Computational Statistics

Guo-Liang TIAN

Department of Statistics and Data Science
Southern University of Science and Technology
Shenzhen, Guangdong Province, P.R. China

Science Press, Beijing

1 September 2021

To Yanli, Margaret and Adam

Contents

Chapter 1

Generation of Random Variables

1• Why is this textbook important to you?

1.1• As a computational toolbox in the frequentist statistics

— In the frequentist statistics, one of the main tasks is to find *maximum likelihood estimates* (MLEs) $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)^\top$ of the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the parameter space, see Chapter 2.

— Next, it is also important to calculate the *standard deviation* (Std) of $\hat{\vartheta}$ or *confidence interval* (CI) $[\hat{\vartheta}_{\mathrm{L}}, \hat{\vartheta}_{\mathrm{U}}]$ of $\vartheta \hat{=} h(\boldsymbol{\theta})$, where $h(\cdot)$ is an arbitrary function of $\boldsymbol{\theta}$, see Chapters 1 and 5.

1.2• As a computational toolbox in the Bayesian statistics

— In the Bayesian statistics, one often needs to compute posterior moments such as $E(\vartheta|Y_{\mathrm{obs}})$ and $\mathrm{Var}(\vartheta|Y_{\mathrm{obs}})$, where ϑ can be viewed as a *random variable* (r.v.) and Y_{obs} denotes the observed data, see Chapter 3.

— More importantly, we would like to generate samples from posterior distributions, see Chapters 1 and 4.

1.3• Benefiting your whole academic career

— This textbook can help you when you write academic papers, research reports, grant proposals, statistical books, thesis and so on.

— This textbook can also serve your other courses including assignments and projects.

2• Why do we need chapter 1?

- We are faced with a dual world.

2.1• The real world: From practice to theory

— Suppose that we have observed $x_1, \ldots, x_n$, which can be viewed as realizations of a set of *independent and identically distributed* (i.i.d.) r.v.'s $X_1, \ldots, X_n$.

— If we could accept the null hypothesis H_0: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ or $f(x; \boldsymbol{\theta})$, where both the population density $f(\cdot)$ and the parameter vector $\boldsymbol{\theta}$ are unknown, then we can do many jobs.

— For example, we can estimate the $f(\cdot)$, $\boldsymbol{\theta}$ and other unknown quantities.

2.2• The statistical world: From theory to practice

— Given a theoretical density function $f(\cdot)$ or a *cumulative distribution function* (cdf) $F(\cdot)$, we want to generate a random sample $x_1, \ldots, x_n$ from $f(\cdot)$ or $F(\cdot)$, which is the topic of Chapter 1.

— Once $x_1, \ldots, x_n$ are produced, we can do many jobs.

— For example, we can use the sample average or the sample variance to estimate the population mean or the population variance, and so on.

3• Aims of chapter 1

- In Chapter 1, we will introduce some basic Monte Carlo simulation techniques for generating random samples from univariate and multivariate distributions with known parameters.
- These techniques also play a critical role in Monte Carlo integration.
- We assume that random numbers or r.v.'s uniformly distributed in the unit interval $(0, 1)$ can be satisfactorily produced on the computer.
- We focus on methods for fast generating non-uniform r.v.'s.

1.1 The Inversion Method

1.1.1 Generating samples from continuous distributions

4• Formulation of the inversion method

4.1• A basic result

— Let X be a r.v. with cdf $F(\cdot)$. Note that $F(\cdot)$ is non-decreasing but possibly discontinuous, the inverse function $F^{-1}(\cdot)$ can be defined by

$$F^{-1}(u) = \inf\{x\colon\ F(x) \geqslant u\}, \quad u \in (0,1).$$

Comment 1: What is the difference between infimum (supremum) and minimum (maximum)? ‖

— We have $F(X) \sim U(0,1)$, denoted by $F(X) \overset{\mathrm{d}}{=} U$, where $U \sim U(0,1)$.

Proof: Note that the cdf of $U \sim U(0,1)$ is given by

$$\begin{aligned}\Pr(U \leqslant u) &= \begin{cases} 0, & \text{if } u \leqslant 0, \\ u, & \text{if } 0 < u < 1, \\ 1, & \text{if } u \geqslant 1 \end{cases} \\ &= 0 \cdot I(u \leqslant 0) + u \cdot I(0 < u < 1) + 1 \cdot I(u \geqslant 1),\end{aligned}$$

where $I(\cdot)$ is the indicator function, then the cdf of $Y \mathrel{\hat{=}} F(X)$ is

$$\Pr(Y \leqslant y) = \Pr\{F(X) \leqslant y\} = \Pr\{X \leqslant F^{-1}(y)\} = F(F^{-1}(y)) = y,$$

implying that $Y = F(X) \sim U(0,1)$. □

Comment 2: (i) The operator '$\overset{\mathrm{d}}{=}$' will be introduced in §1.5.1; (ii) The symbol '$\mathrel{\hat{=}}$' means 'defined as'; (iii) $F(X) \overset{\mathrm{d}}{=} U \sim U(0,1)$ is equivalent to $X \overset{\mathrm{d}}{=} F^{-1}(U) \sim F(x)$. ‖

— Hence, to generate one sample, say x, from the r.v. $X \sim F(\cdot)$, we first draw u from $U \sim U(0,1)$, then compute $F^{-1}(u)$ and set it equal to x.

— Figure 1.1 illustrates the inversion method, which is also called the *inverse transformation* method.

— We summarize the algorithm as follows.

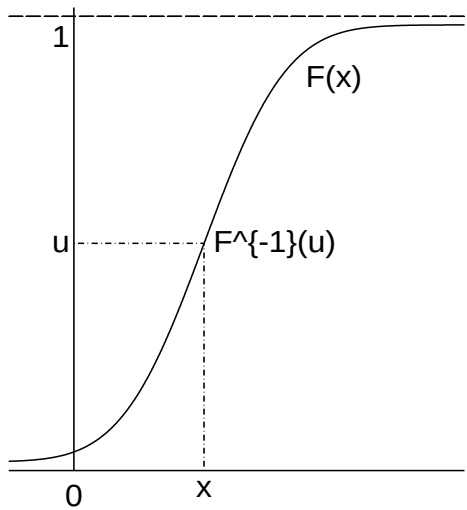

Figure 1.1 Illustration of the inversion method.

4.2• The inversion method

— Step 1: Draw $U = u$ from $U(0, 1)$;

— Step 2: Return $x = F^{-1}(u)$.

4.3• Several examples

Example 1.1 (Exponential distribution). Let X follow the exponential distribution with *probability distribution function* (pdf)

$$f(x) = \beta \, \mathrm{e}^{-\beta x}, \quad x \geqslant 0,$$

where $\beta\,(> 0)$ is the rate parameter, see Appendix A.2.3. Use the inversion method to generate a sample from the X.

Solution: The cdf of X is $F(x) = \int_0^x f(t)\,\mathrm{d}t = 1 - \mathrm{e}^{-\beta x}$ for $x \geqslant 0$. Let $F(x) = u$, then

$$x = F^{-1}(u) = -\frac{\log(1-u)}{\beta}, \quad 0 \leqslant u < 1.$$

Thus, the inversion method for generating x from $X \sim$ Exponential(β) is as follows:

Step 1: Draw $U = u$ from $U[0, 1)$;

Step 2: Return $x = -\log(1 - u)/\beta$.

R code:

```
================================================================
> beta <- 3
> u <- runif(1); x <- -log(1-u)/beta
> x
[1] 0.441
> u <- runif(10); x <- -log(1-u)/beta
> x
 [1] 0.075 0.658 0.429 0.010 0.301
 [5] 0.250 0.299 0.095 0.974 0.202
****************************************************************
```

Equivalent R code:

```
================================================================
> beta <- 3
> x <- rexp(10, beta)
> x
 [1] 0.077 0.055 0.628 0.037 0.622
 [6] 0.226 0.475 0.209 0.113 0.125
> x <- rexp(10, beta)
> x
 [1] 0.160 0.079 0.167 0.002 0.167
 [6] 0.265 0.294 0.449 0.118 0.292
****************************************************************
```

Comment 3: The support of $X \sim f(x)$ is $\mathcal{S}_X = \{x : f(x) > 0\} = [0, \infty)$. ‖

Example 1.2 (Standard Laplace distribution). Let X follow the standard Laplace distribution (or the double exponential distribution) with pdf

$$f(x) = 0.5\, \mathrm{e}^{-|x|}, \quad -\infty < x < \infty,$$

see Appendix A.2.2. Use the inversion method to generate a sample from the distribution.

Solution: The cdf of X is

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \begin{cases} 0.5\,\mathrm{e}^{x}, & \text{if } x < 0, \\ 1 - 0.5\,\mathrm{e}^{-x}, & \text{if } x \geqslant 0. \end{cases}$$

Let $F(x) = u$, then

$$x = F^{-1}(u) = \begin{cases} \log(2u), & \text{if } 0 < u < 0.5, \\ -\log\{2(1-u)\}, & \text{if } 0.5 \leqslant u < 1. \end{cases}$$

Thus, the inversion method for generating $X \sim \text{Laplace}(0, 1)$ is as follows:

Step 1: Draw $U = u$ from $U(0, 1)$;

Step 2: Return $x = \log(2u)$ if $0 < u < 0.5$ and $x = -\log\{2(1-u)\}$ otherwise.

R code: An R code for simulating N i.i.d. samples of X is given in **23.1•** in Appendix A.2.2. ‖

Example 1.3 (Weibull distribution). Use the inversion method to generate a sample from the Weibull distribution with pdf

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{k}\right\},$$

where $x > 0$, $\lambda > 0$ and $k > 0$.

Solution: The cdf of X is

$$F(x) = \int_{0}^{x} \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{t}{\lambda}\right)^{k}\right\} \mathrm{d}t = 1 - \exp\left\{-\left(\frac{x}{\lambda}\right)^{k}\right\}.$$

Let $u = F(x)$, we have $x = F^{-1}(u) = \lambda\{-\log(1-u)\}^{1/k}$. The algorithm is as follows:

Step 1: Draw $U = u$ from $U(0, 1)$;

Step 2: Return $x = \lambda\{-\log(1-u)\}^{1/k}$. ‖

Example 1.4 (Cauchy distribution). Use the inversion method to generate a sample from the Cauchy distribution with pdf

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

where $-\infty < x < +\infty$, $-\infty < \mu < +\infty$ and $\sigma > 0$.

Solution: The cdf of X is

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right) + 0.5.$$

Let $u = F(x)$, then $x = \mu + \sigma\tan(\pi u - 0.5\pi)$. The algorithm is as follows:

Step 1: Generate $U = u$ from $U(0,1)$;

Step 2: Return $x = \mu + \sigma\tan(\pi u - 0.5\pi)$. ‖

4.4• Remarks

— Other distributions with explicit $F^{-1}(\cdot)$ include the logistic, Gumbel–minimum, Gumbel–maximum distributions and so on.

— If $F^{-1}(\cdot)$ is not available analytically, the inversion method may not be efficient.

1.1.2 Generating samples from discrete distributions

5• The issue and basic idea

- Suppose that we want to generate a random number from a discrete r.v. X with *probability mass function* (pmf)

$$\Pr(X = x_i) = p_i, \quad p_i > 0, \quad i = 1, \ldots, d, \quad \sum_{i=1}^{d} p_i = 1,$$

where d is finite or ∞. If $d < \infty$, we write $X \sim \mathrm{FDiscrete}_d(\{x_i\}, \{p_i\})$, see Appendix A.1.1.

- The cdf of $X \sim \text{FDiscrete}_d(\{x_i\}, \{p_i\})$ is

$$\Pr(X \leqslant x) = \sum_{x_i \leqslant x} p_i = \begin{cases} 0, & \text{if } x < x_1, \\ p_1, & \text{if } x_1 \leqslant x < x_2, \\ p_1 + p_2, & \text{if } x_2 \leqslant x < x_3, \\ \vdots & \vdots \\ p_1 + \cdots + p_{d-1}, & \text{if } x_{d-1} \leqslant x < x_d, \\ 1, & \text{if } x \geqslant x_d. \end{cases}$$

- To this end, we first generate $U = u \sim U(0,1)$, and then set

$$X = x = \begin{cases} x_1, & \text{if } u < p_1; \\ x_2, & \text{if } p_1 \leqslant u < p_1 + p_2; \\ \vdots & \vdots \\ x_{d-1}, & \text{if } \sum_{i=1}^{d-2} p_i \leqslant u < \sum_{i=1}^{d-1} p_i; \\ x_d, & \text{if } \sum_{i=1}^{d-1} p_i \leqslant u < 1. \end{cases}$$

5.1• The algorithm

— Step 1: Draw $U = u \sim U(0,1)$;

— Step 2: If $u < p_1$, set $X = x_1$ and stop;

If $u < p_1 + p_2$, set $X = x_2$ and stop;

$\vdots$

If $u < \sum_{i=1}^{d-1} p_i$, set $X = x_{d-1}$ and stop;

If $u < 1$, set $X = x_d$ and stop.

5.2• Remarks

— Let $F(x)$ denote the cdf of the r.v. X.

— If $\{x_i, i \geqslant 1\}$ are ordered as $x_{(1)} < x_{(2)} < \cdots$ and the corresponding probability $\Pr\{X = x_{(i)}\}$ is denoted by $p_{(i)}$, then $F(x_{(i)}) = \sum_{j=1}^{i} p_{(j)}$.

— So $X = x_{(i)}$ if $F(x_{(i-1)}) \leqslant U < F(x_{(i)})$.

5.3• The most efficient procedure

— The most efficient procedure is to rearrange $\{x_i, i \geqslant 1\}$ as $\{x'_1, x'_2, \ldots\}$ such that $\{p'_1 \geqslant p'_2 \geqslant \cdots\}$.

— Then if $U < p'_1$, set $X = x'_1$ and stop;

— If $U < p'_1 + p'_2$, set $X = x'_2$ and stop; ...

5.4• Theoretical justification

— Note that $U \sim U(0,1)$.

— For $0 < a < b < 1$ we have $\Pr(a \leqslant U \leqslant b) = b - a$. Then

$$\Pr(X = x_i) = \Pr\left(\sum_{j=1}^{i-1} p_j \leqslant U < \sum_{j=1}^{i} p_j\right) = p_i, \quad i \geqslant 1.$$

5.5• Several examples

Example 1.5 (Finite discrete distribution). Use the inversion method to generate a sample from a discrete r.v. X with pmf $p_i = \Pr(X = i)$ for $i = 1, 2, 3, 4$, where $p_1 = 0.20$, $p_2 = 0.15$, $p_3 = 0.25$ and $p_4 = 0.40$.

Solution: We can use the following algorithm:

Step 1: Draw $U = u$ from $U(0,1)$;

Step 2: If $u < 0.20$, set $X = 1$ and stop;

If $u < 0.35$, set $X = 2$ and stop;

If $u < 0.60$, set $X = 3$ and stop;

Otherwise, set $X = 4$ and stop.

Alternative solution: However, the most efficient way is to rearrange $\{x_i\}_{i=1}^4$ according to the decreasing order of $\{p_i\}_{i=1}^4$. Then we have the following algorithm:

Step 1: Draw $U = u$ from $U(0,1)$;

Step 2: If $u < 0.40$, set $X = 4$ and stop;

If $u < 0.65$, set $X = 3$ and stop;

If $u < 0.85$, set $X = 1$ and stop;

Otherwise, set $X = 2$ and stop.

R code: The built-in R function sample(x, N, prob= p, replace = F) produces a vector of length N randomly chosen from $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ with corresponding probabilities $\boldsymbol{p} = (p_1, \ldots, p_d)^\top$ without replacement. For the current case, sample(1:4, 100, c(0.20, 0.15, 0.25, 0.40), T) will produce 100 i.i.d. samples from $X \sim \text{FDiscrete}_4(\{i\}, \{p_i\})$. ‖

Example 1.6 (Poisson distribution). Use the inversion method to generate a sample from $X \sim \text{Poisson}(\lambda)$, where $p_i = \Pr(X = i) = \lambda^i \, \mathrm{e}^{-\lambda}/i!$ for $i \geqslant 0$.

Solution: Since $p_0 = \mathrm{e}^{-\lambda}$, the recursive identity between p_{i+1} and p_i is

$$\frac{p_{i+1}}{p_i} = \frac{\lambda^{i+1} \, \mathrm{e}^{-\lambda}/(i+1)!}{\lambda^i \, \mathrm{e}^{-\lambda}/i!} = \frac{\lambda}{i+1}, \quad i \geqslant 0.$$

The algorithm is as follows:

Step 1: Generate $U = u$ from $U(0, 1)$;

Step 2: Let $i = 0$, $p = p_0$ and $F = p$;

Step 3: If $u < F$, set $X = i$ and stop;

Step 4: Otherwise, let $p \leftarrow p\lambda/(i+1)$, $F \leftarrow F + p$, $i \leftarrow i + 1$ and go back to Step 3.

Comment 4: Poisson distribution has infinite possible values. The recursive method solves the problem of calculating the infinite numbers of p_i.

R code: The built-in R function rpois(N, λ) can be used to generate N i.i.d. samples from $X \sim \text{Poisson}(\lambda)$. ‖

Example 1.7 (Binomial distribution). Use the inversion method to generate a sample from $X \sim \text{Binomial}(n, \theta)$, where $p_i = \Pr(X = i) = \binom{n}{i}\theta^i(1-\theta)^{n-i}$ for $i = 0, 1, \ldots, n$.

Solution: Since $p_0 = (1-\theta)^n$, the recursive identity between p_{i+1} and p_i is

$$\frac{p_{i+1}}{p_i} = \frac{\binom{n}{i+1}\theta^{i+1}(1-\theta)^{n-i-1}}{\binom{n}{i}\theta^i(1-\theta)^{n-i}} = \frac{\theta(n-i)}{(1-\theta)(i+1)}, \quad i = 0, 1, \ldots, n-1.$$

The algorithm is as follows:

Step 1: Generate $U = u$ from $U(0, 1)$;

Step 2: Let $i = 0$, $p = p_0$ and $F = p$;

Step 3: If $u < F$, set $X = i$ and stop;

Step 4: Otherwise, let $p \leftarrow p\theta(n-i)/\{(1-\theta)(i+1)\}$, $F \leftarrow F + p$, $i \leftarrow i+1$ and go back to Step 3.

Comment 5: In real application, we can use R program to calculate all $\{p_i\}_{i=1}^n$ simultaneously, and it just a pmf calculation. It will be more efficient than the recursive method.

R code: The binomial generator in R is rbinom(N, n, p). ‖

6• A summary on the inversion method

6.1• A key result

— If X has a continuous cdf $F(\cdot)$, then $F(X) \sim U(0, 1)$.

6.2• Stochastic representation (SR)

— Let $U \sim U(0, 1)$, then $F(X) \overset{\text{d}}{=} U$ or $X \overset{\text{d}}{=} F^{-1}(U)$.

— The SR '$\overset{\text{d}}{=}$' operator will be introduced in §1.5.1.

6.3• Efficiency

— The efficiency of the inversion method depends on the degree of difficulty for finding the root $x = F^{-1}(u)$ from the equation $F(x) = u$ with one unknown quantity.

6.4• Question

— Since we have many algorithms to solve the solution to $F(x) = u$, why do we need other r.v. generation methods?

7• Possible answers to the above question

7.1• A drawback of Newton's method

— If $F^{-1}(\cdot)$ is not available analytically, we could use Newton's method to find the roots of $F(x_i) = u_i$, where $u_1, \ldots, u_n \overset{\text{iid}}{\sim} U(0, 1)$.

— However, Newton's method is sensitive to the choice of the initial values.

— That is, Newton's method may be divergent if a poor initial value is chosen.

7.2• Infeasibility

— When n is very large (say, $n = 100{,}000$), it is not an easy task for the computer to automatically set 100,000 initial values.

7.3• Inefficiency

— Even Newton's method works, say, it takes one second for getting the root of a single equation. Then, $100{,}000/3{,}600 = 27.78$ hours are needed to find 100,000 roots.

— In other words, if the convergence speed is too slow, the inversion method may not be efficient in such cases.

— Therefore, we need to consider other r.v. generation methods.

1.2 The Grid Method

8• Formulation of the grid method

- Let $X \sim f(x)$ be a continuous r.v. with finite support $\mathcal{S}_X = [a, b]$, where $a \neq -\infty$ and $b \neq \infty$.

- To generate X, we first select a set of appropriate grid points $\{x_i\}_{i=1}^d$ covering the support $\mathcal{S}_X$, and then approximate the pdf $f(x)$ by a discrete distribution at $\{x_i\}_{i=1}^d$ with probabilities

$$p_i = \frac{f(x_i)}{\sum_{j=1}^d f(x_j)}, \quad i = 1, \ldots, d.$$

- In other words, we have $X \dot{\sim} \text{FDiscrete}_d(\{x_i\}, \{p_i\})$.

- To generate this finite discrete distribution, we may use the built-in R function sample(x, N, prob= p, replace = F), see Appendix A.1.1.

8.1• Difference between a finite support and an infinite support

— If $X \sim U(0,1)$, its pdf is $f(x) = 1 \cdot I(0 < x < 1)$, then $\mathcal{S}_X = \{x\colon\ f(x) > 0\} = (0,1)$, which is a finite interval.

— If $X \sim \text{Exponential}(\beta)$, its pdf is $f(x) = \beta\, \mathrm{e}^{-\beta x} \cdot I(x \geqslant 0)$, then $\mathcal{S}_X = [0, \infty)$, which is an infinite interval.

8.2• How to choose the grid points in practice?

— Let $\mathcal{S}_X = [a,b]$, we may select the equal-space grid points

$$x_i = a + \frac{i(b-a)}{d} \qquad i = 1, \ldots, d.$$

— When $d \to \infty$, $\{x_i\}_{i=1}^d$ cover the whole interval $[a,b]$.

9• An advantage of the grid method

- Obviously, the grid method also works for an un-normalized density.

- A pdf $f(x)$ is said to be *un-normalized*, if it is of the form

$$f(x) = \frac{g(x)}{c^*},$$

where the normalizing constant c^* is unknown and the kernel $g(x)$ is known.

- In fact, we have

$$p_i = \frac{g(x_i)/c^*}{\sum_{j=1}^d g(x_j)/c^*} = \frac{g(x_i)}{\sum_{j=1}^d g(x_j)}, \quad i = 1, \ldots, d.$$

Example 1.8 (Marginal distribution of the truncated bi-normal distribution). Let $(X, Y)^\top$ follow the truncated two-dimensional normal distribution (see Exercise 1.1) with known parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{a}, \boldsymbol{b})$, then the marginal density of X is $f(x) = g(x)/c^*$, where c^* is the unknown normalizing constant and

$$\begin{aligned} g(x) &= \mathrm{e}^{-(x-\mu_1)^2/(2\sigma_1^2)} \times \left\{ \Phi\left(\frac{b_2 - \mu_2 - \rho\sigma_2\sigma_1^{-1}(x-\mu_1)}{\sigma_2\sqrt{1-\rho^2}} \right) \right. \\ &\quad \left. - \Phi\left(\frac{a_2 - \mu_2 - \rho\sigma_2\sigma_1^{-1}(x-\mu_1)}{\sigma_2\sqrt{1-\rho^2}} \right) \right\}, \ x \in [a_1, b_1]. \end{aligned} \tag{1.1}$$

Use the grid method to generate samples of size $N = 200{,}000$ from X.

Solution: When $-\infty < a_1 < b_1 < \infty$, we may select $x_i = a_1 + (b_1 - a_1)i/d$ for $i = 1, \ldots, d$ and $d = 300$, say. Figure 1.2 shows the histogram based on 200,000 i.i.d. samples generated via the grid method with sample(x, 200,000, prob = p, replace = T).

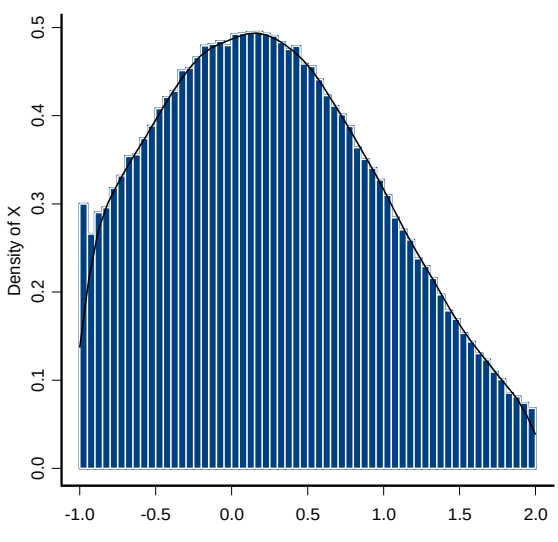

Figure 1.2 The histogram of the marginal density of X based on 200,000 i.i.d. samples generated via the grid method with sample(x, 200,000, prob = p, replace = T), where $(X, Y)^\top \sim \mathrm{TN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{a}, \boldsymbol{b})$, $\mu_1 = \mu_2 = 0$, $\rho = 0.5$, $\sigma_1 = \sigma_2 = 1$, $\boldsymbol{a} = (-1, -1)^\top$, $\boldsymbol{b} = (2, 2)^\top$. ‖

10• Remarks on the grid method

10.1• Applicability

— If (i) the normalizing constant of $f(x)$ is either known or unknown; and (ii) the support of X is a finite interval, i.e., $\mathcal{S}_X = [a, b]$, where $-\infty < a$ and $b < \infty$, then the grid method is an efficient method to generate a random sample of X from $f(\cdot)$.

10.2• Inapplicability

— If $\mathcal{S}_X$ is an infinite interval, the grid method cannot be applied.

1.3 The Rejection Method

1.3.1 Generating samples from continuous distributions

11• The basic idea of the rejection method

- If direct sampling from the target density $f(x)$ with support $\mathcal{S}_X$ is very difficult or inefficient, but the sampling from another density $g(x)$ having same support $\mathcal{S}_X$ is relatively easy.
- Hence, we can transfer the target density to the surrogate density $g(x)$ by first sampling from $g(x)$ and then adjusting the generated samples such that part of them become i.i.d. samples from $f(x)$.

12• Formulation of the rejection method

12.1• Aim

— Suppose that we want to generate random samples from the *target* density $f(x)$ with support $\mathcal{S}_X$.

12.2• Conditions

— If we could find some *envelope* constant $c\,(\geqslant 1)$ and an *envelope* density $g(x)$ having the same support $\mathcal{S}_X$ so that $f(x)$ minorizes $cg(x)$; i.e.,

$$f(x) \leqslant cg(x), \quad \forall\, x \in \mathcal{S}_X, \tag{1.2}$$

as shown in Figure 1.3, then we can apply the following procedure suggested by von Neumann (1951) to generate i.i.d. samples from $f(x)$.

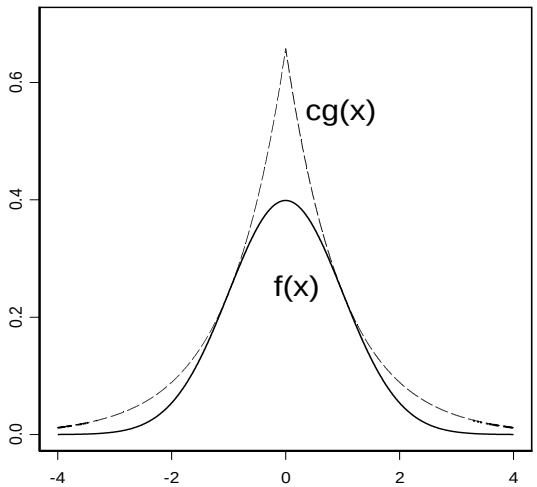

Figure 1.3 Illustration of the rejection method.

12.3• The rejection method

— Step 1: Draw $U \sim U(0,1)$ and independently draw $Y \sim g(\cdot)$;

— Step 2: If $U \leqslant f(Y)/\{cg(Y)\}$, return $X = Y$; Otherwise, go to Step 1.

12.4• Another name

— The rejection method is also called the *acceptance–rejection* (AR) method or algorithm.

13• How to understand the AR algorithm?

13.1• Adjustor

— Let $X \sim f(\cdot)$, $Y \sim g(\cdot)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g(\cdot)$. Furthermore, we define

$$h(\cdot) = \frac{f(\cdot)}{cg(\cdot)}, \tag{1.3}$$

which is called the *adjustor*.

— The samples $\{Y_i\}_{i=1}^n$ can be divided into two groups: say $\{Y_i\}_{i=1}^m$ and $\{Y_i\}_{i=m+1}^n$ such that $\{U_i \leqslant h(Y_i)\}_{i=1}^m$ while $\{U_i > h(Y_i)\}_{i=m+1}^n$, where $U_1, \ldots, U_n \overset{\text{iid}}{\sim} U(0,1)$.

— Thus, we return $X_i = Y_i$ for $i = 1, \ldots, m$. That is $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} f(\cdot)$.

13.2• Understanding Step 2 of the AR algorithm

— In Step 2, $f(Y)$, $g(Y)$ and $h(Y)$ are three r.v.'s.

— The adjustor $h(Y)$ is independent of the U.

— The adjustor should be bounded between 0 and 1, i.e.,

$$0 \leqslant h(Y) \leqslant 1. \tag{1.4}$$

Otherwise, if $h(Y) > 1$, then the condition $\{U \leqslant h(Y)\}$ in Step 2 is always satisfied, we will return all $\{Y_i\}_{i=1}^n$ as $\{X_i\}_{i=1}^n$.

— Step 2 indicates that X has the conditional distribution of Y given that the event $\{U \leqslant h(Y)\}$ occurs. In other words,

$$f(x) = f_{_Y}(x|U \leqslant h(Y)) \tag{1.5}$$

where $f_{_Y}(x|U \leqslant h(Y))$ denotes the conditional density of $Y|\{U \leqslant h(Y)\}$ evaluated at $Y = x$.

14• Theoretical justification of the AR algorithm

14.1• Three forms of Bayes Theorem

— The first form in events: $\Pr(\mathbb{A}|\mathbb{B}) = \Pr(\mathbb{B}|\mathbb{A})\Pr(\mathbb{A})/\Pr(\mathbb{B})$.

— The second form in densities: $f(x|y) = f(y|x)f(x)/f(y)$.

— The third form in mixture: $f_{_Y}(x|\mathbb{B}) = \Pr(\mathbb{B}|x)f_{_Y}(x)/\Pr(\mathbb{B})$.

14.2• Proof of (1.5)

— From (1.3), we have

$$h(x)g(x) = \frac{1}{c} \cdot f(x). \tag{1.6}$$

— In the third form of Bayes Theorem, note that $f_{_Y}(x) = g(x)$ and let $\mathbb{B} = \{U \leqslant h(Y)\}$, we obtain (Rubinstein & Kroese 2004, p.23):

$$f_{_Y}(x|U \leqslant h(Y)) = \frac{\Pr\{U \leqslant h(Y)|Y = x\}g(x)}{\Pr\{U \leqslant h(Y)\}}. \tag{1.7}$$

— Note that the cdf of $U \sim U(0,1)$ is

$$\Pr(U \leqslant u) = 0 \cdot I(u \leqslant 0) + u \cdot I(0 < u < 1) + 1 \cdot I(u \geqslant 1). \tag{1.8}$$

— Since Y is independent of U, we have

$$\begin{aligned} \Pr\{U \leqslant h(Y) | Y = x\} &= \Pr\{U \leqslant h(x) | Y = x\} \\ &= \Pr\{U \leqslant h(x)\} \overset{(1.8)}{=} h(x) \end{aligned} \tag{1.9}$$

and

$$\begin{aligned} \Pr\{U \leqslant h(Y)\} &= \int_{\mathcal{S}_X} \Pr\{U \leqslant h(Y) | Y = x\} g(x)\,\mathrm{d}x \\ &\overset{(1.9)}{=} \int_{\mathcal{S}_X} h(x) g(x)\,\mathrm{d}x \\ &\overset{(1.6)}{=} \int_{\mathcal{S}_X} \frac{1}{c} \cdot f(x)\,\mathrm{d}x = \frac{1}{c}. \end{aligned} \tag{1.10}$$

— Therefore, (1.7) becomes $f_Y(x | U \leqslant h(Y)) = f(x)$, implying (1.5). □

1.3.2 The efficiency of the rejection method

15• The acceptance probability

- From (1.10), the efficiency of the rejection method is determined by the *acceptance probability* $1/c$.
- Thus, a smaller c will result in a better rejection method.

16• Determination of the envelope constant c

- The c can be determined by maximizing the ratio of $f(x)$ over $g(x)$:

$$c = \max_{x \in \mathcal{S}_X} \frac{f(x)}{g(x)}. \tag{1.11}$$

16.1• The motivation of (1.11)

— From (1.4), we have to find a c such that

$$\frac{f(Y)}{g(Y)} \leqslant c. \tag{1.12}$$

— If we specify c by (1.11), then

$$c = \max_{x\in\mathcal{S}_X} \frac{f(x)}{g(x)} = \max_{y\in\mathcal{S}_Y} \frac{f(y)}{g(y)} \geqslant \frac{f(Y)}{g(Y)},$$

indicating that (1.11) is a *sufficient condition* of (1.12).

16.2• A smaller c is desired

— If we have found a c satisfying (1.12), then, for any $c^* \geqslant 1$, we obtain

$$\frac{f(Y)}{g(Y)} \leqslant c \leqslant c \cdot c^* \doteq c_1,$$

i.e., we can find many c_1's to satisfy (1.12). We prefer a smaller c.

16.3• The optimal c

— In practice, we usually consider a family of densities indexed by θ, say $\{g_\theta(x)\colon\ \theta \in \Theta\}$, as the candidate envelopes.

— Define

$$c_\theta = \max_{x\in\mathcal{S}_X} \frac{f(x)}{g_\theta(x)} = \frac{f(\hat{x})}{g_\theta(\hat{x})}, \tag{1.13}$$

where $\hat{x}$ is the mode of the function $f(x)/g_\theta(x)$, i.e.,

$$\hat{x} = \arg\max_{x\in\mathcal{S}_X} \frac{f(x)}{g_\theta(x)}.$$

— Since $\log(\cdot)$ is a monotonic increasing function, we know that both $h(\cdot)$ and $\log\{h(\cdot)\}$ share the mode. Hence we have

$$\hat{x} = \arg\max_{x\in\mathcal{S}_X} \Big\{ \log f(x) - \log g_\theta(x) \Big\}.$$

— Then, the optimal c is $c_{\text{opt}} = \min_{\theta\in\Theta} c_\theta$.

17• The AR algorithm for an un-normalized target density

- The representation of (1.11) implies that the AR algorithm is also available if we replace the normalized pdf $f(\cdot)$ with its un-normalized density $f^*(\cdot)$; that is, the normalizing constant of $f(\cdot)$ is not necessary to be known (see Example 1.13).

- In fact, let $f(x) = K \cdot f^*(x)$, then

$$c = \max_{x \in \mathcal{S}_X} \frac{K \cdot f^*(x)}{g(x)} = K \cdot \max_{x \in \mathcal{S}_X} \frac{f^*(x)}{g(x)} = K \cdot c^*$$

so that

$$U \leqslant \frac{f(Y)}{cg(Y)} = \frac{K \cdot f^*(Y)}{K \cdot c^* g(Y)} = \frac{f^*(Y)}{c^* g(Y)}.$$

- This property is particularly important in Bayesian calculations.

18• Expected iteration number until one sample is accepted

- Let N be the number of iterations in the AR algorithm, i.e., the number of pairs (U, Y) required before a successful pair (U, X) occurs, then N has a geometric distribution (see **12.2•** in Appendix A.1.5):

$$\Pr(N = n) = p(1-p)^{n-1}, \quad n = 1, \ldots, \infty,$$

where $p = \Pr\{U \leqslant h(Y)\} = 1/c$.

- Thus, the expected number of iterations until one sample is accepted is $E(N) = 1/p = c$.

1.3.3 Several examples

Example 1.9 (Beta distribution). Use the uniform pdf $g(x) = 1$ for $x \in (0, 1)$ as the envelope function to generate a random variable having the beta density

$$f(x) = 20x(1-x)^3, \quad 0 < x < 1$$

by the rejection method. What is the acceptance probability for this AR algorithm?

Solution: (i) By differentiating the ratio $f(x)/g(x) = 20x(1-x)^3$ with respect to x and setting the resultant derivative equal to zero, we obtain the maximal value of this ratio at $x = 1/4$. Hence

$$c = \max_{0<x<1} \frac{f(x)}{g(x)} = 20 \times \frac{1}{4} \left(\frac{3}{4}\right)^3 = \frac{135}{64},$$

and

$$\frac{f(x)}{cg(x)} = \frac{256}{27} x(1-x)^3.$$

(ii) The rejection method is as follows:

Step 1: Draw $U, Y \overset{\text{iid}}{\sim} U(0,1)$;

Step 2: If $U \leqslant (256/27)Y(1-Y)^3$, set $X = Y$; Otherwise, go to Step 1.

(iii) The acceptance probability is $1/c = 64/135 \approx 0.4740741$. ‖

Example 1.10 (Gamma distribution). Use a density, selected from the following family of exponential densities

$$g_\theta(x) = \theta\, \mathrm{e}^{-\theta x}, \quad x > 0, \quad 0 < \theta < 1$$

as the optimal envelope function (i.e., with the largest acceptance probability) to generate a random variable having the gamma density

$$f(x) = \frac{1}{\Gamma(3/2)} x^{1/2}\, \mathrm{e}^{-x}, \quad x > 0$$

by the rejection method. Calculate the expected number of iterations until one acceptance and the value of the acceptance probability.

Solution: (i) The ratio is

$$\frac{f(x)}{g_\theta(x)} = \frac{\frac{1}{\Gamma(3/2)} x^{1/2}\, \mathrm{e}^{-x}}{\theta\, \mathrm{e}^{-\theta x}} = \frac{x^{1/2}\, \mathrm{e}^{(\theta-1)x}}{\Gamma(3/2)\theta}.$$

By differentiating the log-ratio with respect to x and setting the resultant derivative equal to zero, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}x} \log\left\{\frac{f(x)}{g_\theta(x)}\right\} = \frac{\mathrm{d}}{\mathrm{d}x}\left\{\frac{1}{2}\log x + (\theta-1)x\right\} = \frac{1}{2x} + \theta - 1,$$

we obtain that the maximum of this ratio is arrived at $x = 0.5/(1-\theta)$. Thus

$$c_\theta = \frac{f(0.5/(1-\theta))}{g_\theta(0.5/(1-\theta))} = \frac{1}{\Gamma(3/2)}\sqrt{\frac{1}{2\,\mathrm{e}}} \cdot \frac{1}{\theta(1-\theta)^{1/2}},$$

and

$$c_{\text{opt}} = \min_{0<\theta<1} c_\theta = \min_{0<\theta<1}\left\{\frac{1}{\Gamma(3/2)}\sqrt{\frac{1}{2\,\mathrm{e}}} \cdot \frac{1}{\theta(1-\theta)^{1/2}}\right\}.$$

Let

$$H(\theta) = \log\left\{\frac{1}{\theta(1-\theta)^{1/2}}\right\} = -\log\theta - \frac{1}{2}\log(1-\theta),$$

and set

$$0 = H'(\theta) = -\frac{1}{\theta} + \frac{1}{2(1-\theta)},$$

we have $\theta = 2/3$. Hence

$$c_{\text{opt}} = \frac{1}{\Gamma(3/2)}\sqrt{\frac{1}{2\,\mathrm{e}}} \cdot \frac{1}{\frac{2}{3}\sqrt{\frac{1}{3}}} = \frac{3}{2\Gamma(3/2)}\sqrt{\frac{3}{2\,\mathrm{e}}} = \frac{3^{3/2}\,\mathrm{e}^{-0.5}}{2^{3/2}\Gamma(3/2)},$$

$$\frac{f(x)}{cg_\theta(x)} = \frac{f(x)}{c_{\text{opt}} g_{\frac{2}{3}}(x)} = \left(\frac{2\,\mathrm{e}x}{3}\right)^{1/2} \mathrm{e}^{-x/3}.$$

On the other hand, the cdf corresponding to $g_\theta(x)$ is

$$G_\theta(x) = \int_0^x g_\theta(t)\,\mathrm{d}t = 1 - \mathrm{e}^{-\theta x}, \quad x > 0.$$

Let $u = G_\theta(x)$, then $x = G_\theta^{-1}(u) = -(1/\theta)\log(1-u)$ for $0 < u < 1$. Especially, when $\theta = 2/3$, we have $x = G_{2/3}^{-1}(u) = -1.5\log(1-u)$.

(ii) The Gamma(3/2, 1) random variable can be generated as follows:

Step 1: Draw $U_1, U_2 \overset{\text{iid}}{\sim} U(0,1)$ and set $Y = -1.5\log(1-U_1)$;

Step 2: If $U_2 \leqslant (2\,\mathrm{e}Y/3)^{1/2}\,\mathrm{e}^{-Y/3}$, set $X = Y$; Otherwise, go to Step 1.

(iii) The expected number of iterations until one sample is accepted is: $c_{\text{opt}} \approx 1.257317$. The acceptance probability is: $1/c_{\text{opt}} \approx 0.7953444$. ‖

Example 1.11 (Generating normal distribution from Laplace distribution). Use a density, selected from the following family of Laplace densities

$$\text{Laplace}(x|0,\sigma^2) = \frac{1}{2\sigma}\,\mathrm{e}^{-|x|/\sigma}, \quad x \in \mathbb{R}, \quad \sigma > 0$$

as the optimal envelope function to generate a random variable from the standard normal distribution by the AR algorithm. Calculate the acceptance probability.

Solution: From (1.13), it is easy to obtain

$$c_\sigma = \max_{x\in\mathbb{R}} \frac{N(x|0,1)}{\text{Laplace}(x|0,\sigma^2)} = \max_{x\in\mathbb{R}} \sqrt{\frac{2}{\pi}}\,\sigma \exp\left(-\frac{x^2}{2} + \frac{|x|}{\sigma}\right)$$

$$= \sqrt{\frac{2}{\pi}}\,\sigma \exp\left(\frac{1}{2\sigma^2}\right) \text{ with } \hat{x} = \frac{\text{sgn}(x)}{\sigma}.$$

The optimal $\sigma = 1$ so that the optimal envelope constant is

$$c_{\text{opt}} = \min_{\sigma>0} c_\sigma = \sqrt{2\,\mathrm{e}/\pi} \approx 1.3155.$$

Thus, the acceptance probability is $1/c_{\text{opt}} \approx 0.7602$. ‖

Example 1.12 (Generating normal distribution from Cauchy distribution). Use a density, selected from the following family of Cauchy densities

$$g_\theta(x) = \frac{\theta}{\pi(x^2+\theta^2)}, \quad x \in \mathbb{R}, \quad \theta > 0$$

as the optimal envelope function to generate a random variable from $N(0,1)$ by the AR algorithm. Calculate the acceptance probability.

Solution: (i) We can show

$$\begin{aligned} c_\theta &= \max_{x\in\mathbb{R}} \frac{N(x|0,1)}{g_\theta(x)} = \sqrt{\frac{\pi}{2}} \cdot \max_{x\in\mathbb{R}} \frac{x^2+\theta^2}{\theta}\,\mathrm{e}^{-x^2/2} \\ &= \begin{cases} \sqrt{2\pi}(\theta\,\mathrm{e})^{-1}\exp(\theta^2/2), & \text{if } 0<\theta\leqslant\sqrt{2}, \\ \theta\sqrt{\pi/2}, & \text{if } \theta>\sqrt{2}. \end{cases} \end{aligned}$$

Note that f/g_θ and $\log(f/g_\theta)$ share the mode. Setting the derivative with respect to x of $\log(f/g_\theta)$ equal to 0 yields the equation

$$-x + \frac{2x}{x^2+\theta^2} = 0.$$

This gives the values $x = 0$ and $x = \pm\sqrt{2-\theta^2}$ (the latter case can happen only when $0 < \theta \leqslant \sqrt{2}$). At $x = 0$, f/g_θ takes the value $\theta\sqrt{\pi/2}$. At $x = \pm\sqrt{2-\theta^2}$, f/g_θ takes the value $\sqrt{2\pi}(\theta\,\mathrm{e})^{-1}\exp(\theta^2/2)$.

It is easy to see that for $0 < \theta \leqslant \sqrt{2}$, the maximum of f/g_θ is attained at $x = \pm\sqrt{2-\theta^2}$. For $\theta > \sqrt{2}$, the maximum is attained at $x = 0$. This concludes the verification of the expression for c_θ.

(ii) The function c_θ has only one minimum at $\theta = 1$. The minimal value is

$$c_{\text{opt}} = \min_{\theta>0} c_\theta = \sqrt{2\pi/\,\mathrm{e}} \approx 1.5203.$$

Thus, the acceptance probability is $1/c_{\text{opt}} \approx 0.6577$.

(iii) The rejection method is as follows:

Step 1: Draw $V \sim U(0,1)$ and independently draw $Y \sim g_1(\cdot)$;

Step 2: If $V \leqslant f(Y)/\{c_{\text{opt}} \cdot g_1(Y)\} = 0.5(Y^2+1)\exp(0.5-0.5Y^2)$, return $X = Y$; Otherwise, go to Step 1.

Comment 6: To generate a random sample of Y from $g_1(\cdot)$, which is the standard Cauchy distribution, we can use the inversion method as shown in Example 1.4 with $\mu = 0$ and $\sigma = 1$. In other words, we first generate $U \sim U(0,1)$ and then set $Y = \tan(\pi U - 0.5\pi)$; or generate $U_1 \sim U(-\pi/2, \pi/2)$ and set $Y = \tan(U_1)$. ‖

1.3.4 Log-concave densities

19• Definition of a log-concave density

- We say a pdf $f(x)$ is *log-concave* if its logarithm is *concave*; i.e.,

$$\frac{\mathrm{d}^2 \log\{f(x)\}}{\mathrm{d}x^2} \leqslant 0.$$

- In other words, $f(x)$ is log-concave *if and only if* (iff) $\log\{f(x)\}$ is concave.

20• Piece-wise exponential envelope densities

- On a log scale, an exponential density is a straight line. In fact, let $f(x) = \beta\exp(-\beta x)$ for $x \geqslant 0$ and $\beta > 0$, then $\log\{f(x)\} = \log(\beta) - \beta x$, which is a straight line with intercept $\log(\beta)$ and slope $-\beta$.
- If a pdf $f(x)$ is log-concave, then any straight line *tangent* to $\log\{f(x)\}$ will lie above $\log\{f(x)\}$.
- Thus, log-concave densities are ideally suited to the rejection method with piece-wise exponential envelopes (Lange 1999, p.273).
- Table 1.1 lists some log-concave densities.

20.1• Remarks

— A strictly log-concave pdf $f(x)$ defined on an interval is unimodal.

— The mode $\tilde{x}$ of $f(x)$ may occur at either endpoint or on the interior of the interval.

Table 1.1 Some log-concave densities

Distribution	Kernel of $f(x)$	$\mathrm{d}^2 \log\{f(x)\}/\mathrm{d}x^2$	Condition
Normal	$\exp\{-(x-\mu)^2/(2\sigma^2)\}$	$-1/\sigma^2$	$\sigma > 0$
Gamma	$x^{\alpha-1}\exp(-\beta x)$	$-(\alpha-1)/x^2$	$\alpha \geqslant 1,\ \beta > 0$
Beta	$x^{a-1}(1-x)^{b-1}$	$-\frac{\alpha-1}{x^2} - \frac{\beta-1}{(1-x)^2}$	$\alpha \geqslant 1,\ \beta \geqslant 1$
Logistic	$\mathrm{e}^{-\frac{x-\mu}{\sigma}}/(1+\mathrm{e}^{-\frac{x-\mu}{\sigma}})^2$	$\frac{-2}{\sigma^2}\mathrm{e}^{-\frac{x-\mu}{\sigma}}/(1+\mathrm{e}^{-\frac{x-\mu}{\sigma}})^2$	$\sigma > 0$
Gumbel	$\mathrm{e}^{-\frac{x-\mu}{\sigma}}\exp(-\mathrm{e}^{-\frac{x-\mu}{\sigma}})$	$-\sigma^{-2}\exp\{-(x-\mu)/\sigma\}$	$\sigma > 0$
Weibull	$x^{\alpha-1}\exp(-\beta x^{\alpha})$	$-(\alpha-1)(x^{-2}+\beta\alpha x^{\alpha-2})$	$\alpha \geqslant 1,\ \beta > 0$

— In the former case, we suggest using a truncated exponential (see Figure 1.4 and Example 1.13).

— In the latter case, we suggest using two truncated exponential envelopes oriented in opposite directions from the mode $\tilde{x}$ (see Figure 1.3).

Example 1.13 (Truncated univariate normal distribution). Use the AR algorithm to generate random samples of $X \sim \mathrm{TN}(\mu, \sigma^2; a, \infty)$ with pdf

$$f(x) \propto \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} I(x > a).$$

Solution: (i) When $a < \mu$, we continuously generate i.i.d. samples from $N(\mu, \sigma^2)$ until one sample satisfying $X > a$ occurs. In the worst case, the efficiency of this method is 50%.

(ii) When $a > \mu$, especially when $a \gg \mu$, the above strategy is very inefficient. Without loss of generality, let $\mu = 0$ and $\sigma = 1$. We consider a truncated exponential envelope family with density $b\,\mathrm{e}^{-b(x-a)} I(x > a)$ indexed by b (Robert 1995). From (1.13), it is easy to obtain

$$c_b = \max_{x>a} \frac{f(x)}{b\,\mathrm{e}^{-b(x-a)}} = \begin{cases} (bc)^{-1}\exp(0.5b^2 - ba), & \text{if} \quad b > a, \\ (bc)^{-1}\exp(-0.5a^2), & \text{if} \quad b \leqslant a, \end{cases}$$

where $c = \sqrt{2\pi}\{1 - \Phi(a)\}$. The first bound is minimized at

$$b^* = 0.5(a + \sqrt{a^2 + 4}\,),$$

whereas $\hat{b} = a$ minimizes the second bound. The optimal choice of b is therefore b^*.

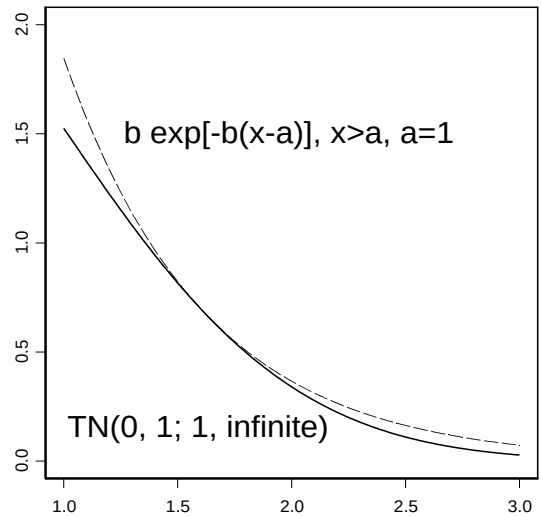

Figure 1.4 A truncated exponential envelope for the truncated normal distribution $\mathrm{TN}(0, 1; 1, \infty)$. ‖

21• Remarks on the AR algorithm

21.1• Requirements for the rejection method

— It requires to find an envelope density $g(x)$ satisfying
∗ ∗ ∗ it has the same support as $f(x)$;
∗ ∗ ∗ it has a larger variance/dispersion than $f(x)$;
∗ ∗ ∗ it is easy to draw a sample from $g(x)$.

— It requires to find an envelope constant

$$c = \max_{x \in \mathcal{S}_X} \frac{f(x)}{g(x)} \quad \text{or} \quad c_{\text{opt}} = \min_{\theta \in \Theta} \max_{x \in \mathcal{S}_X} \frac{f(x)}{g_\theta(x)}. \tag{1.14}$$

21.2• Some merits of the rejection method

— It can generate i.i.d. samples *exactly* from the target density $f(x)$.

