Contents

1 Foundation

1.1 Review on MLE

Central limit theorem: Suppose $\{X_1, X_2, \ldots, X_n\}$ is a sequence of i.i.d. random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$ then as $n \to \infty$, $\bar{X} \to N(\mu, \frac{\sigma^2}{n})$
Score: The partial derivative with respect to θ of the natural logarithm of the likelihood function is called the score

$$Z = l' = \frac{\partial}{\partial \theta} \log f(X; \theta)$$

$$E(Z) = 0 \text{ and } Z \xrightarrow{d} N\left(0, I(\theta_0)\right)$$

under θ_0
Fisher information: The variance of the score is defined to be the Fisher information

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial \theta} \log f(X; \theta)\right)^2 \mid \theta\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \mid \theta\right]$$

Total information and observed total information

$$\mathbf{I}_T(\boldsymbol{\theta}) = -E\left\{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log L(\boldsymbol{\theta} \mid \mathbf{Y})\right\}$$

$$\mathbf{I}_T(\mathbf{Y}, \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log L(\boldsymbol{\theta} \mid \mathbf{Y})$$

Information estimation

- Version 1

$$\begin{aligned}\bar{\mathbf{I}}(\mathbf{Y}, \boldsymbol{\theta}) &= n^{-1} \sum_{i=1}^{n} \left\{-\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{s}\left(Y_i, \boldsymbol{\theta}\right)\right\} \\ &= n^{-1} \sum_{i=1}^{n} \left\{-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f\left(Y_i; \boldsymbol{\theta}\right)\right\}\end{aligned}$$

- Version 2

$$\begin{aligned}\bar{\mathbf{I}}^*(\mathbf{Y}, \boldsymbol{\theta}) &= n^{-1} \sum_{i=1}^{n} \mathbf{s}\left(Y_i, \boldsymbol{\theta}\right)^{\otimes 2} \\ &= n^{-1} \sum_{i=1}^{n} \left\{\frac{\partial}{\partial \boldsymbol{\theta}^T} \log f\left(Y_i; \boldsymbol{\theta}\right)\right\}^{\otimes 2}\end{aligned}$$

Property 1. If $\hat{\theta}$ is the MLE estimate of θ_0, then it has the following property:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \frac{1}{I(\theta_0)})$$

Ways for estimating MLE:

- Newton-Raphson algorithm

$$\begin{aligned}\mathbf{0} = \mathbf{S}(\boldsymbol{\theta}) &\approx \mathbf{S}\left(\boldsymbol{\theta}^{(\nu)}\right) + \left\{\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{S}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\nu)}}\right\}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(\nu)}\right) \\ &= \mathbf{S}\left(\boldsymbol{\theta}^{(\nu)}\right) - \mathbf{I}_T\left(\mathbf{Y}, \boldsymbol{\theta}^{(\nu)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(\nu)}\right) \\ \boldsymbol{\theta}^{(\nu+1)} &= \boldsymbol{\theta}^{(\nu)} + \mathbf{I}_T\left(\mathbf{Y}, \boldsymbol{\theta}^{(\nu)}\right)^{-1} \mathbf{S}\left(\boldsymbol{\theta}^{(\nu)}\right)\end{aligned}$$

1. start with initial $\theta^{(0)}$

2. update from current $\theta^{(v)}$ to obtain $\theta^{(v+1)}$
3. stop if $\left\|\mathbf{S}\left(\boldsymbol{\theta}^{(\nu+1)}\right)\right\|$ or $\left\|\boldsymbol{\theta}^{(\nu+1)}-\boldsymbol{\theta}^{(\nu)}\right\|$ is sufficiently small

Note: Newton-Raphson estimator has local quadratic convergence property

$$\left\|\boldsymbol{\theta}^{(\nu+1)}-\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right\| \leq c\left\|\boldsymbol{\theta}^{(\nu)}-\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right\|^{2} \text{ for some } c>0$$

This property exists for the situation where the estimator is very close to the true parameter. Thus, at least near the solution, convergence is fast for a Newton method.
More specifically, the local quadratic convergence holds under the following conditions

1. $I_T(\mathbf{Y},\theta)\neq 0$ in a neighborhood of $\hat{\theta}_{\text{MLE}}$
2. $S''(\theta)$ is bounded
3. $\theta^{(\nu)}$ is sufficiently close to $\hat{\theta}_{\text{MLE}}$

Proof draft:

$$\begin{aligned}0=S\left(\widehat{\theta}_{\mathrm{MLE}}\right) =& S\left(\theta^{(\nu)}\right)-I_T\left(\mathbf{Y},\theta^{(\nu)}\right)\left(\widehat{\theta}_{\mathrm{MLE}}-\theta^{(\nu)}\right)\\ &+\frac{1}{2}S''\left(\vartheta^{(\nu)}\right)\left(\widehat{\theta}_{\mathrm{MLE}}-\theta^{(\nu)}\right)^{2}\\ \widehat{\theta}_{\mathrm{MLE}}-\theta^{(\nu+1)} =& \frac{1}{2}\frac{S''\left(\vartheta^{(\nu)}\right)}{I_T\left(\mathbf{Y},\theta^{(\nu)}\right)}\left(\widehat{\theta}_{\mathrm{MLE}}-\theta^{(\nu)}\right)^{2}\end{aligned}$$

where $\vartheta^{(\nu)}=\theta^{(\nu)}+k*(\hat{\theta_{\text{MLE}}}-\theta^{(\nu)}),\ k\in[0,1]$. This uses the Taylor expansion with Lagrange remainder. *One-step estimator*: typically, $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}-\boldsymbol{\theta}=O_p\left(n^{-1/2}\right)$. If one starts with $\theta^{(0)}$ such that $\boldsymbol{\theta}^{(0)}-\boldsymbol{\theta}=O_p\left(n^{-1/2}\right)$, then

$$\boldsymbol{\theta}^{(1)}-\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}=O_p\left(n^{-1}\right)$$

under regularity conditions. That is $\theta^{(0)}$ and $\hat{\theta_{\text{MLE}}}$ is asymptotically equivalent.

- Fisher scoring algorithm
  $\mathbf{I}_T\left(\mathbf{Y},\boldsymbol{\theta}^{(\nu)}\right)$ replaced by its expectation $\mathbf{I}_T\left(\boldsymbol{\theta}^{(\nu)}\right)$, which is

$$\boldsymbol{\theta}^{(\nu+1)}=\boldsymbol{\theta}^{(\nu)}+\mathbf{I}_T\left(\boldsymbol{\theta}^{(\nu)}\right)\mathbf{S}\left(\boldsymbol{\theta}^{(\nu)}\right)$$

- EM algorithm
  The basic idea of the EM Algorithm is to view the observed data Y as incomplete, that somehow there is missing data Z that would make the problem simpler if we had it. In some cases Z could truly be missing data, but in others it is just additional data that we wish we had.

  1. The first step is to write down the joint likelihood of the "complete" data $(\boldsymbol{Y},\boldsymbol{Z})$, call it $L_{\mathrm{C}}(\boldsymbol{\theta}\mid\boldsymbol{Y},\boldsymbol{Z})$. The "E" step of the EM Algorithm is to compute the conditional expectation of $\log L_{\mathrm{C}}(\boldsymbol{\theta}\mid\boldsymbol{Y},\boldsymbol{Z})$ given $\boldsymbol{Y}$ assuming the true parameter value is $\boldsymbol{\theta}^{(\nu)}$

$$\begin{aligned}Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(\nu)},\boldsymbol{Y}\right) &= \mathrm{E}_{\boldsymbol{\theta}^{(v)}}\left\{\log L_{\mathrm{C}}(\boldsymbol{\theta}\mid\boldsymbol{Y},\boldsymbol{Z})\mid\boldsymbol{Y}\right\}\\ &= \int\log L_{\mathrm{C}}(\boldsymbol{\theta}\mid\boldsymbol{Y},\boldsymbol{z})f_{Z|Y}\left(\boldsymbol{z}\mid\boldsymbol{Y},\boldsymbol{\theta}^{(v)}\right)dz\end{aligned}$$

  2. M step: calculate $\boldsymbol{\theta}^{(v+1)}$ that maximizes $Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(\nu)},\boldsymbol{Y}\right)$ wrt θ
     The plausibility of EM

     Lemma 1.1. *For* $\mathbf{Y}\sim f(\mathbf{y};\boldsymbol{\theta}_0)$, $E_{\boldsymbol{\theta}_0}\{\log f(\mathbf{Y};\boldsymbol{\theta})\}$ *is maximized at* $\boldsymbol{\theta}=\boldsymbol{\theta}_0$

     Proof: Note that $\psi(x)=-\log(x)$ is a convex function for $x\in(0,\infty)$, since $\psi''(x)=x^{-2}>0$. Then by Jensen's inequality we have

$$-\log\left[E_{\boldsymbol{\theta}_0}\left\{\frac{f(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y};\boldsymbol{\theta}_0)}\right\}\right]\leq -E_{\boldsymbol{\theta}_0}\left[\log\left\{\frac{f(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y};\boldsymbol{\theta}_0)}\right\}\right]$$

     and since $E_{\boldsymbol{\theta}_0}\left\{\frac{f(y;\boldsymbol{\theta})}{f(y;\boldsymbol{\theta}_0)}\right\}=\int\frac{f(y;\boldsymbol{\theta})}{f(y;\boldsymbol{\theta}_0)}f(y;\boldsymbol{\theta}_0)\,dy=1$ we get $E_{\boldsymbol{\theta}_0}\left[\log\left\{\frac{f(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y};\boldsymbol{\theta}_0)}\right\}\right]\leq 0$

1.2 Review on GLM

some materials: on two forms, canonical forms, GLM and exponential family,
Exponential family: Y with distribution with the density of the form

$$\begin{aligned} Y &\sim f(y;\theta,\phi) \\ &= \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \end{aligned}$$

where

- θ is the canonical (natural) parameter – often the parameter of interest
- Φ is dispersion (scale) parameter – often nuisance parameter

Exponential family also has **another** commonly used form A good material here

$$p(x \mid \eta) = h(x)\exp\left\{\eta^T T(x) - A(\eta)\right\}$$

parameter vector η, often referred to as the *canonical parameter* for given functions T and h. The statistic T(X) is referred to as a *sufficient statistic*. The function $A(\eta)$ is known as the *cumulant function* and

$$A(\eta) = \log \int h(x)\exp\left\{\eta^T T(x)\right\}\nu(dx)$$

with respect to the measure ν
For example, for normal distribution,

$$\begin{aligned} f\left(y;\mu,\sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\} \end{aligned}$$

then

- $\theta = \mu$
- $\phi = \sigma^2$
- $a(\phi) = \phi$
- $b(\theta) = \frac{\theta^2}{2}$
- $c(y,\phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$

or

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

then

$$\begin{aligned} \eta &= \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \\ T(x) &= \begin{bmatrix} x \\ x^2 \end{bmatrix} \\ A(\eta) &= \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(-2\eta_2\right) \\ h(x) &= \frac{1}{\sqrt{2\pi}} \end{aligned}$$

We also have the following property

Property 1.

$$\begin{aligned} E(Y) &= b'(\theta) \\ \mathrm{Var}(Y) &= b''(\theta)a(\phi) \end{aligned}$$

Proof

$$\ell(\theta) = \log f(Y;\theta,\phi) = \frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)$$

Then the score function for θ is

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \frac{Y - b'(\theta)}{a(\phi)}$$

Since $E(U(\theta)) = 0$, we have

$$E\left(\frac{Y - b'(\theta)}{a(\phi)}\right) = 0 \Longrightarrow E(Y) = b'(\theta)$$

Since $E(U(\theta)) = 0$, we have

$$\mathrm{Var}(U(\theta)) = E\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right) \Longrightarrow \frac{\mathrm{Var}(Y)}{a^2(\phi)} = \frac{b''(\theta)}{a(\phi)} \Longrightarrow \mathrm{Var}(Y) = b''(\theta)a(\phi)$$

1.2.1 Newly introduced in 711

Exponential family distributions $\{\mathcal{F}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Omega\}$ have densities of the general form

$$f(x;\boldsymbol{\theta}) = h(x)\exp\left\{\sum_{i=1}^{s} g_i(\boldsymbol{\theta})T_i(x) - B(\boldsymbol{\theta})\right\}$$

this representation is not unique. A version with the smallest s is called *minimal exponential family* under this case both $T_i's$ and $g_i\theta's$ are free of linear constraints, which means

$$\sum_{i=1}^{s} c_i T_i(x) = c_{s+1} \text{ a.e. } \nu \Longrightarrow c_i = 0, i = 1, \cdots, s+1$$

With minimal exponential family $(T_1, \cdots, T_s)$ is minimal sufficient. *Natural parameter space*

$$\Omega_0 = \left\{\boldsymbol{\theta} : \int h(x)\exp\left\{\sum_{i=1}^{s} g_i(\boldsymbol{\theta})T_i(x)\right\} d\nu(x) < \infty\right\}$$

Canonical representation

$$f(x;\eta) = h(x)\exp\left\{\sum_{i=1}^{s} \eta_i T_i(x) - A(\boldsymbol{\eta})\right\}$$

by integration equal to 1 we have

$$\int h(x)\exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} d\nu(x) = \exp\left[A(\eta)\right]$$

with the canonical parameter $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ and accordingly

$$\Omega_\eta = \left\{\eta : \int h(x)\exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} d\nu(x) < \infty\right\}$$

is another form of the natural parameter space.
If η is in the interior of Ω_η then the m.g.f of T exists in a neighborhood of $u = 0$

$$\begin{aligned}
M_{\mathbf{T}}(\mathbf{u}) &= Ee^{u_1T_1 + \cdots + u_sT_s} \\
&= \int \exp\{\sum u_iT_i\} \cdot h(x)\exp\left\{\sum \eta_iT_i - A(\eta)\right\} dv(x) \\
&= \int h(x)\exp\left\{\sum (\eta_i + u_i)\, T_i - A(\eta)\right\} dv(x) \\
&= \int h(x)\frac{\exp\{\sum(\eta_i + u_i)T_i\}}{\exp\{A(\eta)\}} dv(x) \\
&= \frac{\exp\{A(\eta + u)\}}{\exp\{A(\eta)\}} = e^{A(\boldsymbol{\eta}+\mathbf{u}) - A(\boldsymbol{\eta})}
\end{aligned}$$

$$K_{\mathbf{T}}(\mathbf{u}) = \log\{M_{\mathbf{T}}(\mathbf{u})\} = A(\boldsymbol{\eta}+\mathbf{u}) - A(\boldsymbol{\eta})$$

therefore

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} = E\{\mathbf{T}(X)\}, \quad \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta \partial \eta^T} = \mathrm{Var}\{\mathbf{T}(X)\}$$

also

$$\mathbf{I}(\boldsymbol{\eta}) = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} = \mathrm{Var}\{\mathbf{T}(X)\}$$

curved exponential family: $\dim(\eta) > \dim(\theta)$
Completeness: a set of statistics $\mathbf{T} = (T_1, \cdots, T_s)$ is complete with respect to the family of their induced distributions indexed by θ if there are no functions $\Phi(T)$ other than $\Phi = 0$ such that $E_{\boldsymbol{\theta}}\phi(\mathbf{T}) = 0$ for all $\theta \in \Omega$

$$\text{complete sufficient statistics} \overset{\rightarrow}{\nleftarrow} \text{minimal sufficient statistics}$$

Lehmann–Scheffe theorem: If a statistic is unbiased, complete and sufficient, then it is the unique best unbiased estimator.
Sufficient statistic $(T_1, \cdots, T_s)$ of a canonical exponential family is complete sufficient provided that the family is minimal and that the parameter space contains an s-dimensional rectangle (which ruled out the curve exponential family).
Distribution of sufficient statistics

$$\begin{aligned} &X \sim \text{ exponential family with } \boldsymbol{\eta} = (\boldsymbol{\zeta}, \boldsymbol{\psi}): \\ &f(x; \boldsymbol{\zeta}, \boldsymbol{\psi}) = h(x) \exp\left\{\textstyle\sum_{i=1}^{r} \zeta_i U_i(x) + \sum_{j=1}^{s} \psi_j T_j(x) - A(\boldsymbol{\zeta}, \boldsymbol{\psi})\right\} \end{aligned}$$

then

1. the distribution of (U, T) is an exponential family
2. marginally T is an exponential family

$$f(\mathbf{t}; \boldsymbol{\zeta}, \boldsymbol{\psi}) = q(\mathbf{t}) C(\boldsymbol{\zeta}) \exp\left\{\sum_{j=1}^{s} \psi_j t_j - A(\boldsymbol{\zeta}, \boldsymbol{\psi})\right\}$$

3. conditional distribution of U given $T = t$ is an exponential family

$$f(\mathbf{u} \mid \mathbf{t}; \boldsymbol{\zeta}) = q_{\mathbf{t}}(\mathbf{u}) \exp\left\{\sum_{i=1}^{r} \zeta_i u_i - A_{\mathbf{t}}(\boldsymbol{\zeta})\right\}$$

1.3 Families of distributions, moments / cumulants, and quantiles / percentiles

Families of distributions

- **Parametric** model or parametric family: a set of distributions indexed by a finite dimensional parameter

$$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_b)^T, \quad \{F(y; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

- **Semiparametric** model: a set of distributions indexed by a finite dimensional θ and some infinite dimensional parameters
  for example

  $\{f(y; \mu, f_0) = f_0(y - \mu) : -\infty < \mu < \infty, f_0 \in C$, where C is a class of continuous unimodeal densities$\}$

  where μ belongs to the parametric part and the f_0 belongs to the nonparametric part

- **Nonparametric** model: a set of distributions indexed by infinite dimensional parameters
  for example

$$Y_i = g(X_i) + e_i, \quad \text{density of } e \in C$$

  where $g(.)$ belongs to a class of functions with certain smoothness

- **jth population moment** of random variable Y

$$\mu_j' = E\left(Y^j\right), j = 1, 2, \cdots$$

  with its sample counterparts as

$$m_j' = n^{-1} \sum_{i=1}^{n} Y_i^j$$

- **jth population central moment** of rv Y

$$\mu_j = E\left[\{Y - E(Y)\}^j\right]$$

with its sample counterparts as

$$m_j = n^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^j$$

- **moment generating function** (MGF)

$$m(t) = E\{\exp(tY)\}$$

a useful trick is to write the MGF as

$$\begin{aligned} M(t) = E\left(e^{tX}\right) &= E\left(1 + tX + \cdots + t^r X^r/r! + \cdots\right) \\ &= \sum_{r=0}^{\infty}\mu_r t^r/r! \end{aligned}$$

which implies that

$$\mu_r' = M^{(r)}(0)$$

- **cumulant generating function** (CGF)

$$k(t) = \log\{m(t)\}$$

it's useful to remember that

$$K(t) = \log M(t) = \sum_r \kappa_r t^r/r!$$

which is just from the Taylor expansion for $K(t)$ (note: the Taylor expansion around 0 is $f(x) = \sum_{n=0}^{\infty}\frac{f^{(n)}(0)}{n!}x^n$) a nice material on this topic

- **jth population cumulant**

$$\kappa_j = k^{(j)}(0)$$

for example

$$\begin{aligned} \kappa_1 &= \mu_1'(\text{ mean }) \\ \kappa_2 &= \mu_2' - \kappa_1^2(\text{ variance }) \end{aligned}$$

proof:

$$\begin{aligned} \kappa_1 = \kappa'(0) &= \left.\frac{m'(t)}{m(t)}\right|_{t=0} \\ &= \left.\frac{E[\exp(tY)Y]}{E[\exp(tY)]}\right|_{t=0} \\ &= E[Y] = \mu_1' \end{aligned}$$

and

$$\begin{aligned} \kappa''(t) &= \left.\frac{m''(t)m(t) - [m'(t)]^2}{m^2(t)}\right|_{t=0} \\ &= E(Y^2) - E(Y)^2 = \mu_2' - \kappa_1^2 \end{aligned}$$

- **skewness**

$$\text{skew } = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \mu_3/\mu_2^{3/2}$$

which measures symmetry. Note that we are using the central moment in the definition

- **kurtosis**

$$\text{kurt } = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^2} = \mu_4/\mu_2^2$$

which measures tail mass. *Note* that for Normal distribution $\text{skew } = 0,\ \text{kurt } = 3$

- **coefficient of variation**

$$\mathrm{CV} = \frac{\sigma}{\mu} = \mu_2^{1/2}/\mu$$

for non-negative random variable, providing a relative standard deviation

- **quantile**

$$\eta_p \equiv \inf\{x : F(x) \geq p\} = F^{-1}(p)$$

the last equality holds only when F is a strictly monotone function and X is continuous

1.4 Transformation of random variables

Jacobian method: Let $f_{X_1X_2}(x_1x_2)$ be the value of the joint probability density of the continuous random variables X_1 and X_2 at (x_1, x_2). If the functions given by $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ are partially differentiable with respect to x_1 and x_2 and represent a one-to-one transformation for all values within the range of X_1 and X_2 for which $f_{X_1X_2}(x_1x_2) \neq 0$, then, for these values of x_1 and x_2, the equations $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ can be uniquely solved for x_1 and x_2 to give $x_1 = w_1(y_1, y_2)$ and $x_2 = w_2(y_1, y_2)$ and for corresponding values of y_1 and y_2, the joint probability density of $Y_1 = u_1(X_1,\ X_2)$ and $Y_2 = u_2(X_1, X_2)$ is given by

$$f_{Y_1Y_2}(y_1y_2) = f_{X_1X_2}[w_1(y_1, y_2), w_2(y_1y_2)] \cdot |J|$$

where J is the Jacobian of the transformation

$$J = \begin{vmatrix} \frac{\partial x_1}{y_1} & \frac{\partial x_1}{y_2} \\ \frac{\partial x_2}{y_1} & \frac{\partial x_2}{y_2} \end{vmatrix}$$

all other points of $f_{Y_1Y_2}(y_1y_2) = 0$
Distribution function method: suppose we have the distribution function for Y as $F_Y(y;\theta) = P(Y \leq y)$. But we want the distribution function of $X = g(Y)$. Then

$$F_X(x;\theta) = P\{X \leq x\} = P\{g(Y) \leq x\}$$

if g is a strictly increasing function which means g^{-1} exists then

$$F_X(x;\boldsymbol{\theta}) = P\{g(Y) \leq x\} = P\left\{Y \leq g^{-1}(x)\right\} = F_Y\left\{g^{-1}(x);\boldsymbol{\theta}\right\}$$

if functions are differentiable then

$$f_X(x;\theta) = f_Y\left\{g^{-1}(x);\theta\right\}\frac{dg^{-1}(x)}{dx}$$

other transformation methods

1.5 Others

Theorem 1.2. ***Singular value decomposition***
*Suppose A is an n*p matrix of rank r, where $r \leq \min(n,p)$. There exists orthogonal matrices $U_{p\times p}$ and $V_{n\times n}$ such that*

$$V'AU = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix} \longrightarrow A = VDU', \text{ where } D = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix}$$

*where $\Delta = \operatorname{diag}(\delta_1, \ldots, \delta_r)$ is an r*r diggonal matrix with $\delta_1 \geq \delta_2 \ldots \geq \delta_r > 0$. The δ_i are called the singular values of A.*

