Estimating the parameters of some probability distributions:

Exemplifications

Estimating the parameter of the Bernoulli distribution: the MLE and MAP approaches

CMU, 2015 spring, Tom Mitchell, Nina Balcan, HW2, pr. 2

Suppose we observe the values of n i.i.d. (independent, identically distributed) random variables $X_1, \ldots, X_n$ drawn from a single Bernoulli distribution with parameter θ. In other words, for each X_i, we know that

$$P(X_i = 1) = \theta \quad \textbf{and} \quad P(X_i = 0) = 1 - \theta.$$

Our *goal* is to estimate the value of θ from the observed values of $X_1, \ldots, X_n$.

Reminder: Maximum Likelihood Estimation

For any hypothetical value $\hat{\theta}$, we can compute the probability of observing the outcome $X_1, \ldots, X_n$ if the true parameter value θ were equal to $\hat{\theta}$.

This probability of the observed data is often called the *data likelihood*, and the function $L(\hat{\theta})$ that maps each $\hat{\theta}$ to the corresponding likelihood is called the *likelihood function*.

A natural way to estimate the unknown parameter θ is to choose the $\hat{\theta}$ that maximizes the likelihood function. Formally,

$$\hat{\theta}_{MLE} = \underset{\hat{\theta}}{\operatorname{argmax}}\, L(\hat{\theta}).$$

a. Write a formula for the likelihood function, $L(\hat{\theta})$.

Your function should depend on the random variables $X_1, \ldots, X_n$ and the hypothetical parameter $\hat{\theta}$.

Does the likelihood function depend on the order of the random variables?

Solution:

Since the X_i are independent, we have

$$\begin{aligned} L(\hat{\theta}) &= P_{\hat{\theta}}(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_{\hat{\theta}}(X_i) = \prod_{i=1}^{n} (\hat{\theta}^{X_i} \cdot (1-\hat{\theta})^{1-X_i}) \\ &= \hat{\theta}^{\#\{X_i=1\}} \cdot (1-\hat{\theta})^{\#\{X_i=0\}}, \end{aligned}$$

where $\#\{\cdot\}$ counts the number of X_i for which the condition in braces holds true.
Note that in the third equality we used the trick $X_i = \mathrm{I}_{\{X_i=1\}}$.

The likelihood function does not depend on the order of the data.

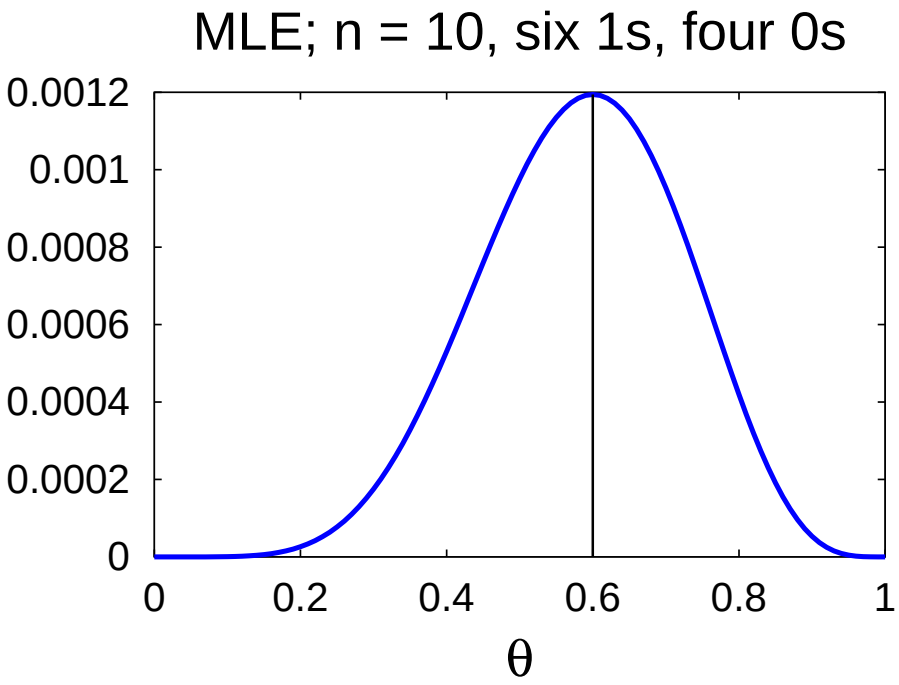
b. Suppose that $n = 10$ and the data set contains six 1s and four 0s.

Write a short computer program that plots the likelihood function of this data.

For the plot, the x-axis should be $\hat{\theta}$, and the y-axis $L(\hat{\theta})$. Scale your y-axis so that you can see some variation in its value.

Estimate $\hat{\theta}_{MLE}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the likelihood.

Solution:



c. Find a closed-form formula for $\hat{\theta}_{MLE}$, the MLE estimate of $\hat{\theta}$. Does the closed form agree with the plot?

Solution:

Let's consider $l(\theta) = \ln(L(\theta))$. Since the $\ln$ function is increasing, the $\hat{\theta}$ that maximizes the log-likelihood is the same as the θ that maximizes the likelihood. Using the properties of the $\ln$ function, we can rewrite $l(\hat{\theta})$ as follows:

$$l(\hat{\theta}) = \ln(\hat{\theta}^{n_1} \cdot (1-\hat{\theta})^{n_0}) = n_1 \ln(\hat{\theta}) + n_0 \ln(1-\hat{\theta}).$$

Assuming that $\hat{\theta} \neq 0$ and $\hat{\theta} \neq 1$, the first and second derivatives of l are given by

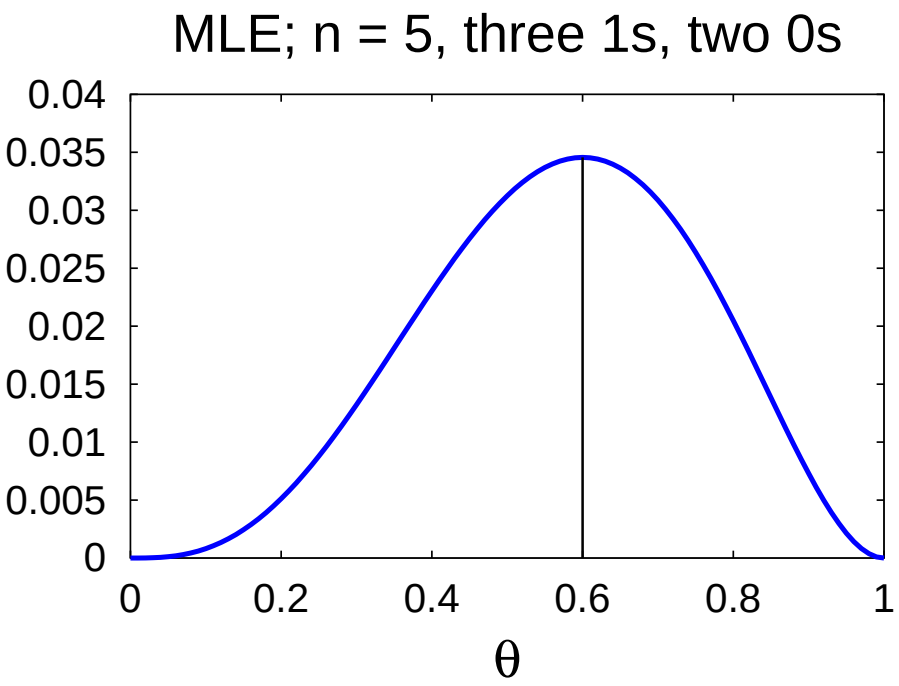
$$l'(\hat{\theta}) = \frac{n_1}{\hat{\theta}} - \frac{n_0}{1-\hat{\theta}} \quad \textbf{and} \quad l''(\hat{\theta}) = -\frac{n_1}{\hat{\theta}^2} - \frac{n_0}{(1-\hat{\theta})^2}$$

