Probability and Statistics

Exercise sheet 12

Exercise 12.1 The goal of this exercise is to show that if $X = (X_1, ..., X_n)^T \sim \mathcal{N}(\mu, \Sigma)$ with Σ invertible, then X admits a density with respect to the Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, given by

$$f(x) = f_X(x) = \frac{1}{(\sqrt{2\pi})^n} \frac{1}{\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{1}$$

for any $x = (x_1, ..., x_n)^T \in \mathbb{R}^n$.

Before showing this, we first settle some questions around the covariance matrix Σ (this is done in the first two parts). In (a) and (b), the random vector X can have any distribution (not necessarily normal).

(a) Recall that the covariance matrix of X, Σ, has entries $\Sigma_{ij} = \text{cov}(X_i, X_j)$ for $1 \leq i, j \leq n$. Show that

$$\Sigma = E[(X - \mu)(X - \mu)^T].$$

Remark: Expectations are evaluated componentwise, i.e. if M is a random matrix,

$$E\left[\begin{pmatrix} M_{11} & \dots & M_{1n} \\ \vdots & \ddots & \vdots \\ M_{n1} & \dots & M_{nn} \end{pmatrix}\right] = \begin{pmatrix} E(M_{11}) & \dots & E(M_{1n}) \\ \vdots & \ddots & \vdots \\ E(M_{n1}) & \dots & E(M_{nn}) \end{pmatrix}.$$

(b) Let $A \in \mathbb{R}^{p \times n}$ be a fixed (deterministic) matrix. Show that the covariance matrix of AX is $A\Sigma A^T$.

If $A = a^T \in \mathbb{R}^{1 \times n}$, what is the covariance of $a^T X$? Conclude that Σ is semi-positive definite.

(c) Now take $X \sim \mathcal{N}(\mu, \Sigma)$. By definition, $X \overset{\text{d}}{=} \mu + AZ$ with $AA^T = \Sigma$ (i.e., A is a square root of Σ), and Z is standard normal, i.e. $Z = (Z_1, ..., Z_n)^T$ for $Z_1, ..., Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

- Check that Σ is indeed the covariance matrix of X.
- Assuming that Σ is invertible, show that A is also invertible. Using the Jacobian formula, show that X has density given by (1) almost everywhere.

(d) Suppose you are given a density in the form (1). Can you find the marginal density of X_i ($i \in \{1, ..., n\}$) without additional calculations?

(e) (optional).

For $d = 2$, if $\sigma_1^2 = \text{var}(X_1) > 0$, $\sigma_2^2 = \text{var}(X_2) > 0$ and $\text{cov}(X_1, X_2) = \sigma_1\sigma_2\rho$ with ρ the correlation between X_1 and X_2.What is the condition on ρ for Σ to be invertible? What is the expression of the density in this case?

Solution 12.1

(a) Put $M := (X-\mu)(X-\mu)^T$. Then, by definition

$$M_{ij} = (X_i - \mu_i)(X_j - \mu_j) = (X_i - E(X_i))(X_j - E(X_j))$$

and

$$E(M_{ij}) = E[(X_i - E(X_i))(X_j - E(X_j))] = \operatorname{cov}(X_i, X_j)$$

for $(i,j) \in \{1, ..., n\}^2$.

Thus, $\Sigma = E[(X-\mu)(X-\mu)^T]$.

(b) The covariance of AX, $\tilde{\Sigma}$, is given by

$$\tilde{\Sigma} := E[(AX - A\mu)(AX - A\mu)^T]$$

since $E(AX) = AE(X) = A\mu$.

Thus,

$$\begin{aligned}\tilde{\Sigma} &= E[A(X-\mu)(X-\mu)^T A^T] \\ &= AE[(X-\mu)(X-\mu)^T]A^T \\ &= A\Sigma A^T.\end{aligned}$$

For $A = a^T$, $AX = a^T X \in \mathbb{R}$ and the covariance boils down to $\operatorname{var}(a^T X)$. Since this covariance is also equal to $a^T \Sigma a$, we conclude that $a^T \Sigma a = \operatorname{var}(a^T X)$. Now, $\operatorname{var}(a^T X) \geq 0 \ \forall a \in \mathbb{R}^n$ implying that Σ is positive semidefinite.

(c)
- We have shown that the covariance is given by

  $$E[(X-\mu)(X-\mu)^T].$$

  Since $X \overset{\mathrm{d}}{=} \mu + AZ$ and μ is a deterministic vector, this implies that

  $$X - \mu \overset{\mathrm{d}}{=} AZ$$

  and hence $X - \mu$ and AZ have the same moments (when these exist).