2 Likelihood construction and estimation

2.1 Introduction

Likelihood is the joint density of the observed data.
For example

- in models for censored or missing data, the likelihood is not the density of the so-called "complete" data that includes the censored or missing values. Rather it is the density of only those components of the data that are observed and used in the statistical analysis
- consider an iid sample $Y_1, \ldots, Y_n$, and a parametric transformation $h(y,\alpha)$ strictly increasing in y for each α. The model assumption is $h(Y_1,\alpha), \ldots, h(Y_n,\alpha)$ are iid with common density $f(y;\boldsymbol{\theta})$ and distribution function $F(y;\boldsymbol{\theta})$, where α and θ are both parameters in the model. To construct the likelihood, we need to find the likelihood of the observed data Y rather than $h(Y)$. The distribution of Y_i is (we can also use Jacobian method here)

  $$P\{Y_i \leq y\} = P\{h(Y_i,\alpha) \leq h(y,\alpha)\} = F\{h(y,\alpha);\theta\}$$

  taking derivative wrt y

  $$f_Y(y;\boldsymbol{\theta},\alpha) = f\{h(y,\alpha);\boldsymbol{\theta}\}\frac{\partial h(y,\alpha)}{\partial y}$$

thus the likelihood is

$$L(\boldsymbol{\theta}, \alpha; \boldsymbol{Y}) = \prod_{i=1}^{n} f\left\{h\left(Y_i, \alpha\right); \boldsymbol{\theta}\right\} \left\{ \left. \frac{\partial h(y, \alpha)}{\partial y} \right|_{y=Y_i} \right\}$$

a common mistake is missing the $\left\{ \left. \frac{\partial h(y,\alpha)}{\partial y} \right|_{y=Y_i} \right\}$ part here

$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ of size $b \times 1$ is derived by maximizing

$$\ell(\boldsymbol{\theta}) = \log\{L(\boldsymbol{\theta} \mid \mathbf{Y})\}$$

under differentiable assumptions we can also solve for

$$\mathbf{S}(\boldsymbol{\theta}) = \{\ell'(\boldsymbol{\theta})\}^T$$

in vector calculus notation

$$\ell'(\boldsymbol{\theta}) = \frac{\partial \ell}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \ell}{\partial \theta_1}, \cdots, \frac{\partial \ell}{\partial \theta_b}\right)$$

$$\ell'(\boldsymbol{\theta})^T = \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right)^T = \frac{\partial \ell}{\partial \boldsymbol{\theta}^T} = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_b} \end{pmatrix}$$

and

$$\ell''(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\right) = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{pmatrix} \frac{\partial^2 \ell}{\partial^2 \theta_1} & \cdots & \frac{\partial^2 \ell}{\partial \theta_b \partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_b} & \cdots & \frac{\partial^2 \ell}{\partial^2 \theta_b} \end{pmatrix} = \left(\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_i}\right)_{i=1,\cdots,b, j=1,\cdots,b}$$

2.2 Likelihood construction

- For discrete IID case, multinomial distribution is widely used cause
  1. any discrete distribution with a finite support $\Longrightarrow$ multinomial model
  2. any discrete distribution with an infinite support (e.g. Poisson) $\xRightarrow{grouped}$ multinomial model

  $$N \sim \text{binomial}(n, p),\ f(N; p) = \begin{pmatrix} n \\ N \end{pmatrix} p^N (1-p)^{n-N}$$

  more generally

  $$(N_1, \cdots, N_k) \sim \text{multinomial}\left(n; p_1, \cdots, p_k\right), k \geq 2,\ f\left(N_1, \cdots, N_k; p_1, \cdots, p_k\right) = \frac{n!}{N_1! \cdots N_k!} p_1^{N_1} \cdots p_k^{N_k}$$

  where $\sum_{i=1}^k p_i = 1, \sum_{i=1}^k N_i = n$, and the model can be interpreted as tossing n balls into k urns. Also

  $$\begin{aligned} E\left(N_i\right) &= np_i \\ \text{var}\left(N_i\right) &= np_i\left(1-p_i\right) \\ \text{cov}\left(N_i, N_j\right) &= -np_i p_j, \quad i \neq j \end{aligned}$$

  hint: to prove the covariance, we can write $N_i = \sum_k^n M_{ki}$ where M_{ki} means toss k-th ball into i-th urn.
- Continuous IID case: skipped (nothing very important here)
- Connection Between Discrete and Continuous Likelihoods:$2h$**-method**
  1. basic formula for continuous variable:

  $$f(y) = \lim_{h \to 0+} \frac{F(y+h) - F(y-h)}{2h} = \lim_{h \to 0+} \frac{P(Y \in (y-h, y+h])}{2h}$$

  2. bivariate data with both X and Y being continuous

  $$\begin{aligned} f_{X,Y}(x,y) = \lim_{h \to 0+} (2h)^{-2} \{ & F_{X,Y}(x+h, y+h) - F_{X,Y}(x-h, y+h) \\ & - F_{X,Y}(x+h, y-h) + F_{X,Y}(x-h, y-h) \\ = \lim_{h \to 0+} & \frac{P(X \in (x-h, x+h], Y \in (y-h, y+h])}{(2h)^2} \end{aligned}$$

3. bivaraite data with X discrete and Y continuous

$$f_{X,Y}(x,y) = \lim_{h\to 0+} \frac{P(X \in (x-h, x+h], Y \in (y-h, y+h])}{2h} = \lim_{h\to 0+} \frac{P(X = x, Y \in (y-h, y+h])}{2h}$$

note: it is $2h$ in the denominator

The likelihood based on the $2h$-method can be summarized as follows

(a) for continuous case

$$\begin{aligned} L(\boldsymbol{\theta} \mid \mathbf{Y}) &= \prod_{i=1}^{n} f\left(Y_i; \boldsymbol{\theta}\right) \\ &= \prod_{i=1}^{n} \lim_{h\to 0+} \frac{F\left(Y_i + h; \boldsymbol{\theta}\right) - F\left(Y_i - h; \boldsymbol{\theta}\right)}{2h} \\ &= \lim_{h\to 0+} \left(\frac{1}{2h}\right)^n \prod_{i=1}^{n} \left\{F\left(Y_i + h; \boldsymbol{\theta}\right) - F\left(Y_i - h; \boldsymbol{\theta}\right)\right\} \end{aligned}$$

(b) for discrete case

$$\begin{aligned} \lim_{h\to 0^+} \prod_{i=1}^{n} \left\{F\left(Y_i + h; \boldsymbol{\theta}\right) - F\left(Y_i - h; \boldsymbol{\theta}\right)\right\} &= \prod_{i=1}^{n} \left\{F\left(Y_i^+; \boldsymbol{\theta}\right) - F\left(Y_i^-; \boldsymbol{\theta}\right)\right\} \\ &= \prod_{i=1}^{n} f\left(Y_i; \boldsymbol{\theta}\right) \\ &= L(\boldsymbol{\theta} \mid \boldsymbol{Y}) \end{aligned}$$

(c) for combination of continuous and discrete random variables

$$L(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \lim_{h\to 0^+} \left(\frac{1}{2h}\right)^m \prod_{i=1}^{n} \left\{F_i\left(Y_i + h; \boldsymbol{\theta}\right) - F_i\left(Y_i - h; \boldsymbol{\theta}\right)\right\}$$

where $1 \le m \le n$ depends on the number of continuous components in the data.

2.3 Proportional likelihoods

Likelihoods are equivalent if they are proportional and the constant of proportionality does not depend on unknown parameters.

- transformation of variables

  $$Y_1, \cdots, Y_n \sim \text{ iid } f_Y(y; \boldsymbol{\theta})$$

  we'd like to focus on transformed data $X_i = g\left(Y_i\right), g(\cdot)$ known, increasing, and continuously differentiable

  $$f_X(x; \boldsymbol{\theta}) = f_Y\{h(x); \boldsymbol{\theta}\} h'(x), \quad h(\cdot) = g^{-1}(\cdot)$$

  $$L(\boldsymbol{\theta} \mid \mathbf{X}) = L(\boldsymbol{\theta} \mid \mathbf{Y}) \prod_{i=1}^{n} h'\left(X_i\right) = L(\boldsymbol{\theta} \mid \mathbf{Y}) \prod_{i=1}^{n} \frac{1}{g'\left(Y_i\right)}$$

  note that $\prod_{i=1}^{n} \frac{1}{g'(Y_i)}$ is constant, thus the estimate of θ will be the same

- sufficient statistic e.g., $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$

  $$L(p \mid \mathbf{Y}) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} = p^S(1-p)^{n-S}$$

  for $S = \sum_{i=1}^{n} Y_i$, which is a sufficient statistic

  $$L(p \mid S) = \begin{pmatrix} n \\ S \end{pmatrix} p^S(1-p)^{n-S} \propto L(p \mid \mathbf{Y})$$

- different sampling plans

  the above two examples give the same sampling plan and data for the two likelihoods. Here we give an example that different sampling plan leads to proportional likelihoods.

1. as in the sufficent statitic example

$$Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$$

which leads to the first likelihood

$$L_1(p \mid S) = \begin{pmatrix} 12 \\ S \end{pmatrix} p^S (1-p)^{12-S}$$

2. negative binomial: Y_i's are observed until 3 0's appear, which leads to likelihood

$$L_2(p \mid S) = \begin{pmatrix} S+2 \\ S \end{pmatrix} p^S (1-p)^3$$

(it's $S+2$ cause the last observation must be 0)
the ratio is

$$\frac{\begin{pmatrix} 12 \\ S \end{pmatrix}}{\begin{pmatrix} S+2 \\ S \end{pmatrix}} (1-p)^{9-S}$$

which depends on p except for $S = 9$. When $S = 9$, the MLE estimate for p will be the same but the inference like hypothesis testing (p-value) is **different** cause null distribution in the two plans are different

2.4 Empirical distribution function as an MLE

$$\begin{aligned} Y_1, \cdots, Y_n &\sim \text{iid } F(y) \\ L_h(F \mid \mathbf{Y}) &= \prod_{i=1}^{n} \{F(Y_i + h) - F(Y_i - h)\} \\ &= \prod_{i=1}^{n} p_{i,h} \text{ assuming no ties} \end{aligned}$$

for the likelihood here we ignored the $(2h)^{-m}$ factor and the $p_{i,h}$ here needs to satisfy that $p_{i,h} \geq 0,\ \sum_{i=1}^{n} p_{i,h} \leq 1$. Obviously, the likelihood increases as $p_{i,h}$ increases until it reaches that

$$\sum_{i=1}^{n} p_{i,h} = 1$$

Thus use Lagrange multipliers, we solve for

$$\log L_h(F \mid \mathbf{Y}) + \lambda \left(\sum_{i=1}^{n} p_{i,h} - 1 \right) = \sum_{i=1}^{n} \log p_{i,h} + \lambda \left(\sum_{i=1}^{n} p_{i,h} - 1 \right)$$

then

$$\begin{aligned} \frac{\partial g}{\partial p_{i,h}} &= \frac{1}{p_{i,h}} + \lambda = 0, \quad i = 1, \ldots, n \\ \frac{\partial g}{\partial \lambda} &= \sum_{i=1}^{n} p_{i,h} - 1 = 0 \end{aligned}$$

we can show that

$$\hat{F}_h \xrightarrow{d} \widehat{F}_{\text{EMP}}(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leq y)$$

therefore we take the MLE of $F(y)$ as the empirical distribution function

$$\widehat{F}_{\text{MLE}}(y) = F_{\text{EMP}}(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq t)$$

2.5 Likelihoods for censored / truncated data

- fixed censoring (for censoring, we know exactly which individuals are censored) e.g. suppose we have underlying random variable $X \sim \mathrm{N}(\mu,\, \sigma^2)$ and we generate Y as follows

$$Y = \begin{cases} 0 & X \le 0 \\ X & X > 0 \end{cases}$$

$$F(y) = \begin{cases} 0 & y < 0 \\ \Phi\{(y-\mu)/\sigma\} & y \ge 0 \end{cases}$$

then the likelihood is

$$Y_1, \cdots, Y_n \sim \text{ iid}$$

$$L(\mu, \sigma \mid \mathbf{Y}) = \left\{ \prod_{i:Y_i=0} \Phi(-\mu/\sigma) \right\} \left[\prod_{i:Y_i>0} \sigma^{-1} \phi\left\{ (Y_i - \mu)/\sigma \right\} \right]$$

- truncation (unlike censoring, here we are unaware of the truncated individuals) e.g. suppose income $X \sim F_X(x;\, \theta)$ and sample $Y_1, \ldots, Y_n$ comes from incomes above L_0 (left truncated). Then

$$\begin{aligned} P(Y \le y) &= P\left(X \le y \mid X > L_0\right) I\left(y > L_0\right) \\ &= \frac{P\left(X \le y, X > L_0\right)}{P\left(X > L_0\right)} I\left(y > L_0\right) \\ &= \frac{F_X(y;\boldsymbol{\theta}) - F_X\left(L_0;\boldsymbol{\theta}\right)}{1 - F_X\left(L_0;\boldsymbol{\theta}\right)} I\left(y > L_0\right) \\ f_Y(y) &= \frac{f_X(y;\boldsymbol{\theta})}{1 - F_X\left(L_0;\boldsymbol{\theta}\right)} I\left(y > L_0\right) \end{aligned}$$

- random censoring: we have underlying random variables X and R (e.g. X: survival time, R: censoring time) and $X \perp R$. What we observed is $Y = \min(X,\, R)$ and $\delta = I(X \le R)$ then

$$\begin{aligned} f_{Y,\delta}(y, \delta = 1) =& f_X(y;\boldsymbol{\theta}) \left\{1 - F_R(y)\right\} \\ f_{Y,\delta}(y, \delta = 0) =& \left\{1 - F_X(y;\boldsymbol{\theta})\right\} f_R(y) \\ L(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\delta}) =& \prod_{i=1}^{n} f_X\left(Y_i;\boldsymbol{\theta}\right)^{\delta_i} \left\{1 - F_R\left(Y_i\right)\right\}^{\delta_i} \times \left\{1 - F_X\left(Y_i;\boldsymbol{\theta}\right)\right\}^{1-\delta_i} f_R\left(Y_i\right)^{1-\delta_i} \\ \propto& \prod_{i=1}^{n} f_X\left(Y_i;\boldsymbol{\theta}\right)^{\delta_i} \left\{1 - F_X\left(Y_i;\boldsymbol{\theta}\right)\right\}^{1-\delta_i} \\ & \text{if } F_R(\cdot) \text{ is noninformative of } \boldsymbol{\theta} \end{aligned}$$

proof: we can use the $2h$-method to get the density in this case.

$$\begin{aligned} \lim_{h\to 0} \frac{P(Y \in (y-h, y+h], \delta = 1)}{2h} &= \lim_{h\to 0} \frac{P(x \in (y-h, y+h), R > x)}{2h} \\ &= \lim_{h\to 0} \frac{\iint I(x \in (y-h, y+h],\, r > x) f_X(x) f_R(r) dx dr}{2h} \\ &= \lim_{h\to 0} \frac{\iint [I(r > x) f_R(r)] dr I(x \in (y-h, y+h]) f_X(x) dx dr}{2h} \\ &= \lim_{h\to 0} \frac{\iint [1 - F_R(x)] I(x \in (y-h, y+h]) f_X(x) dx dr}{2h} \\ &= \lim_{h\to 0} \frac{\int_{y-h}^{y+h} \left\{1 - F_R(x)\right\} f_x(x) dx}{2h} = \left[1 - F_R(y)\right] f_x(y) \end{aligned}$$

2.6 Likelihoods for regression models

- normal linear model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \quad i = 1, \cdots, n, \quad e_1, \cdots, e_n \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \tag{1}$$

1. fixed design: $x_1,\, x_2, \ldots,\, x_n$ are known nonrandom p-vectors

$$L\left(\boldsymbol{\beta}, \sigma \mid \{Y_i, \mathbf{x}_i\}_{i=1}^n\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{ -\sum_{i=1}^{n} \frac{\left(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2}{2\sigma^2} \right\}$$

$$\widehat{\boldsymbol{\beta}}_{\text{MLE}} = \widehat{\boldsymbol{\beta}}_{\text{LS}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ and

$$\widehat{\sigma}^2_{\text{MLE}} = n^{-1}\sum_{i=1}^{n}\widehat{e}_i^2 = n^{-1}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{\text{MLE}}\right)^2$$

whereas

$$\widehat{\sigma}_{\text{unbiased}} = (n-p)^{-1}\sum_{i=1}^{n}\widehat{e}_i^2$$

2. random design: $\{Y_i, \mathbf{x}_i\}_{i=1}^n \overset{\text{iid}}{\sim} (Y, \mathbf{x})$ and marginally $\mathbf{x} \sim f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\tau})$ then

$$L\left(\boldsymbol{\beta}, \sigma, \boldsymbol{\tau} \mid \{Y_i, \mathbf{x}_i\}_{i=1}^n\right) = \text{ above } \times \prod_{i=1}^{n} f_{\mathbf{x}}\left(\mathbf{x}_i; \boldsymbol{\tau}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{\left(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2}{2\sigma^2}\right\} \times \prod_{i=1}^{n} f_{\mathbf{x}}\left(\mathbf{x}_i; \boldsymbol{\tau}\right)$$

- additive errors nonlinear model: it is very similar to 1 but with the following the modification

$$Y_i = g\left(\mathbf{x}_i, \beta\right) + e_i$$

with function g known. Then

$$\widehat{\sigma}^2_{\text{MLE}} = n^{-1}\sum_{i=1}^{n}\widehat{e}_i^2 = n^{-1}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{\text{MLE}}\right)^2$$

but the estimator for β tend to have no closed form and need to solve numerically.

- generalized linear model

Exponential family is of the following form:

$$f(y|\theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

where θ is the *canonical parameter*, ϕ is the *dispersion parameter*, $b(\theta)$ is the *cumulative function* (different from cumulative generating function) and $a(), b(), c()$ are known functions.

To fit into the regression framework, we have

$$\log f\left(y_i; \theta_i, \phi\right) = \frac{y_i\theta_i - b\left(\theta_i\right)}{a_i(\phi)} + c\left(y_i, \phi\right)$$

link function is $g(.)$ such that

$$g\left(\mu_i\right) = \mathbf{x}_i^T\boldsymbol{\beta} \Longrightarrow \mu_i = g^{-1}\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) \Longrightarrow \theta_i = b'^{-1}\left\{g^{-1}\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right\}$$

a special case of the link function is the *canonical link*

$$g\left(\mu_i\right) = \theta_i = \mathbf{x}_i^T\boldsymbol{\beta}$$

in this setting the log-likelihood becomes

$$\log L\left(\boldsymbol{\beta}, \phi \mid \{Y_i, \mathbf{x}_i\}_{i=1}^n\right) = \sum_{i=1}^{n}\left\{\frac{y_i\mathbf{x}_i^T\boldsymbol{\beta} - b\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{a_i(\phi)} + c\left(y_i, \phi\right)\right\}$$

If $Y \sim f(y; \theta, \phi)$ and ϕ is fixed, then

Property 1. $E(Y) = b'(\theta)$

Property 2. $\text{Var}(Y) = b''(\theta)a(\phi)$

Proof

We first prove that

$$E[l'(\theta)] = 0$$

and

$$E[l''(\theta)] = -E\{[l'(\theta)]^2\}$$

$$\ell(\theta) = \log f(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

$$\begin{aligned}
&E\left[\frac{\partial}{\partial\theta}\log f(y;\theta,\phi)\right] \\
&= \int \left\{\frac{\partial}{\partial\theta}\log f(y;\theta,\phi)\right\} f(y;\theta,\phi)dy \\
&= \int f'(y;\theta,\phi)dy \\
&\overset{\text{under some regularity conditions}}{=} \frac{\partial}{\partial\theta}\int f(y;\theta,\phi)dy = 0
\end{aligned}$$

and

$$\begin{aligned}
&E\left[-\frac{\partial^2}{\partial\theta^2}\log f(y;\theta,\phi)\right] \\
=&-\int\left\{\frac{f''(y;\theta,\phi)\cdot f(y;\theta,\phi) - [f'(y;\theta,\phi)]^2}{f(y;\theta,\phi)^2}\right\}\cdot f(y;\theta,\phi)dy \\
=&-\int f''(y;\theta,\phi)dy + \int\left[\frac{f'(y;\theta,\phi)}{f(y;\theta,\phi)}\right]^2 dy \\
=&-\frac{\partial^2}{\partial\theta^2}\int f(y;\theta,\phi)dy + E\left[[(\log f(y;\theta,\phi))']^2\right] \\
=&E\left\{[(\log f(y;\theta,\phi))']^2\right\}
\end{aligned}$$

Then the score function for θ is

$$U(\theta) = \frac{\partial\ell(\theta)}{\partial\theta} = \frac{Y - b'(\theta)}{a(\phi)}$$

Since $E(U(\theta)) = 0$, we have

$$E\left(\frac{Y - b'(\theta)}{a(\phi)}\right) = 0 \Longrightarrow E(Y) = b'(\theta)$$

Since $E(U(\theta)) = 0$, we have

$$\mathrm{Var}(U(\theta)) = E\left(\frac{\partial\ell(\theta)}{\partial\theta}\right)^2 = -E\left(\frac{\partial^2\ell(\theta)}{\partial\theta^2}\right) \Longrightarrow \frac{\mathrm{Var}(Y)}{a^2(\phi)} = \frac{b''(\theta)}{a(\phi)} \Longrightarrow \mathrm{Var}(Y) = b''(\theta)a(\phi)$$

- generalized linear mixed model (GLMM)
  in this model we extend the canonical-link GLM to accommodate random effects

  $$\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\mathbf{U}, \quad \mathbf{U} \sim f_{\mathbf{U}}(\mathbf{u};\boldsymbol{\nu})$$

  then the likelihood is

  $$L\left(\boldsymbol{\beta},\phi,\boldsymbol{\nu} \mid \{Y_i,\mathbf{x}_i,\mathbf{z}_i\}_{i=1}^n\right) = \prod_{i=1}^n \int f_{Y_i|\mathbf{U}}\left(Y_i \mid \mathbf{u},\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\beta},\phi\right) f_{\mathbf{U}}(\mathbf{u};\boldsymbol{\nu})d\mathbf{u}$$

- accelerated failure time model

  $$\log T = \mathbf{x}^T\beta + \sigma e, \quad T \perp R \mid \mathbf{x}$$

  usually $e \sim \mathrm{N}(0,\,1)$
  the observed random variable is $Y = \min(T,R), \delta = I(T \le R)$ then the likelihood is

  $$L\left(\boldsymbol{\beta},\sigma \mid \{Y_i,\delta_i,\mathbf{x}_i\}_{i=1}^n\right) \propto \prod_{i=1}^n \left\{\frac{1}{\sigma}f_e\left(\frac{\log Y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\right\}^{\delta_i}\left\{1 - F_e\left(\frac{\log Y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\right\}^{1-\delta_i}$$

2.7 Marginal and conditional likelihoods

Suppose $\mathbf{Y} \Longleftrightarrow (\mathbf{W}, \mathbf{V})$ and we have

$$\begin{cases} \theta_1\text{: parameter of interest} & (2) \\ \theta_2\text{: nuisance parameter} & (3) \end{cases}$$

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= f_{\mathbf{W},\mathbf{V}}(\mathbf{w}, \mathbf{v}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= f_{\mathbf{W}|\mathbf{v}}(\mathbf{w} \mid \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \end{aligned}$$

1. if $f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}_1)$, it is a marginal likelihood
2. if $f_{\mathbf{W}|\mathbf{V}}(\mathbf{w} \mid \mathbf{v}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_{\mathbf{W}|\mathbf{V}}(\mathbf{w} \mid \mathbf{v}; \boldsymbol{\theta}_1)$, it is a conditional likelihood

Let's illustrate with an example
Neyman–Scott problem

$$\begin{aligned} &Y_{ij} \overset{i.i.d}{\sim} N\left(\mu_i, \sigma^2\right), i = 1, \cdots, n, j = 1, 2 \\ &\boldsymbol{\theta} = \left(\sigma^2, \mu_1, \cdots, \mu_n\right)^T \dim n + 1 \text{ (increases with sample size)} \end{aligned}$$

then we have

$$\begin{aligned} \widehat{\sigma}^2_{\text{MLE}} &= (2n)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \left(Y_{ij} - \widehat{\mu}_{i,\text{MLE}}\right)^2 \\ &= \frac{n^{-1} \sum_{i=1}^{n} \left(Y_{i1} - Y_{i2}\right)^2}{4} \\ E\left(\widehat{\sigma}^2_{\text{MLE}}\right) = \frac{\sigma^2}{2} &\Longrightarrow \widehat{\sigma}^2_{\text{MLE}} \xrightarrow{p} \frac{\sigma^2}{2} \text{ by WLLN} \end{aligned}$$

as we can see, consistency no longer exist for MLE in this case cause the number of parameters increases with sample size. Then how to achieve consistency?