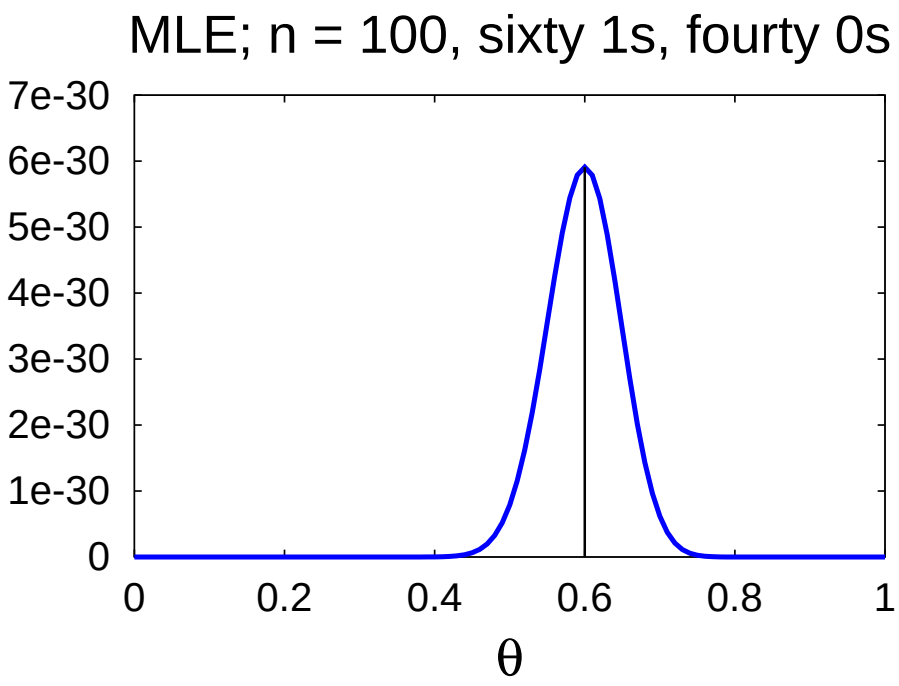
Since $l''(\hat{\theta})$ is always negative, the l function is concave, and we can find its maximizer by solving the equation $l'(\theta) = 0$.

The solution to this equation is given by $\hat{\theta}_{MLE} = \dfrac{n_1}{n_1 + n_0}$.

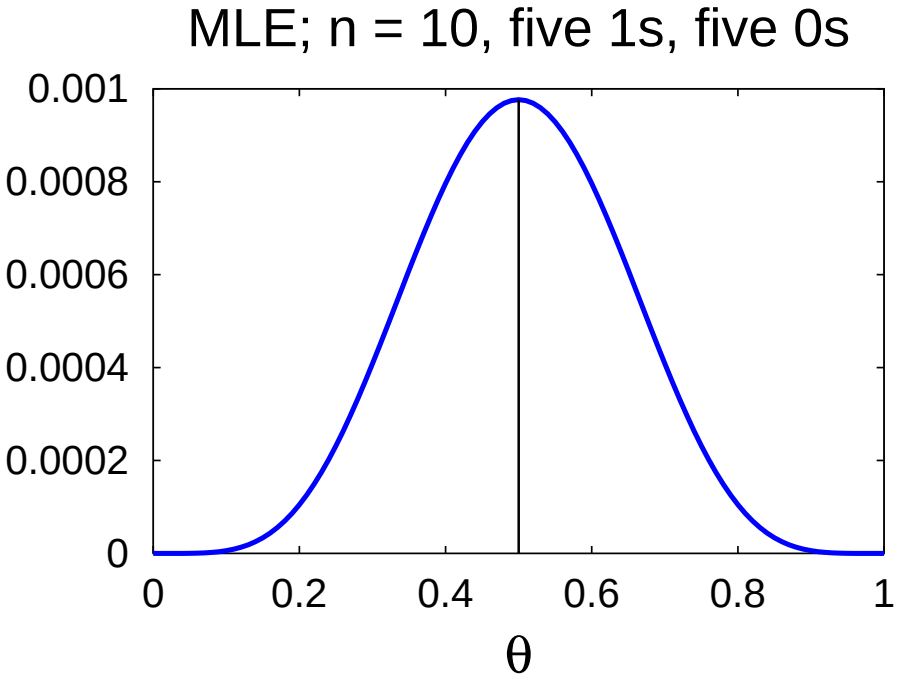
d. Create three more likelihood plots: one where $n = 5$ and the data set contains three 1s and two 0s; one where $n = 100$ and the data set contains sixty 1s and fourty 0s; and one where $n = 10$ and there are five 1s and five 0s.

Solution:





Solution (to part d.):



e. Describe how the likelihood functions and maximum likelihood estimates compare for the different data sets.

Solution (to part e.):

The MLE is equal to the proportion of 1s observed in the data, so for the first three plots the MLE is always at 0.6, while for the last plot it is at 0.5.

As the number of samples n increases, the likelihood function gets more peaked at its maximum value, and the values it takes on decrease.

Reminder: Maximum a Posteriori Probability Estimation

In the maximum likelihood estimate, we treated the true parameter value θ as a fixed (non-random) number. In cases where we have some prior knowledge about θ, it is useful to treat θ itself as a random variable, and express our prior knowledge in the form of a prior probability distribution over θ.

For *example*, suppose that the $X_1, \ldots, X_n$ are generated in the following way:

- **First, the value of θ is drawn from a given prior probability distribution**
- **Second, $X_1, \ldots, X_n$ are drawn independently from a Bernoulli distribution using this value for θ.**

Since both θ and the sequence $X_1, \ldots, X_n$ are random, they have a joint probability distribution. In this setting, a natural way to estimate the value of θ is to simply choose its most probable value given its prior distribution plus the observed data $X_1, \ldots, X_n$.