  Therefore,

  $$E[(X-\mu)(X-\mu)^T] = E[(AZ)(AZ)^T] = A\Sigma_Z A^T$$

  where

  $$\Sigma_Z = E(ZZ^T) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix} = I_{n\times n}$$

  is the identity matrix.

  Hence, the covariance of X is $AA^T = \Sigma$.
- We have

  $$\Sigma = P^T \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix} P$$

  where P is orthogonal ($P^T P = PP^T = I_{n\times n}$). Since Σ is invertible (so that $\det(\Sigma) = \prod_{i=1}^n \lambda_i \neq 0$), and since Σ is semi-positive definite, so that $\lambda_i \geq 0$ for each i, putting these together we get that $\lambda_i > 0$ for each $i = 1, ..., n$.

We have seen that we can take

$$A = P^T \begin{pmatrix} \sqrt{\lambda}_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda}_n \end{pmatrix}$$

as a square root for Σ. A is clearly invertible since

$$B = \begin{pmatrix} \frac{1}{\sqrt{\lambda}_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sqrt{\lambda}_n} \end{pmatrix} P$$

satisfies $AB = I_{n\times n}$.
Letting $g(z) = \mu + Az$ for $z = (z_1, ..., z_n)^T \in \mathbb{R}^n$. Then $g \in C^1(\mathbb{R}^n)$ with $\nabla g(z) = A$ (∇g is an equivalent notation for $\operatorname{grad} g$), and

$$\mathcal{J}_g(z) = \det(\nabla g(z)) = \det(A) \neq 0$$

for any $z \in \mathbb{R}^n$. Also, $g^{-1}(x) = A^{-1}(x - \mu)$.
By the Jacobian theorem, we have

$$f_X(x) = \frac{f_Z \circ g^{-1}(x)}{|\mathcal{J}_g \circ g^{-1}(x)|}$$

(on the open set $\mathcal{O} = \mathbb{R}^n$) with

$$\begin{aligned} f_Z(z) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^n z_i^2\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2} z^T z\right). \end{aligned}$$

It follows that

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{\exp\left(-\frac{1}{2}(x-\mu)^T (A^{-1})^T A^{-1}(x-\mu)\right)}{\det(A)}$$

(note $\det(A) = \prod_{i=1}^n \sqrt{\lambda_i} > 0$).
Now, $(A^{-1})^T A^{-1} = (AA^T)^{-1} = \Sigma^{-1}$ and $\det(AA^T) = (\det(A))^2 = \det(\Sigma)$, so that $\det(A) = \sqrt{\det(\Sigma)}$, yielding the formula in (1).

(d) If we are given a density in the form (1), this means that $X \sim \mathcal{N}(\mu, \Sigma)$. By the general characterisation of Gaussian vectors, this also mean that any linear combination of the components of X is normally distributed, and in particular so are the components themselves.

Thus, for any $i \in \{1, ..., n\}$, $X_i \sim \mathcal{N}(E(X_i), \operatorname{var}(X_i))$. But $E(X_i) = \mu_i$ and $\operatorname{var}(X_i) = \Sigma_{ii}$. Hence, $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$ for $i \in \{1, ..., n\}$.

(e) In this case,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Σ being invertible is equivalent to

$$\det(\Sigma) \neq 0 \Leftrightarrow \sigma_1^2\sigma_2^2(1-\rho^2) \neq 0 \Leftrightarrow |\rho| < 1$$

(note that we must always have $|\rho| \leq 1$).
To find Σ^{-1}, recall that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-cb}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Let $\mu = (\mu_1, \mu_2)^T$ be the mean. For $x = (x_1, x_2)^T \in \mathbb{R}^2$, we compute

$$\begin{aligned}
(x-\mu)^T\Sigma^{-1}(x-\mu) &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}(x_1-\mu_1, x_2-\mu_2)\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}(x_1-\mu_1, x_2-\mu_2)\begin{pmatrix} \sigma_2^2(x_1-\mu_1) - \rho\sigma_1\sigma_2(x_2-\mu_2) \\ \sigma_2^2(x_2-\mu_2) - \rho\sigma_1\sigma_2(x_1-\mu_1) \end{pmatrix} \\
&= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_2^2(x_1-\mu_1)^2 - 2\rho\sigma_1\sigma_2(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2) \\
&= \frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho}{\sigma_1\sigma_2}(x_1-\mu_1)(x_2-\mu_2) + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)
\end{aligned}$$

Finally,

$$f_X(x) = \frac{1}{2\pi}\frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho}{\sigma_1\sigma_2}(x_1-\mu_1)(x_2-\mu_2) + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right).$$

Exercise 12.2 (some training) Let $X_1, ..., X_n$ be i.i.d with density $f(\cdot \mid \theta_0)$, where the true value of θ_0 is unknown.

