Maximum Likelihood Estimation

Let's return to the general problem of estimating a parameter (or set of parameters) of a distribution. The framework, again, is that we observe outcomes of a random variable whose distribution is controlled (parameterized) by one or more variables that we collect in the vector $\boldsymbol{\theta}$. More mathematically, the distribution of our data is assumed to be[1]

$$X \sim f_X(\boldsymbol{x};\boldsymbol{\theta}),$$

for some *unknown* $\boldsymbol{\theta} \in \mathcal{T}$; from a sample of X, we want to estimate the latent $\boldsymbol{\theta}$.

Here are some stylized examples that can be put into this framework.

Example: What is the probability that LeBron James makes a free throw? The model here is that the outcome of each free throw can captured using a binary-valued random variable X_i:

$$X_i = 0 \text{ if he misses}, \quad X_i = 1 \text{ if he makes it}.$$

The distribution of these random variables is controlled by a single parameter θ:

$$X_i = \begin{cases} 1, & \text{with probabilty } \theta, \\ 0, & \text{with probabiity } 1-\theta. \end{cases}$$

(We are assuming here that the value of θ is the same for every free throw.) Given a series of observations $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$, how can we estimate θ?

[1] $f_X(\boldsymbol{x};\boldsymbol{\theta})$ should be thought of a a density with argument $\boldsymbol{x}$ that is parameterized by some $\boldsymbol{\theta}$ — different $\boldsymbol{\theta}$ gives us different probability density functions.

Example: Suppose I represent a social network with D users with a directed graph like this:

[INSERT]

where if there is a directed edge from node i to node j, it means user j is "following" user i. Suppose that every day, a user (the same user everyday) creates and then shares some "fake news" with his followers. With probability 0.9, a follower finds this news credible and passes it on to their followers; with probability 0.1, a follower flags the news as fake and reports it to you (and does not pass it along). At the end of every day, you have a binary vector whose length is D,

$$X_i = \begin{bmatrix} X_i[1] \\ X_i[2] \\ \vdots \\ X_i[D] \end{bmatrix}, \quad X_i[d] = \begin{cases} 1, & \text{if user } d \text{ reported "fake news" on day } i \\ 0, & \text{otherwise} \end{cases}.$$

Given observations $X_1, X_2, \ldots, X_N$ over several days, how can I estimate which user is generating the "fake news"?

Example: Suppose that $X_1, \ldots, X_N$ are independent realizations of a D-dimensional Gaussian random vector with unknown mean $\boldsymbol{\mu}$

and unknown covariance $\boldsymbol{R}$. Given these N realizations, how can I estimate $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{R})$?

In the first example above, the observations are scalars, there is one parameter, and the parameter space $\mathcal{T} = [0, 1]$ was continuous (all reals between zero and one). In the second, the observations are vectors, there is one parameter, and the parameter space $\mathcal{T} = \{1, 2, \ldots, D\}$ is finite. In the third example, the observations are vectors, the unknown parameters are a vector and a matrix, and the parameter space is $\mathcal{T} = \mathbb{R}^D \otimes \mathcal{S}^D_{++}$, where $\mathcal{S}^D_{++}$ is the set of all $D \times D$ symmetric positive-definite matrices. In all three of the examples (as in all problems of this type), the parameter(s) induces different distributions on the observed data.

With a probabilistic model in place for the observations, given a sample[2] $X_1 = \boldsymbol{x}_1, X_2 = \boldsymbol{x}_2, X_N = \boldsymbol{x}_N$, the **likelihood** of a particular set of parameters $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = f_{X_1,\ldots,X_N}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}).$$

We introduce this new notation to emphasize that the likelihood should be thought of as a function of $\boldsymbol{\theta}$. The maximum likelihood estimation is simply the parameters that maximize $L(\boldsymbol{\theta}; \cdot)$:

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}} \; L(\boldsymbol{\theta}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N).$$

Because many times the joint distribution of $X_1, \ldots, X_N$ involves multiplying a bunch of functions together (especially when the observations are independent), it is often convenient to work with the **log likelihood**

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \log f_{X_1,\ldots,X_N}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}).$$

[2] As in the two examples, the observations might be vectors or scalars.