Definition:

$$\hat{\theta}_{MAP} = \underset{\hat{\theta}}{\operatorname{argmax}}\, P(\theta = \hat{\theta} | X_1, \ldots, X_n).$$

This is called the maximum a posteriori probability (MAP) estimate of θ.

Using Bayes rule, we can rewrite the posterior probability as follows:

$$P(\theta = \hat{\theta}|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta})}{P(X_1, \ldots, X_n)}.$$

Since the probability in the denominator does not depend on $\hat{\theta}$, the MAP estimate is given by

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\hat{\theta}}{\operatorname{argmax}}\, P(X_1, \ldots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta}) \\ &= \underset{\hat{\theta}}{\operatorname{argmax}}\, L(\hat{\theta})P(\theta = \hat{\theta}).\end{aligned}$$

In words, the MAP estimate for θ is the value $\hat{\theta}$ that maximizes the likelihood function multiplied by the prior distribution on θ.

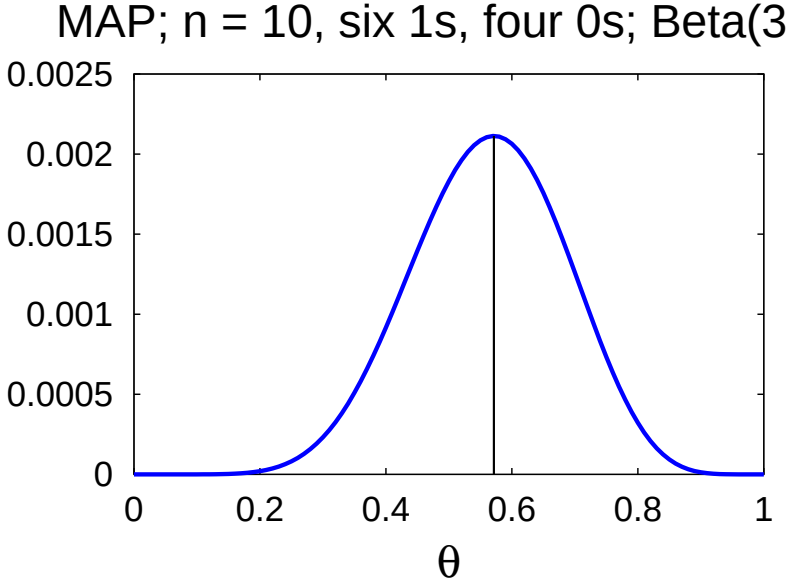
We will consider a $Beta(3,3)$ prior distribution for θ, which has the density function given by $p(\hat{\theta}) = \dfrac{\hat{\theta}^2(1-\hat{\theta})^2}{B(3,3)}$, where $B(\alpha,\beta)$ is the beta function and $B(3,3) \approx 0.0333$.

f. Suppose, as in part c, that $n = 10$ and we observed six 1s and four 0s.

Write a short computer program that plots the function $\hat{\theta} \mapsto L(\hat{\theta})p(\hat{\theta})$ for the same values of $\hat{\theta}$ as in part c.

Estimate $\hat{\theta}_{MAP}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the function.

Solution:



g. Find a closed form formula for $\hat{\theta}_{MAP}$, the MAP estimate of $\hat{\theta}$. Does the closed form agree with the plot?

Solution:

As in the case of the MLE, we will apply the $\ln$ function before finding the maximizer. We want to maximize the function

$$l(\hat{\theta}) = \ln(L(\hat{\theta}) \cdot p(\hat{\theta})) = \ln(\hat{\theta}^{n_1+2} \cdot (1-\hat{\theta})^{n_0+2}) - \ln(B(3,3)).$$

The normalizing constant for the prior appears as an additive constant and therefore the first and second derivatives are identical to those in the case of the MLE (except with $n_1 + 2$ and $n_0 + 2$ instead of n_1 and n_0, respectively).

It follows that the closed form formula for the MAP estimate is given by

$$\hat{\theta}_{MAP} = \frac{n_1 + 2}{n_1 + n_0 + 4}$$

h. Compare the MAP estimate to the MLE computed from the same data in part c. Briefly explain any significant difference.

Solution:

The MAP estimate is equal to the MLE with four additional virtual random variables, two that are equal to 1, and two that are equal to 0. This pulls the value of the MAP estimate closer to the value 0.5, which is why $\hat{\theta}_{MAP}$ is smaller than $\hat{\theta}_{MLE}$.

i. Comment on the relationship between the MAP and MLE estimates as n goes to infinity, while the ratio $\#\{X_i = 1\}/\#\{X_i = 0\}$ remains constant.

Solution:

As n goes to infinity, the influence of the 4 virtual random variables diminishes, and the two estimators become equal.

The MLE estimator for the parameter of the Bernoulli distribution: the bias and [an example of] inadmissibility

CMU, 2004 fall, Tom Mitchell, Ziv Bar-Joseph, HW2, pr. 1

Suppose X is a binary random variable that takes value 0 with probability p and value 1 with probability $1-p$. Let $X_1, \ldots, X_n$ be i.i.d. samples of X.

a. Compute an MLE estimate of p (denote it by $\hat{p}$).

Answer:

By way of definition,

$$\hat{p} = \operatorname*{argmax}_p P(X_1, \ldots, X_n | p) \overset{i.i.d.}{=} \operatorname*{argmax}_p P(X_i | p) = \operatorname*{argmax}_p p^k (1-p)^{n-k}$$

where k is the number of 0's in $x_1, \ldots, x_n$. Furthermore, since $\ln$ is a monotonic (strictly increasing) function,

$$\hat{p} = \operatorname*{argmax}_p \ln p^k (1-p)^{n-k} = \operatorname*{argmax}_p (k \ln p + (n-k) \ln(1-p))$$

Computing the first derivative of $k \ln p + (n-k)\ln(1-p)$ w.r.t. p leads to:

$$\frac{\partial}{\partial p}(k \ln p + (n-k)\ln(1-p)) = \frac{k}{p} - \frac{n-k}{1-p}.$$

Hence,

$$\frac{\partial}{\partial p}(k \ln p + (n-k)\ln(1-p)) = 0 \Leftrightarrow \frac{k}{p} = \frac{n-k}{1-p} \Leftrightarrow \hat{p} = \frac{k}{n}.$$

b. Is $\hat{p}$ an unbiased estimate of p? Prove the answer.

