1 Foundation

1.1 Matrix Derivatives

- Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ and f a real-valued (multivariate) function of $\mathbf{x}$. The derivative of $f(\mathbf{x})$, if exists, is given by

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)^T$$

- The second-order partial derivative of f at $\mathbf{x}$ is given by

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T}_{n\times n} = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial^2 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial^2 x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial^2 x_n} \end{pmatrix}$$

- Let $\mathbf{f}$ be a $p \times 1$ vector of functions of $\mathbf{x}$ as follows,

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_p(\mathbf{x}) \end{pmatrix}$$

  Then the derivative of $\mathbf{f}$ at $\mathbf{x}$, if exists, is given by

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}_{n\times p} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1)} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_p(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_p(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \frac{\partial f_2(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_p(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^T} = \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right)^T$$

- Let $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} = \mathbf{x}^T\mathbf{a}$, then

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$$

- Let $g(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$, then

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A}\mathbf{x} + \left(\mathbf{x}^T\mathbf{A}\right)^T = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}$$

- Let $\mathbf{f}(\mathbf{x}) = \mathbf{A}^T\mathbf{x}$, then

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A}$$

- Let $\mathbf{g}(\mathbf{x}) = \mathbf{x}^T\mathbf{B}$, then

$$\frac{\partial \mathrm{g}(\mathrm{x})}{\partial \mathbf{x}} = \mathbf{B}$$

- Let $\mathbf{A}(\mathbf{x}) = (A_{ij}(\mathbf{x}))_{n\times n}$ which is invertible, then

$$\frac{\partial \mathbf{A}^{-1}}{\partial x_i} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x_i}\mathbf{A}^{-1}$$
$$\frac{\partial |\mathbf{A}|}{\partial x_i} = \operatorname{tr}\left(|\mathbf{A}|\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x_i}\right) = |\mathbf{A}|\operatorname{tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x_i}\right)$$

1.2 Multivariate normal distribution

$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T \sim MVN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a positive definite $\boldsymbol{\Sigma}$ if any of the following definitions hold

- the density function of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}$$

- the moment generating function of $\mathbf{Y}$ is

$$m_{\mathbf{Y}}(\mathbf{t}) \equiv E\left(e^{\mathbf{t}^T\mathbf{Y}}\right) = \exp\left\{\boldsymbol{\mu}^T\mathbf{t} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right\}$$

- Y has the same distribution as

$$\mathbf{Q}^T\mathbf{z} + \boldsymbol{\mu},$$

  where $\mathbf{z} = (z_1, z_2, \ldots, z_n)^T$, z_i 's are iid $N(0,1)$, $\mathbf{Q}^T\mathbf{Q} = \boldsymbol{\Sigma}$.

Property 1. $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\operatorname{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}$

Property 2. If z_i 's are iid standard normal $N(0,1)$, then

$$\mathbf{z} = (z_1, z_2, \ldots, z_n) \sim \mathrm{MVN}(\mathbf{0}_n, \mathbf{I}_{n\times n})$$

Property 3. If $\mathbf{Y} \sim MVN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}_{p\times n}$ is not random, then

$$\mathbf{AY} \sim MVN_p\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right)$$

Property 4. If $\mathbf{Y} \sim MVN_n\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_{n\times n}\right)$ and $\mathbf{A}_{n\times n}$ is a constant orthogonal matrix, then

$$\mathbf{AY} \sim MVN_n\left(\mathbf{A}\boldsymbol{\mu}, \sigma^2\mathbf{I}_{n\times n}\right)$$

1.3 Other relevant distributions

Chi-square Distribution with p degrees of freedom, $\chi^2(p)$: suppose $Z_1, Z_2, \ldots, Z_p$ are iid $N(0,1)$, then

$$\sum_{i=1}^{p} Z_i^2 \sim \chi^2(p)$$

t**-Distribution** with p degrees of freedom, $t(p)$: suppose $Z \sim N(0,1)$ and $V \sim \chi^2(p)$ and Z and V are independent, then

$$\frac{Z}{\sqrt{V/p}} \sim t(p)$$

F**-Distribution** with p and q degrees of freedom, $F(p,q)$: suppose $Z \sim \chi^2(p)$ and $V \sim \chi^2(q)$ and Z and V are independent, then

$$\frac{Z/p}{V/q} \sim F(p,q).$$

1.4 Exponential family

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.
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
Some examples

- $Y\sim N\left(\mu,\sigma^{2}\right)$

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

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

  with another formulation

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

  then

$$\begin{aligned}\eta&=\left[\begin{array}{c}\mu/\sigma^{2}\\-1/2\sigma^{2}\end{array}\right]\\T(x)&=\left[\begin{array}{c}x\\x^{2}\end{array}\right]\\A(\eta)&=\frac{\mu^{2}}{2\sigma^{2}}+\log\sigma=-\frac{\eta_{1}^{2}}{4\eta_{2}}-\frac{1}{2}\log\left(-2\eta_{2}\right)\\h(x)&=\frac{1}{\sqrt{2\pi}}\end{aligned}$$

