Maximum Likelihood Estimation

March 6, 2025

Motto

It is a part of probability that many improbable things will happen.

Aristotle

Outline

1. General MLE

2. MLE for exponential families

3. Algorithms

Estimation Problem

- Consider a sample **X** that comes from an unknown member P_{θ_0} of a family of distributions $\{P_\theta\}_{\theta\in\Theta}$.
- Any statistic $\hat{\theta}$ that is computable from the observed values **x** of **X** and aims at approximating true θ_0 can be called an estimator of this parameter.

Estimation Problem

- Consider a sample **X** that comes from an unknown member P_{θ_0} of a family of distributions $\{P_\theta\}_{\theta\in\Theta}$.
- Any statistic $\hat\theta$ that is computable from the observed values **x** of **X** and aims at approximating true θ_0 can be called an estimator of this parameter.
- We are interested only in ‘good estimators’.
- For example, we want the estimator error $\epsilon = \hat\theta - \theta_0$ to be small, for example its MSE to be small

$$E_{\theta_0}(\epsilon^2) = Var_{\theta_0}(\hat\theta) + \left(E_{\theta_0}\hat\theta - \theta_0\right)^2 = Var_{\theta_0}(\hat\theta) + Bias^2_{\theta_0}.$$

- Ideally, we would like to have the bias to be zero and the variance to be minimal, although, it does not exclude the case when the MSE to be smaller for a biased estimator than the minimal variance for all unbiased estimators.

Estimator maximizing the likelihood

- $L_{\mathbf{x}}(\theta)$ – the likelihood function of θ (the probability density depending on θ with given $\mathbf{x}$).
- We think of $L_{\mathbf{x}}(\theta)$ as a measure of how “likely” θ is to have produced the observed $\mathbf{x}$.
- The method of maximum likelihood finds that value $\hat{\theta}_{MLE}$ of the parameter that is “most likely” to have produced the data.
- That is, if $\mathbf{X} = \mathbf{x}$ is observed

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L_{\mathbf{x}}(\theta).$$

- Maximum likelihood estimates need neither exist nor be unique. They do not even need to be ‘best’.
- But in a wide class of important cases, they are asymptotically (with respect to the sample size) the best possible ones.

Two examples

- An iid sample from normal distribution with $\theta = (\mu, \sigma^2)$. It is easy to check that

$$\hat{\theta} = (\bar{X}, \overline{X^2} - \bar{X}^2)$$

  is the MLE for θ.

- An iid sample from uniform distribution on a sequence $\{1, \dots, \theta\}$ with the size of the population $\theta \in \mathbb{N}$ being an unknown parameter. It is easy to see that the likelihood has the form

$$L_{\mathbf{x}}(\theta) = \frac{1}{\theta^n} \mathbf{1}_{[0,\theta]}(\max_{i=1,\dots,n} x_i)$$

  and the maximum is reached at $\hat{\theta} = \max_{i=1,\dots,n} x_i$. Similar argument leads to the same estimator if the sample is from the continuous uniform distribution on $[0, \theta]$.

Two examples, cont.

The MLE estimators are not necessarily the best estimator, although they are generally good ones.

Homework *For the estimation problem of the normal distribution, show that the MLE estimator of variance is biased and check what is its MSE. How this MSE compares to the MSE of the classical unbiased estimator.*

For the uniform distribution case, the estimator may even seem unnatural. It excludes possibility of exceeding the maximum of the data, which seems to be wrong.

Homework *For the estimation problem of the uniform distribution, show that the MLE estimator of θ is biased and check what is its MSE. Based on your findings suggest the unbiased estimator and evaluat the MSE of both MLE and the proposed unbiased estimator.*

Likelihood equations

- Suppose:
  - $X \sim P_\theta$, with $\theta \in \Theta$, an open set parameter space
  - the likelihood function $L_X(\theta)$ is differentiable in θ
  - $\hat{\theta}_{MLE}$ exists

  Then: $\hat{\theta}_{MLE}$ must satisfy the **Likelihood Equation(s)**

$$\nabla_\theta L_X(\theta) = 0.$$

- Important Case: for independent X_i's with

$$\sum_{i=1}^{n} \nabla \log p_i(x_i|\theta) = 0$$

NOTE: $p_i(\cdot|\theta)$ may vary with i, so that elements in the sample can have different distributions.