Answer:

$$\begin{aligned} E[\hat{p}] &= E\left[\frac{k}{n}\right] = \frac{1}{n}E[k] = \frac{1}{n}E[n - \sum_{i=1}^{n} X_i] = \frac{1}{n}(n - \sum_{i=1}^{n} E[X_i]) \\ &= \frac{1}{n}(n - \sum_{i=1}^{n}(1-p)) = \frac{1}{n}(n - n(1-p)) = \frac{1}{n}np = p. \end{aligned}$$

Therefore, $\hat{p}$ is an unbiased estimator for the parameter p.

c. Compute the expected square error of $\hat{p}$ in terms of p.

Answer:

$$\begin{aligned} E[(\hat{p}-p)^2] &= E[\hat{p}^2] - 2E[\hat{p}]\,p + p^2 = \frac{E[k^2]}{n^2} - 2p^2 + p^2 \\ &= \frac{Var\,(k) + E^2[k]}{n^2} - p^2 = \frac{np(1-p) + (np)^2}{n^2} - p^2 = \frac{p}{n}(1-p) \end{aligned}$$

We used the fact that $Var\,(k) = E[k^2] - E^2[k]$**, and also** $Var\,(k) = np(1-p)$ **because** k **is a binomial random variable.**

Note that $E[\hat{p}] = p$ **(cf. part** b**), therefore**

$$E[(\hat{p}-p)^2] = E[(\hat{p}-E[\hat{p}])^2] = Var\,[\hat{p}] = \frac{p}{n}(1-p).$$

This implies that $Var\,[\hat{p}] \to 0$ **for** $n \to \infty$**.**

d. Prove that if you know that p lies in the interval $[1/4; 3/4]$ and you are given only $n = 3$ samples of X, then $\hat{p}$ is an *inadmissible estimator* of p when minimizing the expected square error of estimation.

***Note*: An estimator δ of a parameter θ is said to be *inadmissible* when there exists a different estimator δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta, \delta') < R(\theta, \delta)$ for some θ, where $R(\theta, \delta)$ is a *risk function* and in this problem it is the expected square error of the estimator.**

Answer:

Consider another estimator, $\tilde{p} = 1/2$.
$E[(\tilde{p} - p)^2] = (1/2 - p)^2$.

For $p = 1/2$ we have $E[(\tilde{p} - p)^2] = 0 < E[(\hat{p} - p)^2] = 1/12$.
We now need to show that $E[(\tilde{p} - p)^2] \leq E[(\hat{p} - p)^2]$ over $p \in [1/4; 3/4]$.

$$E[(\tilde{p} - p)^2] - E[(\hat{p} - p)^2] = \left(\frac{1}{2} - p\right)^2 - \frac{1}{3} \cdot p(1-p) = \frac{1}{4} - \frac{4}{3}p + \frac{4}{3}p^2.$$

This is a parabola going up, so we need to show that it lies below or equal to zero for $p \in [1/4; 3/4]$.

It is equivalent to showing that it is below or equal to 0 at boundary points.
In fact it is: $\frac{1}{4} - \frac{4}{3}p + \frac{4}{3}p^2 = 0$ for both $p = 1/4$ and $p = 3/4$.

Estimating the parameters of the categorical distribution: the MLE approach

CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 2.3

In this problem we will derive the MLE for the parameters of a *categorical distribution* where the variable of interest, X, can take on k values, namely $a_1, a_2, \ldots, a_k$.

a. Given data describing n independent identically distributed *observations* of X, namely $d_1, \ldots, d_n$, each of which can be one of k values, express the *likelihood* of the data given $k-1$ parameters for the distribution over X. Let n_i represent the number of times X takes on value i in the data.

Answer:

Let the k values [of the considered categorial distribution] be $\{a_1, a_2, \ldots, a_k\}$. Since the probability of seeing an event of type j is θ_j — and we are given an ordered list of events and not an unordered bag of events —, the verosimility of the data is:

$$
\begin{aligned}
L(\theta) &= P(d_1, \ldots, d_n|\theta) \overset{i.i.d.}{=} \prod_{j=1}^{n} \sum_{i=1}^{k} (\theta_i I_{d_j=a_i}) && \text{(}I\text{ is the indicator function)} \\
&= \prod_{i=1}^{k} \theta_i^{n_i} && (n_i \overset{not.}{=} \textstyle\sum_{j=1}^{n} I_{d_j=a_i}) \\
&= \underbrace{(1 - \sum_{i=1}^{k-1} \theta_i)}_{\theta_k}{}^{n_k} \prod_{i=1}^{k-1} \theta_i^{n_i} && \text{(since the thetas sum to one)}
\end{aligned}
$$

b. Find the MLE for one of the $k-1$ parameters, θ_j, by setting the partial derivative of the likelihood in part a with respect to θ_j equal to zero and solving for it.

***Hint*: You may want to start by first taking the log of the likelihood from part a before taking its derivative.**

Answer:

$$\ln L(\theta) = n_k \ln(1 - \sum_{i=1}^{k-1} \theta_i) + \sum_{i=1}^{k-1} n_i \ln \theta_i \Rightarrow \frac{\partial \ln L(\theta)}{\partial \theta_j} = -\frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i} + \frac{n_j}{\theta_j}$$

$$\frac{\partial \ln L(\theta)}{\partial \theta_j} = 0 \Leftrightarrow -\frac{n_k}{1 - \hat{\theta}_j - \sum_{i \neq j,k}^{k-1} \theta_i} + \frac{n_j}{\hat{\theta}_j} = 0 \Leftrightarrow \frac{n_j}{\hat{\theta}_j} = \frac{n_k}{1 - \hat{\theta}_j - \sum_{i \neq j,k}^{k-1} \theta_i}$$

$$\Leftrightarrow n_j(1 - \sum_{i \neq j,k}^{k-1} \theta_i) = (n_k + n_j)\hat{\theta}_j$$

$$\Leftrightarrow \hat{\theta}_j = \frac{n_j}{n_j + n_k}(1 - \sum_{i \neq j,k}^{k-1} \theta_i) \text{ for all } j \in \{1, \ldots, k-1\}.$$

c. At this point you should have $k-1$ equations describing MLEs of different parameters. Show how those equations imply that the MLE for a parameter θ_j representing the probability that X takes on value j is equal to $\frac{n_j}{n}$.

