} \exp\left(-\frac{y}{2\theta} - \frac{1}{2}\log\theta\right). \quad (2.26)$$

This density representation means that $C(\theta) = -(2\theta)^{-1}$. The canonical parameter is determined by the equality $v \stackrel{\text{def}}{=} C(\theta)$, i.e. $v = -(2\theta)^{-1}$. This yields

$$p_Y(y,v) = \frac{1}{2\sqrt{2\pi y}} \exp\{yv - d(v)\},$$

where $d(v) = 1/2\log\{-1/(2v)\}$.

3. According to the general theory,

$$I(v) = d''(v) = \frac{1}{2v^2}.$$

Exercise 2.20. *Let (P_v) be a Gaussian shift experiment, that is $P_v = \mathcal{N}(v,1)$, $v \in \mathbb{R}$. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution P_{v^*}.*

1. *Is the parameter v a natural parameter? Is it a canonical parameter?*
2. *Check that*

$$\mathcal{K}(v_1, v_2) = (v_1 - v_2)^2/2.$$

3. *Check that for any v_0 and any $C > 0$, the equation*

$$\mathcal{K}(v_0 + u, v_0) = C \quad (2.27)$$

has only one positive (u^+) and only one negative (u^-) solution.

4. *Compute the maximum likelihood estimator $\tilde{v}$ and check that*

$$L(\tilde{v}, v) = (\tilde{v} - v)^2 n/2.$$

5. *Fix some $\zeta > 0$. Consider the equation (2.27) with $v_0 = v^*$ and $C = \zeta/n$. According to item 3, this equation has two solutions: denote the positive solution by u^+, and the negative solution by u^-. Denote also $v^+ = v^* + u^+$, and $v^- = v^* + u^-$.*

   (a) Compute the sets $\{L(\tilde{v}, v^*) \geq \zeta\}$, $\{L(v^+, v^*) \geq \zeta\}$, $\{L(v^-, v^*) \geq \zeta\}$.
   (b) Check that

$$\{L(\tilde{v}, v^*) \geq \zeta\} \subseteq \{L(v^+, v^*) \geq \zeta\} \cup \{L(v^-, v^*) \geq \zeta\}.$$

   Note that the last item is fulfilled for any v^* (not necessary the true value).

1. Parameter v is a natural parameter, because the expected value of a r.v. with distribution $\mathcal{N}(v, 1)$ is equal to v. The parameter v is also a canonical parameter, because the density function can be represented in the following way

$$p(x, v) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x - v)^2}{2}\right\} = p(x) \exp\{xv - d(v)\},$$

   where

$$p(x) = \varphi(x), \qquad d(v) = \frac{v^2}{2}.$$

2. According to the formula for the canonical parametrization,

$$\mathcal{K}(v_1, v_2) = d'(v_1)(v_1 - v_2) - \{d(v_1) - d(v_2)\}. \tag{2.28}$$

   In the case of a Gaussian shift, (2.28) yields

$$\mathcal{K}(v_1, v_2) = v_1(v_1 - v_2) - \frac{v_1^2 - v_2^2}{2} = (v_1 - v_2)\left(v_1 - \frac{v_1 + v_2}{2}\right) = \frac{(v_1 - v_2)^2}{2}.$$

3. The statement is a straightforward corollary from the previous item:

$$\mathcal{K}(u, v_0) = \frac{(v_0 + u - v_0)^2}{2} = \frac{u^2}{2} = C.$$

   This equation has two solutions: one positive $u^+ = \sqrt{2C}$ and one negative $u^- = -\sqrt{2C}$.

4. The maximum likelihood approach leads to maximizing the sum

$$L(v) = n \log \frac{1}{2\pi} - \sum_{i=1}^{n} \frac{(X_i - v)^2}{2}.$$

Then the maximum likelihood estimator is equal to $\tilde{v} = \sum_i X_i / n$. Consider the difference between $L(\tilde{v})$ and $L(v)$:

$$\begin{aligned} L(\tilde{v}, v) = L(\tilde{v}) - L(v) &= -\sum_i \frac{(X_i - \tilde{v})^2}{2} + \sum_i \frac{(X_i - v)^2}{2} \\ &= \frac{1}{2} \sum_i \left\{ (X_i - v)^2 - (X_i - \tilde{v})^2 \right\} = \frac{1}{2} \sum_i (2X_i - \tilde{v} - v)(\tilde{v} - v) \\ &= \frac{1}{2} (\underbrace{2 \sum_i X_i}_{2n\tilde{v}} - n\tilde{v} - nv)(\tilde{v} - v) = \frac{(\tilde{v} - v)^2 n}{2}. \end{aligned} \tag{2.29}$$

5. (a) Formula (2.29) yields

$$\begin{aligned} \{L(\tilde{v}, v^*) \geq \zeta\} &= \left\{ \frac{(\tilde{v} - v^*)^2 n}{2} \geq \zeta \right\} \\ &= \{\tilde{v} \geq \sqrt{\frac{2\zeta}{n}} + v^*\} \cup \{\tilde{v} \leq -\sqrt{\frac{2\zeta}{n}} + v^*\} \end{aligned}$$

From (2.27) (in item 3) we know that $v^+ = v^* + \sqrt{2\zeta/n}$ and $v^- = v^* - \sqrt{2\zeta/n}$. Then

$$\begin{aligned} \{L(v^+, v^*) \geq \zeta\} &= \{\frac{n}{2}(2\tilde{v} - v^+ - v^*)(\tilde{v} - v^*) \geq \zeta\} \\ &= \{\frac{n}{2}(2\tilde{v} - 2v^* - \sqrt{\frac{2\zeta}{n}})\sqrt{\frac{2\zeta}{n}} \geq \zeta\} \\ &= \{\tilde{v} \geq v^* + \sqrt{\frac{2\zeta}{n}}\}. \end{aligned}$$

Analogously,

$$\{L(v^-, v^*) \geq \zeta\} \supset \{\tilde{v} \leq v^* - \sqrt{\frac{2\zeta}{n}}\}.$$

(b) The required embedding is trivial.