(a) For the following models, find the moment estimator and MLE for $\theta_0 \in \Theta$ as well as the Fisher information $I(\theta_0)$ (you may assume that all regularity conditions are fulfilled).

1. (Geometric)

$$f(x \mid \theta) = (1-\theta)^{x-1}\theta$$

for $x \in \mathbb{N}_{\geq 1}$,where $\theta \in \Theta = (0,1)$.

2. (Bernoulli)

$$f(x \mid \theta) = \theta^x(1-\theta)^{1-x}$$

for $x \in \{0,1\}$, where $\theta \in \Theta = (0,1)$.

3. (Beta$(1,\theta)$)

$$f(x \mid \theta) = \theta(1-x)^{\theta-1}\mathbb{1}_{x\in(0,1)},$$

where $\theta \in \Theta = (0, +\infty)$.

4. (Laplace)

$$f(x \mid \theta) = \frac{\theta}{2} e^{-\theta|x|}$$

for $x \in \mathbb{R}$, where $\theta \in \Theta = (0, +\infty)$.
Hint: Note that for $X \sim \text{Laplace}(\theta)$, $E(X) = 0$ and therefore one needs to use the next order moment.

(b) For the first model $\text{Geo}(\theta)$, construct an asymptotic confidence interval of level $1-\alpha$ for θ_0, based on the asymptotic normality of the MLE $\hat{\theta}$, and approximating $I(\theta_0)$ by $I(\hat{\theta})$.

(c) In a study of feeding behaviors of birds, the number of hops between flights was counted for $n = 130$ birds. The data are given in the following table.

# Hops	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	48	31	20	9	6	5	4	2	1	1	2	1

For example: in 48 occasions, a bird had just 1 hop before flying again, in 20 occasions they had 3 hops, etc. Assume that the number of hops can be modelled as a geometric random variable with unknown success probability $\theta_0 \in (0, 1)$. Compute the MLE based on the data in the table, and find an asymptotic confidence interval of level 95%.

Solution 12.2

(a) 1. Let $X \sim \text{Geo}(\theta_0)$, for some unknown $\theta_0 \in (0, 1)$.

$$E(X) = \frac{1}{\theta_0} \Leftrightarrow \theta_0 = \frac{1}{E(X)}.$$

Approximating $E(X)$ by $\overline{X}_n$ (which we can justify by the strong law of large numbers), we get the moment estimator $\tilde{\theta}_n = \frac{1}{\overline{X}_n}$ for θ_0.
For the MLE, we maximise as usual the log-likelihood.

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta) = \prod_{i=1}^{n} \theta(1-\theta)^{X_i - 1} = \theta^n (1-\theta)^{\sum_{i=1}^{n}(X_i - 1)}$$

$$l(\theta) = \log(L(\theta)) = n \log(\theta) + \sum_{i=1}^{n} (X_i - 1) \log(1-\theta)$$

Differentiating,

$$\begin{aligned}
\frac{\partial l}{\partial \theta}(\hat{\theta}_n) = 0 &\Leftrightarrow \frac{n}{\hat{\theta}_n} - \sum_{i=1}^{n} \frac{X_i - 1}{1 - \hat{\theta}_n} = 0 \\
&\Leftrightarrow n(1 - \hat{\theta}_n) = \hat{\theta}_n \sum_{i=1}^{n} (X_i - 1) \\
&\Leftrightarrow n = n\hat{\theta}_n + \hat{\theta}_n \sum_{i=1}^{n} X_i - n\hat{\theta}_n \\
&\Leftrightarrow \hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}_n}
\end{aligned}$$

is the unique stationary point. We check maximality by taking the second derivative:

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = -\frac{n}{\theta^2} - \sum_{i=1}^{n} \frac{X_i - 1}{(1-\theta)^2} < 0$$