$$\begin{aligned} V_i &= \frac{(Y_{i1} - Y_{i2})}{\sqrt{2}} \sim N\left(0, \sigma^2\right) \\ W_i &= \frac{(Y_{i1} + Y_{i2})}{\sqrt{2}} \sim N\left(\sqrt{2}\mu_i, \sigma^2\right) \end{aligned}$$

1. marginal likelihood

$$\begin{aligned} L(\sigma \mid \mathbf{V}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{V_i^2}{2\sigma^2}\right) \\ \widehat{\sigma}^2_{\text{MMLE}} &= n^{-1} \sum_{i=1}^{n} V_i^2 = \frac{n^{-1} \sum_{i=1}^{n} \left(Y_{i1} - Y_{i2}\right)^2}{2} \end{aligned}$$

   in this case, only V_i needs to be identified. However, W_i may provide a sense of information loss

2. conditional likelihood
   the *key* for this approach is to identify sufficient statistic for nuisance parameter under the assumption that the parameter of interest is known. In our case, $T_i = Y_{i1} + Y_{i2}$ is sufficient for $\mu_i \Longrightarrow (Y_{i1}, Y_{i2}) \mid T_i$ doesn't depend on μ_i

$$\begin{aligned} Y_{i1} \mid T_i &\sim N\left(\frac{T_i}{2}, \frac{\sigma^2}{2}\right) \\ \widehat{\sigma}^2_{\text{CMLE}} &= 2n^{-1} \sum_{i=1}^{n} \left(Y_{i1} - \frac{T_i}{2}\right)^2 \\ &= \frac{n^{-1} \sum_{i=1}^{n} \left(Y_{i1} - Y_{i2}\right)^2}{2} \end{aligned}$$

Example1: Logistic regression measurement error model

$$\begin{aligned} &P(Y = 1 \mid X) = \tfrac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)} \\ &W = X + U, \quad U \sim N\left(0, \sigma_U^2\right) \text{ with } \sigma_U^2 \text{ known} \end{aligned}$$

with $U \perp Y$ and $U \perp X$, σ_U^2 is known

- data: $(Y_i, W_i), i = 1, \cdots, n$
- parameters of interest: α, β

then

$$\begin{aligned} & f(Y_i, W_i \mid \alpha, \beta, X_i) \\ =& \frac{\exp\{Y_i(\alpha+\beta X_i)\}}{1+\exp(\alpha+\beta X_i)} \frac{1}{\sqrt{2\pi}\sigma_U} \exp\left\{-\frac{(W_i-X_i)^2}{2\sigma_U^2}\right\} \\ =& \frac{1}{\sqrt{2\pi}\sigma_U} \frac{1}{1+\exp(\alpha+\beta X_i)} \times \exp\left\{\frac{\left(W_i+Y_i\sigma_U^2\beta\right)X_i}{\sigma_U^2} - \frac{X_i^2}{2\sigma_U^2} + \alpha Y_i - \frac{W_i^2}{2\sigma_U^2}\right\} \end{aligned}$$

$W_i + Y_i\sigma_{UU}^2\beta$ or equivalently $T_i = W_i + (Y_i - 1/2)\sigma_U^2\beta$ is sufficient for X_i, assuming α and β are known (Sufficient statistic depends on the parameter of interest). Then

$$P(Y_i = 1 \mid T_i) = \frac{\exp(\alpha+\beta T_i)}{1+\exp(\alpha+\beta T_i)}$$

and the conditional likelihood score is based on

$$\frac{\partial}{\partial(\alpha,\beta)} \log P(Y_i \mid T_i) = \left\{Y_i - \frac{\exp(\alpha+\beta T_i)}{1+\exp(\alpha+\beta T_i)}\right\} \begin{pmatrix} 1 \\ T_i \end{pmatrix}$$

and the score is

$$\sum_{i=1}^{n} \left\{Y_i - \frac{\exp(\alpha+\beta T_i)}{1+\exp(\alpha+\beta T_i)}\right\} \begin{pmatrix} 1 \\ T_i \end{pmatrix}\Bigg|_{T_i = W_i + (Y_i - 1/2)\sigma_U^2\beta} = \mathbf{0}$$

Note: when taking the differentiation, T_i is treated as given but not as a function of β

Example2: exponential families with canonical parameter

$f(\mathbf{y}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = h(\mathbf{y}) \exp\left\{\sum_i \theta_{1i} W_i + \sum_j \theta_{2j} V_j - A(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right\} \Longrightarrow \mathbf{W} \mid \mathbf{V} \sim$ exponential family indexed by $\boldsymbol{\theta}_1$ only

Example3: conditional logistic regression

$$\begin{aligned} \text{logit}\{P(Y_i = 1)\} &= \mathbf{x}_i^T \boldsymbol{\beta} \\ P(Y_i = 1) &= \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)} \\ L(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) &= \prod_{i=1}^{n} \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)^{Y_i}}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)} = \frac{\exp\left(\sum_{i=1}^n Y_i \mathbf{x}_i^T\boldsymbol{\beta}\right)}{\prod_{i=1}^n \left\{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right\}} \end{aligned}$$

$T_j = \sum_{i=1}^n x_{ij} Y_i$ is sufficient for $\beta_j, j = 1, \cdots, p$. If β_k is the parameter of interest then let $W = T_k$ and $\mathbf{V} = (T_1, \cdots, T_{k-1}, T_{k+1}, \cdots, T_p)^T$

$$P(W \mid \mathbf{V}) = \frac{c(T_1, \cdots, T_p)\exp(\beta_k T_k)}{\sum_u c(T_1, \cdots, T_{k-1}, u, T_{k+1}, \cdots, T_p)\exp(\beta_k u)}$$

where $P(W, V) = c(T_1, \cdots, T_p)\exp(\beta_k T_k)$ and $P(V) = \sum_w P(W, V) = \sum_u c(T_1, \cdots, T_{k-1}, u, T_{k+1}, \cdots, T_p)\exp(\beta_k u)$

2.8 MLE and information matrix

Recall that

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathbf{Y})$$

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta} \mid \mathbf{Y})$$

if $\ell(\boldsymbol{\theta})$ is continuously differentiable, then likelihood score

$$\mathbf{S}(\boldsymbol{\theta}) = \mathbf{S}(\mathbf{Y}, \boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \{\ell'(\boldsymbol{\theta})\}^T$$

exists and $\hat{\theta_{\text{MLE}}}$ satisfies $\mathbf{S}\left(\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right) = \mathbf{0}$. Under most of the cases, $\hat{\theta_{\text{MLE}}}$ is unique, or at least there is a principled strategy for choosing a single solution from among the possibly multiple values. But there are also special cases, for example

$$Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} f(y;\mu) = \begin{cases} e^{-(y-\mu)} & y > \mu \\ 0 & \text{otherwise} \end{cases}$$

any $\mu < \min(Y_i)$ leads to likelihood to be 0.

Property 1. invariant property:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} \text{ for } \boldsymbol{\theta} \Longrightarrow \widehat{\boldsymbol{\tau}}_{\mathrm{MLE}} = g\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) \text{ for } \tau = g(\boldsymbol{\theta})$$

Property 2. asymptotic normality: under regularity conditions

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta}\right) \xrightarrow{d} N\left\{\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})^{-1}\right\}, \quad \text{as } n \to \infty$$

where

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\theta}) &= E\left[\left\{\frac{\partial}{\partial \boldsymbol{\theta}^T} \log f(Y;\boldsymbol{\theta})\right\}^{\otimes 2}\right] \\
&= \operatorname{var}\left\{\frac{\partial}{\partial \boldsymbol{\theta}^T} \log f(Y;\boldsymbol{\theta})\right\} \\
&= -E\left\{\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \log f(Y;\boldsymbol{\theta})\right\}
\end{aligned}$$

How to estimate $I(\theta)$?
$I(\theta)$ has 2 components: $I(.)$ and θ. It is a natural choice for using $\hat{\theta_{\mathrm{MLE}}}$ to estimate θ. But what about $I(.)$?

1. theoretical $I(.) \Longrightarrow \mathbf{I}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)$ is the MLE for $I(\theta)$(invariant property)
2. estimated $I(.)$ version 1

$$\begin{aligned}
\bar{\mathbf{I}}(\mathbf{Y},\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^{n}\left\{-\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{s}\left(Y_i,\boldsymbol{\theta}\right)\right\} \\
&= n^{-1}\sum_{i=1}^{n}\left\{-\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\log f\left(Y_i;\boldsymbol{\theta}\right)\right\}
\end{aligned}$$

3. estimated $I(.)$ version 2 (usually less efficient)

$$\begin{aligned}
\bar{\mathbf{I}}^{*}(\mathbf{Y},\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^{n}\mathbf{s}\left(Y_i,\boldsymbol{\theta}\right)^{\otimes 2} \\
&= n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial \boldsymbol{\theta}^T}\log f\left(Y_i;\boldsymbol{\theta}\right)\right\}^{\otimes 2}
\end{aligned}$$

   Note: when Y_i are independent but not identically distributed. The $f(.;\theta) \Longrightarrow f_i(.;\theta)$

4. estimated $I(.)$ version 1: (Y_i are independent but not identically distributed)

$$\begin{aligned}
\overline{\boldsymbol{I}}(\boldsymbol{Y},\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{s}_i\left(Y_i,\boldsymbol{\theta}\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\log f_i\left(Y_i;\boldsymbol{\theta}\right)\right\}
\end{aligned}$$

   the according average expected information matrix (also called the average Fisher information matrix)

$$\overline{\boldsymbol{I}}(\boldsymbol{\theta}) = \mathrm{E}\{\overline{\boldsymbol{I}}(\boldsymbol{Y},\boldsymbol{\theta})\} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{-\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\log f_i\left(Y_i;\boldsymbol{\theta}\right)\right\} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left\{\frac{\partial}{\partial \boldsymbol{\theta}^T}\log f_i\left(Y_i;\boldsymbol{\theta}\right)\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\log f_i\left(Y_i;\boldsymbol{\theta}\right)\right\}\right]$$

   in the i.i.d case $\bar{I}(\theta) = I(\theta)$

5. estimated $I(.)$ version 2: (Y_i are independent but not identically distributed)

$$\begin{aligned}
\overline{\boldsymbol{I}}^{*}(\boldsymbol{Y},\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}s_i\left(Y_i,\boldsymbol{\theta}\right)\boldsymbol{s}_i\left(Y_i,\boldsymbol{\theta}\right)^T \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial \boldsymbol{\theta}^T}\log f_i\left(Y_i;\boldsymbol{\theta}\right)\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\log f_i\left(Y_i;\boldsymbol{\theta}\right)\right\}
\end{aligned}$$

Actually, we discussed two types of information here

1. the total information:

$$\mathbf{I}_T(\boldsymbol{\theta}) = -E\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta}\mid\mathbf{Y})\right\}$$

2. the observed total information

$$\mathbf{I}_T(\mathbf{Y},\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta}\mid\mathbf{Y})$$

Example 1: $N\left(\mu,\sigma^2\right)$

$$\log f(y;\mu,\sigma) = \text{ constant } -\log\sigma - \frac{1}{2\sigma^2}(y-\mu)^2$$

$$\mathbf{I}(\mu,\sigma) = \frac{1}{\sigma^2}\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix}$$

more specifically

$$\ell'(\theta) = \frac{\partial}{\partial\theta}\log f(y;\mu,\sigma) = \left(-\frac{1}{\sigma} + \frac{(y-\mu)^2}{\sigma^3},\ \frac{y-\mu}{\sigma^2}\right)^T$$

then

$$var(\ell'(\theta)) = \begin{bmatrix} var(-\frac{1}{\sigma} + \frac{(y-\mu)^2}{\sigma^3}) & cov(-\frac{1}{\sigma} + \frac{(y-\mu)^2}{\sigma^3},\ \frac{y-\mu}{\sigma^2}) \\ cov(-\frac{1}{\sigma} + \frac{(y-\mu)^2}{\sigma^3},\ \frac{y-\mu}{\sigma^2}) & var(\frac{y-\mu}{\sigma^2}) \end{bmatrix}$$

then by the results of normal moments we get the answer

Order	Non-central moment	Central moment
1	μ	0
2	$\mu^2+\sigma^2$	σ^2
3	$\mu^3+3\mu\sigma^2$	0
4	$\mu^4+6\mu^2\sigma^2+3\sigma^4$	$3\sigma^4$
5	$\mu^5+10\mu^3\sigma^2+15\mu\sigma^4$	0
6	$\mu^6+15\mu^4\sigma^2+45\mu^2\sigma^4+15\sigma^6$	$15\sigma^6$
7	$\mu^7+21\mu^5\sigma^2+105\mu^3\sigma^4+105\mu\sigma^6$	0
8	$\mu^8+28\mu^6\sigma^2+210\mu^4\sigma^4+420\mu^2\sigma^6+105\sigma^8$	$105\sigma^8$

Example 2: recall for the normal error regression models

$$\begin{aligned} L\left(\beta,\sigma\mid\{Y_i,\boldsymbol{x}_i\}_{i=1}^n\right) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(Y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\sum_{i=1}^n\frac{\left(Y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2}{2\sigma^2}\right\}\end{aligned}$$

then

$$\ell = -\log\sqrt{2\pi} - \log\sigma - \frac{\left(Y_i-x_i^\top\beta\right)^2}{2\sigma^2}$$

it follows that

$$\ell' = (\frac{\partial\ell}{\partial\sigma},\ \frac{\partial\ell}{\partial\beta}) = \left(-\frac{1}{\sigma} + \frac{\left(Y_i-x_i^T\beta\right)^2}{\sigma^3},\ \frac{x_i^\top\left(Y_i-x_i^\top\beta\right)}{\sigma^2}\right)$$

and

$$\ell'' = \begin{bmatrix} \frac{1}{\sigma^2} - 3\frac{\left(Y_i-x_i^\top\beta\right)^2}{\sigma^4} & -2\frac{x_i^\top\left(Y_i-x_i^\top\beta\right)}{\sigma^3} \\ -2\frac{x_i^\top\left(Y_i-x_i^\top\beta\right)}{\sigma^3} & \frac{-x_i'x_i}{\sigma^2} \end{bmatrix}$$

taking expectation and organize we get

$$\bar{\mathbf{I}}(\beta,\sigma) = \frac{1}{\sigma^2}\begin{pmatrix}\mathbf{X}^T\mathbf{X}/n & 0\\ 0 & 2\end{pmatrix}$$

then invert $\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\beta},\sigma) = n\overline{\boldsymbol{I}}(\boldsymbol{\beta},\sigma)$ we get

$$\begin{aligned}\operatorname{avar}(\hat{\beta}) &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\sigma^2 \\ \operatorname{avar}(\hat{\sigma}) &= \frac{\sigma^2}{2n}\end{aligned}$$

Similarly, for $Y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + e_i$

$$\bar{\mathbf{I}}(\boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^2} \begin{pmatrix} \mathbf{G}^T\mathbf{G}/n & 0 \\ 0 & 2 \end{pmatrix}$$

where $[\mathbf{G}]_{ij} = \partial g(\mathbf{x}_i, \boldsymbol{\beta}) / \partial \beta_j$
Example 3: GLM with canonical link ($g(\mu_i) = \theta_i = \mathbf{x}_i^T\boldsymbol{\beta}$)

$$\begin{aligned} \mathbf{S}(\boldsymbol{\beta}, \phi) &= \sum_{i=1}^{n} \frac{\left\{Y_i - b'\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right\}\mathbf{x}_i}{a_i(\phi)} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)\,\mathbf{x}_i}{a_i(\phi)} \\ -\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{S}(\boldsymbol{\beta}, \phi) &= \sum_{i=1}^{n} \frac{b''\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{a_i(\phi)}\mathbf{x}_i^{\otimes 2} = \sum_{i=1}^{n} \frac{\operatorname{var}(Y_i)}{a_i(\phi)^2}\mathbf{x}_i^{\otimes 2} \\ &\equiv \mathbf{X}^T\mathbf{V}\mathbf{X} = n\bar{\mathbf{I}}(\mathbf{Y}, \boldsymbol{\beta}) = n\bar{\mathbf{I}}(\boldsymbol{\beta}) \end{aligned}$$

where $\boldsymbol{V} = \operatorname{diag}\left\{\operatorname{Var}(Y_1)/a_1(\phi)^2, \ldots, \operatorname{Var}(Y_n)/a_n(\phi)^2\right\}$. In this special case, average Fisher information is equal to average observed information (cause $\mathbf{X}^T\mathbf{V}\mathbf{X}$ is irrespective of Y). When Φ is unknown, the full average Fisher information is

$$\bar{\mathbf{I}}(\beta, \phi) = \begin{pmatrix} \mathbf{X}^T\mathbf{V}\mathbf{X}/n & 0 \\ 0 & \bar{I}(\phi) \end{pmatrix}$$

thus estimating Φ does not increase the variability of β estimation asymptotically in canonical link GLM models. Cause the diagonal of $\bar{\mathbf{I}}(\beta, \phi)$ is 0 and therefore we have

$$\bar{I}^{-1}(\beta, \phi) = \begin{pmatrix} n\left[X'VX\right]^{-1} & 0 \\ 0 & \bar{I}^{-1}(\phi) \end{pmatrix}$$

However, in general

$$\begin{pmatrix} I_{11} & \mathbf{I}_{12} \\ & \mathbf{I}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \left(I_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}\right)^{-1} & -I_{11}^{-1}\mathbf{I}_{12}\left(\mathbf{I}_{22} - \mathbf{I}_{21}I_{11}^{-1}\mathbf{I}_{12}\right)^{-1} \\ & \left(\mathbf{I}_{22} - \mathbf{I}_{21}I_{11}^{-1}\mathbf{I}_{12}\right)^{-1} \end{pmatrix}$$

$\mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21} \geq 0 \Longrightarrow \left(I_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}\right)^{-1} \geq I_{11}^{-1}$
which means that when adding parameters to a model, the diagonal elements of the inverse information matrix are always no less than the corresponding elements of the simpler model unless $I_{12} = 0$. This is called *variance inflation*

2.8.1 Transformed and modeled parameters

If we have $f(y;\ \theta)$ where θ: b-dimensional and therefore $I_\theta : b$. We also have β:s-dimensional and $s \leq b,\ \theta = g(\beta)$ then

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \log f\{y; \mathbf{g}(\boldsymbol{\beta})\} \\ \frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{\beta}) &= \left.\frac{\partial}{\partial \boldsymbol{\theta}} \log f(y; \boldsymbol{\theta})\right|_{\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\beta})} \frac{\partial \mathbf{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ \mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) &= \left\{\frac{\partial \mathbf{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^T \mathbf{I}_{\boldsymbol{\theta}}\{\mathbf{g}(\boldsymbol{\beta})\}\frac{\partial \mathbf{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \end{aligned}$$

Example 1: in the normal setting, let $\boldsymbol{\theta} = (\mu, \sigma)^T$ and $\boldsymbol{\beta} = \left(\mu, \sigma^2\right)^T$ then

$$\begin{aligned} \frac{\partial \mathbf{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \operatorname{diag}\{1, 1/(2\sigma)\} \\ \mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) &= \begin{pmatrix} 1 & 0 \\ 0 & 1/(2\sigma) \end{pmatrix} \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1/(2\sigma) \end{pmatrix} \\ &= \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 1/\left(2\sigma^2\right) \end{pmatrix} \end{aligned}$$

2.9 Methods for maximizing the likelihood

1. Profile likelihood: maximize in a sequential way and therefore achieve dimension reduction

$$\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \max_{\boldsymbol{\theta}_1} L\left\{\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\right\}, \quad \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1) = \arg\max_{\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

Example 1: $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$

$$
\begin{aligned}
f_Y(y;\alpha,\beta) =& \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \\
\ell(\alpha,\beta) =& -n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\sum_i \log Y_i - \frac{\sum_i Y_i}{\beta} \\
\frac{\partial}{\partial\beta}\ell(\alpha,\beta) =& -\frac{n\alpha}{\beta} + \frac{\sum_i Y_i}{\beta^2} \Longrightarrow \widehat{\beta}(\alpha) = \bar{Y}/\alpha \\
\ell\{\alpha, \widehat{\beta}(\alpha)\} =& -n\log\Gamma(\alpha) - n\alpha(\log\bar{Y} - \log\alpha) \\
& + (\alpha-1)\sum_i \log Y_i - n\alpha
\end{aligned}
$$

2. Newton methods

$$
\begin{aligned}
\mathbf{0} = \mathbf{S}(\boldsymbol{\theta}) &\approx \mathbf{S}\left(\boldsymbol{\theta}^{(\nu)}\right) + \left\{ \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{S}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\nu)}} \right\}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(\nu)}\right) \\
&= \mathbf{S}\left(\boldsymbol{\theta}^{(\nu)}\right) - \mathbf{I}_T\left(\mathbf{Y}, \boldsymbol{\theta}^{(\nu)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(\nu)}\right) \\
\boldsymbol{\theta}^{(\nu+1)} &= \boldsymbol{\theta}^{(\nu)} + \mathbf{I}_T\left(\mathbf{Y}, \boldsymbol{\theta}^{(\nu)}\right)^{-1}\mathbf{S}\left(\boldsymbol{\theta}^{(\nu)}\right)
\end{aligned}
$$

   (a) start with initial value $\boldsymbol{\theta}^{(0)}$

   (b) update from current $\boldsymbol{\theta}^{(\nu)}$ to obtain $\boldsymbol{\theta}^{(\nu+1)}$

   (c) stop if $\left\|\mathbf{S}\left(\boldsymbol{\theta}^{(\nu+1)}\right)\right\|$ or $\left\|\boldsymbol{\theta}^{(\nu+1)} - \boldsymbol{\theta}^{(\nu)}\right\|$ is sufficiently small. Otherwise keep updating

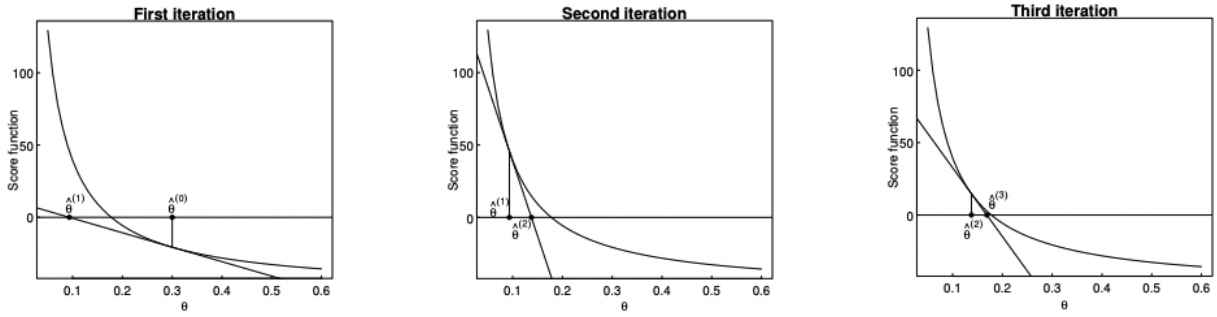


Figure 1: Newton-Raphson method (source)

If the first derivative is not well behaved in the neighborhood of a particular root, the method may overshoot. To prevent overshooting: At each step, ensure improvement of $\boldsymbol{\theta}^{(\nu+1)}$ over $\boldsymbol{\theta}^{(\nu)}$ by repeatedly halving the step size.
Under certain conditions, Newton methods have *local quadratic convergence*, which means that for some $c > 0$

$$
\left\|\boldsymbol{\theta}^{(\nu+1)} - \widehat{\boldsymbol{\theta}}_{\text{MLE}}\right\| \le c\left\|\boldsymbol{\theta}^{(\nu)} - \widehat{\boldsymbol{\theta}}_{\text{MLE}}\right\|^2
$$

proof: (although I am not totally agree with this proof)

$$
\begin{aligned}
0 = S\left(\widehat{\theta}_{\text{MLE}}\right) =& S\left(\theta^{(\nu)}\right) - I_T\left(\mathbf{Y}, \theta^{(\nu)}\right)\left(\widehat{\theta}_{\text{MLE}} - \theta^{(\nu)}\right) + \frac{1}{2}S''\left(\vartheta^{(\nu)}\right)\left(\widehat{\theta}_{\text{MLE}} - \theta^{(\nu)}\right)^2 \\
\widehat{\theta}_{\text{MLE}} - \theta^{(\nu+1)} =& \frac{1}{2}\frac{S''\left(\vartheta^{(\nu)}\right)}{I_T\left(\mathbf{Y}, \theta^{(\nu)}\right)}\left(\widehat{\theta}_{\text{MLE}} - \theta^{(\nu)}\right)^2
\end{aligned}
$$

Thus, the local quadratic convergence holds under the following conditions:

- $I_T(\mathbf{Y}, \theta) \neq 0$ in a neighborhood of $\hat{\theta}_{\text{MLE}}$
- $S''(\theta)$ is bounded
- $\theta^{(\nu)}$ is sufficiently close to $\hat{\theta}_{\text{MLE}}$ (so that we can eliminate $S\left(\theta^{(\nu)}\right)$)

3. Fisher scoring: $\mathbf{I}_T\left(\mathbf{Y}, \boldsymbol{\theta}^{(\nu)}\right)$ in the above method replaced by its expectation $\mathbf{I}_T\left(\boldsymbol{\theta}^{(\nu)}\right)$

4. One-step estimator: typically $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta} = O_p\left(n^{-1/2}\right)$ (from asymptotic normality). If one starts with $\boldsymbol{\theta}^{(0)}$ such that $\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta} = O_p\left(n^{-1/2}\right)$, then

$$\boldsymbol{\theta}^{(1)} - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = O_p\left(n^{-1}\right)$$

under regularity conditions (by the local quadratic convergence property).

