Sketch of proof of asymptotic normality

$y_1, \dots, y_n$ iid from $p(y|\theta)$

$$\theta | y \sim p_n(\theta|y) = \frac{p(\theta) \cdot p_n(y|\theta)}{\int p(\theta') p_n(y|\theta') \, d\theta'}$$

↑ dimensionless

Now, define $\delta = \sqrt{n} \cdot (\theta - \hat{\theta}^*)$, which has posterior density

($\hat{\theta}^*$ ← ML estimate of θ)

$$p_n(\delta|y) = p_n\left(\theta^* + \frac{\delta}{\sqrt{n}} \,|\, y\right) \cdot \left(\frac{1}{\sqrt{n}}\right)^p$$

$$= \frac{p\left(\theta^* + \frac{\delta}{\sqrt{n}}\right) \cdot p_n\left(y|\theta^* + \frac{\delta}{\sqrt{n}}\right) / p_n(y|\theta^*)}{\int p\left(\theta^* + \frac{\delta}{\sqrt{n}}\right) \cdot \frac{p_n\left(y|\theta^* + \frac{\delta}{\sqrt{n}}\right)}{p_n(y|\theta^*)} \, d\delta}$$

Now

$$\log \frac{p_n\left(y|\theta^* + \frac{\delta}{\sqrt{n}}\right)}{p_n(y|\theta^*)} = \sum_{i=1}^{n} \left[\log p\left(y_i|\theta^* + \frac{\delta}{\sqrt{n}}\right) - \log p(y_i|\theta^*)\right]$$

Taylor expansion for $p \geq 2$:

$$h(x) = h(x_0) + \bar{h}(x_0)^T \cdot (x - x_0) + \frac{1}{2}(x - x_0)^T \bar{\bar{h}}(x_0) \cdot (x - x_0) + \dots$$

$\bar{h}$: vector of all 1st order partial derivatives

$\bar{\bar{h}}$: matrix of all 2nd order partial derivatives

okt 20-12:14

Taylor expansion centered at θ^*

$$\approx \sum_{i=1}^{n} \left[\log p(y_i|\theta^*) + \left(\frac{\partial \log p(y_i|\theta^*)}{\partial \theta}\right)^T \cdot \frac{\delta}{\sqrt{n}} + \frac{1}{2}\left(\frac{\delta}{\sqrt{n}}\right)^T \frac{\partial^2 \log p(y_i|\theta^*)}{\partial \theta \, \partial \theta'} \cdot \frac{\delta}{\sqrt{n}}\right.$$

↑ Drop terms of higher order than the 2nd

($\theta^* + \frac{\delta}{\sqrt{n}} - \theta^*$)

$$\left. - \left[\log p(y_i|\theta^*) + 0\right]\right]$$

↑ became $\theta^* - \theta^* = 0$

$$= \sum_{i=1}^{n} \left[\left(\frac{\partial \log p(y_i|\theta^*)}{\partial \theta}\right)^T \cdot \frac{\delta}{\sqrt{n}} + \frac{1}{2}\left(\frac{\delta}{\sqrt{n}}\right)^T \frac{\partial^2 \log p(y_i|\theta^*)}{\partial \theta \, \partial \theta'} \frac{\delta}{\sqrt{n}}\right]$$

$$= -\frac{1}{2} \delta^T J_n \delta \quad \text{where} \quad J_n = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log p(y_i|\theta^*)}{\partial \theta \, \partial \theta'}$$

↑ $\sum_{i=1}^{n}\left(\frac{\partial \log p(y_i|\theta^*)}{\partial \theta}\right)^T = 0$ because $\theta^* =$ ML-estimator

okt 20-12:27

Hence

$$p_n(\delta \mid y) \approx \frac{(\text{Something that does not depend on } \underline{n}) \cdot e^{-\frac{1}{2}\delta^T J_n \delta}}{\int (\text{Something that does not depend on } n) \cdot e^{-\frac{1}{2}\delta^T J_n \delta} \, d\delta}$$