Since the second derivative is negative, l is strictly concave on $(0,1)$ and so $\hat{\theta}_n = \frac{1}{\overline{X}_n}$ is the MLE. In this case it coincides with the moment estimator $\tilde{\theta}_n$.
For the Fisher information, note that

$$\begin{aligned}
\log f(x \mid \theta) &= \log(\theta) + (x-1)\log(1-\theta) \\
\frac{\partial \log f(x \mid \theta)}{\partial \theta} &= \frac{1}{\theta} - \frac{x-1}{1-\theta} \\
\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} &= -\frac{1}{\theta^2} - \frac{x-1}{(1-\theta)^2} \\
\Leftrightarrow I(\theta_0) &= -E\left[\frac{\partial^2 \log f(X_1 \mid \theta)}{\partial \theta^2} \mid_{\theta=\theta_0}\right] \\
&= \frac{1}{\theta_0^2} + \frac{E(X_1) - 1}{(1-\theta_0)^2} \\
&= \frac{1}{\theta_0^2} + \frac{\frac{1}{\theta_0} - 1}{(1-\theta_0)^2} \\
&= \frac{1}{\theta_0^2} + \frac{1}{\theta_0(1-\theta_0)} \\
&= \frac{1}{\theta_0^2(1-\theta_0)}.
\end{aligned}$$

2. If $X \sim \text{Bernoulli}(\theta_0)$, then $E(X) = \theta_0$. Thus we get the moment estimator $\tilde{\theta}_n = \overline{X}_n$ for θ_0.

$$\begin{aligned}
L(\theta) &= \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i} \\
l(\theta) &= \left(\sum_{i=1}^n X_i\right)\log(\theta) + \log(1-\theta)\left(n - \sum_{i=1}^n X_i\right).
\end{aligned}$$

So we get:

$$\begin{aligned}
\frac{\partial l}{\partial \theta}(\hat{\theta}_n) = 0 &\Leftrightarrow \frac{1}{\hat{\theta}_n}\sum_{i=1}^n X_i - \frac{1}{1-\hat{\theta}_n}\left(n - \sum_{i=1}^n X_i\right) = 0 \\
&\Leftrightarrow (1-\hat{\theta}_n)\sum_{i=1}^n X_i = \hat{\theta}_n\left(n - \sum_{i=1}^n X_i\right) \\
&\Leftrightarrow \hat{\theta}_n = \frac{\sum_{i=1}^n X_i}{n} = \overline{X}_n
\end{aligned}$$

as the unique stationary point. We check that

$$\frac{\partial^2 l}{\partial \theta^2}(\theta) = -\frac{1}{\theta^2}\sum_{i=1}^n X_i - \frac{1}{(1-\theta)^2}\left(n - \sum_{i=1}^n X_i\right) < 0$$

for any $X_1, ..., X_n \in \{0,1\}$.

Since l is strictly concave, $\hat{\theta}_n (= \tilde{\theta}_n)$ is the MLE.
For the Fisher information,

$$\begin{aligned}
\log f(x \mid \theta) &= x \log(\theta) + (1-x)\log(1-\theta) \\
\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \\
\Leftrightarrow I(\theta_0) &= -E\left[\frac{\partial^2 \log f(X_1 \mid \theta)}{\partial \theta^2} \mid_{\theta=\theta_0}\right] \\
&= \frac{E(X_1)}{\theta_0^2} + \frac{1-E(X_1)}{(1-\theta_0)^2} \\
&= \frac{\theta_0}{\theta_0^2} + \frac{1-\theta_0}{(1-\theta_0)^2} \\
&= \frac{1}{\theta_0} + \frac{1}{1-\theta_0} \\
&= \frac{1}{\theta_0(1-\theta_0)}.
\end{aligned}$$

3. If $X \sim \text{Beta}(1, \theta_0)$, then

$$E(X) = \frac{1}{1+\theta_0} \Leftrightarrow \theta_0 = \frac{1}{E(X)} - 1$$

and therefore, $\tilde{\theta}_n = \frac{1}{\overline{X}_n} - 1$ is the moment estimator for θ_0.

$$L(\theta) = \prod_{i=1}^{n} \theta(1-X_i)^{\theta-1} = \theta^n \left(\prod_{i=1}^{n}(1-X_i)\right)^{\theta-1}$$

and

$$l(\theta) = n\log(\theta) + (\theta-1)\sum_{i=1}^{n}\log(1-X_i),$$

so we maximise at

$$\begin{aligned}
\frac{\partial l}{\partial \theta}(\hat{\theta}_n) &= \frac{n}{\hat{\theta}_n} + \sum_{i=1}^{n}\log(1-X_i) = 0 \\
\Rightarrow \hat{\theta}_n &= -\frac{n}{\sum_{i=1}^{n}\log(1-X_i)}.
\end{aligned}$$

l is clearly concave (as the sum of a strictly concave and linear functions). Hence $\hat{\theta}_n$ is the MLE.
For the Fisher information,

$$\begin{aligned}
\log f(x \mid \theta) &= \log(\theta) + (\theta-1)\log(1-x) \\
\frac{\partial \log f(x \mid \theta)}{\partial \theta} &= \frac{1}{\theta} + \log(1-x) \\
\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} &= -\frac{1}{\theta^2} \\
\Leftrightarrow I(\theta_0) &= -E\left[\frac{\partial^2 \log f(X_1 \mid \theta)}{\partial \theta^2} \mid_{\theta=\theta_0}\right] = \frac{1}{\theta_0^2}.
\end{aligned}$$