- Bernoulli distribution

$$f(y;\mu)=\mu^{y}(1-\mu)^{1-y}=\exp\left\{y\log\frac{\mu}{1-\mu}+\log(1-\mu)\right\}$$

  where $\mu=\Pr(Y=1)=E(Y)$
  then

$$\begin{aligned}\theta&=\log\frac{\mu}{1-\mu},\phi=1\\a(\phi)&=1\\b(\theta)&=-\log(1-\mu)=\log\left(1+e^{\theta}\right)\\c(y,\phi)&=c(y)=0\end{aligned}$$

  with another formulation

$$\begin{aligned}p(x\mid\pi)&=\pi^{x}(1-\pi)^{1-x}\\&=\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x+\log(1-\pi)\right\}\end{aligned}$$

then

$$
\begin{aligned}
\eta &= \frac{\pi}{1-\pi} \\
T(x) &= x \\
A(\eta) &= -\log(1-\pi) = \log\left(1+e^{\eta}\right) \\
h(x) &= 1
\end{aligned}
$$

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

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

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

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

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

Example: Poisson distribution

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp[y\log\mu - \mu - \log y!]$$

then

$$
\begin{aligned}
\theta &= \log\mu \\
b(\theta) &= \mu = \exp(\theta) \\
a(\phi) &= \phi = 1
\end{aligned}
$$

It follows that

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

1.5 Likelihood theory

Likelihood definition: random variable $Y\tilde{f(Y;\theta)}$, assume n iid observations $y = (y_1, y_2, y_3, \ldots, y_n)$. Then the likelihood function $L(y;\theta) = \Pi f(y;\theta)$, the log-likelihood $logL(y;\theta) = \sum f(y;\theta)$. The MLE estimator $\hat{\theta} = argmaxlogL(y;\theta)$

Fisher's score function:

$$u(\theta) = \frac{\partial logL(y;\theta)}{\partial \theta}$$

Then the MLE estimator $\hat{\theta}$ satisfies $u(\hat{\theta} = 0$.

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

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

Asymptotic theory: the MLE estimator $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, I^{-1}(\theta)$$

1.6 Likelihood for observed data

Suppose we observe a random sample of n independent observations, $O_1 = \left(y_1, \mathbf{x}_1^T\right), \ldots, O_n = \left(y_n, \mathbf{x}_n^T\right)$. Our goal is to estimate the parameters $\boldsymbol{\theta}$ involved in the conditional distribution of Y given $\mathbf{x}$, which is typically the case in regression analysis. Suppose y_i has density $f\left(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right)$. The joint density of $\mathbf{y}$ is

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i\left(y_i; \boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right).$$

The likelihood (function) of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} f\left(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right).$$

The log-likelihood of $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \log(L(\boldsymbol{\theta}; \mathbf{y})) = \sum_{i=1}^{n} \log f\left(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right).$$

A maximum likelihood estimator (MLE) is a maximizer of the likelihood $L(\boldsymbol{\theta}; \mathbf{y})$, denoted by $\widehat{\boldsymbol{\theta}}$,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} L(\boldsymbol{\theta}; \mathbf{y}) = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \ell(\boldsymbol{\theta}; \mathbf{y}),$$

where Θ is the parameter space. $\widehat{\boldsymbol{\theta}}$ is usually obtained by solving the likelihood (score) equations.
The **score function** is defined as

$$\mathbf{U}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\theta}).$$

The **likelihood (score) equations** is defined as

$$\mathbf{U}(\boldsymbol{\theta}) = 0.$$

Then $\widehat{\boldsymbol{\theta}}$ are zeros (solutions) of the score equations.
The **(Fisher) information** matrix is defined as

$$\begin{aligned}
\mathbf{I}_n(\boldsymbol{\theta}) &= -E \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\sum_{i=1}^{n} E \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = n \mathbf{I}(\boldsymbol{\theta}) \\
&= E\left(\mathbf{U}(\boldsymbol{\theta}) \mathbf{U}^T(\boldsymbol{\theta})\right) = \sum_{i=1}^{n} E\left(\mathbf{U}_i(\boldsymbol{\theta}) \mathbf{U}_i^T(\boldsymbol{\theta})\right),
\end{aligned}$$

$\mathbf{I}(\boldsymbol{\theta})$ is the information matrix for a *single* observation.
The **observed information** matrix is defined as

$$\begin{aligned}
\mathbf{i}_n(\boldsymbol{\theta}) &= -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\sum_{i=1}^{n} \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \\
&= \mathbf{U}(\boldsymbol{\theta}) \mathbf{U}^T(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\theta}) \mathbf{U}_i^T(\boldsymbol{\theta}).
\end{aligned}$$

Property 1. Let $\boldsymbol{\theta}_0$ denote the true value. We have

$$E\left(\mathbf{U}\left(\boldsymbol{\theta}_0\right)\right) = \mathbf{0}$$

Property 2. If $O_i = \left(y_i, \mathbf{x}_i^T\right)$ are iid, then

$$\frac{1}{n} \mathbf{i}_n(\boldsymbol{\theta}) \rightarrow_p \mathbf{I}(\boldsymbol{\theta})$$

Property 3. Under some regularity conditions (e.g., the model is correctly specified), the MLE $\widehat{\boldsymbol{\theta}}$ has properties:

$$\text{Consistency: } \widehat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_0$$

$$\text{Asymptotic Normality: } \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \rightarrow_d \operatorname{MVN}\left(0, \mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)\right)$$