It should be absolutely clear[3] that $L(\boldsymbol{\theta};\cdot)$ and $\ell(\boldsymbol{\theta};\cdot)$ are maximized at the same place, so

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{T}}\ \ell(\boldsymbol{\theta};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \ = \ \arg\min_{\boldsymbol{\theta}\in\mathcal{T}} -\ell(\boldsymbol{\theta};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N).$$

Let's see how this works.

Example: In the LeBron James freethrow example above, given $X_1 = x_1, \ldots, X_N = x_N$, and assuming each of the trials are independent of one another, we have

$$L(\theta;x_1,\ldots,x_N) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n}.$$

Notice that the expression inside the product above is θ if $x_n = 1$, and $1-\theta$ if $x_n = 0$. With

$$S_N = \sum_{n=1}^{N} x_n,$$

the expression above becomes

$$L(\theta;x_1,\ldots,x_N) = \theta^{S_N}(1-\theta)^{N-S_N},$$

and so

$$\ell(\theta;x_1,\ldots,x_N) = S_N\log\theta + (N-S_N)\log(1-\theta).$$

[3] If it is not absolutely clear to you, then you have a new homework question: Let $\phi:\mathbb{R}\rightarrow\mathbb{R}$ be a monotonically increases function on $\mathbb{R}$, and $f:\mathbb{R}^D\rightarrow\mathbb{R}$ be a function on $\mathbb{R}^D$. Show that $\arg\max_{\boldsymbol{v}}\phi(f(\boldsymbol{v})) = \arg\max_{\boldsymbol{v}} f(\boldsymbol{v})$.

The first and second derivatives of ℓ are

$$\begin{aligned}\frac{\mathrm{d}\ell(\theta;\cdot)}{\mathrm{d}\theta} &= \frac{S_N}{\theta} - \frac{N-S_N}{1-\theta}\\ \frac{\mathrm{d}^2\ell(\theta;\cdot)}{\mathrm{d}\theta^2} &= -\frac{S_N}{\theta^2} - \frac{N-S_N}{(1-\theta)^2}.\end{aligned}$$

Since $0 \leq S_N \leq N$, the second derivative is ≤ 0 for all $\theta \in \mathcal{T} = [0,1]$, and so we can find the maximizer by setting the first derivative equal to zero. This yields

$$\hat{\theta}_{\text{mle}} = \frac{S_N}{N} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

Example: Suppose that X is a scalar random variable distributed uniformly on the interval $[a,b]$, $X \sim \text{Uniform}([a,b])$, where a and b are unknown. Given $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$, what is the MLE for $\boldsymbol{\theta} = (a,b)$?

Answer:

Example: Suppose we observe $X_1 = \boldsymbol{x}_1, \ldots, X_N = \boldsymbol{x}_N$, where the X_n are independent and identically distributed Gaussian random vectors in $\mathbb{R}^D$:

$$X_n \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R}).$$

From these observations, we want to estimate $\boldsymbol{\mu}$ and $\boldsymbol{R}$. The MLE is the solution to

$$\underset{\boldsymbol{\mu}\in\mathbb{R}^D,\boldsymbol{R}\in\mathcal{S}^D_{++}}{\text{maximize}} \prod_{n=1}^{N}(2\pi)^{-D/2}(\det\boldsymbol{R})^{-1/2}\exp(-(\boldsymbol{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})/2).$$

Taking the log of the likelihood function, this is equivalent to

$$\underset{\boldsymbol{\mu}\in\mathbb{R}^D,\boldsymbol{R}\in\mathcal{S}^D_{++}}{\text{maximize}} \frac{-ND}{2}\log(2\pi)+\frac{N}{2}\log\det\boldsymbol{R}^{-1}-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}).$$

The first terms doesn't depend on either $\boldsymbol{\mu}$ or $\boldsymbol{R}$, so we can drop it, leaving us with

$$\underset{\boldsymbol{\mu}\in\mathbb{R}^D,\boldsymbol{R}\in\mathcal{S}^D_{++}}{\text{maximize}} \frac{N}{2}\log\det\boldsymbol{R}^{-1}-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}).$$