— Its *stopping rule* is very clear. That is, once the sampling process is terminated, we can obtain i.i.d. random samples from $f(x)$; while the stopping rule for an MCMC is not clear.

— It can be applied to both cases: (i) $f(x)$ is completely known; (ii) $f(x)$ could be known up to a normalizing constant, i.e., $f(x) = K \cdot f^*(x)$ with unknown K and known $f^*(x)$.

21.3• Two drawbacks of the rejection method

— It is usually very difficult to find an *automatic* envelope density $g(x)$ or a class of envelope densities $g_\theta(x)$ indexed by a parameter θ.

— Finding a good or optimal envelope constant c is a vital step for the use of rejection method. Sometimes, the optimization (1.14) itself is even more difficult than the sampling from $f(x)$.

1.4 The Sampling/Importance Resampling (SIR) Method

22• Why do we need the SIR method?

- As a *non-iterative* sampling procedure, the *sampling/importance resampling* (SIR) method proposed by Rubin (1987b, 1988) can bypass the *second problem* associated with the rejection method as shown in **21.3•**.

23• Background and basic idea of the SIR method

- If it is very difficult to generate a sample from $f(x)$, but it is relatively easy to draw a sample from $g(x)$, then we can write

$$f(x) = \frac{f(x)}{g(x)} \cdot g(x) \hat{=} \mathrm{w}(x) \cdot g(x).$$

- First, we could generate $X^{(1)}, \ldots, X^{(J)} \overset{\text{iid}}{\sim} g(x)$.

- Second, we adjust the generated samples $\{X^{(j)}\}_{j=1}^{J}$ so that part of them becomes samples from $f(x)$ by comparing their ratios $\mathrm{w}(X^{(j)}) = f(X^{(j)})/g(X^{(j)})$.

1.4.1 The SIR without replacement

24• Formulation of the sir method

- The SIR method generates an i.i.d. sample of size m *approximately* from the target density $f(x)$ with support $\mathcal{S}_X$.
- It consists of a sampling step and an importance resampling step.
- Specifically, it starts by simulating J i.i.d. samples $\{X^{(j)}\}_{j=1}^{J}$ from an *importance sampling density* or *proposal density* $g(x)$ with the same support $\mathcal{S}_X$. Then, it calculates the ratios

$$\mathrm{w}(X^{(j)}) = f(X^{(j)})/g(X^{(j)}) \tag{1.15}$$

and probabilities

$$\omega_j = \frac{\mathrm{w}(X^{(j)})}{\sum_{j'=1}^{J} \mathrm{w}(X^{(j')})}, \quad j = 1, \ldots, J. \tag{1.16}$$

- Finally, a second sample of size I ($I \leqslant J$) is drawn from the discrete distribution on $\{X^{(j)}\}_{j=1}^{J}$ with probabilities $\{\omega_j\}_{j=1}^{J}$. Figure 1.5(a) depicts the relationship between $f(x)$ and $g(x)$.

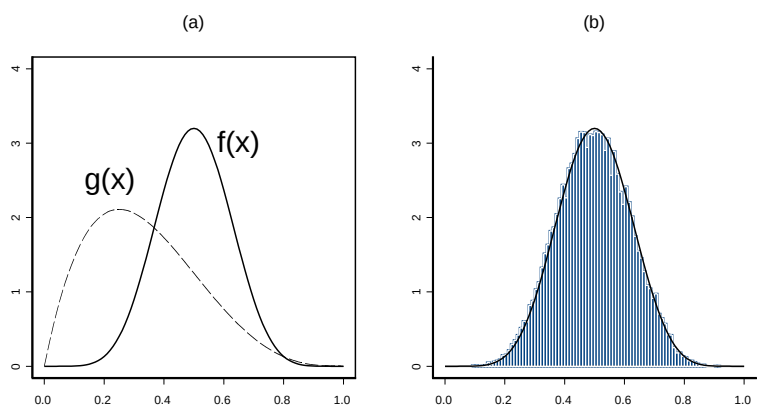

Figure 1.5 Illustration of the SIR method. (a) The target density $f(x)$ is defined by (1.17) with $r = 6$ and the importance sampling density $g(x) = \mathrm{Beta}(x|2, 4)$; (b) The histogram of $f(x)$ is obtained by using the SIR method with $J = 200{,}000$ and $I = 20{,}000$.

24.1• The SIR method without replacement

— Step 1: Generate $X^{(1)}, \ldots, X^{(J)} \overset{\text{iid}}{\sim} g(x)$;

— Step 2: Select a subset $\{X^{(k_i)}\}_{i=1}^{I}$ from $\{X^{(j)}\}_{j=1}^{J}$ via re-sampling *without replacement* from the discrete distribution on $\{X^{(j)}\}_{j=1}^{J}$ with probabilities $\{\omega_j\}_{j=1}^{J}$.

24.2• How to choose the proposal function $g(x)$?

— As expected, the output of the SIR algorithm is good if $g(x)$ is close to $f(x)$ or J/I is large.

— A good choice of J should depend on how close $g(x)$ to $f(x)$.

— If $g(x) \equiv f(x)$, we can set $I = J$.

24.3• How to choose the sample sizes I and J?

— The poorer is $g(x)$ as an approximation of $f(x)$, the larger is J compared to I.

— Rubin (1988) showed that the SIR algorithm is exact when $J/I \to \infty$.

— In practice, Rubin (1987b) suggested $J/I = 20$ and Smith & Gelfand (1992) recommended $J/I \geqslant 10$ in their examples.

Example 1.14 (Simulation from a density defined in the unit interval). Use the SIR algorithm to generate random samples of $X \sim f(x)$ with

$$f(x) = \frac{\pi \sin^r(\pi x)}{B(\frac{1}{2}, \frac{r+1}{2})}, \quad 0 < x < 1, \tag{1.17}$$

where r is a known positive integer.

Solution: For example, when $r = 6$, we consider a skew beta density, say $\text{Beta}(x|2, 4)$, as the importance sampling density $g(x)$, see Figure 1.5(a). Thus, the ratio is

$$\text{w}(x) = \frac{f(x)}{\text{Beta}(x|2, 4)}.$$

We run the SIR algorithm by setting $J = 200{,}000$ and $I = 20{,}000$. Figure 1.5(b) shows that the histogram entirely recovers the target density function $f(x)$. ‖

25• Advantages of the SIR method

- An important feature of the SIR method is that it is non-iterative (Rubin 1988, p.396).
- Another advantage is its simplicity.
- In addition, the SIR method allows f to be known up to a normalizing constant since the probabilities $\{\omega_j\}_{j=1}^J$ do not alter.
- The SIR method can easily be understood.

26• Applicability of the SIR method

- The SIR method can be used as a general tool for full Bayesian analysis (Albert 1993).
- Besides Bayesian computation, the SIR method has been successfully applied in many statistical problems, including weighted likelihood bootstrap (Newton & Raftery 1994), population dynamics model (Raftery *et al.* 1995), stock assessment (McAllister *et al.* 1994; McAllister & Ianelli 1997).
- Givens & Raftery (1996) considered adaptive versions of the SIR.
- Skare *et al.* (2003) proposed an improved SIR algorithm.

1.4.2 Theoretical justification

27• Two preliminaries

27.1• The cdf of a finite discrete distribution

— Let $Y \sim \text{FDiscrete}_J(\{y_j\}, \{\omega_j\})$, then the cdf of Y is

$$\Pr(Y \leqslant y) \quad = \quad \sum_{j=1}^{J} \omega_j \cdot I(y_j \leqslant y) \tag{1.18}$$

$$= \begin{cases} 0, & & y & < y_1, \\ \omega_1 & y_1 \leqslant & y & < y_2, \\ \omega_1 + \omega_2, & y_2 \leqslant & y & < y_3, \\ \vdots & & \vdots & \\ \omega_1 + \omega_2 + \cdots + \omega_{J-1}, & y_{J-1} \leqslant & y & < y_J, \\ 1, & y_J \leqslant & y, & \end{cases}$$

where $y_1 \leqslant y_2 \leqslant \cdots \leqslant y_J$.

27.2• Essence of Monte Carlo integrals

— Let $Y \sim f_Y(y)$ and $Y^{(1)}, \ldots, Y^{(J)} \overset{\text{iid}}{\sim} f_Y(y)$, then we can use the sample mean to estimate the population mean:

$$E(Y) = \int y \cdot f_Y(y) \mathrm{d}x \approx \frac{1}{J} \sum_{j=1}^{J} Y^{(j)} \quad \text{as } J \to \infty.$$

— Let $X \sim g(x)$ and $X^{(1)}, \ldots, X^{(J)} \overset{\text{iid}}{\sim} g(x)$, then

$$E\{\mathrm{w}(X)\} = \int \mathrm{w}(x) \cdot g(x) \, \mathrm{d}x \approx \frac{1}{J} \sum_{j=1}^{J} \mathrm{w}(X^{(j)}) \quad \text{as } J \to \infty. \tag{1.19}$$

28• Verification of the SIR method

- To verify that the SIR method generates samples having approximate pdf $f(x)$, let $X^* \sim \text{FDiscrete}_J(\{X^{(j)}\}, \{\omega_j\})$, then the cdf of X^* is

$$\begin{aligned} \Pr(X^* \leqslant x^*) &\overset{(1.18)}{=} \sum_{j=1}^{J} \omega_j I(X^{(j)} \leqslant x^*) \\ &\overset{(1.16)}{=} \frac{(1/J) \sum_{j=1}^{J} \mathrm{w}(X^{(j)}) \cdot I(X^{(j)} \leqslant x^*)}{(1/J) \sum_{j=1}^{J} \mathrm{w}(X^{(j)})} \\ &\overset{(1.19)}{\approx} \frac{\int \mathrm{w}(x) I(x \leqslant x^*) g(x) \, \mathrm{d}x}{\int \mathrm{w}(x) g(x) \, \mathrm{d}x} \qquad \left[\because \mathrm{w}(x) = \frac{f(x)}{g(x)}\right] \\ &= \frac{\int_{-\infty}^{x^*} f(x) \, \mathrm{d}x}{1} = \int_{-\infty}^{x^*} f(x) \, \mathrm{d}x \quad \text{as } J \to \infty, \end{aligned}$$

which is the cdf with density $f(x)$.

29• Comparison of the rejection method and the SIR method

- The former needs to find an *envelope* density $g(x)$, while the latter needs to find a *proposal* density $g(x)$. They commonly require that
  — $g(x)$ has the same support as $f(x)$;
  — it is easy to draw a sample from $g(x)$.
- However, for the former, it requires that $g(x)$ is *more dispersive* than $f(x)$; while for the latter, it does not have such requirement.
- The former produces i.i.d. samples *exactly* from the target density $f(x)$; while the latter only produces i.i.d. samples *approximately* from $f(x)$.
- The former needs to find an optimal envelope constant c_{opt} through optimization.

1.5 The Stochastic Representation (SR) Method

1.5.1 The '$\overset{\text{d}}{=}$' operator

30• Definition of one-to-many SR

- Let X and $\{Y_j\}_{j=1}^n$ be r.v.'s and $g(\cdot)$ a function.
- If X and $g(Y_1, \ldots, Y_n)$ have the same distribution, denoted by

$$X \overset{\text{d}}{=} g(Y_1, \ldots, Y_n), \tag{1.20}$$

  we say (1.20) is a one-to-many *stochastic representation* (SR) of X.

30.1• Background and usefulness of SR

— If it is very difficult to generate a r.v. X, but we have $X = g(Y_1, \ldots, Y_n)$ with some known $g(\cdot)$ and it is relatively easy to generate $\{Y_j\}_{j=1}^n$, then we can first generate $\{Y_j\}_{j=1}^n$ and second set $X = g(Y_1, \ldots, Y_n)$.

— In other words, the one-to-many SR (1.20) implies that there is a simple approach to generate the r.v. X provided that the generations of $\{Y_j\}_{j=1}^n$ are relatively easy.

30.2• Basic properties of the operator '$\stackrel{\mathrm{d}}{=}$'

— Let $U \sim U(0,1)$, then $U \stackrel{\mathrm{d}}{=} 1-U$.

— Let X and Y be two r.v.'s, then $X|Y \stackrel{\mathrm{d}}{=} X$ iff X and Y are independent.

— The fact $X = Y$ implies $X \stackrel{\mathrm{d}}{=} Y$, but the inverse is not true. For instance, if $X \sim N(0,1)$, we have $X \stackrel{\mathrm{d}}{=} -X$ but $X \neq -X$ in general.

— If the pdf of X is symmetric about y-axis, then we say the cdf of X is symmetric about the point $(0, 1/2)$; i.e., $F(x) = 1 - F(-x)$ for every $x \in \mathcal{X}$, thus $X \stackrel{\mathrm{d}}{=} -X$.

30.3• Remarks

— The inversion method introduced in Section 1.1 can be viewed as a special case of the SR methods. For instance, Example 1.1 shows that if $U \sim U(0,1)$, then $X \stackrel{\mathrm{d}}{=} -\log(U)/\beta \sim \text{Exponential}(\beta)$.

— Sometimes, the SR methods are also known as *transformation* methods.

31• Several examples of SR

- Let $\{Y_j\} \overset{\text{iid}}{\sim} \text{Exponential}(\beta)$, then $X = \sum_{j=1}^{n} Y_j \sim \text{Gamma}(n, \beta)$.
- Let $\{Y_j\} \overset{\text{iid}}{\sim} N(0,1)$, then $X = \sum_{j=1}^{n} Y_j^2 \sim \chi^2(n)$.
- Let $Y_1 \sim N(0,1)$, $Y_2 \sim \chi^2(\nu)$ and $Y_1 \perp\!\!\!\perp Y_2$, then $X = Y_1/\sqrt{Y_2/\nu} \sim t(\nu)$, where $\nu\,(>0)$ is a real number.
- Let $Y_i \sim \chi^2(\nu_i)$ for $i = 1, 2$ and $Y_1 \perp\!\!\!\perp Y_2$, then $X = (Y_1/\nu_1)/(Y_2/\nu_2) \sim F(\nu_1, \nu_2)$, where $\nu_1\,(>0)$ and $\nu_2\,(>0)$ are two real numbers.
- Let $Y_1 \sim \text{Gamma}(\alpha_1, \beta)$, $Y_2 \sim \text{Gamma}(\alpha_2, \beta)$ and $Y_1 \perp\!\!\!\perp Y_2$, then $X = Y_1/(Y_1 + Y_2) \sim \text{Beta}(\alpha_1, \alpha_2)$.
- Let $Y \sim N(\mu, \sigma^2)$, then $X = \exp(Y) \sim \text{Lognormal}(\mu, \sigma^2)$.

1.5.2 Many-to-one SR for univariate case

32• Background

- Let X have a difficult-sampling density $f_X(x)$, but Y has an easy-sampling density $f_Y(y)$, and we have $G(X) \overset{\mathrm{d}}{=} Y$.
- When G is a one-to-one map, then $X \overset{\mathrm{d}}{=} G^{-1}(Y)$.
  - The pdf of $Y = G(X)$ is

$$f_Y(y) = f_X(x) \cdot \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(G^{-1}(y)) \cdot \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|, \tag{1.21}$$

    where $|a|$ denotes the absolute value of a.
  - For example, let $X \sim \text{Exponential}(\beta)$ and $G(x) \hat{=} \Pr(X \leqslant x) = 1 - \exp(-\beta x)$, we know that $G(X) \overset{\mathrm{d}}{=} Y \sim U(0,1)$; i.e., $1 - \exp(-\beta X) \overset{\mathrm{d}}{=} 1 - Y$ or $X \overset{\mathrm{d}}{=} -\log(Y)/\beta = G^{-1}(Y)$.
- However, there are important examples in which the map G is *many-to-one* so that the inverse is not uniquely determined.
  - Without loss of generality, we assume that $G(X) = Y$ has two solutions $X_i = h_i(Y)$, $i = 1, 2$. Hence, the pdf of $Y = G(X)$ is

$$f_Y(y) = \sum_{i=1}^{2} f_X(x_i) \cdot \left|\frac{\mathrm{d}x_i}{\mathrm{d}y}\right| = \sum_{i=1}^{2} f_X(h_i(y)) \cdot |\nabla h_i(y)|, \tag{1.22}$$

    where

$$\begin{aligned} \nabla h_i(y) &= \frac{\mathrm{d}x_i}{\mathrm{d}y} = \left(\frac{\mathrm{d}y}{\mathrm{d}x_i}\right)^{-1} \\ &= \left\{\left.\frac{\mathrm{d}G(x)}{\mathrm{d}x}\right|_{x=x_i}\right\}^{-1} \\ &\hat{=} \{\nabla G(x_i)\}^{-1}. \end{aligned} \tag{1.23}$$

  - For instance, let $X \sim N(0,1)$ and $Y \sim \chi^2(1)$, then $G(X) \hat{=} X^2 \overset{\mathrm{d}}{=} Y$ has two solutions $X_1 = \sqrt{Y}$ and $X_2 = -\sqrt{Y}$.

33• Rationale of the generation method

- Michael *et al.* (1976) suggested the following algorithm: Given a r.v. Y with density $f_Y(y)$, we can obtain a r.v. X with density $f_X(x)$ by choosing $X = h_1(Y)$ with probability

$$\begin{aligned}\frac{f_X(h_1(y))|\nabla h_1(y)|}{f_Y(y)} &\overset{(1.22)}{=} \frac{f_X(h_1(y))|\nabla h_1(y)|}{f_X(h_1(y))|\nabla h_1(y)| + f_X(h_2(y))|\nabla h_2(y)|} \\ &= \left\{1 + \frac{f_X(x_2)}{f_X(x_1)}\left|\frac{\nabla h_2(y)}{\nabla h_1(y)}\right|\right\}^{-1} \\ &\overset{(1.23)}{=} \left\{1 + \frac{f_X(x_2)}{f_X(x_1)}\left|\frac{\nabla G(x_1)}{\nabla G(x_2)}\right|\right\}^{-1}\end{aligned}$$

  and choosing $X = h_2(Y)$ otherwise.

33.1• Many-to-one transformation method

— Step 1: Draw $U = u \sim U(0,1)$ and independently draw $Y = y \sim f_Y(\cdot)$;

— Step 2: Set $x_1 = h_1(y)$ and $x_2 = h_2(y)$;

— Step 3: If $u \leqslant \left\{1 + \frac{f_X(x_2)}{f_X(x_1)}\left|\frac{\nabla G(x_1)}{\nabla G(x_2)}\right|\right\}^{-1}$, return $X = x_1$; Otherwise, return $X = x_2$.

Example 1.15 (Inverse Gaussian distribution). Use the following result

$$G(X) \doteq \frac{\lambda(X-\mu)^2}{\mu^2 X} \sim \chi^2(1)$$

(Shuster 1968) to generate random samples of $X \sim \text{IGaussian}(\mu, \lambda)$ with pdf

$$f_X(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\left\{-\frac{\lambda(x-\mu)^2}{\mu^2 x}\right\} = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\{-G(x)\}, \quad x > 0,$$

where $\mu > 0$ and $\lambda > 0$, see (A.29) in Appendix A.2.5.

Solution: Let $Y \sim \chi^2(1)$, then $G(x) = y$ has two solutions

$$x_1 = \mu + \frac{\mu^2 y}{2\lambda} - \frac{\mu}{2\lambda}\sqrt{4\mu\lambda y + \mu^2 y^2} \quad \text{and} \quad x_2 = \frac{\mu^2}{x_1}.$$

From $G(x_1) = G(x_2) = y$, we have

$$\frac{f_X(x_2)}{f_X(x_1)} = \left(\frac{x_1}{x_2}\right)^{3/2} = \left(\frac{x_1}{\mu}\right)^3.$$

On the other hand,

$$\frac{\mathrm{d}G(x)}{\mathrm{d}x} = \frac{\lambda}{\mu^2} \cdot \frac{x^2 - \mu^2}{x^2}$$

so that

$$\frac{\nabla G(x_1)}{\nabla G(x_2)} = -\left(\frac{\mu}{x_1}\right)^2.$$

Thus, x_1 is chosen with probability $\mu/(\mu + x_1)$. The corresponding R codes for generating X are given in **32.1•** of Appendix A.2.5. ‖

1.5.3 SR for multivariate case

34• Background

- Suppose that we have the following mapping

$$X_i = g_i(Y_1, \ldots, Y_n), \quad i = 1, \ldots, d$$

for a set of known functions $\{g_i\}_{i=1}^d$ and it is easy to generate $\{Y_j\}_{j=1}^n$, then we can first generate $\{Y_j\}_{j=1}^n$ and then set $X_i = g_i(Y_1, \ldots, Y_n)$ for $i = 1, \ldots, d$.

34.1• Two examples

— Let $Y_i \sim \text{Gamma}(a_i, 1)$ for $i = 1, \ldots, d$ and $Y_1, \ldots, Y_d$ are independent, if we define

$$X_i = \frac{Y_i}{Y_1 + \cdots + Y_d}, \quad i = 1, \ldots, d,$$

then $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim \text{Dirichlet}(a_1, \ldots, a_d)$, for more detail, see (A.25) in Appendix A.2.1.

— Let $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{z} = (Z_1, \ldots, Z_d)^\top \sim N_d(\mathbf{0}, \boldsymbol{I}_d)$, i.e., $Z_1, \ldots, Z_d \overset{\text{iid}}{\sim} N(0, 1)$, then we have

$$\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \overset{\text{d}}{=} \mathbf{z} \quad \text{or} \quad \mathbf{x} \overset{\text{d}}{=} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z},$$

where

$$\boldsymbol{\Sigma}^k = \boldsymbol{\Gamma} \begin{pmatrix} \lambda_1^k & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_d^k \end{pmatrix} \boldsymbol{\Gamma}^\top, \quad k = 1, -\frac{1}{2}, \frac{1}{2}$$

are positive (semi)definite matrices, $\boldsymbol{\Gamma}$ is an orthogonal matrix (i.e., $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \boldsymbol{I}_d$) with columns being the eigenvectors of $\boldsymbol{\Sigma}$, and $\lambda_1 \geqslant \cdots \geqslant \lambda_d \geqslant 0$ are the corresponding eigenvalues of $\boldsymbol{\Sigma}$.

Example 1.16 (Uniform distribution on the d-dimensional hyperplane). Let $Z_i \overset{\text{ind}}{\sim} \text{Exponential}(\theta b_i)$ for $i = 1, \ldots, d$, where $\theta > 0$ and $b_i > 0$. Define

$$\begin{aligned} X &= \sum_{i=1}^{d} b_i Z_i, \\ \mathbf{y} &= (Y_1, \ldots, Y_d)^\top = \left(\frac{b_1 Z_1}{X}, \ldots, \frac{b_d Z_d}{X} \right)^\top, \end{aligned}$$

find the joint distribution of $(Y_1, \ldots, Y_{d-1}, X)^\top$, the joint distribution of $\mathbf{y}_{-d} \hat{=} (Y_1, \ldots, Y_{d-1})^\top$ and the distribution of X.

Solution: (i) We have the following transformation:

$$\begin{aligned} z_i &= \frac{y_i x}{b_i}, \quad i = 1, \ldots, d-1, \\ z_d &= \frac{y_d x}{b_d} = \frac{(1 - \sum_{i=1}^{d-1} y_i)x}{b_d}. \end{aligned}$$

The Jacobian determinant is

$$\begin{aligned} J(\boldsymbol{z} \to \boldsymbol{y}_{-d}, x) &= \frac{\partial(z_1, z_2, \ldots, z_{d-1}, z_d)}{\partial(y_1, y_2, \ldots, y_{d-1}, x)} \\ &= \det \begin{pmatrix} \frac{x}{b_1} & 0 & \cdots & 0 & \frac{y_1}{b_1} \\ 0 & \frac{x}{b_2} & \cdots & 0 & \frac{y_2}{b_2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{x}{b_{d-1}} & \frac{y_{d-1}}{b_{d-1}} \\ -\frac{x}{b_d} & -\frac{x}{b_d} & \cdots & -\frac{x}{b_d} & \frac{y_d}{b_d} \end{pmatrix} \end{aligned}$$

$$= \det \begin{pmatrix} \frac{x}{b_1} & 0 & \cdots & 0 & \frac{y_1}{b_1} \\ 0 & \frac{x}{b_2} & \cdots & 0 & \frac{y_2}{b_2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{x}{b_{d-1}} & \frac{y_{d-1}}{b_{d-1}} \\ 0 & 0 & \cdots & 0 & \frac{\sum_{i=1}^d y_i}{b_d} \end{pmatrix} = \frac{x^{d-1}}{\prod_{i=1}^d b_i}.$$

Therefore, we obtain

$$\begin{aligned} f_{(\mathbf{y}_{-d},X)}(\boldsymbol{y}_{-d},x) &= |J(\boldsymbol{z} \to \boldsymbol{y}_{-d},x)| \prod_{i=1}^d f_{Z_i}(z_i) \\ &= \frac{x^{d-1}}{\prod_{i=1}^d b_i} \cdot \prod_{i=1}^d \theta b_i \exp(-\theta b_i z_i) \\ &= \Gamma(d) \cdot \frac{\theta^d}{\Gamma(d)} x^{d-1} \mathrm{e}^{-\theta x} = f_{\mathbf{y}_{-d}}(\boldsymbol{y}_{-d}) \cdot f_X(x). \end{aligned}$$

(ii) The joint distribution of $\mathbf{y}_{-d}$ is $f_{\mathbf{y}_{-d}}(\boldsymbol{y}_{-d}) = (d-1)! \cdot I(\boldsymbol{y}_{-d} \in \mathbb{V}_{d-1})$, where

$$\mathbb{V}_{d-1} = \left\{ (y_1, \ldots, y_{d-1})^\top : \; y_i > 0, \; i = 1, \ldots, d-1, \; \sum_{i=1}^{d-1} y_i \leqslant 1 \right\}. \quad (1.24)$$

In other words, $\mathbf{y}_{-d} \sim U(\mathbb{V}_{d-1})$ and the volume of $\mathbb{V}_{d-1}$ is $v(\mathbb{V}_{d-1}) = 1/(d-1)!$. Equivalently, we can write $\mathbf{y} \sim U(\mathbb{T}_d)$, where

$$\mathbb{T}_d = \left\{ (y_1, \ldots, y_d)^\top : \; y_i > 0, \; i = 1, \ldots, d, \; \sum_{i=1}^{d} y_i = 1 \right\} \quad (1.25)$$

is the d-dimensional hyperplane and $v(\mathbb{T}_d) = \sqrt{d}/(d-1)!$.

(iii) The distribution of X is Gamma(d, θ).

Comment 7: Alternatively, since $Z_i \overset{\text{ind}}{\sim} \text{Exponential}(\theta b_i) = \text{Gamma}(1, \theta b_i)$, we have $b_i Z_i \overset{\text{iid}}{\sim} \text{Gamma}(1, \theta)$ so that $X = \sum_{i=1}^d b_i Z_i \sim \text{Gamma}(d, \theta)$. ‖

Example 1.17 (Uniform distribution in d-dimensional ball). Let random vector $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim U(\mathbb{B}_d(r))$, where

$$\mathbb{B}_d(r) = \{(x_1, \ldots, x_d)^\top \colon x_1^2 + \cdots + x_d^2 \leqslant r^2\} \tag{1.26}$$

is the d-dimensional ball with radius r. Find the SR of $\mathbf{x}$.

Solution: It can be verified that $\mathbf{x}$ has the following SR (see Exercise 1.2)

$$\begin{aligned}
X_1 &\overset{\mathrm{d}}{=} rY_1 \cos(\pi Y_2), \\
X_2 &\overset{\mathrm{d}}{=} rY_1 \sin(\pi Y_2) \cos(\pi Y_3), \\
&\vdots \\
X_{d-2} &\overset{\mathrm{d}}{=} rY_1 \sin(\pi Y_2) \cdots \sin(\pi Y_{d-2}) \cos(\pi Y_{d-1}), \\
X_{d-1} &\overset{\mathrm{d}}{=} rY_1 \sin(\pi Y_2) \cdots \sin(\pi Y_{d-1}) \cos(2\pi Y_d), \\
X_d &\overset{\mathrm{d}}{=} rY_1 \sin(\pi Y_2) \cdots \sin(\pi Y_{d-1}) \sin(2\pi Y_d),
\end{aligned}$$

where $Y_1, \ldots, Y_d$ are mutually independent, and $Y_i \in (0, 1)$ has pdf

$$f_{Y_i}(y) = \begin{cases} dy^{d-1}, & \text{when } i = 1, \\ \dfrac{\pi \sin^{d-i}(\pi y)}{B(\frac{1}{2}, \frac{d-i+1}{2})}, & \text{when } i = 2, \ldots, d-1, \\ 1, & \text{when } i = d. \end{cases}$$

In Example 1.14, we apply the SIR method to draw samples from $f_{Y_i}(y)$. ‖

1.5.4 Mixture representation

35• Augmentation of random variables

- Sometimes, it is quite difficult if we directly generate a random variable (or vector) $X \sim f_X(x)$, but the augmented vector $(X, Y)^\top \sim f_{(X,Y)}(x, y)$ is relatively easy to generate.
- In such cases, we may first generate the augmented vector and then pick up the desired components (Bulter 1958).
- Statistically, we can represent $f_X(x)$ as the marginal distribution of $f_{(X,Y)}(x, y)$ in the form

$$f_X(x) = \int_{\mathbb{Y}} f_{(X,Y)}(x, y)\, \mathrm{d}y. \tag{1.27}$$

- Alternatively, we can rewrite (1.27) in the mixture form

$$f_X(x) = \int_{\mathbb{Y}} f_Y(y) f_{(X|Y)}(x|y)\,\mathrm{d}y \quad \text{or} \quad f_X(x) = \sum_{k\in\mathbb{Y}} p_k f_k(x),$$

  depending on if Y is continuous or discrete.

Example 1.18 (Standard normal distribution). Let $X \sim N(0,1)$. Since the inverse of cdf for X does not have an explicit expression, the inversion method is not a good strategy to simulate X. A well-known approach for generating X is (Box & Muller 1958):

$$X \overset{\mathrm{d}}{=} \sqrt{-2\log U_1}\cos(2\pi U_2) \quad \text{and} \quad Y \overset{\mathrm{d}}{=} \sqrt{-2\log U_1}\sin(2\pi U_2), \tag{1.28}$$

where $U_1, U_2 \overset{\text{iid}}{\sim} U(0,1)$, and $X, Y \overset{\text{iid}}{\sim} N(0,1)$.

Proof: Given $U_1, U_2 \overset{\text{iid}}{\sim} U(0,1)$ and the SR (1.28), we only need to show that $X, Y \overset{\text{iid}}{\sim} N(0,1)$. The Jacobian determinant is

$$\begin{aligned} J(x, y \to u_1, u_2) &= \frac{\partial(x,y)}{\partial(u_1,u_2)} \\ &= \det\begin{pmatrix} \dfrac{-\cos(2\pi u_2)}{u_1\sqrt{-2\log u_1}} & \dfrac{-\sin(2\pi u_2)}{u_1\sqrt{-2\log u_1}} \\ -2\pi\sqrt{-2\log u_1}\sin(2\pi u_2) & 2\pi\sqrt{-2\log u_1}\cos(2\pi u_2) \end{pmatrix} \\ &= -\frac{2\pi}{u_1} \end{aligned}$$

so that the joint pdf of $(X,Y)^\top$ is given by

$$\begin{aligned} f_{(X,Y)}(x,y) &= f_{U_1}(u_1)\cdot f_{U_2}(u_2)\cdot\frac{1}{|J(x,y\to u_1,u_2)|} \\ &= \frac{u_1}{2\pi} = \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}\cdot\frac{\mathrm{e}^{-y^2/2}}{\sqrt{2\pi}} = \phi(x)\cdot\phi(y), \end{aligned}$$

implying that $X, Y \overset{\text{iid}}{\sim} N(0,1)$. □

Generation method: Essentially, this approach is first to augment X with an independent standard Gaussia r.v. Y, and then to generate the joint distribution $(X,Y)^\top \sim N_2(\mathbf{0}, \boldsymbol{I}_2)$ via the SR (1.28).

Motivation: Let $X, Y \overset{\text{iid}}{\sim} N(0,1)$, by making the transformation $X = R\cos(\theta)$ and $Y = R\sin(\theta)$, we know that the joint density of R and θ is

$$\begin{aligned} f_{(R,\theta)}(r,\theta) &= f_X(x) \cdot f_Y(y) \left| \frac{\partial(x,y)}{\partial(r,\theta)} \right| \\ &= \frac{\mathrm{e}^{-r^2/2}}{2\pi} \cdot \left| \det \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix} \right| \\ &= r\,\mathrm{e}^{-r^2/2} \cdot \frac{1}{2\pi} I(0 < \theta < 2\pi), \end{aligned}$$

implying that

$$R \sim f_R(r) = r\,\mathrm{e}^{-r^2/2}, \quad \theta \sim U(0, 2\pi),$$

and $R \perp\!\!\!\perp \theta$. Let $Z = R^2$, then

$$f_Z(z) = f_R(r) \left| \frac{\mathrm{d}r}{\mathrm{d}z} \right| = 0.5\,\mathrm{e}^{-0.5z},$$

i.e., $Z \sim$ Exponential(0.5). From Example 1.1, we have $R^2 = Z \overset{\mathrm{d}}{=} -2\log U_1$ with $U_1 \sim U(0,1)$. On the other hand, $\theta \overset{\mathrm{d}}{=} 2\pi U_2$ with $U_2 \sim U(0,1)$. ‖

36• From mixture representation to SR

- Suppose that we want to generate a positive r.v. X with density

$$f_X(x) = \int_x^\infty f_Y(y) y^{-1}\,\mathrm{d}y, \quad x > 0. \tag{1.29}$$

- From (1.29), the conditional density of $X|(Y=y)$ is $f_{(X|Y)}(x|y) = y^{-1}$, where $0 < x < y$.
- The mixture specified by (1.29) can be represented equivalently by

$$Y \sim f_Y(y),\ y > 0 \quad \text{and} \quad X|(Y=y) \sim U(0,y).$$

- Hence, $\frac{X}{Y}|(Y=y) \sim U(0,1)$, not depending on y, so that $\frac{X}{Y} \overset{\mathrm{d}}{=} U \sim U(0,1)$. Therefore,

$$X \overset{\mathrm{d}}{=} UY \quad \text{and} \quad U \perp\!\!\!\perp Y. \tag{1.30}$$

- The SR (1.30) provides a simple way to draw X. For instance, if $Y \sim \text{Gamma}(a, 1)$, then X follows *gamma–integral* distribution (Devroye 1986, p.191).

36.1• Khintchine theorem

— If $\nabla f_X(\cdot)$ exists and $f_X(\infty) \to 0$, then we have identity

$$f_X(x) = -\int_x^\infty \nabla f_X(y)\,\mathrm{d}y. \tag{1.31}$$

— By comparing (1.29) with (1.31), we obtain

$$f_Y(y) = -y\nabla f_X(y), \tag{1.32}$$

which is the well-known Khintchine's (1938) theorem.

— For example, if $X \sim \text{Exponential}(\beta)$, from (1.32), then Y must be $\text{Gamma}(2, \beta)$.

Example 1.19 (Multivariate t distribution). Let $\tau \sim \text{Gamma}(\nu/2, \nu/2)$ and $\mathbf{x}|\tau \sim N_d(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Sigma})$, then $\mathbf{x} \sim t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Find an SR of $\mathbf{x}$.

Solution: From $\mathbf{x}|\tau \sim N_d(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Sigma})$, we have

$$\sqrt{\tau}(\mathbf{x} - \boldsymbol{\mu})|\tau \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}),$$

not depending on τ, so that $\sqrt{\tau}(\mathbf{x}-\boldsymbol{\mu}) \overset{\mathrm{d}}{=} \mathbf{y} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$. Therefore, we have the following SR:

$$\mathbf{x} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \frac{\mathbf{y}}{\sqrt{\tau}} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \frac{N_d(\mathbf{0}, \boldsymbol{\Sigma})}{\sqrt{\chi^2(\nu)/\nu}},$$

where $\mathbf{y}$ and τ are mutually independent. ‖

1.6 The Conditional Sampling Method

37• Background and idea

- The *conditional sampling method* due to the prominent Rosenblatt transformation is particularly available when the joint distribution of a d-vector is very difficult to generate but one marginal distribution and $d-1$ univariate conditional distributions are easy to simulate.

- Let $\mathbf{x} = (X_1, \ldots, X_d)^\top$ and its density $f_{\mathbf{x}}(\boldsymbol{x})$ can be factorized as

$$f_{\mathbf{x}}(\boldsymbol{x}) = f_1(x_1) \prod_{i=2}^{d} f_i(x_i | x_1, x_2, \ldots, x_{i-1}). \tag{1.33}$$

37.1• Remarks

— The beauty of the conditional sampling method is that it reduces the problem of generating a d-dimensional random vector into the problem of generating d random variables.

— Note that the decomposition of (1.33) is not unique. In fact, there are $d!$ different representations. A better representation will result in a more efficient sampling scheme.

37.2• The conditional sampling method

— Step 1: Draw X_1 from $f_1(x_1)$;

— Step 2: Draw X_2 from $f_2(x_2|x_1)$;

— Step 3: Draw X_3 from $f_3(x_3|x_1, x_2)$;

$\vdots$

— Step d: Draw X_d from $f_d(x_d|x_1, x_2, \ldots, x_{d-1})$.

Example 1.20 (Trinomial distribution). Consider the trinomial distribution and let $\mathbf{x} = (X_1, X_2, X_3)^\top \sim \text{Multinomial}_3(n; p_1, p_2, p_3)$ with pmf

$$\Pr(\mathbf{x} = \boldsymbol{x}) = \binom{n}{x_1, x_2, x_3} \prod_{i=1}^{3} p_i^{x_i},$$

where $x_i \geqslant 0$, $\sum_{i=1}^{3} x_i = n$, $p_i > 0$ and $\sum_{i=1}^{3} p_i = 1$. Find the marginal distribution of X_1 and the conditional distribution $X_2|(X_1 = x_1)$. State the conditional sampling algorithm for generating one random sample for $\mathbf{x}$.

Solution: (i) The joint pmf of $(X_1, X_2)^\top$ can be rewritten as

$$\Pr(X_1 = x_1, X_2 = x_2) = \binom{n}{x_1, x_2, n - x_1 - x_2} p_1^{x_1} p_2^{x_2} p_3^{n - x_1 - x_2},$$

where $x_1 \geqslant 0, x_2 \geqslant 0$, $x_1 + x_2 \leqslant n$; $p_1 > 0, p_2 > 0$, $p_3 = 1 - p_1 - p_2$. The marginal distribution of X_1 is

$$\begin{aligned}
\Pr(X_1 = x_1) &= \sum_{x_2=0}^{n-x_1} \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} p_3^{n-x_1-x_2} \\
&= \frac{n! p_1^{x_1}}{x_1!(n-x_1)!} \sum_{x_2=0}^{n-x_1} \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} p_2^{x_2} p_3^{n-x_1-x_2} \\
&= \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} (1-p_1)^{n-x_1},
\end{aligned}$$

indicating $X_1 \sim \text{Binomial}(n, p_1)$. On the other hand,

$$\begin{aligned}
&\Pr(X_2 = x_2 | X_1 = x_1) \\
= &\frac{\Pr(X_1 = x_1, X_2 = x_2)}{\Pr(X_1 = x_1)} \\
= &\frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} \left(\frac{p_2}{1-p_1}\right)^{x_2} \left(1 - \frac{p_2}{1-p_1}\right)^{n-x_1-x_2},
\end{aligned}$$

i.e., $X_2 | (X_1 = x_1) \sim \text{Binomial}(n - x_1,\ p_2/(1-p_1))$.

(ii) The algorithm is as follows:

Step 1: Draw $X_1 = x_1 \sim \text{Binomial}(n, p_1)$;

Step 2: Draw $X_2 = x_2 | (X_1 = x_1) \sim \text{Binomial}(n - x_1,\ p_2/(1-p_1))$;

Step 3: Set $X_3 = n - x_1 - x_2$. ‖

Example 1.21 (Two-dimensional exponential distribution). Let

$$f_{(X,Y)}(x, y) = c^{-1} x \exp(-xy), \quad 0 < x \leqslant 3,\ y \geqslant 0,$$

where c is the normalizing constant. Generate independent samples from this joint density by using the conditional sampling method.

Solution: (i) The conditional density of $Y | (X = x)$ is

$$\begin{aligned}
f_{(X|Y)}(y|x) &\propto f_{(X,Y)}(x, y) \\
&\propto \exp(-xy) = x \exp(-xy), \quad y \geqslant 0.
\end{aligned}$$

That is, $Y|(X=x) \sim \text{Exponential}(x)$. The marginal density of X is

$$f_X(x) = \frac{f_{(X,Y)}(x,y)}{f_{(X|Y)}(y|x)} = \frac{c^{-1}x\exp(-xy)}{x\exp(-xy)} = c^{-1}, \quad 0 < x \leqslant 3,$$

implying that $X \sim U(0,3]$ with $c = 3$.

(ii) The algorithm is as follows:

Step 1: Draw $X = x$ from $U(0,3]$;

Step 2: Given $X = x$, draw Y from Exponential(x). ‖

Example 1.22 (Dirichlet distribution). Let a bivariate random vector $\mathbf{x} = (X_1, X_2)^\top \sim \text{Dirichlet}(a_1, a_2; a_3)$ with $a_1 = 2$ and $a_2 = a_3 = 1$, its pdf is

$$\begin{aligned} & f_{(X,Y)}(x_1,x_2) \\ = & \begin{cases} \dfrac{\Gamma(a_1+a_2+a_3)}{\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)} x_1^{a_1-1}x_2^{a_2-1}(1-x_1-x_2)^{a_3-1}, & \text{if } (x_1,x_2)^\top \in \mathbb{V}_2, \\ 0, & \text{otherwise} \end{cases} \\ = & \begin{cases} 6x_1, & \text{if } x_1, x_2 > 0,\ x_1 + x_2 \leqslant 1, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

where $\Gamma(1) = 1$ and $\Gamma(n) = (n-1)\Gamma(n-1)$. (i) Find $f_{X_2}(x_2)$ and the conditional distribution of $X_1|(X_2 = x_2)$. (ii) Find $f_{X_1}(x_1)$ and the conditional distribution of $X_2|(X_1 = x_1)$. State the two conditional sampling algorithms for generating one random sample from $\mathbf{x}$ and show the difference.

Solution: (i) We first factorize $f_{(X,Y)}(x_1,x_2)$ into $f_{X_2}(x_2)f_{X_1|X_2}(x_1|x_2)$. The marginal pdf and cdf of X_2 are respectively

$$\begin{aligned} f_{X_2}(x_2) &= \int_0^{1-x_2} f_{(X,Y)}(x_1,x_2)\,\mathrm{d}x_1 = 3x_1^2\Big|_0^{1-x_2} = 3(1-x_2)^2 \\ &= \frac{\Gamma(1+3)}{\Gamma(1)\Gamma(3)} x_2^{1-1}(1-x_2)^{3-1}, \quad 0 < x_2 < 1, \quad \text{and} \\ F_2(x_2) &= \int_0^{x_2} f_{X_2}(u)\,\mathrm{d}u = \int_0^{x_2} 3(1-u)^2\,\mathrm{d}u \\ &= -(1-u)^3\Big|_0^{x_2} = 1-(1-x_2)^3, \quad 0 < x_2 < 1, \end{aligned}$$

indicating that $X_2 \sim \text{Beta}(1,3)$. The conditional pdf and cdf of $X_1|(X_2 = x_2)$ are respectively

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &= \frac{f_{(X,Y)}(x_1,x_2)}{f_{X_2}(x_2)} \\ &= \frac{6x_1}{3(1-x_2)^2}, \quad 0 < x_1 \leqslant 1-x_2, \quad \text{and} \\ F_1(x_1|x_2) &= \int_0^{x_1} f_{X_1|X_2}(u|x_2)\,\mathrm{d}u = \frac{1}{3(1-x_2)^2} \cdot \int_0^{x_1} 6u\,\mathrm{d}u \\ &= \frac{1}{3(1-x_2)^2} \cdot 3u^2\Big|_0^{x_1} = \frac{x_1^2}{(1-x_2)^2}, \quad 0 < x_1 \leqslant 1-x_2. \end{aligned}$$

By the inversion method, i.e., let $F_2(X_2) = U_2 \overset{\mathrm{d}}{=} 1-U_2$ and $F_1(X_1|X_2) = U_1$, we obtain

$$X_2 \overset{\mathrm{d}}{=} 1 - U_2^{1/3} \quad \text{and} \quad X_1 \overset{\mathrm{d}}{=} U_1^{1/2} U_2^{1/3}, \tag{1.34}$$

where $U_1, U_2 \overset{\text{iid}}{\sim} U(0,1)$.

(ii) We factorize $f_{(X,Y)}(x_1,x_2)$ into $f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1)$. Similarly, $X_1 \sim \text{Beta}(2,2)$. The cdf of X_1 and the conditional cdf of $X_2|X_1$ are given by

$$\begin{aligned} F_1(x_1) &= 3x_1^2 - 2x_1^3, \qquad 0 < x_1 < 1, \\ F_2(x_2|x_1) &= (1-x_1)^{-1} x_2, \quad 0 < x_2 \leqslant 1-x_1, \end{aligned}$$

respectively. Hence

$$3X_1^2 - 2X_1^3 \overset{\mathrm{d}}{=} U_1 \quad \text{and} \quad (1-X_1)^{-1} X_2 \overset{\mathrm{d}}{=} U_2, \tag{1.35}$$

where $U_1, U_2 \overset{\text{iid}}{\sim} U(0,1)$. By comparing (1.34) with (1.35), we know that (1.34) is more convenient than (1.35) for generating $\mathbf{x} = (X_1, X_2)^\top$. ‖

Example 1.23 (Truncated bivariate normal distribution). Let $(X,Y)^\top \sim \text{TN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{a}, \boldsymbol{b})$. Generate independent samples from this joint distribution by using the conditional sampling method.

Solution: The marginal density of X is $f(x) = g(x)/c^*$, where c^* is the unknown normalizing constant and $g(x)$ is given by (1.1). The conditional distribution of $Y|(X=x)$ follows truncated univariate normal distribution

$$\text{TN}(\mu_2 + \rho\sigma_2\sigma_1^{-1}(x-\mu_1),\ \sigma_2^2(1-\rho^2);\ a_2, b_2).$$

Thus, we first use the grid method presented in Example 1.8 to simulate X, and then use the rejection method described in Example 1.13 to simulate Y for given $X = x$. ‖

Exercise 1

1.1 (Truncated multivariate normal distribution). An n-dimensional r.v. $\mathbf{w} = (W_1, \ldots, W_n)^\top$ is said to follow a multivariate normal distribution truncated to the rectangle (or Cartesian product) $[\boldsymbol{a}, \boldsymbol{b}] = \prod_{i=1}^n [a_i, b_i]$ if its joint pdf is proportional to (Robert 1995)

$$\exp\left\{-(\boldsymbol{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})/2\right\} \cdot I(\boldsymbol{a} \leqslant \boldsymbol{w} \leqslant \boldsymbol{b}),$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is the location parameter vector, $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite matrix, $I(\cdot)$ is the indicator function, and $\boldsymbol{a} \leqslant \boldsymbol{w} \leqslant \boldsymbol{b}$ means that $a_i \leqslant w_i \leqslant b_i$ for all $i = 1, \ldots, n$. We write $\mathbf{w} \sim \mathrm{TN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{a}, \boldsymbol{b})$ or $\mathbf{w} \sim \mathrm{TN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{a}, \boldsymbol{b}])$.

(Truncated bi-normal). Let $(X, Y)^\top \sim \mathrm{TN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{a}, \boldsymbol{b})$, where

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}, \quad \boldsymbol{a} = \begin{pmatrix}a_1\\ a_2\end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix}b_1\\ b_2\end{pmatrix}.$$

Verify the following facts:

(a) The conditional distributions are truncated uni-normal, i.e.,

$$\begin{aligned} X|(Y=y) &\sim \mathrm{TN}(\mu_1 + \rho\sigma_1\sigma_2^{-1}(y-\mu_2),\ \sigma_1^2(1-\rho^2);\ a_1, b_1),\\ Y|(X=x) &\sim \mathrm{TN}(\mu_2 + \rho\sigma_2\sigma_1^{-1}(x-\mu_1),\ \sigma_2^2(1-\rho^2);\ a_2, b_2). \end{aligned}$$

(b) The marginal density of X is proportional to

$$\mathrm{e}^{-(x-\mu_1)^2/(2\sigma_1^2)} \times \left\{\Phi\left(\frac{b_2-\mu_2-\rho\sigma_2\sigma_1^{-1}(x-\mu_1)}{\sigma_2\sqrt{1-\rho^2}}\right)\right.$$

$$\left. - \Phi\left(\frac{a_2-\mu_2-\rho\sigma_2\sigma_1^{-1}(x-\mu_1)}{\sigma_2\sqrt{1-\rho^2}}\right)\right\} \cdot I(a_1 \leqslant x \leqslant b_1).$$

1.2 Verify the fact on the SR of X stated in Example 1.17.

1.3 Let $\mathbb{B}_d(r)$ be defined by (1.26), prove that the volume of $\mathbb{B}_d(r)$ is $2\pi^{d/2}r^d/\{\Gamma(d/2)d\}$.

1.4 Let $\mathbb{V}_d(r)$ denote the d-dimensional ℓ_1-ball,

$$\mathbb{V}_d(r) = \{(y_1, \ldots, y_d)^\top \colon y_i > 0,\ y_1 + \cdots + y_d \leqslant r\}.$$

Show that the volume of $\mathbb{V}_d(r)$ is $r^d/d!$.

1.5 Use the inversion method to generate a r.v. from the following pdfs:

(a) Logistic density

$$f(x) = \frac{\exp(-\frac{x-\mu}{\sigma})}{\sigma\{1+\exp(-\frac{x-\mu}{\sigma})\}^2}, \quad x \in \mathbb{R},$$

where $\mu \in (-\infty, \infty)$ is the location parameter and $\sigma > 0$ is the scale parameter.

(b) Rayleigh density $f(x) = \sigma^{-2} x \exp(-\frac{x^2}{2\sigma^2})$, $x > 0$, $\sigma > 0$.

(c) Triangular density $f(x) = \frac{2}{a}(1 - \frac{x}{a})$, $0 \leqslant x < a$, $a > 0$.

(d) Pareto density $f(x) = ab^a/x^{a+1}$, $x \geqslant b > 0$, $a > 0$.

(e) Gumbel–minimum density

$$f(x) = \frac{1}{\sigma} \mathrm{e}^{\frac{x-\mu}{\sigma}} \exp(-\mathrm{e}^{\frac{x-\mu}{\sigma}}), \quad x \in \mathbb{R},$$

where $\mu \in (-\infty, \infty)$ is the location parameter and $\sigma > 0$ is the scale parameter.

(f) Gumbel–maximum density

$$f(x) = \frac{1}{\sigma} \mathrm{e}^{-\frac{x-\mu}{\sigma}} \exp(-\mathrm{e}^{-\frac{x-\mu}{\sigma}}), \quad x \in \mathbb{R},$$

where $\mu \in (-\infty, \infty)$ is the location parameter and $\sigma > 0$ is the scale parameter.

(g) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x)$, where $F(\cdot)$ is a continuous cdf. Use the inversion method to generate $X_{(1)} = \min(X_1, \ldots, X_n)$ and $X_{(n)} = \max(X_1, \ldots, X_n)$.