4.

$$\begin{aligned}
E(X^2) &= \frac{\theta_0}{2} \int_{\mathbb{R}} x^2 e^{-\theta_0 |x|} dx \\
&= \theta_0 \int_0^{\infty} x^2 e^{-\theta_0 x} dx \\
&= \theta_0 \frac{\Gamma(3)}{\theta_0^3} \int_0^{\infty} \frac{\theta_0^3}{\Gamma(3)} x^{3-1} e^{-\theta_0 x} dx \\
&= \frac{\Gamma(3)}{\theta_0^2} = \frac{2}{\theta_0^2}.
\end{aligned}$$

(noting that we integrate the density of a $\mathrm{G}(3, \theta_0)$ distribution). Alternatively, we could observe that

$$\int_0^{\infty} \theta_0 x^2 e^{-\theta_0 x} dx = E[Y^2] = E(Y)^2 + \mathrm{var}(Y) = \frac{2}{\theta_0^2}$$

for $Y \sim \mathrm{Exp}(\theta_0)$.

Therefore, $\theta_0^2 = \sqrt{\frac{2}{E(X^2)}}$. We can replace $E(X^2)$ by $\frac{1}{n}\sum_{i=1}^n X_i^2$ to obtain the moment estimator

$$\tilde{\theta}_n = \sqrt{\frac{2}{\frac{1}{n}\sum_{i=1}^n X_i^2}}.$$

$$L(\theta) = \prod_{i=1}^n \frac{\theta}{2} e^{-\theta |X_i|} = \frac{1}{2^n} \theta^n e^{-\theta \sum_{i=1}^n |X_i|}$$

$$l(\theta) = c + n \log(\theta) - \theta \sum_{i=1}^n |X_i|$$

for $c = -n\log(2)$ a constant. Therefore,

$$\begin{aligned}
\frac{\partial l}{\partial \theta}(\hat{\theta}_n) &= \frac{n}{\hat{\theta}_n} - \sum_{i=1}^n |X_i| = 0 \\
\Rightarrow \hat{\theta}_n &= \frac{1}{\frac{1}{n}\sum_{i=1}^n |X_i|}.
\end{aligned}$$

Since the function l is strictly concave, we conclude that $\hat{\theta}_n$ is the MLE.
For the Fisher information,

$$\begin{aligned}
\log f(x \mid \theta) &= -\log(2) + \log(\theta) - \theta |x| \\
\frac{\partial \log f(x \mid \theta)}{\partial \theta} &= \frac{1}{\theta} - |x| \\
\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} &= -\frac{1}{\theta^2} \\
\Rightarrow I(\theta_0) &= \frac{1}{\theta_0^2}.
\end{aligned}$$

(b) Assume that the geometric model satisfies the regularity conditions of Theorem 2 from the lecture. Then, the MLE for θ_0 based on $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathrm{Geo}(\theta_0)$, for some $\theta_0 \in (0, 1)$, is asymptotically normal with

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$$
$$\sqrt{n}(\hat{\theta}_n - \theta_0)\sqrt{I(\theta_0)} \xrightarrow{\text{d}} \mathcal{N}(0,1)$$

with $\sqrt{I(\theta_0)} = \sqrt{\frac{1}{\theta_0^2(1-\theta_0)}}$.

Replacing θ_0 by $\hat{\theta}_n$ results in

$$\sqrt{n}(\hat{\theta}_n - \theta_0)\sqrt{\frac{1}{\hat{\theta}_n^2(1-\hat{\theta}_n)}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

For $\alpha \in (0,1)$, let $z_{1-\frac{\alpha}{2}}$ be the $\left(1-\frac{\alpha}{2}\right)$-quantile of $Z \sim \mathcal{N}(0,1)$. Then,

$$P\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\frac{1}{\hat{\theta}_n\sqrt{1-\hat{\theta}_n}} \in (-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]\right) \overset{n\to\infty}{\to} 1-\alpha$$
$$\Leftrightarrow P(\theta_0 \in I_\alpha) \overset{n\to\infty}{\to} 1-\alpha$$

where

$$\begin{aligned} I_\alpha &= \left[\hat{\theta}_n - \frac{\hat{\theta}_n\sqrt{1-\hat{\theta}_n}}{\sqrt{n}}z_{1-\frac{\alpha}{2}}, \hat{\theta}_n + \frac{\hat{\theta}_n\sqrt{1-\hat{\theta}_n}}{\sqrt{n}}z_{1-\frac{\alpha}{2}}\right) \\ &= \left[\frac{1}{\overline{X}_n} - \frac{1}{\overline{X}_n}\sqrt{1-\frac{1}{\overline{X}_n}}\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \frac{1}{\overline{X}_n} + \frac{1}{\overline{X}_n}\sqrt{1-\frac{1}{\overline{X}_n}}\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right). \end{aligned}$$