5. EM algorithm (nice reading): view observed data Y as incomplete, with Z missing. Write log joint likelihood of "complete" data as $\ell_C(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z})$

   (a) E step: calculate

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y}\right) &= E_{\boldsymbol{\theta}^{(\nu)}}\left\{\ell_C(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}\right\} \\ &= \int \ell_C(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{z}) f_{\mathbf{Z}|\mathbf{Y}}\left(\mathbf{z} \mid \mathbf{Y}, \boldsymbol{\theta}^{(\nu)}\right) d\mathbf{z} \end{aligned}$$

   (b) maximize $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y}\right)$ with respect to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{\nu+1}$

   Example 1: 2-component mixtures

$$Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} f(y; \boldsymbol{\theta}) = p f_1\left(y; \mu_1, \sigma_1\right) + (1-p) f_2\left(y; \mu_2, \sigma_2\right)$$

   where $f_1,\ f_2$ are normal densities

$$\begin{aligned} \boldsymbol{\theta} &= \left(\mu_1, \sigma_1, \mu_2, \sigma_2, p\right)^T \\ \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\left\{p f_1\left(Y_i; \mu_1, \sigma_1\right) + (1-p) f_2\left(Y_i; \mu_2, \sigma_2\right)\right\} \end{aligned}$$

   it's hard to maximize directly. Therefore, turn to the EM algorithm

$$\begin{aligned} Y_i &= Z_i X_{1i} + (1 - Z_i) X_{2i} \\ X_{11}, \cdots, X_{1n} &\overset{\text{iid}}{\sim} N\left(\mu_1, \sigma_1^2\right) \\ X_{21}, \cdots, X_{2n} &\overset{\text{iid}}{\sim} N\left(\mu_2, \sigma_2^2\right) \\ Z_1, \cdots, Z_n &\overset{\text{iid}}{\sim} \text{Bernoulli}(p) \end{aligned}$$

   join likelihood of complete data $(Y,\ Z)$

$$\begin{aligned} L_C(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z}) &= \prod_{i=1}^{n} \left\{p f_1\left(Y_i; \mu_1, \sigma_1\right)\right\}^{Z_i} \times \left\{(1-p) f_2\left(Y_i; \mu_2, \sigma_2\right)\right\}^{(1-Z_i)} \\ \ell_C(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z}) &= \sum_{i=1}^{n} \left\{Z_i \log f_1\left(Y_i; \mu_1, \sigma_1\right) + (1 - Z_i) \log f_2\left(Y_i; \mu_2, \sigma_2\right) + Z_i \log p + (1 - Z_i) \log(1-p)\right\} \end{aligned}$$

   cause

$$\begin{aligned} p(y, z=1) &= p(y \mid z=1) p(z=1) = p f_1 \\ p(y, z=0) &= p(y \mid z=0) p(z=0) = (1-p) f_2 \end{aligned}$$

   - E-step

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y}\right) =& E_{\boldsymbol{\theta}^{(\nu)}}\left\{\ell_C(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}\right\} \\ =& \sum_{i=1}^{n} \left\{w_i^{(\nu)} \log f_1\left(Y_i; \mu_1, \sigma_1\right) + \left(1 - w_i^{(\nu)}\right) \log f_2\left(Y_i; \mu_2, \sigma_2\right) + w_i^{(\nu)} \log p + \left(1 - w_i^{(\nu)}\right) \log(1-p)\right\} \end{aligned}$$

     where

$$\begin{aligned} w_i^{(\nu)} &= E_{\boldsymbol{\theta}^{(\nu)}}\left(Z_i \mid Y_i\right) \\ &= \frac{p^{(\nu)} f_1\left(Y_i; \mu_1^{(\nu)}, \sigma_1^{(\nu)}\right)}{p^{(\nu)} f_1\left(Y_i; \mu_1^{(\nu)}, \sigma_1^{(\nu)}\right) + \left(1 - p^{(\nu)}\right) f_2\left(Y_i; \mu_2^{(\nu)}, \sigma_2^{(\nu)}\right)} \end{aligned}$$

substituting normal densities

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y}\right) = \sum_{i=1}^{n} \Bigg[& w_i^{(\nu)} \left\{ -\log \sigma_1 - \frac{\left(Y_i - \mu_1\right)^2}{2\sigma_1^2} \right\} \\ & + \left(1 - w_i^{(\nu)}\right) \left\{ -\log \sigma_2 - \frac{\left(Y_i - \mu_2\right)^2}{2\sigma_2^2} \right\} \\ & + w_i^{(\nu)} \log p + \left(1 - w_i^{(\nu)}\right) \log(1-p) \Bigg] \\ & + \text{ const.} \end{aligned}$$

- M-step: maximizing $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y}\right)$ is simple. When the "complete" data likelihood has the form of an exponential family, M-step is straightforward.

Example 2: right-censored data
Let $(Y_1, \delta_1), \cdots, (Y_n, \delta_n)$ be i.i.d where $Y_i = \min(X_i, R_i)$ and $\delta_i = I(X_i \leq R_i)$.

$$\begin{aligned} X_i \sim f(x; \boldsymbol{\theta}) &= \sigma^{-1} \exp(-x/\sigma) \\ L(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\delta}) &\propto \prod_{i=1}^{n} f\left(Y_i; \boldsymbol{\theta}\right)^{\delta_i} \left\{1 - F\left(Y_i; \boldsymbol{\theta}\right)\right\}^{1-\delta_i} \\ &= \sigma^{-n_u} \exp\left(-\sum_{i=1}^{n} Y_i/\sigma\right) \end{aligned}$$

cause $F(Y_i;\ \theta) = 1 - \exp(-x/\sigma)$, where n_μ is the number of uncensored observations. The likelihood is easy to maximize. Nonetheless, let $Z_i = X_i$

$$\begin{aligned} \ell(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\delta}, \mathbf{X}) &= \sum_{i=1}^{n} \log f\left(X_i; \boldsymbol{\theta}\right) \\ &= \sum_{i=1}^{n_u} \log f\left(X_i; \boldsymbol{\theta}\right) + \sum_{i=n_u+1}^{n} \log f\left(X_i; \boldsymbol{\theta}\right) \end{aligned}$$

- E-step

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y}, \boldsymbol{\delta}\right) &= \sum_{i=1}^{n_u} \log f\left(Y_i; \boldsymbol{\theta}\right) + \sum_{i=n_u+1}^{n} E_{\boldsymbol{\theta}^{(\nu)}} \left\{\log f\left(X_i; \boldsymbol{\theta}\right) \mid Y_i, \delta_i\right\} \\ &= -n \log \sigma - \sigma^{-1} \sum_{i=1}^{n_u} Y_i - \sigma^{-1} \sum_{i=n_u+1}^{n} E_{\sigma^{(\nu)}}\left(X_i \mid X_i > Y_i\right) \\ &= -n \log \sigma - \sigma^{-1} \sum_{i=1}^{n} Y_i - \sigma^{-1}\left(n - n_u\right) \sigma^{(\nu)} \end{aligned}$$

  things are simplied here by assuming R_i is not random

- M-step

$$\sigma^{(\nu+1)} = n^{-1} \left\{ \sum_{i=1}^{n} Y_i + \left(n - n_u\right) \sigma^{(\nu)} \right\}$$

  as $\nu \to \infty$, we expect $\sigma^{(\nu+1)} \to \widehat{\sigma}_{\text{MLE}}$ and $\sigma^{(\nu)} \to \widehat{\sigma}_{\text{MLE}}$ thus

$$\widehat{\sigma}_{\text{MLE}} = n^{-1} \left\{ \sum_{i=1}^{n} Y_i + \left(n - n_u\right) \widehat{\sigma}_{\text{MLE}} \right\} \Longrightarrow \widehat{\sigma}_{\text{MLE}} = \sum_{i=1}^{n} Y_i / n_u$$

2.9.1 Why does EM work?

- Jensen's inequality: for a random variable X and convex function $\psi(\cdot)$, $\psi\{E(X)\} \leq E\{\psi(X)\}$
- for $\mathbf{Y} \sim f\left(\mathbf{y}; \boldsymbol{\theta}_0\right), E_{\boldsymbol{\theta}_0}\{\log f(\mathbf{Y}; \boldsymbol{\theta})\}$ is maximized at $\theta = \theta_0$ *proof:* $\psi(x) = -\log(x)$ is a convex function for $x \in (0, \infty)$. Therefore, by Jensen's inequality

$$-\log\left[E_{\boldsymbol{\theta}_0}\left\{\frac{f(\mathbf{Y}; \boldsymbol{\theta})}{f\left(\mathbf{Y}; \boldsymbol{\theta}_0\right)}\right\}\right] \leq -E_{\boldsymbol{\theta}_0}\left[\log\left\{\frac{f(\mathbf{Y}; \boldsymbol{\theta})}{f\left(\mathbf{Y}; \boldsymbol{\theta}_0\right)}\right\}\right]$$

note that

$$E_{\theta_0}\left\{\frac{f(Y;\theta)}{f(Y;\theta_0)}\right\}=\int\frac{f(Y;\theta)}{f(Y;\theta_0)}f(Y;\theta_0)\,dy=1$$

therefore the left side is 0

$$-E_{\theta_0}\left[\log\left\{\frac{f(Y;\ \theta)}{f(Y;\theta_0)}\right\}\right]\geqslant 0\Rightarrow E_{\theta_0}\left[\log\left\{\frac{f(Y;\theta_0)}{f(Y;\theta)}\right\}\right]\leq 0$$

$$\Rightarrow E_{\theta_0}\left[\log f(Y;\theta)\right]\leqslant E_{\theta_0}\left[\log f(Y;\theta_0)\right]\Rightarrow E_{\boldsymbol{\theta}_0}\{\log f(\mathbf{Y};\boldsymbol{\theta})\}\text{ is maximized at }\boldsymbol{\theta}=\boldsymbol{\theta}_0$$

- now go back to the EM algorithm

$$\ell(\boldsymbol{\theta}\mid\mathbf{Y})\equiv\log f(\mathbf{Y}\mid\boldsymbol{\theta})=\log f(\mathbf{Y},\mathbf{Z}\mid\boldsymbol{\theta})-\log f(\mathbf{Z}\mid\mathbf{Y},\boldsymbol{\theta})$$

cause $f(y,z;\theta)=f(z\mid y;\theta)f(y;\theta)$. Taking expectation on both sides treating Z as a random variable with density $f\left(\mathbf{Z}\mid\mathbf{Y},\boldsymbol{\theta}^{(\nu)}\right)$

$$\ell(\boldsymbol{\theta}\mid\mathbf{Y})=Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(\nu)},\mathbf{Y}\right)-E_{\boldsymbol{\theta}^{(\nu)}}\{\log f(\mathbf{Z}\mid\mathbf{Y},\boldsymbol{\theta})\mid\mathbf{Y}\}$$

where based on the previous step $E_{\boldsymbol{\theta}^{(\nu)}}\{\log f(\mathbf{Z}\mid\mathbf{Y},\boldsymbol{\theta})\mid\mathbf{Y}\}$ is maximized at $\boldsymbol{\theta}^{(\nu)}$. In performing the EM algorithm, we ensure that $Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(\nu)},\mathbf{Y}\right)\geq Q\left(\boldsymbol{\theta}^{(\nu)},\boldsymbol{\theta}^{(\nu)},\mathbf{Y}\right)$ by maximizing $Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(\nu)},\mathbf{Y}\right)$. Therefore $\ell\left(\boldsymbol{\theta}^{(\nu+1)}\mid\mathbf{Y}\right)\geq\ell\left(\boldsymbol{\theta}^{(\nu)}\mid\mathbf{Y}\right)$.

2.9.2 Calculating observed info matrix after EM

The general way is

$$\begin{aligned}\mathbf{S}_{\mathbf{Y}}(\mathbf{Y},\boldsymbol{\theta})&=\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\theta})=\frac{f'_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\theta})^T}{f_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\theta})}\\ \mathbf{I}_{\mathbf{Y}}(\mathbf{Y},\boldsymbol{\theta})&=-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\theta})\\ &=-\frac{f''_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\theta})}+\mathbf{S}_{\mathbf{Y}}(\mathbf{Y},\boldsymbol{\theta})^{\otimes 2}\end{aligned}$$

Actually, there is a better way than direct computation. Note for the complete data $f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z};\boldsymbol{\theta})$

$$\begin{aligned}\mathbf{S}_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z},\boldsymbol{\theta})&=\frac{f'_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z};\boldsymbol{\theta})^T}{f_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z};\boldsymbol{\theta})}\\ \mathbf{I}_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z},\boldsymbol{\theta})&=-\frac{f''_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z};\boldsymbol{\theta})}{f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y};\boldsymbol{\theta})}+\mathbf{S}_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z},\boldsymbol{\theta})^{\otimes 2}\end{aligned}\tag{4}$$

given that the complete density is related to the density we are interested in, we have and based on 4 we have

$$\begin{aligned}f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})&=\int f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}\\ \mathbf{S}_{\mathbf{Y}}(\mathbf{Y},\boldsymbol{\theta})&=\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\int f_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}\\ &=\frac{\int f'_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})^T d\mathbf{z}}{\int f_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}}\\ &=\frac{\int\mathbf{S}_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z},\boldsymbol{\theta})f_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}}{\int f_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}}\\ &=E_{\boldsymbol{\theta}}\left\{\mathbf{S}_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z},\boldsymbol{\theta})\mid\mathbf{Y}\right\}\\ \mathbf{I}_{\mathbf{Y}}(\mathbf{Y},\boldsymbol{\theta})&=-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log\int f_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}\\ &=-\frac{\int f''_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}}{\int f_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}}+\mathbf{S}_{\mathbf{Y}}(\mathbf{Y},\boldsymbol{\theta})^{\otimes 2}\\ &=E_{\boldsymbol{\theta}}\left\{\mathbf{I}_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z},\boldsymbol{\theta})\mid\mathbf{Y}\right\}\\ &\quad-E_{\boldsymbol{\theta}}\left\{\mathbf{S}_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z},\boldsymbol{\theta})^{\otimes 2}\mid\mathbf{Y}\right\}\\ &\quad+\mathbf{S}_{\mathbf{Y}}(\mathbf{Y},\boldsymbol{\theta})^{\otimes 2}\end{aligned}$$

where $\mathbf{I}_{\mathbf{Y},\mathbf{Z}}$ and $\mathbf{Z}_{\mathbf{Y},\mathbf{Z}}$ only need to be computed at the last iteration of the EM procedure where $\mathbf{S}_{\mathbf{Y}}(\mathbf{Y},\widehat{\boldsymbol{\theta}})=\mathbf{0}$.

2.10 Uniqueness of MLE

Converge to boundary: sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots$ in Θ is said to converge to boundary $\partial\Theta$ if for every compact set (If a closed set A is bounded, then A is a compact set) $K \subset \Theta$, there exists $k_0 \geq 1$ such that $\boldsymbol{\theta}^{(k)} \notin K \quad \forall k \geq k_0$ (if $\Theta = R^b$, then this definition is equivalent to $\lim_{k\to\infty} \left\|\boldsymbol{\theta}^{(k)}\right\| = \infty$)
Constant on boundary: a real-valued function f defined on Θ is said to be constant on boundary $\partial\Theta$ if $\lim_{k\to\infty} f\left(\boldsymbol{\theta}^{(k)}\right) = c$ for every sequence $\boldsymbol{\theta}^{(k)}$ in Θ converging to $\partial\Theta$; c could be $\pm\infty$. Written as $\lim_{\boldsymbol{\theta}\to\partial\Theta} f(\boldsymbol{\theta}) = c$.

Theorem 2.1. *for $\Theta \subset R^b, b \geq 1$ which is a connected open set. If*

1. *$\ell(\boldsymbol{\theta})$ is twice continuously differentiable with $\lim_{\boldsymbol{\theta}\to\partial\Theta} \ell(\boldsymbol{\theta}) = c$, where constant c is either a real number or $-\infty$*
2. *$\mathbf{I}_T(\mathbf{Y}, \boldsymbol{\theta}) = -\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ (observed total information matrix) is positive definite at every point $\boldsymbol{\theta} \in \Theta$ for which $\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T = \mathbf{0}$*

then

1. *the critical point is unique and is the MLE*
2. *$\ell(\boldsymbol{\theta}) > c, \ \forall\boldsymbol{\theta} \in \Theta$*

Recall: Note a matrix A is positive definite if $\mathbf{x}^T A\mathbf{x} > 0, \ \forall x$. A is positive definite if and only if

$$a_{11} > 0; \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0; \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0; \cdots; |A| > 0$$

Note: the constancy on boundary is not really necessary for the scalar parameter but is need for other cases

- with a scalar parameter, the second conditional that minus Hessian is positive definite at all stationary points (x at which $f'(x) = 0$) $\Rightarrow$ absence of local minima $\Rightarrow$ absence of multiple local maxima (multiple local maxima cannot occur without local minima)
- counter-example showing the necessity of "constancy on boundary"

  $$\begin{aligned} \ell = g(x, y) &= -e^{-2y} - e^{-y}\sin x \\ \tfrac{\partial}{\partial x} g(x, y) &= -e^{-y}\cos x \\ \tfrac{\partial}{\partial y} g(x, y) &= 2e^{-2y} + e^{-y}\sin x \end{aligned}$$

  solution: $x = (2k + 1.5)\pi, k = 1, 2, \cdots,$ and $y = \log 2$

  $$-\nabla^2 g(x, y) = \begin{pmatrix} -e^{-y}\sin x & -e^{-y}\cos x \\ -e^{-y}\cos x & 4e^{-2y} + e^{-y}\sin x \end{pmatrix} \overset{\text{at stationary points}}{=} \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$$

  which is positive definite $\Rightarrow$ all solutions corresponding to maxima, which is the solutions are all maxima and there are countably infinite of them cause the "constancy on boundary" is violated

Theorem 2.2. *for $\Theta \subset R^b, b \geq 1$ which is a connected open set and $\ell(\boldsymbol{\theta})$ is twice continuously differentiable. if*

1. *$\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T = \mathbf{0}$ has at least one solution*
2. *$\mathbf{I}_T(\mathbf{Y}, \boldsymbol{\theta})$ is positive definite $\forall\boldsymbol{\theta} \in \Theta$*

then

1. *$\ell(\boldsymbol{\theta})$ is concave*
2. *the solution is unique and is the MLE*

Example 1: normal location-scale model

$$Y_1, \cdots, Y_n \sim \text{iid}\, N\left(\mu, \sigma^2\right), \Theta = (-\infty, \infty) \times (0, \infty)$$

$$\ell(\mu,\sigma) = \text{ const } - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2$$

$$\frac{\partial}{\partial\boldsymbol{\theta}^T}\ell(\mu,\sigma) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu) \\ -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu)^2 \end{pmatrix}$$

$$\mathbf{I}_T(\mu,\sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{2}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu) \\ \frac{2}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu) & -\frac{n}{\sigma^2}+\frac{3}{\sigma^4}\sum_{i=1}^{n}(Y_i-\mu)^2 \end{pmatrix}$$

only solution to $\partial\ell(\mu,\sigma)/\partial\boldsymbol{\theta}^T$ is $\widehat{\mu}=\bar{Y},\widehat{\sigma}=s_n$. Use 2.1 we can show that 1) $\mathbf{I}_T\left(\bar{Y},s_n\right)=\text{diag}\left(n/s_n^2,2n/s_n^2\right)$ is positive definite, and 2) $\lim_{\boldsymbol{\theta}\to 2\Theta}\ell(\boldsymbol{\theta})=-\infty$ (HW3 2.58)
2.2 is not applicable cause $\mathbf{I}_T(\mu,\sigma)$ is not positive definite everywhere. $|\mathbf{I}_T(\mu,\sigma)|=n^2\sigma^{-6}\left\{3s_n^2-\sigma^2-(\bar{Y}-\mu)^2\right\}$ could be negative.
Example 2: Exponential threshold model (combine with the profile likelihood approach)

$$f(y;\mu,\sigma)=\frac{1}{\sigma}\exp\left(-\frac{y-\mu}{\sigma}\right),\quad \mu\le y<\infty$$

with $Y_i's$ are i.i.d and $\Theta=(-\infty,\infty)\times(0,\infty)$

$$L(\mu,\sigma\mid\mathbf{Y})=\sigma^{-n}\exp\left(-\sum_{i=1}^{n}\frac{Y_i-\mu}{\sigma}\right)\prod_{i=1}^{n}I\left(\mu\le Y_i\right)$$

which is not differentiable in μ everywhere. By using the profile likelihood method. For each σ, $L(\mu,\sigma\mid\mathbf{Y})$ is maximized at $\hat{\mu}=Y_{(1)}\Rightarrow$ log-profile likelihood $\ell^*(\sigma)=-n\log\sigma-\sigma^{-1}\sum_i\left(Y_i-Y_{(1)}\right)$

$$\frac{\partial}{\partial\sigma}\ell^*(\sigma)=-n\sigma^{-1}+n\sigma^{-2}\left(\bar{Y}-Y_{(1)}\right)\Longrightarrow\widehat{\sigma}=\bar{Y}-Y_{(1)}\text{ if }\bar{Y}>Y_{(1)}$$

$$-\frac{\partial^2}{\partial\sigma^2}\ell^*(\sigma)=-n\sigma^{-2}+2n\sigma^{-3}\left(\bar{Y}-Y_{(1)}\right)$$

is not always positive thus 2.2 does not apply. Nonetheless, 2.1 does

1. $-\partial^2\ell^*(\sigma)/\left.\partial\sigma^2\right|_{\sigma=\widehat{\sigma}}=n\widehat{\sigma}^{-2}>0$
2. $\lim_{\sigma\to 0}\ell^*(\sigma)=\lim_{\sigma\to\infty}\ell^*(\sigma)=-\infty$