1.6 Let the r.v. $\mathbf{x} = (X_1, \ldots, X_d)^\top$ have density $f(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$, $U \sim U(0,1)$ and $\mathbf{x} \perp\!\!\!\perp U$.

(a) Prove that $\binom{\mathbf{x}}{cUf(\mathbf{x})}$ is uniformly distributed on

$$\mathbb{A} \mathrel{\hat{=}} \left\{ \binom{\boldsymbol{x}}{y} \colon \boldsymbol{x} \in \mathbb{R}^d,\ 0 \leqslant y \leqslant cf(\boldsymbol{x}) \right\},$$

where $c > 0$ is an arbitrary constant.

(b) If $\binom{\mathbf{z}}{W} \sim U(\mathbb{A})$, then $\mathbf{z} = (Z_1, \ldots, Z_d)^\top$ has a density $f(\boldsymbol{z})$ for $\boldsymbol{z} \in \mathbb{R}^d$.

1.7 Use the SIR method to generate a r.v. from the standard normal pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty$$

via the logistic density with known scale parameter $\theta_0\,(>0)$:

$$g(x) = \frac{\exp(-x/\theta_0)}{\theta_0\{1+\exp(-x/\theta_0)\}^2}, \quad -\infty < x < \infty.$$

1.8 Let a r.v. X have the following pdf

$$f_X(x) = \frac{a}{2\sinh(a)} \sin(x) \exp\{a\cos(x)\},$$

where $0 < x < \pi$, $a > 0$, $\sinh(a) = (\mathrm{e}^a - \mathrm{e}^{-a})/2$. Prove that

$$X \stackrel{\mathrm{d}}{=} \arccos\left[a^{-1}\log\left\{(1-U)\,\mathrm{e}^a + U\,\mathrm{e}^{-a}\right\}\right], \quad \text{where} \quad U \sim U(0,1).$$

1.9 Let $X \sim F(\cdot)$ and $Y \sim G(\cdot)$, where

$$G(y) = 0 \cdot I(y < a) + \frac{F(y) - F(a)}{F(b) - F(a)} \cdot I(a \leqslant y \leqslant b) + 1 \cdot I(y > b),$$

$-\infty \leqslant a < b \leqslant \infty$. Show that

$$Y \stackrel{\mathrm{d}}{=} F^{-1}(F(a) + U\{F(b) - F(a)\}), \quad \text{where} \quad U \sim U(0,1).$$

1.10 Use the conditional sampling method to generate the multivariate Cauchy distribution with density

$$f(\boldsymbol{x}) = \frac{\Gamma(b)}{\pi^b(1+\boldsymbol{x}^\top\boldsymbol{x})^b}, \quad \boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d.$$

where $b = (d+1)/2$. [Hint: Use the following integral identity

$$\int_{\mathbb{R}^m} h\left(\sum_{i=1}^m x_i^2\right) \mathrm{d}x_1 \cdots \mathrm{d}x_m = \frac{\pi^{m/2}}{\Gamma(m/2)} \int_0^\infty y^{m/2-1} h(y)\,\mathrm{d}y,$$

where $h(\cdot)$ is an arbitrary non-negative measurable function]

1.11 Let $Z_1, \ldots, Z_d \overset{\text{iid}}{\sim} \text{Exponential}(1)$, prove that

$$\mathbf{y} = (Y_1, \ldots, Y_d)^\top = \left(\frac{Z_1}{\sum_{i=1}^d Z_i}, \ldots, \frac{Z_d}{\sum_{i=1}^d Z_i} \right)^\top \sim U(\mathbb{T}_d),$$

where $\mathbb{T}_d$ is defined by (1.25).

1.12 (Degenerate distribution). If the pmf of a r.v. η is given by $\Pr(\eta = c) = 1$, where c is a constant, we say that η follows a degenerate distribution with all mass at the point c, denoted by $\eta \sim \text{Degenerate}(c)$.

(ZIP distribution). Let $\xi \sim \text{Degenerate}(0)$, $X \sim \text{Poisson}(\lambda)$ and $\xi \perp\!\!\!\perp X$. A discrete r.v. Y is said to follow a *zero-inflated Poisson* (ZIP) distribution if its pmf is

$$\begin{aligned} f(y|\phi, \lambda) &= \phi \Pr(\xi = y) + (1-\phi) \Pr(X = y) \\ &= \begin{cases} \phi + (1-\phi)\, \mathrm{e}^{-\lambda}, & \text{if } y = 0, \\ (1-\phi) \dfrac{\mathrm{e}^{-\lambda} \lambda^y}{y!}, & \text{if } y = 1, \ldots, \infty \end{cases} \\ &= \{\phi + (1-\phi)\, \mathrm{e}^{-\lambda}\} I(y=0) + \left\{ (1-\phi) \frac{\mathrm{e}^{-\lambda} \lambda^y}{y!} \right\} I(y > 0), \quad (1.36) \end{aligned}$$

where $\phi \in [0, 1)$ denotes the unknown proportion for incorporating extra zeros than those permitted by the original Poisson distribution. We will write $Y \sim \text{ZIP}(\phi, \lambda)$.

Let $Y \sim \text{ZIP}(\phi, \lambda)$.

(a) Prove that the ZIP r.v. Y has the following SR:

$$Y \overset{\text{d}}{=} ZX = \begin{cases} 0, & \text{with probability } \phi, \\ X, & \text{with probability } 1-\phi, \end{cases} \quad (1.37)$$

where $Z \sim \text{Bernoulli}(1-\phi)$, $X \sim \text{Poisson}(\lambda)$ and $Z \perp\!\!\!\perp X$. Thus (1.37) provides a simple way to generate $Y \sim \text{ZIP}(\phi, \lambda)$.

(b) Prove that $Y \sim \text{ZIP}(\phi, \lambda)$ can be generated alternatively by the following conditional sampling method:

$$W \sim \text{Bernoulli}(1-\phi) \quad \text{and} \quad Y|W \sim \text{Poisson}(\lambda W).$$

[Hint: Show that $\int f_W(w) \cdot f_{Y|W}(y|w)\, \mathrm{d}w$ is the same as (1.36)]

1.13 Let $X_i \sim f_{X_i}(x)$ for $i = 1, \ldots, n$.

(a) Define

$$X = \begin{cases} X_1, & \text{with probability } \phi, \\ X_2, & \text{with probability } 1 - \phi, \end{cases} \tag{1.38}$$

where $\phi \in (0, 1)$. Find an SR of X by mimicking (1.37) and find the pdf of X.

(b) Define

$$X = \begin{cases} X_1, & \text{with probability } \phi_1, \\ X_2, & \text{with probability } \phi_2, \\ \vdots & \vdots \\ X_n, & \text{with probability } \phi_n, \end{cases} \tag{1.39}$$

where $\phi_i \in (0, 1)$ and $\sum_{i=1}^n \phi_i = 1$. Find an SR of X and find the pdf of X.

1.14 Use two methods to generate a r.v. from the following pdf

$$f_X(x) = \frac{5}{12}\Big\{1 + (x-1)^4\Big\}, \quad 0 \leqslant x \leqslant 2.$$

[Hint: Use the result in Q1.13(a)]

1.15 Use the method of mixture representation (see §1.5.4) to generate a r.v. from the following pdf

$$f_X(x) = n \int_1^\infty y^{-n}\, \mathrm{e}^{-xy} \mathrm{d}y, \quad x > 0.$$

1.16 (Positively correlated bivariate beta distribution). A two-dimensional r.v. $(X_1, X_2)^\top$ is said to have a *positively correlated bivariate beta distribution*, if the joint density is (Albert & Gupta 1983, 1985)

$$\int_0^1 \frac{z^{a-1}(1-z)^{b-1}}{B(a,b)} \prod_{i=1}^2 \frac{x_i^{\gamma_i z - 1}(1-x_i)^{\gamma_i(1-z)-1}}{B(\gamma_i z, \gamma_i(1-z))}\, \mathrm{d}z.$$

Show how to use the mixture method to simulate $(X_1, X_2)^\top$. [Hint: Consider the augmented vector $(X_1, X_2, Z)^\top$ with $Z \sim \mathrm{Beta}(a, b)$ and $X_i|(Z = z) \sim \mathrm{Beta}(\gamma_i z, \gamma_i(1-z))$ for $i = 1, 2$]

Chapter 2

Optimization

1• Two motivations for chapter 2

- Computationally, finding the MLEs of parameters or the posterior mode is an optimization problem.

1.1• The core of the frequentist statistics

— Let Y_{obs} denote the observed (or observable) data, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the parameter vector of interest and $\boldsymbol{\Theta}$ the parameter space.

— A central subject of statistics is to make inference on $\boldsymbol{\theta}$ based on Y_{obs}.

— Frequentist/classical method arrives its inferential statements by combining point estimators of parameters with their standard errors.

— In the *parametric* statistics, the observed data $Y_{\text{obs}} = \{x_1, \ldots, x_n\}$ are viewed as realizations of the r.v.'s $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \boldsymbol{\theta})$.

— The joint pdf of $X_1, \ldots, X_n$ is $f(Y_{\text{obs}}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$, while the *likelihood function* is usually denoted by $L(\boldsymbol{\theta}|Y_{\text{obs}})$.

— Among all estimation approaches, the method of MLE is the most popular one, where $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$ that maximizes $L(\boldsymbol{\theta}|Y_{\text{obs}})$ or equivalently maximizes its logarithm $\log\{L(\boldsymbol{\theta}|Y_{\text{obs}})\} \triangleq \ell(\boldsymbol{\theta}|Y_{\text{obs}})$.

— In other words, we want to calculate

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|Y_{\text{obs}}).$$

1.2• The posterior mode in Bayesian statistics

— When sample sizes are *small to moderate*, a useful alternative to MLE is to utilize prior knowledge on the parameter vector $\boldsymbol{\theta}$, and thus incorporate a *prior distribution* $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ into the likelihood function; and then compute the observed posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

— The *posterior mode* is defined as an argument $\tilde{\boldsymbol{\theta}}$ that maximizes $p(\boldsymbol{\theta}|Y_{\text{obs}})$ or equivalently $\log\{p(\boldsymbol{\theta}|Y_{\text{obs}})\}$.

— In other words, we want to calculate

$$\tilde{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log\{p(\boldsymbol{\theta}|Y_{\text{obs}})\}.$$

2.1 A Review of Some Standard Concepts

2• Aims of this section

- With the advent of differential calculus, it becomes possible to solve optimization problems more systematically.
- Before discussing concrete examples, it is helpful to review some standard concepts.
- We restrict our attention to real–valued functions defined on intervals.
- Intervals in questions can be finite or infinite in extent and open or closed at either end.

2.1.1 Order relations

3• Basic assumptions for order relations

- Suppose that we have two functions $f(x)$ and $g(x)$ defined on a common interval $\mathbb{I}$.
- Let x_0 be either an internal point or a boundary point of $\mathbb{I}$ with $g(x) \neq 0$ for x close (but not equal) to x_0.

3.1• Big O

— The function $f(x)$ is said to be $O(g(x))$; i.e., $f(x) = O(g(x))$, if there exists a constant M such that

$$|f(x)| \leqslant M|g(x)| \quad \text{as} \quad x \to x_0.$$

— Especially, if $f(x)$ is bounded in a neighborhood of x_0, then we write $f(x) = O(1)$ as $x \to x_0$.

— The notation $f(x) = g(x) + O(h(x))$ implies $f(x) - g(x) = O(h(x))$.

3.2• Small o

— If $\lim_{x\to x_0}\{f(x)/g(x)\} = 0$, then we write $f(x) = o(g(x))$.

— Obviously, the relation $f(x) = o(g(x))$ implies the weaker relation

$$f(x) = O(g(x)).$$

Proof: $\lim_{x\to x_0}\{f(x)/g(x)\} = 0$ implies that for any small real number $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever

$$|x - x_0| < \delta,$$

the inequality

$$\left|\frac{f(x)}{g(x)} - 0\right| < \varepsilon$$

is true; i.e., $|f(x)| < \varepsilon|g(x)|$ or $f(x) = O(g(x))$. □

— In addition,

$$\lim_{x\to x_0} \frac{o(g(x))}{g(x)} = 0 \quad \text{and} \quad \lim_{x\to x_0} o(1) = 0.$$

In other words, if $\lim_{x\to x_0} f(x) = 0$, we write $f(x) = o(1)$ as $x \to x_0$.

— Finally, if $f(x)$ is differentiable at the point x_0, then

$$f(x_0 + \Delta) - f(x_0) = f'(x_0)\Delta + o(\Delta),$$

which is the first–order Taylor expansion.

Proof: The definition

$$\begin{aligned} f'(x_0) &= \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0} && [\text{let } x = x_0+\Delta] \\ &= \lim_{\Delta\to 0} \frac{f(x_0+\Delta)-f(x_0)}{\Delta} \end{aligned}$$

becomes

$$0 = \lim_{\Delta\to 0} \frac{f(x_0+\Delta)-f(x_0)-f'(x_0)\Delta}{\Delta};$$

i.e., $f(x_0+\Delta)-f(x_0)-f'(x_0)\Delta = o(\Delta)$ as $\Delta\to 0$. □

3.3• Asymptotic equality $\asymp$

— If $\lim_{x\to x_0}\{f(x)/g(x)\} = 1$, then $f(x)$ is said to be asymptotic to $g(x)$, written as $f(x)\asymp g(x)$.

3.4• Their relations

— There is a lot of miniature theorems dealing with order relations.

— Among these are

$$\begin{aligned} O(g)+O(g) &= O(g), \\ o(g)+o(g) &= o(g), \\ O(g_1)O(g_2) &= O(g_1g_2), \\ o(g_1)O(g_2) &= o(g_1g_2), \\ |O(g)|^\lambda &= O(|g|^\lambda), \quad \lambda>0, \\ |o(g)|^\lambda &= o(|g|^\lambda), \quad \lambda>0. \end{aligned}$$

Proof of the second formula: Let $f_i(x) = o(g(x))$ for $i=1,2$, then

$$\lim_{x\to x_0}\frac{f_i(x)}{g(x)} = \lim_{x\to x_0}\frac{o(g(x))}{g(x)} = 0, \quad i=1,2.$$

We need to prove $f_1(x)+f_2(x) = o(g(x))$, which is equivalent to proving

$$0 = \lim_{x\to x_0}\frac{f_1(x)+f_2(x)}{g(x)} = \lim_{x\to x_0}\frac{f_1(x)}{g(x)} + \lim_{x\to x_0}\frac{f_2(x)}{g(x)} = 0+0.$$ □

2.1.2 Stationary points

4• Weierstrass theorem

- A continuous function $f(x)$ defined on a closed finite interval $[a, b]$ attains its *minimum* and *maximum* values on the interval.
- These extremal values are necessarily finite.

4.1• Principle of Fermat

— The extremal points could occur at the endpoints a or b or at an interior point c.

— In the later case, when $f(x)$ is differentiable, an even older *principle of Fermat* requires that $f'(c) = 0$, see Figure 2.1.

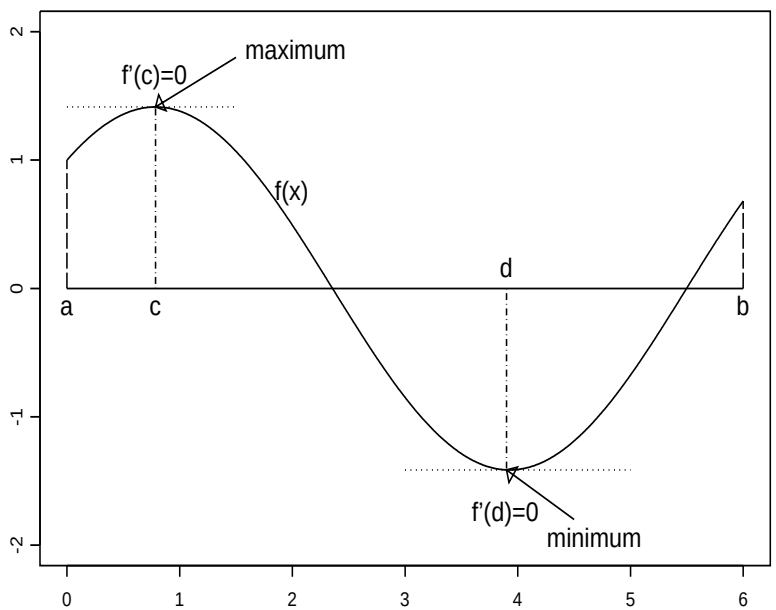

Figure 2.1 Illustration of Weierstrass theorem and Fermat's principle.

5• Saddle points

- An *optimal point* means a minimum or a maximum.
- The condition $f'(c) = 0$ is called *stationary condition*.
- The *stationary condition* $f'(c) = 0$ is no guarantee that c is optimal.
- A point c is said to be a *saddle point* if $f'(c) = 0$ but c is not an optimal point.
- It is possible for c to be a local rather than a global minimum or maximum (see Figure 2.2) or even to be a saddle point, see Figure 2.3(b).

5.1• Stationary points and critical points

— Stationary points = $\{c\colon\ f'(c) = 0\ \&\ c$ is an interior point$\}$.

— Critical points = $\{a, b,$ or any stationary point $c\}$, where a and b are endpoints.

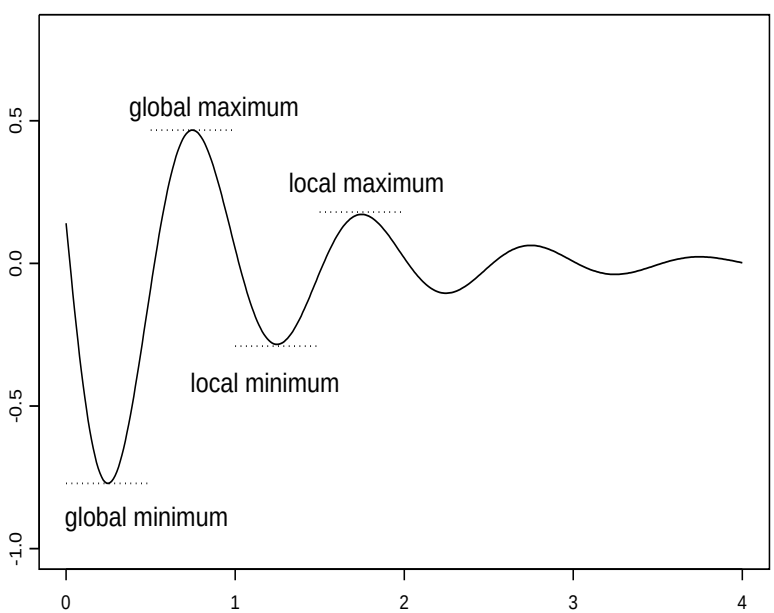

Figure 2.2 Illustration of global/local minimum or maximum.

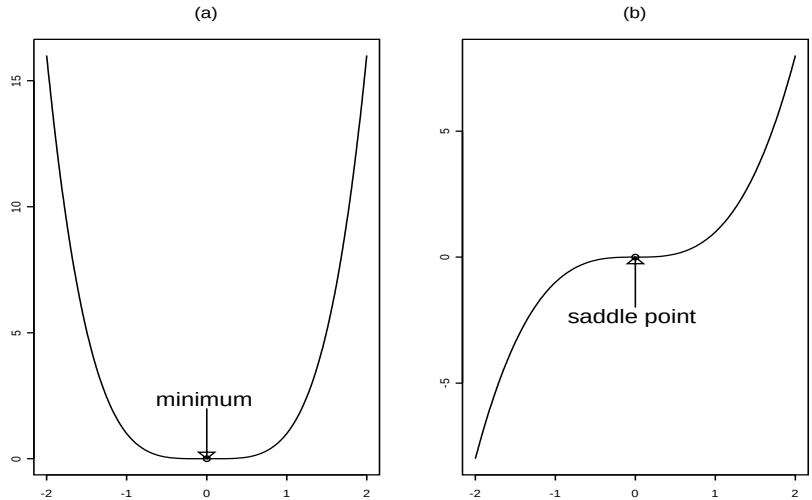

Figure 2.3 (a) $f(x) = x^4$ attains its minimum at $x = 0$. (b) $f(x) = x^3$ has a saddle point at $x = 0$. In both cases, $f''(0) = 0$.

6• The case that the minimum or maximum may not exist

- If the domain of $f(x)$ is not a closed finite interval $[a, b]$, then the minimum or maximum of $f(x)$ may not exist.
- For example, the maximum of $f(x) = x^2$ on $(-1, 1)$ does not exist, where there is one stationary point (i.e., a minimum) $x = 0$ and there are three critical points $\{-1, 0, 1\}$.
- One can usually rule out such behavior by examining the limit of $f(x)$ as x approaches an open boundary.
- For example on the interval $[a, \infty)$, if $\lim_{x\to\infty} f(x) = \infty$, then we can be sure that $f(x)$ possesses a minimum on the interval, and we can find it by comparing the values of $f(x)$ at a and any stationary points c.
- On a half open interval such as $(a, b]$, we can likewise find a minimum whenever $\lim_{x\to a} f(x) = \infty$.
- Similar considerations apply to finding a maximum.

7• Local minimum, local maximum and saddle points

- The nature of a stationary point c can be determined by testing the second derivative $f''(c)$.
- If $f''(c) > 0$, then c at least qualifies as a *local minimum*.
- Similarly, If $f''(c) < 0$, then c at least qualifies as a *local maximum*.
- The ambiguous case $f''(c) = 0$ is consistent with c being a local minimum, maximum, or saddle point.
- Figure 2.3 shows that $f(x) = x^4$ attains its minimum at $x = 0$ while $f(x) = x^3$ has a saddle point there. In both cases, $f''(0) = 0$.
- Higher–order derivatives or other qualitative features of $f(x)$ must be invoked to discriminate among these possibilities.

2.1.3 Convex and concave functions

8• Convex and strictly convex functions

- If $f''(x) \geqslant 0$ for all x, then $f(x)$ is said to be *convex*.
- Any stationary point of a convex function is a minimum, see Figure 2.4(a).
- If $f''(x) > 0$ for all x, then $f(x)$ is *strictly convex*, and there is *at most* one stationary point.
- Whenever it exists, the stationary point furnishes the *global minimum*, see Figure 2.4(b).

9• Concave and strictly concave functions

- A *concave* function satisfies $f''(x) \leqslant 0$ for all x, see Figure 2.4(c).
- If $f''(x) < 0$ for all x, then $f(x)$ is *strictly concave*, see Figure 2.4(d), and there is *at most* one stationary point.
- Whenever it exists, the stationary point is the *global maximum*.
- Concavity bears the same relation to maxima as convexity does to minima.

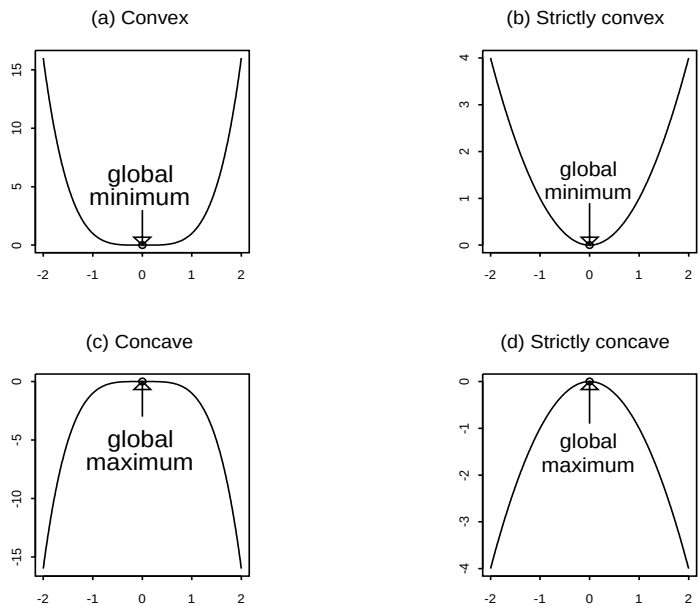

Figure 2.4 (a) $f_1(x) = x^4$ with $f_1''(x) = 12x^2 \geqslant 0$ for all x. (b) $f_2(x) = x^2$ with $f_2''(x) = 2 > 0$ for all x. (c) $f_3(x) = -x^4$ with $f_3''(x) = -12x^2 \leqslant 0$ for all x. (d) $f_4(x) = -x^2$ with $f_4''(x) = -2 < 0$ for all x.

2.1.4 Mean value theorem

10• MEAN VALUE THEOREM AND ITS GEOMETRIC INTERPRETATION

- Fermat's principle has some surprising consequences. Among these is the mean value property.

Theorem 2.1 (Mean value theorem). Suppose $f(x)$ is continuous on $[a, b]$ and differentiable on (a, b). Then there exists a point $c \in (a, b)$ such that

$$f(b) - f(a) = f'(c)(b - a). \tag{2.1}$$

As a consequence:

(1) If $f'(x) \geqslant 0$ for all $x \in (a, b)$, then $f(x)$ is increasing;

(2) If $f'(x) = 0$ for all $x \in (a, b)$, then $f(x)$ is constant;

(3) If $f'(x) \leqslant 0$ for all $x \in (a, b)$, then $f(x)$ is decreasing. ‖

Proof: Consider the function

$$g(x) = f(b) - f(x) + \frac{f(b) - f(a)}{b - a}(x - b).$$

Clearly, $g(x)$ is also continuous on $[a, b]$ and differentiable on (a, b). We have

$$g'(x) = -f'(x) + \frac{f(b) - f(a)}{b - a}.$$

Furthermore, $g(a) = g(b) = 0$. According to Weierstrass theorem, $g(x)$ attains either a maximum or a minimum at some $c \in (a, b)$. At this point, $g'(c) = 0$, i.e.,

$$0 = g'(c) = -f'(c) + \frac{f(b) - f(a)}{b - a},$$

which implies (2.1). □

Comment 1: This kind of proof is called *constructive* proof. The key is how to define the function $g(\cdot)$. ‖

10.1• A geometric interpretation

— For any two points $A(a, f(a))$ and $B(b, f(b))$, we can have a line AB. Then, there is a line CD which is parallel to line AB, and the line CD is tangent to the curve $f(x)$ at the tangent point $(c, f(c))$.

— We have

$$\tan(\theta) = \frac{f(b) - f(a)}{b - a} = f'(c).$$

Example 2.1 (A trigonometric identity). Show that

$$\cos^2(x) + \sin^2(x) = 1 \quad \text{for all } x \in (-\infty, \infty).$$

Proof: The function $f(x) = \cos^2(x) + \sin^2(x)$ has derivative

$$f'(x) = -2\cos(x)\sin(x) + 2\sin(x)\cos(x) = 0.$$

Therefore, according to Property (2) of Theorem 2.1, we know that $f(x)$ is constant for all x. Since $f(0) = 1$, we have $f(x) = 1$ for all $x \in \mathbb{R}$. □

2.1.5 Taylor theorem

11• Taylor expansions

- The n-order derivative of $f(x)$ is denoted by $f^{(n)}(x)$.
- The mean value theorem is the first–order Taylor expansion.
- The next theorem makes this clear and offers an explicit estimate of the error in a finite Taylor expansion of $f(x)$.

Theorem 2.2 (Taylor theorem). Suppose $f(x)$ has a derivative of order $n+1$ on an open interval (a,b) around the point $x_0 \in (a,b)$. Then for all $x \in (a,b)$, we have

$$f(x) = f(x_0) + \sum_{j=1}^{n} \frac{(x-x_0)^j}{j!} f^{(j)}(x_0) + \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(x^*),$$

for some x^* between x_0 and x. If $f^{(n+1)}(x^*)$ is bounded, then the remainder

$$R_{n+1}(x) = \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(x^*) = O(|x-x_0|^{n+1}). \qquad \|$$

Proof of the remainder formula: If $f^{(n+1)}(x^*)$ is bounded; i.e., there is an $M\,(>0)$ such that $|f^{(n+1)}(x^*)| < M$, then

$$\frac{|R_{n+1}(x)|}{|x-x_0|^{n+1}} = \frac{|f^{(n+1)}(x^*)|}{(n+1)!} < \frac{M}{(n+1)!} < M,$$

implying that $R_{n+1}(x) = O(|x-x_0|^{n+1})$. □

11.1• Two versions of the second–order Taylor expansion

— The first one is

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{1}{2}(x-x_0)^2 f''(x^*)$$

for some $x^* = \alpha x + (1-\alpha)x_0$, where $\alpha \in [0,1]$.

— The second one is

$$f(x) \approx f(x_0) + (x-x_0)f'(x_0) + \frac{1}{2}(x-x_0)^2 f''(x_0).$$

2.1.6 Rates of convergence

12• Convergence at rate q

- During the process of optimization, a sequence in the form of $\{x^{(t)}\}_{t=1}^{\infty}$ is produced and converges to x^*.
- If there is a constant $c \in [0, 1)$ such that

$$\lim_{t\to\infty} \frac{|x^{(t+1)} - x^*|}{|x^{(t)} - x^*|^q} = c, \tag{2.2}$$

  we say that the sequence $\{x^{(t)}\}_{t=1}^{\infty}$ converges to x^* at rate q.

12.1• Quadratic convergence

— When $q = 2$, we say that $\{x^{(t)}\}_{t=1}^{\infty}$ converges *quadratically*.

12.2• Linear convergence

— When $q = 1$, we say that the sequence $\{x^{(t)}\}_{t=1}^{\infty}$ converges *linearly*.

12.3• Superlinear convergence

— If there exists a sequence $\{\beta^{(t)}\}_{t=1}^{\infty}$ with $\lim_{t\to\infty} \beta^{(t)} = 0$ such that

$$\lim_{t\to\infty} \frac{|x^{(t+1)} - x^*|}{\beta^{(t)}|x^{(t)} - x^*|} = 0,$$

it is said that the sequence $\{x^{(t)}\}_{t=1}^{\infty}$ converges *superlinearly*.

2.1.7 The case of multiple dimensions

13• Gradient vector and hessian matrix

- The standard vocabulary and symbolism adopted here stress the minor adjustments necessary in going from one dimension to multiple dimensions.

13.1• Differential, gradient vector and score vector

— In mathematics, for a real–valued function $f(\boldsymbol{x})$ defined on the n-dimensional Euclidean space $\mathbb{R}^n$, the *differential* $\mathrm{d}f(\boldsymbol{x})$ is the generalization of the derivative $f'(x)$ with $x \in \mathbb{R}$. For our purposes,

$$\mathrm{d}f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_n} \right)_{1\times n}$$

is the row vector of partial derivatives.

— In optimization, its transpose is called the *gradient vector*

$$\nabla f(\boldsymbol{x}) = \{\mathrm{d}f(\boldsymbol{x})\}^\top = \begin{pmatrix} \dfrac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\boldsymbol{x})}{\partial x_n} \end{pmatrix}.$$

— In statistics, $\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is called *score vector*, where $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is the log-likelihood function and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

13.2• $\mathrm{d}^2 f(\boldsymbol{x})$, Hessian matrix and observed information matrix

— In mathematics and optimization, the symmetric matrix of second partial derivatives constitutes the *second differential* $\mathrm{d}^2 f(\boldsymbol{x})$ or *Hessian matrix*

$$\nabla^2 f(\boldsymbol{x}) = \mathrm{d}^2 f(\boldsymbol{x}) = \begin{pmatrix} \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_2} & \cdots & \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n} \end{pmatrix}.$$

— In statistics, $-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is called *observed information matrix*.

14• Stationary point and Fermat's principle

- A stationary point $\boldsymbol{x}$ satisfies $\nabla f(\boldsymbol{x}) = \mathbf{0}_n$.

- Fermat's principle says that all local maxima and minima on the interior of the domain of $f(\boldsymbol{x})$ are stationary points.

14.1• Local minimum and convexity

— If $\mathrm{d}^2 f(\boldsymbol{x})$ is positive definite at a stationary point $\boldsymbol{x}$ (denoted by $\mathrm{d}^2 f(\boldsymbol{x}) > 0$), then $\boldsymbol{x}$ furnishes a *local minimum*. If $\mathrm{d}^2 f(\boldsymbol{x})$ is negative definite (denoted by $\mathrm{d}^2 f(\boldsymbol{x}) < 0$), then $\boldsymbol{x}$ furnishes a *local maximum*.

— The function $f(\boldsymbol{x})$ is said to be *convex* if $\mathrm{d}^2 f(\boldsymbol{x})$ is positive semi-definite for all $\boldsymbol{x}$ (i.e., $\mathrm{d}^2 f(\boldsymbol{x}) \geqslant 0$); it is *strictly convex* if $\mathrm{d}^2 f(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}$ (i.e., $\mathrm{d}^2 f(\boldsymbol{x}) > 0$).

14.2• Global minimum and concave function

— Every stationary point of a convex function represents a *global minimum*. At most one stationary point exists per strictly convex function.

— Similar considerations apply to concave functions and global maxima, provided we substitute "negative" for "positive" throughout these definitions.

2.2 Newton's Method and Its Variants

15• Why is newton's method so important?

- From the viewpoint of applications, Newton's method is well known and much widely applied in statistics.
- From the viewpoint of speed of convergence, Newton's method has a quadratic rate of convergence.
- From the viewpoint of importance, it forms the basis of most modern optimization algorithms. Its many variants seek to retain its fast convergence while taming its defects.

16• The main idea of newton's method

- The key idea of Newton's method is to *locally* approximate the objective function by a strictly convex/concave quadratic function.
- Then, at each iteration, the quadratic function is optimized.

2.2.1 Newton's method and root finding

17• Formulation of the algorithm

- Newton's method can be motivated by the mean value theorem.
- Let $x^{(t)}$ approximate the root $x^{(\infty)}$ of the equation

$$g(x) = 0.$$

  Of course, we have $g(x^{(\infty)}) = 0$.

- According to the mean value theorem, since $g(x^{(\infty)}) = 0$, we have

$$g(x^{(t)}) = g(x^{(t)}) - g(x^{(\infty)}) = g'(z)(x^{(t)} - x^{(\infty)}) \tag{2.3}$$

  for some z between $x^{(t)}$ and $x^{(\infty)}$.

- If we replace z by the current iteration $x^{(t)}$, and replace $x^{(\infty)}$ by the next iteration $x^{(t+1)}$, then the equality (2.3) can be rewritten as

$$x^{(t+1)} = x^{(t)} - \frac{g(x^{(t)})}{g'(x^{(t)})}, \tag{2.4}$$

  which defines Newton's method.

17.1• Alternative derivation of (2.4)

— Applying the first–order Taylor expansion to $g(x)$, we have

$$g(x) = l(x) + R_2(x),$$

where

$$l(x) = g(x^{(t)}) + g'(x^{(t)})(x - x^{(t)})$$

is a straight line tangent to $g(x)$ at the point $(x^{(t)}, g(x^{(t)}))$.

— The basic idea is to locally approximate $g(x)$ by $l(x)$. By solving $l(x) = 0$, we obtain (2.4). □

Example 2.2 (An illustrative example). Use Newton's method to find the unique root of the following equation

$$0 = g(x) = 1.95 - \exp(-2/x) - 2\exp(-x^4), \quad x > 0.$$

Solution: It is easy to verify that

$$g'(x) = -2x^{-2}\exp(-2/x) + 8x^3\exp(-x^4).$$

Let $x^{(0)} = 1$, then from (2.4), we obtain

$$\begin{aligned}
x^{(1)} &= x^{(0)} - \frac{g(x^{(0)})}{g'(x^{(0)})} = 1 - \frac{1.0789}{2.6724} = 0.596273,\\
x^{(2)} &= x^{(1)} - \frac{g(x^{(1)})}{g'(x^{(1)})} = x^{(1)} - \frac{0.1526}{1.2981} = 0.478749,\\
x^{(3)} &= x^{(2)} - \frac{g(x^{(2)})}{g'(x^{(2)})} = x^{(2)} - \frac{0.0370}{0.6991} = 0.425798,\\
x^{(4)} &= x^{(3)} - \frac{g(x^{(3)})}{g'(x^{(3)})} = x^{(3)} - \frac{0.0056}{0.4969} = 0.414628,\\
x^{(5)} &= x^{(4)} - \frac{g(x^{(4)})}{g'(x^{(4)})} = x^{(4)} - \frac{2.0768\times 10^{-4}}{0.460135} = 0.414177,\\
x^{(6)} &= x^{(5)} - \frac{g(x^{(5)})}{g'(x^{(5)})} = x^{(5)} - \frac{3.2684\times 10^{-7}}{0.458687} = 0.414176, \quad \text{and}\\
x^{(7)} &= x^{(6)} - \frac{g(x^{(6)})}{g'(x^{(6)})} = x^{(6)} - \frac{8.1334\times 10^{-13}}{0.458685} = 0.414176.
\end{aligned}$$

Thus, $x^{(\infty)} = x^{(7)} = 0.414176$ is the unique root of $g(x) = 0$.

Newton's method is sensitive to initial values: In this example, when the initial value $x^{(0)} \in [0.182, 1.199]$, Newton's method defined by (2.4) converges to $x^{(\infty)} = 0.414176$. In other words, when $x^{(0)} \notin [0.182, 1.199]$, Newton's method diverges.

A geometric interpretation: Figure 2.5 gives a geometric interpretation for Newton's method.

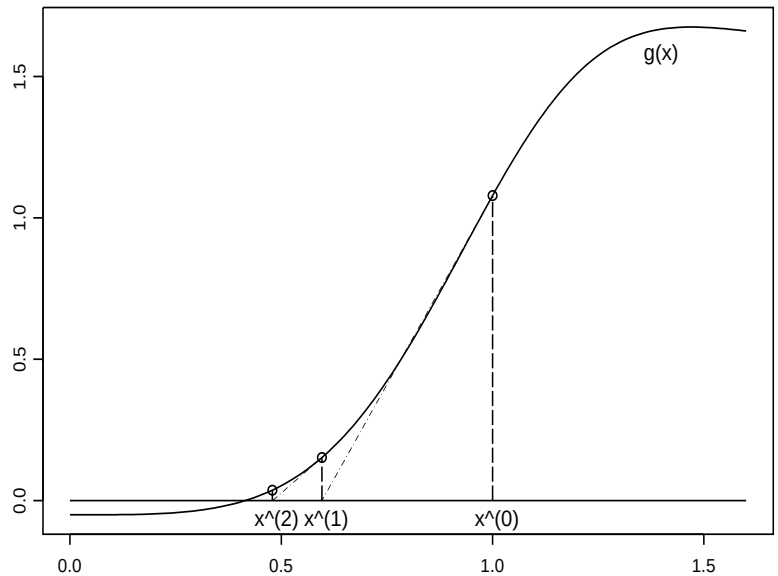

Figure 2.5 Two steps of Newton's method starting from $x^{(0)} = 1$ and moving toward the unique root of $g(x) = 0$ on $(0, \infty)$. The iterate $x^{(t+1)}$ is taken as the point of intersection of the x-axis and the tangent drawn through $(x^{(t)}, g(x^{(t)}))$. Newton's method fails to converge if $x^{(0)}$ is chosen too far to the left or right. ‖

18• Convergence properties

- From the perspective of functional iteration, Newton's method defined by (2.4) can be rephrased as

$$x^{(t+1)} = f(x^{(t)}), \tag{2.5}$$

where

$$f(x) = x - \frac{g(x)}{g'(x)}. \tag{2.6}$$

- The local convergence property of Newton's method is determined by

$$\begin{aligned} f'(x^{(\infty)}) &= f'(x)|_{x=x^{(\infty)}} \\ &= 1 - \frac{\{g'(x)\}^2 - g(x)g''(x)}{\{g'(x)\}^2}\bigg|_{x=x^{(\infty)}} \\ &= \frac{g(x^{(\infty)})g''(x^{(\infty)})}{\{g'(x^{(\infty)})\}^2} = 0, \end{aligned} \tag{2.7}$$

since $g(x^{(\infty)}) = 0$. In other words, $x^{(\infty)}$ is a stationary point of $f(\cdot)$.

- Let $\varepsilon^{(t+1)} = x^{(t+1)} - x^{(\infty)}$ be the current error in approximating $x^{(\infty)}$, then a second-order Taylor expansion around $x^{(\infty)}$ yields

$$\begin{aligned}
\varepsilon^{(t+1)} &= x^{(t+1)} - x^{(\infty)} \\
&\overset{(2.5)}{=} f(x^{(t)}) - x^{(\infty)} + \frac{g(x^{(\infty)})}{g'(x^{(\infty)})} \qquad [\because g(x^{(\infty)}) = 0] \\
&\overset{(2.6)}{=} f(x^{(t)}) - f(x^{(\infty)}) \\
&= f'(x^{(\infty)})(x^{(t)} - x^{(\infty)}) + \frac{1}{2} f''(z)(x^{(t)} - x^{(\infty)})^2 \\
&= f'(x^{(\infty)})\varepsilon^{(t)} + \frac{1}{2} f''(z)(\varepsilon^{(t)})^2 \\
&\overset{(2.7)}{=} \frac{1}{2} f''(z)(\varepsilon^{(t)})^2, \qquad (2.8)
\end{aligned}$$

where z again lies between $x^{(t)}$ and $x^{(\infty)}$.

18.1• Newton's method has quadratic convergence

— Provided $f''(z)$ is continuous, $|f''(x^{(\infty)})| < 2$, and the initial value $x^{(0)}$ is close enough to $x^{(\infty)}$, the error representation (2.8) makes it clear that Newton's method converges quadratically since

$$\lim_{t\to\infty} \frac{|\varepsilon^{(t+1)}|}{(\varepsilon^{(t)})^2} = \frac{1}{2}|f''(x^{(\infty)})| < 1.$$

18.2• An algorithm with linear convergence

— Now we consider some algorithm (say, the EM algorithm to be introduced in §2.3) represented by $x^{(t+1)} = h(x^{(t)})$, where $h(\cdot)$ is the iteration function.

— If $h(\cdot)$ satisfies

$$0 < |h'(x^{(\infty)})| < 1,$$

then the first–order Taylor expansion implies

$$\lim_{t\to\infty} \frac{|x^{(t+1)} - x^{(\infty)}|}{|x^{(t)} - x^{(\infty)}|} = |h'(x^{(\infty)})|, \qquad (2.9)$$

which is referred to as *linear convergence*.

Example 2.3 (Division without dividing). Use Newton's method to calculate the reciprocal of a number $a\,(>0)$, which is equivalent to solving the equation $g(x) = a - x^{-1} = 0$.

Solution: According to Newton's method (2.4), we have

$$x^{(t+1)} = x^{(t)} - \frac{a - (x^{(t)})^{-1}}{(x^{(t)})^{-2}} = x^{(t)}(2 - ax^{(t)}),$$

which involves multiplication and subtraction but no division. If $x^{(t+1)} > 0$, then $x^{(t)}(2 - ax^{(t)}) > 0$, so that $x^{(t)}$ must lie on the interval $(0, 2/a)$.

If $x^{(t)} \in (0, 2/a)$, then $x^{(t+1)}$ will reside on the shorter interval $(0, 1/a)$. Thus, starting on $(0, 1/a)$, the iterate $x^{(t+1)}$ monotonically increase to their limit $1/a$. Starting on $[1/a, 2/a)$, the first iterate satisfies $x^{(1)} \leqslant 1/a$, and subsequent iterates monotonically increase to $1/a$. Finally, starting outside $(0, 2/a)$, leads either to fixation at 0 or divergence to $-\infty$.

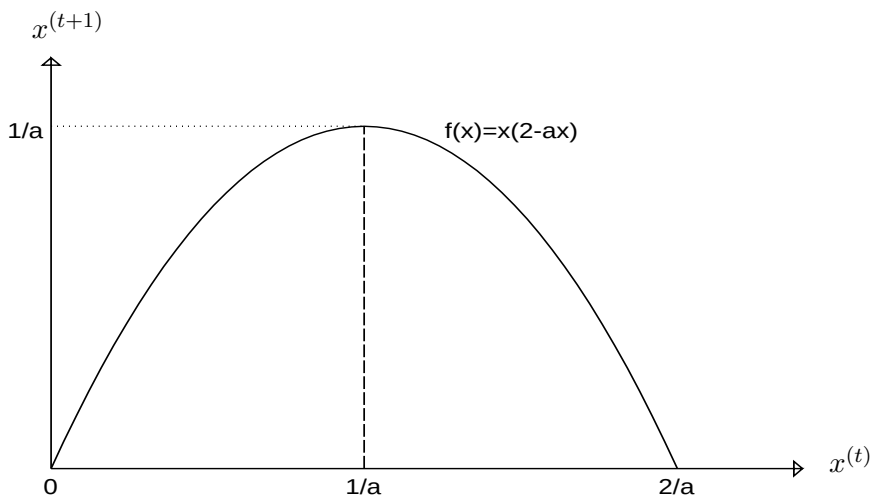

Figure 2.6 $x^{(t+1)} = f(x^{(t)})$, where $f(x) \hat{=} x(2 - ax)$ and $a > 0$. $f(x)$ attains its maximum $1/a$ at $x = 1/a$. ‖

2.2.2 Newton's method and optimization

19• Optimization with newton's method

- One of the advantages of Newton's method is that it is a root–finding technique as well as an optimization technique.

- Suppose we want to maximize the real–valued function $g(x)$, which is twice differentiable. That is, the goal is to find

$$\hat{x} = \arg\max_{x \in \mathbb{X}} g(x).$$

- Any stationary point x of $g(x)$ satisfies

$$g'(x) = 0. \tag{2.10}$$

  That is, we need to find a root of the equation (2.10).

- According to (2.4), Newton's method is defined by

$$x^{(t+1)} = x^{(t)} - \frac{g'(x^{(t)})}{g''(x^{(t)})}. \tag{2.11}$$

19.1• An alternative derivation of (2.11)

— We consider a second–order Taylor expansion of $g(x)$ around $x^{(t)}$,

$$\begin{aligned} g(x) &\approx g(x^{(t)}) + (x - x^{(t)})g'(x^{(t)}) + 0.5(x - x^{(t)})^2 g''(x^{(t)}) \\ &\hat{=} Q(x|x^{(t)}), \end{aligned}$$

and use the quadratic function $Q(x|x^{(t)})$ to approximate $g(x)$ in the neighbourhood of $x^{(t)}$.

— Let $\mathrm{d}Q(x|x^{(t)})/\mathrm{d}x = 0$, we have

$$g'(x^{(t)}) + 0.5 \times 2(x - x^{(t)})g''(x^{(t)}) = 0.$$

Solving x, we obtain (2.11). □

2.2.3 The Newton–Raphson algorithm

20• Aim of this subsection

- The main goal of this subsection is to find the MLEs of the parameter vector $\boldsymbol{\theta}$; i.e., to calculate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|Y_{\text{obs}}), \tag{2.12}$$

  where $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is the log-likelihood function, which is twice continuously differentiable and concave.

21• Score vector and information matrices

- Mathematically, for the case of one dimension, $f'(x)$ and $f''(x)$ denote the first–order and second–order derivatives, respectively. For the case of multiple dimensions, $\mathrm{d}f(\boldsymbol{x})$ denotes the first partial derivatives or the differential (a row vector), and $\mathrm{d}^2 f(\boldsymbol{x})$ the second partial derivatives (a matrix).

- Computationally/optimizationally, we call $\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}})$ as the gradient vector and $\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}})$ as the Hessian matrix, where ∇ denotes the derivative operator.

- Statistically, $\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is called the *score vector*,

$$\boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}}) = -\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}})$$

  is called the *observed information matrix* and

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{\theta}) &= E\{\boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}})\} \\ &= -\int \frac{\partial^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} f(Y_{\text{obs}};\boldsymbol{\theta})\, \mathrm{d}Y_{\text{obs}} \end{aligned}$$

  is called the *Fisher/expected information matrix*, where

$$f(Y_{\text{obs}};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$

  is the joint density of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\boldsymbol{\theta})$.

22• Derivation of the Newton–Raphson algorithm

- Statistics, like other scientific disciplines, has a special vocabulary. For example, Newton's method is usually called the *Newton–Raphson* (NR) algorithm.

- When the log-likelihood function $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is well behaved, a natural candidate for finding the MLEs is the NR algorithm or the *Fisher scoring* algorithm because they converge quadratically (Tanner 1996, p.26–27; Little & Rubin 2002, Ch.8).

22.1• Second–order Taylor expansion and the NR iteration

— Consider a second–order Taylor expansion of $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ around $\boldsymbol{\theta}^{(t)}$:

$$\begin{aligned}\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &\approx \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + (\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top \nabla^2 \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}).\end{aligned}$$

— Any stationary points of $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ satisfy (see Exercise 2.1)

$$\mathbf{0}_q = \nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}) \approx \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + \nabla^2 \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}).$$

— Let $\boldsymbol{\theta}^{(0)}$ be an initial value of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{(t)}$ the t-th approximation of $\hat{\boldsymbol{\theta}}$, the NR algorithm is defined by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{I}^{-1}(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}). \tag{2.13}$$

22.2• Asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$

— Let $\text{Se}(\hat{\boldsymbol{\theta}})$ and $\text{Cov}(\hat{\boldsymbol{\theta}})$ denote the standard error and the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$, respectively.

— Under certain regularity conditions (Rao 1973, p.364; Little & Rubin 2002, p.105–108), it can be shown by the Central Limit Theorem that

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \dot{\sim} N_q(\mathbf{0}, \boldsymbol{J}^{-1}(\boldsymbol{\theta}))$$

so that $\text{Cov}(\hat{\boldsymbol{\theta}}) = \boldsymbol{J}^{-1}(\boldsymbol{\theta})$.

22.3• Estimation of $\text{Cov}(\hat{\boldsymbol{\theta}})$ and the delta method

— It is important to note that $\boldsymbol{J}^{-1}(\boldsymbol{\theta})$ is the covariance of the asymptotic distribution, not the limit of the exact covariance.

— The standard errors are the square roots of the diagonal elements of the inverse Fisher information matrix. We estimate $\text{Cov}(\hat{\boldsymbol{\theta}})$ by

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{J}^{-1}(\hat{\boldsymbol{\theta}}) \doteq \begin{pmatrix} \hat{\sigma}_1^2 & \cdots & * \\ \vdots & & \vdots \\ * & \cdots & \hat{\sigma}_q^2 \end{pmatrix}$$

and estimate $\text{Se}(\hat{\boldsymbol{\theta}})$ by

$$\widehat{\text{Se}}(\hat{\boldsymbol{\theta}}) = (\hat{\sigma}_1, \ldots, \hat{\sigma}_q)^\top.$$

— Sometimes, when calculating $\boldsymbol{J}(\boldsymbol{\theta})$ is quite difficult, we could set

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{\theta}) &\approx \boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}}) \quad \text{and} \\ \boldsymbol{J}(\hat{\boldsymbol{\theta}}) &\approx \boldsymbol{I}(\hat{\boldsymbol{\theta}}|Y_{\text{obs}}). \end{aligned}$$

— In addition, the *delta method* can be used to derive the asymptotic distribution of $h(\hat{\boldsymbol{\theta}})$, where h is differentiable (see Exercise 2.2).

22.4• Three potential problems with the NR algorithm

— First, for complicated incomplete data or when the dimension of $\boldsymbol{\theta}$ is (very) large, it requires tedious calculations of the Hessian matrix at each iteration.

— Second, when the observed information evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ is apparently singular owing to collinearities among covariates, the NR does not work, for example, see Example 2.5.

— Third, $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ does not necessarily increase at each iteration for the NR algorithm, which may sometimes be divergent (Cox & Oakes 1984, p.172).

— Böhning & Lindsay (1988, p.645–646) provided an example of a concave function for which the NR algorithm does not converge if a poor initial value is chosen.