(c) With $n = 130$,

$$\sum_{i=1}^{n} X_i = 1 \times 48 + 2 \times 31 + 3 \times 20 + ... + 12 \times 1 = 363$$

and so $\overline{X}_n = 2.792$. Using $\alpha = 0.05$, $z_{1-\frac{\alpha}{2}} = z_{0.975} \approx 1.964$ and we get the confidence interval

$$P(\theta_0 \in [0.308, 0.407]) \approx 0.95.$$

Exercise 12.3

(a) Find a sufficient statistic for the parameters generating the following models:

1. $$X_1, ..., X_n \overset{\text{iid}}{\sim} \text{U}([0,\theta]), \quad \theta \in (0,+\infty).$$

2. $$X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda), \quad \lambda \in (0,+\infty).$$

3. $$X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu,\sigma^2), \quad \theta = (\mu,\sigma)^T \in \mathbb{R} \times (0,+\infty).$$

4. $$X_1, ..., X_n \overset{\text{iid}}{\sim} \text{U}([\theta,\theta+1]), \quad \theta \in \mathbb{R}.$$

(b) Show that in general, if $T(X_1, ..., X_n)$ is a sufficient statistic for $\theta \in \Theta$ (where $X_1, ..., X_n \overset{\text{iid}}{\sim} f(\cdot \mid \theta)$), then for any $c \in \mathbb{R} \setminus \{0\}$, $cT(X_1, ..., X_n)$ is also sufficient for θ.

Hint: Use the factorisation theorem.

Solution 12.3

(a) We use the factorisation theorem.

1. $f(x \mid \theta) = \frac{1}{\theta}\mathbb{1}_{x\in[0,\theta]}$, so

$$\begin{aligned}\prod_{i=1}^n f(x_i \mid \theta) &= \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{x_i\in[0,\theta]}\\ &= \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{x_i\geq 0}\mathbb{1}_{x_i\leq\theta}\\ &= \frac{1}{\theta^n}\mathbb{1}_{\min_i x_i\geq 0}\mathbb{1}_{\max_i x_i\leq\theta}\\ &= g(T(x_1, ..., x_n), \theta)h(x_1, ..., x_n)\end{aligned}$$

with $T(x_1, ..., x_n) = \max_{1\leq i\leq n} x_i$, $g(t, \theta) = \frac{1}{\theta^n}\mathbb{1}_{t\leq\theta}$ and $h(x_1, ..., x_n) = \mathbb{1}_{\min_{1\leq i\leq n} x_i\geq 0}$. Hence, $T(X_1, ..., X_n) = \max_{1\leq i\leq n} X_i$ is sufficient for θ.

2.

$$\begin{aligned}\prod_{i=1}^n f(x_i \mid \lambda) &= \prod_{i=1}^n \lambda e^{-\lambda x_i}\mathbb{1}_{x_i>0}\\ &= \lambda^n e^{-\lambda\sum_{i=1}^n x_i}\mathbb{1}_{\min_i x_i\geq 0}\\ &= g(T(x_1, ..., x_n), \lambda)h(x_1, ..., x_n)\end{aligned}$$

with $g(t, \lambda) = \lambda^n e^{-\lambda t}$, $h(x_1, ..., x_n) = \mathbb{1}_{\min_{1\leq i\leq n} x_i\geq 0}$ and $T(x_1, ..., x_n) = \sum_{i=1}^n x_i$. Therefore, $T(X_1, ..., X_n) = \sum_{i=1}^n X_i$ is sufficient for λ.

3.

$$\begin{aligned}\prod_{i=1}^n f(x_i \mid \theta) &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}\\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2+\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i-\frac{n\mu^2}{2\sigma^2}}\\ &= g(T(x_1, ..., x_n), \theta)h(x_1, ..., x_n)\end{aligned}$$

with

$$g(t, \theta) = \frac{1}{(2\pi)^{\frac{n}{2}}\theta_2^n}e^{-\frac{1}{2\theta_2^2}t_2+\frac{\theta_1}{\theta_2^2}t_1-n\frac{\theta_1^2}{2\theta_2^2}},$$

$h(x_1, ..., x_n) = 1$, $T(x_1, ..., x_n) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)^T$. Thus $T(X_1, ..., X_n) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)^T$ is sufficient for $\theta = (\mu, \sigma)^T$.