Natural parametrization has some "nice" properties:

1.

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

2.

$$L(\tilde{\theta}, \theta) = n\mathcal{K}(\mathbb{P}_{\tilde{\theta}}, \mathbb{P}_{\theta}).$$

The following exercise shows, that the choice of parametrization is crucial for the first property, but the second one is fulfilled for any parametrization.

Exercise 2.21. *Let (P_θ) be an exponential family (θ – **any** parameter). Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from distribution that belongs to (P_θ), and X be a random variable with the same distribution.*

Show that the maximum likelihood estimator $\tilde{\theta}$ has the following properties:

1.

$$\mathbb{E}_{\tilde{\theta}} X = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

2.

$$L(\tilde{\theta}, \theta) = n\mathcal{K}(\mathbb{P}_{\tilde{\theta}}, \mathbb{P}_{\theta}).$$

1. $\tilde{\theta}$ is a point of maximum of the function

$$L(\theta) = \sum_{i=1}^{n} \log p(X_i, \theta) = C(\theta)\sum_{i=1}^{n} X_i - nB(\theta).$$

Differentiating w.r.t θ yields the equation for $\tilde{\theta}$:

$$C'(\tilde{\theta})\sum_{i=1}^{n} X_i - nB'(\tilde{\theta}) = 0. \tag{2.30}$$

On the other hand, differentiating both sides of the equality

$$\int p(x, \theta)dx = 1$$

w.r.t. θ yields

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta}\{p(x,\theta)\}dx = \int \frac{\partial}{\partial\theta}\{\log p(x,\theta)\}\, p(x,\theta)dx \\ &= \int \{xC'(\theta) - B'(\theta)\}\, p(x,\theta)dx \\ &= C'(\theta)\underbrace{\int xp(x,\theta)dx}_{=\mathbb{E}_\theta X} - B'(\theta)\underbrace{\int p(x,\theta)dx}_{=1}. \end{aligned}$$

This means that the equality

$$C'(\theta)\mathbb{E}_\theta X - B'(\theta) = 0$$

holds for any parameter θ, in particular for $\theta = \tilde{\theta}$:

$$C'(\tilde{\theta})\mathbb{E}_{\tilde{\theta}} X - B'(\tilde{\theta}) = 0. \tag{2.31}$$

Comparison of the equations (2.30) and (2.31) (using positivity of the first derivative of function $C(\theta)$) completes the proof.

2. Transformation of the left-hand side yields:

$$\begin{aligned} L(\tilde{\theta},\theta) &= \sum_{i=1}^{n} \Big\{\log p(X_i,\tilde{\theta}) - \log p(X_i,\theta)\Big\} \\ &= \{C(\tilde{\theta}) - C(\theta)\}\sum_{i=1}^{n} X_i - n\{B(\tilde{\theta}) - B(\theta)\}. \end{aligned} \tag{2.32}$$

The Kullback-Leibler divergence in the right-hand side can be transformed in the following way:

$$\begin{aligned} \mathcal{K}(\mathbb{P}_{\tilde{\theta}}, \mathbb{P}_\theta) &= \int \log\{\frac{p(x,\tilde{\theta})}{p(x,\theta)}\} P_{\tilde{\theta}}(dx) \\ &= \{C(\tilde{\theta}) - C(\theta)\}\int xP_{\tilde{\theta}}(dx) - \{B(\tilde{\theta}) - B(\theta)\} \\ &= \{C(\tilde{\theta}) - C(\theta)\}\mathbb{E}_{\tilde{\theta}} X - \{B(\tilde{\theta}) - B(\theta)\}. \end{aligned} \tag{2.33}$$

Comparison of the equalities (2.32) and (2.33) using the first item completes the proof.

Exercise 2.22 (Suhov and Kelbert 2005). *There is widespread agreement amongst the managers of the Reliable Motor Company that the number x of faulty cars produced in a month has a binomial distribution*

$$\mathbb{P}(x = s) = \binom{n}{s} p^s (1-p)^{n-s}, s = 0, 1, \ldots, n; 0 \le p \le 1.$$

There is, however, some dispute about the parameter p. The general manager has a prior distribution for p which is uniform (i.e. with the pdf $f_p(x) = \mathbf{1}(0 \le x \le 1)$), while the more pessimistic production manager has a prior distribution with density $f_p(x) = 2x\mathbf{1}(0 \le x \le 1)$. Both pdfs are concentrated on $(0, 1)$.