Multinomial trials

- an experiment with n i.i.d. trials in which each trial can produce a result in one of k categories
- $X_i = j$ if the ith trial produces a result in the jth category
- let $\theta_j = P(X_i = j)$ be the probability of the jth category, and let

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{\{X_i = j\}},$$

  i.e. N_j is the number of observations in the jth category.
- for an experiment in which we observe n_j, $j = 1, \ldots, k$ we have

$$l_{\mathbf{x}}(\theta) = \sum_{j=1}^{k} n_j \log \theta_j,$$

  with $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$ and the normal (likelihood) equations are for $j < k$

$$\frac{\partial l_{\mathbf{x}}}{\partial \theta_j} = n_j/\theta_j - n_k/\left(1 - \sum_{i=1}^{k-1} \theta_i\right) = 0.$$

  Thus, assuming w.l.g. that $n_k \neq 0$, $n_j/n_k = \theta_j/\theta_k$, which implies that $\hat{\theta}_j = n_j/n$.

Outline

1 General MLE

2 MLE for exponential families

3 Algorithms

Natural parameter

Questions of existence and uniqueness of maximum likelihood estimates in canonical exponential families can be answered completely and elegantly. This is largely a consequence of the strict concavity of the log likelihood in the natural parameter η.

Recall that $\{P_\theta\}$, $\theta \in \Theta$ is a k-parameter exponential family if

$$p(x|\theta) = h(x) \exp\left(\sum_{j=1}^{k} \eta_j(\theta) T_j(x) - B(\theta)\right), \quad x \in \mathbb{R}^q$$

- $\eta_1, \ldots, \eta_k$ and B are real-valued functions mapping $\Theta \mapsto \mathbb{R}$,
- $T_1, \ldots, T_k$ and h are real-valued functions mapping $\mathbb{R}^q \mapsto \mathbb{R}$.

Canonical form of exponential families

Canonical parameter

Consider the canonical form of the exponential density

$$q(x|\eta) = h(x)\exp(\mathbf{T}^\top(x)\eta - A(\eta)).$$

In the continuous case

$$A(\eta) = \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x)\exp(\mathbf{T}^\top(x)\eta) dx.$$

In the discrete case, $A(\eta)$ is defined in the same way except integrals are replaced by sums. In either case, we define the natural parameter space as

$$\mathcal{E} = \{\eta \in \mathbf{R}^k : -\infty < A(\eta) < \infty\},$$

where $k \in \mathbb{N}$ is the rank of the exponential family (the smallest possible dimension).

A sample from canonical exponential family

- If a distribution is from exponential family, then independent sample $(X_1, \dots, X_n)$ from this distribution has the distribution from exponential family.

Homework *Find the canonical exponential form of the distribution of $(X_1, \dots, X_n)$ given that you know this form for the individual X_i.*

Homework *Argue directly from the definition of $A(\eta)$ on the previous page that $\dot{A}(\eta) = E_\eta(\mathbf{T}(X))$.*

Homework *Argue convexity of the loglikelihood of the exponential family from the fact the matrix of the second derivatives of $\ddot{A}(\theta) = Var_\eta(\mathbf{T}(X))$ which is a positive definite matrix (what is a connection between convexity of a multivariate function and the matrix of its second derivatives?).*

MLE in the convexity (concavity) context

Proposition 2.3.1.

Suppose that the domain of the parameters is open in $\mathbf{R}^p$. If the log-likelihood is the strictly concave function of the parameter and is unbounded on the boundary of the parameter space, then the MLE exists and is unique.

Some clarificiations:

- an open set is a set that with each point contains also a disk (ball) that is in this set – topological definition
- the boundary of a set U is a set ∂U that in any disk (ball) contains both the points from the set U and outside of the set U
- the strictly concave is a function like $-x^2$, ('sad face')

A necessary and sufficient condition for existence and uniqueness of the MLE

Theorem 2.3.1.

Suppose $\mathcal{P}$ is the canonical exponential family generated by $(\mathbf{T}, h)$ and that

- the natural parameter space, $\mathcal{E}$, is open,
- the family is of rank k.

Then the MLE of η exists and is unique for every x such that $\mathbf{t}_0 = \mathbf{T}(x)$ satisfies for all $\mathbf{c} \neq \mathbf{0}$:

$$P(\mathbf{c}^\top \mathbf{T}(X) > \mathbf{c}^\top \mathbf{t}_0) > 0. \tag{1}$$

and the MLE is a solution to

$$\dot{A}(\eta)(= E_\eta(\mathbf{T}(X))) = \mathbf{t}_0.$$

Conversely, if t_0 doesn't satisfy (1), then the MLE doesn't exist and the above equation has no solution.