4.

$$\begin{aligned}\prod_{i=1}^{n} f(x_i \mid \theta) &= \prod_{i=1}^{n} \mathbb{1}_{\theta \leq x_i \leq \theta+1} \\ &= \prod_{i=1}^{n} \mathbb{1}_{\theta \leq x_i} \mathbb{1}_{\theta \geq x_i - 1} \\ &= \mathbb{1}_{\theta \leq \min_i x_i} \mathbb{1}_{\theta \geq \max_i x_i - 1} \\ &= \mathbb{1}_{\max_i x_i - 1 \leq \theta \leq \min_i x_i} \\ &= g(T(x_1, ..., x_n), \theta) h(x_1, ..., x_n)\end{aligned}$$

with $g(t, \theta) = \mathbb{1}_{t_2 - 1 \leq \theta \leq t_1}$, $h(x_1, ..., x_n) = 1$ and $T(x_1, ..., x_n) = (\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i)^T$, and therefore $T(X_1, ..., X_n) = (\min_{1 \leq i \leq n} X_i, \max_{1 \leq i \leq n} X_i)^T$ is sufficient for θ.

(b) If $T(X_1, ..., X_n)$ is sufficient then

$$\prod_{i=1}^{n} f(x_i \mid \theta) = g(T(x_1, ..., x_n), \theta) h(x_1, ..., x_n)$$

for some measurable functions g and h. If we define $\tilde{g}(t, \theta) = g(\frac{t}{c}, \theta)$, it will follow that

$$\prod_{i=1}^{n} f(x_i \mid \theta) = \tilde{g}(\tilde{T}(x_1, ..., x_n), \theta) h(x_1, ..., x_n)$$

where $\tilde{T}(x_1, ..., x_n) = cT(x_1, ..., x_n)$.

This shows that $\tilde{T}(X_1, .., X_n) = cT(X_1, ..., X_n)$ is sufficient for θ. Replacing c by $\frac{1}{c}$ gives the equivalence.

Remark: This implies, for example, that if $\sum_{i=1}^{n} X_i$ is sufficient for θ, then so is $\overline{X}_n$.

Exercise 12.4 Let $(X, Y)^T$ be a random vector. We want to show that $\text{var}(X \mid Y) = 0$ with probability 1, if and only if there is a measurable function h such that $P(X = h(Y)) = 1$.

We consider only the case where the vector is discrete (takes either finitely many or countably many different values).

(a) State the definition of $\text{var}(X \mid Y = y)$.

(b) Show that $\text{var}(X \mid Y) = 0$ with probability 1 if and only if $P(X = E(X \mid Y)) = 1$.

(c) Conclude.

Solution 12.4

(a)

$$\text{var}(X \mid Y = y) := \sum_{x} (x - E(X \mid Y = y))^2 p(x \mid y)$$

for any y such that $p_Y(y) > 0$.

(b) Let $Z = X - E(X \mid Y)$. First, assume that $P(\text{var}(X \mid Y) = 0) = 1$, or in other words, $P(E(Z^2 \mid Y) = 0) = 1$. For any y with $p_Y(y) > 0$, we have that

$$0 = P(E(Z^2 \mid Y) \neq 0) \geq p_Y(y) \mathbb{1}_{E(Z^2 \mid Y = y) \neq 0}$$

and thus $E(Z^2 \mid Y = y) = 0$. Then we get:

$$\begin{aligned} E(Z^2) &= E(E(Z^2 \mid Y)) \\ &= \sum_{y:p_Y(y)>0} E(Z^2 \mid Y = y)p_Y(y) \\ &= 0. \end{aligned}$$

Since $Z^2 \geq 0$, this implies that $Z = 0$ almost surely, or

$$P(Z = 0) = 1 \Leftrightarrow P(X = E(X \mid Y)) = 1.$$

For the other direction, we start from $P(Z = 0) = 1$. It will be convenient to use the joint probability of Z and Y:

$$q_{Z,Y}(z, y) = P(Z = z, Y = y) = \sum_{x:x-E(X|Y=y)=z} p_{X,Y}(x, y) = p_{X,Y}(E(X \mid Y = y) + z, y).$$

Note that the resulting marginal pmf for Y is $q_Y(y) = P(Y = y) = p_Y(y)$.Then, for y with $p_Y(y) > 0$, and denoting by $q(z \mid y)$ the conditional pmf of Z given $Y = y$:

$$q(z \mid y) = \frac{q(z, y)}{p_Y(y)} \leq \frac{\sum_{y'} q(z, y')}{p_Y(y)} = \frac{p_Z(z)}{p_Y(y)} = 0$$

unless $z = 0$ (since $P(Z = 0) = 1$).