(i) In a particular month, s faulty cars are produced. Show that if the general manager's loss function is $(\hat{p} - p)^2$, where $\hat{p}$ is her estimate and p is the true value, then her best estimate of p is

$$\hat{p} = \frac{s+1}{n+2}$$

(ii) The production manager has responsibilities different from those of the general manager, and a different loss function given by $(1-p)(\hat{p} - p)^2$. Find his best estimator of p and show that it is greater than that of the general manager unless $s \ge n/2$.
You may assume that, for non-negative integers α, β,

$$\int_0^1 p^\alpha (1-p)^\beta \, dp \approx \frac{\alpha!\beta!}{(\alpha+\beta+1)!}$$

As $\mathbb{P}_p(X = s) = \alpha p^s (1-p)^{n-s}$, $s = 0, 1, \ldots, n$, the posterior for the general manager (GM) is

$$\pi^{GM}(p\, s) = \alpha p^s (1-p)^{n-s} \mathbf{1}(0 < p < 1),$$

and for the production manager (PM)

$$\pi^{PM}(p\, s) = \alpha p p^s (1-p)^{n-s} \mathbf{1}(0 < p < 1).$$

Then the expected loss for the GM is minimized at the posterior mean:

$$\begin{aligned} \hat{p}^{GM} &= \frac{\int_0^1 p p^s (1-p)^{n-s} \, dp}{\int_0^1 p^s (1-p)^{n-s} \, dp} \\ &= \frac{(s+1)!(n-s)!}{(n-s+s+2)!} \frac{(n-s+s+1)!}{s!(n-s)!} = \frac{s+1}{n+2}. \end{aligned}$$

For the PM, the expected loss

$$\int_0^1 (1-p)\,(p-a)^2\, \pi^{PM}(p\, s)\, dp$$

is minimized at

$$a = \frac{\int_0^1 p(1-p)\pi^{PM}(p\,s)\,dp}{\int_0^1 (1-p)\pi^{PM}(p\,s)\,dp},$$

which yields

$$\begin{aligned}\hat{p}^{PM} &= \frac{\int_0^1 p(1-p)pp^s(1-p)^{n-s}\,dp}{\int_0^1 p(1-p)pp^s(1-p)^{n-s}\,dp} \\ &= \frac{(s+2)!(n-s+1)!}{(n-s+s+4)!}\frac{(n-s+s+3)!}{(s+1)!(n-s+1)!} = \frac{s+2}{n+4}.\end{aligned}$$

We see that $(s+2)/(n+4) > (s+1)/(n+2)$, i.e., $s < n/2$.

Exercise 2.23. *Denote the number of incoming telecom signals between* $[0,t]$ *as* $C(0,t)$. *Assume that* $C(0,t)$ *satisfies*

(a) The number of arrivals in disjoint time intervals are independent;
(b) The distribution of $C(s,t)$ *depends on* $t-s$;
(c) For $h>0$ *small,* $P\{C(0,h)=1\} = \lambda h + o(h)$, *where* $\lambda > 0$ *is a constant;*
(d) $P\{C(0,h) \geq 2\} = o(h)$.

Please answer the following questions:

1. *Prove that* $C(0,t)$ *follows a Poisson distribution with mean* λt.
2. *Find the function* $p(y)$, $C(\theta)$ *and* $B(\theta)$ *of the natural parametrization*

$$p(y,\theta) \stackrel{\text{def}}{=} p(y)e^{yC(\theta)-B(\theta)}$$

and function $d(\theta)$ *of the canonical parametrization*

$$p(y,\theta) \stackrel{\text{def}}{=} e^{y\theta-d(\theta)}$$

for this Poisson distribution with mean λt.
3. *Find an estimator for constant* λ.

1. Let $X_m^n = C\{(m-1)t/n, mt/n\}$, $1 \leq m \leq n$, X_m^n are i.i.d. by assumption (a). Define Y_m^n be i.i.d. Bernoulli random variable such that $Y_m^n = 1$ with probability $/n$, $1 \leq m \leq n$. Define

$$S_n = X_1^n + \ldots + X_n^n$$

and

$$T_n = Y_1^n + \ldots + Y_n^n.$$

Suppose $\mathrm{P}\{C(0,h) = 1\} = \lambda h + g_1(h)$ and $\mathrm{P}\{C(0,h) \geq 2\} = g_2(h)$ where $g_1(h)$ and $g_2(h)$ are of order $o(h)$. We claim the following lemma:

Lemma 2.1. *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be complex numbers with modulus $\leq c$, then*

$$\left| \prod_{m=1}^{n} a_m - \prod_{m=1}^{n} b_m \right| \leq c^{n-1} \sum_{m=1}^{n} |a_m - b_m|.$$

The proof of this simple lemma is left to the reader (hint: use induction). The modulus of $\varphi_Y(\xi) = \exp(\mathbf{i}Y_m^n\xi)$ and $\varphi_{X_m^n}(\xi) = \exp(\mathbf{i}X_m^n\xi)$ are less than 1, $|\varphi_{X_m^n}(\xi) - \varphi_Y(\xi)| \leq 2g_1(t/n) + 2g_2(t/n)$ (verify!). By the lemma,

$$\begin{aligned}
&|\mathbb{E}\exp(\mathbf{i}T_n\xi) - \mathbb{E}\exp(\mathbf{i}S_n\xi)| \\
&\qquad = \left| \prod_{m=1}^{n} \varphi_{X_m^n}(\xi) - \prod_{m=1}^{n} \varphi_{Y_m}(\xi) \right| \\
&\qquad \leq \sum_{m=1}^{n} |\varphi_{X_m^n}(\xi) - \varphi_Y(\xi)| \\
&\qquad \leq \sum_{m=1}^{n} 2\left\{ \left| g_1\left(\frac{t}{n}\right) \right| + \left| g_2\left(\frac{t}{n}\right) \right| \right\} \\
&\qquad \to 0, \qquad \text{as } n \to \infty.
\end{aligned}$$