2.2.4 The Fisher scoring algorithm

23• A variant of the NR algorithm

- A variant of the NR algorithm is the Fisher scoring algorithm or scoring algorithm, where the observed information in (2.13) is replaced by the expected information:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{J}^{-1}(\boldsymbol{\theta}^{(t)})\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}).$$

- An extra benefit of the scoring algorithm is that $\boldsymbol{J}^{-1}(\hat{\boldsymbol{\theta}})$ provides an estimated asymptotic covariance of $\hat{\boldsymbol{\theta}}$ (Lange 1999, Ch.11).

2.2.5 Application to logistic regression

24• Logistic distribution

- The pdf of $X \sim \text{Logistic}(\mu, \sigma^2)$ is defined by

$$\text{Logistic}(x|\mu, \sigma^2) = \frac{\exp(-\frac{x-\mu}{\sigma})}{\sigma\{1+\exp(-\frac{x-\mu}{\sigma})\}^2}, \quad x \in \mathbb{R},$$

  where $\mu \in (-\infty, \infty)$ is the location parameter and $\sigma > 0$ is the scale parameter.

- The mean and variance of the X are given by

$$E(X) = \mu \quad \text{and} \quad \text{Var}(X) = \frac{\pi^2\sigma^2}{3}.$$

- The logistic is another symmetric and uni-modal distribution, more similar to the normal in appearance than the Laplace, but with even heavier tails.

- Both the cdf and its inverse function have closed–form expressions:

$$\begin{aligned} F_{\text{L}}(x|\mu, \sigma^2) &= \left(1 + \text{e}^{-\frac{x-\mu}{\sigma}}\right)^{-1}, \quad x \in \mathbb{R}, \\ F_{\text{L}}^{-1}(x|\mu, \sigma^2) &= \mu + \sigma \log\left(\frac{x}{1-x}\right), \quad x \in (0, 1). \end{aligned}$$

- The logistic generator in R is rlogis(N, location $= \mu$, scale $= \sigma$).

25• Logistic regression

- Consider the following logistic regression

$$\begin{cases} Y_i \overset{\text{ind}}{\sim} \text{Binomial}(n_i, p_i), \quad i = 1, \ldots, m, \\ \text{logit}(p_i) \hat{=} \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta} \hat{=} \eta_i, \end{cases} \tag{2.14}$$

  where Y_i denotes the number of subjects with positive response in the i-th group with n_i trials, p_i is the probability of a subject in the i-th group with positive response, $\boldsymbol{x}_{(i)} = (x_{i1}, \ldots, x_{iq})^\top$ is a known vector of covariates, $\boldsymbol{\theta}$ is an unknown $q \times 1$ vector of parameters, and η_i is the linear predictor for the i-th group.

25.1• The observed-data log-likelihood function

— Let $Y_{\text{obs}} = \{y_i\}_{i=1}^m$, where y_i denotes the realization of Y_i.

— The observed-data likelihood function of $\boldsymbol{\theta}$ is

$$\begin{aligned} L(\boldsymbol{\theta}|Y_{\text{obs}}) &= \prod_{i=1}^m \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i} \\ &= \prod_{i=1}^m \binom{n_i}{y_i} \left(\frac{p_i}{1-p_i}\right)^{y_i} (1-p_i)^{n_i}. \end{aligned}$$

— From (2.14), we obtain

$$\begin{aligned} p_i &= \frac{\mathrm{e}^{\eta_i}}{1+\mathrm{e}^{\eta_i}} = \frac{1}{1+\mathrm{e}^{-\eta_i}} = F_{\text{L}}(\eta_i|0,1) \quad \text{and} \\ 1-p_i &= \frac{1}{1+\mathrm{e}^{\eta_i}}. \end{aligned}$$

— Thus, the observed-data log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \text{constant} + \sum_{i=1}^m \left\{ y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i) \right\} \\ &\overset{(2.14)}{=} \text{constant} + \sum_{i=1}^m \left\{ y_i\eta_i - n_i \log(1+\mathrm{e}^{\eta_i}) \right\}. \end{aligned}$$

25.2• The Newton–Raphson algorithm

— The score vector, Hessian and observed information matrices are

$$\begin{aligned} \nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \sum_{i=1}^m \left(y_i\boldsymbol{x}_{(i)} - n_i \frac{\mathrm{e}^{\eta_i}}{1+\mathrm{e}^{\eta_i}} \boldsymbol{x}_{(i)} \right) \\ &= \sum_{i=1}^m \boldsymbol{x}_{(i)}(y_i - n_i p_i) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{p}), \\ \nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \frac{\partial \nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \boldsymbol{\theta}^\top} = -\sum_{i=1}^m \boldsymbol{x}_{(i)} n_i \frac{\mathrm{e}^{\eta_i}\,(1+\mathrm{e}^{\eta_i}) - \mathrm{e}^{2\eta_i}}{(1+\mathrm{e}^{\eta_i})^2} \boldsymbol{x}_{(i)}^\top \\ &= -\sum_{i=1}^m n_i p_i(1-p_i)\boldsymbol{x}_{(i)}\boldsymbol{x}_{(i)}^\top = -\boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{P}\boldsymbol{X}, \quad \text{and} \\ -\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{P}\boldsymbol{X}, \end{aligned} \tag{2.15}$$

respectively, where

$$\begin{aligned} \underset{m\times q}{\boldsymbol{X}} &= (\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(m)})^\top, \qquad \boldsymbol{y} = (y_1,\ldots,y_m)^\top, \\ \boldsymbol{N} &= \operatorname{diag}(n_1,\ldots,n_m), \qquad \boldsymbol{p} = (p_1,\ldots,p_m)^\top, \\ \boldsymbol{P} &= \operatorname{diag}\Big(p_1(1-p_1),\ldots,p_m(1-p_m)\Big). \end{aligned}$$

— Thus, the Newton–Raphson algorithm is defined by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + (\boldsymbol{X}^\top \boldsymbol{N}\boldsymbol{P}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{p}^{(t)}). \tag{2.16}$$

25.3• The Fisher scoring algorithm

— The observed information matrix does not depend on the observed data Y_{obs} so that $\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}})$.

— Thus, the Fisher scoring algorithm is also defined by (2.16). Finally,

$$\widehat{\operatorname{Cov}}(\hat{\boldsymbol{\theta}}) = (\boldsymbol{X}^\top \boldsymbol{N}\hat{\boldsymbol{P}}\boldsymbol{X})^{-1}.$$

Table 2.1 Mice exposure data

i	d_i	t_i	y_i	n_i	i	d_i	t_i	y_i	n_i
1	1.5	96.0	44	120	10	3.5	7.0	152	280
2	1.5	168.0	37	80	11	3.5	14.0	55	80
3	1.5	336.0	43	80	12	3.5	24.0	98	140
4	1.5	504.0	35	60	13	3.5	48.0	121	160
5	3.5	0.5	29	100	14	7.0	0.5	52	120
6	3.5	1.0	53	200	15	7.0	1.0	62	120
7	3.5	2.0	13	40	16	7.0	1.5	61	120
8	3.5	3.0	75	200	17	7.0	2.0	86	120
9	3.5	5.0	23	40					

Example 2.4 (Mice exposure data). We consider the mice exposure data in Table 2.1 reported by Larsen *et al.* (1979) and previously analyzed by Hasselblad *et al.* (1980), where $m = 17$, y_i denotes the number of dead mice among n_i mice in the i-th group, and p_i is the probability of death

for any mouse exposured to nitrous dioxide (NO_2) in the i-th group. We wish to investigate whether or not the probability of death depends on two further variables d_i (degree of exposure to NO_2), t_i (exposure time) and their interaction. We first take logarithms, $x_{i1} = \log(d_i)$ and $x_{i2} = \log(t_i)$, and then standardize x_{i1} and x_{i2} by subtracting the sample mean and then dividing by the sample standard deviation (Leonard 2000, p.147).

Solution: The considered model is

$$\text{logit}(p_i) = \theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + (x_{i1} \times x_{i2})\theta_3.$$

Use $\boldsymbol{\theta}^{(0)} = \mathbf{1}_4$ as the initial values, the NR algorithm (2.16) converged in 4 iterations. The MLEs of parameters are $\hat{\boldsymbol{\theta}} = (0.1852,\ 1.0384,\ 1.2374,\ 0.2287)^\top$. The estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} 0.00391 & 0.00017 & 0.00136 & 0.00249 \\ 0.00017 & 0.00827 & 0.00725 & 0.00047 \\ 0.00136 & 0.00725 & 0.00964 & 0.00175 \\ 0.00249 & 0.00047 & 0.00175 & 0.00356 \end{pmatrix},$$

and estimated standard errors are $0.0624, 0.0909, 0.0981$ and 0.0597. ‖

25.4• The problem of collinearity

— The next example shows that the NR algorithm does not work because of the problem of collinearity.

Example 2.5 (Cancer remission data). We consider the cancer remission data (Lee 1974) listed in Table 2.2. The binary outcome $y_i = 1$ implies that the i-th patient's remission is occurred and $y_i = 0$ otherwise. There are six explanatory variables, where $x_{i1} = \text{CELL}/100$, $x_{i2} = \text{SMEAR}/100$, $x_{i3} = \text{ABS.INFIL}/100$, $x_{i4} = \text{LI}/20$, $x_{i5} = \log(\text{ABS.BLAST} + 1)$, and $x_{i6} = \text{Temperature}/100$.

Solution: The logistic regression model is

$$\text{logit}(p_i) = \theta_0 + x_{i1}\theta_1 + \cdots + x_{i6}\theta_6.$$

Use $\boldsymbol{\theta}^{(0)} = \mathbf{1}_7$ as the initial values, the NR algorithm (2.16) does not work because the observed information matrix evaluated at $\boldsymbol{\theta}^{(0)}$ is apparently singular. The problem persists with other initial values. In fact, this is not surprising since the correlation coefficients for (x_1, x_2), (x_1, x_3), (x_2, x_3) are 0.98, -0.97, and -0.998, respectively. ‖

Table 2.2 Cancer remission data

i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	y_i	i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	y_i
1	0.80	0.88	0.70	0.8	0.176	0.982	0	15	1.00	0.33	0.33	0.4	0.176	1.010	0
2	1.00	0.87	0.87	0.7	1.053	0.986	0	16	0.90	0.93	0.84	0.6	1.591	1.020	0
3	1.00	0.65	0.65	0.6	0.519	0.982	0	17	0.95	0.32	0.30	1.6	0.886	0.988	0
4	0.95	0.87	0.83	1.9	1.354	1.020	0	18	1.00	0.73	0.73	0.7	0.398	0.986	0
5	1.00	0.45	0.45	0.8	0.322	0.999	0	19	0.80	0.83	0.66	1.9	1.100	0.996	1
6	0.95	0.36	0.34	0.5	0.000	1.038	0	20	0.90	0.36	0.32	1.4	0.740	0.992	1
7	0.85	0.39	0.33	0.7	0.279	0.988	0	21	0.90	0.75	0.68	1.3	0.519	0.980	1
8	0.70	0.76	0.53	1.2	0.146	0.982	0	22	0.95	0.97	0.92	1.0	1.230	0.992	1
9	0.80	0.46	0.37	0.4	0.380	1.006	0	23	1.00	0.84	0.84	1.9	2.064	1.020	1
10	0.20	0.39	0.08	0.8	0.114	0.990	0	24	1.00	0.63	0.63	1.1	1.072	0.986	1
11	1.00	0.90	0.90	1.1	1.037	0.990	0	25	1.00	0.58	0.58	1.0	0.531	1.002	1
12	0.65	0.42	0.27	0.5	0.114	1.014	0	26	1.00	0.60	0.60	1.7	0.964	0.990	1
13	1.00	0.75	0.75	1.0	1.322	1.004	0	27	1.00	0.69	0.69	0.9	0.398	0.986	1
14	0.50	0.44	0.22	0.6	0.114	0.990	0								

2.3 The Expectation–Maximization (EM) Algorithm

26• Why do we need alternatives to the NR algorithm?

- There are three potential drawbacks with the NR algorithm as shown in **22.4•**.
- For *missing/incomplete data problems*, the NR algorithm is often difficult to apply while the EM algorithm (Dempster *et al.* 1977) could be a useful alternative.

26.1• Missing data problems

— For example, let

$$\binom{1}{1}, \binom{1}{-1}, \binom{-1}{1}, \binom{-1}{-1}, \binom{2}{*}, \binom{-2}{*}, \binom{*}{2}, \binom{*}{-2}$$

be realizations of a set of random vectors from $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where "$*$" denotes missing data.

26.2• Incomplete data problems

— For example, let

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_m \\ y_m \end{pmatrix}, \begin{pmatrix} x^*_{m+1} \\ y_{m+1} \end{pmatrix}, \begin{pmatrix} x_{m+2} \\ y^*_{m+2} \end{pmatrix}$$

be realizations of a set of random vectors from $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where x^*_{m+1} and y^*_{m+2} are unobservable. But, we know that $x^*_{m+1} < \delta_1$ and $y^*_{m+2} < \delta_2$ with known δ_1 and δ_2.

27• What is the EM algorithm?

- The EM algorithm is an *iterative deterministic* method for finding the MLEs or the posterior mode, and is remarkably simple both conceptually and computationally in many important problems.

27.1• Application range of the EM Algorithm

— The EM algorithm can be applied to missing/incomplete data problems, and *latent variables problems*.

— In fact, there are no missing data in latent variables problems. However, when the optimization problem is very complicated, we could simplify the complicated optimization by introducing some latent variables/data, see Examples 2.6 and 2.7 below.

27.2• Basic idea of the EM

— The basic idea of the EM is that instead of performing a complicated optimization, one augments the observed data with latent data to perform a series of simple optimizations.

2.3.1 The formulation of the EM algorithm

28• What is the original EM?

- Let $\ell(\boldsymbol{\theta}|Y_{\text{obs}}) \hat{=} \log\{L(\boldsymbol{\theta}|Y_{\text{obs}})\}$ denote the log-likelihood function.
- Usually, directly solving the MLEs $\hat{\boldsymbol{\theta}}$ defined by (2.12) is extremely difficult.

- We augment the observed data Y_{obs} with a latent variable Z (or latent vector $\mathbf{z}$) so that both the complete-data log-likelihood $\ell(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and the conditional predictive distribution $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ are available, where z is the realization of Z and "*available*" means
  - the MLEs of $\boldsymbol{\theta}$ based on the complete-data $Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\} = \{Y_{\text{obs}}, z\}$ are given by
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|Y_{\text{com}}),$$
    which have closed–form expressions, and
  - the $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ is a standard density or a product of standard densities.

28.1• Description of the EM algorithm

— Each iteration of the EM algorithm consists of an *expectation step* (E-step) and a *maximization step* (M-step).

— Specifically, let $\boldsymbol{\theta}^{(t)}$ be the t-th approximation of the MLEs $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

— The E-step is to compute the Q-function defined by
$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\Big\{\ell(\boldsymbol{\theta}|Y_{\text{obs}}, Z)\Big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\Big\} \\ &= \int_{\mathbb{Z}} \ell(\boldsymbol{\theta}|Y_{\text{obs}}, z) \times f(z|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})\,\mathrm{d}z. \end{aligned} \tag{2.17}$$

— The M-step is to maximize the Q-function with respect to $\boldsymbol{\theta}$ to obtain
$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{2.18}$$

— The two-step process is repeated until convergence occurs.

28.2• Remarks on the Q-function

— The Q-function is sometimes called *surrogate* function.

— The Q-function is the expectation of the complete-data log-likelihood function with respect to the conditional predictive density given Y_{obs} and $\boldsymbol{\theta}^{(t)}$.

Example 2.6 (Genetic linkage model). In this study (Rao 1973; Tanner 1996, p.66 & p.113; Lange 2002), AB/ab animals are crossed to measure the recombination fraction (r) between loci with alleles A and a at the first locus and alleles B and b at the second locus. Then the offspring of an $AB/ab \times AB/ab$ mating fall into the four categories AB, Ab, aB and ab with cell probabilities

$$\frac{\theta+2}{4}, \ \frac{1-\theta}{4}, \ \frac{1-\theta}{4}, \ \frac{\theta}{4}, \quad 0 \leqslant \theta \leqslant 1,$$

where $\theta = (1-r)^2$. Observed frequencies $\{Y_1, \ldots, Y_4\}$ follow a multinomial distribution with above cell probabilities; i.e.,

$$(Y_1, \ldots, Y_4)^\top \sim \text{Multinomial}\left(n; \frac{\theta+2}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right),$$

where $n = \sum_{i=1}^4 y_i$ and $Y_{\text{obs}} = \{y_i\}_{i=1}^4$ denote the corresponding observed values of $\{Y_i\}_{i=1}^4$. The observed-data likelihood function of θ is given by

$$\begin{aligned} L(\theta|Y_{\text{obs}}) &= \binom{n}{y_1, \ldots, y_4} \left(\frac{\theta+2}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} \\ &\propto \underbrace{\left(\frac{\theta+2}{4}\right)^{y_1}}_{\text{difficult part}} \times \underbrace{\left(\frac{1-\theta}{4}\right)^{y_2+y_3} \left(\frac{\theta}{4}\right)^{y_4}}_{\text{binomial density kernel}}. \end{aligned}$$

The aim is to find the MLE of θ.

Why do we choose "genetic linkage model" as the first example?

— Though the likelihood function $L(\theta|Y_{\text{obs}}) \propto (\theta+2)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}$ is very simple, and the NR algorithm can be employed to find the MLE $\hat{\theta}$, we want to show how the EM algorithm apply.

— We want to show how we formulate a non-missing data problem into a latent variable problem; i.e., how to introduce latent variables, which is called *data augmentation* (DA).

Solution: Augmenting the observed data Y_{obs} with a latent variable Z by splitting y_1 and the corresponding cell probability as follows:

$$\begin{array}{ccccc} y_1 & = & Z & + & (y_1 - Z), \\ \updownarrow & & \updownarrow & & \updownarrow \\ \dfrac{\theta+2}{4} & = & \dfrac{\theta}{4} & + & \dfrac{2}{4} \end{array} \quad \text{or} \quad 1 = \frac{\theta}{\theta+2} + \frac{2}{\theta+2}.$$

Thus, the distribution of $Z|(Y_{\text{obs}}, \theta)$ is

$$\begin{aligned} f(z|Y_{\text{obs}}, \theta) &= \text{Binomial}\left(z \middle| y_1, \frac{\theta}{\theta+2}\right) \\ &= \binom{y_1}{z}\left(\frac{\theta}{\theta+2}\right)^z\left(\frac{2}{\theta+2}\right)^{y_1-z}, \end{aligned} \tag{2.19}$$

for $z = 0, 1, \ldots, y_1$. We have

$$E(Z|Y_{\text{obs}}, \theta) = E(Z|y_1, \theta) = y_1 \times \frac{\theta}{\theta+2}.$$

Calculating the Q-function: The likelihood of the complete-data $\{Y_{\text{obs}}, z\} = \{z, y_1 - z, y_2, y_3, y_4\}$ is

$$L(\theta|Y_{\text{obs}}, z) = \binom{n}{z, y_1 - z, y_2, y_3, y_4}\left(\frac{\theta}{4}\right)^z\left(\frac{2}{4}\right)^{y_1-z}\left(\frac{1-\theta}{4}\right)^{y_2+y_3}\left(\frac{\theta}{4}\right)^{y_4}$$

so that the complete-data log-likelihood function is given by

$$\ell(\theta|Y_{\text{obs}}, z) = c(z) + (z + y_4)\log\theta + (y_2 + y_3)\log(1-\theta), \tag{2.20}$$

where $c(z)$ is a function of z, not depending on θ. From (2.17), we obtain

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \int_{\mathbb{Z}} \ell(\theta|Y_{\text{obs}}, z) \times f(z|Y_{\text{obs}}, \theta^{(t)})\,\mathrm{d}z \\ &= \sum_{z=0}^{y_1}\{c(z) + (z + y_4)\log\theta + (y_2 + y_3)\log(1-\theta)\} \\ &\quad \times \binom{y_1}{z}\left(\frac{\theta^{(t)}}{\theta^{(t)}+2}\right)^z\left(\frac{2}{\theta^{(t)}+2}\right)^{y_1-z} \\ &= E\{c(Z)|Y_{\text{obs}}, \theta^{(t)}\} + \{E(Z|Y_{\text{obs}}, \theta^{(t)}) + y_4\}\log\theta \\ &\quad + (y_2 + y_3)\log(1-\theta). \end{aligned}$$

The EM iteration: From (2.18), we have

$$\begin{aligned} \theta^{(t+1)} &= \frac{E(Z|Y_{\text{obs}}, \theta^{(t)}) + y_4}{E(Z|Y_{\text{obs}}, \theta^{(t)}) + y_2 + y_3 + y_4} \\ &= \frac{y_1\theta^{(t)}/(\theta^{(t)}+2) + y_4}{y_1\theta^{(t)}/(\theta^{(t)}+2) + y_2 + y_3 + y_4}. \end{aligned} \tag{2.21}$$

An illustrative real data set: Consider $Y_{\text{obs}} = \{y_1, y_2, y_3, y_4\} = \{125, 18, 20, 34\}$. If let $\theta^{(0)} = 0.5$, by using (2.21), we obtain

$$\begin{aligned} \theta^{(1)} &= 0.608247, \quad \theta^{(2)} = 0.624321, \\ \theta^{(3)} &= 0.626489, \quad \theta^{(4)} = 0.626777, \\ \theta^{(5)} &= 0.626816, \quad \theta^{(6)} = 0.626821, \\ \theta^{(7)} &= 0.626821. \end{aligned}$$

That is, the EM algorithm converged in 7 iterations with precision 10^{-6}. ‖

29• What is the modern EM?

- Step 1: Find the complete-data log-likelihood function $\ell(\theta|Y_{\text{obs}}, z)$ given by (2.20);
- Step 2: Find the conditional predictive distribution $f(z|Y_{\text{obs}}, \theta)$ given by (2.19);
- E-step: From (2.19), compute the conditional expectation

$$E(Z|Y_{\text{obs}}, \theta) = y_1\theta/(\theta + 2);$$

- M-step: From (2.20), find the complete-data MLE

$$\hat{\theta} = \frac{z + y_4}{z + y_2 + y_3 + y_4}$$

and update $\hat{\theta}$ by replacing z with $E(Z|Y_{\text{obs}}, \theta)$.

Example 2.7 (Two-parameter multinomial model). Gelfand & Smith (1990) extend the genetic-linkage model in Example 2.6 to a two-parameter multinomial model:

$$(Y_1, \ldots, Y_5)^\top \sim \text{Multinomial}(n; a_1\theta_1 + b_1, a_2\theta_1 + b_2, a_3\theta_2 + b_3, a_4\theta_2 + b_4, c\theta_3),$$

where $a_i, b_i > 0$ are known, $0 < c = 1 - \sum_{i=1}^4 b_i = a_1 + a_2 = a_3 + a_4 \leqslant 1$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top \in \mathbb{T}_3$. Use the EM algorithm to find the MLEs of $\boldsymbol{\theta}$.

Solution: The likelihood function for the observed-data $Y_{\text{obs}} = \{y_i\}_{i=1}^5$ is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) \propto (a_1\theta_1 + b_1)^{y_1}(a_2\theta_1 + b_2)^{y_2}(a_3\theta_2 + b_3)^{y_3}(a_4\theta_2 + b_4)^{y_4}(c\theta_3)^{y_5}.$$

We introduce a latent vector $\mathbf{z} = (Z_1, \ldots, Z_4)^\top$ by splitting $y_i = Z_i + (y_i - Z_i)$ for $i = 1, \ldots, 4$ so that the augmented sampling distribution is

$$\text{Multinomial}_9(n; a_1\theta_1, b_1, a_2\theta_1, b_2, a_3\theta_2, b_3, a_4\theta_2, b_4, c\theta_3).$$

Thus, the complete-data likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}) &\propto \left\{\prod_{i=1}^2 (a_i\theta_1)^{z_i} b_i^{y_i - z_i}\right\} \left\{\prod_{i=3}^4 (a_i\theta_2)^{z_i} b_i^{y_i - z_i}\right\} (c\theta_3)^{y_5} \\ &\propto \theta_1^{z_1+z_2}\theta_2^{z_3+z_4}\theta_3^{y_5}, \end{aligned} \tag{2.22}$$

where $\boldsymbol{z} = (z_1, \ldots, z_4)^\top$. In addition, the conditional predictive density is

$$f(\boldsymbol{z}|Y_{\text{obs}}, \boldsymbol{\theta}) = \prod_{i=1}^4 \text{Binomial}(z_i|y_i, p_i), \tag{2.23}$$

where

$$p_i = \frac{a_i\theta_1}{a_i\theta_1 + b_i} \text{ for } i = 1, 2 \quad \text{and} \quad p_i = \frac{a_i\theta_2}{a_i\theta_2 + b_i} \text{ for } i = 3, 4. \tag{2.24}$$

Complete-data MLEs: Now suppose we are interested in finding the MLEs of $\boldsymbol{\theta}$. From (2.22), the complete-data MLEs are given by

$$\hat{\theta}_1 = \frac{z_1 + z_2}{\Delta}, \quad \hat{\theta}_2 = \frac{z_3 + z_4}{\Delta}, \quad \hat{\theta}_3 = \frac{y_5}{\Delta}, \tag{2.25}$$

where $\Delta = \sum_{i=1}^4 z_i + y_5$. Thus, the E-step of the EM algorithm computes

$$E(Z_i|Y_{\text{obs}}, \boldsymbol{\theta}) = y_i p_i$$

with p_i being defined by (2.24), and the M-step updates (2.25) by replacing z_i with $E(Z_i|Y_{\text{obs}}, \boldsymbol{\theta})$ for $i = 1, \ldots, 4$.

An illustrative real data set: For illustrative purpose, we consider the same dataset as that of Gelfand & Smith (1990):

$$\begin{aligned} \{y_1, \ldots, y_5\} &= \{14, 1, 1, 1, 5\}, \quad a_1 = \cdots = a_4 = 0.25, \\ b_1 = 1/8, &\quad b_2 = b_3 = 0, \quad b_4 = 3/8. \end{aligned} \tag{2.26}$$

Use $\boldsymbol{\theta}^{(0)} = \mathbf{1}_3/3$ as initial values, we obtain

$$\begin{aligned}
\boldsymbol{\theta}^{(1)} &= (0.516358,\ 0.092461,\ 0.391181)^\top,\\
\boldsymbol{\theta}^{(2)} &= (0.572495,\ 0.074665,\ 0.352840)^\top,\\
\boldsymbol{\theta}^{(3)} &= (0.583528,\ 0.072133,\ 0.344339)^\top,\\
\boldsymbol{\theta}^{(4)} &= (0.585487,\ 0.071707,\ 0.342806)^\top,\\
\boldsymbol{\theta}^{(5)} &= (0.585828,\ 0.071633,\ 0.342538)^\top,\\
\boldsymbol{\theta}^{(6)} &= (0.585888,\ 0.071620,\ 0.342492)^\top,\\
\boldsymbol{\theta}^{(7)} &= (0.585898,\ 0.071618,\ 0.342484)^\top,\\
\boldsymbol{\theta}^{(8)} &= (0.585900,\ 0.071618,\ 0.342482)^\top,\\
\boldsymbol{\theta}^{(9)} &= (0.585900,\ 0.071618,\ 0.342482)^\top.
\end{aligned}$$

The EM algorithm converged in 8 iterations with precision 10^{-6}. The obtained MLEs are

$$\hat{\boldsymbol{\theta}} = (0.585900,\ 0.0716178,\ 0.342482)^\top. \qquad \|$$

Example 2.8 (Right censored regression model). Consider the linear regression model (Wei & Tanner 1990; Chib 1992):

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\,\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{I}_n),$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$ is the response random vector, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the realization vector of $\mathbf{y}$, $\boldsymbol{X} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})^\top$ the covariate matrix, $\boldsymbol{\beta}$ and σ^2 the unknown parameters. Suppose that the first r components of $\boldsymbol{y}$ are uncensored and the remaining $n - r$ are right censored (c_i denotes a censored time). Use the EM algorithm to find the MLEs of $\boldsymbol{\beta}$ and σ^2.

Solution: We augment the observed data $Y_{\text{obs}} = \{y_1, \ldots, y_r; c_{r+1}, \ldots, c_n\}$ with the unobserved failure times $\mathbf{z} = (Z_{r+1}, \ldots, Z_n)^\top$. If we had observed the value of $\mathbf{z}$, say $\boldsymbol{z} = (z_{r+1}, \ldots, z_n)^\top \equiv (y_{r+1}, \ldots, y_n)^\top$ with $z_i > c_i$ $(i = r+1, \ldots, n)$, we could have the complete-data likelihood function

$$L(\boldsymbol{\beta}, \sigma^2 | Y_{\text{obs}}, \boldsymbol{z}) = N_n(\boldsymbol{y} | \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n). \tag{2.27}$$

The conditional predictive density is the product of $(n - r)$ independent truncated normal densities:

$$f(\boldsymbol{z} | Y_{\text{obs}}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=r+1}^{n} \text{TN}(z_i | \boldsymbol{x}_{(i)}^\top \boldsymbol{\beta}, \sigma^2;\ c_i, \infty). \tag{2.28}$$

Table 2.3 Insulation life data with censoring

i	Y_i	T_i	i	Y_i	T_i	i	c_i	T_i	i	c_i	T_i
1	3.2464	2.2563	13	2.6107	2.0276	18	3.9066	2.3629	30	3.7362	2.2563
2	3.4428	2.2563	14	2.6107	2.0276	19	3.9066	2.3629	31	3.2253	2.1589
3	3.5371	2.2563	15	2.7024	2.0276	20	3.9066	2.3629	32	3.2253	2.1589
4	3.5492	2.2563	16	2.7024	2.0276	21	3.9066	2.3629	33	3.2253	2.1589
5	3.5775	2.2563	17	2.7024	2.0276	22	3.9066	2.3629	34	3.2253	2.1589
6	3.6866	2.2563				23	3.9066	2.3629	35	3.2253	2.1589
7	3.7157	2.2563				24	3.9066	2.3629	36	2.7226	2.0276
8	2.6107	2.1589				25	3.9066	2.3629	37	2.7226	2.0276
9	2.6107	2.1589				26	3.9066	2.3629	38	2.7226	2.0276
10	3.1284	2.1589				27	3.9066	2.3629	39	2.7226	2.0276
11	3.1284	2.1589				28	3.7362	2.2563	40	2.7226	2.0276
12	3.1584	2.1589				29	3.7362	2.2563			

The E-step requires to calculate (see Exercise 2.3)

$$E(Z_i|Y_{\text{obs}},\boldsymbol{\beta},\sigma^2) \;=\; \boldsymbol{x}_{(i)}^\top\boldsymbol{\beta} + \sigma\Psi\left(\frac{c_i - \boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sigma}\right) \quad \text{and}$$

$$E(Z_i^2|Y_{\text{obs}},\boldsymbol{\beta},\sigma^2) \;=\; (\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta})^2 + \sigma^2 + \sigma(c_i + \boldsymbol{x}_{(i)}^\top\boldsymbol{\beta})\Psi\left(\frac{c_i - \boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sigma}\right),$$

for $i = r+1,\ldots,n$, where

$$\Psi(x) \;\hat{=}\; \frac{\phi(x)}{1-\Phi(x)}, \tag{2.29}$$

$\phi(x)$ and $\Phi(x)$ denote the pdf and cdf of $N(0,1)$, respectively. The M-step is to find the complete-data MLEs:

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \quad \sigma^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{n},$$

where $\boldsymbol{y} = (y_1,\ldots,y_r;z_{r+1},\ldots,z_n)^\top$.

smidata!insulation life

Insulation life data: We consider the data set of accelerated life tests on electrical insulation in 40 motorettes (Schmee & Hahn 1979). Ten motorettes

were tested at each of the four temperatures: 150°C, 170°C, 190°C and 220°C. Testing was terminated at different times at each temperature level. The linear model is

$$Y_i = \beta_0 + \beta_1 T_i + \sigma\varepsilon_i, \quad \varepsilon_i \sim N(0,1),$$

where $Y_i = \log_{10}(\text{failuretime})$ and $T_i = 1000/(\text{Temperature} + 273.2)$. The dataset is reordered so that the first $r = 17$ units are uncensored and the remaining $n - r = 40 - 17 = 23$ are censored, see Table 2.3. Use $\beta_0^{(0)} = \beta_1^{(0)} = \sigma^{(0)} = 1$ as initial values, the EM algorithm converged to

$$\hat{\beta}_0 = -6.019, \quad \hat{\beta}_1 = 4.311 \quad \text{and} \quad \hat{\sigma} = 0.2592$$

in 40 iterations with precision 10^{-5}. ‖

2.3.2 The ascent property of the EM algorithm

30• Why does the em algorithm work?

- One of the advantages of the EM algorithm is its *numerical stability.*
- That is, the EM algorithm increases the observed-data log-likelihood $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at each iteration, see this subsection for more detail.
- Thus, it avoids wildly overshooting or undershooting the maximum of the likelihood along its current direction of search like the NR.

31• The ascent property of the em

31.1• Starting from Bayes theorem

— The Bayes theorem gives

$$\frac{f(Y_{\text{obs}}, z|\boldsymbol{\theta})}{f(Y_{\text{obs}}|\boldsymbol{\theta})} = f(z|Y_{\text{obs}}, \boldsymbol{\theta}) \quad \text{or} \quad \frac{L(\boldsymbol{\theta}|Y_{\text{obs}}, z)}{L(\boldsymbol{\theta}|Y_{\text{obs}})} = f(z|Y_{\text{obs}}, \boldsymbol{\theta})$$

so that the log-likelihood satisfies the following identity

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) - \ell(\boldsymbol{\theta}|Y_{\text{obs}}, z) = -\log\{f(z|Y_{\text{obs}}, \boldsymbol{\theta})\}. \tag{2.30}$$

— Integrating both sides of (2.30) with respect to $f(z|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$, we obtain

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = -H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \tag{2.31}$$

where the Q-function is defined by (2.17) and

$$\begin{aligned} H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \int_{\mathbb{Z}} \log\{f(z|Y_{\text{obs}},\boldsymbol{\theta})\} \times f(z|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})\,\mathrm{d}z \qquad (2.32)\\ &= E\Big[\log\{f(Z|Y_{\text{obs}},\boldsymbol{\theta})\}\Big|Y_{\text{obs}},\boldsymbol{\theta}^{(t)}\Big]. \end{aligned}$$

31.2• The non-negativity of the Kullback–Leibler divergence

— It follows that

$$\begin{aligned} & H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\\ \overset{(2.32)}{=}\ & \int_{\mathbb{Z}} \log\{f(z|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})\} \times f(z|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})\,\mathrm{d}z\\ & - \int_{\mathbb{Z}} \log\{f(z|Y_{\text{obs}},\boldsymbol{\theta})\} \times f(z|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})\,\mathrm{d}z\\ =\ & \int_{\mathbb{Z}} f(z|Y_{\text{obs}},\boldsymbol{\theta}^{(t)}) \log\left\{\frac{f(z|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})}{f(z|Y_{\text{obs}},\boldsymbol{\theta})}\right\}\mathrm{d}z\\ \overset{(2.62)}{=}\ & \mathrm{KL}\Big(f(z|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})\Big\|f(z|Y_{\text{obs}},\boldsymbol{\theta})\Big) \overset{(2.63)}{\geqslant} 0, \end{aligned}$$

and the equality holds iff $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

31.3• The essence of the EM algorithm

— Thus, for all $\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}$, we have

$$\begin{aligned} \ell(\boldsymbol{\theta}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &\overset{(2.31)}{=} -H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geqslant -H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})\\ &\overset{(2.31)}{=} \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}). \qquad (2.33) \end{aligned}$$

— In other words, $\ell(\boldsymbol{\theta}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ attains its minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

— Inequality (2.33) is the essence of the EM algorithm because if we choose $\boldsymbol{\theta}^{(t+1)}$ as the maximizer of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, then it follows that

$$\begin{aligned} & \ell(\boldsymbol{\theta}^{(t+1)}|Y_{\text{obs}}) - \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})\\ \overset{(2.33)}{\geqslant}\ & Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \geqslant 0. \qquad (2.34) \end{aligned}$$

— Namely, the EM algorithm holds the *ascent property* that increasing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ causes an increase in $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$.

32• Generalized EM

- Generally, a *generalized EM* (GEM) algorithm chooses $\boldsymbol{\theta}^{(t+1)}$ so that

$$Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geqslant Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}).$$

- From (2.34), a GEM algorithm also has the ascent property.

33• Convergence of the EM algorithm

- Suppose the a sequence of EM iterates $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ satisfies
  - $\left.\dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t+1)}} = \mathbf{0}$;
  - $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ converge to some value $\boldsymbol{\theta}^{*}$.
- Furthermore, let $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ be sufficiently smooth.
- Then, it follows that $\left.\dfrac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*}} = \mathbf{0}$.

33.1• Local convergence of the EM algorithm

— In other words, if the iterates $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ converge, they converge to a *stationary point of* $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$.

— This implies that when there are multiple stationary points (local maxima or saddle points), the EM algorithm may not converge to the global maximum.

— For more detail, see Wu (1983).

33.2• How to reach the global maximum?

— The global maximum can usually be reached by starting the parameters at good but *suboptimal* estimates such as moment estimates, or by choosing multiple starting points.

— In general, almost all algorithms have trouble distinguishing global from local maximum points.

2.3.3 Missing information principle and standard errors

34• Another identity similar to (2.31)

- Integrating both sides of (2.30) with respect to $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$, similar to (2.31), we have
$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}), \tag{2.35}$$
where $Q(\cdot|\cdot)$ and $H(\cdot|\cdot)$ are defined by (2.17) and (2.32), respectively.
- Differentiating (2.35) twice with respect to the left argument $\boldsymbol{\theta}$ yields
$$-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = -\nabla^{20} Q(\boldsymbol{\theta}|\boldsymbol{\theta}) + \nabla^{20} H(\boldsymbol{\theta}|\boldsymbol{\theta}), \tag{2.36}$$
where $\nabla^{ij} Q(x|y)$ means $\partial^{i+j} Q(x|y)/\partial x^i \partial y^j$.

34.1• Definition of three information matrices

— We call
$$\begin{aligned} \boldsymbol{I}_{\text{obs}} &= -\nabla^2 \ell(\hat{\boldsymbol{\theta}}|Y_{\text{obs}}) = \boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \\ \boldsymbol{I}_{\text{com}} &= -\nabla^{20} Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) \\ &= E\{-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}, Z)|Y_{\text{obs}}, \boldsymbol{\theta}\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad \text{and} \\ \boldsymbol{I}_{\text{mis}} &= -\nabla^{20} H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) \\ &= E\{-\nabla^2 \log f(Z|Y_{\text{obs}}, \boldsymbol{\theta})|Y_{\text{obs}}, \boldsymbol{\theta}\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \end{aligned} \tag{2.37}$$
the *observed information*, the *complete information* and the *missing information*, respectively.

34.2• Missing information principle

— If we evaluate the functions in (2.36) at the converged value $\hat{\boldsymbol{\theta}}$, then (2.36) becomes
$$\boldsymbol{I}_{\text{obs}} = \boldsymbol{I}_{\text{com}} - \boldsymbol{I}_{\text{mis}}. \tag{2.38}$$

— Namely, the observed information equals the complete information minus the missing information.

35• Calculation of the estimated standard errors

- Louis (1982) showed that

$$-\nabla^{20} H(\boldsymbol{\theta}|\boldsymbol{\theta}) = E[\{\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}, Z)\}^{\otimes 2}|Y_{\text{obs}}, \boldsymbol{\theta}] - \{\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}})\}^{\otimes 2},$$

  where $\boldsymbol{a}^{\otimes 2} \triangleq \boldsymbol{a}\boldsymbol{a}^\top$.

- Since $\nabla \ell(\hat{\boldsymbol{\theta}}|Y_{\text{obs}}) = \mathbf{0}$, the equation (2.38) becomes

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}}|Y_{\text{obs}}) = -\nabla^{20} Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - E[\{\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}, Z)\}^{\otimes 2}|Y_{\text{obs}}, \boldsymbol{\theta}]|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \quad (2.39)$$

- The estimated standard errors are the square roots of the diagonal elements for the inverse information matrix; i.e., $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}|Y_{\text{obs}})$.

Example 2.9 (Two-parameter multinomial model revisited). In Example 2.7, we derived the MLEs of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top$ by using an EM algorithm. Use (2.39) to calculate the estimated standard errors.

Solution: From (2.22), the complete-data log-likelihood function is

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}) = (z_1 + z_2)\log\theta_1 + (z_3 + z_4)\log\theta_2 + y_5 \log\theta_3.$$

It is easy to obtain

$$\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}) = \begin{pmatrix} \dfrac{z_1 + z_2}{\theta_1} \\ \dfrac{z_3 + z_4}{\theta_2} \end{pmatrix} - \frac{y_5}{\theta_3}\mathbf{1}_2 \quad \text{and}$$

$$-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}) = \begin{pmatrix} \dfrac{z_1 + z_2}{\theta_1^2} & 0 \\ 0 & \dfrac{z_3 + z_4}{\theta_2^2} \end{pmatrix} + \frac{y_5}{\theta_3^2}\mathbf{1}_2\mathbf{1}_2^\top.$$

From (2.37), we have

$$\begin{aligned} -\nabla^{20} Q(\boldsymbol{\theta}|\boldsymbol{\theta}) &= E\{-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}, \mathbf{z})|Y_{\text{obs}}, \boldsymbol{\theta}\} \\ &= \begin{pmatrix} \dfrac{z_1^* + z_2^*}{\theta_1^2} & 0 \\ 0 & \dfrac{z_3^* + z_4^*}{\theta_2^2} \end{pmatrix} + \frac{y_5}{\theta_3^2}\mathbf{1}_2\mathbf{1}_2^\top, \end{aligned}$$

where $z_i^* = E(Z_i|Y_{\text{obs}}, \boldsymbol{\theta}) = y_i p_i$ with p_i being defined by (2.24). Substituting the data set given by (2.26) and

$$\hat{\boldsymbol{\theta}} = (0.585900,\ 0.0716178,\ 0.342482)^\top$$

in these expressions yields

$$-\nabla^{20} Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) = \begin{pmatrix} 67.5457 & 42.628 \\ 42.6280 & 246.478 \end{pmatrix}.$$

Similarly, we have

$$E[\{\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}, \mathbf{z})\}^{\otimes 2}|Y_{\text{obs}}, \boldsymbol{\theta}] = \frac{y_5^2}{\theta_3^2}\mathbf{1}_2\mathbf{1}_2^\top + \begin{pmatrix} \frac{z_{12}^*}{\theta_1^2} - 2y_5\frac{z_1^*+z_2^*}{\theta_1\theta_3}, & \frac{(z_1^*+z_2^*)(z_3^*+z_4^*)}{\theta_1\theta_2} - \frac{y_5}{\theta_3}\left(\frac{z_1^*+z_2^*}{\theta_1} + \frac{z_3^*+z_4^*}{\theta_2}\right) \\ * & \frac{z_{34}^*}{\theta_2^2} - 2y_5\frac{z_3^*+z_4^*}{\theta_2\theta_3} \end{pmatrix},$$

where

$$\begin{aligned} z_{12}^* &\hat{=} E\{(Z_1+Z_2)^2|Y_{\text{obs}}, \boldsymbol{\theta}\}, \\ z_{34}^* &\hat{=} E\{(Z_3+Z_4)^2|Y_{\text{obs}}, \boldsymbol{\theta}\} \quad \text{and} \\ E(Z_i^2|Y_{\text{obs}}, \boldsymbol{\theta}) &= y_i p_i (1 - p_i + y_i p_i). \end{aligned}$$

Numerically, we obtain

$$E[\{\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}, \mathbf{z})\}^{\otimes 2}|Y_{\text{obs}}, \boldsymbol{\theta}]|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 10.1320 & 0 \\ 0 & 8.47962 \end{pmatrix}.$$

Hence, from (2.39), the observed information matrix is

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}}|Y_{\text{obs}}) = \begin{pmatrix} 57.4137 & 42.628 \\ 42.6280 & 237.998 \end{pmatrix},$$

and the estimated standard errors are $\widehat{\text{Se}}(\hat{\theta}_1) = 0.1417$ and $\widehat{\text{Se}}(\hat{\theta}_2) = 0.0696$. Using the delta method (see Exercise 2.2), we obtain $\widehat{\text{Se}}(\hat{\theta}_3) = 0.1332$.

Comment 2: The above method of estimating standard errors is so tedious, do we have any alternatives? Yes, we have. Please see Chapter 5: Bootstrap Methods. ‖

36• Advantages of the em algorithm

- The first advantage of the EM algorithm is its numerical stability; i.e., it increases the observed-data likelihood function at each iteration.
- The second advantage of the EM algorithm is that it sufficiently uses the structure of complete data such that both the E- and M-steps have closed–form expressions in many problems.

37• Limitations of the em algorithm

- The EM holds the ascent property at the cost of slow convergence; i.e., the linear convergence rate when comparing to the quadratic convergence rate associated with the NR algorithm.
- For some problems, the E-step has no explicit expressions.
- Sometimes, the M-step may be difficult to achieve.

37.1• How to overcome/alleviate these limitations with the EM?

— For the problem of slow convergence, we could speed EM-type algorithms via a so-called "working parameter" method (Meng & van Dyk 1997).

— For the problem of difficult E-step, we could use the Monte Carlo EM algorithm (Wei & Tanner 1990).

— For the problem of difficult M-step, we could use the ECM algorithm (Meng & Rubin 1993).

2.4 The ECM Algorithm

38• What is the ecm algorithm?

- The *expectation/conditional maximization* (ECM) algorithm replaces a complicated M-step with several computationally simpler conditional maximization steps.
- As a consequence, it typically converges more slowly than the EM algorithm in terms of number of iterations, but can be faster in total computer time.

- Importantly, the ECM algorithm preserves the monotone convergence property of the EM algorithm.

39• Basic idea of the ECM algorithm

- More precisely, the ECM replaces each M-step of the EM by a sequence of K conditional maximization steps; i.e., CM-steps.

- Each CM-step maximizes the Q-function defined in (2.17) over $\boldsymbol{\theta}$ but with some vector function of $\boldsymbol{\theta}$, $g_k(\boldsymbol{\theta})$ $(k = 1, \ldots, K)$ fixed at its previous value.

- For example, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, where $\boldsymbol{\theta}_1 = (\theta_1, \ldots, \theta_r)^\top$ and $\boldsymbol{\theta}_2 = (\theta_{r+1}, \ldots, \theta_q)^\top$.

- The M-step of calculating $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta} | Y_{\text{com}})$ is replaced by

  — CM-Step 1: Given $\boldsymbol{\theta}_2^{(t)}$, to calculate

$$\boldsymbol{\theta}_1^{(t+1)} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell \left(\begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2^{(t)} \end{pmatrix} \middle| Y_{\text{com}} \right), \quad \text{and}$$

  — CM-Step 2: Given $\boldsymbol{\theta}_1^{(t+1)}$, to calculate

$$\boldsymbol{\theta}_2^{(t+1)} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell \left(\begin{pmatrix} \boldsymbol{\theta}_1^{(t+1)} \\ \boldsymbol{\theta}_2 \end{pmatrix} \middle| Y_{\text{com}} \right).$$

39.1• Aim of Example 2.10

— Example 2.10 below illustrates the ECM algorithm in a simple but rather general model.

— In this model, we partition the parameters into a parameter vector of regression coefficients and a covariance matrix, leading to a simple ECM with two CM-steps.

— Each CM-step involves closed–form solutions while holding the other fixed.

Example 2.10 (A multivariate normal regression model with missing data). Suppose that we have n independent observations from the following m-dimensional normal distribution:

$$\mathbf{y}_j = (Y_{1j}, \ldots, Y_{mj})^\top \sim N_m(\boldsymbol{X}_j\boldsymbol{\beta}, \boldsymbol{\Sigma}), \quad j = 1, \ldots, n,$$

where $\boldsymbol{X}_j$ is a known $m \times p$ design matrix for the j-th observation, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{\Sigma}$ is an unknown $m \times m$ covariance matrix. Let $Y_{\text{com}} = \{\boldsymbol{y}_j\}_{j=1}^n$ be the complete data, where $\boldsymbol{y}_j$ is the realization of $\mathbf{y}_j$. Assume that some components in $\boldsymbol{y}_j$ are missing, use the ECM algorithm to calculate the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

Solution: Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$, the complete-data likelihood function is

$$L(\boldsymbol{\theta}|Y_{\text{com}}) = \prod_{j=1}^n \frac{1}{(2\pi)^{\frac{m}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})\right\},$$

so that the complete-data log-likelihood function is

$$\begin{aligned}
\ell(\boldsymbol{\theta}|Y_{\text{com}}) &= c - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{j=1}^n (\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}) \\
&= c - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{j=1}^n \boldsymbol{y}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_j \\
&\quad - \frac{1}{2}\sum_{j=1}^n \left(-2\boldsymbol{y}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_j\boldsymbol{\beta}\right),
\end{aligned}$$

where $c = -(mn/2)\log(2\pi)$ is a constant. By using Exercise 2.1, we have

$$\frac{\partial\ell(\boldsymbol{\theta}|Y_{\text{com}})}{\partial\boldsymbol{\beta}} = \left(\sum_{j=1}^n \boldsymbol{X}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_j\right) - \left(\sum_{j=1}^n \boldsymbol{X}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_j\right)\boldsymbol{\beta}.$$

If $\boldsymbol{\Sigma}$ was given, say $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(t)}$, then the conditional MLEs of $\boldsymbol{\beta}$ would be simply the weighted least–squares estimate:

$$\boldsymbol{\beta}^{(t+1)} = \left\{\sum_{j=1}^n \boldsymbol{X}_j^\top(\boldsymbol{\Sigma}^{(t)})^{-1}\boldsymbol{X}_j\right\}^{-1}\left\{\sum_{j=1}^n \boldsymbol{X}_j^\top(\boldsymbol{\Sigma}^{(t)})^{-1}\boldsymbol{y}_j\right\}. \tag{2.40}$$

On the other hand, from $\operatorname{tr}(\boldsymbol{A}_1\boldsymbol{A}_2) = \operatorname{tr}(\boldsymbol{A}_2\boldsymbol{A}_1)$, we can rewrite

$$\begin{aligned}\ell(\boldsymbol{\theta}|Y_{\text{com}}) &= c - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})^\top\right\}\\ &= c + \frac{n}{2}\log(|\boldsymbol{\Sigma}^{-1}|) - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{B})\\ &= c_1(\boldsymbol{B}) + \log(|\boldsymbol{\Sigma}^{-1}\boldsymbol{B}|)^{\frac{n}{2}} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}),\end{aligned}$$

where $c_1(\boldsymbol{B})$ is a function of

$$\boldsymbol{B} \doteq \sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})^\top.$$

Given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t+1)}$, from Exercise 2.14, then the conditional MLEs of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma}^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}^{(t+1)})(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}^{(t+1)})^\top. \tag{2.41}$$

Without loss of generality, let

$$\mathbf{y}_j = \begin{pmatrix}\mathbf{y}_{j,\,\text{obs}}\\ \mathbf{y}_{j,\,\text{mis}}\end{pmatrix} \quad \text{and} \quad \boldsymbol{y}_j = \begin{pmatrix}\boldsymbol{y}_{j,\,\text{obs}}\\ \boldsymbol{y}_{j,\,\text{mis}}\end{pmatrix}, \quad j = 1, \ldots, n.$$

The E-step of the ECM algorithm is to compute

$$\begin{aligned}E(\mathbf{y}_j|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}) &= \begin{pmatrix}\boldsymbol{y}_{j,\,\text{obs}}\\ E(\mathbf{y}_{j,\,\text{mis}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})\end{pmatrix} \quad \text{and}\\ E(\mathbf{y}_j\mathbf{y}_j^\top|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}) &= \begin{pmatrix}\boldsymbol{y}_{j,\,\text{obs}}\boldsymbol{y}_{j,\,\text{obs}}^\top & \boldsymbol{y}_{j,\,\text{obs}}E(\mathbf{y}_{j,\,\text{mis}}^\top|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})\\ * & E(\mathbf{y}_{j,\,\text{mis}}\mathbf{y}_{j,\,\text{mis}}^\top|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})\end{pmatrix},\end{aligned}$$

where $\boldsymbol{\theta}^{(t)} = \{\boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)}\}$. The two CM-steps are as follows.