It turns out that we can tease out an estimate for $\boldsymbol{\mu}$ from the expression above that is independent of $\boldsymbol{R}$. The maximizer of the second term above is the same as solving the minimization program[4]

$$\underset{\boldsymbol{\mu}\in\mathbb{R}^D}{\text{minimize}} \sum_{n=1}^{N}\|\boldsymbol{R}^{-1/2}(\boldsymbol{x}_n-\boldsymbol{\mu})\|_2^2.$$

This is a least-squares problem equivalent to minimize$_{\boldsymbol{\mu}}$ $\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\|_2^2$ with

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{R}^{-1/2}\\ \boldsymbol{R}^{-1/2}\\ \vdots\\ \boldsymbol{R}^{-1/2}\end{bmatrix},\quad \boldsymbol{y}=\begin{bmatrix}\boldsymbol{R}^{-1/2}\boldsymbol{x}_1\\ \boldsymbol{R}^{-1/2}\boldsymbol{x}_2\\ \cdots\\ \boldsymbol{R}^{-1/2}\boldsymbol{x}_N\end{bmatrix}.$$

[4] We know that the expression $\boldsymbol{R}^{-1/2}$ makes sense thanks to the fact that $\boldsymbol{R}\in\mathcal{S}^D_{++}$ and the Sylvester theorem.

Thus

$$\hat{\boldsymbol{\mu}}_{\mathrm{mle}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n.$$

and so the estimator for the mean does not depend on the covariance. We can now estimate $\boldsymbol{R}$ as the solution to

$$\underset{\boldsymbol{R}\in\mathcal{S}^D_{++}}{\text{maximize}} \;\; \frac{N}{2}\log\det\boldsymbol{R}^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}).$$

Using the easily checked fact that $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{w} = \operatorname{trace}(\boldsymbol{S}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})$ for vectors $\boldsymbol{w}$ and sym+def matrices $\boldsymbol{S}$, and the fact that trace is a linear operator, and the fact that an inverse of a sym+def matrix is again sym+def, this is equivalent to

$$\underset{\boldsymbol{S}\in\mathcal{S}^D_{++}}{\text{maximize}} \;\; \log\det\boldsymbol{S} - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is the sample covariance,

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})^{\mathrm{T}}.$$

It is a fact that log det is concave in its matrix argument $\boldsymbol{S}$ over $\mathcal{S}^D_{++}$, and its gradient is

$$\nabla\log\det\boldsymbol{S} = \boldsymbol{S}^{-1}.$$

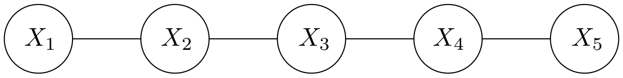
The function $\operatorname{trace}(\boldsymbol{S}\boldsymbol{\Sigma})$ is linear in $\boldsymbol{S}$, and hence both concave and convex; its gradient is

$$\nabla\operatorname{trace}(\boldsymbol{S}\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}.$$

So setting the gradient equal to zero yields:

$$\hat{\boldsymbol{S}} = \boldsymbol{\Sigma}^{-1}, \quad \text{or} \quad \hat{\boldsymbol{R}}_{\mathrm{mle}} = \boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_{\mathrm{mle}})(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_{\mathrm{mle}})^{\mathrm{T}}.$$

Example: Suppose that we have some information about the covariance matrix $\boldsymbol{R}$. One for this information could take is the the conditional independence structure. For instance, if this were quantified with the graph



we would know that certain entries in the inverse covariance matrix were zero. In this case, we would set up a constrained optimization program

$$\begin{aligned}
&\underset{\boldsymbol{S}\in\mathcal{S}^5_{++}}{\text{maximize}} \quad \log\det\boldsymbol{S} - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Sigma}) \quad \text{subject to} \\
&\qquad\qquad S[1,3]=0,\ S[1,4]=0,\ S[1,5]=0,\ S[2,4]=0,\ S[2,5]=0 \\
&\qquad\qquad S[3,1]=0,\ S[3,5]=0,\ S[4,1]=0,\ S[4,2]=0.
\end{aligned}$$

This program does not have an explicit solution, but it is a concave program with linear constraints, so there is an established methodology for solving it.