Now we show that $\mathbb{E}\exp(\mathbf{i}T_n\xi) \to \exp\{\lambda t(\exp(\mathbf{i}\xi) - 1)\}$, the characteristic function of the Poisson distribution with mean λt and finish the proof. Observe that $|\mathbb{E}\exp(\mathbf{i}Y_m^n\xi)| = (1 - \lambda t/n) + (\lambda t/n)\exp(\mathbf{i}\xi) = 1 + (\lambda t/n)\{\exp(\mathbf{i}\xi) - 1\}$ and $|\exp(\mathbf{i}\xi) - 1| \leq 2$. When n large, $\lambda t/n \leq 1/2$. Using the lemma again,

$$\begin{aligned}
&\left| \exp(\lambda t\{\exp(\mathbf{i}\xi) - 1\}) - \prod_{m=1}^{n} [1 + (\lambda t/n)\{\exp(\mathbf{i}\xi) - 1\}] \right| \\
&\qquad \leq \sum_{m=1}^{n} \left| \exp\left[\frac{\lambda t}{n}\{\exp(\mathbf{i}\xi) - 1\} \right] - \left[1 + \frac{\lambda t}{n}\{\exp(\mathbf{i}\xi) - 1\} \right] \right| \\
&\qquad \leq \sum_{m=1}^{n} \left(\frac{\lambda t}{n} \right)^2 |\exp(\mathbf{i}\xi) - 1|^2 \\
&\qquad \leq 4\left(\frac{\lambda t}{n} \right) \lambda t \\
&\qquad \to 0,
\end{aligned}$$

as $n \to \infty$. This finishes the proof.

2. The Poisson density with mean λt is

$$p(y, \lambda t) = \exp(-\lambda t)(\lambda t)^y / y!.$$

The $p(y)$, $C(\lambda)$, $B(\lambda)$ of the natural parametrization is

$$\begin{aligned} p(y) &= \frac{1}{y!}; \\ C(\lambda t) &= \log(\lambda t); \\ B(\lambda t) &= \lambda t. \end{aligned}$$

The $d(\lambda t)$ for the canonical parametrization is

$$d(\lambda t) = -\lambda t(y+1) + y \log(\lambda t) - \log y!.$$

3. Suppose we have an observation of the number of signal y between time 0 and t. The maximizer for the log natural parametrization is $\hat{\lambda} = y/t$.

Exercise 2.24. *Let $\boldsymbol{Y}$ be an i.i.d. sample from $P_{\boldsymbol{\theta}^*} \in (P_{\boldsymbol{\theta}})$, where $(P_{\boldsymbol{\theta}})$ is a regular parametric family. The fundamental exponential bound for the maximum likelihood is given by the fact that for any $0 < \varrho < 1$, $0 < s < 1$, $\mu > 0$, the log-likelihood process $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ fulfills for a fixed constant $\mathfrak{Q}(\varrho, s)$*

$$\mathbb{E} \exp\Big[\varrho \sup_{\boldsymbol{\theta} \in \Theta} \big\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + s\, \mathcal{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\big\}\Big] \le \mathfrak{Q}(\varrho, s), \tag{2.34}$$

see Spokoiny and Dickhaus (2014). Denote the set $\mathcal{A}\left(\mathfrak{z}, \boldsymbol{\theta}^*\right) = \{\boldsymbol{\theta} : \mathcal{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \mathfrak{z}\}$, where $\mathfrak{z}$ is positive, and $\mathcal{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the rate function defined for $\mu > 0$ by

$$\mathcal{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} -\log \mathbb{E}_{\boldsymbol{\theta}^*} \exp\left\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\right\}.$$

Using (2.34), prove that for any $\varrho' < \varrho$,

1.

$$\mathbb{E}\left[\exp\Big\{\varrho' s\, \mathcal{M}(\mu, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\Big\} \mathbf{1}\Big\{\tilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}^*)\Big\}\right] \le \mathfrak{Q}(\varrho, s) \exp\big\{-(\varrho - \varrho')s\mathfrak{z}\big\};$$

in particular,

$$\mathbb{P}\big\{\tilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}^*)\big\} \le \mathfrak{Q}(\varrho, s) \exp\big(-\varrho s \mathfrak{z}\big).$$

2.

$$\mathbb{E}\Big[\mathcal{M}(\mu, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \mathbf{1}\big(\tilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}^*)\big)\Big] \le \frac{1}{\varrho' s} \mathfrak{Q}(\varrho, s) \exp\big\{-(\varrho - \varrho')s\mathfrak{z}\big\}.$$

1. The inequalities $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq 0$ and $\mathcal{M}(\mu, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > \mathfrak{z}$ for $\tilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}^*)$ imply

$$\begin{aligned}
&\mathbb{E}\Big[\exp\{(\varrho - \varrho')s\mathfrak{z}\}\ \exp\Big\{\varrho' s\, \mathcal{M}(\mu, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\Big\}\ \mathbf{1}\Big\{\tilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}^*)\Big\}\Big] \\
&\qquad \leq \mathbb{E}\Big[\exp\Big\{\varrho s\, \mathcal{M}(\mu, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\Big\}\mathbf{1}\Big\{\tilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}^*)\Big\}\Big] \\
&\qquad \leq \mathbb{E}\Big[\exp\Big\{\varrho s\, \mathcal{M}(\mu, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\Big\}\Big] \\
&\qquad \leq \mathbb{E}\Big[\exp\Big\{\varrho\mu L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) + \varrho s\, \mathcal{M}(\mu, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\Big\}\Big] \\
&\qquad \leq \mathfrak{Q}(\varrho, s),
\end{aligned}$$

   and the assertion follows.
2. The second item directly follows from the first one, because $x < e^x$ for any positive x.