***Hint*: In order to remove the k-th parameter from the likelihood in part a you had to represent it with an equation, $\theta_k = f(\ )$. At this point you may find it helpful to replace all occurrences of $f(\ )$ with θ_k. After replacing $f(\ )$ with θ_k you can substitute all occurrences of each other parameter in $f(\ )$ with its MLE from part b. This should allow you to solve for the MLE of θ_k, which can then be used to simplify all of the other equations.**

Answer

As the likelihood function is uniquely optimal for the vector θ, the last equation in part b can be written as:

$$\hat{\theta}_j = \frac{n_j}{n_j + n_k}\left(1 - \sum_{i \neq j,k}^{k-1} \hat{\theta}_i\right) \Leftrightarrow \hat{\theta}_j = \frac{n_j}{n_j + n_k}(\hat{\theta}_j + \hat{\theta}_k) \qquad (\textbf{because } \hat{\theta}_k = 1 - \sum_{i=1}^{k-1} \hat{\theta}_i)$$

$$\Leftrightarrow \hat{\theta}_j\left(1 - \frac{n_j}{n_j + n_k}\right) = \frac{n_j}{n_j + n_k}\hat{\theta}_k \Leftrightarrow \hat{\theta}_j \frac{n_k}{n_j + n_k} = \frac{n_j}{n_j + n_k}\hat{\theta}_k$$

$$\Leftrightarrow \hat{\theta}_j n_k = n_j \hat{\theta}_k$$

$$\Leftrightarrow \hat{\theta}_j = \frac{n_j}{n_k}\hat{\theta}_k \textbf{ for all } j \in \{1, \ldots, k-1\}.$$

Finally,

$$\hat{\theta}_k = 1 - \hat{\theta}_1 - \ldots - \hat{\theta}_{k-1} = 1 - \frac{n_1}{n_k}\hat{\theta}_k - \ldots - \frac{n_{k-1}}{n_k}\hat{\theta}_k$$

$$\Rightarrow n_k\hat{\theta}_k = n_k - (n_1 + \ldots + n_{k-1})\hat{\theta}_k$$

$$\Rightarrow \hat{\theta}_k(\underbrace{n_1 + \ldots + n_{k-1} + n_k}_{n}) = n_k$$

$$\Rightarrow \hat{\theta}_k = \frac{n_k}{n}$$

$$\Rightarrow \hat{\theta}_j = \frac{n_j}{n_k} \cdot \frac{n_k}{n} = \frac{n_j}{n} \textbf{ for all } j \in \{1, \ldots, k-1\}.$$

***Note*: Even though [here] we can go from the non-hatted to hatted form of the equation in the first step of c, this will generally not be possible. To solve for a maximum likelihood criterion under additional constraints like the thetas summing to one, a generic and useful method is the method of Lagrange multipliers.**

The Gaussian [uni-variate] distribution: estimating μ when σ^2 is known

CMU, 2011 fall, Tom Mitchell, Aarti Singh, HW2, pr. 1

CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 1.2-3

Assume we have n samples, $x_1, \ldots, x_n$, independently drawn from a normal distribution with *known* variance σ^2 and *unknown* mean μ.

a. Derive the MLE estimator for the mean μ.

Solution:

$$P(x_1, \ldots, x_n|\mu) = \prod_{i=1}^{n} P(x_i|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow \ln P(x_1, \ldots, x_n|\mu) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

$$\Rightarrow \frac{\partial}{\partial \mu} P(x_1, \ldots, x_n|\mu) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}$$

$$\frac{\partial}{\partial \mu} P(x_1, \ldots, x_n|\mu) = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Leftrightarrow \sum_{i=1}^{n} (x_i - \mu) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i = n\mu$$

$$\Rightarrow \mu_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

***Remark*: It can be easily shown that $\ln P(x_1, \ldots, x_n|\mu)$ indeed reaches its maximum for $\mu = \mu_{MLE}$.**

b. Show that $E[\mu_{MLE}] = \mu$.

Solution:

The sample $x_1, \ldots, x_n$ can be seen as the realization of n independent random variables $X_1, \ldots, X_n$ of Gaussian distribution of mean μ and variance σ^2. Then, due to the property of linearity for the expectation of random variables, we get:

$$E[\mu_{MLE}] = E\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{E[X_1] + \ldots + E[X_n]}{n} = \frac{n\mu}{n} = \mu$$

Therefore, the μ_{MLE} estimator is unbiased.

c. What is $Var[\mu_{MLE}]$?

Solution:

$$Var[\mu_{MLE}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \overset{(\star)}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \overset{i.i.d.}{=} n\frac{1}{n^2} Var[X_1] = \frac{\sigma^2}{n}$$

Therefore, $Var[\mu_{MLE}] \to 0$ as $n \to \infty$.

($\star$) **Remember that** $Var[aX] = a^2\, Var[X]$.

d. Now derive the MAP estimator for the mean μ. Assume that the prior distribution for the mean is itself a normal distribution with mean ν and variance β^2.

Solution 1:

$$P(\mu|x_1,\ldots,x_n) \overset{T.\ Bayes}{=} \frac{P(x_1,\ldots,x_n|\mu)\,P(\mu)}{P(x_1,\ldots,x_n)} \tag{1}$$

$$= \frac{\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\right)e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\;\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{(\mu-\nu)^2}{2\beta^2}}}{C} \tag{2}$$

where $C \overset{not.}{=} P(x_1,\ldots,x_n)$**.**

$$\Rightarrow \quad \ln P(\mu|x_1,\ldots,x_n) = -\sum_{i=1}^{n}\left(\ln\sqrt{2\pi}\sigma + \frac{(x_i-\mu)^2}{2\sigma^2}\right) - \ln\sqrt{2\pi}\beta - \frac{(\mu-\nu)^2}{2\beta^2} - \ln C$$

$$\Rightarrow \quad \frac{\partial}{\partial\mu}\ln P(\mu|x_1,\ldots,x_n) = \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2} - \frac{\mu-\nu}{\beta^2}$$

$$\frac{\partial}{\partial\mu}\ln P(\mu|x_1,\ldots,x_n) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2} = \frac{\mu-\nu}{\beta^2} \Leftrightarrow \mu\left(\frac{1}{\beta^2}+\frac{n}{\sigma^2}\right) = \frac{\sum_{i=1}^{n}x_i}{\sigma^2} + \frac{\nu}{\beta^2}$$

$$\Rightarrow \quad \mu_{MAP} = \frac{\sigma^2\nu + \beta^2\sum_{i=1}^{n}x_i}{\sigma^2+n\beta^2}$$

Solution 2:

Instead of computing the derivative of the posterior distribution $P(\mu|x_1,\ldots,x_n)$,
we will first show that the right hand side of (2) is itself a Gaussian, and then we will use the fact that the mean of a Gaussian is where it achieves its maximum value.