CM-step 1: Calculate $\boldsymbol{\beta}^{(t+1)}$ by using (2.40), where $\boldsymbol{y}_j$ is replaced with its conditional expectation $E(\mathbf{y}_j|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$.

CM-step 2: Calculate $\boldsymbol{\Sigma}^{(t+1)}$ via (2.41) where $\boldsymbol{y}_j$ and $\boldsymbol{y}_j\boldsymbol{y}_j^\top$ being replaced by $E(\mathbf{y}_j|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$ and $E(\mathbf{y}_j\mathbf{y}_j^\top|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$, respectively. ‖

39.2• Aim of Example 2.11

— Example 2.11 below illustrates that even if some CM-steps do not have analytical solutions, ECM may still be computationally simpler and more stable because it only involves lower–dimensional maximization than EM.

Example 2.11 (A gamma model with incomplete data). Let

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$$

(see Appendix A.2.3) and we are interested in finding the MLEs of α and β based on the observed data $Y_{\text{obs}} = \{y_1, \ldots, y_r; c_{r+1}, \ldots, c_n\}$. The first r components of $\boldsymbol{y} = (y_1, \ldots, y_r, y_{r+1}, \ldots, y_n)^\top$ are completely observed and the rest $n - r$ are left censoring (i.e., $y_i = z_i < c_i$ for $i = r + 1, \ldots, n$), so that the complete data are $Y_{\text{com}} = \{y_i\}_{i=1}^n$.

Solution: The complete-data likelihood function is

$$L(\alpha, \beta | Y_{\text{com}}) = \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} \, \mathrm{e}^{-\beta y_i},$$

so that the complete-data log-likelihood function is

$$\ell(\alpha, \beta | Y_{\text{com}}) = (\alpha - 1) \left(\sum_{i=1}^n \log y_i \right) - n\bar{y}\beta + n\{\alpha \log \beta - \log \Gamma(\alpha)\},$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$. The E-step of the ECM computes

$$E(Y_i | Y_{\text{obs}}, \alpha^{(t)}, \beta^{(t)}) \quad \text{and} \quad E(\log Y_i | Y_{\text{obs}}, \alpha^{(t)}, \beta^{(t)})$$

for $i = r + 1, \ldots, n$. Given $\alpha = \alpha^{(t)}$, the first CM-step is to calculate the conditional MLE of β as

$$\beta^{(t+1)} = \alpha^{(t)} / \bar{y}.$$

On the other hand, given $\beta = \beta^{(t+1)}$, the second CM-step is to find the conditional MLE of α, $\alpha^{(t+1)}$, which satisfies the following equation

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{\mathrm{d}\Gamma(\alpha)/\mathrm{d}\alpha}{\Gamma(\alpha)} = \frac{1}{n} \left(\sum_{i=1}^n \log y_i \right) + \log \beta^{(t+1)}. \tag{2.42}$$

Although (2.42) does not provide an analytic solution for $\alpha^{(t+1)}$, a value can be easily obtained by a one–dimensional NR algorithm. ‖

2.5 Minorization–Maximization (MM) Algorithms

40• Why do we need mm algorithms?

- Although EM-type algorithms (Dempster *et al.* 1977; Meng & Rubin 1993) possess the ascent property that ensures monotone convergence, they may not be applied to generalized linear models (e.g., logistic regression, log-linear models) and Cox proportional models due to the absence of a missing-data structure in these models.

- Therefore, for problems in which the missing-data structure does not exist or is not readily available, *minorization–maximization* (MM) algorithms (Lange *et al.* 2000; Hunter & Lange 2004) are often useful alternatives. In the sequel, MM refers to a class of algorithms.

2.5.1 A brief review of MM algorithms

41• Literature review

- According to Hunter & Lange (2004), the general principle of MM algorithms is first enunciated by numerical analysts Ortega & Rheinboldt (1970, p.253–255) in the context of linear search methods.

- De Leeuw & Heiser (1977) presented an MM algorithm for multidimensional scaling in parallel to the classic Dempster *et al.* (1977) article on EM algorithms.

- Although the work of De Leeuw & Heiser did not draw the same degree of attention from the statistical community as did the Dempster *et al.* (1977) article, development of MM algorithms has continued.

- The MM principle reappears, among other places, in robust regression (Huber 1981), in correspondence analysis (Heiser 1987), in the quadratic lower–bound principle of Böhning & Lindsay (1988), in the psychometrics literature on least squares (Bijleveld & De Leeuw 1991; Kiers & Ten Berge 1992), and in medical imaging (De Pierro 1995; Lange & Fessler 1995).

- The survey articles of De Leeuw (1994), Heiser (1995), Becker *et al.* (1997), and Lange *et al.* (2000) deal with the general principle, but

it is not until the rejoinder of Hunter & Lange (2000) that the acronym MM first appears.

- This acronym pays homage to the earlier names "majorization" and "iterative majorization" of the MM principle, emphasizes its crucial link to the well–known EM principle, and diminishes the possibility of confusion with the distinct subject in mathematics known as majorization (Marshall & Olkin 1979).

2.5.2 The MM idea

42• Minorization and optimal lower–bound

- Let $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ denote the log-likelihood function and we want to find the MLEs $\hat{\boldsymbol{\theta}}$ defined by (2.12).
- Let $\boldsymbol{\theta}^{(t)}$ represent the t-th approximation to $\hat{\boldsymbol{\theta}}$ and $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ a real–valued function of $\boldsymbol{\theta}$ whose form depends on $\boldsymbol{\theta}^{(t)}$.
- The function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is said to *minorize* $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ if

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \leqslant \ell(\boldsymbol{\theta}|Y_{\text{obs}}) \quad \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}, \quad \text{and} \tag{2.43}$$

$$Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}). \tag{2.44}$$

- Sometimes, the $Q(\cdot|\boldsymbol{\theta}^{(t)})$ function satisfying both (2.43) and (2.44) is also called the *optimal lower–bound* (OLB) of $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ with the tangent point being $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

42.1• The ascent property

— In an MM algorithm, we maximize the *minorizing* function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ instead of the target function $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$.

— If $\boldsymbol{\theta}^{(t+1)}$ is the maximizer of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$; i.e.,

$$\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \tag{2.45}$$

then from (2.43) and (2.44), we have

$$\begin{aligned} \ell(\boldsymbol{\theta}^{(t+1)}|Y_{\text{obs}}) &\geqslant Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \\ &\geqslant Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}). \end{aligned} \tag{2.46}$$

— Consequently, an increase in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ forces $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ uphill.

42.2• The monotone convergence

— Under appropriate conditions of additional compactness and continuity, the ascent property (2.46) guarantees convergence of the MM algorithm, and lends the algorithm monotone convergence.

42.3• The generalized MM algorithm

— From (2.46), it is clear that it is not necessary to actually maximize the minorizing function.

— It suffices to find $\boldsymbol{\theta}^{(t+1)}$ such that

$$Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geqslant Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}).$$

43• EM is a special case of MM algorithms

- From (2.33), we can have an important conclusion: EM is a special case of MM.
- The inequality (2.33) can be rewritten as

$$\begin{aligned} \ell(\boldsymbol{\theta}|Y_{\text{obs}}) &\geqslant Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \\ &\triangleq Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \quad \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}. \end{aligned}$$

43.1• Further remarks on why EM is a special MM

— $\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})$ and $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ are constant, independent of $\boldsymbol{\theta}$.

— Given $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ defined by (2.17), we immediately have $Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

— It is easy to verify that $Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ minorizes $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

— We have

$$\begin{aligned} \boldsymbol{\theta}^{(t+1)} &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \end{aligned}$$

2.5.3 The quadratic lower–bound algorithm

44• Key idea for the QLB algorithm

- Let $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ denote the log-likelihood function and we want to find the MLEs $\hat{\boldsymbol{\theta}}$ defined by (2.12).
- The *quadratic lower–bound* (QLB) algorithm developed by Böhning & Lindsay (1988) is a special MM algorithm.
- The key idea is to transfer the optimization from the intractable $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ to a quadratic *surrogate* function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

44.1• A key condition for the QLB algorithm

— A key for the QLB is to find a positive definite matrix $\boldsymbol{B}\,(>0)$ that does not depend on $\boldsymbol{\theta}$ such that

$$\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) + \boldsymbol{B} \geqslant 0 \qquad \forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{2.47}$$

— In other words, $\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) + \boldsymbol{B}$ is a non-negative definite matrix.

44.2• Constructing a surrogate function Q

— The second–order Taylor expansion of $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ around $\boldsymbol{\theta}^{(t)}$ is

$$\begin{aligned}\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + (\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top\nabla^2\ell(\boldsymbol{\theta}^*|Y_{\text{obs}})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})\end{aligned}$$

for some $\boldsymbol{\theta}^* = \alpha\boldsymbol{\theta} + (1-\alpha)\boldsymbol{\theta}^{(t)}$ with $\alpha \in [0,1]$.

— Replacing $\nabla^2\ell(\boldsymbol{\theta}^*|Y_{\text{obs}})$ above by $-\boldsymbol{B}$, we can construct a quadratic surrogate function

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + (\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) \\ &\quad - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top\boldsymbol{B}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}), \quad \boldsymbol{\theta},\boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}.\end{aligned} \tag{2.48}$$

— It is easy to verify that this Q-function minorizes $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ (see Exercise 2.6).

44.3• Definition of the QLB algorithm

— The QLB algorithm is defined by (2.45); i.e.,

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{B}^{-1}\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}). \tag{2.49}$$

— Just like an EM algorithm, the QLB algorithm also holds the ascent property (2.46), leading to monotone convergence.

— The algorithm has the advantage of requiring a single matrix inversion instead of repeated matrix inversions as in the NR algorithm.

Example 2.12 (Logistic regression model). In §2.2.5, we have applied the NR algorithm to logistic regression, however it does not work for the cancer remission data in Example 2.5. Perform a QLB algorithm for the logistic regression.

A simple inequality: Let $p \in (0,1)$, then $p(1-p) \leqslant 1/4$.

Proof: From $(\sqrt{p} - \sqrt{1-p})^2 \geqslant 0$, we have $p + (1-p) - 2\sqrt{p(1-p)} \geqslant 0$. That is,

$$2\sqrt{p(1-p)} \leqslant 1 \quad \text{or} \quad p(1-p) \leqslant 1/4. \qquad \square$$

Solution: The key to the application of the QLB algorithm is to find a positive definite matrix $\boldsymbol{B}$ satisfying the condition (2.47). Since $p_i(1-p_i) \leqslant 1/4$, from (2.15), we have

$$\boldsymbol{B} \triangleq \frac{1}{4}\boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{X} \geqslant -\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}).$$

From (2.49), we obtain

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + 4(\boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{p}^{(t)}), \tag{2.50}$$

where

$$p_i^{(t)} = \frac{\exp\left\{\boldsymbol{x}_{(i)}^\top\boldsymbol{\theta}^{(t)}\right\}}{1 + \exp\left\{\boldsymbol{x}_{(i)}^\top\boldsymbol{\theta}^{(t)}\right\}}$$

is the i-th component of $\boldsymbol{p}^{(t)}$. Applying the QLB algorithm (2.50) to the cancer remission data in Example 2.5, we obtain

$$\hat{\boldsymbol{\theta}} = (58.039, 24.661, 19.293, -19.601, 3.896, 0.151, -87.434)^\top.$$

after 1500 iterations starting from $\boldsymbol{\theta}^{(0)} = \mathbf{1}_7$. $\|$

Example 2.13 (Cox's proportional hazards model). Cox's regression (Cox 1972) is a semi-parametric approach to survival analysis. The hazard function is

$$h(t) = h_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\theta}),$$

where $h_0(t)$ is the baseline hazard function and can be viewed as nuisance parameters, $\boldsymbol{x}$ a vector of covariates and $\boldsymbol{\theta}_{q\times 1}$ regression parameters. Suppose that among a total of N subjects there are m ordered *distinct* survival times $\{y_{(i)}\}_{i=1}^m$ (i.e., there are no ties in the data) and $N-m$ right censored survival times. We also assume that censoring is non-informative in the sense that inferences do not depend on the censoring process. Let $\mathbb{R}_i$ denote the risk set at $y_{(i)}$ so that $\mathbb{R}_i$ is the set of individuals who are event–free and uncensored at a time just prior to $y_{(i)}$. The parameter vector $\boldsymbol{\theta}$ are estimated by maximizing the partial log-likelihood function

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^m \log\left\{\frac{\exp(\boldsymbol{x}_{(i)}^\top\boldsymbol{\theta})}{\sum_{j\in\mathbb{R}_i}\exp(\boldsymbol{x}_j^\top\boldsymbol{\theta})}\right\},$$

where $\boldsymbol{x}_{(i)}$ denotes the vector of covariates for individual who has an event at $y_{(i)}$. Derive the QLB algorithm to calculate the MLEs of $\boldsymbol{\theta}$.

Solution: The score vector and the observed information matrix are

$$\begin{aligned}
\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \sum_{i=1}^m\left(\boldsymbol{x}_{(i)} - \sum_{j\in\mathbb{R}_i} p_{ij}\boldsymbol{x}_j\right) \quad \text{and}\\
-\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \sum_{i=1}^m\left\{\sum_{j\in\mathbb{R}_i} p_{ij}\boldsymbol{x}_j\boldsymbol{x}_j^\top - \left(\sum_{j\in\mathbb{R}_i} p_{ij}\boldsymbol{x}_j\right)^{\otimes 2}\right\},
\end{aligned}$$

respectively, where

$$p_{ij} = \frac{\exp(\boldsymbol{x}_j^\top\boldsymbol{\theta})}{\sum_{k\in\mathbb{R}_i}\exp(\boldsymbol{x}_k^\top\boldsymbol{\theta})}, \qquad j\in\mathbb{R}_i.$$

Note that, for given i, $\{p_{ij}\}$ is a probability distribution with support points in the set $\{\boldsymbol{x}_j\colon\, j\in\mathbb{R}_i\}$. Let $\mathbf{u}_i$ denote a random vector taking the value $\boldsymbol{x}_j$ with probability p_{ij}, then we have

$$\begin{aligned}
E(\mathbf{u}_i) &= \sum_{j\in\mathbb{R}_i} p_{ij}\boldsymbol{x}_j \quad \text{and}\\
\text{Var}(\mathbf{u}_i) &= \sum_{j\in\mathbb{R}_i} p_{ij}\{x_j - E(\mathbf{u}_i)\}\{x_j - E(\mathbf{u}_i)\}^\top.
\end{aligned}$$

Thus, $-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_i \text{Var}(\mathbf{u}_i)$ is positive definite for all $\boldsymbol{\theta}$, which implies the concavity of the log-likelihood $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$. By Exercise 2.7, we have

$$\begin{aligned} -\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) &\leqslant \sum_{i=1}^{m} \frac{1}{2}\left\{ \sum_{j\in\mathbb{R}_i} \boldsymbol{x}_j \boldsymbol{x}_j^\top - \frac{1}{n_i}\left(\sum_{j\in\mathbb{R}_i} \boldsymbol{x}_j \right)^{\otimes 2} \right\} \\ &\hat{=} \boldsymbol{B}, \quad \text{for any } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \end{aligned}$$

where n_i denotes the number of individuals in $\mathbb{R}_i$. Thus, the QLB algorithm can be applied to find the MLEs of $\boldsymbol{\theta}$. ‖

45• Applicability of the QLB algorithm

- Although the QLB algorithm is elegant, its applicability depends on the *existence* of a positive definite $\boldsymbol{B}$ satisfying (2.47).
- Up to now, such a $\boldsymbol{B}$ matrix is already identified only for the logistic regression model (Example 2.12), Cox's proportional hazards models (Example 2.13), the multinomial logistic regression model (Exercise 2.8), and the probit regression model:

$$Y_i \overset{\text{ind}}{\sim} \text{Binomial}(n_i, p_i), \quad i = 1, \ldots, m,$$

  where $p_i = \Phi(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta})$ and $\Phi(\cdot)$ is the cdf of $N(0,1)$.

2.5.4 The De Pierro algorithm

46• Why do we need the DP algorithm

- In examples such as Poisson regression models (or log-linear models), the QLB algorithm fails because of the absence of such a matrix $\boldsymbol{B}$.
- The De Pierro (DP) algorithm originally proposed by De Pierro (1995) is also a special member of MM algorithms.

46.1• Application range of the DP algorithm

— The normal linear model is assumed to be

$$Y_i \overset{\text{ind}}{\sim} N(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}, \sigma^2) \quad \text{or} \quad E(Y_i) = \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}, \quad i = 1, \ldots, m.$$

— *Generalized linear models* (GLMs) extend the normal linear model from two parts: Y_i may come from a distribution in the exponential family other than the normal distribution; and the link function $g(\cdot)$ may be any monotonic differentiable function in the form of

$$g(E(Y_i)) = \boldsymbol{x}_{(i)}^{\top}\boldsymbol{\theta}.$$

— The exponential family includes binomial, Poisson, exponential, gamma, inverse Gaussian, normal distributions, and so on.

— The DP algorithm can be applied to GLMs.

46.2• Main idea of the DP algorithm

— The main idea of the DP algorithm is to transfer the optimization of a high–dimensional function $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ to the optimization of a low dimensional *surrogate* function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in the sense that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is a sum of linear combinations of a set of one–dimensional concave functions.

— Maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ can be implemented via the one–step Newton–Raphson method.

47• Basic assumptions for the DP algorithm

- Let the log-likelihood function be of the form

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}_{(i)}^{\top}\boldsymbol{\theta}), \tag{2.51}$$

  where $\boldsymbol{x}_{(i)}$ is the covariate vector for individual i,

$$\boldsymbol{X}_{m\times q} = \begin{pmatrix} \boldsymbol{x}_{(1)}^{\top} \\ \vdots \\ \boldsymbol{x}_{(i)}^{\top} \\ \vdots \\ \boldsymbol{x}_{(m)}^{\top} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1q} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{iq} \\ \vdots & & \vdots & & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mq} \end{pmatrix}$$

  is the design matrix, and $\boldsymbol{\theta}$ is the parameter vector of interest.

- Let $\{f_i(\cdot)\}_{i=1}^{m}$ be twice continuously differentiable and *strictly concave* functions (i.e., $f_i''(\cdot) < 0$) defined in the real line $\mathbb{R}$.

47.1• Score vector and observed information matrix

— From (2.51), the score vector and the observed information matrix are

$$\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \frac{\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{m} f_i'(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) \boldsymbol{x}_{(i)} \quad \text{and}$$

$$-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = -\frac{\partial\{\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}})\}}{\partial \boldsymbol{\theta}^\top} = \sum_{i=1}^{m} \Big\{ - f_i''(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) \Big\} \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top,$$

where

$$f_i'(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) \doteq f_i'(\eta)\Big|_{\eta = \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}} = \frac{\mathrm{d} f_i(\eta)}{\mathrm{d}\eta}\Big|_{\eta = \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}}.$$

— Thus, $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is strictly concave provided that at least one $f_i''(\cdot) < 0$.

47.2• Choice of the weight vectors

— To construct a surrogate function, we first define two index sets:

$$\mathbb{J}_i = \{j\colon x_{ij} \neq 0\}, \quad 1 \leqslant i \leqslant m \quad \text{and}$$

$$\mathbb{I}_j = \{i\colon x_{ij} \neq 0\}, \quad 1 \leqslant j \leqslant q.$$

— In other words, $\mathbb{J}_i$ is a set of column–subscripts of non-zero elements in the i-th row of $\boldsymbol{X}$, while $\mathbb{I}_j$ is a set of row–subscripts of non-zero elements in the j-th column of $\boldsymbol{X}$.

— Furthermore, for a given $r \in \mathbb{R}_+ = \{r\colon r \geqslant 0\}$ and a fixed i, we define weights

$$\lambda_{ij}(r) = \frac{|x_{ij}|^r}{\sum_{j' \in \mathbb{J}_i} |x_{ij'}|^r}, \quad j \in \mathbb{J}_i.$$

47.3• Some remarks on weights $\{\lambda_{ij}\}$

— Obviously,

$$\lambda_{ij}(r) > 0 \quad \text{and} \quad \sum_{j \in \mathbb{J}_i} \lambda_{ij}(r) = 1.$$

— It can be shown that when $r = 1$, the corresponding MM algorithm converges most quickly.

— Note that when $r = 0$, $\lambda_{ij}(0) = 1/n_i$, where n_i denotes the number of elements in $\mathbb{J}_i$.

48• Construction of a surrogate function

- Now, we can construct a surrogate function as follows:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{m} \sum_{j \in \mathbb{J}_i} \lambda_{ij} f_i\Big(\lambda_{ij}^{-1} x_{ij}(\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)}\Big), \tag{2.52}$$

  for $\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}$, where $\lambda_{ij} \hat{=} \lambda_{ij}(1)$.

- It can be shown that this $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ function minorizes $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ (see Exercise 2.9).

48.1• Definition of an additive separable function

— Suppose that g is a function of n variables $z_1, \ldots, z_n$.

— The function g is said to be *additive separable* if there exist functions $g_1, \ldots, g_n$, each g_k is of one variable, such that

$$g(z_1, \ldots, z_n) = g_1(z_1) + \cdots + g_n(z_n).$$

48.2• Understanding the Q-function

— Note that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ defined in (2.52) is an additive separable function with the following expression

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{j=1}^{q} \sum_{i \in \mathbb{I}_j} \lambda_{ij} f_i\Big(\lambda_{ij}^{-1} x_{ij}(\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)}\Big) \\ &= \sum_{j=1}^{q} Q_j(\theta_j|\boldsymbol{\theta}^{(t)}), \end{aligned}$$

where

$$Q_j(\theta_j|\boldsymbol{\theta}^{(t)}) = \sum_{i \in \mathbb{I}_j} \lambda_{ij} f_i\Big(\lambda_{ij}^{-1} x_{ij}(\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)}\Big)$$

is a linear combination of $\#\{\mathbb{I}_j\}$ concave functions

$$\Big\{ f_i\Big(\lambda_{ij}^{-1} x_{ij}(\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^{\top} \boldsymbol{\theta}^{(t)}\Big) \Big\}_{i \in \mathbb{I}_j}.$$

— In this sense, essentially, Q is a one–dimensional function, which is much easier to optimize.

49• Definition of the DP algorithm

- The DP algorithm is defined by

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

- The $(t+1)$-th iteration can be obtained as the solution to the system of equations: For $j = 1, \ldots, q$,

$$\begin{aligned} 0 &= \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\theta_j} = \frac{\partial Q_j(\theta_j|\boldsymbol{\theta}^{(t)})}{\partial\theta_j} \\ &= \sum_{i\in\mathbb{I}_j} f_i'\Big(\lambda_{ij}^{-1}x_{ij}(\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^\top\boldsymbol{\theta}^{(t)}\Big)\, x_{ij}, \end{aligned} \tag{2.53}$$

- When (2.53) cannot be solved explicitly, the one–step NR algorithm provides an approximate solution

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \tau_j^2(\boldsymbol{\theta}^{(t)}) \sum_{i\in\mathbb{I}_j} f_i'(\boldsymbol{x}_{(i)}^\top\boldsymbol{\theta}^{(t)})\, x_{ij}, \quad j = 1, \ldots, q, \tag{2.54}$$

where

$$\tau_j^2(\boldsymbol{\theta}) = \left[\sum_{i\in\mathbb{I}_j}\{-f_i''(\boldsymbol{x}_{(i)}^\top\boldsymbol{\theta})\}\, x_{ij}^2/\lambda_{ij}\right]^{-1}. \tag{2.55}$$

Example 2.14 (Exponential family distribution). Consider a GLM with canonical link function. Let $Y_{\text{obs}} = \{y_i\}_{i=1}^m$ be m independent observations from the exponential family distribution (McCullagh & Nelder 1989):

$$f_e(y;\psi) = \exp\left\{\frac{y\psi - b(\psi)}{a(\gamma)} + c(y,\gamma)\right\}, \tag{2.56}$$

where ψ denotes the canonical or natural parameter, $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are known real–valued functions, γ is the dispersion parameter and we assume that γ is known. For index i, let $\psi_i = \boldsymbol{x}_{(i)}^\top\boldsymbol{\theta}$, where $\boldsymbol{x}_{(i)}$ is the covariate vector and $\boldsymbol{\theta}$ is the parameter vector. Derive the DP algorithm to calculate the MLEs of $\boldsymbol{\theta}$.

Member 1: For the $N(\mu, \sigma^2)$ distribution, its pdf can be written as

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

so that $\psi = \mu$, $\gamma = \sigma^2$, and $a(\gamma) = \sigma^2$,

$$b(\psi) = \frac{\mu^2}{2} \quad \text{and} \quad c(y, \gamma) = -\frac{y^2/\sigma^2 + \log(2\pi\sigma^2)}{2}.$$

Member 2: For the Binomial(n, p) distribution, its pmf can be written as

$$\binom{n}{y} p^y (1-p)^{n-y} = \exp\left\{y \log\left(\frac{p}{1-p}\right) + n \log(1-p) + \log\binom{n}{y}\right\},$$

so that $\psi = \log\{p/(1-p)\}$, $\gamma = 1$, and $a(\gamma) = 1$,

$$b(\psi) = -n \log(1-p) = n \log(1 + \mathrm{e}^{\psi}) \quad \text{and} \quad c(y, \gamma) = \log\binom{n}{y}.$$

Member 3: For the Poisson(λ) distribution, its pmf can be written as

$$\frac{\lambda^y \mathrm{e}^{-\lambda}}{y!} = \exp\{y \log(\lambda) - \lambda - \log(y!)\},$$

so that $\psi = \log(\lambda)$, $\gamma = 1$, and $a(\gamma) = 1$,

$$b(\psi) = \lambda = \mathrm{e}^{\psi} \quad \text{and} \quad c(y, \gamma) = -\log(y!).$$

Result 1: Let $Y \sim f_e(y; \psi)$ and define $\ell(\psi; Y) = \log\{f_e(Y; \psi)\}$, we have the following well–known formulae:

$$E\left(\frac{\mathrm{d}\ell}{\mathrm{d}\psi}\right) = 0 \quad \text{and} \quad E\left(\frac{\mathrm{d}^2\ell}{\mathrm{d}\psi^2}\right) + E\left(\frac{\mathrm{d}\ell}{\mathrm{d}\psi}\right)^2 = 0. \tag{2.57}$$

Proof: Since $1 = \int f_e(y; \psi)\, \mathrm{d}y$, by differentiating both sides of this identity with respect to ψ, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}\psi} \int f_e(y; \psi)\, \mathrm{d}y.$$

When the support of Y does not depend on the parameter ψ, we can interchange the differentiation and integration with respect to ψ, yielding

$$0 = \int \frac{\mathrm{d}f_e}{\mathrm{d}\psi}\,\mathrm{d}y = \int \frac{\mathrm{d}f_e/\,\mathrm{d}\psi}{f_e}\cdot f_e\,\mathrm{d}y = \int \frac{\mathrm{d}\ell}{\mathrm{d}\psi}\cdot f_e\,\mathrm{d}y = E\left(\frac{\mathrm{d}\ell}{\mathrm{d}\psi}\right), \quad (2.58)$$

which is the first formula in (2.57). By differentiating both sides of (2.58) with respect to ψ again, we have

$$\begin{aligned} 0 &= \int \left(\frac{\mathrm{d}^2\ell}{\mathrm{d}\psi^2}\cdot f_e + \frac{\mathrm{d}\ell}{\mathrm{d}\psi}\cdot\frac{\mathrm{d}f_e}{\mathrm{d}\psi}\right)\mathrm{d}y \\ &= \int \left\{\frac{\mathrm{d}^2\ell}{\mathrm{d}\psi^2}\cdot f_e + \left(\frac{\mathrm{d}\ell}{\mathrm{d}\psi}\right)^2\cdot f_e\right\}\mathrm{d}y \\ &= E\left(\frac{\mathrm{d}^2\ell}{\mathrm{d}\psi^2}\right) + E\left(\frac{\mathrm{d}\ell}{\mathrm{d}\psi}\right)^2, \end{aligned}$$

which is the second formula in (2.57). □

Result 2: Let $Y \sim f_e(y;\psi)$, show that

$$E(Y) = b'(\psi) \quad \text{and} \quad \mathrm{Var}(Y) = b''(\psi)a(\gamma), \quad (2.59)$$

respectively.

Proof: From (2.56), we have $\ell(\psi;Y) = \{Y\psi - b(\psi)\}/a(\gamma) + c(Y,\gamma)$, so that

$$\frac{\mathrm{d}\ell}{\mathrm{d}\psi} = \frac{Y - b'(\psi)}{a(\gamma)} \quad \text{and} \quad \frac{\mathrm{d}^2\ell}{\mathrm{d}\psi^2} = -\frac{b''(\psi)}{a(\gamma)}.$$

From the first formula in (2.57), we have

$$0 = E\left(\frac{\mathrm{d}\ell}{\mathrm{d}\psi}\right) = \frac{E(Y) - b'(\psi)}{a(\gamma)} \quad \Longrightarrow \quad E(Y) = b'(\psi).$$

From the second formula in (2.57), we have

$$0 = -\frac{b''(\psi)}{a(\gamma)} + \frac{\mathrm{Var}(Y)}{\{a(\gamma)\}^2} \quad \Longrightarrow \quad \mathrm{Var}(Y) = b''(\psi)a(\gamma). \qquad \square$$

Solution: Let $Y_i \overset{\text{ind}}{\sim} f_e(y_i;\psi_i)$, $i = 1,\ldots,m$, where $\psi_i = \boldsymbol{x}_{(i)}^\top\boldsymbol{\theta}$, then the log-likelihood function of $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{m}\left\{\frac{y_i\psi_i - b(\psi_i)}{a(\gamma)} + c(y_i,\gamma)\right\}$$

$$= \quad \text{constant} + \frac{1}{a(\gamma)} \sum_{i=1}^{m} \Big\{ y_i \cdot \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta} - b(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) \Big\}$$

$$\hat{=} \quad \text{constant} + \frac{1}{a(\gamma)} \sum_{i=1}^{m} f_i(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}),$$

where $f_i(\psi) = y_i \psi - b(\psi)$. Noting that

$$f_i'(\psi) = y_i - b'(\psi) \quad \text{and} \quad -f_i''(\psi) = b''(\psi),$$

from (2.54), we have

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \frac{\sum_{i \in \mathbb{I}_j} \Big\{ y_i - b'(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)}) \Big\} x_{ij}}{\sum_{i \in \mathbb{I}_j} b''(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)})\, x_{ij}^2 / \lambda_{ij}} \tag{2.60}$$

for $j = 1, \ldots, q$. ‖

Example 2.15 (Log-linear model for lymphocyte data). Consider the following Poisson regression model:

$$\begin{cases} Y_i & \overset{\text{ind}}{\sim} & \text{Poisson}(\mu_i), \quad i = 1, \ldots, m, \\ \log(\mu_i) & = & \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}, \end{cases}$$

where $\boldsymbol{x}_{(i)}$ is the covariate vector and $\boldsymbol{\theta}$ is the parameter vector. Derive the DP algorithm to calculate the MLEs of $\boldsymbol{\theta}$.

Solution: Let $Y_{\text{obs}} = \{y_i\}_{i=1}^m$ be the observed data, then the observed-data likelihood function of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta} | Y_{\text{obs}}) = \prod_{i=1}^{m} \frac{\mu_i^{y_i}}{y_i!} \, \mathrm{e}^{-\mu_i}.$$

The log-likelihood function is given by

$$\ell(\boldsymbol{\theta} | Y_{\text{obs}}) = c + \sum_{i=1}^{m} \Big\{ y_i (\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) - \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) \Big\} \hat{=} \, c + \sum_{i=1}^{m} f_i(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}),$$

where c is a constant and $f_i(\psi) = y_i \psi - \mathrm{e}^{\psi}$. Set $b(\psi) = \mathrm{e}^{\psi}$ in (2.60), then the DP algorithm (2.60) reduces to

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \frac{\sum_{i \in \mathbb{I}_j} \Big\{ y_i - \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)}) \Big\} x_{ij}}{\sum_{i \in \mathbb{I}_j} \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)})\, x_{ij}^2 / \lambda_{ij}}, \quad j = 1, \ldots, q. \tag{2.61}$$

The DP algorithm (2.61) can be re-expressed in the form of matrix (see Exercise 2.10).

Table 2.4 Lymphocyte data

i	y_i	c_i	d_i
1	109	269	0.50
2	47	78	0.75
3	94	115	1.00
4	114	90	1.50
5	138	84	2.00
6	125	59	2.50
7	97	37	3.00

Lymphocyte data: We consider the lymphocyte data in Table 2.4, the medical background is reported by Groer & Pereira (1987). Here $m = 7$, y_i denotes the number of dicentrics for individual i, c_i is the number of cells (in thousands), and d_i the dose level. It is required to relate the expectation μ_i with the explanatory variables c_i and d_i. The considered model is as follows:

$$\log(\mu_i) = \theta_0 + \theta_1 \log(c_i) + \theta_2 \log(d_i).$$

Applying the DP algorithm (2.61), we obtain $\hat{\boldsymbol{\theta}} = (-0.125,\ 0.985,\ 1.022)^\top$ after 12,000 iterations starting from $\boldsymbol{\theta}^{(0)} = \mathbf{1}_3$. The estimated standard errors are 0.751, 0.156 and 0.148, respectively.

R codes for the Poisson regression:

```
lymphocyte <- function(ind, MMn)
{ # ========= Aim ============================================
  # y_i ~ind Poisson(\mu_i), i=1,...,m
  # log(mu_i) = x_i^T \theta_{q x 1}
  # Find the MLEs of \theta in Example 2.15
  # ========= Input ===========================================
  # ind = 1: use the built-in R function glm()
  # ind = 2: use the DP algorithm
  # MMn    : the number of iterations in MM algorithm
  # ======== Some Notations ==================================
  # X_{m x q} = (x_1, ..., x_q)
  #           = (x_(1), ..., x_(m))^T
  # Y = |X|, Z = diag(Y 1_q) %*% Y
  # th^(t+1) = th^(t) +
```

```
  # X^T(y-e^{X th^(t)})/Z^T e^{X th^(t)}
  # ======== Output ========================================
  # TH: storing the approximation MLEs of \theta
  # ========================================================
  m <- 7
  q <- 2
  y <- c(109, 47, 94, 114, 138, 125, 97)
  cell <- c(269, 78, 115, 90, 84, 59, 37)
  dose <- c(0.5, 0.75, 1.0, 1.5, 2., 2.5, 3)
  x1 <- log(cell)
  x2 <-log(dose)
  if(ind == 1) {
     model <- glm(y ~ x1 + x2, family=poisson)
     print(summary(model))
  }
  if(ind == 2) {
    X <- cbind(rep(1, m), c(x1), c(x2))
    q <- 3
    Y <- abs(X)
    Z <- diag(c(Y %*% rep(1, q))) %*% Y
    th <- c(1, 1, 1)
    TH <- matrix(0, MMn, q)
    for(tt in 1:MMn) {
      p <- exp(c(X %*% th))
      th <- th+(t(X)%*%(y - p))/(t(Z)%*%p)
      TH[tt,  ] <- th }
  return(TH[MMn,  ])
  }
}
```

‖

Exercise 2

2.1 (Derivative of a vector). Let $\boldsymbol{x}$ and $\boldsymbol{a}$ be two $n \times 1$ vectors, $\boldsymbol{b}$ an $m \times 1$ vector, $\boldsymbol{A}$ an $m \times n$ matrix, and $\boldsymbol{B}$ an $n \times n$ matrix. Define

$$\frac{\partial \boldsymbol{b}^\top}{\partial x} = \left(\frac{\partial b_1}{\partial x}, \ldots, \frac{\partial b_m}{\partial x} \right).$$

Show that

$$\begin{cases} \dfrac{\partial(\boldsymbol{a}^\top\boldsymbol{x})}{\partial\boldsymbol{x}} = \boldsymbol{a}, & \dfrac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial\boldsymbol{x}^\top} = \boldsymbol{A}, & \dfrac{\partial(\boldsymbol{A}\boldsymbol{x})^\top}{\partial\boldsymbol{x}} = \boldsymbol{A}^\top, \\ \dfrac{\partial(\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x})}{\partial\boldsymbol{x}} = (\boldsymbol{B}+\boldsymbol{B}^\top)\boldsymbol{x}, & & \dfrac{\partial^2(\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x})}{\partial\boldsymbol{x}\partial\boldsymbol{x}^\top} = \boldsymbol{B}+\boldsymbol{B}^\top. \end{cases}$$

2.2 (The delta method). Let $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \dot{\sim} N_d(\mathbf{0}, \boldsymbol{\Sigma})$ and $h(\cdot)$ be a differentiable function. Prove that $h(\hat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta}) \dot{\sim} N\big(0, \{\nabla h(\hat{\boldsymbol{\theta}})\}^\top \boldsymbol{\Sigma} \nabla h(\hat{\boldsymbol{\theta}})\big)$.

2.3 Let $X \sim \mathrm{TN}(\mu, \sigma^2; a, b)$, define $a_1 = (a-\mu)/\sigma$ and $b_1 = (b-\mu)/\sigma$. Show that

$$\begin{cases} E(X) = \mu - \sigma \dfrac{\phi(b_1) - \phi(a_1)}{\Phi(b_1) - \Phi(a_1)}, \\ \mathrm{Var}(X) = \sigma^2 \left\{ 1 - \dfrac{b_1\phi(b_1) - a_1\phi(a_1)}{\Phi(b_1) - \Phi(a_1)} \right\} - \{E(X) - \mu\}^2. \end{cases}$$

2.4 (Kullback–Leibler divergence). Let $g(x)$ and $h(x)$ be two densities with the same support $\mathbb{X} = \{x\colon\, g(x) > 0\} = \{x\colon\, h(x) > 0\}$. Let a r.v. $X \sim g(x)$. The KL divergence between $g(x)$ and $h(x)$ is defined as

$$\mathrm{KL}(g\|h) = E\left[\log\left\{\frac{g(X)}{h(X)}\right\}\right] = \int_{\mathbb{X}} g(x) \log\left\{\frac{g(x)}{h(x)}\right\} \mathrm{d}x. \tag{2.62}$$

$\mathrm{KL}(g\|h)$ is also called the KL cross–entropy or the cross–entropy or the relative entropy. Note that $\mathrm{KL}(g\|h)$ is not a "distance" between $g(x)$ and $h(x)$ since $\mathrm{KL}(g\|h) \neq \mathrm{KL}(h\|g)$. From Jensen's inequality (see Exercise 2.5), we have

$$\mathrm{KL}(g\|h) \geqslant 0 \quad \text{and} \quad \mathrm{KL}(g\|h) = 0 \ \text{ iff } \ g(x) = h(x). \tag{2.63}$$

In fact, we have

$$\mathrm{KL}(g\|h) = E\left[-\log\left\{\frac{h(X)}{g(X)}\right\}\right] \geqslant -\log\left[E\left\{\frac{h(X)}{g(X)}\right\}\right] = -\log 1 = 0.$$

2.5 (Convexity and Jensen's inequality). Let $\psi(x)$ be a twice differentiable function defined on a convex set $\mathbb{S}$. The following statements are equivalent:

(a) $\psi(x)$ is convex.

(b) $\psi''(x) \geqslant 0$ for all $x \in \mathbb{S}$.

(c) $\psi(x) \geqslant \psi(x_0) + \psi'(x_0)(x - x_0)$ for any $x, x_0 \in \mathbb{S}$.

(d) $\psi(\alpha x_1 + (1-\alpha)x_2) \leqslant \alpha\psi(x_1) + (1-\alpha)\psi(x_2)$ for any $x_1, x_2 \in \mathbb{S}$ and $\alpha \in (0,1)$.

(e) $\psi(\sum_{i=1}^n \alpha_i x_i) \leqslant \sum_{i=1}^n \alpha_i \psi(x_i)$ for any $\{x_i\}_{i=1}^n \in \mathbb{S}$, where $\alpha_i > 0$ and $\sum_{i=1}^n \alpha_i = 1$.

Let X be a r.v. and ψ a convex function, then

$$\psi(E(X)) \leqslant E\{\psi(X)\},$$

provided that both expectations exist. In addition, For a strictly convex function ψ, equality holds in Jensen's inequality iff $X = E(X)$ almost surely. [Hint: Set $u_0 = E(X)$. The convexity yields $\psi(u) \geqslant \psi(u_0) + \psi'(u_0)(u - u_0)$. Substitute X for u and take expectations]

2.6 Prove that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ defined by (2.48) minorizes $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

2.7 Let $\boldsymbol{p} = (p_1, \ldots, p_n)^\top \in \mathbb{T}_n$, then (Böhning & Lindsay 1988)

$$\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top \leqslant 0.5(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^\top/n).$$

2.8 (Multinomial logistic regression). Consider the baseline category logit model for nominal responses. Let Y be a categorical response with K categories and $p_k(\boldsymbol{x}) = \Pr(Y = k|\boldsymbol{x})$ the response probability at a fixed vector $\boldsymbol{x}_{n\times 1}$ of covariates so that $\sum_{k=1}^K p_k(\boldsymbol{x}) = 1$. A baseline-category logit model often assumes $\log\{p_k(\boldsymbol{x})/p_K(\boldsymbol{x})\} = \boldsymbol{x}^\top\boldsymbol{\theta}_k$; i.e.,

$$p_k(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\theta}_k)}{1 + \sum_{h=1}^{K-1}\exp(\boldsymbol{x}^\top\boldsymbol{\theta}_h)}, \qquad k = 1, \ldots, K,$$

where $\boldsymbol{\theta}_K = \mathbf{0}$ and $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kn})^\top$ denote parameters for the k-th category. Consider m independent observations and let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})^\top$ denote the observation for subject $i\,(i = 1, \ldots, m)$, where $y_{ik} = 1$ when the response is in category k and $y_{ik} = 0$ otherwise so that $\sum_{k=1}^K y_{ik} = 1$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})^\top$ denote covariates for subject i and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_{K-1}^\top)^\top$ the $n(K-1)$-vector of parameters, then the log-likelihood is

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^m \left\{ \sum_{k=1}^{K-1} y_{ik}(\boldsymbol{x}_i^\top\boldsymbol{\theta}_k) - \log\left(1 + \sum_{k=1}^{K-1} \mathrm{e}^{\boldsymbol{x}_i^\top\boldsymbol{\theta}_k}\right)\right\}.$$

Prove that (Böhning 1992)

(a) The score vector is

$$\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{m} \left[\left\{ \boldsymbol{y}_i^{(-K)} - \boldsymbol{p}^{(-K)}(\boldsymbol{x}_i) \right\} \otimes \boldsymbol{x}_i \right],$$

and the observed information $-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is given by

$$\sum_{i=1}^{m} \left[\left\{ \text{diag}\left(\boldsymbol{p}^{(-K)}(\boldsymbol{x}_i) \right) - \left(\boldsymbol{p}^{(-K)}(\boldsymbol{x}_i) \right)^{\otimes 2} \right\} \otimes \boldsymbol{x}_i \boldsymbol{x}_i^\top \right],$$

where $\boldsymbol{y}_i^{(-K)} = (y_{i1}, \ldots, y_{i,K-1})^\top$ and

$$\boldsymbol{p}^{(-K)}(\boldsymbol{x}_i) = (p_1(\boldsymbol{x}_i), \ldots, p_{K-1}(\boldsymbol{x}_i))^\top.$$

(b) $-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) \leqslant \frac{1}{2}(\boldsymbol{I}_{K-1} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top) \otimes \sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}_i^\top \equiv \boldsymbol{B}$ and

$$\boldsymbol{B}^{-1} = 2(\boldsymbol{I}_{K-1} + \mathbf{1}\mathbf{1}^\top) \otimes \left(\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1}.$$

(c) Devise a QLB algorithm to find the MLEs of $\boldsymbol{\theta}$.

2.9 Prove that the function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ defined in (2.52) minorizes $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. [Hint: Note that

$$\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta} = \sum_{j \in \mathbb{J}_i} \lambda_{ij} \left\{ \lambda_{ij}^{-1} x_{ij} (\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)} \right\}$$

and use the concavity inequality; i.e., the reverse of inequality (e) in Exercise 2.5]

2.10 Show that (2.61) can be rewritten in matrix form as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{X}^\top (\boldsymbol{y} - \mathrm{e}^{\boldsymbol{X}\boldsymbol{\theta}^{(t)}}) / \boldsymbol{Z}^\top \mathrm{e}^{\boldsymbol{X}\boldsymbol{\theta}^{(t)}},$$

where $\boldsymbol{X} = (x_{ij}) = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(m)})^\top$, $\boldsymbol{Y} = (|x_{ij}|) = \text{abs}(\boldsymbol{X})$, $\boldsymbol{Z} = \text{diag}(\boldsymbol{Y}\mathbf{1}_q)\boldsymbol{Y}$.

2.11 Apply the DP algorithm to the logistic regression (2.14).

2.12 Apply the DP algorithm to the probit regression as shown in 45•.

2.13 Consider the following normal linear model

$$Y_i \overset{\text{ind}}{\sim} N(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}, \sigma^2), \quad i = 1, \ldots, m,$$

where $\boldsymbol{\theta}_{q\times 1}$ is the vector of regression coefficients, σ^2 is the variance, and $\boldsymbol{x}_{(i)}$ is the vector of covariates for subject i.

(a) Derive the DP algorithm for finding the MLEs of $\boldsymbol{\theta}$ and σ^2.

(b) Compare the advantages and disadvantages between the DP algorithm for finding the MLEs of $\boldsymbol{\theta}$ and the closed–form solution

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$ and $\boldsymbol{X} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(m)})^\top$, for two cases (i) q is very large; and (ii) $\boldsymbol{X}^\top \boldsymbol{X}$ is almost singular.

2.14 Let $\boldsymbol{A} > 0$ be an $m \times m$ positive definite matrix. Define

$$f(\boldsymbol{A}) = c|\boldsymbol{A}|^{\frac{n}{2}} \exp(-0.5 \operatorname{tr} \boldsymbol{A}),$$

where c is a positive constant. Show that $n\boldsymbol{I}_m = \arg\max_{\boldsymbol{A}>0} f(\boldsymbol{A})$.

2.15 For one sample size, to find the log-likelihood function $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$, the score vector $\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}})$, the observed information matrix $\boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}})$, and the expected information matrix $\boldsymbol{J}(\boldsymbol{\theta})$, where $Y_{\text{obs}} = \{y\}$ or $\{\boldsymbol{y}\}$, y is the realization of Y and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the realization of $\mathbf{y}$.

(a) $Y \sim \text{Binomial}(n, \theta)$, where $0 < \theta < 1$.

(b) $Y \sim \text{Poisson}(\theta)$, where $\theta > 0$.

(c) $Y \sim \text{Exponential}(1/\theta)$, where $\theta > 0$.

(d) $\mathbf{y} = (Y_1, \ldots, Y_n)^\top \sim \text{Multinomial}(N, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{T}_n$.

2.16 Let $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} N(\mu_i(\boldsymbol{\theta}), \sigma^2/w_i)$, where $\mu_i(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ that is an unknown parameter vector, and $\{w_i\}_{i=1}^n$ are known constants. Derive the Fisher scoring algorithm to find the MLEs of $\boldsymbol{\theta}$ and σ^2.

2.17 Let $\{Y_i\}_{i=1}^n \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i)$ and $p_i = \Phi(\boldsymbol{x}_{(i)}^\top \boldsymbol{\beta})$, where $\Phi(\cdot)$ is the cdf of $N(0, 1)$, $\boldsymbol{x}_{(i)}$ is a known vector of covariates for subject i, and $\boldsymbol{\beta}_{(p+1)\times 1}$ is an unknown vector of parameters.

(a) Derive the score vector and the observed information matrix.

(b) Use the Newton–Raphson algorithm to find the MLEs $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

2.18 Assume that the likelihood function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top \in \mathbb{T}_4$ based on the observed-data $Y_{\text{obs}} = \{n_1, \ldots, n_4; n_{12}, n_{34}\}$ is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) = \left(\prod_{i=1}^{4} \theta_i^{n_i}\right) \times (\theta_1 + \theta_2)^{n_{12}} (\theta_3 + \theta_4)^{n_{34}}.$$

Use the EM algorithm to find the MLEs of $\boldsymbol{\theta}$.

2.19 Assume that the likelihood function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top \in \mathbb{T}_4$ based on the observed-data $Y_{\text{obs}} = \{n_1, \ldots, n_4; n_{12}, n_{34}; n_{13}, n_{24}\}$ is

$$\begin{aligned} L(\boldsymbol{\theta}|Y_{\text{obs}}) &= \left(\prod_{i=1}^{4} \theta_i^{n_i}\right) \times (\theta_1 + \theta_2)^{n_{12}} (\theta_3 + \theta_4)^{n_{34}} \\ &\quad \times (\theta_1 + \theta_3)^{n_{13}} (\theta_2 + \theta_4)^{n_{24}}. \end{aligned}$$

Use the EM algorithm to find the MLEs of $\boldsymbol{\theta}$.

2.20 Let $Y = 1$ if a respondent is a drug user and $Y = 0$ otherwise. Let U denote the number of travel out of Hong Kong per year for the same respondent in a population in Hong Kong. Obviously, Y is a sensitive binary r.v. (thus it is not observable if the question is asked directly) and U is a non-sensitive random variable. Define $X = Y + U$. Let $Y \sim \text{Bernoulli}(\theta)$, $U \sim \text{Poisson}(\lambda)$ and $Y \perp\!\!\!\perp U$. The interviewer could ask the i-th respondent to report the sum $X_i = Y_i + U_i$ according to his/her truthful answer, $i = 1, \ldots, n$. Let the observed data be $x_1, \ldots, x_n$.

(a) Find the moment estimators of θ and λ.

(b) Find the distribution of X.

(c) Find the conditional distribution of Y given X.

(d) Find the MLEs of θ and λ via the EM algorithm.

2.21 Let $Z \sim \text{Bernoulli}(1-\phi)$, $X \sim \text{Poisson}(\lambda)$ and $Z \perp\!\!\!\perp X$. From Exercise 1.12, we know that $Y \sim \text{ZIP}(\phi, \lambda)$ has the SR $Y \overset{\text{d}}{=} ZX$.

(a) Find the conditional distribution of $Z|(Y = y)$.

(b) Find the conditional distribution of $X|(Y = y)$.

(c) Find $E(Y)$.

(d) Let $Y_{\text{obs}} = \{y_i\}_{i=1}^n$ be the observed data, where y_i is the realization of Y_i, and $\{Y_i\}_{i=1}^n \overset{\text{iid}}{\sim} \text{ZIP}(\phi, \lambda)$. Based on the SR $Y_i \overset{\text{d}}{=} Z_i X_i$, we can introduce the latent variables $Z_i \sim \text{Bernoulli}\,(1-\phi)$, $X_i \sim \text{Poisson}\,(\lambda)$ and $Z_i \perp\!\!\!\perp X_i$ for $i = 1, \ldots, n$. The corresponding latent data are denoted by $Y_{\text{mis}} = \{z_i, x_i\}_{i=1}^n$, which are equal to the complete data Y_{com}. Find the MLEs of ϕ and λ via an EM algorithm.

