AMS 207: Intermediate Bayesian Modeling

2: Asymptotics (continued);
Multivariate Models

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
http://www.ams.ucsc.edu/~draper

© 2010 David Draper (all rights reserved)
Asymptotics (continued)

In part 1 of the lecture notes we saw, in parametric problems where the sampling model is \( p(y|\theta) \) — in which \( y = (y_1, \ldots, y_n) \), \( \theta = (\theta_1, \ldots, \theta_k) \) and the \( y_i \) are (conditionally) IID from \( p(y_i|\theta) \) — that for large \( n \), the repeated-sampling (frequentist) distribution of the MLE \( \hat{\theta}_{MLE} \) is approximately Gaussian:

\[
\hat{\theta}_{MLE} \sim_{RS} N_k(\theta, \hat{I}^{-1}), \tag{1}
\]

where \( \sim_{RS} \) means "is distributed as in repeated sampling," \( N_k \) is the \( k \)-dimensional Gaussian distribution, and the \( k \times k \) information matrix \( \hat{I} \) is minus the Hessian of the log likelihood — i.e., the \((i, j)\) entry of this matrix is

\[
\hat{I}_{ij} = - \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log l(\theta|y) \right]_{\hat{\theta}_{MLE}}, \tag{2}
\]

in which the likelihood function \( l(\theta|y) = c p(y|\theta) \) for any positive \( c \).

**Example:** Gaussian sampling model with unknown mean and variance: \( (y_i|\mu, \sigma^2) \) IID \( N(\mu, \sigma^2) \) \( (i = 1, \ldots, n) \), so that \( \theta = (\mu, \sigma^2) \).

Basic template for large-sample likelihood inference:

**Step 1:** the joint sampling distribution is

\[
p(y|\theta) = p(y|\mu, \sigma^2) = \prod_{i=1}^n p(y_i|\mu, \sigma^2)
= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ - \frac{(y_i - \mu)^2}{2\sigma^2} \right], \tag{3}
\]
Asymptotics (continued)

so the **likelihood function** is

\[
l(\theta | y) = l(\mu, \sigma^2 | y) = cp(y | \mu, \sigma^2) \\
= c \sigma^{-n} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \mu)^2 \right]
\]  

and the **log likelihood function** is

\[
ll(\theta | y) = ll(\mu, \sigma^2 | y) = c - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \mu)^2.
\]

**Step 2**: if the problem is **nice**, the MLE can be found by computing the **first partial derivatives** of \(ll(\mu, \sigma^2 | y)\) with respect to \(\mu\) and \(\sigma\), setting them to 0 and **solving** (here, "nice" means that the **maximum** occurs somewhere other than on a **boundary** of the **parameter space**):

\[
\frac{\partial}{\partial \mu} ll(\mu, \sigma^2 | y) = \frac{1}{\sigma^2} \sum_{i=1}^{n}(y_i - \mu) = 0 \quad \text{iff} \quad \mu = \hat{\mu}_{MLE} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i;
\]

\[
\frac{\partial}{\partial \sigma} ll(\mu, \sigma^2 | y) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n}(y_i - \mu) = 0 \quad \text{iff} \quad \sigma^2 = \hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}_{MLE})^2
\]

\[
= \frac{1}{n} \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{n-1}{n} s^2,
\]

where \(s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2\) is the usual (repeated-sampling-unbiased) **sample variance**.
Asymptotics (continued)

(Here you can readily check that these are the **global maxima** by verifying that the **second partial derivatives at the MLE** are **negative**.)

With the usual analysis-of-variance trick

\[
\sum_{i=1}^{n} (y_i - \mu)^2 = \sum_{i=1}^{n} [ (y_i - \bar{y}) + (\bar{y} - \mu) ]^2 \\
= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\bar{y} - \mu)^2 \\
= (n - 1)s^2 + n(\bar{y} - \mu)^2
\]  

we can now rewrite the **log likelihood function** as

\[
ll(\mu, \sigma^2 | y) = c - n \log \sigma - \frac{1}{2\sigma^2} \left[ (n - 1)s^2 + n(\bar{y} - \mu)^2 \right]
\]

and see directly (if we didn't know it already from the **MLEs**) that \((\bar{y}, s^2)\) are **sufficient** for \((\mu, \sigma^2)\) in this model.