Example – the two-parameter gamma family

The density proportional to $\exp(-\lambda x)x^{p-1}$. This is a rank 2 canonical exponential family with

$$\begin{aligned} \mathbf{T} &= \left(\sum \log X_i, \sum X_i\right) \\ h(x) &= 1/x \\ \eta &= (p, -\lambda) \\ A(\eta) &= n\left(\log \Gamma(\eta_1) - \eta_1 \log(-\eta_2)\right). \end{aligned}$$

We have

$$\dot{A}(\eta) = n(\Gamma'(p)/\Gamma(p) - \log \lambda, p/\lambda) = n(\overline{\log X}, \bar{X}).$$

We conclude from Theorem 2.3.2 that the equations have a unique solution with probability 1. How to find such nonexplicit solutions is discussed in Section 2.4 to which we turn next.

Outline

1 General MLE

2 MLE for exponential families

3 Algorithms

MLE as an optimization problem

- As we have seen, even in the context of canonical multiparameter exponential families, such as the two-parameter gamma, MLEs may not be given explicitly by formulae but only implicitly as the solutions of systems of nonlinear equations.
- In the classical regression model with design matrix of full rank, the formula for the parameter estimator is easy to write down symbolically but not easy to evaluate if the dimension of the parameter is large. It is because the number of operations for inversion of a matrix is on the order of the third power with respect to the parameter dimension. For example, if one looks at the expression of 10000 genes through a linear model, the inverse of the matrix would involve 10^{12} operations.
- We will discuss three algorithms of a type used in different statistical contexts both for their own sakes and to illustrate what kinds of things can be established about the black boxes to which we all, at various times, entrust ourselves.

Bisection algorithm

By the calculus, if the likelihood is differentiable and the location of the global maximum is in the interior of the parameter space finding the solution reduces to finding the solution to the likelihood (normal) equations (the equations for zeros of the derivative functions).

Algorithm

- Finding two points, one that the derivative is below zero and one that it is above.
- Take the middle of the interval made of these points and cut it in half.
- Evaluate the value of the derivative at the middle, and replace by the middle point the one of the original points that has the same sign of the value of the derivative.
- Continue until the length of the interval has the desired accuracy or the zero is reached before that.

Example: the shape for gamma distribution

- Recall that the density of a gamma distribution is given by

$$f(x;\beta,\tau)=\frac{x^{\tau-1}}{\beta^{\tau}\Gamma(\tau)}e^{-x/\beta}.$$

- The expected value of a gamma variable X is $E(X)=\tau\beta$.
- Suppose that we have a sample $x_1,\dots,x_n$ from this distribution and it is known that the expected value is equal to a certain known value, say a.
- Loglikelihood is then equal to

$$l(\tau;x_1,\dots,x_n)=$$

$$=(\tau-1)\sum_{i=1}^{n}\log(x_i)+\tau\sum_{i=1}^{n}x_i/a-\log n\tau\left(\log a-\log\tau\right)\log\Gamma(\tau)=$$

$$=A+\tau B-\tau\log\Gamma(\tau)\log n(\log a-\log\tau).$$

Hybrid estimation the shape for gamma distribution

- The derivative of the loglikelihood is

$$
\begin{aligned}
&l'(\tau; x_1, \dots, x_n) = \\
&\qquad = B - (\tau \log \Gamma(\tau))' \log n(\log a - \log \tau) + \log \Gamma(\tau) \log n \\
&\qquad = B - [\log(a/e) \log \Gamma(\tau) - \tau\psi(\tau) \log(\tau/a) - \log \tau \log \Gamma(\tau)] \log n,
\end{aligned}
$$

where $\psi(\tau) = \Gamma'(\tau)/\Gamma(\tau)$ is the digammma function.

Homework *Verify the above calculations. Based on them propose an algorithm that uses the sample mean estimator of the mean of gamma distribution and the maximizing the likelihood with respect τ (the hybrid estimation method). Design a Monte Carlo study that examines the performance of this algorithm. Summarize the results of a study in a table.*

Example: the shape for gamma distribution

- The bisection algorithm applied to a strictly increasing derivative (concave function) finds the solution that is unique.
- For the canonical exponential family this is the case since $f(\eta) = E_\eta T(X) - t_0$ is strictly increasing:

$$f'(\eta) = A''(\eta) = Var_\eta T(X) > 0.$$

- **Example:** For the shape parameter gamma family with the density proportional to $x^{\theta-1}e^{-x}$, one has to find the solution