2.22 A discrete r.v. X is said to follow a *zero-truncated Poisson* (ZTP) distribution, denoted by $X \sim \text{ZTP}(\lambda)$ with $\lambda > 0$, if its pmf is

$$\Pr(X = x) = c \cdot \frac{\lambda^x \, \mathrm{e}^{-\lambda}}{x!}, \quad c \,\hat{=}\, \frac{1}{1 - \mathrm{e}^{-\lambda}}, \quad x = 1, \ldots, \infty. \tag{2.64}$$

(a) Let $X \sim \text{ZTP}(\lambda)$ and $Y \sim \text{Poisson}(\lambda)$. Show that Y has the following SR:

$$Y \overset{\text{d}}{=} ZX = \begin{cases} 0, & \text{with probability } 1 - 1/c, \\ X, & \text{with probability } 1/c, \end{cases} \tag{2.65}$$

where $Z \perp\!\!\!\perp X$ and $Z \sim \text{Bernoulli}(1/c)$ with c specified by (2.64); i.e., $Z \sim \text{Bernoulli}(1 - \mathrm{e}^{-\lambda})$.

(b) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{ZTP}(\lambda)$ and $Y_{\text{obs}} = \{x_i\}_{i=1}^n$.

(b1) Based on the SR (2.65), please create an EM algorithm to calculate the MLE of λ.

(b2) Show that the observed-data log-likelihood is proportional to $n\{\bar{x} \log \lambda - \lambda + g(\lambda)\}$, where $g(\lambda) = -\log(1 - \mathrm{e}^{-\lambda})$.

(b3) Prove the following inequality:

$$g(\lambda) \geqslant g(\lambda_0) + (\lambda - \lambda_0) g'(\lambda_0), \quad \forall \lambda, \lambda_0 > 0.$$

(b4) In the above inequality, let $\lambda_0 = \lambda^{(t)}$ and then construct a Q-function $Q(\lambda|\lambda^{(t)})$. Based on this Q-function, create an MM algorithm to calculate the MLE of λ.

2.23 For $i = 1, \ldots, m$, let $\mathbf{x}_i = (X_{i1}, \ldots, X_{in})^\top \overset{\text{iid}}{\sim} \text{Dirichlet}_n(\boldsymbol{a})$ on $\mathbb{T}_n$. Use the Newton–Raphson algorithm to estimate the parameter vector $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$. [Hint: See (A.24) in Appendix A.2.1]

2.24 From (A.4) in Appendix A.1.3, let $\{X_i\}_{i=1}^m \overset{\text{ind}}{\sim} \text{BBinomial}(n_i, \alpha, \beta)$, where all $\{n_i\}_{i=1}^m$ are known positive integers and $\alpha, \beta\,(>0)$ are two parameters.

(a) Show that the log-likelihood function of (α, β) is

$$\begin{aligned}\ell(\alpha,\beta) &= c+\sum_{i=1}^{m}\left\{\sum_{j=0}^{x_i-1}\log(\alpha+j)+\sum_{j=0}^{n_i-x_i-1}\log(\beta+j)\right\}\\ &\quad-\sum_{i=1}^{m}\sum_{j=0}^{n_i-1}\log(\alpha+\beta+j),\end{aligned}$$

where c is a constant free from (α, β).

(b) Apply the discrete Jensen's inequality, see Exercise 2.5(d), to the concave function $\log(\cdot)$, show that

$$\begin{aligned}\log(\alpha+j) &\geqslant \frac{\alpha^{(t)}}{\alpha^{(t)}+j}\log\left(\frac{\alpha^{(t)}+j}{\alpha^{(t)}}\alpha\right)\\ &\quad+\frac{j}{\alpha^{(t)}+j}\log\left(\frac{\alpha^{(t)}+j}{j}j\right),\end{aligned}\tag{2.66}$$

where $\alpha^{(t)}$ denotes the t-th approximate of the MLE $\hat{\alpha}$.

(c) Apply the support superplane inequality, see Exercise 2.5(c) to the convex function $-\log(\cdot)$, show that

$$\begin{aligned}-\log(\alpha+\beta+j) &\geqslant -\log(\alpha^{(t)}+\beta^{(t)}+j)\\ &\quad-\frac{\alpha+\beta-\alpha^{(t)}-\beta^{(t)}}{\alpha^{(t)}+\beta^{(t)}+j},\end{aligned}\tag{2.67}$$

where $\beta^{(t)}$ denotes the t-th approximate of the MLE $\hat{\beta}$.

(d) Design an MM algorithm to find the MLEs of α and β. [Hint: Combining (2.66) with (2.67) can construct a minorizing function $Q(\alpha, \beta|\alpha^{(t)}, \beta^{(t)})$]

2.25 Let the log-likelihood function be given by

$$\ell(\theta) = -\frac{\sqrt{a^2+\theta^2}}{s_1}-\frac{\sqrt{b^2+(c-\theta)^2}}{s_2},\quad \theta\geqslant 0,$$

where $a>0$, $b<0$, $c>0$, $s_1>0$, $s_2>0$ are known constants.

(a) Prove the following inequality:

$$-\sqrt{x} \geqslant -\sqrt{x_0} - (x - x_0)/(2\sqrt{x_0}), \quad \forall\, x, x_0 > 0.$$

(b) Based on the above inequality, construct a Q-function $Q(\theta|\theta^{(t)})$ and create an MM algorithm to calculate the MLE of θ.

(c) Let $a = 3$, $b = -1$, $c = 2$, $s_1 = 1$, $s_2 = 1.5$, and the initial value $\theta^{(0)} = 0$, calculate $\theta^{(t)}$ for $t = 1, \ldots, 5$.

2.26 Let $Y_{\text{obs}} = \{y_{ij}\colon 1 \leqslant i, j \leqslant n\}$ denote the observed data and the likelihood function be given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{n} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{y_{ij}},$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ is the parameter vector.

(a) Prove the following inequality:

$$-\log x \geqslant -\log x_0 + (x - x_0)(-x_0^{-1}), \quad \forall\, x, x_0 > 0.$$

(b) In the above inequality, let $x = \theta_i + \theta_j$ and $x_0 = \theta_i^{(t)} + \theta_j^{(t)}$. Construct a Q-function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ and create an MM algorithm to calculate the MLEs of $\boldsymbol{\theta}$.

2.27 Let $X_1, \ldots, X_m$ be a random sample from the Weibull density with two parameters $\theta > 0$ and $\lambda > 0$:

$$f(x|\theta, \lambda) = \frac{\theta}{\lambda} \left(\frac{x}{\lambda} \right)^{\theta - 1} \exp\left\{ -\left(\frac{x}{\lambda} \right)^{\theta} \right\}, \quad x > 0.$$

(a) If θ is known, show that the MLE of λ satisfies $\lambda^\theta = \sum_{i=1}^{m} x_i^\theta / m$.

(b) Show that the MLE of θ is given by $\hat{\theta} = \arg\max_{\theta > 0} \ell_1(\theta)$, where

$$\ell_1(\theta) = m \log(\theta) - m \log\left(\sum_{i=1}^{m} x_i^\theta \right) + (\theta - 1) \sum_{i=1}^{m} \log(x_i).$$

Use the Newton method to find the MLE $\hat{\theta}$.

Chapter 3

Integration

1• A motivation from Bayesian computation

- There are two major challenges in advanced Bayesian computation:
  - How to compute posterior quantities of interest?
  - How to sample from a posterior distribution?
- For example, the computation of normalizing constant is closely related to the marginal likelihood, Bayes factor and Bayes model selection.

2• Aims of chapter 3

- In this chapter, we introduce three basic tools to evaluate integrals often encountered in statistical computation.
- The Laplace approximation (§3.1) is an *analytic* approach to approximate integrals based on the Taylor expansion around the mode of the log-integrand.
  - Therefore, algorithms for finding mode of a uni-modal function presented in Chapter 2 would be helpful here.
- The Riemannian simulation (§3.2) and the importance sampling (§3.3) are two *simulation-based* approaches to approximate integrals based on i.i.d. samples from the target or proposal density.
  - Chapter 1 has introduced a number of methods for the generation of r.v.'s with any given distribution, and hence provides a basis for the two simulation-based approaches.

3.1 Laplace Approximations

3• The issue

- Suppose that we are interested in evaluating the integral

$$I(f) = \int_{\mathbb{X}} f(x)\,\mathrm{d}x, \tag{3.1}$$

where the $f(\cdot)$ is non-negative and integrable (Robert & Casella 1999).

3.1• Third–order Taylor expansion

— Let n be the sample size or a parameter which can go to infinity. We define

$$h(x) \hat{=} \frac{1}{n}\log\{f(x)\}, \tag{3.2}$$

which has a mode denoted by $\tilde{x}$.

— A third–order Taylor expansion of $h(x)$ around its mode $\tilde{x}$ gives

$$\begin{aligned} h(x) &= h(\tilde{x}) + (x-\tilde{x})h'(\tilde{x}) + \frac{(x-\tilde{x})^2}{2!}h''(\tilde{x}) \\ &\quad + \frac{(x-\tilde{x})^3}{3!}h'''(\tilde{x}) + \frac{(x-\tilde{x})^4}{4!}h''''(x^*) \\ &\approx h(\tilde{x}) + (x-\tilde{x})h'(\tilde{x}) + \frac{(x-\tilde{x})^2}{2!}h''(\tilde{x}) + R_3, \end{aligned} \tag{3.3}$$

where $x^* = \alpha x + (1-\alpha)\tilde{x}$ for some $\alpha \in [0,1]$, and

$$R_3 \hat{=} \frac{(x-\tilde{x})^3}{3!}h'''(\tilde{x}).$$

3.2• General approximation formulae of $I(f)$

— Note that $h'(\tilde{x}) = 0$ and

$$\mathrm{e}^y = 1 + y + \frac{y^2}{2!} + r_3, \tag{3.4}$$

we have

$$\begin{aligned} f(x) &\overset{(3.2)}{=} \exp\{nh(x)\} \\ &\overset{(3.3)}{\approx} \exp\left\{nh(\tilde{x}) + \frac{n(x-\tilde{x})^2}{2!}h''(\tilde{x})\right\} \times \exp(nR_3). \end{aligned} \tag{3.5}$$

— Therefore, we obtain

$$\begin{aligned} I(f) &\overset{(3.1)}{=} \int_{\mathbb{X}} f(x)\,\mathrm{d}x \overset{(3.5)}{\approx} \mathrm{e}^{nh(\tilde{x})} \int_{\mathbb{X}} \mathrm{e}^{0.5n(x-\tilde{x})^2 h''(\tilde{x})}\,\mathrm{e}^{nR_3}\,\mathrm{d}x \\ &\overset{(3.4)}{=} \mathrm{e}^{nh(\tilde{x})} \int_{\mathbb{X}} \mathrm{e}^{0.5n(x-\tilde{x})^2 h''(\tilde{x})} \Big[1 + \frac{n(x-\tilde{x})^3}{6} h'''(\tilde{x}) \\ &\quad + \frac{n^2(x-\tilde{x})^6}{72}\{h'''(\tilde{x})\}^2 + r_3\Big]\mathrm{d}x. \end{aligned}$$

3.3• First– and second–order approximations of $I(f)$

— The first– and second–order approximations of $I(f)$ are

$$\begin{aligned} I_1(f) &= \mathrm{e}^{nh(\tilde{x})} \int_{\mathbb{X}} \mathrm{e}^{0.5n(x-\tilde{x})^2 h''(\tilde{x})}\,\mathrm{d}x \quad \text{and} \qquad (3.6) \\ I_2(f) &= \mathrm{e}^{nh(\tilde{x})} \int_{\mathbb{X}} \mathrm{e}^{0.5n(x-\tilde{x})^2 h''(\tilde{x})} \Big\{1 + \frac{n(x-\tilde{x})^3}{6} h'''(\tilde{x})\Big\}\,\mathrm{d}x, \end{aligned}$$

respectively.

— We note that the integrand in (3.6) is the kernel of a normal density with mean $\tilde{x}$ and variance $\sigma^2 = -1/\{nh''(\tilde{x})\}$.

— Let $\mathbb{X} = [a, b]$, then

$$\begin{aligned} \int_a^b f(x)\,\mathrm{d}x &\approx \mathrm{e}^{nh(\tilde{x})} \int_a^b \exp\Big\{-\frac{(x-\tilde{x})^2}{2\sigma^2}\Big\}\,\mathrm{d}x \\ &\overset{(3.2)}{=} f(\tilde{x})\sqrt{2\pi}\sigma\Big\{\Phi\Big(\frac{b-\tilde{x}}{\sigma}\Big) - \Phi\Big(\frac{a-\tilde{x}}{\sigma}\Big)\Big\}. \qquad (3.7) \end{aligned}$$

— Tierney & Kadane (1986) further considered the extension to the case of vector.

Example 3.1 (The first–order approximation to incomplete beta integral). To illustrate the Laplace approximation, we consider evaluating the following incomplete beta integral

$$\int_a^b \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}x.$$

Solution: Let

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},$$

then, $\log\{f(x)\} = \text{constant} + (\alpha-1)\log(x) + (\beta-1)\log(1-x)$. Solving

$$0 = \frac{\mathrm{d}\log\{f(x)\}}{\mathrm{d}x} = \frac{\alpha-1}{x} - \frac{\beta-1}{1-x},$$

we obtain the mode of $f(x)$, given by

$$\tilde{x} = \frac{\alpha-1}{\alpha+\beta-2}.$$

From (3.2), we have

$$\begin{aligned} h(x) &= \frac{\log\{f(x)\}}{n} = \frac{\text{constant} + (\alpha-1)\log(x) + (\beta-1)\log(1-x)}{n}, \\ h'(x) &= \frac{\alpha-1}{nx} - \frac{\beta-1}{n(1-x)}, \\ h''(x) &= -\frac{\alpha-1}{nx^2} - \frac{\beta-1}{n(1-x)^2}. \end{aligned}$$

Since

$$\sigma^2 = -\frac{1}{nh''(\tilde{x})},$$

we obtain

$$\sigma = 1 \Big/ \sqrt{\frac{\alpha-1}{\tilde{x}^2} + \frac{\beta-1}{(1-\tilde{x})^2}}.$$

Numerical illustration: For $\alpha = \beta = 3$, we obtain $\tilde{x} = 0.5$ and $\sigma = 0.25$. In Table 3.1, we see that the first–order Laplace approximation (3.7) works better in the central interval of the density, but the accuracy is not very high when the interval is enlarged. ‖

Table 3.1 Accuracy of the first–order Laplace approximation

Interval $[a, b]$	Exact result	Approximation
$[0.50, 0.55]$	0.09312	0.09313
$[0.45, 0.65]$	0.35795	0.35837
$[0.30, 0.70]$	0.67384	0.67712
$[0.20, 0.80]$	0.88416	0.90457
$[0.10, 0.90]$	0.98288	1.04621

4• Two limitations with the Laplace approximation

- First, the integrand $f(\cdot)$ must be *uni-modal* or nearly so.
- Second, although the second– or third–order approximation provides better accuracy than the first–order approximation, numerical computation of the associated Hessian matrices will be prohibitively difficult anyway, especially, for $\boldsymbol{x}$ of moderate– to high–dimension (e.g., greater than 10).
- For these reasons, practitioners often turn to Monte Carlo methods.

3.2 Riemannian Simulation

3.2.1 Classical Monte Carlo integration

5• Formula of the classical Monte Carlo integration

- Let $f(x)$ be a density and the r.v. $X \sim f(x)$. Suppose that we are interested in evaluating

$$\mu \triangleq E\{h(X)\} = \int_{\mathbb{X}} h(x) \cdot f(x)\, \mathrm{d}x, \tag{3.8}$$

  where $\mathbb{X}$ is the support of X and $h(x) \geqslant 0$.

- If we can generate $X^{(1)}, \ldots, X^{(m)} \overset{\text{iid}}{\sim} f(x)$, an approximation to μ is obtained as the empirical average

$$\bar{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} h(X^{(i)}). \tag{3.9}$$

- This estimator is often referred to as the *Monte Carlo integration* following Metropolis & Ulam (1949).

5.1• Understanding Monte Carlo integration

— First, in (3.8) let $h(x) = x$, then $\mu = E(X)$ is the X population mean; while $\bar{\mu}_m = (1/m) \sum_{i=1}^{m} X^{(i)}$ denotes the X sample mean.

— Thus, the essence of Monte Carlo integration is just to estimate the population mean with the sample mean.

— Next, define $Y = h(X)$ and $Y^{(i)} = h(X^{(i)})$ for $i = 1, \ldots, m$, then $\mu = E(Y)$ is the Y population mean; while

$$\bar{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} Y^{(i)}$$

denotes the Y sample mean.

6• Other names of monte carlo integration

- Sometimes, the Monte Carlo integration is also called
  - — classical,
  - — crude,
  - — standard,
  - — native,
  - — regular, or
  - — naïve.

  Monte Carlo integration in order to distinguish it from Monte Carlo integrations via *importance sampling* to be introduced in §3.3.

7• Basic properties of $\bar{\mu}_m$

7.1• Unbiasedness

— The expectation and variance of $\bar{\mu}_m$ are given by

$$E(\bar{\mu}_m) = E(Y) = \mu \quad \text{and} \quad \text{Var}(\bar{\mu}_m) = \frac{\sigma^2}{m}, \tag{3.10}$$

respectively, where

$$\sigma^2 = \text{Var}(Y) = \text{Var}\{h(X)\} = \int_{\mathbb{X}} \{h(x) - \mu\}^2 \cdot f(x)\, \mathrm{d}x. \tag{3.11}$$

7.2• Consistency

— The strong law of large numbers states that $\bar{\mu}_m$ converges *almost surely* (a.s.) to μ, denoted by $\bar{\mu}_m \xrightarrow{\text{a.s.}} \mu$. That is

$$\Pr\left(\lim_{m\to\infty} \bar{\mu}_m = \mu\right) = 1.$$

— The weak law of large numbers states that $\bar{\mu}_m$ converges *in probability* to μ, denoted by $\bar{\mu}_m \xrightarrow{\text{P}} \mu$. That is, for any $\varepsilon > 0$,

$$\lim_{m\to\infty} \Pr\left(|\bar{\mu}_m - \mu| \geqslant \varepsilon\right) = 0.$$

7.3• Asymptotic normality

— The central limit theorem states that $\bar{\mu}_m$ converges *in distribution* to normal distribution; i.e.,

$$\frac{\bar{\mu}_m - \mu}{\sqrt{\text{Var}(\bar{\mu}_m)}} \overset{(3.10)}{=} \frac{\bar{\mu}_m - \mu}{\sigma/\sqrt{m}} \xrightarrow{\text{D}} N(0,1),$$

or equivalently,

$$\Pr\left(|\bar{\mu}_m - \mu| \leqslant 1.96\,\sigma m^{-1/2}\right) = 0.95. \tag{3.12}$$

7.4• The convergence rate of $\bar{\mu}_m$

— The convergence rate of $\bar{\mu}_m$ can be assessed by (3.12).

— That is, its *theoretical rate of convergence* is $O(m^{-1/2})$, regardless of the dimensionality of x.

— The theoretical rate of convergence of $\bar{\mu}_m$ is $O(m^{-1/2})$ in the sense of *probability* and not the usual sense of *absolute error*. The law of the iterated logarithm shows that with probability one,

$$\lim_{m\to\infty} \sup \sqrt{\frac{m}{2\log(\log m)}} |\bar{\mu}_m - \mu| = \sigma^2.$$

Therefore, the theoretical rate of convergence of the Monte Carlo integration is in no case worse than $O(\sqrt{\log(\log m)/m}\,)$.

7.5• The convergence speed of $\bar{\mu}_m$

— Note that the variance of $\bar{\mu}_m$ can be estimated by

$$\widehat{\text{Var}}(\bar{\mu}_m) \overset{(3.10)}{=} \frac{\hat{\sigma}^2}{m} \overset{(3.11)}{=} \frac{1}{m^2} \sum_{i=1}^{m} \left\{h(X^{(i)}) - \bar{\mu}_m\right\}^2,$$

we say the *convergence speed* of $\bar{\mu}_m$ is of order $O(m^{-1})$.

3.2.2 Motivation for Riemannian simulation

8• Advantages of riemannian simulation

- Although (3.9) is rather attractive for practical user because of the simplicity, its convergence speed is very low.
- The approach of *Riemannian simulation* or *simulation by Rieman sums* (Yakowitz *et al.* 1978; Philippe 1997a, 1997b) shares the same simplicity as $\bar{\mu}_m$, while speeding up the convergence from $O(m^{-1})$ to $O(m^{-2})$ for the one–dimensional setting.

8.1• Rieman sum

— To motivate the approach, we first consider the one–dimensional case of (3.8).

— Let $\mathbb{X} = [a, b]$ and $a = a_1 < \cdots < a_{m+1} = b$, then when $m \to \infty$, the Rieman sum

$$\sum_{i=1}^{m} h(a_i) f(a_i)(a_{i+1} - a_i) \to \int_a^b h(x) \cdot f(x)\, \mathrm{d}x.$$

8.2• Riemannian sum estimator

— Replacing the fixed points $\{a_i\}_{i=1}^{m+1}$ by stochastic points $\{X_{(i)}\}_{i=1}^{m+1}$, which are order statistics of i.i.d. samples $X^{(1)}, \ldots, X^{(m+1)}$ from $f(x)$, the approach of Riemannian simulation approximates μ in (3.8) by

$$\hat{\mu}^{\mathrm{R}} = \sum_{i=1}^{m} h(X_{(i)}) f(X_{(i)})(X_{(i+1)} - X_{(i)}). \tag{3.13}$$

We call $\hat{\mu}^{\mathrm{R}}$ the *Riemannian sum estimator* of μ.

— When $f(\cdot)$ is known only up to a normalizing constant; that is, $f(x) = c^{-1} f^*(x)$ with c being unknown, (3.13) can be replaced by

$$\hat{\mu}^{\mathrm{R}} = \frac{\sum_{i=1}^{m} h(X_{(i)}) f^*(X_{(i)})(X_{(i+1)} - X_{(i)})}{\sum_{i=1}^{m} f^*(X_{(i)})(X_{(i+1)} - X_{(i)})}. \tag{3.14}$$

Proof: From (3.8), we have

$$\mu = \int_{\mathbb{X}} h(x) \cdot c^{-1} f^*(x)\, \mathrm{d}x = \frac{\int_{\mathbb{X}} h(x) \cdot f^*(x)\, \mathrm{d}x}{\int_{\mathbb{X}} f^*(x)\, \mathrm{d}x}.$$

Using (3.13) twice, we obtain (3.14). □

3.2.3 Variance of the Riemannian sum estimator

9• Derivation of $\mathrm{Var}(\hat{\mu}^{\mathrm{R}})$

- Let $F(\cdot)$ denote the cdf of the r.v. $X \sim f(x)$ and $F^{-1}(\cdot)$ the inverse function of $F(\cdot)$.
- Let $u = F(x)$, then $x = F^{-1}(u)$ and (3.8) can be rewritten as

$$\mu = \int_{\mathbb{X}} h(x)\,\mathrm{d}F(x) = \int_0^1 H(u)\,\mathrm{d}u,$$

  where $H(u) = h(F^{-1}(u))$.
- Furthermore, let $U_{(1)}, \ldots, U_{(m+1)}$ be an ordered sample from $U(0,1)$, then we have

$$\begin{aligned} X_{(i+1)} - X_{(i)} &= F^{-1}(U_{(i+1)}) - F^{-1}(U_{(i)}) \\ &= (U_{(i+1)} - U_{(i)})\nabla F^{-1}(U_{(i)}) + \text{Remainder} \\ &\approx (U_{(i+1)} - U_{(i)})\nabla F^{-1}(U_{(i)}), \end{aligned}$$

  as the remainder is negligible.
- Since $\nabla F^{-1}(x) = 1/f(F^{-1}(x))$, from (3.13), we obtain

$$\begin{aligned} \hat{\mu}^{\mathrm{R}} &= \sum_{i=1}^{m} h(F^{-1}(U_{(i)}))f(F^{-1}(U_{(i)}))(X_{(i+1)} - X_{(i)}) \\ &\approx \sum_{i=1}^{m} H(U_{(i)})(U_{(i+1)} - U_{(i)}) \\ &= \delta(U) - H(0)U_{(1)} - H(U_{(m+1)})(1 - U_{(m+1)}), \end{aligned}$$

  where $\delta(U)$ is defined by (3.22).
- Exercise 3.3 shows that

$$\mathrm{Var}\{\delta(U)\} = O(m^{-2}).$$

- Similarly, we can verify that (See Philippe 1997a, 1997b for more detail)

$$\begin{aligned} \mathrm{Var}\{H(0)U_{(1)}\} &= O(m^{-2}) \quad \text{and} \\ \mathrm{Var}\{H(U_{(m+1)})[1 - U_{(m+1)}]\} &= O(m^{-2}) \end{aligned}$$

  so that $\mathrm{Var}(\hat{\mu}^{\mathrm{R}}) = O(m^{-2})$.

9.1• Remarks

— When compared with the classical Monte Carlo estimator $\bar{\mu}_m$, the Riemannian sum estimator $\hat{\mu}^{\mathrm{R}}$ improves the approximation by reducing the variance from $O(m^{-1})$ to $O(m^{-2})$.

— Unfortunately, this improvement fails to extend to the case of multidimensional integrals due to the "curse of dimensionality".

Example 3.2 (Comparison of $\bar{\mu}_m$ with Riemannian sum estimator). Let $X \sim \mathrm{Beta}(a, b)$ with $a = 3$, $b = 7$ and $h(x) = x^2 + \log(x + 1)$ be the function of interest. Let μ be defined by (3.8), please compare $\bar{\mu}_m$ with $\hat{\mu}^{\mathrm{R}}$.

Solution: We generate $m = 4{,}000$ i.i.d. samples $X^{(1)}, \ldots, X^{(m)}$ from $\mathrm{Beta}(a, b)$ and compare the empirical average $\bar{\mu}_m$ defined by (3.9) with the Riemannian sum estimator given by

$$\hat{\mu}^{\mathrm{R}} = \sum_{i=1}^{m-1} h(X_{(i)}) \frac{X_{(i)}^2 (1 - X_{(i)})^6}{B(3, 7)} (X_{(i+1)} - X_{(i)}).$$

Figure 3.1 illustrates that $\hat{\mu}^{\mathrm{R}}$ has much greater stability and faster speed of convergence than $\bar{\mu}_m$.

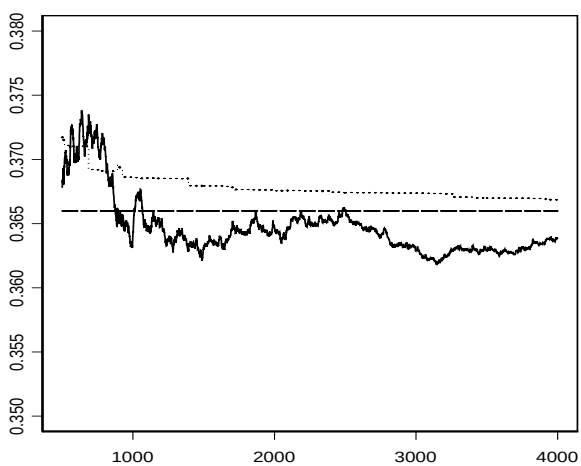

Figure 3.1 Comparison between the convergence of the empirical average $\bar{\mu}_m$ (solid line) and the Riemannian sum estimator $\hat{\mu}^{\mathrm{R}}$ (dotted line) for $a = 3$ and $b = 7$. The final values are 0.363785 and 0.366862, respectively, for a true value of 0.366 (dashed line). ‖

3.3 The Importance Sampling Method

10• Why does $\bar{\mu}_m$ usually have a low efficiency

- In §3.2.1, we estimate $\mu = \int_{\mathbb{X}} h(x) \cdot f(x)\,\mathrm{d}x$ by the empirical average $\bar{\mu}_m = (1/m)\sum_{i=1}^{m} h(X^{(i)})$.
- Let $\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1$, where $\mathbb{X}_0 = \{x\colon\ h(x) \approx 0\}$ and

$$\mathbb{X}_1 = \{x\colon \text{The value of } h(x) \text{ is very large}\}.$$

- The main reason for low efficiency associated with the classical Monte Carlo integration $\bar{\mu}_m$ is that it wastes a lot of effort in evaluating such $h(X^{(i)})$'s, where $X^{(i)} \in \mathbb{X}_0$.
- Especially, when $h(\cdot)$ is very complicated so that the computation of $h(X^{(i)})$ is expensive, which does not contribute too much to $\bar{\mu}_m$.

3.3.1 The formulation of the importance sampling method

11• The idea of the importance sampling

- The *importance sampling* idea (Marshall 1956) suggests that one should focus on the regions of "importance" (i.e., $\mathbb{X}_1$) so as to save computational resources.
- The importance sampling method has a close relationship with the SIR method in §1.4.

11.1• A motivation

— Suppose that we want to evaluate the integral (3.8), but now $f(x)$ is not necessary to be a density.

— Let $H(x) \hat{=} h(x)f(x)$ be defined on $\mathbb{X}$.

— If we could find an easy–sampling density function $g(\cdot)$ with support $\mathbb{X}$, we can write

$$\mu = \int_{\mathbb{X}} H(x)\,\mathrm{d}x = \int_{\mathbb{X}} \frac{H(x)}{g(x)} \cdot g(x)\,\mathrm{d}x = \int_{\mathbb{X}} \mathrm{w}(x) \cdot g(x)\,\mathrm{d}x,$$

where $\mathrm{w}(x) \hat{=} H(x)/g(x)$ is called ratio function.

— Let $X^{(1)}, \ldots, X^{(m)} \overset{\text{iid}}{\sim} g(x)$, then μ can be estimated by

$$\tilde{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \mathrm{w}(X^{(i)}), \tag{3.15}$$

which is called the *importance sampling* (IS) estimator.

11.2• The importance sampling method

— Step 1: Generate i.i.d. samples $\{X^{(i)}\}_{i=1}^{m}$ from a *proposal density* $g(\cdot)$;

— Step 2: Calculate ratios $\mathrm{w}(X^{(i)}) = H(X^{(i)})/g(X^{(i)})$ for $i = 1, \ldots, m$, and approximate μ with $\tilde{\mu}_m$ given by (3.15).

11.3• Other names of the proposal density

— The proposal density $g(\cdot)$ has various names such as generating density, trial density, importance (sampling) density, instrumental density, and so on.

12• Basic properties of $\tilde{\mu}_m$

12.1• Unbiasedness

— Let $X \sim g(x)$, the expectation of $\tilde{\mu}_m$ is given by (Gamerman 1997)

$$\begin{aligned} E(\tilde{\mu}_m) &= \frac{1}{m} \sum_{i=1}^{m} E\{\mathrm{w}(X^{(i)})\} = E\{\mathrm{w}(X)\} \\ &= \int_{\mathbb{X}} \mathrm{w}(x) \cdot g(x)\, \mathrm{d}x = \int_{\mathbb{X}} H(x)\, \mathrm{d}x = \mu, \end{aligned}$$

indicating that $\tilde{\mu}_m$ is an unbiased estimator.

— The variance of $\tilde{\mu}_m$ is $\mathrm{Var}(\bar{\mu}_m) = \sigma^2/m$, where

$$\sigma^2 = \mathrm{Var}\{\mathrm{w}(X)\} = \int_{\mathbb{X}} \{\mathrm{w}(x) - \mu\}^2 \cdot g(x)\, \mathrm{d}x.$$

12.2• Consistency

— The strong law of large numbers states that $\tilde{\mu}_m \xrightarrow{\text{a.s.}} \mu$; i.e.,

$$\Pr\left(\lim_{m \to \infty} \tilde{\mu}_m = \mu\right) = 1.$$

— The weak law of large numbers states that $\tilde{\mu}_m \xrightarrow{\text{P}} \mu$; i.e., for any $\varepsilon > 0$,

$$\lim_{m\to\infty} \Pr\left(|\tilde{\mu}_m - \mu| \geqslant \varepsilon\right) = 0.$$

12.3• Asymptotic normality

— The central limit theorem states that

$$\frac{\tilde{\mu}_m - \mu}{\sigma/\sqrt{m}} \xrightarrow{\text{D}} N(0,1),$$

13• Choice of the proposal density

- By properly choosing $g(\cdot)$, one can reduce the variance of the estimator $\tilde{\mu}_m$ substantially.
- A good candidate for $g(\cdot)$ is the one that is close to the shape of $H(x)$.

Example 3.3 (Comparison of classical MC integration with IS estimator). Consider to evaluate the integral $\mu = \int_{\mathbb{X}} H(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, x_2)^\top$, $\mathbb{X} = [-1,1]^2$ and

$$H(\boldsymbol{x}) = \exp\{-90(x_1 - 0.5)^2 - 10(x_2 + 0.1)^4\}.$$

Please compare the classical MC integration $\bar{\mu}_m$ with the IS estimator $\tilde{\mu}_m$.

Solution: Let the density of the uniform distribution over $\mathbb{X}$ (denoted by $U(\mathbb{X})$) be $f(\boldsymbol{x}) = 1/4$ for $\boldsymbol{x} \in \mathbb{X}$. First we can write

$$\mu = 4\int_{\mathbb{X}} H(\boldsymbol{x}) \cdot f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

Second, we generate $m = 6{,}000$ random samples $\{\boldsymbol{u}^{(i)}\}_{i=1}^m \overset{\text{iid}}{\sim} U(\mathbb{X})$ and estimate μ by the empirical mean

$$\bar{\mu}_m = \frac{4}{m}\sum_{i=1}^m H(\boldsymbol{u}^{(i)}).$$

On the other hand, we choose the proposal density $g(\boldsymbol{x})$ proportional to

$$\exp\{-90(x_1 - 0.5)^2 - 10(x_2 + 0.1)^2\}, \quad \boldsymbol{x} \in \mathbb{X}.$$

In fact, $g(\boldsymbol{x})$ is a product of two independent truncated normal densities (see Exercise 1.1):

$$g(\boldsymbol{x}) = \mathrm{TN}\left(x_1 \middle| 0.5, \frac{1}{180}; -1, 1\right) \times \mathrm{TN}\left(x_2 \middle| -0.1, \frac{1}{20}; -1, 1\right).$$

We also generate $m = 6{,}000$ random samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^{m} \overset{\text{iid}}{\sim} g(\boldsymbol{x})$ and estimate μ by the importance sampling estimator

$$\tilde{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \frac{H(\boldsymbol{x}^{(i)})}{g(\boldsymbol{x}^{(i)})}.$$

Figure 3.2 illustrates that $\tilde{\mu}_m$ has much greater stability and faster speed of convergence than $\bar{\mu}_m$.

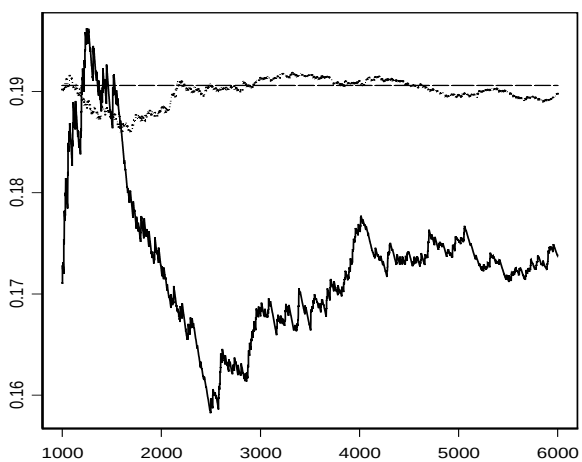

Figure 3.2 Comparison between the convergence of the empirical mean $\bar{\mu}_m$ (solid line) and the importance sampling estimator $\tilde{\mu}_m$ (dotted line). The final values are 0.1739 and 0.1898, respectively, for a true value of 0.1906 (dashed line).

Comment 1: The chosen $g(\boldsymbol{x})$ is close to the shape of $H(\boldsymbol{x})$. Next, the support of $g(\boldsymbol{x})$ is the same as the domain of $H(\boldsymbol{x})$. ‖

3.3.2 The weighted estimator

14• The issue

- Let $h(x) \geqslant 0$ be completely known, two densities $f(x) = c_1^{-1} f^*(x)$ and $g(x) = c_2^{-1} g^*(x)$ have the same support $\mathbb{X}$, where c_1, c_2 are unknown and $f^*(x), g^*(x)$ are known.

- The aim is to compute $\mu = \int_{\mathbb{X}} h(x) \cdot f(x)\,\mathrm{d}x$.

15• The definition of the weighted estimator

- A practical alternative to the importance sampling estimator $\tilde{\mu}_m$ is to use the *weighted estimator*:

$$\hat{\mu}_m = \sum_{i=1}^{m} \omega_i h(X^{(i)}), \tag{3.16}$$

where

$$\omega_i = \frac{f(X^{(i)})/g(X^{(i)})}{\sum_{j=1}^{m} f(X^{(j)})/g(X^{(j)})} = \frac{f^*(X^{(i)})/g^*(X^{(i)})}{\sum_{j=1}^{m} f^*(X^{(j)})/g^*(X^{(j)})},$$

and $X^{(1)}, \ldots, X^{(m)} \overset{\text{iid}}{\sim} g(x)$.

15.1• Proof of (3.16)

— Since $c_1 = \int_{\mathbb{X}} f^*(x)\,\mathrm{d}x$, we have

$$\begin{aligned}
\mu &= \int_{\mathbb{X}} h(x) \cdot c_1^{-1} f^*(x)\,\mathrm{d}x = \frac{\int_{\mathbb{X}} h(x) f^*(x)\,\mathrm{d}x}{\int_{\mathbb{X}} f^*(x)\,\mathrm{d}x} \\
&= \frac{\int_{\mathbb{X}} \frac{h(x) f^*(x)}{g(x)} \cdot g(x)\,\mathrm{d}x}{\int_{\mathbb{X}} \frac{f^*(x)}{g(x)} \cdot g(x)\,\mathrm{d}x} \qquad [\text{let } X^{(1)}, \ldots, X^{(m)} \overset{\text{iid}}{\sim} g(x)] \\
&\approx \frac{\frac{1}{m}\sum_{i=1}^{m} h(X^{(i)}) \frac{f^*(X^{(i)})}{g(X^{(i)})}}{\frac{1}{m}\sum_{j=1}^{m} \frac{f^*(X^{(j)})}{g(X^{(j)})}} \\
&= \frac{\sum_{i=1}^{m} h(X^{(i)}) \cdot f^*(X^{(i)})/g^*(X^{(i)})}{\sum_{j=1}^{m} f^*(X^{(j)})/g^*(X^{(j)})} \\
&= \sum_{i=1}^{m} \omega_i h(X^{(i)}),
\end{aligned}$$

which means (3.16). □

15.2• A major advantage of the weighted estimator

— A major advantage of using $\hat{\mu}_m$ instead of the unbiased estimator $\tilde{\mu}_m$ is that we only need to know the ratio function $f(x)/g(x)$; i.e., both $f(x)$ and $g(x)$ could be known up to a constant.

16• Basic properties of the weighted estimator

- The weighted estimator $\hat{\mu}_m$ is only asymptotically unbiased, see (3.17), but it often has a smaller *mean square error* (MSE) than the unbiased estimator $\tilde{\mu}_m$ (Casella & Robert 1998).

16.1• Proof of above facts

— Let $Y = h(X)f(X)/g(X) = \mathrm{w}(X)$ and $W = f(X)/g(X)$, and let $\bar{Y}$ and $\bar{W}$ be the corresponding sample means. We rewrite $\hat{\mu}_m$ as

$$\hat{\mu}_m = \frac{\bar{Y}}{\bar{W}} = \frac{\tilde{\mu}_m}{\bar{W}} = \tilde{\mu}_m\Big\{1 - (\bar{W}-1) + (\bar{W}-1)^2 - \cdots\Big\}.$$

— Hence, we obtain (Liu 2001, p.35–36)

$$\begin{aligned} E_g(\hat{\mu}_m) &\approx \mu - \frac{\mathrm{Cov}_g(W,Y)}{m} + \frac{\mu \mathrm{Var}_g(W)}{m}, \\ \mathrm{Var}_g(\hat{\mu}_m) &\approx \frac{1}{m}\Big\{\mu^2 \mathrm{Var}_g(W) - 2\mu \mathrm{Cov}_g(W,Y) + \mathrm{Var}_g(Y)\Big\}. \end{aligned} \tag{3.17}$$

— Note that the MSE of $\tilde{\mu}_m$ is

$$\mathrm{MSE}(\tilde{\mu}_m) = E_g(\tilde{\mu}_m - \mu)^2 = \frac{\mathrm{Var}_g(Y)}{m},$$

so that we have

$$\begin{aligned} \mathrm{MSE}(\hat{\mu}_m) &= \mathrm{Var}_g(\hat{\mu}_m) + \{E_g(\hat{\mu}_m) - \mu\}^2 \\ &= \mathrm{MSE}(\tilde{\mu}_m) + \frac{1}{m}\Big\{\mu^2 \mathrm{Var}_g(W) - 2\mu \mathrm{Cov}_g(W,Y)\Big\} \\ &\quad + O(m^{-2}). \end{aligned}$$

— Without loss of generality, we assume that $\mu > 0$, then

$$\mathrm{MSE}(\hat{\mu}_m) < \mathrm{MSE}(\tilde{\mu}_m)$$

when

$$\mu < \frac{2\mathrm{Cov}_g(W,Y)}{\mathrm{Var}_g(W)}.$$

□

3.4 Variance Reduction Techniques

3.4.1 Antithetic variables

17• An alternative to the importance sampling estimator

- Apart from the importance sampling estimator, using *antithetic variables* is an alternative to reduce the variance of the Monte Carlo integration.
- In §3.2.1, we introduced the classical Monte Carlo integration, where i.i.d. random samples were generated.
- In fact, the variance of the estimator can be reduced by generating correlated random samples.

18• The issue

- Let $X^{(1)}, \ldots, X^{(m)} \overset{\text{iid}}{\sim} F(x)$ and $Y^{(1)}, \ldots, Y^{(m)} \overset{\text{iid}}{\sim} F(x)$.
- It is easy to show that

$$\hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m} h(X^{(i)}) \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{m}\sum_{i=1}^{m} h(Y^{(i)}) \tag{3.18}$$

  are two unbiased estimators of

$$\mu = \int_{\mathbb{X}} h(x)\,\mathrm{d}F(x). \tag{3.19}$$

- Based on both $\hat{\mu}_1$ and $\hat{\mu}_2$, we wonder if there exists another unbiased estimator $\hat{\mu}_3$, which has smaller variance than $\hat{\mu}_1$ or $\hat{\mu}_2$.
- If yes, how to find it?
- What are the conditions for achieving a smaller variance?

19• The basic principle

- Let X and Y be two r.v.'s and they have the same distribution function $F(\cdot)$; i.e., $X \overset{\text{d}}{=} Y$.

- Let
$$\mu = \int_{\mathbb{X}} x \, \mathrm{d}F(x) \quad \text{and} \quad \sigma^2 = \int_{\mathbb{X}} (x-\mu)^2 \, \mathrm{d}F(x)$$
denote the population mean and population variance, respectively, then, $E(X) = E(Y) = \mu$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = \sigma^2$.

- Define $Z = (X+Y)/2$, we have the following variance formula
$$\mathrm{Var}(Z) = \frac{1}{4}\mathrm{Var}(X) + \frac{1}{4}\mathrm{Var}(Y) + \frac{1}{2}\mathrm{Cov}(X,Y).$$

- If $\mathrm{Cov}(X,Y) < 0$, then
$$E(Z) = \mu, \quad \text{and} \quad \mathrm{Var}(Z) < \frac{\sigma^2}{4} + \frac{\sigma^2}{4} = \frac{\sigma^2}{2} < \sigma^2.$$

19.1• Remarks

— If we treat X as $\hat{\mu}_1$ and Y as $\hat{\mu}_2$, we have the following interpretation.

— Starting from two unbiased estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ with a common variance, we can find the third unbiased estimator $\hat{\mu}_3 = (\hat{\mu}_1 + \hat{\mu}_2)/2$ having smaller variance than $\hat{\mu}_1$ or $\hat{\mu}_2$, provided that $\hat{\mu}_1$ and $\hat{\mu}_2$ are negatively correlated.

— This idea can be utilized to reduce the variance of a Monte Carlo estimator of an integral.

20• A sufficient condition for achieving negative correlation

- The following theorem provides a sufficient condition for achieving negative correlation.

- Its proof exploits coupled r.v.'s.

Theorem 3.1 (Sufficient conditions for positive/negative correlation). Let X be a r.v. and the r.v.'s $g_1(X)$ and $g_2(X)$ have finite second moments.

(1) If the functions $g_1(x)$ and $g_2(x)$ are both increasing or both decreasing, then $\mathrm{Cov}\{g_1(X), g_2(X)\} \geqslant 0$.

(2) If the function $g_1(x)$ is increasing while the function $g_2(x)$ is decreasing, or vice versa, then $\mathrm{Cov}\{g_1(X), g_2(X)\} \leqslant 0$. ‖

Proof: (1) Consider another r.v. Y independent of X and $Y \overset{\mathrm{d}}{=} X$. Thus

$$E\{g_i(Y)\} = E\{g_i(X)\}, \quad i = 1, 2.$$

If the functions $g_1(x)$ and $g_2(x)$ are both increasing or both decreasing, then the product

$$\{g_1(X) - g_1(Y)\}\{g_2(X) - g_2(Y)\} \geqslant 0.$$

Hence,

$$\begin{aligned} 0 &\leqslant E\Big[\{g_1(X) - g_1(Y)\}\{g_2(X) - g_2(Y)\}\Big] \\ &= E\{g_1(X)g_2(X)\} - E\{g_1(X)\}E\{g_2(Y)\} \\ &\quad + E\{g_1(Y)g_2(Y)\} - E\{g_1(Y)\}E\{g_2(X)\} \\ &= 2\mathrm{Cov}\{g_1(X), g_2(X)\}. \end{aligned}$$

Similarly, we can prove (2). □

Example 3.4 (Antithetic uniform estimators). Let $h(x)$ in (3.19) be increasing. If $U^{(1)}, \ldots, U^{(m)} \overset{\text{iid}}{\sim} U(0,1)$, then $1-U^{(1)}, \ldots, 1-U^{(m)} \overset{\text{iid}}{\sim} U(0,1)$. Define

$$X^{(i)} = F^{-1}(U^{(i)}) \quad \text{and} \quad Y^{(i)} = F^{-1}(1 - U^{(i)}), \quad i = 1, \ldots, m,$$

we know that $X^{(1)}, \ldots, X^{(m)} \overset{\text{iid}}{\sim} F(x)$ and $Y^{(1)}, \ldots, Y^{(m)} \overset{\text{iid}}{\sim} F(x)$. Let

$$\begin{aligned} g_1(U^{(i)}) &= h(X^{(i)}) = h(F^{-1}(U^{(i)})) \quad \text{and} \\ g_2(U^{(i)}) &= h(Y^{(i)}) = h(F^{-1}(1 - U^{(i)})), \end{aligned}$$

then $g_1(\cdot)$ is increasing while $g_2(\cdot)$ is decreasing. According to Theorem 3.1(2), we have

$$\mathrm{Cov}\{h(X^{(i)}), h(Y^{(i)})\} = \mathrm{Cov}\{g_1(U^{(i)}), g_2(U^{(i)})\} \leqslant 0. \tag{3.20}$$

On the other hand, since $U^{(i)}$ is independent of $U^{(j)}$ $(j \neq i)$, we obtain

$$\mathrm{Cov}\{h(X^{(i)}), h(Y^{(j)})\} = \mathrm{Cov}\{h(F^{-1}(U^{(i)})), h(F^{-1}(1 - U^{(j)}))\} = 0 \tag{3.21}$$

for all $i \neq j$. Let $\hat{\mu}_1$ and $\hat{\mu}_2$ be defined by (3.18), please find an unbiased estimator of μ with a smaller variance than $\hat{\mu}_1$ or $\hat{\mu}_2$.

Solution: We have

$$\begin{aligned}
\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) &= \mathrm{Cov}\left\{ \frac{1}{m}\sum_{i=1}^{m} h(X^{(i)}),\ \frac{1}{m}\sum_{j=1}^{m} h(Y^{(j)}) \right\} \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{Cov}\{h(X^{(i)}), h(Y^{(j)})\} \\
&\overset{(3.21)}{=} \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Cov}\{h(X^{(i)}), h(Y^{(i)})\} \\
&\overset{(3.20)}{\leqslant} 0.
\end{aligned}$$

That is, $\hat{\mu}_1$ and $\hat{\mu}_2$ are negatively correlated. According to remarks in **19.1•**, $\hat{\mu}_3 = (\hat{\mu}_1 + \hat{\mu}_2)/2$ is an unbiased estimator of μ, having smaller variance than $\hat{\mu}_1$ or $\hat{\mu}_2$. ‖

21• How to choose r.v.'s with negative correlation?

- How to choose random samples $\{X^{(i)}\}$ and $\{Y^{(i)}\}$ such that (3.20) holds is a key step for the application of the technique of antithetic variables.
- When $U \sim U(0,1)$, we know that $U \overset{\mathrm{d}}{=} 1 - U$.
- This property of uniform r.v.'s is utilized in Example 3.4 to induce the desired random samples $\{X^{(i)}\}$ and $\{Y^{(i)}\}$.
- However, this technique only applies to the case that $F^{-1}(\cdot)$ is easy to be computed.
- Let $f(\cdot) = F'(x)$ be the pdf, if $f(\cdot)$ is symmetric around the point θ, the antithetic variable Y can be chosen to be $Y = 2\theta - X$ since $Y - \theta = -(X - \theta)$.
- For example, when $f(x) = N(x|\theta, \sigma^2)$, we have $X - \theta \overset{\mathrm{d}}{=} -(X - \theta) \sim N(0, \sigma^2)$.
- This approach can be extended to the case where $f(\cdot)$ is not symmetric and θ is replaced by the mode or median of the distribution function.

3.4.2 Control variables

22• The method of control variables

- Suppose that there exists a function $h_0(\cdot)$ such that the expectation of $h_0(X)$ under the density $f(x)$ is known and $h_0(\cdot)$ is usually related to $h(\cdot)$.
- The method of control variables incorporates this additional information into Monte Carlo integration to reduce the variance of the estimator of
$$I = \int_{\mathbb{X}} h(x) \cdot f(x)\, \mathrm{d}x.$$
- Note that the Monte Carlo estimator of I is
$$\hat{I}_1 = \frac{1}{m} \sum_{i=1}^{m} h(X^{(i)}),$$
where $X^{(1)}, \ldots, X^{(m)} \overset{\text{iid}}{\sim} f(x)$.

23• Construction of a class of estimators

- Let $X \sim f(x)$. The expectation of $h_0(X)$ under the density $f(x)$ is denoted by $E_f\{h_0(X)\}$.
- An unbiased estimator of $E_f\{h_0(X)\}$ is
$$\hat{I}_2 = \frac{1}{m} \sum_{i=1}^{m} h_0(X^{(i)}).$$
- Now, we consider a class of estimators
$$\hat{I}_3(\beta) = \hat{I}_1 + \beta[\hat{I}_2 - E_f\{h_0(X)\}],$$
indexed by a real number β.

23.1• An optimal estimator

— We need to select an optimal β, say β^*, such that $\mathrm{Var}\{\hat{I}_3(\beta^*)\} \leqslant \mathrm{Var}(\hat{I}_1)$.

— Note that

$$E\{\hat{I}_3(\beta)\} = E(\hat{I}_1) + \beta[E(\hat{I}_2) - E_f\{h_0(X)\}] = I,$$

i.e., $\hat{I}_3(\beta)$ is also an unbiased estimator of I.