Hence,

$$E[Z^2 \mid Y = y] = \sum_z z^2 q(z \mid y) = 0$$

for any y with $p_Y(y) > 0$, since $z^2 q(z \mid y) = 0$ for any z.

Therefore,

$$\begin{aligned} P(\mathrm{var}(X \mid Y) = 0) &= P(E[Z^2 \mid Y] = 0) \\ &= \sum_{y:p_Y(y)>0} \mathbb{1}_{E[Z^2|Y=y]=0} p_Y(y) \\ &= \sum_{y:p_Y(y)>0} p_Y(y) \\ &= 1 \end{aligned}$$

as we wanted.

(c) We have shown that

$$P(X = E(X \mid Y)) = 1 \Leftrightarrow P(\mathrm{var}(X \mid Y) = 0) = 1.$$

Thus,

$$P(\mathrm{var}(X \mid Y) = 0) = 1 \Rightarrow P(X = \mu_X(Y)) = 1$$

where $\mu_X(y) = E(X \mid Y = y)$.

If there is a measurable function Ψ such that $P(X = \Psi(Y)) = 1$, then $P(X = E(X \mid Y)) \geq P(X = \Psi(Y))$ since we know that $X = \Psi(Y) \Rightarrow E(X \mid Y) = \Psi(Y)$. Therefore, $P(X = E(X \mid Y)) = 1$ which implies that $P(\mathrm{var}(X \mid Y) = 0) = 1$.

Exercise 12.5 (optional).

The goal here is to justify why the idea of maximising the likelihood is a good one.

(a) For $X \sim f(\cdot \mid \theta_0)$ and $\theta \in \Theta$, assume that $E[\log f(X \mid \theta)]$ exists.

Show that $E[\log f(X \mid \theta)] \leq E[\log f(X \mid \theta_0)]$.

Hint: Show that $E\left[\log\left(\frac{f(X|\theta_0)}{f(X|\theta)}\right)\right] \geq 0$ by using Jensen's inequality for the convex function $t \mapsto -\log t, t \in (0, +\infty)$.

(b) Recall the weak law of large numbers: if $Y_1, ..., Y_n$ are i.i.d. such that $E(|Y_1|) < \infty$, then

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{\mathbb{P}} E(Y_1) \quad (n \to \infty).$$

Using the WLLN, explain why the MLE would be a reasonable estimator.

Solution 12.5

(a) We have that

$$\begin{aligned} E\left[\log \frac{f(X \mid \theta_0)}{f(X \mid \theta)}\right] &= E\left[-\log \frac{f(X \mid \theta)}{f(X \mid \theta_0)}\right] \\ &\geq -\log\left(E\left[\frac{f(X \mid \theta)}{f(X \mid \theta_0)}\right]\right) \end{aligned}$$

by Jensen's inequality applied to the convex function $t \mapsto -\log(t)$, for $t \in (0, +\infty)$.

But since $X \sim f(\cdot \mid \theta_0)$,

$$\begin{aligned} E\left[\frac{f(X \mid \theta)}{f(X \mid \theta_0}\right] &= \int \frac{f(x \mid \theta)}{f(x \mid \theta_0)} f(x \mid \theta_0) d\mu(x) \\ &= \int f(x \mid \theta) d\mu(x) \\ &= 1 \end{aligned}$$

since we are integrating a density. Note that $-\log(1) = 0$, and thus, for $\theta \in \Theta$,

$$E[\log f(X \mid \theta_0)] \geq E[\log f(X \mid \theta)].$$

Therefore, the true parameter θ_0 maximises the function $\theta \mapsto E[\log f(X \mid \theta)]$.

(b) By the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta) \underset{n\to\infty}{\xrightarrow{\mathbb{P}}} E[\log f(X \mid \theta)].$$

Hence, by maximising the log-likelihood

$$l(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

over Θ, we are also maximising

$$\frac{1}{n}l(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta).$$

We "hope" (under some technical conditions) that as $n \to \infty$, we will manage to get closer to the maximal value of $E[\log f(X \mid \theta)]$, which is $E[\log f(X \mid \theta_0)]$ by part (a). This gives a heuristic argument for why the MLE should converge to θ_0, as $n \to \infty$.