$$
\begin{aligned}
P(\mu|x_1,\ldots,x_n) &= \frac{1}{C}\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\right)e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\,\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= const\ \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= const\ \cdot e^{-\frac{\beta^2\sum_{i=1}^{n}(x_i-\mu)^2+\sigma^2(\mu-\nu)^2}{2\sigma^2\beta^2}} \\
&= const\ \cdot e^{-\frac{n\beta^2+\sigma^2}{2\sigma^2\beta^2}\mu^2+\frac{\beta^2\sum_{i=1}^{n}x_i+\nu\sigma^2}{\sigma^2\beta^2}\mu-\frac{\beta^2\sum_{i=1}^{n}x_i^2+\nu^2\sigma^2}{2\sigma^2\beta^2}}
\end{aligned}
$$

$$
\begin{aligned}
P(\mu|x_1,\ldots,x_n) &= \\
&= const \cdot \exp\left(-\frac{\mu^2 - 2\mu\frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2} + \frac{\beta^2\sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2+\sigma^2}}{\frac{2\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \\
&= const \cdot \exp\left(-\frac{(\mu - \frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2})^2 - \left(\frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2 + \frac{\beta^2\sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2+\sigma^2}}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \\
&= const \cdot \exp\left(-\frac{\left(\mu - \frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \cdot \exp\left(\frac{\left(\frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2 - \frac{\beta^2\sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2+\sigma^2}}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \\
&= const' \cdot \exp\left(-\frac{\left(\mu - \frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right)
\end{aligned}
$$

The exp **term in the last equality being a Gaussian of mean** $\frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2 + \sigma^2}$ **and variance** $\frac{\sigma^2\beta^2}{n\beta^2 + \sigma^2}$**, it follows that its maximum is obtained for** $\mu = \frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2 + \sigma^2} = \mu_{MAP}$**.**

e. Please comment on what happens to the MLE and MAP estimators for the mean μ as the number of samples n goes to infinity.

Solution:

$$\begin{aligned}
\mu_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\mu_{MAP} &= \frac{\sigma^2\nu + \beta^2 \sum_{i=1}^{n} x_i}{\sigma^2 + n\beta^2} = \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} + \frac{\beta^2 \sum_{i=1}^{n} x_i}{\sigma^2 + n\beta^2} \\
&= \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} + \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{1 + \frac{\sigma^2}{n\beta^2}} = \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} + \frac{\mu_{MLE}}{1 + \frac{\sigma^2}{n\beta^2}}
\end{aligned}$$

$$n \to \infty \Rightarrow \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} \to 0 \textbf{ and } \frac{\sigma^2}{n\beta^2} \to 0 \Rightarrow \mu_{MAP} \to \mu_{MLE}$$

The Gaussian [uni-variate] distribution: estimating σ^2 when $\mu = 0$

CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 2.1

Let X be a random variable distributed according to a Normal distribution with 0 mean, and σ^2 variance, i.e. $X \sim N(0, \sigma^2)$.

a. Find the maximum likelihood estimate for σ^2, i.e. σ^2_{MLE}.

Solution:

Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim N(0, \sigma^2)$. Let f be the density function corresponding to X. Then we can write the likelihood function as:

$$
\begin{aligned}
L(X_1, X_2, \ldots, X_n | \sigma^2) &= \prod_{i=1}^{n} f(X_i; \mu = 0, \sigma^2) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(X_i - 0)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n} X_i^2}{2\sigma^2}\right) \\
\Rightarrow \ln L &= \text{constant} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} X_i^2 \\
\Rightarrow \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} X_i^2. \text{ Therefore, } \frac{\partial \ln L}{\partial \sigma^2} = 0 \Leftrightarrow \sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i^2
\end{aligned}
$$

Note: **It can be easily shown that $L(X_1, X_2, \ldots, X_n | \sigma^2)$ indeed reaches its maximum for $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.**

b. Is the estimator you obtained biased?

Solution:

It is unbiased, since:

$$
\begin{aligned}
E[\tfrac{1}{n}\sum_{i=1}^{n} X_i^2] &= \frac{n}{n}E[X^2] && \text{since i.i.d.} \\
&= Var\,[X] + (E[X])^2 && \\
&= Var\,[X] = \sigma^2 && \text{since } E[X] = 0
\end{aligned}
$$

The Gaussian [uni-variate] distribution: estimating σ^2 (without restrictions on μ)

CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 2.1.1-2

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be observed i.i.d. samples from a Gaussian distribution $N(x|\mu, \sigma^2)$.

a. Derive σ^2_{MLE}, the MLE for σ^2.

Solution:

The p.d.f. for $N(x|\mu, \sigma^2)$ has the form $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

The log likelihood function of the data x is:

$$\begin{aligned} \ln \mathcal{L}(\mathbf{x} \mid \mu, \sigma^2) &= \ln \prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

The partial derivative of $\ln \mathcal{L}$ w.r.t. σ^2: $\frac{\partial \ln \mathcal{L}(\mathbf{x} \mid \mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$.

Solving the equation $\frac{\partial \ln \mathcal{L}(\mathbf{x} \mid \mu, \sigma^2)}{\partial \sigma^2} = 0$, **we get:** $\sigma^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{MLE})^2$.

Note that we had to take into account the optimal value of μ (see problem CMU, 2011 fall, T. Mitchell, A. Singh, HW2, pr. 1)

b. Show that $E[\sigma^2_{MLE}] = \dfrac{n-1}{n}\sigma^2$.

Solution:

$$
\begin{aligned}
E[\sigma^2_{MLE}] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2\right] = E[(x_1 - \mu_{MLE})^2] = E\left[(x_1 - \frac{1}{n}\sum_{i=1}^{n} x_i)^2\right] \\
&= E\left[x_1^2 - \frac{2}{n}x_1\sum_{i=1}^{n} x_i + \frac{1}{n^2}(\sum_{i=1}^{n} x_i)^2\right] \\
&= E\left[x_1^2 - \frac{2}{n}x_1\sum_{i=1}^{n} x_i + \frac{1}{n^2}\sum_{i=1}^{n} x_i^2 + \frac{2}{n^2}\sum_{i<j} x_i x_j\right] \\
&= E[x_1^2] + \frac{1}{n^2}\sum_{i=1}^{n} E[x_i^2] - \frac{2}{n}\sum_{i=1}^{n} E[x_1 x_i] + \frac{2}{n^2}\sum_{i<j} E[x_i x_j] \\
&= E[x_1^2] + \frac{1}{n^2}nE[x_1^2] - \frac{2}{n}E[x_1^2] - \frac{2}{n}(n-1)E[x_1 x_2] + \frac{2}{n^2}\frac{n(n-1)}{2}E[x_1 x_2] \\
&= \frac{n-1}{n}E[x_1^2] - \frac{n-1}{n}E[x_1 x_2]
\end{aligned}
$$

$$\sigma^2 = Var(x_1) = E[x_1^2] - (E[x_1])^2 = E[x_1^2] - \mu^2 \Rightarrow E[x_1^2] = \sigma^2 + \mu^2$$

Because x_1 and x_2 are independent, it follows that $Cov(x_1, x_2) = 0$. Therefore,

$$\begin{aligned} 0 &= Cov(x_1, x_2) = E[(x_1 - E[x_1])(x_2 - E[x_2])] = E[(x_1 - \mu)(x_2 - \mu)] \\ &= E[x_1 x_2] - \mu E[x_1 + x_2] + \mu^2 = E[x_1 x_2] - \mu(E[x_1] + E[x_2]) + \mu^2 \\ &= E[x_1 x_2] - \mu(2\mu) + \mu^2 = E[x_1 x_2] - \mu^2 \end{aligned}$$

So, $E[x_1 x_2] = \mu^2$.

