Maximum likelihood asymptotics

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 13, Hilary Term 2009

March 4, 2009

Let $X_1, \ldots, X_n$ be independent and identically distributed with density f and moment generating function $M(t) = \mathbf{E}e^{tX}$. The *cumulant generating function* of X is

$$K(t) = \log M(t) = \sum_{r=1}^{\infty} \frac{\kappa_r}{r!} t^r,$$

and the coefficient

$$\kappa_r = \frac{\partial^r K(0)}{\partial^r t}$$

is the *cumulant of order r*. *The first two cumulants are the mean and variance*

$$\kappa_1 = \mu = \mathbf{E}(X), \quad \kappa_2 = \sigma^2 = \mathbf{V}(X).$$

If X and Y are independent random variable, their cumulants satisfy

$$\kappa_r(aX + bY) = a^r \kappa_r(X) + b^r \kappa^r(Y).$$

The *standardized cumulants*

$$\rho_r = \kappa_r / \kappa_2^{r/2}, r = 3, 4, \ldots$$

are thus invariant under translations and scaling

$$\rho_r(aX + b) = \rho_r$$

and therefore determine the shape of the density.

In the normal distribution, $\kappa_r = 0$ for $r > 2$ and cumulants ρ_r for $r > 2$ therefore indicate departures from normality.

The third standardized cumulant ρ_3 is is known as the *skewness*, and the fourth ρ_4 as the *kurtosis* of the distribution.

F. Y. Edgeworth (1845-1926), Professor of Political Economy at Oxford, showed that the density of

$$S_n^* = \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}}$$

could be approximated as

$$f_{S_n^*}(x) = \phi(x)\left\{1 + \frac{\rho_3 H_3(x)}{6\sqrt{n}} + \frac{3\rho_4 H_4(x) + \rho_3^2 H_6(x)}{72n}\right\} + O(n^{-3/2})$$

where ϕ is standard normal density and the omitted terms are $O(n^{-3/2})$ and H_r are *Hermite polynomials*

$$H_r(x) = (-1)^r \phi^{(r)}(x)/\phi(x).$$

For example, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, $H_6(x) = x^6 - 15x^4 + 45x^2 - 15$.

In terms of the original variable S we get

$$f_{S_n}(s) = \frac{e^{-x^2/2}}{\sigma\sqrt{2\pi n}}\left\{1 + \frac{\rho_3 H_3(x)}{6\sqrt{n}} + \frac{3\rho_4 H_4(x) + \rho_3^2 H_6(x)}{72n}\right\} + O(n^{-3/2}),$$

where $x = (s - n\mu)/(\sigma\sqrt{n})$.

Since $H_3(0) = 0$ this is particularly accurate when s is close to $n\mu$, as the first correction term then disappears.

If we wish a similar accuracy for other values of s we use the idea of *tilting* the distribution by shifting the log-density with a linear term, as we shall see next.

Associate an exponential family of densities with the originial density f as

$$f(x;\gamma) = f(x)e^{x\gamma - K(\gamma)},$$

where K is the cumulant generating function of f. Clearly, $f(x;0) = f(x)$. We say that $f(x;\gamma)$ is obtained by *tilting* f by γ.

If X_i have density $f(x;\gamma)$, the sum S_n has density

$$f_{S_n}(s;\gamma) = f_{S_n}(s)e^{s\gamma - nK(\gamma)},$$

implying that

$$f_{S_n}(s) = e^{nK(\gamma) - s\gamma} f_{S_n}(s;\gamma).$$

Since this equation holds for all γ *we can now choose γ freely to suit our purpose*.

If we use an Edgeworth expansion to approximate $f_{S_n}(s;\gamma)$ we can thus choose γ so that the expectation $\mathbf{E}_\gamma(S_n) = s$.

Since the mean of S_n in the tilted distribution is $nK'(\gamma)$ we should choose $nK'(\hat{\gamma}) = s$. As the variance of S_n in the tilted distribution is $nK''(\gamma)$, the resulting *saddle-point* approximation is

$$f_{S_n}(s) \approx e^{nK(\hat{\gamma})-s\hat{\gamma}} \frac{1}{\{2\pi nK''(\hat{\gamma})\}^{-1/2}},$$

which can be extremely accurate.

Note that the Edgeworth approximation uses a normal approximation around the *mean* of the distribution whereas Laplace's method uses its *mode*. The tilting technique can be useful in both cases.

Use this approximation for a natural exponential family with canonical parameter θ, i.e. with density

$$f(x;\theta) = b(x)e^{\theta x - c(\theta)}$$

then $K(t) = c(\theta + t) - c(\theta)$, $K'(t) = c'(\theta + t)$ and thus when

$$nK'(\hat{\gamma}) = nc'(\theta + \hat{\gamma}) = s$$

we have that $\hat{\gamma} = \hat{\theta} - \theta$, where $\hat{\theta}$ is the MLE, yielding

$$f_{S_n}(s;\theta) \approx e^{n\{K(\hat{\theta}) - K(\theta)\} - s(\hat{\theta} - \theta)} \frac{1}{\{2\pi n K''(\hat{\theta})\}^{-1/2}} \propto \frac{L(\theta)}{L(\hat{\theta})} |j(\hat{\theta})|^{-1/2}.$$

Since $nK'(\hat{\theta}) = s$ we have

$$\frac{\partial \hat{\theta}}{\partial s} = \frac{1}{nK''(\hat{\theta})} = \frac{1}{nj(\hat{\theta})}$$

so a change of variables leads to the following approximate formula for the density of the MLE

$$f(\hat{\theta}; \theta) \approx\propto \frac{L(\theta)}{L(\hat{\theta})} |j(\hat{\theta})|^{1/2}.$$

Similar methods can be used to show that, in wide generality, if A is ancillary so that $(\hat{\theta}, A)$ is minimal sufficient, then approximately, and quite often exactly,

$$f(\hat{\theta} \mid A = a; \theta) \approx\propto \frac{L(\theta)}{L(\hat{\theta})} |j(\hat{\theta})|^{1/2},$$

which is known as *Barndorff–Nielsen's formula.* Note that normalization constant may depend on θ and a.

Note similarity to the approximate Bayesian posterior:

$$\pi^*(\theta) \approx\propto \frac{L(\theta)}{L(\hat{\theta})} \left| j(\hat{\theta}) \right|^{1/2}$$

where we have ignored the contribution $\pi(\theta)/\pi(\hat{\theta})$ from the prior. Only the interpretations are different!