Exercise 2.25. *Consider a multivariate normal rv $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta}^*, \Sigma)$, where $\Sigma = (nD^2)^{-1}$ for some matrix D. In other words, $\boldsymbol{Y} = \boldsymbol{\theta}^* + \boldsymbol{\zeta}$ with $\boldsymbol{\zeta} \sim \mathcal{N}\{0, (nD^2)^{-1}\}$.*

1. Check that the log-likelihood ratio computed on one observation of $\boldsymbol{Y}$ is equal to

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D^2 \boldsymbol{\zeta} - n\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2. \tag{2.35}$$

2. Prove that the r.v. $\boldsymbol{\zeta}$ is equal to

$$\boldsymbol{\zeta} = (nD^2)^{-1} \nabla L(\boldsymbol{\theta}^*).$$

1. The log-likelihood is equal to

$$\begin{aligned}
&L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \\
&= L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \\
&= -\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\theta}^*)^\top \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}^*) \\
&= -\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* - \boldsymbol{\theta}) \\
&\qquad\qquad + \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\theta}^*)^\top \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}^*) \\
&= -\big(\boldsymbol{\theta}^* - \boldsymbol{\theta}\big)^\top \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}^*) - \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta}).
\end{aligned}$$

   To conclude the proof, it is sufficient to note that

$$\Sigma^{-1}\big(\boldsymbol{Y} - \boldsymbol{\theta}^*\big) = nD^2\boldsymbol{\zeta},$$

and

$$\frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top nD^2(\boldsymbol{\theta}^* - \boldsymbol{\theta}) = n\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2.$$

2. The proof is straightforward:

$$\nabla L(\boldsymbol{\theta}^*) = \nabla\Big\{-\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\theta}^*)^\top \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}^*)\Big\} = \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}^*) = nD^2\boldsymbol{\zeta}.$$

Exercise 2.26. *Consider the model from the previous exercise, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta}^*, \Sigma)$ with $\Sigma = (nD^2)^{-1}$ for some matrix D.*

Using the formula (2.35), simulate the log-likelihood ratio for $D^2 = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$, $\boldsymbol{\theta}^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^2$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\boldsymbol{\theta} = \boldsymbol{\theta}_2 \stackrel{\text{def}}{=} \begin{pmatrix} 1.2 \\ 1 \end{pmatrix}$.

Draw a plot for $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ as a function of $\boldsymbol{\zeta}$ and a plot for an estimator of the density function of $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

Define $\mu \stackrel{\text{def}}{=} -n\big\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\big\|^2/2$, and note that $nD^2\boldsymbol{\zeta} \sim \mathcal{N}(0, nD^2)$. Therefore, by formula (2.35), the rv $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ has the distribution

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \sim \mathcal{N}\{\mu, (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top (nD^2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}.$$

The square root of D^2 can be found via the Jordan decomposition $D^2 = \Gamma\Lambda\Gamma^\top$, where Γ is the eigenvector matrix and Λ is the diagonal matrix of eigenvalues of D^2. In our case, the diagonal entries of the matrix Λ are $\lambda_1 = (5 + \sqrt{5})/2$ and $\lambda_2 = (5 - \sqrt{5})/2$.

Figure 2.5 describes the simulation of the r.v. $\boldsymbol{\zeta}$ for $n = 1{,}000$, $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_2$.

Exercise 2.27 (Shao 2005). *Let $(X_1, \ldots, X_n)$ be a random sample from a distribution on $\mathbb{R}$ with the Lebesgue density $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$, where $f(x) > 0$ is a known Lebesgue density and $f'(x)$ exists for all $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, and $\sigma > 0$. Let $\boldsymbol{\theta} = (\mu, \sigma)$. Show that the Fisher information about $\boldsymbol{\theta}$ contained in $X_1, \ldots, X_n$ is*

$$I(\boldsymbol{\theta}) = \frac{n}{\sigma^2}\begin{pmatrix} \int \frac{\{f'(x)\}^2}{f(x)}\,dx & \int \frac{f'(x)\{xf'(x)+f(x)\}}{f(x)}\,dx \\ \int \frac{f'(x)\{xf'(x)+f(x)\}}{f(x)}\,dx & \int \frac{\{xf'(x)+f(x)\}^2}{f(x)}\,dx \end{pmatrix},$$

assuming that all integrals are finite.