Example 3: example to show that existence and uniqueness of the MLE is not always a given

$$Y_1,\cdots,Y_n\overset{i.i.d}{\sim}\text{mixture of normals}$$

$$f(y;\mu,\sigma,p)=p\phi(y-\mu)+(1-p)\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)$$

the log-likelihood assuming $\mu=Y_1$

$$\ell=\log\left\{p\phi(0)+(1-p)\frac{1}{\sigma}\phi(0)\right\}+\sum_{i=2}^{n}\log\left\{p\phi\left(Y_i-Y_1\right)+(1-p)\frac{1}{\sigma}\phi\left(\frac{Y_i-Y_1}{\sigma}\right)\right\}$$

as $\sigma\to 0$ the first term $\to\infty$ and the rest are bounded. Thus, MLE does not exist in the strict sense. Furthermore, there exist multiple local maxima.
Nonetheless, local maxima satisfying the likelihood score tend to behave well. EM algorithm can find at least one of them.
Uniqueness of the MLE for exponential family
X $\sim$ minimal exponential family

$$f(x;\boldsymbol{\theta})=h(x)\exp\left\{\sum_{i=1}^{s}g_i(\boldsymbol{\theta})T_i(x)-B(\boldsymbol{\theta})\right\}=h(x)\exp\left\{\sum_{i=1}^{k}\eta_iT_i(\eta)-A(\eta)\right\}$$

where $g(\boldsymbol{\theta})$ is 1-to-1, twice differentiable in Θ and Θ is an open subset of R^s. Then if there is at least one solution to the transformed likelihood equation $E_{\boldsymbol{\theta}}\{\mathbf{T}(X)\}=\mathbf{T}(x)$ then with 2.2 the solution is unique and is the MLE.
With the canonical representation, the likelihood score is

$$\frac{\partial}{\partial\eta^T}A(\boldsymbol{\eta})-\mathbf{T}(x)=0\Rightarrow\frac{\partial}{\partial\eta^T}A(\boldsymbol{\eta})\equiv E_{\boldsymbol{\eta}}\mathbf{T}(X)=\mathbf{T}(x)$$

the minus Hessian is

$$\frac{\partial^2}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} A(\boldsymbol{\eta}) \equiv \mathbf{I}(\boldsymbol{\eta}) = \mathrm{Var}_{\boldsymbol{\eta}}\{\mathbf{T}(X)\}$$

which is positive definite as T is affinely independent. Then based on invariant property so is $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = g^{-1}\left(\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}}\right)$

Families with truncation or threshold parameters

$$x_1, \cdots, x_n \overset{i.i.d}{\sim} f\left(x; \boldsymbol{\theta}, \mu_1, \mu_2\right)$$

$$f\left(x; \boldsymbol{\theta}, \mu_1, \mu_2\right) = c\left(\mu_1, \mu_2, \boldsymbol{\theta}\right) d(x, \boldsymbol{\theta}) \quad \mu_1 < x < \mu_2$$

for fixed θ, $\left(X_{(1)}, X_{(n)}\right)$ is minial sufficent for (μ_1, μ_2). Conditional on $\left(X_{(1)}, X_{(n)}\right) = \left(x_{(1)}, x_{(n)}\right)$, sample values between $X_{(1)}$ and $X_{(n)}$ say

$$z_1, \cdots, z_{n-2} \overset{i.i.d}{\sim} q(z, \boldsymbol{\theta}) = \frac{d(z, \boldsymbol{\theta}) I\left(x_{(1)} < z < x_{(n)}\right)}{\int_{x_{(1)}}^{x_{(n)}} d(z, \boldsymbol{\theta}) dz}$$

if $d(x, \boldsymbol{\theta})$ has an exponential family form then so does $q(z, \boldsymbol{\theta})$
Similar results for

$$f(x; \boldsymbol{\theta}, \mu) = c_1(\mu, \boldsymbol{\theta}) d_1(x, \boldsymbol{\theta}) \quad \mu < x$$
$$f(x; \boldsymbol{\theta}, \mu) = c_2(\mu, \boldsymbol{\theta}) d_2(x, \boldsymbol{\theta}) \quad x < \mu$$

Example 1:

$$f(x; \theta, \mu) = \theta e^{-\theta(x-\mu)} I(\mu < x)$$

conditional on $X_{(1)} = x_{(1)}$, observations larger than $x_{(1)}$ have density

$$q(z, \theta) = \theta e^{-\theta\left(z - x_{(1)}\right)} I\left(x_{(1)} < z\right) = e^{-\theta z + \theta x_{(1)} + \log(\theta)} I\left(x_{(1)} < z\right)$$

3 Likelihood-based tests and confidence regions

Let's first look at the simplest scalar parameter case

$$H_0 : \theta = \theta_0 \text{ vs. } H_a : \theta \neq \theta_0$$

- Wald test

$$T_{\mathrm{W}} = \frac{\left(\widehat{\theta}_{\mathrm{MLE}} - \theta_0\right)^2}{\left\{I_{\mathrm{T}}\left(\widehat{\theta}_{\mathrm{MLE}}\right)\right\}^{-1}}$$

- Likelihood ratio test

$$T_{\mathrm{LR}} = -2\left\{\ell\left(\theta_0\right) - \ell\left(\widehat{\theta}_{\mathrm{MLE}}\right)\right\}$$

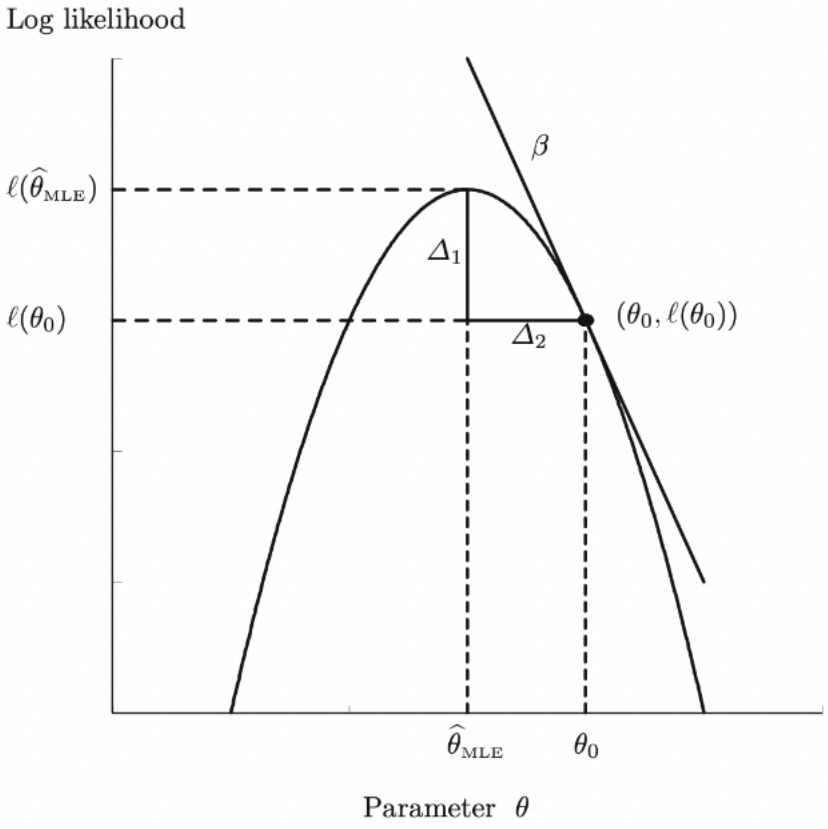
- Score test

$$T_{\mathrm{S}} = \frac{S\left(\theta_0\right)^2}{I_{\mathrm{T}}\left(\theta_0\right)}$$

  Under H_0, asymptotically they are all χ_1^2. Under local alternatives, they have identical asymptotic non-central χ^2 distribution.

The idea behind those tests are graphically illustrated below

Figure 2: Graphical representation of the relationships between Wald, Score and Likelihood Ratio test statistics



The Likelihood Ratio test statistic is a multiple of the difference,Δ_1; the Wald test statistic is a multiple of the squared difference, Δ_2^2; and the Score test statistic is a multiple of the squared slope β^2 (this figure is from the textbook)

3.1 Simple null hypothesis

By "simple" we mean don't have nuisance parameter which is we are interested in the whole parameter vector.

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ vs. } H_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \quad b\text{-dimensional}$$

- Wald test

  $$T_{\mathrm{W}} = \left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta}_0\right)^T \mathbf{I}_{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta}_0\right)$$

  where $\mathbf{I}_{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)$ may be replace by $\mathbf{I}_{\mathrm{T}}\left(\boldsymbol{\theta}_0\right)$ or $\mathbf{I}_{\mathrm{T}}\left(\mathbf{Y}, \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)$

- Likelihood ratio test

  $$T_{\mathrm{LR}} = -2\left\{\ell\left(\boldsymbol{\theta}_0\right) - \ell\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)\right\}$$

  By taylor expansion

  $$\ell\left(\boldsymbol{\theta}_0\right) = \ell\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) + \mathbf{S}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)^T\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) - \frac{1}{2}\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)^T \mathbf{I}_{\mathrm{T}}\left(\mathbf{Y}, \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) + \text{residual}$$

  here $\mathbf{S}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right)^T = 0$ cause this is how MLE is derived. Therefore

  $$T_{\mathrm{W}} + \delta_n \text{ with } \delta_n \xrightarrow{p} 0$$

  where δ_n is the residual

- Score test

  $$T_{\mathrm{S}} = \mathbf{S}\left(\boldsymbol{\theta}_0\right)^T \mathbf{I}_{\mathrm{T}}\left(\boldsymbol{\theta}_0\right)^{-1} \mathbf{S}\left(\boldsymbol{\theta}_0\right)$$

  Under H_0, $E\left\{\mathbf{S}\left(\boldsymbol{\theta}_0\right)\right\} = 0$ and $\mathrm{var}\left\{\mathbf{S}\left(\boldsymbol{\theta}_0\right)\right\} = \mathbf{I}_{\mathrm{T}}\left(\boldsymbol{\theta}_0\right)$. By central limit theorem and continuous mapping theorem $T_{\mathrm{S}} \xrightarrow{d} \chi_b^2$.

Example 1:

$$Y_1, \cdots, Y_n \overset{\text{i.i.d}}{\sim} N(\mu, 1). H_0 : \mu = \mu_0$$

$$\ell(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_i (Y_i - \mu)^2$$

$$S(\mu) = \frac{\partial}{\partial \mu}\ell(\mu) = \sum_i (Y_i - \mu) \quad \Longrightarrow \widehat{\mu} = \bar{Y}$$

$$I_{\mathrm{T}}(\mathbf{Y}, \mu) = -\frac{\partial}{\partial \mu} S(\mu) = n$$

we got the test statistics as follows

$$T_{\mathrm{W}} = \left(\bar{Y} - \mu_0\right)(n)\left(\bar{Y} - \mu_0\right) = n\left(\bar{Y} - \mu_0\right)^2$$

$$T_{\mathrm{S}} = \left\{\sum_i (Y_i - \mu_0)\right\} n^{-1} \left\{\sum_i (Y_i - \mu_0)\right\} = T_{\mathrm{W}}$$

$$T_{\mathrm{LR}} = -2\left\{-\frac{1}{2}\sum_i (Y_i - \mu_0)^2 + \frac{1}{2}\sum_i \left(Y_i - \bar{Y}\right)^2\right\} = T_{\mathrm{W}}$$

Example 2: $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. $H_0 : p = p_0$. Let $X = \sum_i Y_i$

$$L = \prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)}$$

$$\ell = \sum_{i=1}^{n} y_i \log(p) + (1 - y_i)\log(1-p) = X\log(p) + (n - X)\log(1-p)$$

$$S(p) = \frac{X}{p} - \frac{n-X}{1-p} = \frac{X - np}{p(1-p)} \quad \Longrightarrow \widehat{p} = \frac{X}{n} \Longrightarrow X = n\hat{p}$$

$$I_{\mathrm{T}}(\mathbf{Y}, p) = \frac{X}{p^2} + \frac{n-X}{(1-p)^2} \quad I_{\mathrm{T}}(p) = \frac{n}{p(1-p)}$$

$$T_{\mathrm{W}} = (\widehat{p} - p_0)\frac{n}{\widehat{p}(1-\widehat{p})}(\widehat{p} - p_0) = \frac{n(\widehat{p} - p_0)^2}{\widehat{p}(1-\widehat{p})}$$

$$T_{\mathrm{S}} = S(p_0)\, I_{\mathrm{T}}(p_0)^{-1} S(p_0) = \frac{n(\widehat{p} - p_0)^2}{p_0(1-p_0)} \; X \text{ is replaced by } n\hat{p}$$

$$T_{\mathrm{LR}} = -2\left[X\log\left(p_0/\widehat{p}\right) + (n-X)\log\left\{\left(1-p_0\right)/\left(1-\widehat{p}\right)\right\}\right]$$

3.2 Composite null hypothesis

In this case, $\boldsymbol{\theta}$ is partitioned and we are only interested in part of it.

$$\underset{b\times 1}{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ r \times 1 \\ \boldsymbol{\theta}_2 \\ (b-r)\times 1 \end{pmatrix}$$

$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ vs. $H_a : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$, with $\boldsymbol{\theta}_2$ as nuisance.

- Wald test

  $$\mathbf{I}_{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) \equiv \widehat{\mathbf{I}}_{\mathrm{T}} = \begin{pmatrix} \widehat{\mathbf{I}}_{\mathrm{T},11} & \widehat{\mathbf{I}}_{\mathrm{T},12} \\ \widehat{\mathbf{I}}_{\mathrm{T},21} & \widehat{\mathbf{I}}_{\mathrm{T},22} \end{pmatrix}$$

  Under $H_0, \mathrm{Avar}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)$ is the upper $(1,1)$ element of $\widehat{\mathbf{I}}_{\mathrm{T}}^{-1}$, given by

  $$\left(\widehat{\mathbf{I}}_{\mathrm{T},11} - \widehat{\mathbf{I}}_{\mathrm{T},12}\widehat{\mathbf{I}}_{\mathrm{T},22}^{-1}\widehat{\mathbf{I}}_{\mathrm{T},21}\right)^{-1}$$

  therefore

  $$T_{\mathrm{W}} = \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)^T \left(\widehat{\mathbf{I}}_{\mathrm{T},11} - \widehat{\mathbf{I}}_{\mathrm{T},12}\widehat{\mathbf{I}}_{\mathrm{T},22}^{-1}\widehat{\mathbf{I}}_{\mathrm{T},21}\right)\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right) \overset{d}{\to} \chi_r^2$$

  be careful here, don't write the middle part as $(\widehat{\mathbf{I}}_{\mathrm{T},11})^{-1}$

- Likelihood ratio test

$$T_{\mathrm{LR}} = -2\log\left\{\frac{\sup_{\boldsymbol{\theta}\in H_0} L(\boldsymbol{\theta}\mid \mathbf{Y})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}\mid \mathbf{Y})}\right\} = -2\{\ell(\widetilde{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}})\}$$

  where $\sup_{\boldsymbol{\theta}\in H_0} L(\boldsymbol{\theta}\mid \mathbf{Y})$ is the restricted MLE and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}\mid \mathbf{Y})$ is the unrestricted MLE

- Score test
  First derive the H_0-restricted MLE

$$\widetilde{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{\theta}_{10} \\ \widetilde{\boldsymbol{\theta}}_2 \end{pmatrix} \quad \widetilde{\boldsymbol{\theta}}_2 = \arg\max_{\boldsymbol{\theta}_2} \ell\left(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2\right)$$

$$\mathbf{I}_{\mathrm{T}}\left(\tilde{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) \equiv \tilde{\mathbf{I}}_{\mathrm{T}}$$

$$\begin{aligned}
\mathbf{S}(\boldsymbol{\theta}) &= \begin{pmatrix} \mathbf{S}_1(\boldsymbol{\theta}) \\ \mathbf{S}_2(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}_1^T}\ell(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \boldsymbol{\theta}_2^T}\ell(\boldsymbol{\theta}) \end{pmatrix} \\
T_{\mathrm{S}} &= \mathbf{S}(\widetilde{\boldsymbol{\theta}})^T \widetilde{\mathbf{I}}_{\mathrm{T}}^{-1} \mathbf{S}(\widetilde{\boldsymbol{\theta}}) \\
&= \left(\mathbf{S}_1\left(\widetilde{\boldsymbol{\theta}}\right)^T, \mathbf{0}^T\right) \begin{pmatrix} \widetilde{\mathbf{I}}_{\mathrm{T},11} & \widetilde{\mathbf{I}}_{\mathrm{T},12} \\ \widetilde{\mathbf{I}}_{\mathrm{T},21} & \widetilde{\mathbf{I}}_{\mathrm{T},22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}_1(\widetilde{\boldsymbol{\theta}}) \\ \mathbf{0} \end{pmatrix} \\
&= \mathbf{S}_1\left(\widetilde{\boldsymbol{\theta}}\right)^T \left(\widetilde{\mathbf{I}}_{\mathrm{T},11} - \widetilde{\mathbf{I}}_{\mathrm{T},12}\widetilde{\mathbf{I}}_{\mathrm{T},22}^{-1}\widetilde{\mathbf{I}}_{\mathrm{T},21}\right)^{-1} \mathbf{S}_1(\widetilde{\boldsymbol{\theta}})
\end{aligned}$$

  where $\mathbf{S}_2(\widetilde{\boldsymbol{\theta}}) = 0$ by definition of MLE

Composite null hypotheses of general form $H_0 : \underset{r\times 1}{\mathbf{h}(\boldsymbol{\theta})} = \mathbf{0}$

$$\underset{r\times b}{\mathbf{H}}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{h}(\boldsymbol{\theta}), \quad r \le b$$

which needs to be of *full rank*
Below are some examples to help understand the definition of h and H *Example 1:* bi-variate normal data

$$\begin{aligned}
\boldsymbol{\theta} &= (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)^T \\
H_0 &: \mu_1 = \mu_2 \\
h(\boldsymbol{\theta}) &= \mu_1 - \mu_2 \\
\mathbf{H}(\boldsymbol{\theta}) &= (1, -1, 0, 0, 0)
\end{aligned}$$

Example 2: linear hypotheses $\mathbf{K}^T\beta = \mathbf{m}$

$$\mathbf{h}(\beta) = \mathbf{K}^T\beta - \mathbf{m}$$

Example 3: partitioned-vector hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$

$$\mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}$$

Now let's jump in to the test statistics

- Wald test

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &\sim AN\left(\boldsymbol{\theta}, \mathbf{I}_{\mathrm{T}}(\boldsymbol{\theta})^{-1}\right) \\
\mathbf{h}(\widehat{\boldsymbol{\theta}}) &\sim AN\left\{\mathbf{h}(\boldsymbol{\theta}), \mathbf{H}(\boldsymbol{\theta})\mathbf{I}_{\mathrm{T}}(\boldsymbol{\theta})^{-1}\mathbf{H}(\boldsymbol{\theta})^T\right\} \\
T_{\mathrm{W}} &= \mathbf{h}(\widehat{\boldsymbol{\theta}})^T \left\{\mathbf{H}(\widehat{\boldsymbol{\theta}})\mathbf{I}_{\mathrm{T}}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{H}(\widehat{\boldsymbol{\theta}})^T\right\}^{-1} \mathbf{h}(\widehat{\boldsymbol{\theta}})
\end{aligned}$$

  from the first line to second line we used Delta method one problem is that the test statistic varies with reparameterization and choice of h, e.g., $h\left(\mu_1, \mu_2\right) = \mu_1 - \mu_2$ and $h\left(\mu_1, \mu_2\right) = \mu_1/\mu_2 - 1$.

- Score test

$$T_{\mathrm{S}} = \mathbf{S}(\widetilde{\boldsymbol{\theta}})^T \widetilde{\mathbf{I}}_{\mathrm{T}}^{-1} \mathbf{S}(\widetilde{\boldsymbol{\theta}})$$

  where $\tilde{\boldsymbol{\theta}}$ maximizes the likelihood subject to $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$.
  Actually a better way to do this is using the Lagrange multiplier for

$$\max \ell(\boldsymbol{\theta}) \text{ subject to } \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$$

we will focus on $\ell(\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta})^T_{r\times 1}\boldsymbol{\lambda}$

$$\mathbf{S}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})^T\boldsymbol{\lambda} = \mathbf{0}$$
$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$$

Denote the solution by $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\lambda}}$ then we get $\mathbf{S}(\tilde{\boldsymbol{\theta}}) = \mathbf{H}(\widetilde{\boldsymbol{\theta}})^T\tilde{\boldsymbol{\lambda}}$ and the score test statistic turns into

$$T_{\mathrm{S}} = \tilde{\boldsymbol{\lambda}}^T\mathbf{H}(\widetilde{\boldsymbol{\theta}})\tilde{\mathbf{I}}_{\mathrm{T}}^{-1}\mathbf{H}(\widetilde{\boldsymbol{\theta}})^T\tilde{\boldsymbol{\lambda}}$$

Below is a comparison of the three method

	invariant to re-parameterization and the choice of $\mathbf{h}(\cdot)$	computing MLE's	
	unrestricted	restricted	
Wald	N	Y	N
Score	Y	N	Y
LR	Y	Y	Y

Nonetheless, LR can be convenient for nested models. Typically, Wald does not have as good type I error as score and LR. And Wald and score are easier to adjust in the case of model misspecification.

Example 1: Normal location-scale model $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)$ and $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ σ unrestricted

$$\ell(\mu, \sigma) = \text{constant} - n\log\sigma - \frac{1}{2\sigma^2}\sum_i \left(Y_i - \mu\right)^2$$

$$\mathbf{S}(\mu,\sigma) = \sum_i \begin{bmatrix} \sigma^{-2}\left(Y_i - \mu\right) \\ \sigma^{-3}\left\{\left(Y_i-\mu\right)^2 - \sigma^2\right\} \end{bmatrix} \Rightarrow \widehat{\mu} = \bar{Y} \quad \widehat{\sigma}^2 = s_n^2 = n^{-1}\sum_i \left(Y_i - \bar{Y}\right)^2,\ \mathbf{I}_{\mathrm{T}}(\mu,\sigma) = \mathrm{diag}\left(n/\sigma^2, 2n/\sigma^2\right)$$

$$\widetilde{\mu} = \mu_0 \quad \tilde{\sigma}^2 = n^{-1}\sum_i \left(Y_i - \mu_0\right)^2 = s_n^2 + \left(\bar{Y} - \mu_0\right)^2$$

$$T_{\mathrm{w}} = \left(\bar{Y} - \mu_0\right)\frac{n}{s_n^2}\left(\bar{Y} - \mu_0\right) = \frac{n\left(\bar{Y} - \mu_0\right)^2}{s_n^2} = \frac{n}{n-1}\frac{n\left(\bar{Y} - \mu_0\right)^2}{ns_n^2/(n-1)} = \frac{n}{n-1}t^2 \text{ where } t \text{ means t-distribution with } df = 1$$

$$T_{\mathrm{S}} = \left\{\frac{1}{\widetilde{\sigma}^2}\sum_i \left(Y_i - \mu_0\right)\right\}\frac{\tilde{\sigma}^2}{n}\left\{\frac{1}{\widetilde{\sigma}^2}\sum_i \left(Y_i - \mu_0\right)\right\} = \frac{n\left(\bar{Y} - \mu_0\right)^2}{\tilde{\sigma}^2} = \frac{nT_{\mathrm{W}}}{n + T_{\mathrm{W}}}$$

$$T_{\mathrm{LR}} = n\log\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = n\log\left\{1 + \frac{\left(\bar{Y} - \mu_0\right)^2}{s_n^2}\right\} = n\log\left\{1 + \frac{T_{\mathrm{W}}}{n}\right\}$$

Since $\frac{x}{1+x} \leq \log(1+x) \leq x$ for $x > -1$

$$T_{\mathrm{S}} \leq T_{\mathrm{LR}} \leq T_{\mathrm{W}}$$

Using exact distributions, they are all equivalent to t test (can be transformed to t test). Using the asymptotic χ_1^2 critical values, T_{W} is more liberal.