By substituting $E[x_1^2] = \sigma^2 + \mu^2$ and $E[x_1 x_2] = \mu^2$ into the previously obtained equality ($E[\sigma_{MLE}] = \frac{n-1}{n} E[x_1^2] - \frac{n-1}{n} E[x_1 x_2]$), we get:

$$E[\sigma_{MLE}] = \frac{n-1}{n}(\sigma^2 + \mu^2) - \frac{n-1}{n}\mu^2 = \frac{n-1}{n}\sigma^2$$

c. Find an unbiased estimator for σ^2.

Solution:

It can be immediately proven that $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2$ is an unbiased estimator of σ^2.

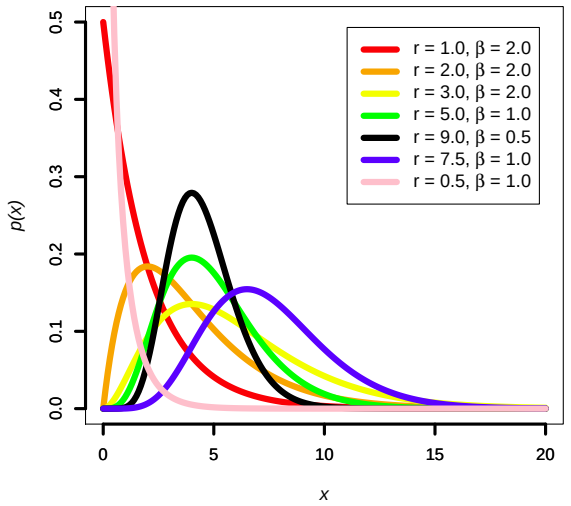
The Gamma distribution:
Maximum Likelihood Estimation of parameters

Liviu Ciortuz, 2017

The *Gamma distribution* of parameters $r > 0$ and $\beta > 0$ has the following *density function*:

$$Gamma(x|r,\beta) = \frac{1}{\beta^r \Gamma(r)} x^{r-1} e^{-\frac{x}{\beta}}, \text{ for all } x > 0,$$

where the Γ symbol designates Euler's *Gamma function*.



***Notes*:**

1. In the above definition, $\frac{1}{\beta^r \Gamma(r)}$ is the so-called *normalization factor*, since it does not depend on x, and $\int_{x=-\infty}^{+\infty} x^{r-1} e^{-\frac{x}{\beta}} dx = \beta^r \Gamma(r)$.

2. Euler's Gamma function is defined as follows: $\Gamma(r) = \int_0^{+\infty} t^{r-1} e^{-t} dt$, for all $r \in \mathbb{R}$, except for the negative integers. Starting from the definition of Γ, it can be easily shown that $\Gamma(r+1) = r\Gamma(r)$ for any $r > 0$, and also $\Gamma(1) = 1$. Therefore, $\Gamma(r+1) = r \cdot \Gamma(n) = r \cdot (r-1) \cdot \Gamma(r-1) = \ldots = r \cdot (r-1) \cdot \ldots \cdot 2 \cdot 1 = r!$, which means that the Γ function generalizes the *factorial function*.

3. The exponential distribution is a member of the Gamma family of distributions. (Just set r to 1 in Gamma's density function.)

Consider $x_1, \ldots, x_n \in \mathbb{R}^+$, each one of them being generated by one component of the above mixture, denoted by $z_i \in \{1, \ldots, K\}$.

Find the maximum likelihood estimation of the parameters r and β.

Solution

- **The verosimility function:**

$$L(r, \beta) \overset{def.}{=} P(x_1, \ldots, x_n | r, \beta) \overset{i.i.d.}{=} \prod_{i=1}^{n} P(x_i | r, \beta)$$

$$= \beta^{-rn} (\Gamma(r))^{-n} \left(\prod_{i=1}^{n} x_i \right)^{r-1} e^{-\frac{1}{\beta} \sum_{i=1}^{n} x_i}$$

- **The log-verosimility function:**

$$\ell(r, \beta) \overset{def.}{=} \ln L(r, \beta) = -rn \ln \beta - n \ln \Gamma(r) + (r-1) \sum_{i=1}^{n} \ln x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i.$$

Now we will calculate the partial derivative of $\ell(r,\beta)$ w.r.t. β, and then equate it to 0:

$$\frac{\partial}{\partial\beta}\ell(r,\beta) = -\frac{rn}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i = \frac{1}{\beta^2}\left[\sum_{i=1}^{n} x_i - rn\beta\right]$$

$$\frac{\partial}{\partial\beta}\ell(r,\beta) = 0 \Leftrightarrow \hat{\beta} = \frac{1}{rn}\sum_{i=1}^{n} x_i > 0.$$

By substituting $\hat{\beta}$ into $\ell(r,\beta)$, we will get:

$$\begin{aligned}
\ell(r,\hat{\beta}) &= -rn\ln\hat{\beta} - n\ln\Gamma(r) + (r-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{\hat{\beta}}\sum_{i=1}^{n} x_i \\
&= rn\ln(rn) - rn\ln\sum_{i=1}^{n} x_i - n\ln\Gamma(r) + (r-1)\sum_{i=1}^{n}\ln x_i - \frac{rn}{\sum_{i=1}^{n} x_i}\cdot\sum_{i=1}^{n} x_i \\
&= rn\ln(rn) - rn\left(\ln\sum_{i=1}^{n} x_i + 1\right) - n\ln\Gamma(r) + (r-1)\sum_{i=1}^{n}\ln x_i
\end{aligned}$$

Therefore, by computing the partial derivative of $\ell(r, \hat{\beta})$ with respect to r, and then equating this derivative to 0, we will get:

$$\frac{\partial}{\partial r}\ell(r,\hat{\beta}) = 0 \Leftrightarrow n\ln(nr) + n - n\left(\ln\sum_{i=1}^{n} x_i + 1\right) - n\cdot\frac{\Gamma'(r)}{\Gamma(r)} + \sum_{i=1}^{n}\ln x_i = 0 \Leftrightarrow$$

$$n(\ln r - \psi(r)) = -n\ln n - \sum_{i=1}^{n}\ln x_i + n\ln\sum_{i=1}^{n} x_i \Leftrightarrow$$

$$\ln r - \psi(r) = -\ln n - \frac{1}{n}\sum_{i=1}^{n}\ln x_i + \ln\sum_{i=1}^{n} x_i.$$

The solution of the last equation is $\hat{r}$, the maximum likelihood estimation of the parameter r.