Equation (8) makes it easy to **visualize** the **likelihood** and **log likelihood functions** in this problem with \(k = 2\):

```maple
>> ll := (mu, sigma, n, y_bar, s) -> -n * log(sigma) - 
   ((n - 1) * s^2 + n * (y_bar - mu)^2) / 
   (2 * sigma^2);

ll := (mu, sigma, n, y_bar, s) ->

\[
\frac{2}{2}
\]

\[
- n \log(\sigma) - \frac{1}{2} \frac{(n - 1)s^2 + n(y_bar - \mu)^2}{\sigma^2}
\]
```
Asymptotics (continued)

With $n = 10$ the plots look like this:

> plotsetup(x11);

> plot3d(ll(mu, sigma, 10, 0, 1), mu = -2 .. 2, sigma = 0.5 .. 3, axes = boxed);

> plot3d(exp(ll(mu, sigma, 10, 0, 1)), mu = -2 .. 2, sigma = 0.5 .. 3, axes = boxed);
Asymptotics (continued)

> plot( ll( mu, 1, 10, 0, 1 ), mu = -2 .. 2, 
   color = black );

> plot( exp( ll( mu, 1, 10, 0, 1 ) ), mu = -2 .. 2, 
   color = black );

> plot( ll( 0, sigma, 10, 0, 1 ), sigma = 0.5 .. 3, 
   color = black );

> plot( exp( ll( 0, sigma, 10, 0, 1 ) ), sigma = 0.5 .. 3, 
   color = black );
Asymptotics (continued)

With $n = 500$ the plots look like this:

```plaintext
> plotsetup( x11 );

> plot3d( ll( mu, sigma, 500, 0, 1 ), mu = -2 .. 2,
      sigma = 0.5 .. 3, axes = boxed );

> plot3d( exp( ll( mu, sigma, 500, 0, 1 ) ), mu = -2 .. 2,
      sigma = 0.5 .. 3, axes = boxed );
```
Asymptotics (continued)

```R
> plot( ll( mu, 1, 500, 0, 1 ), mu = -2 .. 2,
        color = black);

> plot( exp( ll( mu, 1, 500, 0, 1 ) ), mu = -2 .. 2,
        color = black);

> plot( ll( 0, sigma, 500, 0, 1 ), sigma = 0.5 .. 3,
        color = black);

> plot( exp( ll( 0, sigma, 500, 0, 1 ) ), sigma = 0.5 .. 3,
        color = black);
```
Asymptotics (continued)

We’re not actually directly checking the asymptotic behavior of the MLE in these plots (to do that we’d have to write a program that simulated the repeated-sampling distributional behavior of the MLE as a function of \( n \), which would empirically verify the frequentist Central Limit Theorem); what we’re noticing (and what we noticed in part 1 of the lecture notes) is that as \( n \) increases the likelihood function (interpreted in a Bayesian way, as a density) looks more and more Gaussian (which empirically verifies the Bayesian Central Limit Theorem (discussed below)).

But we ARE INDIRECTLY checking the asymptotic behavior of the MLE in these plots, because of a result called the Bernstein-von Mises Theorem, which informally says that in nice problems, frequentist and Bayesian numerical inferential conclusions in parametric models should closely agree when \( n \) is large.

To see why this must be true, consider an even simpler setting, a Gaussian sampling model with unknown mean:
\[
(y_i|\mu) \overset{iid}{\sim} N(\mu, \sigma^2) \quad (i = 1, \ldots, n)
\]

with \( \sigma \) known, so that \( \theta = \mu \).