$e^{-\frac{1}{2}\delta^T J_n \delta}$ — kernel of a Normal distr. with mean 0 and covariance matrix J_n^{-1}

$$\approx (2\pi)^{-p/2} \cdot |J_n|^{1/2} \cdot e^{-\frac{1}{2}\delta^T J_n \delta} = N_p(0, J_n^{-1})$$

↑ For θ close to θ^*

$$J_n \xrightarrow{p} J(\theta_0)$$

Hence

$$[\sqrt{n}(\theta - \theta^*) \mid y] \xrightarrow{d} N_p(0, J(\theta_0)^{-1})$$

$$\hookrightarrow \theta \mid y \approx N(\theta^*, \tfrac{1}{n} J(\theta_0)^{-1})$$

In practice, we approximate θ_0 by θ^*

$$\theta \mid y \approx N(\theta^*, \tfrac{1}{n} J(\theta^*)^{-1})$$

Remember frequentist result:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N_p(0, J(\theta_0)^{-1})$$

okt 20-12:35

Remarks

- Conditions: – The likelihood must be continuous as a function of θ
  – θ_0 cannot be on the boundary of the parameter space
  – The result assumes a fixed number of parameters!
- If the true, underlying data distribution $f(y)$ is included in the parametric family $p(y|\theta)$, i.e. $f(y) = p(y|\theta)$ for some $\theta = \theta_0$, then we have consistency: $p(y|\theta) \xrightarrow[n\to\infty]{}$ point mass at θ_0
- If $f(y) \neq p(y|\theta), \forall \theta$, as long as all the data come from $f(y)$, the asymptotic posterior normality result still holds
- Unidentified models, i.e. $p_n(y|\theta)$ has max for a rang of θ-values
  ↳ problematic
  – except if you have strong enough prior information

okt 20-12:43

Hierarchical models

Example

Data $y_1, \dots, y_{71}$

y_i: Number of rats with tumour in experiment i, which had n_i rats in total

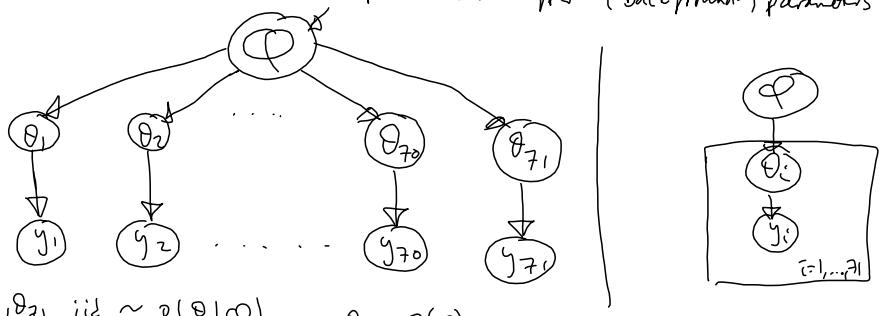
Reasonable sampling distr.: $y_i \sim Bin(n_i, \theta_i)$, $i = 1, \dots, 71$

(θ_i: Probability of tumour in experiment i)

$\theta_1, \dots, \theta_{71}$ are probably not identical

Same type of rats in each experiment, same kind of tumour considered

$\theta_1, \dots, \theta_{71}$ are likely to be "similar", in the sense that they come from the same distribution, which depends on hyper- (background-) parameters φ (vector)



$\theta_1, \dots, \theta_{71}$ iid $\sim p(\theta|\varphi)$, $\varphi \sim p(\varphi)$

okt 20-12:53

Advantages of this way of modelling over e.g. fitting $y_i \sim Bin(n_i, \theta_i)$ separately for each i

- Avoid overfitting
- "Borrow strength" for observational units with little data
- Allow for better predictions for new data

  ($\tilde{y}_{72} \sim Bin(\tilde{n}_{72}, \tilde{\theta}_{72})$ with $\tilde{\theta}_{72} \sim p(\tilde{\theta}_{72} | y_1, \dots, y_{71})$)