The Gaussian multi-variate distribution:

ML estimation of
the mean and the *precision matrix*, Λ
(Λ is the inverse of the *covariance matrix*, Σ)

CMU, 2010 fall, Aarti Singh, HW1, pr. 3.2.1

The density function of a d-dimensional Gaussian distribution is as follows:

$$\mathcal{N}(x \mid \mu, \Lambda^{-1}) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^\top \Lambda (x-\mu)\right)}{(2\pi)^{d/2}\sqrt{|\Lambda^{-1}|}},$$

where Λ is the inverse of the covariance matrix, or the so-called precision matrix. Let $\{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample from a d-dimensional Gaussian distribution.

Suppose that $n \gg d$. Derive the MLE estimates $\hat{\mu}$ and $\hat{\Lambda}$.

Hint

You may find useful the following formulas (taken from *Matrix Identities*, by Sam Roweis, 1999):

(2b) $|A^{-1}| = \dfrac{1}{|A|}$

(2e) $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$;[a]
more generally, $\mathrm{Tr}(ABC\ldots) = \mathrm{Tr}(BC\ldots A) = \mathrm{Tr}(C\ldots AB) = \ldots$

(3b) $\dfrac{\partial}{\partial X}\mathrm{Tr}(XA) = \dfrac{\partial}{\partial X}\mathrm{Tr}(AX) = A^\top$

(4b) $\dfrac{\partial}{\partial X}\ln|X| = (X^{-1})^\top = (X^\top)^{-1}$

(5c) $\dfrac{\partial}{\partial X}a^\top Xb = ab^\top$

(5g) $\dfrac{\partial}{\partial X}(Xa+b)^\top C(Xa+b) = (C+C^\top)(Xa+b)a^\top$

$\mathrm{Tr}(A)$**, the *trace* of an n-by-n square matrix** A**, is defined as the sum of the elements on the main diagonal (the diagonal from the upper left to the lower right) of** A**, i.e.,** $a_{11} + \ldots + a_{nn}$**.**

[a] See Theorem 1.3.d from *Matrix Analysis for Statistics*, 2017, James R. Schott.

Given the $x_1, \ldots, x_n$ data, the log-likelihood function is:

$$\begin{aligned} l(\mu, \Lambda) &\overset{i.i.d.}{=} \ln \prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \Lambda^{-1}) = \sum_{i=1}^{n} \ln \mathcal{N}(x_i|\mu, \Lambda^{-1}) \\ &= -\frac{nd}{2} \ln(2\pi) - \frac{n}{2} \ln |\Lambda^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Lambda (x_i - \mu) \\ &\overset{(2b)}{=} -\frac{nd}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Lambda (x_i - \mu). \end{aligned}$$

For any fixed positive definite precision matrix Λ, the log-likelihood is a quadratic function of μ with a negative leading coefficient, hence a strictly concave function of μ. We then solve

$$\nabla_\mu l(\mu, \Lambda) = 0 \overset{(5g)}{\iff} \Lambda \sum_{i=1}^{n} (x_i - \mu) = 0 \iff \Lambda \sum_{i=1}^{n} x_i = n\Lambda\mu. \tag{3}$$

From (3) we get, by the assumption that Λ is invertible, the following estimate of μ:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n},$$

which coincides with the *sample mean* $\bar{x}$ and is constant w.r.t. Λ.

Now that we have

$$l(\mu, \Lambda) \leq l(\hat{\mu}, \Lambda) \quad \forall \mu \in \mathbb{R}^d,\ \Lambda \textbf{ being positive definite,}$$

we continue to consider Λ by first plugging $\hat{\mu}$ back in the log-likelihood function (3):

$$l(\hat{\mu}, \Lambda) = -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln|\Lambda| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) \tag{4}$$

$$\overset{(2e)}{=} -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}(\ln|\Lambda| - \mathrm{Tr}(S\Lambda)), \tag{5}$$

where S is the *sample covariance matrix*: $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$.

***Explanation*:**

$(x_i - \bar{x})^\top \Lambda (x_i - \bar{x})$ is a 1-by-1 matrix, therefore $(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) = \mathrm{Tr}((x_i - \bar{x})^\top \Lambda (x_i - \bar{x}))$, and using the (2e) rule, it can be further written as $\mathrm{Tr}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)$.
Using another simple rule, $\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$ (which can be easily proven), it follows that

$$\sum_{i=1}^{n}(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) = \sum_{i=1}^{n}\mathrm{Tr}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)$$
$$= \mathrm{Tr}(\sum_{i=1}^{n}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)) = \mathrm{Tr}((\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top)\Lambda) = \mathrm{Tr}((nS)\Lambda) = n\mathrm{Tr}(S\Lambda).$$

By the fact that $\ln|\Lambda|$ is strictly concave on the domain of positive definite Λ,[a] and that $\mathrm{Tr}(S\Lambda)$ is linear in Λ, we are able to find the maximum of the expression (5) by solving

$$\nabla_\Lambda l(\hat{\mu}, \Lambda) = 0,$$

which can be proven[b] to be equivalent to

$$\Lambda^{-1} - S = 0. \qquad (\textbf{Therefore, } \hat{\Sigma} = \hat{\Lambda}^{-1} = S.)$$

Since $n \gg d$, we can safely assume that S is invertible and get the following estimate:

$$\hat{\Lambda} = S^{-1}.$$

---

[a] See, for example, Section 3.1.5, *Convex Optimization*: `http://www.stanford.edu/~boyd/cvxbook/`.

[b] Use (4b) and (3b) on (5): . $(\lambda^\top)^{-1} - S^\top = 0 \overset{(4b)}{\Leftrightarrow} (\lambda^{-1})^\top = S^\top \Leftrightarrow \lambda^{-1} = S$.

Notes

1. In the above derivation, we have ensured that the estimates μ and $\hat{\Lambda}$ are in the parameter space and satisfy

$$l(\mu, \Lambda) \leq l(\hat{\mu}, \Lambda) \leq l(\hat{\mu}, \hat{\Lambda}) \quad \forall \mu \in \mathbb{R}^d, \ \Lambda \text{ being positive definite,}$$

so they are the MLE estimates.

2. Instead of using the relation (5), i.e., working with the Tr functional, one could directly compute the partial derivative of $l(\hat{\mu}, \Lambda)$ in (4):

$$\begin{aligned} \nabla_\Lambda l(\hat{\mu}, \Lambda) &\overset{(4b),(5c)}{=} \frac{n}{2}(\Lambda^\top)^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top = \frac{n}{2}\Lambda^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top \\ &= \frac{n}{2}\Lambda^{-1} - \frac{n}{2}S. \end{aligned}$$

So,

$$\nabla_\Lambda l(\hat{\mu}, \Lambda) = 0 \Leftrightarrow \hat{\Lambda}^{-1} \overset{not.}{=} \hat{\Sigma} = S \overset{not.}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top.$$