— The variance of $\hat{I}_3(\beta)$ is

$$\mathrm{Var}\{\hat{I}_3(\beta)\} = \mathrm{Var}(\hat{I}_1) + \beta^2\mathrm{Var}(\hat{I}_2) + 2\beta\mathrm{Cov}(\hat{I}_1, \hat{I}_2),$$

which is minimized at

$$\beta^* = -\frac{\mathrm{Cov}(\hat{I}_1, \hat{I}_2)}{\mathrm{Var}(\hat{I}_2)}.$$

— Thus,

$$\mathrm{Var}\{\hat{I}_3(\beta^*)\} = \mathrm{Var}(\hat{I}_1)\left\{1 - \frac{\mathrm{Cov}^2(\hat{I}_1, \hat{I}_2)}{\mathrm{Var}(\hat{I}_1)\mathrm{Var}(\hat{I}_2)}\right\} \leqslant \mathrm{Var}(\hat{I}_1).$$

— Therefore, the control variable estimator is

$$\hat{I}_3(\beta^*) = \frac{1}{m}\sum_{i=1}^{m}\{h(X^{(i)}) + \beta^* h_0(X^{(i)})\} - \beta^* E_f\{h_0(X)\}.$$

Exercise 3

3.1 Use the first–order Laplace approximation (3.7) to evaluate the gamma integral

$$\int_a^b \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\,\mathrm{e}^{-\beta x}\,\mathrm{d}x.$$

Compare the approximate results with the exact results for $\alpha = 5$, $\beta = 0.5$, $(a, b) = (7, 9)$, $(6, 10)$, $(2, 14)$, $(15.987, \infty)$, respectively.

3.2 Use the classical Monte Carlo integration (3.9) to evaluate the following integrals

$$\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,\mathrm{e}^{-(x-x_0)^2/2}\,\mathrm{d}x \quad \text{and} \quad \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x,$$

where x_0 is a known constant.

3.3 Let $U = \{U_{(1)}, \ldots, U_{(m+1)}\}$ be an ordered sample from $U(0,1)$. If the derivative $H'(x)$ is bounded on $(0,1)$, the estimator (Yakowitz *et al.* 1978)

$$\delta(U) = \sum_{i=0}^{m+1} H(U_{(i)})(U_{(i+1)} - U_{(i)}) \tag{3.22}$$

has a variance of order $O(m^{-2})$, where $U_{(0)} \hat{=} 0$ and $U_{(m+2)} \hat{=} 1$.

3.4 Consider the integral $\int_0^1 \cos(\pi x/2)\,\mathrm{d}x = 2/\pi$. Interpreting this as an expectation relative to the uniform distribution on $(0,1)$, show that

$$\mathrm{Var}\,\{\cos(\pi X/2)\} = 1/2 - (2/\pi)^2 \approx 0.095.$$

The importance density $g(x) = 3(1-x^2)/2$ roughly resembles the integrand $\cos(\pi x/2)$. Demonstrate numerically that

$$\mathrm{Var}\left\{\frac{2\cos(\pi Y/2)}{3(1-Y^2)}\right\} \approx 0.00099,$$

where Y is sampled from $g(y)$. Thus, the importance sampling reduces the variance of the classical Monte Carlo estimator by almost a factor of 100.

Chapter 4

Markov Chain Monte Carlo Methods

1• Difference of a probability model and a statistical model

- In a probability model, all parameters are *known*.
  - For example, let $X \sim N(\mu, \sigma^2)$, where $\mu = 0.1$ and $\sigma = 0.2$.
  - We can compute $\Pr(-0.1 < X < 0.9)$, $E(X)$, $\mathrm{Var}(X)$, median, quantiles, and so on.
- In a statistical model, all parameters are *unknown*.
  - For example, let $X \sim f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
  - Based on a random sample $X_1, \ldots, X_n$, we can find the estimates of $\boldsymbol{\theta}$.

2• Difference of frequentist statistics and bayesian statistics

- In the frequentist statistics, all parameters are *fixed* but *unknown*.
- In the Bayesian statistics, all parameters are *random variables*.
  - Bayesian statistics transfers a statistical model into a probability model.
  - Thus, in Bayesian statistics, it only involves *computation issues*, but does not involve estimation issues.

3• Core of modern bayesian statistics

3.1• Case I: Complete-data problems

— In the Bayesian statistics, all unknown parameters $\boldsymbol{\theta}$ are treated as random variables.

— Let Y_{com} be the completely observed data. Given $\boldsymbol{\theta}$, the likelihood function is $L(\boldsymbol{\theta}|Y_{\text{com}}) = f(Y_{\text{com}}|\boldsymbol{\theta})$.

— The prior distribution of $\boldsymbol{\theta}$ is specified by $\pi(\boldsymbol{\theta})$.

— The posterior distribution of $\boldsymbol{\theta}$ is then given by

$$p(\boldsymbol{\theta}|Y_{\text{com}}) = \frac{f(Y_{\text{com}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(Y_{\text{com}})} \propto L(\boldsymbol{\theta}|Y_{\text{com}})\pi(\boldsymbol{\theta}).$$

Aim 1: Finding the exact expression for each marginal posterior density $p(\theta_i|Y_{\text{com}})$, $i = 1, \ldots, d$.

Aim 2: Generating posterior samples from each $p(\theta_i|Y_{\text{com}})$, which is equivalent to generating posterior samples from the joint posterior density $p(\theta_1, \ldots, \theta_d|Y_{\text{com}})$.

3.2• Case II: Incomplete-data problems

— Now, both the unknown parameters $\boldsymbol{\theta}$ and the missing data Y_{mis} are treated as random variables.

— Let Y_{obs} be the observed data.

Aim 1: Finding the observed posterior density $p(\theta_1, \ldots, \theta_d|Y_{\text{obs}})$ or the exact expression of each marginal observed posterior density $p(\theta_i|Y_{\text{obs}})$, $i = 1, \ldots, d$.

Aim 2: Generating samples from the joint posterior distribution, say, $p(\theta_1, \ldots, \theta_d, Y_{\text{mis}}|Y_{\text{obs}})$.

4• Problem transfer

- Usually, direct sampling from $p(\theta_1, \ldots, \theta_d, Y_{\text{mis}}|Y_{\text{obs}})$ is very difficult.

- However, it is relatively easy to sample from all *full* conditional distributions:
  - $p(\theta_1|Y_{\text{obs}}, \theta_2, \theta_3, \ldots, \theta_d, Y_{\text{mis}})$;
  - $p(\theta_2|Y_{\text{obs}}, \theta_1, \theta_3, \ldots, \theta_d, Y_{\text{mis}})$;
  - ……
  - $p(\theta_d|Y_{\text{obs}}, \theta_1, \theta_2, \ldots, \theta_{d-1}, Y_{\text{mis}})$;
  - $f(Y_{\text{mis}}|Y_{\text{obs}}, \theta_1, \theta_2, \ldots, \theta_{d-1}, \theta_d)$.
- What is the definition of a full conditional distribution?
- If we treat the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ as one *block* and note that Y_{mis} could be a vector, sometimes, it is more convenient to sample
  - from the complete-data posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}}, Y_{\text{mis}})$ and
  - from the conditional predictive distribution $f(Y_{\text{mis}}|Y_{\text{obs}}, \boldsymbol{\theta})$.
- Because Y_{obs} is constant in the whole sampling process, the problem can be reformulated as:
  - Given both conditional pdfs $f(x|y)$ and $f(y|x)$, how to determine the joint pdf $f(x, y)$ or how to sample from $f(x, y)$.

4.1 Bayes Formulae and Inverse Bayes Formulae (IBF)

5• Bayes formulae

- The Bayes formula, Bayes rule, or Bayes theorem, was published posthumously and named after Reverend Thomas Bayes (1763) and has been the foundation of Bayesian inference.

5.1• Three forms of Bayes theorem

- The first form in events: $\Pr(\mathbb{A}|\mathbb{B}) = \Pr(\mathbb{B}|\mathbb{A}) \Pr(\mathbb{A}) / \Pr(\mathbb{B})$.
- The second form in densities: $f_{(X|Y)}(x|y) = f_{(Y|X)}(y|x) f_X(x) / f_Y(y)$.
- The third form in mixture: $f_{(X|\mathbb{B})}(x|\mathbb{B}) = \Pr(\mathbb{B}|x) f_X(x) / \Pr(\mathbb{B})$.

6• Statistics: classical and modern

- Historically, the Bayesian inference is the first default paradigm of statistical inference and was referred to as the "classical procedure" by earlier European literature in contrast to the "modern formulation" of R.A. Fisher; see the later English translation of Gnedenko (1962, p.409, p.419).
- Many modern writings seem to ignore this historical fact and call the latter approach "classical" while referring to the older one as "modern".

6.1• Bayesian paradigm

— The Bayesian paradigm starts with a prior distribution of the parameter that reflects our information about, or subjective evaluation of, competing values of the parameter before collecting data, see §4.2.1 for more detail.

— And then update the prior distribution to the posterior distribution with available data according to the Bayes formula, see, e.g., (4.1) or (4.14).

4.1.1 The point-wise, function-wise and sampling-wise IBF

7• Objective of this subsection

- We know that one marginal density $f_X(x)$ and one conditional density $f_{(Y|X)}(y|x)$ can uniquely determine the joint density $f_{(X,Y)}(x,y)$.
- Second, it is well known that two marginal densities $f_X(x)$ and $f_Y(y)$ are not sufficient to determine the joint density $f_{(X,Y)}(x,y)$ uniquely.

  — For example, for any given $\alpha \in [-1,1]$,

  $$f_{(X,Y)}(x,y) = f_X(x)f_Y(y)\Big[1+\alpha\{2F_X(x)-1\}\{2F_Y(y)-1\}\Big]$$

  is a joint pdf of (X,Y) with marginal pdfs $f_X(x)$ and $f_Y(y)$, where $F_X(x)$ and $F_Y(y)$ are corresponding cdfs of X and Y.
- Given both conditional pdfs $f_{(X|Y)}(x|y)$ and $f_{(Y|X)}(y|x)$, we wonder if there exists a joint pdf $f_{(X,Y)}(x,y)$ with them as its conditional pdfs.

- If yes, we wonder under what kind of conditions, the $f_{(X,Y)}(x,y)$ exists.
- And then, how to find the marginal $f_X(x)$ or $f_{(X,Y)}(x,y)$?

8• Support, product space and non-product space

- Since the derivation of the three IBF from the Bayes formula with rigor involving integration in the support of a r.v., we first need to distinguish two important notions: The product measurable space and the non-product measurable space.

8.1• Marginal supports and joint support

— Let $f_X(x)$ be the pdf of the r.v. X taking values in the space $\mathcal{X}$, then $\mathcal{S}_X = \{x\colon\ f_X(x) > 0,\ x \in \mathcal{X}\}$ is called the *marginal support* of X.

— For example, if $X \sim U(0,1)$, then the pdf of X is

$$f_X(x) = \begin{cases} 1, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, $\mathcal{X} = \mathbb{R}$ and $\mathcal{S}_X = (0,1)$.

— Similarly, let $f_Y(y)$ be the pdf of the r.v. Y taking values in the space $\mathcal{Y}$, then $\mathcal{S}_Y = \{y\colon\ f_Y(y) > 0,\ y \in \mathcal{Y}\}$ is the marginal support of Y.

— Let $f_{(X,Y)}(x,y)$ denote the joint density of (X,Y), then

$$\mathcal{S}_{(X,Y)} = \{(x,y)\colon\ f_{(X,Y)}(x,y) > 0,\ (x,y) \in (\mathcal{X},\mathcal{Y})\}$$

is called the *joint support* of (X,Y).

8.2• Conditional supports

— The conditional supports of $X|(Y=y)$ and $Y|(X=x)$ are defined by

$$\begin{aligned} \mathcal{S}_{(X|Y)}(y) &= \{x\colon\ f_{(X|Y)}(x|y) > 0,\ x \in \mathcal{X}\} \quad \forall\, y \in \mathcal{S}_Y \ \text{ and} \\ \mathcal{S}_{(Y|X)}(x) &= \{y\colon\ f_{(Y|X)}(y|x) > 0,\ y \in \mathcal{Y}\} \quad \forall\, x \in \mathcal{S}_X. \end{aligned}$$

— In practice, we usually have

$$\mathcal{S}_{(Y|X)}(x) \subseteq \mathcal{S}_Y \quad \forall\, x \in \mathcal{S}_X \quad \text{and} \quad \mathcal{S}_{(X|Y)}(y) \subseteq \mathcal{S}_X \quad \forall\, y \in \mathcal{S}_Y.$$

8.3• Product space and non-product space

— If $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y$, we say $\mathcal{S}_{(X,Y)}$ is a *product space*; otherwise, it is called a *non-product space*.

— For example, let $(X, Y) \sim N_2(\mathbf{0}, \boldsymbol{I})$, then $\mathcal{S}_{(X,Y)} = \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \mathcal{S}_X \times \mathcal{S}_Y$; that is, $\mathcal{S}_{(X,Y)}$ is a product space.

— If $(X, Y) \sim U(\mathbb{B}_2)$, where $\mathbb{B}_2 = \{(x, y)\colon x^2 + y^2 \leqslant 1\}$, then $\mathcal{S}_{(X,Y)} = \mathbb{B}_2 \neq [-1, 1] \times [-1, 1] = \mathcal{S}_X \times \mathcal{S}_Y$; that is, $\mathcal{S}_{(X,Y)}$ is a non-product space.

9• Three IBF formulae and the conditions

- We always have the following identity

$$f_{(X|Y)}(x|y) f_Y(y) = f_{(Y|X)}(y|x) f_X(x), \quad (x, y) \in \mathcal{S}_{(X,Y)}, \tag{4.1}$$

  provided that the joint density $f_{(X,Y)}(x, y)$ exists.

  — A counter example is as follows. Let $X \sim N(0, 1)$ and $Y \,\hat{=}\, X^2$, where the joint density $f_{(X,Y)}(x, y)$ does not exist (see Exercise 4.1).

- Under the condition of *product space*, we have $\mathcal{S}_{(Y|X)}(x) = \mathcal{S}_Y$ for all $x \in \mathcal{S}_X$, and $\mathcal{S}_{(X|Y)}(y) = \mathcal{S}_X$ for all $y \in \mathcal{S}_Y$.

9.1• Point-wise formula

— From (4.1), by division, we obtain

$$f_Y(y) = \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)} \cdot f_X(x), \qquad x \in \mathcal{S}_X, \;\; y \in \mathcal{S}_Y. \tag{4.2}$$

— Integrating this identity with respect to y on support $\mathcal{S}_Y$; i.e.,

$$\int_{\mathcal{S}_Y} f_Y(y)\,\mathrm{d}y = \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)} \cdot f_X(x)\,\mathrm{d}y = f_X(x) \cdot \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)}\,\mathrm{d}y,$$

we immediately have the following *point-wise formula*:

$$f_X(x) = \left\{ \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)}\,\mathrm{d}y \right\}^{-1}, \quad \text{for any } x \in \mathcal{S}_X. \tag{4.3}$$

— By symmetry, we obtain

$$f_Y(y) = \left\{ \int_{\mathcal{S}_X} \frac{f_{(X|Y)}(x|y)}{f_{(Y|X)}(y|x)} \, \mathrm{d}x \right\}^{-1}, \quad \text{for any } y \in \mathcal{S}_Y. \tag{4.4}$$

— Thus, two conditional pdfs can uniquely determine marginal densities provided that the integrals in (4.3) and (4.4) exist.

9.2• Function-wise formula

— From (4.2), we obtain the following *function-wise formula*:

$$\begin{aligned} f_X(x) &= f_Y(y) \cdot \frac{f_{(X|Y)}(x|y)}{f_{(Y|X)}(y|x)} \qquad \forall\, y \in \mathcal{S}_Y \\ &= f_Y(y_0) \cdot \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)} \\ &\overset{(4.4)}{=} \left\{ \int_{\mathcal{S}_X} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)} \, \mathrm{d}x \right\}^{-1} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)}, \end{aligned} \tag{4.5}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

— The function-wise formula (4.5) shows that two fully specified conditional densities $f_{(X|Y)}(x|y)$ and $f_{(Y|X)}(y|x)$ *overly* determine (i.e., requiring more than enough information to characterize) the marginal densities.

— In fact, $f_{(X|Y)}(x|y_0)$ and $f_{(Y|X)}(y_0|x)$ are enough to determine the $f_X(x)$.

9.3• Sampling-wise formula

— By dropping the normalizing constant in (4.5), we obtain the so-called *sampling-wise formula*:

$$f_X(x) \propto \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)}, \tag{4.6}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

9.4• One merit of function-wise and sampling-wise IBF

— Often in practice, we know $f_{(X|Y)}(x|y)$ only up to a normalizing constant.

— In other words, let $f_{(X|Y)}(x|y) = c(y) \cdot g(x|y)$, where $c(y)$ is unknown and $g(x|y)$ is completely known, then the function-wise IBF (4.5) and sampling-wise IBF (4.6) still hold if we replace $f_{(X|Y)}(x|y_0)$ by $g(x|y_0)$.

Proof: We have $f_{(X|Y)}(x|y_0) = c(y_0) \cdot g(x|y_0)$ so that

$$\begin{aligned} f_X(x) &= \left\{ \int_{\mathcal{S}_X} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)} \, \mathrm{d}x \right\}^{-1} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)} \\ &= \left\{ \int_{\mathcal{S}_X} \frac{c(y_0) \cdot g(x|y_0)}{f_{(Y|X)}(y_0|x)} \, \mathrm{d}x \right\}^{-1} \frac{c(y_0) \cdot g(x|y_0)}{f_{(Y|X)}(y_0|x)} \\ &= \left\{ \int_{\mathcal{S}_X} \frac{g(x|y_0)}{f_{(Y|X)}(y_0|x)} \, \mathrm{d}x \right\}^{-1} \frac{g(x|y_0)}{f_{(Y|X)}(y_0|x)} \\ &\propto \frac{g(x|y_0)}{f_{(Y|X)}(y_0|x)}, \end{aligned}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$. □

Example 4.1 (Bivariate normal distribution). Assume that

$$\begin{aligned} X|(Y = y) &\sim N(\mu_1 + \rho(y - \mu_2), 1 - \rho^2) \quad \text{and} \\ Y|(X = x) &\sim N(\mu_2 + \rho(x - \mu_1), 1 - \rho^2). \end{aligned}$$

Find the marginal distribution of X and the joint distribution of (X, Y).

Solution: Note that $\mathcal{S}_{(X,Y)} = \mathcal{S}_{(X|Y)}(y) \times \mathcal{S}_Y$. Since $\mathcal{S}_{(X|Y)}(y) = \mathcal{S}_X = \mathbb{R}$, we have $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y = \mathbb{R}^2$. From (4.3), we obtain

$$\{f_X(x)\}^{-1} = \sqrt{2\pi} \exp\{(x - \mu_1)^2/2\},$$

which means $X \sim N(\mu_1, 1)$. Therefore, the joint distribution of (X, Y) exists and is bivariate normal with means μ_1 and μ_2, unit variances and correlation coefficient ρ.

Alternative solution: When using (4.3), we need to evaluate an integral. In contrast, using (4.6), the integration can be avoided. In fact, let the arbitrary y_0 be μ_2, then

$$f_X(x) \propto \exp\{-(x - \mu_1)^2/2\}.$$ ‖

Example 4.2 (Bivariate exponential distribution). Let $\delta \geqslant 0$,

$$f_{(X|Y)}(x|y) = (\alpha + \delta y)\,\mathrm{e}^{-(\alpha+\delta y)x}, \quad x \in \mathbb{R}_+, \quad \alpha > 0, \quad \text{and}$$

$$f_{(Y|X)}(y|x) = (\beta + \delta x)\,\mathrm{e}^{-(\beta+\delta x)y}, \quad y \in \mathbb{R}_+, \quad \beta > 0.$$

Find the marginal distribution of X and the joint distribution of (X, Y).

Solution: Note that $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y = \mathbb{R}_+^2$. Setting $y_0 = 0$ in the sampling-wise IBF (4.6), we obtain $f_X(x) \propto (\beta+\delta x)^{-1} \exp(-\alpha x)$. This is a univariate distribution, from which it is easy to generate i.i.d. samples by using the rejection method (see §1.3) since

$$(\beta + \delta x)^{-1}\,\mathrm{e}^{-\alpha x} \leqslant (\alpha\beta)^{-1} \cdot \alpha\,\mathrm{e}^{-\alpha x}.$$

Then, (X, Y) follows a bivariate exponential distribution with

$$f_{(X,Y)}(x, y) \propto \exp\{-(\alpha x + \beta y + \delta x y)\}.$$

Arnold & Strauss (1988) give more detail on this joint density. ‖

9.5• Dimension reduction in high–dimensional integrals

— By utilizing the structure of the low–dimensional conditional densities, both the point-wise and function-wise IBF can effectively reduce the dimensionality in Monte Carlo integral.

— For example, given a joint density $f_{(X,\mathbf{y})}(x, \boldsymbol{y})$, where X is a r.v. and $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$ an n-dimensional random vector. Suppose we are interested in obtaining the marginal density

$$f_X(x) = \int \cdots \int f_{(X,\mathbf{y})}(x, \boldsymbol{y})\,\mathrm{d}y_1 \cdots \mathrm{d}y_n.$$

— There are many cases where the above n-fold integration is extremely difficult to perform, either analytically or numerically.

— In such cases the function-wise IBF (4.5) provides an alternative method to obtain $f_X(x)$.

— Note that both the conditional pdfs $f_{(X|\mathbf{y})}(x|\boldsymbol{y})$ and $f_{(\mathbf{y}|X)}(\boldsymbol{y}|x)$ are easy to obtain and are usually available, we only need to perform a one-dimension integration to calculate the normalizing constant with formula (4.5) in order to obtain $f_X(x)$.

— On the other hand, if Y is a r.v. while $\mathbf{x} = (X_1, \ldots, X_n)^\top$ is a random vector, then the point-wise IBF (4.3) can be used to obtain $f_X(x)$. Therefore, we can choose between the point-wise and function-wise IBF depending on the practical problem at hand.

10• Discrete versions of IBF

- When both X and Y are discrete r.v.'s, we have

$$\Pr(X = x) = \left\{ \sum_{y \in \mathcal{S}_Y} \frac{\Pr(Y = y|X = x)}{\Pr(X = x|Y = y)} \right\}^{-1},$$

  for any $x \in \mathcal{S}_X$, which is called the discrete version of the point-wise formula.

- The discrete version of the sampling-wise formula is

$$\Pr(X = x) \propto \frac{\Pr(X = x|Y = y_0)}{\Pr(Y = y_0|X = x)}, \tag{4.7}$$

  for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

Example 4.3 (Bivariate discrete distribution). Let X be a discrete r.v. with pmf $p_i = \Pr(X = x_i)$ for $i = 1, 2, 3$ and Y be a discrete r.v. with pmf $q_j = \Pr(Y = y_j)$ for $j = 1, 2, 3$. Given two conditional distribution matrices

$$\boldsymbol{A} = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1/6 & 0 & 3/14 \\ 0 & 1/4 & 4/14 \\ 5/6 & 3/4 & 7/14 \end{pmatrix}$$

and

$$\boldsymbol{B} = (b_{ij}) = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} 1/4 & 0 & 3/4 \\ 0 & 1/3 & 2/3 \\ 5/18 & 6/18 & 7/18 \end{pmatrix},$$

where the (i, j) element of $\boldsymbol{A}$ is $a_{ij} = \Pr\{X = x_i|Y = y_j\}$ and the (i, j) element of $\boldsymbol{B}$ is $b_{ij} = \Pr\{Y = y_j|X = x_i\}$.

1) Find the marginal distributions of X and Y.

2) Find the joint distribution of (X, Y).

Solution: 1) The support of X and Y are $\mathcal{S}_X = \{x_1, x_2, x_3\}$ and $\mathcal{S}_Y = \{y_1, y_2, y_3\}$. By using (4.7) with $y_0 = y_3$, the X-marginal is given by

$$\begin{aligned}
p_1 &\hat{=} \Pr(X = x_1) = f_X(x_1) \\
&\propto \frac{f_{(X|Y)}(x_1|y_0)}{f_{(Y|X)}(y_0|x_1)} = \frac{\Pr(X = x_1|Y = y_3)}{\Pr(Y = y_3|X = x_1)} \\
&= \frac{a_{13}}{b_{13}} = \frac{3/14}{3/4} = \frac{4}{14}, \\
p_2 &\hat{=} \Pr(X = x_2) = f_X(x_2) \\
&\propto \frac{f_{(X|Y)}(x_2|y_0)}{f_{(Y|X)}(y_0|x_2)} = \frac{\Pr(X = x_2|Y = y_3)}{\Pr(Y = y_3|X = x_2)} \\
&= \frac{a_{23}}{b_{23}} = \frac{4/14}{2/3} = \frac{6}{14}, \\
p_3 &\hat{=} \Pr(X = x_3) = f_X(x_3) \\
&\propto \frac{f_{(X|Y)}(x_3|y_0)}{f_{(Y|X)}(y_0|x_3)} = \frac{\Pr(X = x_3|Y = y_3)}{\Pr(Y = y_3|X = x_3)} \\
&= \frac{a_{33}}{b_{33}} = \frac{7/14}{7/18} = \frac{18}{14}.
\end{aligned}$$

Note that $p_1 + p_2 + p_3 = 1$, we obtain

$$\begin{aligned}
p_1 &= \frac{4/14}{4/14 + 6/14 + 18/14} = \frac{4}{4 + 6 + 18} = \frac{4}{28} = \frac{2}{14}, \\
p_2 &= \frac{6/14}{4/14 + 6/14 + 18/14} = \frac{6}{4 + 6 + 18} = \frac{6}{28} = \frac{3}{14}, \\
p_3 &= \frac{18/14}{4/14 + 6/14 + 18/14} = \frac{18}{4 + 6 + 18} = \frac{18}{28} = \frac{9}{14},
\end{aligned}$$

which are summarized into

X	x_1	x_2	x_3
$p_i = \Pr(X = x_i)$	2/14	3/14	9/14

Similarly, letting $x_0 = x_3$ in (4.7) yields the following Y-marginal

Y	y_1	y_2	y_3
$q_j = \Pr(Y = y_j)$	3/14	4/14	7/14

2) The joint distribution of (X, Y) is given by

$$\boldsymbol{P} = \begin{pmatrix} 1/28 & 0 & 3/28 \\ 0 & 2/28 & 4/28 \\ 5/28 & 6/28 & 7/28 \end{pmatrix}. \qquad \|$$

4.1.2 Monte Carlo versions of the IBF

11• Harmonic mean formula

- In some applications, the integrals in the right-hand side of (4.3) and (4.5) may be difficult to evaluate analytically.
- However, since the IBF already reduces the dimensionality by the conditional pdfs, we may use Monte Carlo methods to numerically evaluate the integral in (4.3) or to numerically calculate the normalizing constant in (4.5).
- For any given $x \in \mathcal{S}_X$, we have

$$\begin{aligned} f_X(x) &= \left\{ \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)} \, \mathrm{d}y \right\}^{-1} \\ &\approx \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{f_{(X|Y)}(x|y^{(i)})} \right\}^{-1} \hat{=} \hat{f}_{1,X}(x), \end{aligned} \tag{4.8}$$

where $\{y^{(i)}\}_{i=1}^n$ are realizations of $\{Y^{(i)}\}_{i=1}^n \overset{\text{iid}}{\sim} f_{(Y|X)}(y|x)$.

11.1• Definition of harmonic mean

— Given a sample $z_1, \ldots, z_n$, the mean is $\bar{z} = (1/n) \sum_{i=1}^n z_i$.

— The harmonic mean is

$$\tilde{z} = \left\{ \frac{1}{n} \sum_{i=1}^{n} z_i^{-1} \right\}^{-1} \quad \text{or} \quad \frac{1}{\tilde{z}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{z_i}.$$

12• Weighted point-wise IBF

- Similar to the importance function in importance sampling method in §3.3, we can derive a weighted version of the point-wise IBF (4.3).

- Let $w(y)$ be a density with the same support as $Y \sim f_Y(y)$.
- From the identity (4.2), we have

$$w(y) = \frac{f_{(Y|X)}(y|x)w(y)}{f_{(X|Y)}(x|y)f_Y(y)} \cdot f_X(x), \quad \forall\, x \in \mathcal{S}_X,\ \ y \in \mathcal{S}_Y. \tag{4.9}$$

- Integrating (4.9) with respect to y over the support $\mathcal{S}_Y$ gives

$$f_X(x) = \left\{ \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)w(y)}{f_{(X|Y)}(x|y)f_Y(y)}\, \mathrm{d}y \right\}^{-1}, \quad \begin{array}{l}\text{for any given} \\ x \in \mathcal{S}_X,\end{array} \tag{4.10}$$

  which is called the *weighted point-wise IBF*.

12.1• The first Monte Carlo version of (4.10)

— The first Monte Carlo version of (4.10) is given by

$$f_X(x) \approx \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{w(y^{(i)})}{f_{(X|Y)}(x|y^{(i)})f_Y(y^{(i)})} \right\}^{-1} \ \hat{=}\ \hat{f}_{2,X}(x), \tag{4.11}$$

where x is a given point in $\mathcal{S}_X$, $\{y^{(i)}\}_{i=1}^n$ are realizations of $\{Y^{(i)}\}_{i=1}^n \overset{\text{iid}}{\sim} f_{(Y|X)}(y|x)$, and $f_Y(\cdot)$ is calculated by the point-wise IBF (4.4).

— In particular, (4.10) and (4.11) are reduced to (4.3) and (4.8), respectively, if we set $w(y) = f_Y(y)$ for all $y \in \mathcal{S}_Y$.

12.2• The second Monte Carlo version of (4.10)

— If the generation from $f_{(Y|X)}(y|x)$ is very difficult but the evaluations of both $f_{(X|Y)}(x|y)$ and $f_{(Y|X)}(y|x)$ is relatively simple, then we may use the following Monte Carlo version of (4.10),

$$f_X(x) \approx \left\{ \frac{1}{m} \sum_{j=1}^{m} \frac{f_{(Y|X)}(y^{(j)}|x)}{f_{(X|Y)}(x|y^{(j)})f_Y(y^{(j)})} \right\}^{-1} \ \hat{=}\ \hat{f}_{3,X}(x), \tag{4.12}$$

where $x \in \mathcal{S}_X$ is fixed, $\{y^{(j)}\}_{j=1}^m$ are realizations of $\{Y^{(j)}\}_{j=1}^m \overset{\text{iid}}{\sim} w(y)$, and $f_Y(\cdot)$ is computed by the point-wise IBF (4.4).

13• Rao–blackwellized formula

- Let the joint samples $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ be available; i.e., $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ are realizations of $\{X^{(i)}, Y^{(i)}\}_{i=1}^n \overset{\text{iid}}{\sim} f_{(X,Y)}(x, y)$.

- Then, the Rao–Blackwellized estimator is defined by (Arnold 1993)

$$\begin{aligned} f_X(x) &= \int_{\mathcal{S}_Y} f_{(X,Y)}(x, y)\, \mathrm{d}y \\ &= \int_{\mathcal{S}_Y} f_{(X|Y)}(x|y) \cdot f_Y(y)\, \mathrm{d}y \\ &\approx \frac{1}{n} \sum_{i=1}^{n} f_{(X|Y)}(x|y^{(i)}) \doteq \hat{f}_{4,X}(x), \quad \forall\, x \in \mathcal{S}_X. \end{aligned} \tag{4.13}$$

13.1• Sampling from the joint density

— Starting with two conditional pdfs $f_{(X|Y)}(x|y)$ and $f_{(Y|X)}(y|x)$, we can obtain i.i.d. joint samples of (X, Y) through the following two steps:
$***$ First to draw x from $f_X(x)$ based on the sampling-wise IBF (4.6);
$***$ Second to draw y from the conditional $f_{(Y|X)}(y|x)$.

— Any procedures such as the grid method (§1.2), the rejection method (§1.3), and the adaptive rejection method (Gilks & Wild 1992) can be employed to the key first step.

— For instance, with the grid method, we select an appropriate grid of values $\{x^{(i)}\}_{i=1}^n$ that cover the support $\mathcal{S}_X$, and compute $f_X(x)$ by (4.3) at each point on the grid.

— Then, the density $f_X(x)$ can be approximated by the discrete distribution at $\{x^{(i)}\}_{i=1}^n$ with probabilities $p_i \doteq f_X(x^{(i)}) / \sum_{j=1}^n f_X(x^{(j)})$, $i = 1, \ldots, n$.

— The built-in R function sample(x, N, prob= p, replace = F) can be used to generate a sample from $X \sim \text{FDiscrete}_n(\{x^{(i)}\}, \{p_i\})$.

— Obviously, this method also works for an un-normalized density.

4.1.3 Generalization to the case of three random variables

14• Assumptions and goal

- We now extend the IBF from two random variables/vectors to the case of three random variables/vectors.
- Consider three random variables X_1, X_2 and X_3 and let $\mathcal{S}_{(X_1,X_2,X_3)} = \mathcal{S}_{X_1} \times \mathcal{S}_{X_2} \times \mathcal{S}_{X_3}$.
- Assume that three conditional pdfs $f_1(x_1|x_2,x_3)$, $f_2(x_2|x_1,x_3)$ and $f_3(x_3|x_1,x_2)$ are given and are positive. The goal is to find the joint density.

14.1• Formulae

— We only need to derive $f_{X_1}(x_1)$ and $f_{(X_2|X_1)}(x_2|x_1)$, since

$$f_{(X_1,X_2,X_3)}(x_1,x_2,x_3) = f_{X_1}(x_1) f_{(X_2|X_1)}(x_2|x_1) f_3(x_3|x_1,x_2).$$

— By (4.3), we have

$$f_{(X_2|X_1)}(x_2|x_1) = \left\{ \int \frac{f_3(x_3|x_1,x_2)}{f_2(x_2|x_1,x_3)} \, \mathrm{d}x_3 \right\}^{-1} \quad \text{and}$$

$$f_{(X_1|X_2)}(x_1|x_2) = \left\{ \int \frac{f_3(x_3|x_1,x_2)}{f_1(x_1|x_2,x_3)} \, \mathrm{d}x_3 \right\}^{-1}.$$

— Hence, by using (4.3) again, we obtain

$$f_{X_1}(x_1) = \left\{ \int \frac{f_{(X_2|X_1)}(x_2|x_1)}{f_{(X_1|X_2)}(x_1|x_2)} \, \mathrm{d}x_2 \right\}^{-1}.$$

4.2 The Bayesian Methodology

15• Likelihood principle

- In the likelihood-based (or frequentist/classical) statistics, the observed data $Y_{\text{obs}} = \{y_i\}_{i=1}^n$ are realizations of $\{Y_i\}_{i=1}^n$, which are r.v.'s having a joint distribution $f(y_1,\ldots,y_n;\boldsymbol{\theta})$; while the parameter vector $\boldsymbol{\theta}$ is *non-random* and *unknown*.

- The aim is to find the MLEs of $\boldsymbol{\theta}$ by maximizing the likelihood function $L(\boldsymbol{\theta}|Y_{\text{obs}}) = f(Y_{\text{obs}};\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where Y_{obs} are treated as fixed values.

16• Bayes principle

- In the Bayesian statistics, the $\boldsymbol{\theta}$ is treated as a *random vector* having a prior distribution $\pi(\boldsymbol{\theta})$.

- For a given $\boldsymbol{\theta}$, $Y_1, \ldots, Y_n$ have a conditional/sampling distribution $f(Y_{\text{obs}}|\boldsymbol{\theta}) = f(y_1, \ldots, y_n|\boldsymbol{\theta})$, which is equivalent to the observed-data likelihood in the frequentist statistics.

- The fundamentals of the Bayesian analysis is to combine the prior distribution with the observed-data likelihood to yield the posterior distribution:
$$p(\boldsymbol{\theta}|Y_{\text{obs}}) = \frac{f(Y_{\text{obs}}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{f(Y_{\text{obs}})},$$
which is called *Bayes formula*.

17• Purpose of this section

- First, we introduce some basic ideas in Bayesian approach.

- Second, we look at some fundamental Bayesian issues from the perspective of the IBF.

- In this section, we only consider the complete data problems. Let Y_{com} denote the completely observed data.

18• Bayesian approach with/without missing data

- Bayesian approach to data analysis consists of the following three main steps (Gelman *et al.* 1995, p.3)

  Step 1: Construct a full probability model summarized by a joint distribution for all observable and unobservable quantities (e.g., observed data Y_{obs}, missing data z or Y_{mis}, unknown parameters $\boldsymbol{\theta}$);

Step 2: Summarize the findings (e.g., mean, median, modes, posterior quantiles and intervals) for the unobserved quantities of interest based on the derived conditional distributions of these quantities given the observed data; i.e., $p(\boldsymbol{\theta}, z|Y_{\text{obs}})$;

Step 3: Assess the appropriateness of the model (i.e., model checking via goodness-of-fit test) and to suggest improvements (e.g., model selection such as variable selection).

4.2.1 The posterior distribution

19• Bayesian approach for complete data problems

- In order to carry out Step 1, we must provide the joint distribution of Y_{com} and $\boldsymbol{\theta}$:
$$f(Y_{\text{com}}, \boldsymbol{\theta}) = f(Y_{\text{com}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$
where $f(Y_{\text{com}}|\boldsymbol{\theta})$ is called the *sampling distribution*.

- Step 2 is completed by the Bayes formula to obtain the posterior distribution:
$$p(\boldsymbol{\theta}|Y_{\text{com}}) = \frac{f(Y_{\text{com}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(Y_{\text{com}})} \propto f(Y_{\text{com}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{4.14}$$
where
$$f(Y_{\text{com}}) = \int f(Y_{\text{com}}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = \int f(Y_{\text{com}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{4.15}$$
is the normalizing constant of $p(\boldsymbol{\theta}|Y_{\text{com}})$.

Example 4.4 (The multinomial model). Let N items be classified into n categories, $Y_{\text{com}} = \{\boldsymbol{y}\}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ denote the completely observed counts of the n cells, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ the cell probabilities. The multinomial sampling distribution (see (A.20) in Appendix A.1.7) is
$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \text{Multinomial}_n(\boldsymbol{y}|N, \boldsymbol{\theta}) = \binom{N}{y_1, \ldots, y_n} \prod_{i=1}^{n} \theta_i^{y_i},$$
where
$$\boldsymbol{y} \in \mathbb{T}_n(N) \triangleq \left\{ (y_1, \ldots, y_n)^\top \colon\ y_i \geqslant 0,\ \sum_{i=1}^{n} y_i = N \right\}.$$

If the prior distribution of $\boldsymbol{\theta}$ is taken as the *conjugate* Dirichlet distribution (see (A.24) in Appendix A.2.1)

$$\pi(\boldsymbol{\theta}) = \text{Dirichlet}_n(\boldsymbol{\theta}|\boldsymbol{a}) = \frac{\prod_{i=1}^n \theta_i^{a_i-1}}{B_n(\boldsymbol{a})}, \quad \boldsymbol{\theta} \in \mathbb{T}_n \equiv \mathbb{T}_n(1),$$

where $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$ and

$$B_n(\boldsymbol{a}) = \frac{\Gamma(a_1)\cdots\Gamma(a_n)}{\Gamma(a_1+\cdots+a_n)}$$

is the multivariate beta function. Find the posterior distribution of $p(\boldsymbol{\theta}|\boldsymbol{y})$ and its normalizing constant $f(\boldsymbol{y})$.

Solution: The posterior of $\boldsymbol{\theta}$ and its normalizing constant are given by

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \text{Dirichlet}_n(\boldsymbol{\theta}|\boldsymbol{y}+\boldsymbol{a}), \qquad \boldsymbol{\theta} \in \mathbb{T}_n, \quad \text{and} \\ f(\boldsymbol{y}) &= \binom{N}{y_1, \ldots, y_n} \frac{B_n(\boldsymbol{y}+\boldsymbol{a})}{B_n(\boldsymbol{a})}, \quad \boldsymbol{y} \in \mathbb{T}_n(N), \end{aligned} \tag{4.16}$$

respectively. The density (4.16) is called the *Dirichlet–multinomial* density (see (A.21) in Appendix A.1.7):

$$\mathbf{y} = (Y_1, \ldots, Y_n)^\top \sim \text{DMultinomial}_n(N, \boldsymbol{a}), \qquad \boldsymbol{y} \in \mathbb{T}_n(N).$$

From (4.15), it is easy to see that (4.16) is a *mixture* of multinomial distribution with rates, $\boldsymbol{\theta}$, and it follows a Dirichlet distribution. Therefore, the Dirichlet–multinomial distribution is a *robust* alternative to the multinomial distribution.

Conjugate prior: A prior density $\pi(\boldsymbol{\theta}) \in \mathcal{F}(\cdot)$ is called a conjugate prior if the resultant posterior distribution $p(\boldsymbol{\theta}|Y_{\text{com}})$ still belongs to the same distribution family $\mathcal{F}(\cdot)$.

Mixture representation: Recall §1.5.4, we know that

$$f_X(x) = \int_{\mathbb{Y}} f_{(X|Y)}(x|y) \cdot f_Y(y)\, \mathrm{d}y$$

is called the mixture of $f_{(X|Y)}(x|y)$.

Robustness and heavy–tail distribution: The variance of X is always larger than the variance of $X|(Y=y)$; i.e., $\text{Var}(X) \geqslant \text{Var}(X|Y=y)$. We say the distribution of X is more robust than the distribution of $X|(Y=y)$, or the tail of the density $f_X(x)$ is heavier than that of the conditional density $f_{(X|Y)}(x|y)$. ‖

4.2.2 Nuisance parameters

20• The issue

- A major difficulty in the classical likelihood approach is with nuisance parameters.
- Let $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_{-1}^\top)^\top$, where θ_1 is the parameter of interest and $\boldsymbol{\theta}_{-1} \hat{=} (\theta_2, \ldots, \theta_d)^\top$. Then, $\boldsymbol{\theta}_{-1}$ becomes a *nuisance parameter* vector.

20.1• Profile likelihood method

— A common approach to deal with the nuisance parameter is by the *profile likelihood method*, in which $\boldsymbol{\theta}_{-1}$ is treated as known and is fixed usually at its MLE.

— The profile likelihood is defined as $L_p(\theta_1, \hat{\boldsymbol{\theta}}_{-1}|Y_{\text{com}})$, where $\hat{\boldsymbol{\theta}}_{-1}$ is the MLEs of $\boldsymbol{\theta}_{-1}$ and $L(\theta_1, \boldsymbol{\theta}_{-1}|Y_{\text{com}})$ denotes the original likelihood.

— However, this method *underestimates* the uncertainty in the estimation of $\boldsymbol{\theta}_{-1}$, often leading to biased estimate of θ_1 and incorrect conclusion, especially, when the dimension of $\boldsymbol{\theta}_{-1}$ is high (Liu 2001, p.304).

20.2• Traditional Bayesian method

— In Bayesian framework, $\boldsymbol{\theta}_{-1}$ can be removed by integration:

$$\begin{aligned} p(\theta_1|Y_{\text{com}}) &= \int p(\theta_1, \boldsymbol{\theta}_{-1}|Y_{\text{com}})\, \mathrm{d}\boldsymbol{\theta}_{-1} = \int p(\boldsymbol{\theta}|Y_{\text{com}})\, \mathrm{d}\boldsymbol{\theta}_{-1} \\ &\overset{(4.14)}{=} \int \frac{f(Y_{\text{com}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(Y_{\text{com}})}\, \mathrm{d}\boldsymbol{\theta}_{-1} \\ &\overset{(4.15)}{=} \frac{\int f(Y_{\text{com}}|\theta_1, \boldsymbol{\theta}_{-1})\pi(\theta_1, \boldsymbol{\theta}_{-1})\, \mathrm{d}\boldsymbol{\theta}_{-1}}{\int\int f(Y_{\text{com}}|\theta_1, \boldsymbol{\theta}_{-1})\pi(\theta_1, \boldsymbol{\theta}_{-1})\, \mathrm{d}\boldsymbol{\theta}_{-1}\, \mathrm{d}\theta_1}. \end{aligned}$$

— This is a ratio of two high–dimensional integrals.

21• Looking at nuisance parameters from the viewpoint of IBF

- If we could obtain both $p(\theta_1|Y_{\text{com}}, \boldsymbol{\theta}_{-1})$ and $p(\boldsymbol{\theta}_{-1}|Y_{\text{com}}, \theta_1)$ in closed form, according to the sampling-wise IBF (4.6), we would have

$$p(\theta_1|Y_{\text{com}}) \propto \frac{p(\theta_1|Y_{\text{com}}, \boldsymbol{\theta}_{-1}^*)}{p(\boldsymbol{\theta}_{-1}^*|Y_{\text{com}}, \theta_1)}, \tag{4.17}$$

  where $\boldsymbol{\theta}_{-1}^*$ is some arbitrary value in $\mathcal{S}_{\boldsymbol{\theta}_{-1}}$.

- Therefore, the IBF provides a way to obtain the marginal distribution of interest and it does not require integration.
- Otherwise, sampling-based methods such as the Gibbs sampler can be utilized to evaluate such integrals.

Example 4.5 (The normal model). Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and $Y_{\text{com}} = \{y_i\}_{i=1}^n$ be realizations of $\{Y_i\}_{i=1}^n$. The normal sampling distribution is

$$f(Y_{\text{com}}|\mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{ - \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \right\}.$$

Consider a *non-informative prior* distribution $\pi(\mu, \sigma^2) \propto 1/\sigma^2$, please find the marginal posterior distribution of $\mu|Y_{\text{com}}$.

Solution: The joint posterior density is

$$p(\mu, \sigma^2|Y_{\text{com}}) \propto \sigma^{-n-2} \exp\left\{ - \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}, \tag{4.18}$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

are joint sufficient statistics for (μ, σ^2). We are often interested in μ, so σ^2 may be considered as a nuisance parameter. From (4.18), we readily have (see **26•** in Appendix A.2.3)

$$\begin{aligned} \sigma^2|(Y_{\text{com}}, \mu) &\sim \text{IGamma}\left(\frac{n}{2}, \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2} \right), \\ \mu|(Y_{\text{com}}, \sigma^2) &\sim N(\bar{y}, \sigma^2/n). \end{aligned}$$

Using (4.17) and letting the arbitrary value of nuisance parameter $\sigma^2 = 1$, we have (see **34•** in Appendix A.2.7) the posterior of interest:

$$\mu|Y_{\text{com}} \sim t(\bar{y}, s^2/n, n-1).$$

Non-informative prior of μ: A non-informative distribution itself may not a proper density. For example, let $X \sim N(\mu, \sigma_0^2)$, then

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma_0^2} \right\}, \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}.$$

If we treat x as a fixed value and μ as a r.v. , from above formula, we have $\mu \sim N(x, \sigma_0^2)$. When $\sigma_0 \to \infty$, we obtain $\pi(\mu) \propto 1$, which is not a proper density.

Inverse gamma and non-informative prior of σ^2: The pdf of IGamma(α, β) is

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \, \mathrm{e}^{-\beta/x} \propto x^{-(\alpha+1)} \, \mathrm{e}^{-\beta/x}, \quad x \geqslant 0,$$

where $\alpha > 0$ and $\beta > 0$. As $\alpha \to 0$ and $\beta \to 0$, we have $f(x|\alpha, \beta) \to x^{-1}$. The conjugate prior for the normal variance σ^2 is IGamma(α, β), so the non-informative prior for σ^2 is proportional to $1/\sigma^2$.

Non-informative joint prior of (μ, σ^2): If we assume that $\mu \perp\!\!\!\perp \sigma^2$, then the non-informative joint prior distribution of (μ, σ^2) is $\pi(\mu, \sigma^2) \propto 1/\sigma^2$. ‖

4.2.3 Posterior predictive distribution

22• Why do we need the posterior predictive distribution?

- Model checking is crucial to Bayesian data analysis and is a major aspect of implementing Step 3 presented in the beginning of §4.2.
- One popular technique for model checking is to draw samples from the *posterior predictive distribution* of replicated data and compare these samples to the observed data.
- Any systematic discrepancies between the simulations and the data imply potential lack of fit of the model.

22.1• Prior predictive distribution

— Let $\tilde{y}$ denote a future observation or an unknown observable quantity.

— Predictive inference is to make inference on $\tilde{y}$ or its function.

— Before Y_{com} are observed, the distribution of the unknown but observable Y_{com} is $f(Y_{\text{com}})$ specified by (4.15).

— It is sometimes called the *prior predictive distribution*.

22.2• Posterior predictive distribution

— After Y_{com} were observed, we can predict or forecast $\tilde{y}$.

— The posterior predictive distribution of $\tilde{y}$ given the data Y_{com} is defined as (see, e.g., Aitchison & Dunsmore 1975, p.24)

$$\begin{aligned} f(\tilde{y}|Y_{\text{com}}) &= \int f(\tilde{y}, \boldsymbol{\theta}|Y_{\text{com}})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int f(\tilde{y}|Y_{\text{com}}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|Y_{\text{com}})\, \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{4.19}$$

22.3• An assumption of conditional independence

— Most frequently, the future observation $\tilde{y}$ and Y_{com} are conditionally independent given $\boldsymbol{\theta}$.

— In this case, we have

$$f(\tilde{y}|Y_{\text{com}}, \boldsymbol{\theta}) = f(\tilde{y}|\boldsymbol{\theta}). \tag{4.20}$$

22.4• Alternative formula of (4.19) via the function-wise IBF

— On the other hand, the function-wise IBF (4.5) can be used to derive the following alternative formula:

$$f(\tilde{y}|Y_{\text{com}}) = f(\tilde{y}|Y_{\text{com}}, \boldsymbol{\theta}_0) \times \frac{p(\boldsymbol{\theta}_0|Y_{\text{com}})}{p(\boldsymbol{\theta}_0|Y_{\text{com}}, \tilde{y})}, \tag{4.21}$$

where $\boldsymbol{\theta}_0$ is an arbitrary value of $\boldsymbol{\theta}$ in its support $\mathcal{S}_{\boldsymbol{\theta}}$ and $p(\boldsymbol{\theta}|Y_{\text{com}}, \tilde{y})$ is the posterior density of $\boldsymbol{\theta}$, with Y_{com} augmented by an additional observation $\tilde{y}$.

— Note that (4.21) gives $f(\tilde{y}|Y_{\text{com}})$ directly without integration.

— The formula (4.21) appeared without explanation in a Durham University undergraduate final examination script of 1984 and was reported by Besag (1989).

— However, the writings of Besag (1989) indicates that they seem to be unaware that (4.21) could come from an exact formula; i.e., the function-wise IBF.

Example 4.6 (The Bernoulli model). Let $\{Y_i\}_{i=1}^n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. The Bernoulli sampling distribution of $\{Y_i\}_{i=1}^n$ is

$$f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i} = \theta^y(1-\theta)^{n-y},$$

where $y \hat{=} \sum_{i=1}^n y_i$ is a sufficient statistic of θ. Let the prior of θ follow $\text{Beta}(a, b)$, please find the posterior predictive distribution of Y_{n+1}.