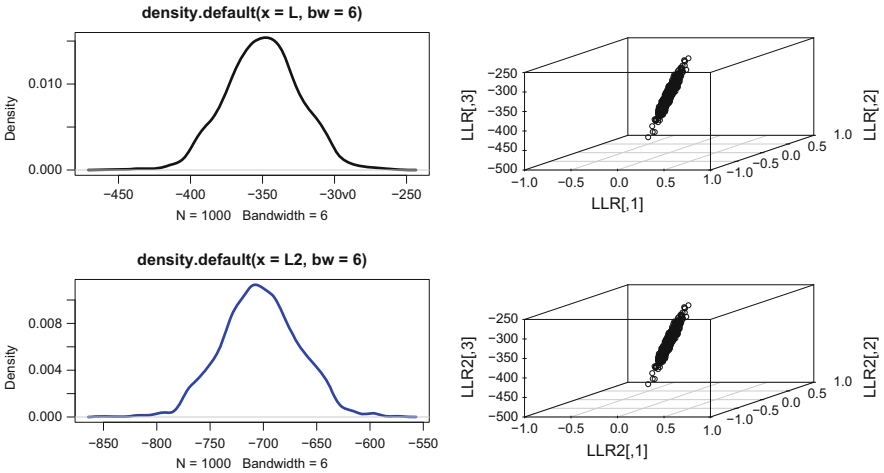

Fig. 2.5 Plots of density estimator and log-likelihood ratio function. MSEloglikelihood

Denote $g(\mu, \sigma, x) \stackrel{\text{def}}{=} \log \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$. Then

$$\frac{\partial}{\partial \mu} g(\mu, \sigma, x) = -\frac{f'\left(\frac{x-\mu}{\sigma}\right)}{\sigma f\left(\frac{x-\mu}{\sigma}\right)}$$

$$\frac{\partial}{\partial \sigma} g(\mu, \sigma, x) = -\frac{(x-\mu) f'\left(\frac{x-\mu}{\sigma^2}\right)}{\sigma f\left(\frac{x-\mu}{\sigma}\right)} - \frac{1}{\sigma}.$$

By the direct computation,

$$\begin{aligned}
\mathbb{E}\left\{\frac{\partial}{\partial \mu} g(\mu, \sigma, X_1)\right\}^2 &= \frac{1}{\sigma^2} \int \left\{\frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)}\right\}^2 \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx \\
&= \frac{1}{\sigma^2} \int \frac{\{f'\left(\frac{x-\mu}{\sigma}\right)\}^2}{f\left(\frac{x-\mu}{\sigma}\right)} d\left(\frac{x-\mu}{\sigma}\right) \\
&= \frac{1}{\sigma^2} \int \frac{\{f'(x)\}^2}{f(x)} dx,
\end{aligned}$$

$$\mathbb{E}\left\{\frac{\partial}{\partial \sigma} g(\mu, \sigma, X_1)\right\}^2 = \frac{1}{\sigma^2} \int \left\{\frac{x-\mu}{\sigma} \frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)} + 1\right\}^2 \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx$$

$$= \frac{1}{\sigma^2} \int \left\{ x \frac{f'(x)}{f(x)} + 1 \right\}^2 f(x)\, dx$$
$$= \frac{1}{\sigma^2} \int \frac{\{x f'(x) + f(x)\}^2}{f(x)}\, dx,$$

and

$$\mathbb{E}\left\{ \frac{\partial}{\partial \mu} g(\mu, \sigma, x) \frac{\partial}{\partial \sigma} g(\mu, \sigma, x) \right\}$$
$$= \frac{1}{\sigma^2} \int \frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)} \left\{ \frac{x-\mu}{\sigma} \frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)} + 1 \right\} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx$$
$$= \frac{1}{\sigma^2} \int \frac{f'(x)\{x f'(x) + f(x)\}}{f(x)}\, dx.$$

The result follows since

$$I(\boldsymbol{\theta}) = n\mathbb{E}\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log \frac{1}{\sigma} f\left(\frac{X_1 - \mu}{\sigma}\right) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log \frac{1}{\sigma} f\left(\frac{X_1 - \mu}{\sigma}\right) \right\}^\top.$$

Exercise 2.28 (Shao 2005). *Let X be a random variable having a cumulative distribution function F. Show that if $\mathbb{E}X$ exists, then*

$$\mathbb{E}X = \int_0^\infty \{1 - F(x)\}\, dx - \int_{-\infty}^0 F(x)\, dx.$$

By Fubini's theorem,

$$\int_0^\infty \{1 - F(x)\}\, dx = \int_0^\infty \int_{(x,\infty)} dF(y) dx$$
$$= \int_0^\infty \int_{(0,y)} dx dF(y)$$
$$= \int_0^\infty y\, dF(y).$$

Similarly,

$$\int_{-\infty}^0 F(x)\, dx = \int_{-\infty}^0 \int_{(-\infty,x]} dF(y) dx = -\int_{-\infty}^0 y\, dF(y).$$

If $\mathbb{E}X$ exists, then at least one of $\int_0^\infty y\, dF(y)$ and $\int_{-\infty}^0 y\, dF(y)$ is finite and

$$\mathbb{E}X = \int_{-\infty}^{\infty} yF(y) = \int_0^\infty \{1 - F(x)\}\, dx - \int_{-\infty}^0 F(x)\, dx.$$

Exercise 2.29 (Shao 2005). *Let $(X_1, \ldots, X_n)$ be a random sample from the exponential distribution on (a, ∞) with scale parameter 1, where $a \in \mathbb{R}$ is unknown.*