The MLE for \( \mu \) in this model is of course just \( \hat{\mu}_{\text{MLE}} = \bar{y} \), whose repeated-sampling distribution is \( N(\mu, \frac{\sigma^2}{n}) \):

\[
\bar{y} \sim_{RS} N\left(\mu, \frac{\sigma^2}{n}\right).
\]  \hspace{1cm} (9)

You’ll recall from AMS 206 that the conjugate prior for \( \mu \) in this model is \( \mu \sim N(\mu_0, \sigma_0^2) \), and with this prior the posterior for \( \mu \) is Gaussian with mean given by the precision-weighted average of the prior and data means,

\[
E_B(\mu|y) = \frac{\frac{1}{\sigma_0^2} \mu_0 + \frac{n}{\sigma^2} \bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n},
\]  \hspace{1cm} (10)
Asymptotics (continued)

where \( n_0 = \frac{\sigma^2}{\sigma_0^2} \) is the prior sample size, and variance given by the reciprocal of the sum of the prior and data precisions,

\[
V_B(\mu | y) = \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} = \frac{\sigma^2}{n_0 + n}.
\]  

(11)

For large \( n \) the effect of the prior information is small, and the posterior distribution for \( \mu \) given the data is Gaussian with approximate posterior mean \( \bar{y} \) and approximate posterior variance \( \frac{\sigma^2}{n} \):

\[
(\mu | y) = (\mu | \bar{y}) \sim_B N\left( \bar{y}, \frac{\sigma^2}{n} \right).
\]  

(12)

In pictures, equations (9) and (12) look like this:

A 95% frequentist confidence interval for \( \mu \) will be \( \bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}} \), and a 95% Bayesian central posterior interval for \( \mu \) given the data will be the same; the reason this works is that the normal densities in the left and right plots above are

\[
\frac{\sqrt{n}}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{n}{2\sigma^2} (\bar{y} - \mu)^2 \right] \quad \text{and} \quad \frac{\sqrt{n}}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{n}{2\sigma^2} (\mu - \bar{y})^2 \right],
\]

(13)

and these are of course the same (but the left-hand plot is a normal curve for \( \bar{y} \) centered at \( \mu \), and the right-hand plot is the same normal curve for \( \mu \) centered at \( \bar{y} \)).
Asymptotics (continued)

Back to the main line of the large-sample likelihood analysis:

**Step 3**: compute the information matrix by evaluating the second partial derivatives of \( ll(\mu, \sigma^2 | y) \) at the MLE:

\[
ll(\mu, \sigma^2 | y) = c - n \log \sigma - \frac{1}{2\sigma^2} [(n - 1)s^2 + n(\bar{y} - \mu)^2];
\]

\[
\frac{\partial}{\partial \mu} ll(\mu, \sigma^2 | y) = \frac{n(\bar{y} - \mu)}{\sigma^2};
\]

\[
\frac{\partial^2}{\partial \mu^2} ll(\mu, \sigma^2 | y) = -\frac{n}{\sigma^2};
\]

\[
\frac{\partial^2}{\partial \mu \partial \sigma} ll(\mu, \sigma^2 | y) = -\frac{2n(\bar{y} - \mu)}{\sigma^3};
\]

\[
\frac{\partial}{\partial \sigma} ll(\mu, \sigma^2 | y) = -\frac{n}{\sigma} + \frac{(n - 1)s^2 + n(\bar{y} - \mu)^2}{\sigma^3};
\]

\[
\frac{\partial^2}{\partial \sigma^2} ll(\mu, \sigma^2 | y) = \frac{n}{\sigma^2} - \frac{3 [(n - 1)s^2 + n(\bar{y} - \mu)^2]}{\sigma^4}.
\]

Now \(- \left[ \frac{\partial^2}{\partial \mu^2} ll(\mu, \sigma^2 | y) \right]_{\hat{\theta}_{MLE}} = \frac{n}{\hat{\sigma}_{MLE}^2}, - \left[ \frac{\partial^2}{\partial \mu \partial \sigma} ll(\mu, \sigma^2 | y) \right]_{\hat{\theta}_{MLE}} = 0, \) and

\[- \left[ \frac{\partial^2}{\partial \sigma^2} ll(\mu, \sigma^2 | y) \right]_{\hat{\theta}_{MLE}} = \frac{2n}{\hat{\sigma}_{MLE}^2}, \) so the information matrix is

\[
\hat{I} = \begin{bmatrix}
\frac{n}{\hat{\sigma}_{MLE}^2} & 0 \\
0 & \frac{2n}{\hat{\sigma}_{MLE}^2}
\end{bmatrix}, \text{ and its inverse is evidently}
\]