- ...

okt 20-13:18

A simple hierarchical normal model (a one-way random effects model) (≈ the model in Ch. 5.4)

Example: Concentration of aldrin at 3 depths of a river in the US, 10 measurements per depth, hence $J = 3$, $n_j = 10$, $\forall j$

$$y_{ij} \mid \theta_j, \sigma^2 \sim N(\theta_j, \sigma^2), \quad i = 1, \dots, n_j, \quad j = 1, \dots, J, \quad y = \{y_{ij}\}_{\forall i,j}$$

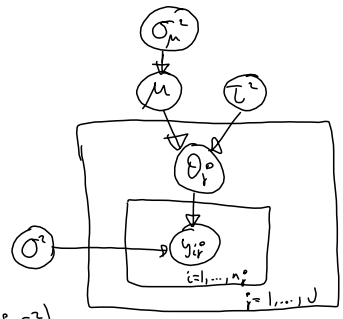
$$\theta_j \mid \mu, \tau^2 \sim N(\mu, \tau^2), \quad j = 1, \dots, J, \quad \theta = \{\theta_j\}_{j=1}^{J}$$

$$\mu \sim N(\mu_0, \sigma_\mu^2) \quad \leftarrow \text{Fixed value}$$

$$\sigma^2 \sim \text{Inverse-Gamma}(a_1, b_1)$$

$$\tau^2 \sim \text{Inverse-Gamma}(a_2, b_2)$$

$$\sigma_\mu^2 \sim \text{Inverse-Gamma}(a_3, b_3)$$



The joint posterior distribution:

$$p(\theta, \mu, \sigma^2, \tau^2, \sigma_\mu^2 \mid y) \propto p(\sigma_\mu^2) \cdot p(\tau^2) \cdot p(\sigma^2) \cdot p(\mu \mid \sigma_\mu^2) \cdot \prod_{j=1}^{J} p(\theta_j \mid \mu, \tau^2) \cdot \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij} \mid \theta_j, \sigma^2)$$

All we need in order to perform simple MCMC (Gibbs) are something called the full conditional distributions:

- for each parameter, the distribution conditional on all other parameters and all the data

Because of Bayes: $P(A \mid B) = \frac{P(A,B)}{P(B)}$

this is proportional to the joint posterior distribution, e.g.:

okt 20-13:22

$$p(\sigma_\mu^2 \mid \theta, \mu, \sigma^2, \tau^2, y) = \frac{p(\theta, \mu, \sigma^2, \tau^2, \sigma_\mu^2 \mid y)}{p(\theta, \mu, \sigma^2, \tau^2 \mid y)} \leftarrow \text{Constant w.r.t. } \sigma_\mu^2$$

$$\underset{\text{w.r.t } \sigma_\mu^2}{\propto} p(\theta, \mu, \sigma^2, \tau^2, \sigma_\mu^2 \mid y)$$

$$\underset{\text{w.r.t } \sigma_\mu^2}{\propto} p(\sigma_\mu^2)\, p(\mu \mid \sigma_\mu^2)$$

$$\propto (\sigma_\mu^2)^{-(a_3+1)} e^{-b_3/\sigma_\mu^2} \cdot \frac{1}{\sigma_\mu} \cdot \exp\left\{-\frac{1}{2\sigma_\mu^2}(\mu - \mu_0)^2\right\}$$

$$\propto \text{Inverse-Gamma}\left(a_3 + \tfrac{1}{2},\ b_3 + \tfrac{1}{2}(\mu - \mu_0)^2\right)$$