Example 2: Score test for multinomial data = Pearson χ^2 test

$$
\begin{aligned}
&(N_1, \cdots, N_k) \sim \text{Multinomial}\,(n, p_1, \cdots, p_k) \\
&\mathbf{p} = (p_1, \cdots, p_{k-1})^T \quad p_k = 1 - p_1 - \cdots - p_{k-1} \\
&f\,(n_1, \cdots, n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k} \\
&\ell(\mathbf{p}) = \sum_{i=1}^{k-1} N_i \log p_i + N_k \log\left(1 - \sum_{i=1}^{k-1} p_i\right) + \text{ const.} \\
&\mathbf{S}(\mathbf{p}) = \left(\frac{N_1}{p_1} - \frac{N_k}{p_k}, \cdots, \frac{N_{k-1}}{p_{k-1}} - \frac{N_k}{p_k}\right)^T \\
&\mathbf{I}_{\mathrm{T}}(\mathbf{p}) = n\,\mathrm{diag}\left(\frac{1}{p_1}, \cdots, \frac{1}{p_{k-1}}\right) + \frac{n}{p_k}\mathbf{1}^{\otimes 2} \\
&\mathbf{I}_{\mathrm{T}}(\mathbf{p})^{-1} = \left\{\mathrm{diag}(\mathbf{p}) - \mathbf{p}^{\otimes 2}\right\}/n \\
&T_{\mathrm{S}} = \mathbf{S}(\widetilde{\mathbf{p}})^T \mathbf{I}_{\mathrm{T}}(\widetilde{\mathbf{p}})^{-1} \mathbf{S}(\widetilde{\mathbf{p}}) \\
&= n^{-1} \sum_{i=1}^{k} \left(\frac{N_i}{\widetilde{p}_i} - \frac{N_k}{\widetilde{p}_k}\right)^2 \widetilde{p}_i \\
&- n^{-1} \left\{\sum_{i=1}^{k} \left(\frac{N_i}{\widetilde{p}_i} - \frac{N_k}{\widetilde{p}_k}\right) \widetilde{p}_i\right\}^2 \\
&= n^{-1} \sum_{i=1}^{k} \left\{\frac{N_i}{\widetilde{p}_i} - \frac{N_k}{\widetilde{p}_k} - \left(n - \frac{N_k}{\widetilde{p}_k}\right)\right\}^2 \widetilde{p}_i \\
&= \sum_{i=1}^{k} \frac{(N_i - n\widetilde{p}_i)^2}{n\widetilde{p}_i}
\end{aligned}
$$

in the last step, the variance equality

$$\sum a_i^2 p_i - \left(\sum a_i p_i\right)^2 = \sum \left(a_i - \sum a_i p_i\right)^2 p_i \text{ is used.}$$

Example 3: Testing for Hardy-Weinberg equilibrium
Multinomial $k = 3 : p_{AA}, p_{Aa}, p_{aa}$
Under the equilibrium, $p_{AA} = p_A^2, p_{Aa} = 2p_A\,(1 - p_A)$,
$p_{aa} = (1 - p_A)^2$
restricted MLE: $\widetilde{p}_A = (2N_{AA} + N_{Aa})\,/(2n)$

$$
\begin{aligned}
T_{\mathrm{S}} = &\frac{\left(N_{AA} - n\widetilde{p}_A^2\right)^2}{n\widetilde{p}_A^2} + \frac{\left\{N_{Aa} - 2n\widetilde{p}_A\,(1 - \widetilde{p}_A)\right\}^2}{2n\widetilde{p}_A\,(1 - \widetilde{p}_A)} \\
&+ \frac{\left\{N_{aa} - n\,(1 - \widetilde{p}_A)^2\right\}^2}{n\,(1 - \widetilde{p}_A)^2}
\end{aligned}
$$

the degree of freedom is 1 cause

$$h\,(P_{AA}, P_{Aa}) = P_{Aa} - 2P_{AA}^{1/2}\left(1 - P_{AA}^{1/2}\right) \Rightarrow r = 1$$

3.3 Confidence interval

$$
\underset{b \times 1}{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ r \times 1 \\ \boldsymbol{\theta}_2 \\ (b - r) \times 1 \end{pmatrix}
$$

$T\,(\boldsymbol{\theta}_1) =$ test stat with $\boldsymbol{\theta}_1$ as the null value (actually the true value)

$$C_{1-\alpha} = \left\{\boldsymbol{\theta}_1 : T\,(\boldsymbol{\theta}_1) \leq \chi_r^2(1 - \alpha)\right\}$$

Asymptotically, (random) confidence region contains the true parameter value θ_{10} with probability $1 - \alpha$ if $\boldsymbol{\theta}_{10}$ is the true value. If $\boldsymbol{\theta}_{10} \notin C_{1-\alpha}$ then we reject the null hypothesis

$$P\,(\boldsymbol{\theta}_{10} \in C_{1-\alpha}) = P\left\{T\,(\boldsymbol{\theta}_{10}) \leq \chi_r^2(1 - \alpha)\right\}$$

Example 1: Binomial model
We previously derived the test statistics for the binomial model as $T_{\mathrm{W}} = \frac{n(\widehat{p}-p)^2}{\widehat{p}(1-\widehat{p})}$ the solving

$$T_{\mathrm{W}} = \frac{n(\widehat{p}-p)^2}{\widehat{p}(1-\widehat{p})} \leq \chi_1^2(1-\alpha) = z_{1-\alpha/2}^2 \Longrightarrow \left(\widehat{p} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$$

3.4 Nonstandard hypothesis testing problems

If $\frac{\partial}{\partial \boldsymbol{\theta}}\ell(\boldsymbol{\theta} \mid \mathbf{Y})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}} \neq \mathbf{0}$ in probability, typically the asymptotic normality of $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is no longer true and the 3 likelihood-based tests do not have a limiting χ_r^2 null distribution. *Example 1:* Exponential threshold model

$$Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} f(y;\mu) = \begin{cases} e^{-(y-\mu)} & y > \mu \\ 0 & \text{otherwise} \end{cases}$$

and the MLE is $\widehat{\mu}_{\mathrm{MLE}} = Y_{(1)}$. In this case, the score equation is not 0.
Null hypotheses on the boundary of the parameter space

$$\begin{aligned}
&Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} N(\mu, 1) \quad \mu \geq \mu_0 \\
&\widehat{\mu}_{\mathrm{MLE}} = \max\left(\bar{Y}, \mu_0\right) \\
&H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0 \\
&T_{\mathrm{W}} = n\left(\widehat{\mu}_{\mathrm{MLE}} - \mu_0\right)^2 \\
&= \begin{cases} n\left(\bar{Y} - \mu_0\right)^2 & \bar{Y} \geq \mu_0 \\ 0 & \text{otherwise} \end{cases} \\
&T_{\mathrm{S}} = n\left(\bar{Y} - \mu_0\right)^2 \sim \chi_1^2 \text{ under } H_0 \\
&T_{\mathrm{LR}} = T_{\mathrm{W}}
\end{aligned}$$

When a null hypothesis value, say $\boldsymbol{\theta}_0$ lies on the boundary of the parameter space, then maximum likelihood estimators are often truncated at that boundary because by definitionb $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ must lie in the parameter space of $\boldsymbol{\theta}$. Thus $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ equal to the boundary value $\boldsymbol{\theta}_0$ with positive probability and correspondingly T_{LR} zero for those cases. The result is that the limiting distribution of T_{LR} is a mixture of a point mass at zero and a chi-squared distribution.
In this case, $\bar{Y}$ used to be greater than μ_0 with probability $\frac{1}{2}$. Therefore T_{w} or T_{LR} has a null distribution that is an equal mixture of a point mass at 0 and a χ_1^2 :

$$Z^2 I(Z > 0), \quad Z \sim N(0,1)$$

with this modified limiting distribution. The confidence interval shall be modified accordingly: level- α test: reject H_0 with critical value being $1-2\alpha$ quantile of χ_1^2

$$\begin{aligned}
P\left\{Z^2 I(Z>0) \geq \chi_1^2(1-2\alpha)\right\} &= P\left\{Z^2 \geq \chi_1^2(1-2\alpha), Z > 0\right\} \\
&= \frac{1}{2} P\left\{Z^2 \geq \chi_1^2(1-2\alpha)\right\} = \alpha
\end{aligned}$$

Although the score test statistics still follows the chi-square distribution. By following standard approach we can reject if

$$\bar{Y} < \mu_0 - \sqrt{\chi_1^2(\alpha)/n} \text{ or } \bar{Y} > \mu_0 + \sqrt{\chi_1^2(\alpha)/n}$$

however, the $\bar{Y} < \mu_0 - \sqrt{\chi_1^2(\alpha)/n}$ part doesn't make sense cause the parameter space is $[\boldsymbol{\theta}_0, \infty)$ thus a natural solution is to reject when $\sqrt{n}\left(\bar{Y} - \mu_0\right) > z_\alpha$
For cases that are not in the same form as this example, we can try to transfer the test statistics to be like that in this example.
Example 2: Isotonic regression
There are k independent normal samples of size $n_1, \cdots, n_k$, each i.i.d with means $\mu_1, \cdots, \mu_k$, respectively, and common variance σ^2, satisfying

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$$

The MLE minimizes $\sum_{i=1}^k n_i\left(\bar{Y}_i - \mu_i\right)^2$ subject to $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$·. We'd like to test

$$\begin{aligned}
H_0 :&\quad \mu_1 = \cdots = \mu_k \\
H_a :&\quad \mu_1 \leq \cdots \leq \mu_k \text{ with at least 1 strict inequality}
\end{aligned}$$

we can reparameterize as

$$\begin{aligned} &\mu_1 \\ &\Delta\mu_2 = \mu_2 - \mu_1 \\ &\vdots \\ &\Delta\mu_k = \mu_k - \mu_{k-1} \end{aligned}$$

Then, this results in null hypothesis on the boundary:

$$\begin{aligned} H_0: &\quad \Delta\mu_2 = \cdots = \Delta\mu_k = 0 \\ H_a: &\quad \Delta\mu_2 \geq 0, \cdots, \Delta\mu_k \geq 0 \text{ with at least } 1 \text{ strict inequality} \end{aligned}$$

4 Bayesian methods

⋆ this part will not be tested
⋆ this part of the note is combined with notes from *BIS 567 Bayesian Statistics* at Yale to enhance understanding

4.1 Introduction

- Frequentist approach
  - the unknown parameter θ is assume to be constant
  - data $\mathbf{Y} \sim f(\mathbf{y}; \boldsymbol{\theta})$ is considered random
  - estimation can be via MLE, M-estimation, ...
  - hypothesis testing can be likelihood-based,...
  - confidence interval is derived by inverting a test statistic
- Bayesian approach
  - the unknown parameter θ is from $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, the prior distribution
  - data is from $\mathbf{Y} \sim f(\mathbf{y} \mid \boldsymbol{\theta})$
  - estimation is via the posterior density

    $$\pi(\boldsymbol{\theta} \mid \mathbf{Y} = \mathbf{y}) = \frac{f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

    point estimation can use ,e.g., posterior mean
  - $1 - \alpha$ credible region: parameter space with posterior probability $1 - \alpha$

Marginal density (prior predictive density) of Y

$$\mathbf{Y}, m(\mathbf{y}) = \int f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Posterior predictive density

$$\begin{aligned} m\left(\mathbf{y}_{new} \mid \mathbf{Y}\right) &= \int f\left(\mathbf{y}_{new} \mid \boldsymbol{\theta}, \mathbf{Y}\right) \pi(\boldsymbol{\theta} \mid \mathbf{Y}) d\boldsymbol{\theta} \\ &= \int f\left(\mathbf{y}_{new} \mid \boldsymbol{\theta}\right) \pi(\boldsymbol{\theta} \mid \mathbf{Y}) d\boldsymbol{\theta} \text{ when } \mathbf{Y}_{\text{new}} \mathbf{Y} \end{aligned}$$

Where does the prior $\pi(\boldsymbol{\theta})$ come from?

- Subjective Bayesian: personal uncertainty about $\boldsymbol{\theta}$
- From previous knowledge: previous information about $\boldsymbol{\theta}$ (Bayesian analysis is used to combine previous info with current data)
- For convenience: convenient technical density to employ the Bayesian machinery

Example 1: $Y \sim \text{binomial}(n; p)$

$$
\begin{aligned}
f(y \mid p) &= \left(\begin{array}{c} n \\ p \end{array} \right) p^y (1-p)^{n-y} \quad y = 0, 1, \cdots, n \\
\pi(p) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} (\text{ Beta }) \\
m(y) &= \int f(y \mid p)\pi(p)dp = \frac{B(y+\alpha, n-y+\beta)}{B(\alpha,\beta)} \left(\begin{array}{c} n \\ y \end{array} \right) \\
\pi(p \mid \mathbf{Y}) &= \frac{f(y \mid p)\pi(p)}{m(y)} = \frac{\left(\begin{array}{c} n \\ p \end{array} \right) p^y (1-p)^{n-y} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{m(y)} \\
&= \frac{\left(\begin{array}{c} n \\ p \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}}{m(y)} \sim \text{beta}(Y+\alpha, n-Y+\beta)
\end{aligned}
$$

the posterior mean is

$$
E(p \mid \mathbf{Y}) = \frac{Y+\alpha}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\frac{Y}{n}
$$

which shrinks the MLE toward prior mean. Bayesian estimator may have smaller MSE than the MLE.
When $\alpha = \beta = 0$, the prior is improper

$$
\Longrightarrow \text{ posterior beta } (Y, n-Y) \text{ is proper if } 1 \le Y \le n-1 \Longrightarrow \text{ posterior mean } Y/n
$$

Example 2: $\text{Normal}\left(\theta, \sigma_Y^2\right)$

$$
\begin{aligned}
&Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} N\left(\theta, \sigma_Y^2\right) \quad \sigma_Y^2 \text{ known} \\
&\text{prior } \theta \sim N\left(\mu_0, \sigma_0^2\right) \quad \mu_0, \sigma_0^2 \text{ known} \\
&\text{posterior } \theta \mid \mathbf{Y} \sim N\left(\frac{\tau_0\mu_0 + \tau_n\bar{Y}}{\tau_0+\tau_n}, \frac{1}{\tau_0+\tau_n}\right)
\end{aligned}
$$

where precisions $\tau_0 = 1/\sigma_0^2$ and $\tau_n = n/\sigma_Y^2$
95% credible interval

$$
\left(\frac{\tau_0\mu_0 + \tau_n\bar{Y}}{\tau_0+\tau_n} - \frac{1.96}{\sqrt{\tau_0+\tau_n}}, \frac{\tau_0\mu_0 + \tau_n\bar{Y}}{\tau_0+\tau_n} + \frac{1.96}{\sqrt{\tau_0+\tau_n}}\right)
$$

A nice feature of independent data is that one can sequentially update the prior for each additional datum, or all at once. For example, suppose $Y_1, \ldots, Y_n$ are independent with respective densities $f_i\left(y_i \mid \boldsymbol{\theta}\right), i = 1, \ldots, n$
Then, the posterior is

$$
\begin{aligned}
\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) =& \frac{\pi(\boldsymbol{\theta}) \prod_{i=1}^n f_i\left(Y_i \mid \boldsymbol{\theta}\right)}{\int \pi(\boldsymbol{\theta}) \prod_{i=1}^n f_i\left(Y_i \mid \boldsymbol{\theta}\right) d\boldsymbol{\theta}} \\
=& \frac{\pi\left(\boldsymbol{\theta} \mid Y_1\right) \prod_{i=2}^n f_i\left(Y_i \mid \boldsymbol{\theta}\right)}{\int \pi\left(\boldsymbol{\theta} \mid Y_1\right) \prod_{i=2}^n f_i\left(Y_i \mid \boldsymbol{\theta}\right) d\boldsymbol{\theta}} \\
=& \frac{\pi\left(\boldsymbol{\theta} \mid Y_1, \ldots Y_k\right) \prod_{i=k+1}^n f_i\left(Y_i \mid \boldsymbol{\theta}\right)}{\int \pi\left(\boldsymbol{\theta} \mid Y_1, \ldots, Y_k\right) \prod_{i=k+1}^n f_i\left(Y_i \mid \boldsymbol{\theta}\right) d\boldsymbol{\theta}}
\end{aligned}
$$

Here $\pi\left(\theta \mid Y_1\right)$ is the posterior from using the prior and Y_1. It then is used as the prior for the remaining data.
Sufficient statistics often make calculations easier. Because of the factorization theorem, if a sufficient statistic for θ exists, then the posterior depends on the data only through the sufficient statistic.
Estimation and inference: Frequentist uses sampling distribution, whereas Bayesian uses posterior density.

4.2 Bayesian estimator from decision theory perspective

Loss function $L\{\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{Y})\}$: non-negative function of the true parameter and an estimator $\delta(\mathbf{Y})$. For example:

1. squared error loss: $L\{\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{Y})\} = \|\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{Y})\|^2$
2. absolute error loss: $L\{\theta, \delta(\mathbf{Y})\} = |\theta - \delta(\mathbf{Y})|$
3. 0-1 loss: $L\{\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{Y})\} = \left\{ \begin{array}{ll} 0 & \text{if } \boldsymbol{\theta} = \boldsymbol{\delta}(\mathbf{Y}) \\ 1 & \text{otherwise} \end{array} \right.$

Risk: is the average loss over Y

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = \int L\{\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{y})\} f(\mathbf{y} \mid \boldsymbol{\theta}) d\mathbf{y}$$

which is not a single number, but rather a function of θ Several standard frequentist approaches regarding the risk:

1. minimize $R(\boldsymbol{\theta}, \boldsymbol{\delta})$, with squared error loss, over the class of unbiased estimators $\Longrightarrow$ if exists, minimum variance unbiased estimator (MVUE)
2. minimax estimator

   $$\inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \boldsymbol{\delta})$$

   maximize the risk over $\boldsymbol{\theta}$ first then seek an estimator that minimizes the maximum risk

Bayesian approach (take another average over θ)

$$R_{\text{Bayes}}(\pi, \boldsymbol{\delta}) = \int R(\boldsymbol{\theta}, \boldsymbol{\delta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Bayesian estimator δ_{Bayes} minimizes $R_{\text{Bayes}}(\pi, \boldsymbol{\delta})$
We can also minimize the *posterior risk* instead

$$\rho\{\pi, \boldsymbol{\delta}(\mathbf{Y})\} = \int L\{\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{Y})\} \pi(\boldsymbol{\theta} \mid \mathbf{Y}) d\boldsymbol{\theta}$$

But typically $\arg\min \rho\{\pi, \boldsymbol{\delta}(\mathbf{Y})\} \equiv \arg\min R_{\text{Bayes}}(\pi, \boldsymbol{\delta})$

$$\begin{aligned} R_{\text{Bayes}}(\pi, \boldsymbol{\delta}) &= \iint L\{\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{y})\} f(\mathbf{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta} \\ &= \iint L\{\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{y})\} \pi(\boldsymbol{\theta} \mid \mathbf{y}) m(\mathbf{y}) d\mathbf{y} d\boldsymbol{\theta} \\ &= \int \rho\{\pi, \boldsymbol{\delta}(\mathbf{y})\} m(\mathbf{y}) d\mathbf{y} \end{aligned}$$

thus minimize $\rho\{\pi, \boldsymbol{\delta}(\mathbf{Y})\}$ usually minimize $R_{\text{Bayes}}(\pi, \boldsymbol{\delta})$
Different loss function leads to different parameter estimate

- squared error loss $\Longrightarrow$ posterior mean
- absolute error loss $\Longrightarrow$ posterior median
- 0-1 loss $\Longrightarrow$ posterior mode

Bayes estimators with proper priors are generally not unbiased in the frequentist sense. However, they typically have good risk behavior in the frequentist sense. Interestingly, a key technique for finding minimax estimators starts with a Bayes estimator (see Lehmann and Casella, 1998, Ch. 5). Moreover, an admissible estimator (an estimator not uniformly larger in risk compared to any other estimator) must be a Bayes estimator or the limit of Bayes estimators. Thus, Bayes estimators are not only good in terms of Bayes risk but are often of interest to frequentists willing to sacrifice unbiasedness.

4.3 Credible intervals

$1 - \alpha$ *credible interval/ region*: a region of $\boldsymbol{\theta}$ of posterior prob $1 - \alpha$
highest posterior density (HPD) region: the region with minimized volume
region with equal tail probability is often used for a scalar parameter

4.4 Conjugate prior

Conjugate prior: when the data Y has density $f(y \mid \theta)$ and the prior and posterior are from the same family of densities, we say that the prior is conjugate.
$\pi(\boldsymbol{\theta})$ governed by fixed hyperparameters
γ_{prior} $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$ has updated hyperparameters γ_{post} , via a known function of γ_{prior} and $\mathbf{Y}$
Some examples:

- $\text{beta}(\alpha, \beta)$ prior for binomial (n, p) data
- $N\left(\mu_0, \sigma_0^2\right)$ prior for $N\left(\theta, \sigma_Y^2\right)$ data (known σ_Y^2)

- $\text{beta}(\alpha, \beta)$ prior for negative binomial data
- gamma prior for Poisson data
- gamma prior for gamma data
- Pareto prior for Uniform $(0, \theta)$ data

Example 1:

$$\begin{aligned}
\mathbf{Y} &\sim \text{multinomial}\left(n; p_1, \cdots, p_k\right) \\
\pi\left(p_1, \cdots, p_k\right) &\sim \text{Dirichlet}\left(\alpha_1, \cdots, \alpha_k\right) \\
\pi(\mathbf{p}) &= \frac{\prod_{i=1}^k p_i^{\alpha_i - 1}}{B(\boldsymbol{\alpha})}, \quad B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^k \Gamma\left(\alpha_i\right)}{\Gamma\left(\sum_{i=1}^k \alpha_i\right)} \\
\pi(\mathbf{p} \mid \mathbf{Y}) &\propto \prod_{i=1}^k p_i^{N_i} \prod_{i=1}^k p_i^{\alpha_i - 1} \\
\implies \pi(\mathbf{p} \mid \mathbf{Y}) &\sim \text{Dirichlet}\left(\alpha_1 + N_1, \cdots, \alpha_k + N_k\right)
\end{aligned}$$

Any data density having a sufficient statistic of fixed dimension $\forall n$ has a conjugate prior.
Example 2: exponential family canonical form

$$\begin{aligned}
f(\mathbf{y}; \boldsymbol{\eta}) &= h(\mathbf{y}) \exp\left\{\sum_{i=1}^s \eta_i T_i(\mathbf{y}) - A(\boldsymbol{\eta})\right\} \\
\text{conjugate } \pi(\boldsymbol{\eta} \mid \boldsymbol{\gamma}, \lambda) &= K(\boldsymbol{\gamma}, \lambda) \exp\left\{\sum_{i=1}^s \gamma_i \eta_i - \lambda A(\boldsymbol{\eta})\right\} \\
\text{posterior } &\sim \pi(\boldsymbol{\eta} \mid \boldsymbol{\gamma} + \mathbf{T}(\mathbf{y}), \lambda + 1)
\end{aligned}$$

4.5 Noninformative prior

Truly noninformative may not be possible. For example, assigning prior probability to events A, B, and $C = C_1 \cup C_2$. The choice between the following two cases reflects some kind of information

$$\begin{aligned}
&P(A) = P(B) = P(C) = 1/3 \text{ or} \\
&P(A) = P(B) = P\left(C_1\right) = P\left(C_2\right) = 1/4?
\end{aligned}$$

For the case of a location parameter μ taking values on $(-\infty, \infty)$: improper prior $\pi(\mu) = 1$ giving equal weight to all values can be justified on a variety of grounds. An improper prior does not have a finite integral. However, there seems to be no philosophical problem with improper priors as long as they lead to proper posteriors.
For the case of a scale parameter σ taking values in $(0, \infty)$: Jeffreys' suggestion $\pi(\sigma) = 1/\sigma$ which has an invariance argument. . His invariance argument is that any power transformation of σ, say $\gamma = \sigma^a$ has via a change-of-variables, the improper density

$$\gamma = \sigma^a \implies \pi(\gamma) = \frac{1}{\gamma^{1/a}} \left| \frac{\gamma^{1/a - 1}}{a} \right| = \frac{1}{a\gamma}$$

which is similar in form to $1/\sigma$
Combining these last two improper priors, a location-scale family with $(\mu, \sigma) \in (-\infty, \infty) \times (0, \infty)$, suggests using the improper prior

$$\pi(\mu, \sigma) \propto \frac{1}{\sigma}$$

Now, moving to the case of general parameters on continuous parameter spaces. For Jeffreys prior

- Scalar parameter $\pi(\theta) \propto I(\theta)^{1/2}$. For transformed parameter $\gamma = g(\theta)$,

$$\pi(\gamma) \propto \frac{1}{\left|g'\left\{g^{-1}(\gamma)\right\}\right|} I\left\{g^{-1}(\gamma)\right\}^{1/2} = I(\gamma)^{1/2}$$

  Example 1: for $\text{binomial}(n, p)$. $I(p) = n/\{p(1-p)\}$

$$\pi(p) \propto \{p(1-p)\}^{-1/2} \sim \text{beta}(1/2, 1/2)$$

- Vector parameter $\pi(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2}$; $|\cdot|$ is determinant
  Example 1: for multinomial$(n; p_1, \cdots, p_k)$

$$\mathbf{I}_T(\mathbf{p}) = n\left\{\text{diag}\left(1/p_1, \cdots, 1/p_{k-1}\right) + 11^T/p_k\right\}$$

$$\pi(\mathbf{p}) \propto \frac{1}{\sqrt{p_1 \cdots p_k}} \sim \text{Dirichlet}(1/2, \cdots, 1/2)$$

For $N\left(\mu, \sigma^2\right)$ case with $\boldsymbol{\theta} = (\mu, \sigma), \mathbf{I}(\boldsymbol{\theta}) = \text{diag}\left(\sigma^{-2}, 2\sigma^{-2}\right) \Longrightarrow \pi(\mu, \sigma) \propto \sigma^{-2}$ which is different from σ^{-1} given earlier. Thus, Jeffreys modified his original proposal in the presence of location parameters say $\mu_1, \ldots, \mu_k$ to

$$\pi\left(\mu_1, \ldots, \mu_k, \boldsymbol{\theta}\right) \propto |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2}$$

where $\mu_i' s$ are held as fixed.