\[
\hat{I}^{-1} = \begin{bmatrix}
\frac{1}{n} & 0 \\
0 & \frac{2n}{\hat{\sigma}_{MLE}^2}
\end{bmatrix}; \text{ thus for large } n
\]

\[
\hat{\theta}_{MLE} = \begin{pmatrix} \hat{\mu}_{MLE} \\ \hat{\sigma}_{MLE}^2 \end{pmatrix} \sim_{RS} N_2 \left\{ \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{bmatrix} \frac{\hat{\sigma}_{MLE}^2}{n} & 0 \\ 0 & \frac{\hat{\sigma}_{MLE}^2}{2n} \end{bmatrix} \right\}. \quad (15)
\]
Asymptotics (continued)

So large-sample approximate 95% confidence intervals for \( \mu \) and \( \sigma^2 \) in this model would be of the form \( \bar{y} \pm 1.96 \frac{\hat{\sigma}_{MLE}}{\sqrt{n}} \) and \( \hat{\sigma}_{MLE}^2 \pm 1.96 \frac{\hat{\sigma}_{MLE}}{\sqrt{2n}} \), respectively; with small \( n \) we know we can do better (the interval for \( \mu \) is based not on the Gaussian but on \( t_{n-1} \), and in this model \( \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \)).

Bayesian asymptotics. Same setup: in parametric problems where the sampling model is \( p(y|\theta) \) — in which \( y = (y_1, \ldots, y_n) \), \( \theta = (\theta_1, \ldots, \theta_k) \) and the \( y_i \) are (conditionally) IID from \( p(y_i|\theta) \) — informally (from the plots in parts 1 and 2 of the lecture notes and the Bernstein-von Mises story above) we know what happens for large \( n \):

(1) the likelihood function (viewed in a Bayesian way, as a density for \( \theta \)) looks more and more like a Gaussian with mode (and therefore mean) \( \hat{\theta}_{MLE} \) and covariance matrix \( \hat{I}^{-1} \);

(2) the prior information becomes negligible in relation to the likelihood information, meaning that it becomes locally close to constant in the region where the likelihood is appreciable; and therefore

(3) the posterior distribution becomes more and more like a \( k \)-dimensional Gaussian centered at \( \hat{\theta}_{MLE} \) (in both a mode and mean sense) with covariance matrix \( \hat{I}^{-1} \).

GCSR (Gelman et al.) sharpen this a bit (chapter 4) by making a Taylor expansion of the log posterior around the posterior mode \( \hat{\theta} \) (assumed to be in the interior of the parameter space):

\[
\log p(\theta|y) = \log p(\hat{\theta}|y) + (\theta - \hat{\theta}) \left[ \frac{d}{d\theta} \log p(\theta|y) \right]_{\theta=\hat{\theta}} + \\
\frac{1}{2} (\theta - \hat{\theta})' \left[ \frac{d^2}{d\theta^2} \log p(\theta|y) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \ldots .
\]
Asymptotics (continued)

Now by definition the log posterior has zero derivative at the posterior mode \( \hat{\theta} \), so the linear term in this Taylor expansion vanishes:

\[
\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2} (\theta - \hat{\theta})' \left[ \frac{d^2}{d\theta^2} \log p(\theta|y) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \ldots .
\]

They show in section 4.2 that the remainder (third- and higher-order) term is small relative to the quadratic term when \( \theta \) is close to \( \hat{\theta} \) and \( n \) is large; thinking of this as a function of \( \theta \), the first term is constant and the quadratic term is proportional to the logarithm of a Gaussian density in \( \theta \) with mean \( \hat{\theta} \) and covariance matrix \( \hat{I}^{-1} \) (because the derivative of the log posterior and the derivative of the log likelihood are approximately the same with large \( n \) (in that case, as noted above, the prior information is negligible in relation to the likelihood information)), so the result is that for large \( n \)

\[
(\theta|y) \sim_B N(\hat{\theta}, \hat{I}^{-1}). \tag{16}
\]

This is why likelihood methods can be useful to Bayesians — the MLE and the information matrix help in creating large-sample approximations to posterior distributions — and it also shows when Fisher's approach to inference is approximately reasonable (when the sample size is large and the MLE is not at a boundary of the parameter space).