$$\frac{\Gamma'(\theta)}{\Gamma(\theta)} = \overline{\log X}.$$

Extension to multidimensional parameter – coordinate ascent

We need to find a solution to the normal equation

$$\dot{\mathbf{A}}(\eta_1, \ldots, \eta_k) = \mathbf{T}$$

- The coordinate ascent method is the bisection algorithm applied to each coordinate in the iterative recycling manner.
    - Start with an arbitrary $(\hat{\eta}_1^0, \ldots, \hat{\eta}_k^0)$ (some at hoc estimation, method of moments or plug-in estimation are recommended)
    - Solve the equation for the first derivative with respect to η_1 leading, say, to η_1^1 and replace η_1^0 by it.
    - Repeat the previous step for the first derivative with respect to η_2 and so on until all η_i^0's are replaced by the corresponding η_i^1's.
    - Repeat the above series of steps with $(\hat{\eta}_1^{j-1}, \ldots, \hat{\eta}_k^{j-1})$ replaced by $(\hat{\eta}_1^j, \ldots, \hat{\eta}_k^j)$ and leading to $(\hat{\eta}_1^{j+1}, \ldots, \hat{\eta}_k^{j+1})$ until the desired accuracy is attained.

Why does it work? – mathematics

Theorem 2.4.2.

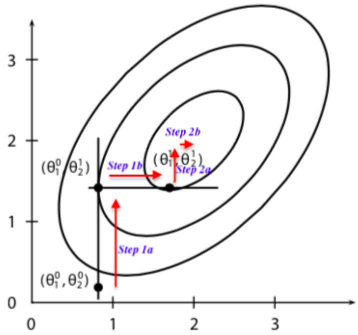
If the conditions of Theorem 2.3.1 hold, then $(\hat{\eta}_1^r, \ldots, \hat{\eta}_k^r) \overset{r\to\infty}{\to} (\hat{\eta}_1, \ldots, \hat{\eta}_k)$.

A sketch of the proof.

- The log-likelihood is $\mathbf{t}_0^\top \eta - A(\eta) + \log h(\mathbf{x})$ and it is a concave function.
- The algorithm finds the maximum along one coordinates while all other are fixed so that the value of the likelihood is increasing at each step of the algorithm.
- The log-likelihood is bounded from the above in the interior of the set of the parameters thus there has to be a limit $l(\hat{\eta}^r) \to \lambda \in (-\infty, \infty)$
- This implies that the $\hat{\eta}^r$ for large r must reside in some closed ball in the interior of the parameter space and thus needs to have a convergent subsequence.
- The limit of this subsequence must be the same due to the strict concavity and the continuity of the derivatives of A, that take the value zero at each ‘loop’ of the algorithm so in the limit they have to be zero at all the coordinates.

□

Why does it work? – a picture


3
2
1
0
0
1
2
3
(θ_1^0, θ_2^1)
(θ_1^0, θ_2^0)
(θ_1^1, θ_2^1)
Step 1a
Step 1b
Step 2a
Step 2b

The coordinate ascent algorithm. The graph shows log likelihood contours, that is values of (θ_1, θ_2) where the loglikelihood is constant. At each stage with one coordinate fixed, find that member of the family of contours to which the vertical (or horizontal) line is tangent.

Example: gamma distribution

As an alternative (better?) method to the hybrid one from Homework, one can solve the problem of estimation of the shape τ and scale β using Theorem 2.3.1, see the previous section. This is a rank 2 canonical exponential family with

$$\begin{aligned} \mathbf{T} &= \left(\sum \log X_i, \sum X_i\right) \\ h(x) &= 1/x \\ \eta &= (\tau, -1/\beta) \\ A(\eta) &= n\left(\log \Gamma(\eta_1) + \eta_1 \log(-\eta_2)\right). \end{aligned}$$

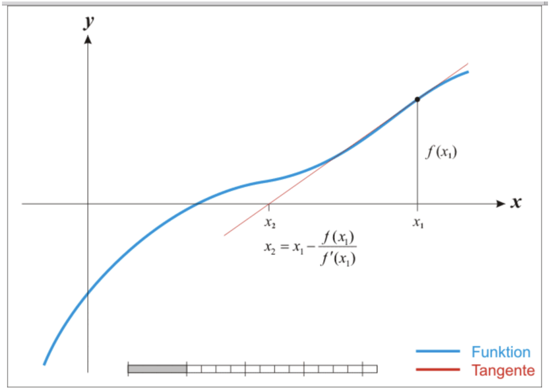
We have

$$\dot{A}(\eta) = n(\Gamma'(\tau)/\Gamma(\tau) + \log \beta, \tau\beta) = n(\overline{\log X}, \bar{X}).$$

Homework *Implement the coordinate ascent method to estimate τ and β. Perform simulation study of the performance and compare to the hybrid method discussed in the previous homework. Draw the conclusions on the performance of the two methods.*

The Newton-Raphson Algorithm

Let us start with a visualization of the algorithm



The algorithm that, in general, can be shown to be faster than coordinate ascent. This method requires computation of the inverse of the Hessian, which may counterbalance its advantage in speed of convergence (when it does converge).