4.6 Normal data examples

4.6.1 One sample with unknown mean and variance

Data: $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right), \boldsymbol{\theta} = \left(\mu, \tau = 1/\sigma^2\right)$ unknown

Prior: $\pi(\mu \mid \tau) \sim N\left(\mu_0, \left(\tau n_0\right)^{-1}\right)$

$\pi(\tau) \sim \text{gamma}\left(\alpha_0, 1/\beta_0\right)$, mean $= \alpha_0/\beta_0$

a rough way to estimate the prior variance is treating $\tau = \alpha_0/\beta_0 \Longrightarrow$ prior variance of $\mu \approx \left(\beta_0/\alpha_0\right) n_0^{-1}$

$\mu \sim t$ with center $\mu_0, \text{scale}\, e^2 = \left(\beta_0/\alpha_0\right)/n_0$, and df $2\alpha_0$

then the likelihood is

$$\left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum\left(Y_i - \bar{Y}\right)^2 + n(\bar{Y} - \mu)^2\right\}\right]$$

then the posterior is

$$\pi(\mu, \tau \mid \mathbf{Y}) \propto \tau^{1/2} \exp\left\{-\frac{\tau n'}{2}\left(\mu - \mu'\right)^2\right\} \tau^{\alpha' - 1} e^{-\tau\beta'}$$

$$\mu' = \frac{n_0\mu_0 + n\bar{Y}}{n_0 + n}$$

$$n' = n_0 + n$$

$$\alpha' = n/2 + \alpha_0$$

$$\beta' = \beta_0 + \frac{\sum\left(Y_i - \bar{Y}\right)^2}{2} + \frac{1}{2}\left(\frac{n_0 n}{n_0 + n}\right)\left(\mu_0 - \bar{Y}\right)^2$$

$\Longrightarrow \mu, \tau \mid \mathbf{Y} \sim \text{N}\left(\mu', \left(\tau n'\right)^{-1}\right) \times \text{gamma}\left(\alpha', 1/\beta'\right)$ (conjugate prior)

$\mu \mid \mathbf{Y} \sim t$ with center μ', scale $e^2 = \left(\beta'/\alpha'\right)/n'$, and $\text{df} 2\alpha'$.

posterior predictive density

$$m\left(y_{n+1} \mid Y_1, \cdots, Y_n\right) = \int f\left(y_{n+1} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta} \mid Y_1, \cdots, Y_n\right) d\boldsymbol{\theta}$$

which is a t distribution with mean μ', scale $e^2 = \left(\beta'/\alpha'\right)\left(1 + 1/n'\right)$, and $df = 2\alpha'$.
Letting $\alpha_0 \to -1/2, \beta_0 \to 0$, and $n_0 \to 0$ s.t. $\left(n_0/\beta_0\right)^{1/2} \to 1$
$\Longrightarrow \pi(\mu, \tau) \to 1/\tau$, an improper prior

$\Longrightarrow \mu \mid \mathbf{Y} \sim t$ with mean $\bar{Y}$, scale $^2 = s_{n-1}^2/n$, and $\text{df} n - 1$

$\Longrightarrow$ Bayes estimator (with squared error loss) is $\bar{Y}$ credible interval $\equiv$ usual frequentist t interval

4.6.2 Two samples

$$X_1, \cdots, X_m \overset{\text{iid}}{\sim} N\left(\mu_1, \sigma^2\right), \quad Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} N\left(\mu_2, \sigma^2\right), \boldsymbol{\theta} = \left(\mu_1, \mu_2, \tau = 1/\sigma^2\right)^T$$

$\Delta = \mu_1 - \mu_2$ is of interest
prior

$$\begin{aligned}
\pi\left(\mu_1 \mid \tau\right) &\sim N\left\{\mu_{10}, \left(\tau m_0\right)^{-1}\right\} \\
\pi\left(\mu_2 \mid \tau\right) &\sim N\left\{\mu_{20}, \left(\tau n_0\right)^{-1}\right\} \\
\pi(\tau) &\sim \text{gamma}\left(\alpha_0, 1/\beta_0\right)
\end{aligned}$$

with similar derivations as in the last example, we can get the posterior

$$\theta \mid \mathbf{X}, \mathbf{Y} \sim \mathrm{N}\left\{\mu_1', \left(\tau m'\right)^{-1}\right\} \times \mathrm{N}\left\{\mu_2', \left(\tau n'\right)^{-1}\right\} \times \text{gamma}\left(\alpha', 1/\beta'\right)$$

$$\begin{aligned}
m' &= m_0 + m\mu_1' = \left(m_0\mu_{10} + m\bar{X}\right) / \left(m_0 + m\right) \\
n' &= n_0 + n\mu_2' = \left(n_0\mu_{20} + n\bar{Y}\right) / \left(n_0 + n\right) \\
\alpha' &= \alpha_0 + m/2 + n/2 \\
\beta' &= \beta_0 + \frac{\sum\left(X_i - \bar{X}\right)^2}{2} + \frac{1}{2}\left(\frac{m_0 m}{m_0 + m}\right)\left(\mu_{10} - \bar{X}\right)^2 + \frac{\sum\left(Y_i - \bar{Y}\right)^2}{2} + \frac{1}{2}\left(\frac{n_0 n}{n_0 + n}\right)\left(\mu_{20} - \bar{Y}\right)^2 \\
&\Longrightarrow \Delta, \tau \mid \mathbf{X}, \mathbf{Y} \sim \mathrm{N}\left\{\Delta' = \mu_1' - \mu_2', 1/\left(\tau m'\right) + 1/\left(\tau n'\right)\right\} \times \text{gamma}\left(\alpha', 1/\beta'\right) \\
&\Longrightarrow \Delta \mid \mathbf{X}, \mathbf{Y} \sim t \text{ with center } \Delta', \text{ scale }^2 = \left(\beta'/\alpha'\right)\left(m' + n'\right) / \left(m'n'\right), \text{ and df } 2\alpha'
\end{aligned}$$

Letting $\alpha_0 \to -1, \beta_0 \to 0, m_0 \to 0, n_0 \to 0$ such that $\left(m_0 n_0\right)^{1/2} / \beta_0 \to 1$
$\longrightarrow$ improper prior $\pi(\boldsymbol{\theta}) = 1/\tau$ (Jeffrey's prior)
$\Longrightarrow \Delta \mid \mathbf{X}, \mathbf{Y} \sim t$ with mean $\bar{X} - \bar{Y}$, scale $^2 = s_p^2(1/m + 1/n)$, and $df = m + n - 2$, where s_p^2 is the usual pooled estimated variance.

4.6.3 Normal linear model

$$\mathbf{Y}_{(n\times 1)} = \mathbf{X}_{(n\times p)}\boldsymbol{\Delta}_{(p)} + \mathbf{e}_{(n\times 1)}, \quad e_1, \cdots, e_n \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$$

Normal-gamma prior:

$$\boldsymbol{\Delta} \mid \tau = 1/\sigma^2 \sim N\left(\boldsymbol{\Delta}_0, \boldsymbol{\Sigma}_0^{-1}/\tau\right), \quad \tau \sim \text{gamma}\left(\alpha_0, 1/\beta_0\right)$$

Posterior is also Normal-gamma:

$$\boldsymbol{\Delta} \left| \tau, \mathbf{Y} \sim N\left(\boldsymbol{\Delta}', \left(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0\right)^{-1}/\tau\right), \quad \tau \right| \mathbf{Y} \sim \text{gamma}\left(\alpha', 1/\beta'\right)$$

where

$$\begin{aligned}
\Delta' &= \left(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0\right)^{-1}\left(\mathbf{X}^T\mathbf{Y} + \boldsymbol{\Sigma}_0\boldsymbol{\Delta}_0\right) \\
\alpha' &= \alpha_0 + n/2 \\
\beta' &= \beta_0 + \frac{1}{2}\left\{\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\Delta}'\right)^T\mathbf{Y} + \left(\boldsymbol{\Delta}_0 - \boldsymbol{\Delta}'\right)^T\boldsymbol{\Sigma}_0\boldsymbol{\Delta}_0\right\}
\end{aligned}$$

Δ' is a weighted average of the OLS estimator $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ and prior mean Δ_0
$\Delta \mid \mathbf{Y}$ is multivariate t with center Δ', scale matrix $\left(\beta'/\alpha'\right)\left(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0\right)^{-1}$, and df $2\alpha'$. For a subset of $\boldsymbol{\Delta}$, the marginal posterior is also a multivariate t.
let $\Delta_0 = \mathbf{0}, \boldsymbol{\Sigma}_0 = d\mathbf{I}_p$ with constant $d \Longrightarrow \Delta' = \left(\mathbf{X}^T\mathbf{X} + d\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{Y}$, the ridge regression estimator.
Letting $\boldsymbol{\Sigma}_0 \to \mathbf{0}, \alpha_0 \to -p/2$, and $\beta_0 \to 0$
$\Longrightarrow \pi(\Delta, \tau) = 1/\tau$
$\Longrightarrow \Delta' = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$, and $\boldsymbol{\Delta} \mid \mathbf{Y}$ is a p-dimensional t with df $n - p$ and scale matrix $s^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$, where $s^2 = (n-p)^{-1}\sum\left(Y_i - \mathbf{x}_i^T\Delta'\right)^2$

4.7 Hierarchical Bayes and empirical Bayes

Previously we discussed that case where prior $\pi(\boldsymbol{\theta})$ has been specified fully, say, $\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$ with hyperparameter $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ given. What if unsure about $\boldsymbol{\alpha}_0$? We turn to Hierarchical Bayes.
Hierarchical Bayes: Apart from $\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$, we also specify a hyperprior $h(\boldsymbol{\alpha}) = h(\boldsymbol{\alpha} \mid \gamma_0)$ with γ_0 given. The resulting prior

$$\begin{aligned}
\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{\gamma}_0) &= \int \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) h(\boldsymbol{\alpha} \mid \gamma_0)\, d\boldsymbol{\alpha} \\
m(\mathbf{y} \mid \boldsymbol{\alpha}) &= \int f(\mathbf{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) d\boldsymbol{\theta} \\
\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \mathbf{Y}) &= \frac{f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha})}{m(\mathbf{Y} \mid \boldsymbol{\alpha})} \\
\pi(\boldsymbol{\alpha} \mid \mathbf{Y}) &= \frac{m(\mathbf{Y} \mid \boldsymbol{\alpha}) h(\boldsymbol{\alpha})}{\iint f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) h(\boldsymbol{\alpha}) d\boldsymbol{\alpha} d\boldsymbol{\theta}} \\
\pi(\boldsymbol{\theta} \mid \mathbf{Y}) &= \frac{\int f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) h(\boldsymbol{\alpha}) d\boldsymbol{\alpha}}{\iint f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) h(\boldsymbol{\alpha}) d\boldsymbol{\alpha} d\boldsymbol{\theta}} \\
&= \int \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \mathbf{Y}) \pi(\boldsymbol{\alpha} \mid \mathbf{Y}) d\boldsymbol{\alpha}
\end{aligned}$$

Empirical Bayes: use MLE $\widehat{\boldsymbol{\alpha}}$ from marginal likelihood $m(\mathbf{Y} \mid \boldsymbol{\alpha})$ (other estimators may also be used).

$$\pi(\boldsymbol{\alpha} \mid \mathbf{Y}) \text{ highly peaked at } \widehat{\boldsymbol{\alpha}} \Longrightarrow \pi(\boldsymbol{\theta} \mid \mathbf{Y}) \approx \pi(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\alpha}}, \mathbf{Y}) \text{ (taking } \pi(\hat{\boldsymbol{\alpha}} \mid \mathbf{Y}) \approx 1)$$

Empirical Bayes posterior has a similar mean, but smaller variability than the full Bayes.
Example 1: one-way normal random effects model

$$Y_{ij}, i = 1, \cdots, k; j = 1, \cdots, n_i$$

Frequentist methods: fixed and random effects ANOVA.
Bayesian with fixed effects: earlier normal linear model. A Bayesian analogue of the random effects model:

$$\begin{aligned}
Y_{ij} \mid \theta_i, \sigma_e^2 &\sim N\left(\theta_i, \sigma_e^2\right) \text{ given } \theta_1, \cdots, \theta_k, \sigma_e^2,\ Y_{ij} \text{ 's are mutually independent} \\
\theta_1, \cdots, \theta_k \mid \boldsymbol{\alpha} = (\mu, \sigma_a) &\overset{\text{i.i.d}}{} N\left(\mu, \sigma_a^2\right) \\
\pi\left(\sigma_e^2\right) &\propto 1/\sigma_e^2 (\text{ Jeffrey's} \\
h(\mu) &\propto 1 \\
h(\sigma_a) &\propto 1
\end{aligned}$$

Of interest:

1. random effects population: μ and σ_a
2. individual θ_i

For the empirical Bayes approach, consider a simplified model with σ_e^2 known. Obtain $\widehat{\boldsymbol{\alpha}}$ from the marginal likelihood $m(\mathbf{y} \mid \boldsymbol{\alpha})$. The posterior is $\pi(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\alpha}}, \mathbf{Y})$.

4.7.1 James–Stein estimation

$Y_1, \cdots, Y_b$ are independent: $Y_i \sim N\left(\theta_i, \sigma_0^2\right)$ with σ_0^2 known.
While the parameters are unconnected in the model, the inferences regarding them are connected through the squared error loss $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{Y})\|^2$. Consider $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \mathbf{Y}$, $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = b\sigma_0^2$.
Stein (1955) proved the remarkable result that for $b \geq 3$, $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is inadmissible. James and Stein (1961) provided a dominating estimator,

$$\widehat{\boldsymbol{\theta}}_{\mathrm{JS}} = \left\{1 - \frac{(b-2)\sigma_0^2}{\sum_{i=1}^b Y_i^2}\right\} \mathbf{Y}$$

shrinking $\mathbf{Y}$ toward $\mathbf{0}$.
Expected squared error loss

$$\sum_{i=1}^b E\left(\widehat{\theta}_{\mathrm{JS},i} - \theta_i\right)^2 = b\sigma_0^2 - (b-2)^2 \sigma_0^4 E\left(\frac{1}{\sum_{i=1}^b Y_i^2}\right)$$

which is less than the risk for $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ provided $b \geq 3$
Empirical Bayes interpretation:

prior: $\boldsymbol{\theta} \sim \mathrm{MN}\left(\mathbf{0}, \sigma_a^2 \mathbf{I}_b\right)$

$\Longrightarrow$ posterior: $\boldsymbol{\theta} \mid \mathbf{Y} \sim$ normal with mean $\left(1 - \dfrac{\sigma_0^2}{\sigma_0^2 + \sigma_a^2}\right) \mathbf{Y}$ marginal $\mathbf{Y} \sim MN\left(\mathbf{0}, \left(\sigma_0^2 + \sigma_a^2\right) \mathbf{I}_b\right)$ σ_a^2 is unknown, but can be estimated from the fact

$$E\left\{\frac{(b-2)\sigma_0^2}{\sum_{i=1}^b Y_i^2}\right\} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_a^2}$$

Thus, $\hat{\boldsymbol{\theta}}_{\mathrm{JS}}$ may be viewed as an empirical Bayes estimator (Efron and Morris, 1972).
Alternative prior

$$\boldsymbol{\theta} \sim MN\left(\mu, \sigma_a^2 \mathbf{I}_b\right)$$

Bayes posterior mean $B\mu + (1-B)\mathbf{Y}, B = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_a^2}$

$$\text{Marginal } \mathbf{Y} \sim \sim \mathrm{N}\left(\mu, \left(\sigma_0^2 + \sigma_a^2\right) \mathbf{I}_b\right)$$

unbiased estimators $\widehat{\mu} = \bar{Y} \quad \widehat{B} = (b-3)\sigma_0^2 / \sum_{i=1}^b \left(Y_i - \bar{Y}\right)^2$

$$\text{James-Stein estimator } \widehat{B}\widehat{\mu} + (1 - \widehat{B})\mathbf{Y},$$

shrinking towards sample mean $\bar{Y}$ and having expected squared error loss less than $\mathbf{Y}$ if $b \geq 4$.

4.7.2 Meta-analysis applications of hierarchical and empirical Bayes

Meta analysis: analysis of a group of studies related to the same question of interest. Each study has a point effect estimate and its associated standard error.
Hierarchical Bayes: k studies. i th study - effect parameter θ_i and data Y_i. Ignoring first-stage nuisance parameters, 3 levels of the model are

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = f\left(y_1, \cdots, y_k \mid \theta_1, \cdots, \theta_k\right) = \prod_{i=1}^k f\left(y_i \mid \theta_i\right)$$

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \prod_{i=1}^k f\left(\theta_i \mid \boldsymbol{\alpha}\right)$$

$$h(\boldsymbol{\alpha})$$

The 2 nd and 3 rd stages $\Longrightarrow$ prior for $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}) = \int \prod_{i=1}^k f\left(\theta_i \mid \boldsymbol{\alpha}\right) h(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$$

Empirical Bayes:
marginal density of $\mathbf{Y} \Longrightarrow$ estimator $\widehat{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}$
$\Longrightarrow$ estimated prior $\pi(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\alpha}})$
$\Longrightarrow$ estimated posterior $\pi(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\alpha}}, \mathbf{Y})$
$\Longrightarrow$ empirical Bayes estimate close to posterior mean from the hierarchical Bayes, but the variance is smaller.
Example 1: Normal models with known variance
Data: $\left(Y_1, V_1\right), \cdots, \left(Y_k, V_k\right)$, where Y_i is approximately normal with variance estimated by V_i (but treated as known in analysis)
Frequentist fixed effect approach: all Y_i 's are approximately unbiased estimators of true effect $\mu \Longrightarrow$ weighted average $\sum_{i=1}^k V_i^{-1} Y_i / \sum_{i=1}^k V_i^{-1}$ is approximately optimal. Assumption is not quite realistic!
Frequentist random effect approach: $\theta_i \sim N\left(\mu, \sigma_a^2\right)$. The likelihood of $\mathbf{Y}$ or other methods can be used for the estimation of μ and σ_a.
Hierarchical Bayesian:

$$Y_i \mid \theta_i \sim N\left(\theta_i, V_i\right)$$
$$\theta_i \mid \mu, \sigma_a \sim N\left(\mu, \sigma_a^2\right)$$
$$h\left(\mu, \sigma_a\right) \propto 1 \text{ (noninformative prior)}$$

Empirical Bayes: use estimator of $\left(\mu, \sigma_a\right)$ (same as those from the frequentist random effect approach) without the hyperprior.
Full Bayes: with the hyperprior, MCMC is needed for the estimation.

4.8 Monte Carlo estimation of a posterior

(most materials from Carlin and Louis, 1996)
Main technical problem in Bayesian analysis: obtaining the posterior density and computing summary quantities such as posterior mean, standard deviation, and quantiles.
Which means we need to calculate integrals! As soon as we move away from conjugate priors and/or to hierarchical models.
One approach is to resort to asymptotic results. More focus is on Monte Carlo methods.

