Maximum Likelihood in Exponential Families

BS2 Statistical Inference, Lecture 6 Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 7, 2004

Maximum likelihood in canonical families

Recall that a canonical exponential family has the form

$$f(x;\theta) = b(x)e^{\theta^\top t(x) - c(\theta)}, \theta \in \Theta \subseteq \mathcal{R}^d,$$

where Θ is open and connected.

To find the MLE of θ based on observing $X = x$ we write

$$l_x(\theta) = \log L_x(\theta) = \theta^\top t(x) - c(\theta)$$

and equate partial derivatives w.r.t. θ_r to zero to get the maximum likelihood equations

$$s_r(\theta) = \frac{\partial}{\partial \theta_r} l_x(\theta) = 0 \iff t_r(x) = \mathbf{E}_\theta\{t_r(X)\},$$

where we have used that

$$\frac{\partial}{\partial \theta_r} c(\theta) = \mathbf{E}_\theta\{t_r(X)\}.$$

Taking second derivatives we further get

$$\frac{\partial^2}{\partial \theta_r \theta_s} l_x(\theta) = -i_{rs}(\theta) = -\operatorname{Cov}_\theta\{t_r(X), t_s(X)\}.$$

Since the latter is negative definite, any stable point of the log-likelihood is a maximum and there is therefore also at most one of them.

Another way of expressing the latter is to say that the *log-likelihood function is strictly concave.*

Moment equations for the MLE

What we have just shown can be expressed as follows:

In canonical exponential families the log-likelihood function has at most one local maximum within Θ. This is then equal to the global maximum and determined by the unique solution to the equation

$$\mathbf{E}_\theta\{t(X)\} = t(x).$$

In this sense the method of MLE for linear exponential families is similar to the method of moments, just that general functions $t_1(X), t_2(X), \ldots, t_d(X)$ are used rather than the powers $X, X^2, \ldots, X^d$.

It is less trivial to identify when a solution to the likelihoood equation exists, but the problem is well understood.

Example

Consider again the linear and canonical exponential family of Gamma distributions with

$$f(x;\alpha,\beta) \quad = \quad \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x} = \frac{1}{x} e^{\alpha \log x - \beta x + \alpha \log \beta - \log \Gamma(\alpha)},$$

where $\alpha > 0$ and $\beta > 0$ are unknown.

Assume we have a sample $x = (x_1, \dots, x_n)$ from this distribution.

The canonical sufficient statistics for the sample are then

$$\sum t(x_i) = \left(\sum \log x_i, -\sum x_i \right)$$

and the corresponding likelihood equations become

$$\sum \log x_i = n\mathbf{E}_{\alpha,\beta}(\log X) = n\{\psi(\alpha) - \log \beta\}$$

and

$$-\sum x_i = -n\mathbf{E}_{\alpha,\beta}(X) = -n\alpha/\beta.$$

Solving the second equation for β and inserting the result into the first yields

$$\overline{\log x} - \log \overline{x} = \psi(\alpha) - \log \alpha, \quad \beta = \alpha/\bar{x}.$$

The first of these equations must be solved numerically.

This is in contrast to the simple moment estimators where $\mathbf{E}(\log X)$ is replaced with $\mathbf{E}(X^2)$ to yield the explicit solution $\tilde{\alpha} = n\bar{x}^2/SSD$.

The mean value mapping

Define the *mean value mapping* τ as

$$\tau(\theta) = \mathbf{E}_\theta\{t(X)\}.$$

The likelihood equation can then be compactly written as

$$\tau(\theta) = t(x)$$

and since the likelihood equation always has at most one solution, the mapping τ is one-to-one so we can write

$$\hat{\theta} = \tau^{-1}\{t(x)\}$$

provided a solution exists, i.e. if $t(x)$ is in the *image* of τ.

The mean value parameter

This then leads to the idea of using $\eta = \tau(\theta)$ as an alternative parametrization of the exponential family.

The parameter η is known as the *mean value parameter* whereas the parameter θ is known as the *canonical* parameter.

The literature is a little confused concerning the terminology. Many authors use the term *natural parameter* for θ, but others use the same term for η, so beware when you read about exponential families elsewhere.

Example

Consider the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 > 0$ and both μ and σ^2 unknown. From the expression

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2} + x\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2}$$

we identify the canonical parameters as

$$\theta_1 = \frac{-1}{2\sigma^2}, \quad \theta_2 = \frac{\mu}{\sigma^2}$$

whereas the mean value parameters are

$$\eta_1 = \mathbf{E}(X^2) = \sigma^2 + \mu^2, \quad \eta_2 = \mathbf{E}(X) = \mu.$$

Note that both parametrizations are different from the usual (μ, σ^2).

Estimation of the mean value parameter

It follows from the invariance of the MLE under reparametrizations that we simply have

$$\hat{\eta} = t(X)$$

and therefore that *the MLE for the mean value parameter is unbiased*. Since $t(X)$ is also complete and sufficient, *the MLE for is MVUE for the mean value parameter*.

But in addition it holds that *the MLE of the mean value parameter is efficient* in the sense that it attains the Cramér–Rao bound.

To see the latter in the one-dimensional case, we just use that for $\eta = \tau(\theta)$ we have

$$\tau(\theta) = c'(\theta), \quad i(\theta) = c''(\theta) = \tau'(\theta)$$

so the score statistic has the form

$$s(x;\theta) = t(x) - c'(\theta) = \frac{i(\theta)\{t(x) - \tau(\theta)\}}{\tau'(\theta)},$$

which was the condition derived in Lecture 2.

The converse also holds: the Cramér–Rao bound is *only* attained for affine transformations of the mean value parameter.

Note also that *the mean value mapping τ is continuously differentiable* (in fact it is analytic) with derivatives

$$\frac{\partial}{\partial\theta_s}\tau_r(\theta) = \frac{\partial^2}{\partial\theta_r\theta_s}c(\theta) = i_{rs}(\theta).$$