1. *Construct $(1 - \alpha)$ – confidence interval for a using the cumulative distribution function of the smallest order statistic $X_{(1)}$.*
2. *Show that the confidence interval in (i) can also be obtained using a pivotal quantity.*

1. The cumulative distribution function of $X_{(1)}$ is

$$F_a(t) = \begin{cases} 0 & t \le a \\ 1 - \exp^{-n(t-a)} & t > a\,, \end{cases}$$

which is decreasing in a for fixed $t > a$. A $(1 - \alpha)$ – confidence interval for a has upper limit equal to the unique solution of $F_a(T) = \alpha_1$ and lower limit equal to the unique solution of $F_a(T) = 1 - \alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$. Then, $[T + n^{-1}\log(\alpha_2), T + n^{-1}\log(1 - \alpha_1)]$ is the resulting confidence interval.
2. Note that $W(a) = n(X_{(1)} - a)$ has the exponential distribution on (0,1) with scale parameter 1. Therefore the distribution of $W(a)$ doesn't depend on the parameter and, hence, $W(a)$ it is a pivotal quantity. The $1 - \alpha$ confidence interval for a constructed this random variable is the same as that derived in item (i).

Exercise 2.30 (Shao 2005). *Let F_n be the edf based on a random sample of size n from cdf F on $\mathbb{R}$ having Lebesgue density f. Let $\varphi_n(t)$ be the Lebesgue density of the pth sample quantile $F_n^{-1}(p)$.*

Denote by m_p the integer part of np. Introduce also the quantity ℓ_p, which is equal to m_p if np is an integer and is equal to $m_p + 1$ if np is not an integer. Prove that

$$\varphi_n(t) = n \binom{n-1}{l_p - 1} \{F(t)\}^{l_p - 1} \{1 - F(t)\}^{n - l_p} f(t),$$

1. *Using the fact that $nF_n(t)$ has a binomial distribution;*
2. *Using the Lebesque density of the j-th order statistic.*

1. Since $nF_n(t)$ has the binomial distribution with size n and probability $F(t)$, for any $t \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}\{F_n^{-1}(p) \le t\} &= \mathbb{P}\{F_n(t) \ge p\} \\ &= \sum_{i=l_p}^{n} \binom{n}{i} \{F(t)\}^i \{1 - F(t)\}^{n-i}. \end{aligned}$$

Differentiating term by term leads to

$$\begin{aligned}
\varphi_n(t) &= \sum_{i=l_p}^{n} \binom{n}{i} i\{F(t)\}^{i-1}\{1-F(t)\}^{n-i} f(t) \\
&\quad - \sum_{i=l_p}^{n} \binom{n}{i} (n-i)\{F(t)\}^{i}\{1-F(t)\}^{n-i-1} f(t) \\
&= \binom{n}{l_p} l_p\{F(t)\}^{l_p-1}\{1-F(t)\}^{n-l_p} f(t) \\
&\quad + n \sum_{i=l_p+1}^{n} \binom{n-1}{i-1} \{F(t)\}^{i-1}\{1-F(t)\}^{n-i} f(t) \\
&\quad - n \sum_{i=l_p}^{n-1} \binom{n-1}{i} \{F(t)\}^{i}\{1-F(t)\}^{n-i-1} f(t) \\
&= n \binom{n-1}{l_p-1} \{F(t)\}^{l_p-1}\{1-F(t)\}^{n-l_p} f(t).
\end{aligned}$$

2. The Lebesgue density of the j-th order statistic is

$$n \binom{n-1}{j-1} \{F(t)\}^{j-1}\{1-F(t)\}^{n-j} f(t).$$

Then, the result follows from the fact that

$$F_n^{-1}(p) = \begin{cases} X_{(m_p)} & \text{if } np \text{ is an integer,} \\ X_{(m_p+1)} & \text{if } np \text{ is not an integer.} \end{cases}$$

Exercise 2.31. *Consider samples $\{Y_i\}_{i=1}^n$, where Y_i are i.i.d. with distribution function $F_Y(y)$. We want to estimate the τth quantile of the distribution function $F_Y^{-1}(\tau)$:*

$$F_Y^{-1}(\tau) \stackrel{\text{def}}{=} \inf\{y \in \mathbb{R} : \tau \le F_Y(y)\}.$$

This problem can be seen as in a location model:

$$Y_i = \theta^* + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim ALD(\tau),$$

where $F_{\boldsymbol{\varepsilon}}^{-1}(\tau) = 0$ and $\boldsymbol{\varepsilon}_i$'s are i.i.d. The QMLE estimation follows the framework with ALD likelihood, where ALD stands for "Asymmetric Laplace Distribution", and has probability density function

$$f(u|\tau) = \tau(1-\tau)exp\left\{-\rho_\tau(u)\right\},$$

with $\rho_\tau(u) = u\{\tau \mathbf{1}(u \geq 0) - (1-\tau)\mathbf{1}(u < 0)\}$.

