Asymptotic properties of posterior

- Normal approximation to the posterior dist.
- When the sample size $n$ is large enough the posterior $p(\theta \mid y)$ is often unimodal and roughly symmetric. It can be approximated by normal centered at its mode.
- Consider the Taylor expansion of $\log[p(\theta \mid y)]$ at $\hat{\theta}$ (posterior mode) ($\theta$ may be a vector, $\hat{\theta}$ is in interior)
Asymptotic properties of posterior—

\[
\log p(\theta \mid y) = \log p(\hat{\theta} \mid y) + \frac{\partial}{\partial \theta} \log p(\hat{\theta} \mid y)(\theta - \hat{\theta}) \\
+ \frac{1}{2}(\theta - \hat{\theta})\left[\frac{\partial^2 \log p(\hat{\theta} \mid y)}{\partial \theta_i \partial \theta_j}\right](\theta - \hat{\theta}) + \ldots
\]

So, \( p(\theta \mid y) \sim N(\hat{\theta}, I_n(\hat{\theta})^{-1}) \) where

\[
l_n(\theta) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\theta \mid y).
\]
Asymptotic properties of posterior

b) Likelihood dominating the prior dist.
In practice \( y = (y_1, \ldots, y_n) \) a random sample of size \( n \).

\[
\log p(\theta | y) = \log p(\theta) + \sum \log p(y_i | \theta)
\]

\[
I_n(\theta) = - \sum \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\theta | y_i) \approx nI_1(\theta)
\]

As \( n \to \infty \), the posterior mode \( \hat{\theta} \) is \( \approx \hat{\theta}_M(MLE) \)
Thus, when \( n \to \infty \), \( p(\theta | y) \approx N(\hat{\theta}_{MLE}, I_n(\hat{\theta}_{MLE})^{-1}) \)
When \( n \) is small, the prior dist. is important in the posterior computation.
Example 1

\( y_1, \ldots, y_n \sim N(\mu, \sigma^2), (\mu, \sigma^2) : \) both unknown.

\( p(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \) or \( p(\mu, \log \sigma) \propto 1 \)

\( t = \log \sigma = \frac{1}{2} \log \sigma^2, \sigma^2 = e^{2t}, \frac{d\sigma^2}{dt} = 2e^{2t} \)

\[ J = \begin{vmatrix} 1 & 0 \\ 0 & 2e^{2t} \end{vmatrix} = 2e^{2t} \]

\( \Rightarrow p(\mu, t) = \frac{1}{e^{2t}}(2e^{2t}) = 2 \propto 1 \)

\( \Rightarrow \log p(\mu, t|y) = \text{const.} - n \cdot t - \frac{1}{2}e^{-2t}[(n - 1)S^2 + n(\bar{y} - \mu)^2] \)

\( \Rightarrow \begin{align*} 
\frac{\partial}{\partial \mu} \log p(\mu, t|y) &= \frac{n(\bar{y} - \mu)}{\sigma^2} = ne^{-2t}(\bar{y} - \mu) \\
\frac{\partial}{\partial t} \log p(\mu, t|y) &= -n + e^{-2t}[(n - 1)S^2 + n(\bar{y} - \mu)^2] 
\end{align*} \)
Example 1

The posterior mode is \((\hat{\mu}, \log \hat{\sigma}) = (\bar{y}, \frac{1}{2} \log(\frac{n-1}{n} S^2))\)

\[
\Rightarrow \begin{cases}
\frac{\partial^2}{\partial \mu^2} \log p(\mu, t | y) = -\frac{n}{\sigma^2} \\
\frac{\partial^2}{\partial \mu \partial t} \log p(\mu, t | y) = 0 \\
\frac{\partial^2}{\partial t^2} \log p(\mu, t | y) = -2e^{-2t}[(n - 1)S^2 + n(\bar{y} - \mu)^2]
\end{cases}
\]

\[
\Rightarrow I(\hat{\mu}, \log \hat{\sigma}) = \begin{pmatrix}
\frac{n}{\hat{\sigma}^2} & 0 \\
0 & 2n
\end{pmatrix}
\]

\[
\Rightarrow p(\mu, \log \sigma | y) \approx N \left( \left( \begin{array}{c} \bar{y} \\ \log \hat{\sigma} \end{array} \right), \left( \begin{array}{cc} \frac{\hat{\sigma}^2}{n} & 0 \\
0 & \frac{1}{2n} \end{array} \right) \right)
\]

Given \(y\), \(\mu\) and \(\log \sigma\) are indep.
Example 1

If we write back to the parameter $(\mu, \sigma^2)$: given $y$, $\mu$ and $\sigma^2$ are indep.,

$$
\begin{align*}
\mu &\sim N(\bar{y}, \frac{\sigma^2}{n}) \\
\sigma^2 &\sim N(\tilde{\sigma}^2, \frac{2\tilde{\sigma}^4}{n+2}), \tilde{\sigma}^2 = \frac{n}{n+2} \hat{\sigma}^2
\end{align*}
$$
Counter examples to the theorem

1) Under identified model and parameters
\[
\begin{pmatrix} u \\ v \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)
\]
only one of \( u \) or \( v \) is observed for each pair \( (u, v) \)
\( \Rightarrow \rho \) is nonidentified. Data contain no information on \( \rho \).
\( \Rightarrow \) posterior of \( \rho \) = prior of \( \rho \)

2) Number of parameters increasing with sample size
\( y_{ij} \sim N(\theta, \sigma^2), \ j = 1, 2, \ i = 1, \ldots, n \)
No consistent estimator of \( \theta_i \)
Counter examples to the theorem–

3) Aliasing (special case of underidentified parameters)

- Assume that $y_1, ..., y_n \ iid \sim p(y_i \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)$,

$$
= \lambda \Phi(y_i\mid\mu_1, \sigma_1^2) + (1 - \lambda)\Phi(y_i\mid\mu_2, \sigma_2^2)
$$

$$
= \lambda \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp \left\{ -\frac{1}{2\sigma_1^2} (y_i - \mu_1)^2 \right\}
$$

$$
+ (1 - \lambda) \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp \left\{ -\frac{1}{2\sigma_2^2} (y_i - \mu_2)^2 \right\}.
$$

- Replace $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)$ by $(\mu_2, \mu_1, \sigma_2^2, \sigma_1^2, 1 - \lambda)$, the likelihood remains the same.
- It is (50%, 50%) of two modes.
Counter examples to the theorem—

- Improper posterior.
- Unbounded likelihood
- No limit of likelihood
  \[ y_i \sim \text{Bernoulli}(p), \ p = \frac{1}{1 + e^{-(x-\mu)}} \]
  only \( x_i = 1, \)
  \[ L(\mu) = \frac{1}{1 + e^{-(x-\mu)}} = \frac{1}{1 + e^{\mu-x}} \]
  \( \hat{\mu} = -\infty? \)
- Prior dists. exclude the point of convergence
- Convergence to the edge of parameter space
  \[ y_1, \ldots, y_n \sim N(\theta, 1), \theta \geq 0 \]
  \( \theta = 0 \) is true value, \( \bar{y} \to 0 \)
- Tail of dist.
§4.4 Frequentist evaluation of Bayesian inference

Frequentist statistics provide a useful approach for evaluating the properties of Bayesian inference.

1. Large sample correspondence: asymptotic normality
2. Point estimation: consistency, efficiency, asymptotic unbiased
3. Confidence coverage.