NRA for MLE and exponential family

- For a general log-likelihood

$$\hat{\theta}_{new} = \hat{\theta}_{old} - \ddot{l}^{-1}(\hat{\theta}_{old})\dot{l}(\hat{\theta}_{old}).$$

- For the canonical exponential family

$$\hat{\eta}_{new} = \hat{\eta}_{old} - \ddot{A}^{-1}(\hat{\eta}_{old})\left(\dot{A}(\hat{\eta}_{old}) - \mathbf{t}_0\right).$$

When likelihoods are non-concave, methods such as bisection, coordinate ascent, and Newton–Raphson's are still employed, though there is a distinct possibility of nonconvergence or convergence to a local rather than global maximum.

The EM algorithm – foundation

There are many models that have the following structure.

- There are complete observations, X with density $p(x, \theta)$ with log likelihood $l_x(\theta)$ that is easy to maximize.
- Unfortunately, only incomplete data are observed and they are given by $S = S(X)$ with density $q(s, \theta)$, where log-likelihood $l_s(\theta)$ is difficult to maximize.
- A fruitful way of thinking of such problems is in terms of S being incomplete data and representing a part of X, the rest of X is "missing" and its "reconstruction" is part of the process of estimating θ by maximum likelihood.
- The algorithm was formalized with many examples in Dempster, Laird, and Rubin (1977), though an earlier general form goes back to Baum, Petrie, Soules, and Weiss (1970).

Example – normal distribution with missing data

- Let $(Z_1, Y_1), \dots, (Z_n, Y_n)$ be i.i.d. as $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.
- Suppose that some of the Z_i and some of the Y_i are missing as follows:
  - For $1 \le i \le n_1$ we observe both Z_i and Y_i,for $n_1 + 1 \le i \le n_2$, we observe only Z_i, and for $n_2 + 1 \le i \le n$, we observe only Y_i. The observed data are denoted by S.
  - In this case a set of sufficient statistics for the complete data is

$$T = (\bar{Z},\ \bar{Y},\ \overline{Z^2}, \overline{Y^2}, \overline{ZY}).$$

- One wants to reconstruct the missing parts needed for T from the incomplete data given in S.

Homework *An intuitive way to approach to the missing data is by replacing statistics in T by their conditional expectation with respect to $S(X)$. Formalize this approach and give its explicit description.*

Formulation of the EM algorithm

- Let
$$J(\theta|\theta_0) = E_{\theta_0}\left(\log \frac{p(X,\theta)}{p(X,\theta_0)}\middle| S = s\right).$$
- **E-step** – evaluate $J(\theta|\theta_0)$ as a function of θ.
- **M-step** – maximize $J(\theta|\theta_0)$ over θ to get a ‘new’ θ_0.
- repeat the above until the convergence.

If θ_{new} and θ_{old} the values at the end and at the beginning of the loop in the algorithm, respectively. Then

$$q(s,\theta_{new}) \geq q(s,\theta_{old}).$$

The formal argument is interesting (not very intuitive), see Lemma 2.4.1.

The EM algorithm for exponential families

Theorem 2.4.3

For a canonical exponential family generated by (T, h) satisfying the conditions of Theorem 2.3.1. Let $S(X)$ be any statistic, then the EM algorithm consists of the alternation

$$\dot{A}(\theta_{new}) = E_{\theta_{old}}(T(X)|S(X) = s).$$

Example (cont.): We can see that

$$\dot{A}(\theta) = E_\theta T = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \sigma_1\sigma_2\rho + \mu_1\mu_1)$$

so the left-hand side is straightforward. The right hand side in Theorem 2.4.3 can be easily derived by noting the well-known relations

$$E_\theta(Y|Z) = \mu_2 + \rho\sigma_2(Z - \mu_1)/\sigma_1$$
$$E_\theta(Y^2|Z) = (\mu_2 + \rho\sigma_2(Z - \mu_1)/\sigma_1)^2 + (1 - \rho^2)\sigma_2^2.$$

Example of normal distribution continued

The new paremater values being simply regular functions of the sufficient statistic T but evaluated at the expected values in the E-step of the algorithm. See Example 2.4.6 for details.

Homework *Compare critically the above method to the one that you described in the previous homework. How are they similar and how they differ? Can you provide some analysis of the performance, theoretical, or by simulations?*