1. *Prove that*

$$\underset{\theta}{\operatorname{argmin}}\, \mathbb{E}\rho_\tau(Y_i - \theta) = F_Y^{-1}(\tau) = \theta^*. \tag{2.37}$$

2. *Please write the empirical loss function for the estimation of $F_Y^{-1}(\tau)$.*

1. To prove (2.37),

$$\begin{aligned}
&\frac{\partial \mathbb{E}\rho_\tau(Y_i - \theta)}{\partial \theta} \\
&= \frac{\partial \int \{\tau(Y_i - \theta)\mathbf{1}(Y_i - \theta > 0) dF_Y(u)\} - (1-\tau)\int \{(Y_i - \theta)\mathbf{1}(Y_i - \theta \leq 0)\} dF_Y(u)}{\partial \theta} \\
&= -\tau\theta f_Y(\theta) - \tau\{1 - F_Y(\theta)\} + \tau\theta f(\theta) - (1-\tau) f_Y(\theta) + (1-\tau)(F_Y(\theta) + \theta f_Y(\theta)) \\
&= (1-\tau)F_Y(\theta) - \tau\{1 - F_Y(\theta)\} \\
&= F_Y(\theta) - \tau
\end{aligned}$$

Solve

$$\frac{\partial \mathbb{E}\rho_\tau(Y_i - \theta)}{\partial \theta} = 0,$$

we get

$$F(\theta^*) = \tau.$$

Thus, $\theta^* = F_Y^{-1}(\tau)$.

2. An estimator of θ^* would be

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{\tau \mathbf{1}\left(Y_i > \theta\right) - (1-\tau)\mathbf{1}\left(Y_i < \theta\right)\right\}.$$

Exercise 2.32. *Consider samples $\{(X_i, Y_i)\}_{i=1}^n$ i.i.d., in a regression framework, we now want to estimate the conditional τth quantile of the conditional distribution function $F^{-1}_{Y|X}(\tau)$. If we believe in the following linear model:*

$$Y_i = X_i^\top \theta^* + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim ALD(\tau),$$

where $F^{-1}_{\boldsymbol{\varepsilon}|X}(\tau) = 0$ and $\boldsymbol{\varepsilon}_i$s are i.i.d. Similarly we take a QMLE in an ALD likelihood.

1. Prove that

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}_{Y|X} \rho_\tau (Y_i - X_i^\top \theta) \tag{2.38}$$

$$F^{-1}_{Y|X_i}(\tau) = X_i^\top \theta^* \tag{2.40}$$

2. Suppose now $\{(X_i, Y_i)\}_{i=1}^n$ is a bivariate i.i.d. sequence from a joint normal distribution $N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Please write down the theoretical form of $F^{-1}_{Y|X}(\tau)$. (Hint: Observe that the conditional distribution is again normally distributed, with $\mu_{Y|X=x} = \mu_1 + \sigma_{12}\sigma_{22}^{-1}(x - \mu_2)$ and $\sigma_{Y|X} = \sigma_{11} - \sigma_{12}^2/\sigma_{22}$.)

1. To prove (2.40),

$$\begin{aligned}
&\frac{\partial \mathbb{E}\rho_\tau(Y_i - X_i^\top \theta)}{\partial \theta_j} \\
&= \frac{\partial \int \{\tau(Y_i - X_i^\top \theta)\mathbf{1}(Y_i - X_i^\top \theta > 0) dF_{Y|X}(u)\}}{\partial \theta_j} \\
&- \frac{(1-\tau)\int \{(Y_i - X_i^\top \theta)\mathbf{1}(Y_i - X_i^\top \theta \le 0)\} dF_{Y|X}(u)}{\partial \theta_j} \\
&= -\tau X_{ij} X_i^\top \theta f_Y(X_i^\top \theta) - X_{ij}\tau\{1 - F_Y(X_i^\top \theta)\} + X_{ij}\tau X_i^\top \theta f(X_i^\top \theta) \\
&-(1-\tau)X_{ij} f_Y(X_i^\top \theta) + (1-\tau)X_{ij}(F_Y(X_i^\top \theta) + X_i^\top \theta f_Y(X_i^\top \theta)) \\
&= (1-\tau)X_{ij} F_{Y|X}(X_i^\top \theta) - \tau X_{ij}\{1 - F_{Y|X}(X^\top \theta)\} \\
&= X_{ij} F_{Y|X}(X^\top \theta) - \tau X_{ij}
\end{aligned}$$

Solve

$$\frac{\partial \mathbb{E}_{Y|X}\rho_\tau(Y_i - X_i^\top \theta)}{\partial \theta_j} = 0, \forall\, j \in 1, \ldots, d$$

we get

$$F_{Y|X}(X_i^\top \theta^*) = \tau, \forall i, 1, \ldots, n$$

Thus, $F^{-1}_{Y|X_i}(\tau) = X_i^\top \theta^*$.

2. Use the hint, we have the normal conditional distribution. Given $X = x$, $(Y_i - u_{Y|X=x})/\sigma_{Y|X} \sim \mathbf{N}(0, 1)$. Denote $\Phi^{-1}(\tau)$ as the τth quantile of a standard normal distribution. Then we have,

$$F^{-1}_{Y|X=x}(\tau) = \sigma_{Y|X}\Phi^{-1}(\tau) + u_{Y|X=x}$$

References

Shao, J. (2005). *Mathematical statistics: Exercises and solutions*. New York: Springer

Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.

Suhov, Y., & Kelbert, M. (2005). *Probability and statistics by example, 1 basic probability and statistics*. New York: Cambridge University Press.

http://www.springer.com/978-3-642-36849-3

Basics of Modern Mathematical Statistics
Exercises and Solutions

Härdle, W.K.; Spokoiny, V.; Panov, V.; Wang, W.
2014, XXV, 185 p. 123 illus., 81 illus. in color.,
Hardcover
ISBN: 978-3-642-36849-3