4.8.1 Noniterative Monte Carlo methods

Direct sampling:
Suppose that $\boldsymbol{\theta} \sim f(\boldsymbol{\theta})$ and $\gamma = E\{h(\boldsymbol{\theta})\} = \int h(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is of interest. Generate $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N \overset{\text{iid}}{\sim} f(\boldsymbol{\theta})$ and obtain

$$\widehat{\gamma} = N^{-1} \sum_{j=1}^{N} h\left(\boldsymbol{\theta}_j\right)$$

Indirect sampling:

1. Importance sampling
   Consider a posterior expectation

   $$E\{h(\boldsymbol{\theta}) \mid \mathbf{Y}\} = \frac{\int h(\boldsymbol{\theta}) f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

   Suppose that we can roughly approximate the normalized likelihood times prior, $cf(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$, by some density $g(\boldsymbol{\theta})$ from which we can easily sample. Define weight function $w(\boldsymbol{\theta}) = f(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})/g(\boldsymbol{\theta})$

   $$E\{h(\boldsymbol{\theta}) \mid \mathbf{Y}\} = \frac{\int h(\boldsymbol{\theta}) w(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int w(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}} \approx \frac{N^{-1} \sum_{j=1}^{N} h\left(\boldsymbol{\theta}_j\right) w\left(\boldsymbol{\theta}_j\right)}{N^{-1} \sum_{j=1}^{N} w\left(\boldsymbol{\theta}_j\right)}$$

   where $\boldsymbol{\theta}_j \overset{\text{iid}}{\sim} g(\boldsymbol{\theta})$, the *importance function*.
   The performance depends on how close $g(\boldsymbol{\theta})$ resembles $cf(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

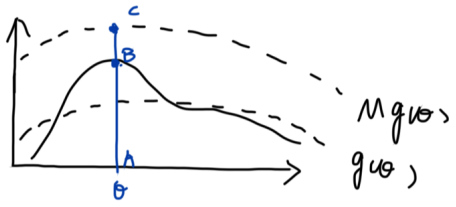
2. Rejection sampling
   Consider posterior sampling

   $$\pi(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{f(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

   Suppose $\exists M > 0$ and a smooth density $g(\boldsymbol{\theta})-$ envelope function - s.t. $f(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) < Mg(\boldsymbol{\theta})$ then

   (a) Generate $\boldsymbol{\theta}_j \sim g(\boldsymbol{\theta})$
   (b) Generate $U \sim \text{Uniform}(0, 1)$
   (c) If $MUg\left(\boldsymbol{\theta}_j\right) < f\left(\mathbf{Y} \mid \boldsymbol{\theta}_j\right) \pi\left(\boldsymbol{\theta}_j\right)$, accept $\boldsymbol{\theta}_j$; otherwise reject $\boldsymbol{\theta}_j$.
   (d) Return to step (i) and repeat until a desired size is obtained. The members of this sample is an iid sample from $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$

Figure 3: Graphical representation of the rejection sampling method



The third condition ensures that the density of the sampled θ is within $[A,\ B]$. Therefore, more likely to be from the posterior distribution

3. Weighted bootstrap
   Again consider the posterior density. Suppose that we have a sample $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N$ from approximating

density $g(\boldsymbol{\theta})$.
Define

$$w_i = \frac{f(\mathbf{Y} \mid \boldsymbol{\theta}_i)\,\pi(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}, \quad q_i = \frac{w_i}{\sum_{j=1}^{N} w_j}$$

Draw $\boldsymbol{\theta}^*$ from the discrete distribution over $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N\}$ which places mass q_i at $\boldsymbol{\theta}_i$.
In these indirect sampling methods, the prior $\pi(\boldsymbol{\theta})$, if proper, can play a role.
Example 1: Suppose $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ and $\pi(\theta) = \text{Cauchy}(\mu, \tau)$ with known σ^2, μ, and τ. Likelihood $f(\mathbf{Y} \mid \theta)$ is maximized at $\widehat{\theta} = \bar{Y} \Longrightarrow M = f(\mathbf{Y} \mid \widehat{\theta})$ in the rejection method, and $g(\theta) = \pi(\theta)$.
However, $\pi(\boldsymbol{\theta})$ is often very flat relative to $f(\mathbf{Y} \mid \theta) \Rightarrow$ quite inefficient.

4.9 MCMC methods

4.9.1 Substitution sampling

Consider the 3 -stage hierarchical model

$$\begin{gathered}\text{likelihood } p(Y \mid \theta) \\ \text{prior } p(\theta \mid \eta) \\ \text{hyperprior } p(\eta)\end{gathered}$$

Assuming the prior is conjugate with the likelihood $\Longrightarrow$ marginal distribution $p(Y \mid \eta) = \int p(Y \mid \theta) p(\theta \mid \eta) d\theta$ easily computed $\Longrightarrow$ closed-form posterior $p(\theta \mid Y, \eta) = \frac{p(Y|\theta)p(\theta|\eta)}{p(Y|\eta)}$
Also assuming the hyperprior is conjugate with the prior $\Longrightarrow$ closed-form $p(\eta \mid \theta)$ For inference, we seek the marginal posterior,

$$p(\theta \mid Y) = \int p(\theta \mid Y, \eta) p(\eta \mid Y) d\eta$$

and we also have

$$p(\eta \mid Y) = \int p(\eta \mid \theta) p(\theta \mid Y) d\theta$$

Switching to generic notation, we have a system of two linear integral equations

$$\begin{aligned} p(x) &= \int p(x \mid y) p(y) dy \\ p(y) &= \int p(y \mid x) p(x) dx \end{aligned}$$

where $p(x \mid y)$ and $p(y \mid x)$ are known, and we seek $p(x)$. This is a fixed point system:

$$\begin{aligned} p(x) &= \int p(x \mid y) \int p(y \mid x')\, p(x')\, dx' dy \\ &= \int h(x, x')\, p(x')\, dx' \end{aligned}$$

where $h(x, x') = \int p(x \mid y) p(y \mid x')\, dy$. *Sampling-based algorithm:*
Draw $X^{(0)} \sim p_0(x)$
Draw $Y^{(1)} \sim p\left(y \mid X^{(0)}\right) \sim p_1(y) = \int p(y \mid x) p_0(x) dx$
Draw $X^{(1)} \sim p\left(x \mid Y^{(1)}\right) \sim p_1(x) = \int h(x, x')\, p_0(x')\, dx'$
Repeat this process:

$$X^{(i)} \xrightarrow{d} X \sim p(x) \quad Y^{(i)} \xrightarrow{d} Y \sim p(y)$$

A variant: multiple $Y_1^{(i)}, \cdots, Y_m^{(i)} \overset{\text{iid}}{\sim} p\left(y \mid X^{(i-1)}\right)$ and single

$$X^{(i)} \sim \widehat{p}_i(x) = \frac{1}{m} \sum_{j=1}^{m} p\left(x \mid Y_j^{(i)}\right)$$

$\Longrightarrow$ automatically produce a smooth estimate of $p(x)$
Parallel sampling with multiple chains: Marginally independent replicates $\Longrightarrow$ presumably better $\widehat{p}_i(x)$. But wasteful.
Ergodic sampling with a single chain: Continue for an additional $m - 1$ iterations after convergence at iteration i.

$$\widehat{p}_i(x) = \frac{1}{m} \sum_{j=1}^{m} p\left(x \mid Y_j^{(i+j-1)}\right)$$

To reduce high autocorrelation, retain only every k th iteration:

$$\widehat{p}_i(x) = \frac{1}{m}\sum_{j=1}^{m} p\left(x \mid Y_j^{(i+(j-1)k)}\right)$$

Although X and Y can conceivably be vectors, sampling from complex multivariate distributions is difficult $\Longrightarrow$ Require a K-variate extension. Consider $K = 3$:

$$p(x) = \int p(x, z \mid y)p(y)dydz$$
$$p(y) = \int p(y, x \mid z)p(z)dxdz$$
$$p(z) = \int p(z, y \mid x)p(x)dxdy$$

Closed-form bivariate conditional distributions are unlikely available. With univariate distributions only:

$$p(x) = \int p(x \mid z, y)p(z \mid y)p(y)dydz$$
$$p(y) = \int p(y \mid x, z)p(x \mid z)p(z)dxdz$$
$$p(z) = \int p(z \mid y, x)p(y \mid x)p(x)dxdy$$

6 of them!
General K-dimensional problem $\Longrightarrow$ $K(K-1)$ distributions $\Longrightarrow$ impractical for large K.

4.9.2 Gibbs sampling

Gibbs sampler can be viewed as a special case of Metropolis-Hastings. And Metropolis-Hastings algorithm is a generalization of the Metropolis algorithm.
Under mild conditions, between full or complete conditional distributions and joint distribution

$$\{p_i\left(U_i \mid U_{j\neq i}\right), i = 1, \cdots, K\} \quad \Longleftrightarrow \quad p\left(U_1, \cdots, U_K\right)$$

Suppose these full conditional distributions are available for sampling. Given an arbitrary set of starting values $\left(U_1^{(0)}, \cdots, U_K^{(0)}\right)$, the algorithm proceeds as follows:

$$\begin{aligned}
&\text{Draw } U_1^{(1)} \sim p_1\left(U_1 \mid U_2^{(0)}, \cdots, U_K^{(0)}\right), \\
&\text{Draw } U_2^{(1)} \sim p_2\left(U_2 \mid U_1^{(1)}, U_3^{(0)}, \cdots, U_K^{(0)}\right), \\
&\text{Draw } U_K^{(1)} \sim p_K\left(U_K \mid U_1^{(1)}, \cdots, U_{K-1}^{(1)}\right),
\end{aligned}$$

completing one iteration of the scheme. After t such iterations, we obtain $\left(U_1^{(t)}, \cdots, U_K^{(t)}\right)$.

$$\left(U_1^{(t)}, \cdots, U_K^{(t)}\right) \xrightarrow{d} \left(U_1, \cdots, U_K\right) \sim p\left(U_1, \cdots, U_K\right)$$

Hierarchical models with conjugate priors and hyperpriors $\Longrightarrow$ closed form $\{p_i\left(U_i \mid U_{j\neq i}, \mathbf{Y}\right), i = 1, \cdots, K\}$ $\Longrightarrow$ joint posterior $p\left(U_1, \cdots, U_K \mid \mathbf{Y}\right)$ by Gibbs sampler.
Marginal posterior $p\left(U_i \mid \mathbf{Y}\right)$ can be obtained the same manner as before. For example, with parallel sampling,

$$\widehat{p}_t\left(U_i \mid \mathbf{Y}\right) = \frac{1}{m}\sum_{j=1}^{m} p\left(U_i \mid U_{1,j}^{(t)}, \cdots, U_{i-1,j}^{(t)}, U_{i+1,j}^{(t)}, \cdots, U_{K,j}^{(t)}, \mathbf{Y}\right)$$

which is less variable than a kernel-smoothed estimate.
What about non-conjugate priors? One may consider, for example, an indirect sampling method. But such a solution is not ideal.

4.9.3 Metropolis-Hastings algorithm

Target: joint posterior $p(u)$ for (possibly vector-valued) U
Candidate or proposal density: $q(v,u)$ such that $q(\cdot,u)$ is a $\mathrm{pdf}\,\forall u$, and $q(u,v)=q(v,u)\forall u,v$.
The Metropolis algorithm:

1. Draw $v\sim q\left(\cdot,U^{(t-1)}\right)$
2. Compute the density ratio $r=p(v)/p(u)$
3. If $r\geq 1$, set $U^{(t)}=v$

If $r<1$, set $U^{(t)}=\begin{cases} v \text{ with prob } r \\ u \text{ with prob } 1-r \end{cases}$
Mild conditions $\Longrightarrow U^{(t)}\xrightarrow{d}U\sim p(\cdot)$.
$p(\cdot)$ is needed only up to proportionality constant:

$$\pi(\boldsymbol{\theta}\mid\mathbf{Y})\propto f(\mathbf{Y}\mid\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

in Bayesian applications.
Why symmetric q ?
Consider finite state spaces, where the transition kernel can be represented by matrix $\mathbf{P}$; P_{ij} is the prob of moving from state i to j.
The Markov chain has equilibrium distribution

$$\mathbf{d}=(d_1,\cdots,d_k)^T$$

if and only if

$$\mathbf{d}^T\mathbf{P}=\mathbf{d}^T.$$

Symmetric $q\Longrightarrow$

$$\begin{aligned} d_iP_{ij}&=d_i\left[\min\left(1,\frac{d_j}{d_i}\right)Q_{ij}\right]=\min\left(d_i,d_j\right)Q_{ij} \\ &=\min\left(d_i,d_j\right)Q_{ji}=d_jP_{ji} \end{aligned}$$

$\Longrightarrow$ chain is reversible $\Longrightarrow$

$$\left(\mathbf{d}^TP\right)_j=\sum_{i=1}^k d_iP_{ij}=\sum_{i=1}^k d_jP_{ji}=d_j\sum_{i=1}^k P_{ji}=d_j.$$

For the continuous parameter settings, a convenient choice for q is a $N\left(\boldsymbol{\theta}^{(i-1)},\widetilde{\boldsymbol{\Sigma}}\right)$. In theory, any positive-definite $\widetilde{\boldsymbol{\Sigma}}$ suffices. Care is required in practice:

- too large $\Longrightarrow$ large jumps $\Longrightarrow$ many candidates far from posterior support and rejected $\Longrightarrow$ tendency to "get stuck"
- too small $\Longrightarrow$ "baby-stepping"

A simple but important generalization of the Metropolis algorithm was due to Hastings, by dropping the symmetry requirement of q and redefining

$$r=\frac{p(v)q(u,v)}{p(u)q(v,u)}$$

From Yale BIS 567
Metropolis Algorithm
Define $h(\theta)\equiv f(y\mid\theta)f(\theta)$ and $q\left(\boldsymbol{\theta}^*\mid\boldsymbol{\theta}^{(t-1)}\right)$ is a candidate or proposal density. It must satisfy

1. valid density function for every possible value of $\boldsymbol{\theta}^{(t-1)}$
2. be symmetric $q\left(\boldsymbol{\theta}^*\mid\boldsymbol{\theta}^{(t-1)}\right)=q\left(\boldsymbol{\theta}^{(t-1)}\mid\boldsymbol{\theta}^*\right)$. For example
   - $q\left(\boldsymbol{\theta}^*\mid\boldsymbol{\theta}^{(t-1)}\right)=$ Uniform $\left(\boldsymbol{\theta}^{(t-1)}-\delta,\boldsymbol{\theta}^{(t-1)}+\delta\right)$
   - $q\left(\boldsymbol{\theta}^*\mid\boldsymbol{\theta}^{(t-1)}\right)=\mathrm{Normal}\left(\boldsymbol{\theta}^{(t-1)},\delta^2\right)$

δ is chosen to make the algorithm run more efficiently. A poor choice of δ leads to high correlation in the Markov chain.

(a) δ too big: Low acceptance of a new θ and high correlation between samples

(b) δ too small: High acceptance of a new θ and high correlation between samples

Generally, acceptance around 20% to 50% is desired. Pilot runs are often required to properly tune the algorithm and select a reasonable δ.

The algorithm: for $t = 1, \ldots, T$:

1. draw θ^* from $q\left(. \mid \theta^{(t-1)}\right)$
2. compute $r = h\left(\boldsymbol{\theta}^*\right) / h\left(\boldsymbol{\theta}^{(t-1)}\right) = \exp\left[\log\left\{h\left(\boldsymbol{\theta}^*\right)\right\} - \log\left\{h\left(\boldsymbol{\theta}^{(t-1)}\right)\right\}\right]$
3. if $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$; if $r < 1$, set

$$\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } r \\ \boldsymbol{\theta}^{(t-1)} & \text{with probability } 1 - r \end{cases}$$

Under mild conditions (similar to those for the Gibbs sampler), a draw $\boldsymbol{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $f(\boldsymbol{\theta} \mid \boldsymbol{Y} = \boldsymbol{y})$.
Intuition:
Suppose we have a working set of posterior samples $\left\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(s)}\right\}$. We want to add a new sample to the set, θ^*, which is nearby $\theta^{(t)}$. If $f\left(\boldsymbol{\theta}^* \mid \boldsymbol{Y} = \boldsymbol{y}\right) > f\left(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{Y} = \boldsymbol{y}\right)$, then we want to add θ^* to the set. If $f\left(\boldsymbol{\theta}^* \mid \boldsymbol{Y} = \boldsymbol{y}\right) < f\left(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{Y} = \boldsymbol{y}\right)$, then we don't necessarily want to add it to the set. So we accept it with probability r

$$r = \frac{f\left(\boldsymbol{\theta}^* \mid \boldsymbol{Y} = \boldsymbol{y}\right)}{f\left(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{Y} = \boldsymbol{y}\right)} = \frac{f\left(y \mid \theta^*\right) f\left(\theta^*\right)}{f(y)} \frac{f(y)}{f\left(y \mid \theta^{(s)}\right) f\left(\theta^{(s)}\right)} = \frac{f\left(y \mid \theta^*\right) f\left(\theta^*\right)}{f\left(y \mid \theta^{(s)}\right) f\left(\theta^{(s)}\right)}$$

4.9.4 Hybrid forms

Several MCMC algorithms may be combined in a single problem, to take advantage of the strengths of each. Markov kernels $P_1, \cdots, P_m$ all have stationary distribution p; they may correspond to, say, one Gibbs sampler and $m - 1$ Metropolis algorithms.
Mixer: At each iteration, kernel P_i is chosen with prob α_i, where $\sum_i \alpha_i = 1$.
Cycle: Each kernel P_i is used in a prespecified order.
Metropolis within Gibbs: In the case that, say, all full conditional distributions are available in closed form except for one, a Metropolis subalgorithm may be embedded in the Gibbs sampler.

5 Large sample theory

6 M-Estimation (Estimating Equations)

6.1 Introduction

$$\sum_{i=1}^{n} \boldsymbol{\Psi}\left(Y_i; \boldsymbol{\theta}\right) = \mathbf{0}$$

$Y_1, \cdots, Y_n$: independent but not necessarily identically distributed
$\boldsymbol{\theta}$: $b \times 1$ parameter
$\boldsymbol{\Psi}$: $b \times 1$ function

Example 1. the mean estimator $\widehat{\theta} = n^{-1} \sum_{i=1}^{n} Y_i$ is from

$$\sum_{i=1}^{n} \left(Y_i - \theta\right) = 0$$

the deviance from the mean $\widehat{\theta}_1 = n^{-1}\sum_{i=1}^n \left|Y_i - \bar{Y}\right|$ is from

$$\sum_{i=1}^n \begin{pmatrix} |Y_i - \theta_2| - \theta_1 \\ Y_i - \theta_2 \end{pmatrix} = \mathbf{0}$$

6.2 Basic approach

Suppose $Y_1, \cdots, Y_n \overset{\text{iid}}{\sim} F$ and the true parameter θ_0 is defined as a solution to

$$E_F \mathbf{\Psi}(Y_1, \boldsymbol{\theta}) = \int \mathbf{\Psi}(y, \boldsymbol{\theta}) dF(y) = \mathbf{0}$$

if $\mathbf{\Psi}$ is suitably smooth then by Taylor expansion of

$$\mathbf{G}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \mathbf{\Psi}(Y_i, \boldsymbol{\theta})$$

we have

$$0 = \mathbf{G}_n(\widehat{\boldsymbol{\theta}}) = \mathbf{G}_n(\boldsymbol{\theta}_0) + \mathbf{G}_n'(\boldsymbol{\theta}_0)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + o_p\left(n^{-1/2}\right)$$
$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \left\{-\mathbf{G}_n'(\boldsymbol{\theta}_0)\right\}^{-1}\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) + o_p(1)$$

by WLLN

$$-\mathbf{G}_n'(\boldsymbol{\theta}_0) = -n^{-1}\sum_{i=1}^n \mathbf{\Psi}'(Y_i, \boldsymbol{\theta}_0) \xrightarrow{p} E_F\left\{-\mathbf{\Psi}'(Y_1, \boldsymbol{\theta}_0)\right\} \equiv \mathbf{A}(\boldsymbol{\theta}_0)$$

by CLT

$$\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{B}(\boldsymbol{\theta}_0)) \text{ where } \mathbf{B}(\boldsymbol{\theta}_0) = E_F\left\{\mathbf{\Psi}(Y_1, \boldsymbol{\theta}_0)^{\otimes 2}\right\}$$

then by Slutsky's theorem

$$\widehat{\boldsymbol{\theta}} \sim AN(\boldsymbol{\theta}_0, \mathbf{V}(\boldsymbol{\theta}_0)/n) \text{where} \mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\left\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\right\}^T$$

More generally, the estimating equation does not have to be exactly 0, for example

$$\sum_{i=1}^n \mathbf{\Psi}(Y_i, \boldsymbol{\theta}) = o_p\left(n^{1/2}\right) \Rightarrow \frac{1}{n}\sum_{i=1}^n \mathbf{\Psi}(Y_i, \boldsymbol{\theta}) = o_p\left(n^{-1/2}\right)$$

The asymptotic distribution of $\hat{\theta}$ remains the same.

6.2.1 Estimation for A, B, and V

Empirical estimator of $\mathbf{A}(\boldsymbol{\theta}_0)$:

$$\mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) = -n^{-1}\sum_{i=1}^n \mathbf{\Psi}'\left(Y_i, \widehat{\boldsymbol{\theta}}\right)$$

Empirical estimator of $\mathbf{B}(\boldsymbol{\theta}_0)$:

$$\mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) = n^{-1}\sum_{i=1}^n \mathbf{\Psi}\left(Y_i, \widehat{\boldsymbol{\theta}}\right)^{\otimes 2}$$

then

$$\mathbf{V}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) = \mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})^{-1}\mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})\left\{\mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})^{-1}\right\}^T$$

A **special** case is the MLE, where

$$\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{B}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0) \quad \mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0)^{-1}$$

Example 2. Sample mean and variance

$$\begin{aligned}
\boldsymbol{\Psi}\left(Y_i, \boldsymbol{\theta}\right) &= \begin{pmatrix} Y_i - \theta_1 \\ \left(Y_i - \theta_1\right)^2 - \theta_2 \end{pmatrix} \\
\mathbf{A}\left(\boldsymbol{\theta}_0\right) &= E\left\{-\boldsymbol{\Psi}'\left(Y_1, \boldsymbol{\theta}_0\right)\right\} = E\begin{pmatrix} 1 & 0 \\ 2\left(Y_1 - \theta_{10}\right) & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\
\mathbf{B}\left(\boldsymbol{\theta}_0\right) &= E\left\{\boldsymbol{\Psi}\left(Y_1, \boldsymbol{\theta}_0\right)^{\otimes 2}\right\} = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix} \\
\mathbf{V}\left(\boldsymbol{\theta}_0\right) &= \mathbf{B}\left(\boldsymbol{\theta}_0\right) \\
\mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) &= n^{-1}\sum_{i=1}^{n}\begin{pmatrix} \left(Y_i - \bar{Y}\right)^2 & \left(Y_i - \bar{Y}\right)^3 \\ \left(Y_i - \bar{Y}\right)^3 & \left\{\left(Y_i - \bar{Y}\right)^2 - s_n^2\right\}^2 \end{pmatrix}
\end{aligned}$$

the estimated $\widehat{\boldsymbol{\theta}} = \left(\bar{Y}, s_n^2\right)^T$ is the MLE for the normal density $f(y;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\theta_2}}\exp\left\{-\frac{(y-\theta_1)^2}{2\theta_2}\right\}$

Example 3. Ratio estimator
For i.i.d $\left(Y_1, X_1\right), \cdots, \left(Y_n, X_n\right)$ and $\mu_X \neq 0$

$$\begin{aligned}
\Psi\left(Y_i, X_i, \theta\right) &= Y_i - \theta X_i \\
A\left(\theta_0\right) &= E\left(X_1\right) = \mu_X \\
B\left(\theta_0\right) &= E\left\{\left(Y_1 - \theta_0 X_1\right)^2\right\} \\
V\left(\theta_0\right) &= E\left\{\left(Y_1 - \theta_0 X_1\right)^2\right\}/\mu_X^2 \\
A_n(\mathbf{Y}, \widehat{\theta}) &= \bar{X} \\
B_n(\mathbf{Y}, \widehat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{\bar{Y}}{\bar{X}}X_i\right)^2 \\
V_n(\mathbf{Y}, \widehat{\theta}) &= \frac{1}{\bar{X}^2}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{\bar{Y}}{\bar{X}}X_i\right)^2
\end{aligned}$$

The Ψ functions are not unique, it can even be of different dimensions

$$\begin{aligned}
\boldsymbol{\Psi}\left(Y_i, X_i, \boldsymbol{\theta}\right) &= \begin{pmatrix} Y_i - \theta_1 \\ X_i - \theta_2 \\ \theta_1 - \theta_3\theta_2 \end{pmatrix} \\
\mathbf{A}\left(\boldsymbol{\theta}_0\right) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & \theta_{30} & \theta_{20} \end{pmatrix} \\
\mathbf{B}\left(\boldsymbol{\theta}_0\right) &= \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} & 0 \\ \sigma_{YX} & \sigma_X^2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{singular} \\
\mathbf{V}\left(\boldsymbol{\theta}_0\right) &= \mathbf{A}\left(\boldsymbol{\theta}_0\right)^{-1}\mathbf{B}\left(\boldsymbol{\theta}_0\right)\left\{\mathbf{A}\left(\boldsymbol{\theta}_0\right)^{-1}\right\}^T \\
v_{33} &= \frac{1}{\theta_{20}^2}\left(\sigma_Y^2 - 2\theta_{30}\sigma_{YX} + \theta_{30}^2\sigma_X^2\right)
\end{aligned}$$

6.3 Delta method via M-estimation