Similarly

$$p(\tau^2 \mid \theta, \mu, \sigma^2, \sigma_\mu^2, y) \propto (\tau^2)^{-(a_2+1)} e^{-b_2/\tau^2} \cdot \prod_{j=1}^{J} \frac{1}{\tau} \exp\left\{-\frac{1}{2\tau^2}(\theta_j - \mu)^2\right\}$$

$$\propto \text{Inv-Gamma}\left(a_2 + \frac{J}{2},\ b_2 + \frac{1}{2}\sum_{j=1}^{J}(\theta_j - \mu)^2\right\}$$

$$p(\sigma^2 \mid \theta, \mu, \tau^2, \sigma_\mu^2, y) = \text{Inv-Gamma}\left(a_1 + \tfrac{1}{2}\sum_{j=1}^{J} n_j,\ b_1 + \tfrac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2\right)$$

$$p(\mu \mid \theta, \sigma^2, \tau^2, \sigma_\mu^2, y) \propto p(\mu \mid \sigma_\mu^2) \cdot \prod_{j=1}^{J} p(\theta_j \mid \mu, \tau^2)$$

$$\propto N\left(\frac{\tau^2}{\tau^2 + J\sigma_\mu^2}\mu_0 + \frac{J\sigma_\mu^2}{\tau^2 + J\sigma_\mu^2} \cdot \bar{\theta},\ \frac{\sigma_\mu^2 \cdot \tau^2}{\tau^2 + J \cdot \sigma_\mu^2}\right)$$

← Can be found in a similar way that we did for the Normal model with one parameter (mean) θ (Ch. 2)

$$p(\theta_j \mid \theta_{-j}, \mu, \sigma^2, \tau^2, \sigma_\mu^2, y) \propto p(\theta_j \mid \mu, \tau^2) \cdot \prod_{i=1}^{n_j} p(y_{ij} \mid \theta_j, \sigma^2)$$

All θ_j's except θ_j

$$\propto N\left(\frac{\sigma^2}{\sigma^2 + n_j\tau^2}\mu + \frac{n_j\tau^2}{\sigma^2 + n_j\tau^2} \cdot \bar{y}_j,\ \frac{\sigma^2\tau^2}{\sigma^2 + n_j\tau^2}\right), \quad \bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

okt 20-13:42

For this model, we have local conjugacy for all the parameters and hence all full conditional distributions are known distributions

↳ Gibbs sampler very easy!

The MCMC algorithm called the Gibbs sampler is then simply

Initialization $\mu^{(0)}$, $\theta_j^{(0)}$, $j = 1, \dots, J$ (starting values)

For i in $1:M$ ← number of simulations

- Draw $\sigma_\mu^{2(i)}$ from $p(\sigma_\mu^2 \mid \theta^{(i-1)}, \mu^{(i-1)}, \sigma^{2(i-1)}, \tau^{2(i-1)}, y)$
- Draw $\tau^{2(i)}$ from $p(\tau^2 \mid \theta^{(i-1)}, \mu^{(i-1)}, \sigma^{2(i-1)}, \sigma_\mu^{2(i)}, y)$
- Draw $\sigma^{2(i)}$ from $p(\sigma^2 \mid \theta^{(i-1)}, \mu^{(i-1)}, \tau^{2(i)}, \sigma_\mu^{2(i)}, y)$
- Draw $\mu^{(i)}$ from $p(\mu \mid \theta^{(i-1)}, \sigma^{2(i)}, \tau^{2(i)}, \sigma_\mu^{2(i)}, y)$
- For $j = 1, \dots, J$ draw θ_j from $p(\theta_j \mid \theta_j^{(i-1)}, \mu^{(i)}, \sigma^{2(i)}, \tau^{2(i)}, \sigma_\mu^{2(i)})$

okt 20-13:59

Depending on how good the initial values are, it takes more or less time (simulation steps) to reach convergence, hence must disregard the first B simulated values for each parameter

↳ look at e.g. traceplots

Now, for the nice result (more on this later):

$\sigma_\mu^{2(i)}, \tau^{2(i)}, \sigma^{2(i)}, \mu^{(i)}, \theta^{(i)}$, $i = B+1, \dots, M$

are samples from the posterior distributions, and posterior summaries (approximate) can be calculated from them.

okt 20-14:25