Solution: Note that $Y \hat{=} \sum_{i=1}^n Y_i \sim \text{Binomial}(n, \theta)$, so the complete-data $\{y_i\}_{i=1}^n$ can be written as $Y_{\text{com}} = \{y\}$. The posterior of θ is $p(\theta|Y_{\text{com}}) = \text{Beta}(\theta|y + a, n - y + b)$ for $\theta \in (0, 1)$. From (4.15), the prior predictive distribution is a beta–binomial distribution (see (A.4) in Appendix A.1.3):

$$f(Y_{\text{com}}) = \text{BBinomial}(y|n, a, b) = \binom{n}{y} \frac{B(y + a, n - y + b)}{B(a, b)}.$$

Since $Y_{n+1} = \tilde{y}$ is the result of a new trial, from (4.20), we obtain

$$\begin{aligned} f(\tilde{y}|Y_{\text{com}}, \theta) &= f(\tilde{y}|\theta) = \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}} \quad \text{and} \\ p(\theta|Y_{\text{com}}, \tilde{y}) &= \text{Beta}(\theta|y + \tilde{y} + a, n + 1 - y - \tilde{y} + b). \end{aligned}$$

In (4.21) letting $\theta_0 = 0.5$, then the posterior predictive density is

$$f(\tilde{y}|Y_{\text{com}}) = \text{BBinomial}(\tilde{y}|1, y + a, n - y + b),$$

which is a special beta–binomial distribution; i.e., a Bernoulli distribution with pmf

$$\Pr(Y_{n+1} = 1|Y_{\text{com}}) = \frac{y + a}{n + a + b}, \quad \Pr(Y_{n+1} = 0|Y_{\text{com}}) = \frac{n - y + b}{n + a + b}. \quad \|$$

22.5• The case of conditional dependency

— The next example shows that sometimes the assumption of conditional independence (4.20) is not valid.

Example 4.7 (Homogeneous Poisson process). Let $\{Y_i, i \geqslant 1\} \sim \text{HPP}(\lambda)$ (see Appendix A.3.1) and $Y_{\text{com}} = \{y_i\}_{i=1}^n$. According to **38.1•** Property (3) in Appendix A.3.1, the joint pdf of the successive event times $Y_1, \ldots, Y_n$ is

$$f(y_1, \ldots, y_n|\lambda) = \lambda^n \, \mathrm{e}^{-\lambda y_n}, \quad 0 < y_1 < \cdots < y_n.$$

Let the prior of the rate λ be Gamma(a, b), please find the posterior predictive distribution of Y_{n+1}.

Solution: The posterior of λ is $p(\lambda|Y_{\text{com}}) = \text{Gamma}(\lambda|n + a, y_n + b)$. From (4.15), the prior predictive distribution is given by

$$f(Y_{\text{com}}) = \frac{b^a}{\Gamma(a)} \cdot \frac{\Gamma(n + a)}{(y_n + b)^{n+a}}, \quad 0 < y_1 < \cdots < y_n.$$

Let $Y_{n+1} = y_{n+1} = \tilde{y}$, we obtain

$$\begin{aligned} f(y_{n+1}|Y_{\text{com}}, \lambda) &= \frac{f(Y_{\text{com}}, y_{n+1}|\lambda)}{f(Y_{\text{com}}|\lambda)} = \frac{f(y_1, \ldots, y_n, y_{n+1}|\lambda)}{f(y_1, \ldots, y_n|\lambda)} \\ &= \frac{\lambda^{n+1} \, \mathrm{e}^{-\lambda y_{n+1}}}{\lambda^n \, \mathrm{e}^{-\lambda y_n}} \\ &= \lambda \, \mathrm{e}^{-\lambda(y_{n+1} - y_n)}, \quad y_{n+1} > y_n, \\ p(\lambda|Y_{\text{com}}, y_{n+1}) &= \text{Gamma}(\lambda|n + 1 + a, y_{n+1} + b). \end{aligned}$$

In (4.21) letting $\lambda_0 = 1$, then the posterior predictive density is

$$f(y_{n+1}|Y_{\text{com}}) = f(y_{n+1}|y_n) = \frac{(n + a)(y_n + b)^{n+a}}{(y_{n+1} + b)^{n+a+1}}, \quad y_{n+1} > y_n.$$

Stationary increments: A continuous–time stochastic process $\{X(t)\colon t \in \mathbb{T}\}$ is said to have stationary increments, if the r.v. $X(t + s) - X(t)$ has the same distribution for all $t \in \mathbb{T}$.

Independent increments: It is said to possess independent increments for all $t_0 < t_1 < t_2 < \cdots < t_n$, if the r.v.'s

$$X(t_1) - X(t_0), \; X(t_2) - X(t_1), \; \ldots, \; X(t_n) - X(t_{n-1})$$

are independent. ‖

4.2.4 Bayes factor

23• Marginal likelihood

- The normalizing constant $f(Y_{\text{com}})$ defined by (4.15) is sometimes called the *marginal likelihood* of the data and is often denoted by $m(Y_{\text{com}})$ in the statistical literature.

- The marginal likelihood is closely related to *Bayes factor*, which can be used for model selection (Kass & Raftery 1995).

23.1• Usefulness of the Bayes factor

— The Bayes factor is defined as the ratio of posterior odds versus prior odds, which is simply a ratio of two marginal likelihoods.

— To compare two models, say M_1 and M_2, the Bayes factor for model M_1 versus model M_2 is

$$B_{12} = \frac{m(Y_{\text{com}}|M_1)}{m(Y_{\text{com}}|M_2)}.$$

— Jeffreys (1961) suggested interpreting B_{12} in half–units on the $\log_{10}$ scale; i.e., when B_{12} falls in intervals $(1, 3.2)$, $(3.2, 10)$, $(10, 100)$ and $(100, \infty)$, the evidence against M_2 is considered not worth more than a bare mention, substantial, strong and decisive, respectively.

4.2.5 Estimation of marginal likelihood

24• Four estimators of marginal likelihood

- Estimating the marginal likelihood $m(Y_{\text{com}})$ is crucial to the calculation of Bayes factor.

24.1• First estimator based on harmonic mean formula

— If we replace X and Y in (4.8) by Y_{com} and $\boldsymbol{\theta}$, respectively, then from (4.8) we immediately obtain the first estimator of $m(Y_{\text{com}})$:

$$\hat{m}_1(Y_{\text{com}}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{f(Y_{\text{com}}|\boldsymbol{\theta}^{(i)})} \right\}^{-1}, \tag{4.22}$$

where Y_{com} is a given point in its support, and $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{n} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta}|Y_{\text{com}})$.

24.2• Second estimator based on (4.11)

— Similarly, from (4.11), the second estimator of $m(Y_{\text{com}})$ is

$$\hat{m}_2(Y_{\text{com}}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{w(\boldsymbol{\theta}^{(i)})}{f(Y_{\text{com}}|\boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})} \right\}^{-1}, \tag{4.23}$$

where Y_{com} is a given point in its support, $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{n} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta}|Y_{\text{com}})$.

— In particular, the formula (4.23) will be reduced to (4.22) if we let $w(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

— The formula (4.23) was mentioned by Gelfand & Dey (1994) in the context of importance sampling. Kass & Raftery (1995) showed that $\hat{m}_2(Y_{\text{com}})$ is an unbiased and consistent estimator of $m(Y_{\text{com}})$, and satisfies a Gaussian central limit theorem if the tails of $w(\cdot)$ are thin enough.

— Therefore, $\hat{m}_2(Y_{\text{com}})$ does not have the instability of $\hat{m}_1(Y_{\text{com}})$.

24.3• Third estimator based on (4.12)

— Alternatively, if generating samples from $p(\boldsymbol{\theta}|Y_{\text{com}})$ is very difficult but evaluating both $f(Y_{\text{com}}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|Y_{\text{com}})$ is relatively simple, from (4.12), then we can obtain the third estimator of $m(Y_{\text{com}})$:

$$\hat{m}_3(Y_{\text{com}}) = \left\{ \frac{1}{m} \sum_{j=1}^{m} \frac{p(\boldsymbol{\theta}^{(j)}|Y_{\text{com}})}{f(Y_{\text{com}}|\boldsymbol{\theta}^{(j)})\pi(\boldsymbol{\theta}^{(j)})} \right\}^{-1},$$

where Y_{com} is fixed, $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{m} \overset{\text{iid}}{\sim} w(\boldsymbol{\theta})$.

24.4• Fourth estimator based on Rao–Blackwellized formula

— Finally, let $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{n} \overset{\text{iid}}{\sim} \pi(\boldsymbol{\theta})$, from (4.13), then the fourth estimator of $m(Y_{\text{com}})$ is the Rao–Blackwellized estimator given by

$$\hat{m}_4(Y_{\text{com}}) = \frac{1}{n} \sum_{i=1}^{n} f(Y_{\text{com}}|\boldsymbol{\theta}^{(i)}).$$

— As mentioned by Gelfand *et al.* (1992), $\hat{m}_4(Y_{\text{com}})$ is better than the kernel density estimator under a wide range of loss functions.

— The disadvantages of the Rao–Blackwellized estimator are that (i) the closed form of $f(Y_{\text{com}}|\boldsymbol{\theta})$ must be known, and (ii) $\hat{m}_4(Y_{\text{com}})$ is a mixture density, which is relatively difficult to treat when n is large enough.

4.3 The Data Augmentation (DA) Algorithm

25• Missing data problems

- In the previous section, we assumed that the desired data are completely observed.
- However, incomplete observations arise in many applications. For example,
  - a survey with multiple questions may include non-responses to some personal questions.
  - In an industrial experiment some results are missing because of mechanical breakdowns unrelated to the experimental process.
  - A pharmaceutical experiment on the after–effects of a toxic product may skip some doses for a given patient.
- In addition, many other problems can be treated as a missing data problem (e.g, latent–class model, mixture model, some constrained parameter models, etc).
- The missing data formulation is an important tool for modeling to address a specific scientific question.

25.1• Some notations

— In this section, let $Y_{\text{obs}} = (y_1, \ldots, y_r)^\top$ denote the observed values, $Y_{\text{mis}} = (y_{r+1}, \ldots, y_n)^\top$ the missing values, $\boldsymbol{\theta}$ the parameter vector of interest, and write

$$Y_{\text{com}} = \begin{pmatrix} Y_{\text{obs}} \\ Y_{\text{mis}} \end{pmatrix} = \boldsymbol{y}.$$

4.3.1 Missing data mechanism

26• Two important concepts in missing data problems

- MAR: Missing at random.
- MCAR: Missing completely at random.

26.1• Missing-data indicator

— We define a *missing-data indicator* vector $\boldsymbol{m}$, whose elements take value 1 if the corresponding components of $\boldsymbol{y}$ are observed and 0 if the corresponding components of $\boldsymbol{y}$ are missing; that is

$$\boldsymbol{m} = \begin{pmatrix} \mathbf{1}_r \\ \mathbf{0}_{n-r} \end{pmatrix}.$$

26.2• Missing-data mechanism

— The joint distribution of $(\mathbf{y}, \mathbf{m})$ is $f(\boldsymbol{y}, \boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}|\boldsymbol{\theta}) \times f(\boldsymbol{m}|\boldsymbol{y}, \boldsymbol{\psi})$.

— The missing-data mechanism is characterized by the conditional distribution of $\mathbf{m}$ given $\mathbf{y}$, which is indexed by unknown parameters $\boldsymbol{\psi}$.

— The joint distribution of observed data and $\mathbf{m}$ is obtained by integrating over the distribution of Y_{mis}:

$$f(Y_{\text{obs}}, \boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\boldsymbol{\theta}) \cdot f(\boldsymbol{m}|Y_{\text{obs}}, Y_{\text{mis}}, \boldsymbol{\psi}) \, \mathrm{d}Y_{\text{mis}}.$$

26.3• Missing at random

— Missing data are said to be MAR if

$$f(\boldsymbol{m}|Y_{\text{obs}}, Y_{\text{mis}}, \boldsymbol{\psi}) = f(\boldsymbol{m}|Y_{\text{obs}}, \boldsymbol{\psi}).$$

— That is, the distribution of the missing-data mechanism does not depend on the missing values, but depend on the observed values (including fully observed covariates) and the parameter vector $\boldsymbol{\psi}$.

— Under the assumption of MAR, we have

$$f(Y_{\text{obs}}, \boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{m}|Y_{\text{obs}}, \boldsymbol{\psi}) \times f(Y_{\text{obs}}|\boldsymbol{\theta}).$$

26.4• Ignorable missing-data mechanism

— If the joint prior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\psi}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{\psi})$, then Bayesian inferences on $\boldsymbol{\theta}$ can be obtained by considering only the observed-data likelihood $f(Y_{\text{obs}}|\boldsymbol{\theta})$.

— In this case, the missing-data mechanism is said to be *ignorable*.

26.5• Missing completely at random

— Missing data are said to be MCAR if the distribution of the missing-data mechanism is completely independent of Y_{com}:

$$f(\boldsymbol{m}|Y_{\text{obs}}, Y_{\text{mis}}, \boldsymbol{\psi}) = f(\boldsymbol{m}|\boldsymbol{\psi}).$$

4.3.2 The idea of data augmentation

27• DA is a common idea for both EM and DA algorithms

- Assume that we want to make statistical inference on the parameter vector $\boldsymbol{\theta}$ based on the observed posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

- The EM algorithm (Dempster *et al.* 1977; see §2.3 for more detail) and the DA algorithm are designed for obtaining the mode of $p(\boldsymbol{\theta}|Y_{\text{obs}})$ and for simulating from $p(\boldsymbol{\theta}|Y_{\text{obs}})$, respectively.

- The EM and the DA share a simple idea that rather than performing a complicated optimization or simulation, the observed data is augmented with latent data so that a series of simple optimizations or simulations can be performed.

27.1• DA algorithm is a stochastic version of the EM algorithm

— As a stochastic version of the deterministic EM algorithm, the DA algorithm was originally proposed by Tanner & Wong (1987).

— The basic idea is to introduce latent data Z so that the complete-data posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and the conditional predictive distribution $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ are available, where by "available" it means either the samples can be easily generated from both, or each can be easily evaluated at any given point.

27.2• Purpose of the DA algorithm

— Given both $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$, the aim is to find the observed-data posterior $p(\boldsymbol{\theta}|Y_{\text{obs}})$ or to generate posterior samples from $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

— If $f(z|Y_{\text{obs}})$ was known, then we have

$$\begin{aligned} p(\boldsymbol{\theta}|Y_{\text{obs}}) &= \int p(\boldsymbol{\theta}, z|Y_{\text{obs}})\,\mathrm{d}z \\ &= \int_{\mathcal{S}_{(Z|Y_{\text{obs}})}} p(\boldsymbol{\theta}|Y_{\text{obs}}, z)\cdot f(z|Y_{\text{obs}})\,\mathrm{d}z \qquad (4.24) \\ &\approx \frac{1}{m}\sum_{j=1}^{m} p(\boldsymbol{\theta}|Y_{\text{obs}}, z^{(j)}), \end{aligned}$$

where $\mathcal{S}_{(Z|Y_{\text{obs}})}$ is the support of $Z|Y_{\text{obs}}$, and $\{z^{(j)}\}_{j=1}^{m} \overset{\text{iid}}{\sim} f(z|Y_{\text{obs}})$.

— The key is how to generate samples from the unknown $f(z|Y_{\text{obs}})$.

4.3.3 The original DA algorithm

28• The fixed point iteration

- The DA algorithm is motivated by (4.24) and the integral identity:

$$f(z|Y_{\text{obs}}) = \int_{\mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}} f(z|Y_{\text{obs}}, \boldsymbol{\phi})\cdot p(\boldsymbol{\phi}|Y_{\text{obs}})\,\mathrm{d}\boldsymbol{\phi},$$

  where $\mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$ denotes the support of $\boldsymbol{\theta}|Y_{\text{obs}}$.

- By substitution, we have $p(\boldsymbol{\theta}|Y_{\text{obs}}) = \int K(\boldsymbol{\theta}, \boldsymbol{\phi})\cdot p(\boldsymbol{\phi}|Y_{\text{obs}})\,\mathrm{d}\boldsymbol{\phi}$, where the kernel function is

$$K(\boldsymbol{\theta}, \boldsymbol{\phi}) = \int p(\boldsymbol{\theta}|Y_{\text{obs}}, z)\cdot f(z|Y_{\text{obs}}, \boldsymbol{\phi})\,\mathrm{d}z. \qquad (4.25)$$

- The fixed point iteration in functional analysis is thus

$$p_{k+1}(\boldsymbol{\theta}|Y_{\text{obs}}) = \int K(\boldsymbol{\theta}, \boldsymbol{\phi})\cdot p_k(\boldsymbol{\phi}|Y_{\text{obs}})\,\mathrm{d}\boldsymbol{\phi}, \qquad k \in \mathbb{N}. \qquad (4.26)$$

29• Sufficient conditions for convergence

- For the convergence of $p_{k+1}(\boldsymbol{\theta}|Y_{\text{obs}})$ to $p(\boldsymbol{\theta}|Y_{\text{obs}})$ in ℓ_1-norm; i.e.

$$\int |p_{k+1}(\boldsymbol{\theta}|Y_{\text{obs}}) - p(\boldsymbol{\theta}|Y_{\text{obs}})|\,\mathrm{d}\boldsymbol{\theta} \to 0 \quad \text{as} \quad k \to \infty,$$

  Tanner & Wong (1987) established the following sufficient conditions:

(1) $K(\boldsymbol{\theta}, \boldsymbol{\phi})$ is uniformly bounded;

(2) $K(\boldsymbol{\theta}, \boldsymbol{\phi})$ is equicontinuous in $\boldsymbol{\theta}$;

(3) For any $\boldsymbol{\theta}_0 \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$, there is an open neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$ such that $K(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ in $\mathcal{N}$;

(4) The starting function $p_0(\boldsymbol{\theta}|Y_{\text{obs}})$ is such that $\sup_{\boldsymbol{\theta}} \frac{p_0(\boldsymbol{\theta}|Y_{\text{obs}})}{p(\boldsymbol{\theta}|Y_{\text{obs}})} < \infty$.

- Note that, in addition to the integral (4.25) for obtaining the kernel function, an integral (4.26) is to be evaluated for each update from $p_k(\boldsymbol{\theta}|Y_{\text{obs}})$ to $p_{k+1}(\boldsymbol{\theta}|Y_{\text{obs}})$.

- Tanner & Wong (1987) adopt the method of Monte Carlo to perform the integration in (4.25).

30• The original DA algorithm

- The DA algorithm consists of iterates between the *imputation step* (I-step) and the *posterior step* (P-step), which is summarized as follows:

  I-step: Draw $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^m$ from the current $p_k(\boldsymbol{\theta}|Y_{\text{obs}})$; For each $\boldsymbol{\theta}^{(j)}$, draw $z^{(j)}$ from $f(z|Y_{\text{obs}}, \boldsymbol{\theta}^{(j)})$;

  P-step: Update the posterior as

$$p_{k+1}(\boldsymbol{\theta}|Y_{\text{obs}}) = \frac{1}{m}\sum_{j=1}^{m} p(\boldsymbol{\theta}|Y_{\text{obs}}, z^{(j)}). \tag{4.27}$$

30.1• Multiple imputation

— The $\{z^{(j)}\}_{j=1}^m$ so produced are often called *multiple imputation* (Rubin 1987a).

30.2• Remarks

— The convergence rate and other aspects of the augmentation scheme (4.27) were further considered by Schervish & Carlin (1992) and Liu *et al.* (1994, 1995).

— It was stated without substantiation by Gelfand & Smith (1990) that the fixed point iteration can be extended to conditionals for more than two components.

— Since Conditions (1)–(4) and the proof involve considerable mathematics, the extension is not that obvious.

31• The modern DA algorithm (time: 1990)

- Particularly, in the original DA algorithm letting $m = 1$, we obtain

  I-step: Draw $z^{(t)} \sim f(z|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$;

  P-step: Draw $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|Y_{\text{obs}}, z^{(t)})$.

- As $t \to \infty$, $\{\boldsymbol{\theta}^{(t+1)}\}_{t=1}^{\infty}$ has a stationary distribution $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

4.3.4 Connection with the IBF

32• The connection with the point-wise IBF (time: 1996)

- In the DA scheme, we assumed that both $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ are available.
- Typically, in practice, we have $\mathcal{S}_{(\boldsymbol{\theta}, Z|Y_{\text{obs}})} = \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})} \times \mathcal{S}_{(Z|Y_{\text{obs}})}$.
- Corresponding to the point-wise IBF (4.3), for any given $\boldsymbol{\theta} \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$, we obtain

$$p(\boldsymbol{\theta}|Y_{\text{obs}}) = \left\{ \int_{\mathcal{S}_{(Z|Y_{\text{obs}})}} \frac{f(z|Y_{\text{obs}}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}|Y_{\text{obs}}, z)} \, \mathrm{d}z \right\}^{-1}.$$

32.1• Other two formulae

— Corresponding to the sampling-wise IBF (4.6) and the function-wise IBF (4.5), for some arbitrary $z_0 \in \mathcal{S}_{(Z|Y_{\text{obs}})}$ and all $\boldsymbol{\theta} \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$, we have

$$p(\boldsymbol{\theta}|Y_{\text{obs}}) \propto \frac{p(\boldsymbol{\theta}|Y_{\text{obs}}, z_0)}{f(z_0|Y_{\text{obs}}, \boldsymbol{\theta})} \quad \text{and} \tag{4.28}$$

$$p(\boldsymbol{\theta}|Y_{\text{obs}}) = c(z_0) \times \frac{p(\boldsymbol{\theta}|Y_{\text{obs}}, z_0)}{f(z_0|Y_{\text{obs}}, \boldsymbol{\theta})},$$

where the normalizing constant of $p(\boldsymbol{\theta}|Y_{\text{obs}})$ is given by

$$c(z_0) = f(z_0|Y_{\text{obs}}) = \left\{ \int_{\mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}} \frac{p(\boldsymbol{\theta}|Y_{\text{obs}}, z_0)}{f(z_0|Y_{\text{obs}}, \boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta} \right\}^{-1}.$$

— Therefore, the IBF gives explicit solutions to the fixed point iteration of the original DA algorithm.

Example 4.8 (Incomplete categorical data analysis). Let the observed frequencies be denoted by $Y_{\text{obs}} = \{(n_1, \ldots, n_4); m\}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top \in \mathbb{T}_4$ be the cell probability vector. Suppose that the observed-data likelihood function of $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) \propto \left(\prod_{i=1}^{4} \theta_i^{n_i}\right) \times (\theta_1 + \theta_2 + \theta_3)^m.$$

Use the DA algorithm to obtain posterior samples from the observed-data posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

Solution: We introduce a latent vector $\mathbf{z} = (Z_1, Z_2, Z_3)^\top$ to split the term $(\theta_1 + \theta_2 + \theta_3)^m$ so that the conditional predictive distribution is

$$\mathbf{z}|(Y_{\text{obs}}, \boldsymbol{\theta}) \sim \text{Multinomial}\left(m; \frac{\theta_1}{\theta_{123}}, \frac{\theta_2}{\theta_{123}}, \frac{\theta_3}{\theta_{123}}\right),$$

where $\theta_{123} \hat{=} \theta_1 + \theta_2 + \theta_3$. Thus, the I-step is to draw $\boldsymbol{z}^{(t)} \sim f(\boldsymbol{z}|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$, where $\boldsymbol{z} = (z_1, z_2, z_3)^\top$.

The complete-data likelihood function is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}) \propto \prod_{i=1}^{4} \theta_i^{n_i + z_i}, \quad z_4 \hat{=} 0.$$

We take $\text{Dirichlet}_4(a_1, a_2, a_3, a_4)$ as the prior of $\boldsymbol{\theta}$, then the complete-data posterior is

$$\boldsymbol{\theta}|(Y_{\text{obs}}, \boldsymbol{z}) \sim \text{Dirichlet}(n_1 + z_1 + a_1, \ldots, n_4 + z_4 + a_4).$$

Thus, the P-step is to draw $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}^{(t)})$. ‖

4.4 The Gibbs sampler

33• Purpose of the Gibbs sampler

- Given all full univariate conditional distributions

$$f(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d), \qquad i = 1, \ldots, d,$$

how to generate samples from the joint distribution of $(X_1, \ldots, X_d)^\top$.

33.1• Background

— The Gibbs sampler is one of the best known MCMC sampling algorithms in the Bayesian computation literature.

— It was originally developed by Geman & Geman (1984) for simulating posterior samples in image reconstruction.

— The seminal paper of Gelfand & Smith (1990) introduces the Gibbs sampler to the main stream statistical literature and has had far-reaching impact on the application of Bayesian method.

33.2• A brief review

— The enormous potential of Gibbs sampling in complex statistical modeling is now being realized although there are still issues regarding convergence and the speed of computation.

— Various aspects of Gibbs sampler and MCMC methods are summarized by Smith & Roberts (1993), Besag & Green (1993), Gilks *et al.* (1993), and the discussions on all these papers.

— Casella & George (1992) and Arnold (1993) provide excellent tutorials on the Gibbs sampler.

4.4.1 The formulation of the Gibbs sampling

34• The method

- Suppose that we want to simulate a random sample from the random vector $\mathbf{x} = (X_1, \ldots, X_d)^\top$ with a joint cdf $F(\cdot)$.
- Assume that $F(\cdot)$ is either unknown or very complicated, but for each i, the full univariate conditional distribution

$$f(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$$

  is known and relatively easy to simulate.
- Choose a starting point $\boldsymbol{x}^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})^\top$ and set $t = 0$, the Gibbs sampling iterates the following loop:

  — Draw $x_1^{(t+1)} \sim f(x_1 | x_2^{(t)}, \ldots, x_d^{(t)})$;

— Draw $x_2^{(t+1)} \sim f(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_d^{(t)})$;

— ……

— Draw $x_d^{(t+1)} \sim f(x_d|x_1^{(t+1)}, \ldots, x_{d-1}^{(t+1)})$.

34.1• Issue of convergence

— Gelfand & Smith (1990) showed that under mild conditions, the vector sequence $\{\boldsymbol{x}^{(t)}\}_{t=1}^{\infty}$ has a stationary distribution $F(\cdot)$.

— Schervish & Carlin (1992) provided a sufficient condition that guarantees geometric convergence.

— Roberts & Polson (1994) discussed other properties regarding geometric convergence.

Example 4.9 (Dirichlet distribution). Use the Gibbs sampler to draw samples from $(X_1, X_2, X_3, 1 - \sum_{i=1}^3 X_i)^\top \sim \text{Dirichlet}(a_1, a_2, a_3, a_4)$.

Solution: Let $X_4 = 1 - \sum_{i=1}^3 X_i$, then the joint density of $(X_1, X_2, X_3)^\top$ is

$$f(x_1, x_2, x_3) = \frac{1}{B(a_1, \ldots, a_4)} \prod_{i=1}^4 x_i^{a_i - 1},$$

where $x_i > 0$, $i = 1, 2, 3$, and $x_1 + x_2 + x_3 \leqslant 1$. The conditional density of $X_1|(X_2 = x_2, X_3 = x_3)$ is

$$\begin{aligned} f(x_1|x_2, x_3) &\propto x_1^{a_1-1}(1 - x_1 - x_2 - x_3)^{a_4-1} \\ &\propto x_1^{a_1-1}(1 - x_2 - x_3 - x_1)^{a_4-1} \\ &\propto \left(\frac{x_1}{1 - x_2 - x_3}\right)^{a_1-1} \left(1 - \frac{x_1}{1 - x_2 - x_3}\right)^{a_4-1}. \end{aligned}$$

Define

$$Y_1 \triangleq \frac{X_1}{1 - x_2 - x_3},$$

then $f(y_1|x_2, x_3) \propto y_1^{a_1-1}(1 - y_1)^{a_4-1}$. That is, $Y_1|(x_2, x_3) \sim \text{Beta}(a_1, a_4)$. Since $\text{Beta}(a_1, a_4)$ is independent of (x_2, x_3), we have $Y_1 \sim \text{Beta}(a_1, a_4)$ and Y_1 is independent of (X_2, X_3). Hence, we obtain

$$X_1|(x_2, x_3) \overset{\text{d}}{=} (1 - x_2 - x_3)Y_1, \quad Y_1 \sim \text{Beta}(a_1, a_4).$$

Similarly, we can derive the following full conditional distributions:

$$X_2|(x_1, x_3) \stackrel{\mathrm{d}}{=} (1 - x_1 - x_3)Y_2, \quad Y_2 \sim \mathrm{Beta}(a_2, a_4),$$

$$X_3|(x_1, x_2) \stackrel{\mathrm{d}}{=} (1 - x_1 - x_2)Y_3, \quad Y_3 \sim \mathrm{Beta}(a_3, a_4),$$

where Y_2 is independent of (X_1, X_3) and Y_3 is independent of (X_1, X_2).

Thus, first we need to choose a starting vector $(x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^\top$ such that $x_i^{(0)} > 0$, $i = 1, 2, 3$, $\sum_{i=1}^3 x_i^{(0)} \leqslant 1$. Then we independently simulate $y_1^{(1)} \sim \mathrm{Beta}(a_1, a_4)$, $y_2^{(1)} \sim \mathrm{Beta}(a_2, a_4)$ and $y_3^{(1)} \sim \mathrm{Beta}(a_3, a_4)$. Let

$$\begin{aligned} x_1^{(1)} &= (1 - x_2^{(0)} - x_3^{(0)})y_1^{(1)}, \\ x_2^{(1)} &= (1 - x_1^{(1)} - x_3^{(0)})y_2^{(1)}, \\ x_3^{(1)} &= (1 - x_1^{(1)} - x_2^{(1)})y_3^{(1)}, \end{aligned}$$

then $(x_1^{(1)}, x_2^{(1)}, x_3^{(1)})^\top$ is the first iteration of the Gibbs sampler. We can compute the second, third and higher iterations in a similar fashion. As $t \to \infty$, the distribution of $(X_1^{(t)}, X_2^{(t)}, X_3^{(t)}, 1 - \sum_{i=1}^3 X_i^{(t)})^\top$ converges to the desired Dirichlet distribution. ‖

4.4.2 The two–block Gibbs sampling

35• How to improve the efficiency of the Gibbs sampler?

- Since the rate of convergence of the Gibbs sampler is controlled by the maximal correlation between the states of two consecutive Gibbs iterations, Liu *et al.* (1994) and Liu (1994) argued that grouping (or blocking) highly correlated components together in the Gibbs sampler can greatly improve its efficiency.

36• Two–block Gibbs sampler

- For example, we can split $\mathbf{x} = (X_1, \ldots, X_d)^\top$ into two blocks: $\mathbf{x}_1 = (X_1, \ldots, X_r)^\top$ and $\mathbf{x}_2 = (X_{r+1}, \ldots, X_d)^\top$, resulting in the following two–block Gibbs sampler:
  - Draw $\boldsymbol{x}_1^{(t+1)} \sim f(\boldsymbol{x}_1|\boldsymbol{x}_2^{(t)})$;
  - Draw $\boldsymbol{x}_2^{(t+1)} \sim f(\boldsymbol{x}_2|\boldsymbol{x}_1^{(t+1)})$.

36.1• The modern DA algorithm

— Particularly, in the original DA algorithm letting $m = 1$ and treating Z as the first block and $\boldsymbol{\theta}$ the second block, we obtain

$$\begin{aligned} &\text{I-step:} \quad \text{Draw} \quad z^{(t)} \sim f(z|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}); \\ &\text{P-step:} \quad \text{Draw} \quad \boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|Y_{\text{obs}}, z^{(t)}). \end{aligned} \tag{4.29}$$

Example 4.10 (Two-parameter multinomial model). Consider the following two-parameter multinomial model (Gelfand & Smith 1990):

$$\mathbf{y} \sim \text{Multinomial}_5(n; a_1\theta_1 + b_1, a_2\theta_1 + b_2, a_3\theta_2 + b_3, a_4\theta_2 + b_4, c\theta_3),$$

where $a_i, b_i > 0$ are known, $0 \leqslant c = 1 - \sum_{i=1}^4 b_i = a_1 + a_2 = a_3 + a_4 \leqslant 1$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top \in \mathbb{T}_3$. Use the Gibbs sampler to generate the posterior samples from the observed-data posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}})$, where $Y_{\text{obs}} = \{y_i\}_{i=1}^5$ is the realization of $\mathbf{y} = (Y_1, \ldots, Y_5)^\top$.

Solution: We introduce a latent vector $\mathbf{z} = (Z_1, \ldots, Z_4)^\top$ so that the conditional predictive density is

$$f(\boldsymbol{z}|Y_{\text{obs}}, \boldsymbol{\theta}) = \prod_{i=1}^4 \text{Binomial}(z_i|y_i, p_i),$$

where

$$p_i \mathrel{\hat{=}} \frac{a_i\theta_1 I(1 \leqslant i \leqslant 2)}{a_i\theta_1 + b_i} + \frac{a_i\theta_2 I(3 \leqslant i \leqslant 4)}{a_i\theta_2 + b_i}.$$

Thus, the I-step is to draw $\boldsymbol{z}^{(t)} \sim f(\boldsymbol{z}|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})$.

The augmented sampling distribution is

$$\text{Multinomial}_9(n; a_1\theta_1, b_1, a_2\theta_1, b_2, a_3\theta_2, b_3, a_4\theta_2, b_4, c\theta_3).$$

Equivalently, the complete-data likelihood is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}) \propto \theta_1^{z_1+z_2}\theta_2^{z_3+z_4}\theta_3^{y_5}.$$

A natural prior on $\boldsymbol{\theta}$ is the Dirichlet distribution $\text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ so that the augmented posterior is given by

$$\boldsymbol{\theta}|(Y_{\text{obs}}, \boldsymbol{z}) \sim \text{Dirichlet}(z_1 + z_2 + \alpha_1, z_3 + z_4 + \alpha_2, y_5 + \alpha_3).$$

Hence, the P-step is to draw $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z}^{(t)})$. ‖

37• Difference between DA algorithm and Gibbs sampler

- The DA algorithm is a special case of the Gibbs sampler.
- The DA algorithm is only applied to missing data problems with EM structure; i.e., there is a latent–variable structure.
- The DA algorithm is a stochastic version of the deterministic EM algorithm.
- The Gibbs sampler is much universal, it can be applied to both the complete-data cases and the incomplete-data cases.
- The Gibbs sampler is a stochastic version of the deterministic ECM algorithm.

38• Gibbs sampler in the complete-data cases

- Now $Y_{\text{obs}} = Y_{\text{com}}$.

  Step 1: Find the posterior distribution:

$$p(\theta_1, \ldots, \theta_d | Y_{\text{obs}}) \propto L(\boldsymbol{\theta} | Y_{\text{obs}}) \cdot \pi(\boldsymbol{\theta}).$$

  Step 2: Draw posterior samples from the full conditional distributions: $p(\theta_i | Y_{\text{obs}}, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$ for $i = 1, \ldots, d$.

39• Gibbs sampler in the incomplete-data cases

Step 1: Augment the observed data Y_{obs} with the latent data Z.

Step 2: Find the complete-data posterior distribution:

$$p(\theta_1, \ldots, \theta_d | Y_{\text{obs}}, z) \propto L(\boldsymbol{\theta} | Y_{\text{obs}}, z) \cdot \pi(\boldsymbol{\theta})$$

Step 3: Draw $Z = z \sim f(z | Y_{\text{obs}}, \boldsymbol{\theta})$.

Step 4: Draw posterior samples from the full conditional distributions:

$$p(\theta_i | Y_{\text{obs}}, z, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d), \quad i = 1, \ldots, d.$$

4.5 The Exact IBF Sampling

40• Purpose of the exact IBF sampling

- Let Y_{obs} be the observed data, Z the missing or latent data and $\boldsymbol{\theta}$ the parameter vector.
- Assume that both the complete-data posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and conditional predictive distribution $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ are available.
- The objective is to obtain i.i.d. samples from the observed posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

41• Conditional sampling method

- Note that $p(\boldsymbol{\theta}, z|Y_{\text{obs}}) = f(z|Y_{\text{obs}}) \times p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$.
- The conditional sampling method in §1.6 states that: If we could obtain independent samples $\{z^{(\ell)}\}_{\ell=1}^{L}$ from $f(z|Y_{\text{obs}})$ and generate $\boldsymbol{\theta}^{(\ell)} \sim p(\boldsymbol{\theta}|Y_{\text{obs}}, z^{(\ell)})$, $\ell = 1, \ldots, L$, then $\{\boldsymbol{\theta}^{(\ell)}\}_{\ell=1}^{L}$ are i.i.d. samples from $p(\boldsymbol{\theta}|Y_{\text{obs}})$.
- Thus, the key is to be able to generate independent samples from $f(z|Y_{\text{obs}})$.

42• Use of the sampling-wise IBF

- Let $\mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$ and $\mathcal{S}_{(Z|Y_{\text{obs}})}$ denote the conditional supports of $\boldsymbol{\theta}|Y_{\text{obs}}$ and $Z|Y_{\text{obs}}$, respectively.
- By exchanging the role of $\boldsymbol{\theta}$ and Z, the sampling-wise IBF (4.28) becomes
$$f(z|Y_{\text{obs}}) \propto \frac{f(z|Y_{\text{obs}}, \boldsymbol{\theta}_0)}{p(\boldsymbol{\theta}_0|Y_{\text{obs}}, z)}, \tag{4.30}$$
for an arbitrary $\boldsymbol{\theta}_0 \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$ and all $z \in \mathcal{S}_{(Z|Y_{\text{obs}})}$.

43• Formulation of the exact IBF sampling

- Consider a special case that Z is a discrete r.v. or discrete random vector taking *finite* values on the domain.

43.1• Identification of the conditional support

— We use $\mathcal{S}_{(Z|Y_{\text{obs}},\boldsymbol{\theta})} = \{z_1, \ldots, z_K\}$ to denote the conditional support of $Z|(Y_{\text{obs}}, \boldsymbol{\theta})$.

— Since the $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ is available, we can directly identify $\{z_k\}_{k=1}^K$ and thus all $\{z_k\}_{k=1}^K$ are known.

— Next, we assume that $\{z_k\}_{k=1}^K$ do not depend on the value of $\boldsymbol{\theta}$, thus

$$\mathcal{S}_{(Z|Y_{\text{obs}})} = \mathcal{S}_{(Z|Y_{\text{obs}},\boldsymbol{\theta})} = \{z_1, \ldots, z_K\}.$$

43.2• Calculating the probability weights

— Because of the discreteness of Z, the notation $f(z_k|Y_{\text{obs}})$ will be used to denote the pmf; i.e., $f(z_k|Y_{\text{obs}}) = \Pr(Z = z_k|Y_{\text{obs}})$.

— Therefore, the key is to find $p_k = f(z_k|Y_{\text{obs}})$ for $k = 1, \ldots, K$.

— For some $\boldsymbol{\theta}_0 \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$, let

$$q_k = q_k(\boldsymbol{\theta}_0) = \frac{\Pr(Z = z_k|Y_{\text{obs}}, \boldsymbol{\theta}_0)}{p(\boldsymbol{\theta}_0|Y_{\text{obs}}, z_k)}, \quad k = 1, \ldots, K. \tag{4.31}$$

— As both $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ are available, the computation of (4.31) is straightforward.

— Observing that all $\{q_k\}_{k=1}^K$ depend on $\boldsymbol{\theta}_0$, we then denote q_k by $q_k(\boldsymbol{\theta}_0)$ to emphasize its dependency on $\boldsymbol{\theta}_0$.

— From the sampling-wise IBF (4.30), we immediately obtain

$$p_k = \frac{q_k(\boldsymbol{\theta}_0)}{\sum_{k'=1}^K q_{k'}(\boldsymbol{\theta}_0)}, \quad k = 1, \ldots, K, \tag{4.32}$$

where $\{p_k\}_{k=1}^K$ do not depend on $\boldsymbol{\theta}_0$.

Proof of (4.32): Since $p_k \propto q_k = q_k(\boldsymbol{\theta}_0)$, we can write $p_k = c^{-1} \cdot q_k$, where c is the normalizing constant. From

$$1 = \sum_{k=1}^K p_k = c^{-1} \sum_{k=1}^K q_k,$$

we obtain $c = \sum_{k=1}^K q_k$. □

43.3• Generation of i.i.d. samples

— We can use the built-in R function

$$\mathsf{sample}(\boldsymbol{z},\ \mathsf{N},\ \mathsf{prob} = \boldsymbol{p},\ \mathsf{replace} = \mathsf{F})$$

to generate i.i.d. samples from $f(z|Y_{\text{obs}})$ since it is a discrete distribution with probability p_k on z_k for $k = 1, \ldots, K$, where $\boldsymbol{z} = (z_1, \ldots, z_K)^\top$ and $\boldsymbol{p} = (p_1, \ldots, p_K)^\top$.

44• The exact IBF sampling

- Given both the complete-data posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and the conditional predictive distribution $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$:

  Step 1: Identify $\mathcal{S}_{(Z|Y_{\text{obs}})} = \{z_1, \ldots, z_K\}$ from $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ and calculate $\{p_k\}_{k=1}^K$ according to (4.32) and (4.31);

  Step 2: Generate i.i.d. samples $\{z^{(\ell)}\}_{\ell=1}^L$ of Z from the pmf $f(z|Y_{\text{obs}})$ with probabilities $\{p_k\}_{k=1}^K$ on $\{z_k\}_{k=1}^K$;

  Step 3: Generate $\boldsymbol{\theta}^{(\ell)} \sim p(\boldsymbol{\theta}|Y_{\text{obs}}, z^{(\ell)})$ for $\ell = 1, \ldots, L$, then $\{\boldsymbol{\theta}^{(\ell)}\}_{\ell=1}^L$ are i.i.d. samples from the posterior distribution $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

Example 4.11 (Genetic linkage model revisited). We consider the following small sample dataset for the genetic linkage model presented in Example 2.6: $Y_{\text{obs}} = \{y_1, y_2, y_3, y_4\} = \{14, 0, 1, 5\}$. Use the exact IBF sampling to generate i.i.d. samples from $p(\theta|Y_{\text{obs}})$.

Solution: Use Beta(a, b) as the prior distribution of θ, we have

$$\theta|(Y_{\text{obs}}, z) \sim \text{Beta}(a + y_4 + z, b + y_2 + y_3). \tag{4.33}$$

From (2.19), obviously Z is discrete and takes values on

$$\mathcal{S}_{(Z|Y_{\text{obs}},\theta)} = \mathcal{S}_{(Z|Y_{\text{obs}})} = \{0, 1, \ldots, y_1\}.$$

Let $a = b = 1$ (i.e., the uniform prior) and $\theta_0 = 0.5$, then $q_k(\theta_0)$ and p_k can be calculated according to (4.31) and (4.32). These results are listed in Table 4.1. If we choose $\theta_0 = 0.8$, then $\{q_k(\theta_0)\}_{k=1}^K$ will vary but $\{p_k\}_{k=1}^K$ remain the same. Hence, the exact IBF sampling is as follows:

(i) Draw $L = 30{,}000$ independent samples $\{z^{(\ell)}\}_{\ell=1}^L$ of Z from the discrete distribution $f(z|Y_{\text{obs}})$ with $p_k = \Pr(Z = z_k|Y_{\text{obs}})$ given in Table 4.1;

(ii) Generate $\theta^{(\ell)} \sim p(\theta|Y_{\text{obs}}, z^{(\ell)})$ defined in (4.33) for $\ell = 1, \ldots, L$, then $\{\theta^{(\ell)}\}_{\ell=1}^{L}$ are i.i.d. samples from the observed-data posterior distribution $p(\theta|Y_{\text{obs}})$.

Table 4.1 **The values of $\{q_k(\theta_0)\}$ with $\theta_0 = 0.5$ and $\{p_k\}$**

k	z_k	$q_k(\theta_0)$	p_k	k	z_k	q_k	p_k
1	0	0.0670	0.0094	9	8	1.572×10^{-1}	2.200×10^{-2}
2	1	0.3518	0.0493	10	9	4.580×10^{-2}	6.400×10^{-3}
3	2	0.8894	0.1245	11	10	1.012×10^{-2}	1.416×10^{-3}
4	3	1.4230	0.1992	12	11	1.635×10^{-3}	2.289×10^{-4}
5	4	1.6010	0.2241	13	12	1.829×10^{-4}	2.560×10^{-5}
6	5	1.3340	0.1868	14	13	1.266×10^{-5}	1.772×10^{-6}
7	6	0.8466	0.1185	15	14	4.090×10^{-7}	5.727×10^{-8}
8	7	0.4147	0.0581				

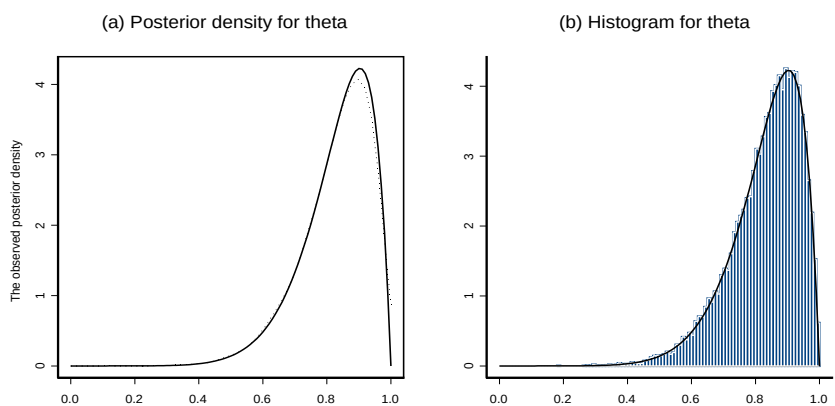

Figure 4.1 (a) The comparison between the observed posterior density of θ (solid curve) exactly given by (4.34) with the dotted curve estimated by a kernel density smoother based on $L = 30{,}000$ i.i.d. samples generated via the exact IBF sampling. (b) The histogram of θ based on $L = 30{,}000$ i.i.d. samples generated via the exact IBF sampling.

The accuracy of the exact IBF sampling is shown in Figure 4.1(a), where the hardly visible dotted curve is estimated by a kernel density smoother

based on the i.i.d. IBF samples and the solid line is exactly given by

$$p(\theta|Y_{\text{obs}}) = \frac{(\theta+2)^{y_1}\theta^{a+y_4-1}(1-\theta)^{b+y_2+y_3-1}}{\sum_{z=0}^{y_1}\binom{y_1}{z}B(a+y_4+z, b+y_2+y_3)2^{y_1-z}}. \tag{4.34}$$

In addition, the histogram based on these samples is plotted in Figure 4.1(b), which shows that the exact IBF sampling has recovered the density completely as expected. ‖

4.6 The IBF sampler

4.6.1 Background and the basic idea

45• Background

- In the previous section, we introduced the exact IBF sampling to obtain i.i.d. samples *exactly* from an observed posterior distribution for discrete missing-data problems.
- Generally, the exact IBF sampling is feasible only when the discrete latent vector (or missing data) Z is of low dimension.
- For more general cases, (e.g., when Z is discrete but of high dimension, or when Z is continuous and so on), the non-iterative sampling method may require to be modified.

46• Purpose of this section

- The purpose of this section is to develop such a non-iterative sampling method (called IBF sampler), as opposed to the iterative sampling in an MCMC, for computing posteriors based on IBF and SIR to obtain i.i.d. samples *approximately* from the observed posterior distribution while using the posterior mode and structure from an EM algorithm.

47• Basic idea of the IBF sampler

- The basic idea of the IBF sampler in the EM framework is to use EM algorithm to obtain an optimal importance sampling density and the sampling-wise IBF to get i.i.d. samples from the posterior distribution. Specifically,

— Firstly, we augment the observed data with latent data and obtain the structure of augmented posterior/conditional predictive distributions as in the EM or the DA algorithm;

— Secondly, in the class of built-in IS densities provided by the sampling-wise IBF, we choose the best IS density by using preliminary estimates from the EM algorithm so that the overlap area under the target density and the IS density is large;

— Thirdly, the sampling-wise IBF and SIR are combined to generate i.i.d. samples approximately from the observed posterior distribution.

48• Application range of the IBF sampler

- The synergy of IBF, EM and SIR creates an attractive sampling approach for Bayesian computation.
- Since the sampling-wise IBF and the EM share the DA structure, the IBF sampler via the EM does not require extra derivations, and can be applied to problems where the EM is applicable while obtaining the whole posterior.

4.6.2 The formulation of the IBF sampler

49• Three basic assumptions

- Let $\tilde{\boldsymbol{\theta}}$ denote the mode of the observed posterior density $p(\boldsymbol{\theta}|Y_{\text{obs}})$.
- In this subsection, we make the following three basic assumptions:
  — Both $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ are available, indicating that both densities have closed–form expressions or they can easily be evaluated or sampled;
  — The posterior mode $\tilde{\boldsymbol{\theta}}$ is already obtained via an EM algorithm;
  — The joint support is a product space; i.e.,

$$\mathcal{S}_{(\boldsymbol{\theta},Z|Y_{\text{obs}})} = \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})} \times \mathcal{S}_{(Z|Y_{\text{obs}})}.$$

49.1• Goal–driven motivations

— The goal is to generate posterior samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(I)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta}|Y_{\text{obs}})$, which can be achieved if we could generate

$$\{\boldsymbol{\theta}^{(i)}, z^{(i)}\}_{i=1}^{I} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta}, z|Y_{\text{obs}}) = p(\boldsymbol{\theta}|Y_{\text{obs}}, z) \cdot f(z|Y_{\text{obs}}).$$

— Since $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ is available, the key is to be able to generate samples $\{z^{(i)}\}_{i=1}^{I} \overset{\text{iid}}{\sim} f(z|Y_{\text{obs}})$, which can be achieved by using the sampling-wise IBF (4.30):

$$f(z|Y_{\text{obs}}) \propto \frac{f(z|Y_{\text{obs}}, \boldsymbol{\theta}_0)}{p(\boldsymbol{\theta}_0|Y_{\text{obs}}, z)},$$

or

$$f(z|Y_{\text{obs}}) = \frac{c_0}{p(\boldsymbol{\theta}_0|Y_{\text{obs}}, z)} \cdot f(z|Y_{\text{obs}}, \boldsymbol{\theta}_0) \,\hat{=}\, \text{w}(z) \cdot g(z).$$

— Considering the conditional predictive density $f(z|Y_{\text{obs}}, \boldsymbol{\theta}_0)$ as an approximation to the marginal predictive density $f(z|Y_{\text{obs}})$, the IBF sampling is realized via SIR as follows.

50• The IBF sampler

- Find the posterior mode $\tilde{\boldsymbol{\theta}}$ via an EM algorithm based on $p(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$, and set $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$.

  Step 1: Draw J i.i.d. samples $\{z^{(j)}\}_{j=1}^{J}$ of Z from $f(z|Y_{\text{obs}}, \boldsymbol{\theta}_0)$;

  Step 2: Calculate the reciprocals of the augmented posterior densities to obtain the weights

$$\omega_j = \frac{\text{w}(z^{(j)})}{\sum_{\ell=1}^{J} \text{w}(z^{(\ell)})} = \frac{p^{-1}(\boldsymbol{\theta}_0|Y_{\text{obs}}, z^{(j)})}{\sum_{\ell=1}^{J} p^{-1}(\boldsymbol{\theta}_0|Y_{\text{obs}}, z^{(\ell)})} \tag{4.35}$$

  for $j = 1, \ldots, J$;

  Step 3: Choose a subset from $\{z^{(j)}\}_{j=1}^{J}$ via resampling *without replacement* from the discrete distribution on $\{z^{(j)}\}_{j=1}^{J}$ with probabilities $\{\omega_j\}_{j=1}^{J}$ to obtain an i.i.d. sample of size $I\,(< J)$ approximately from $f(z|Y_{\text{obs}})$, denoted by $\{z^{(k_i)}\}_{i=1}^{I}$;

  Step 4: Generate $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|Y_{\text{obs}}, z^{(k_i)})$ for $i = 1, \ldots, I$, then $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{I}$ are i.i.d. samples from $p(\boldsymbol{\theta}|Y_{\text{obs}})$.

4.6.3 Theoretical justification for choosing $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$

51• How to choose $\boldsymbol{\theta}_0$?

- It is worth noting that only one pre-specified $\boldsymbol{\theta}_0$ is needed for the whole IBF sampling process although the sampling-wise IBF (4.30) holds for any $\boldsymbol{\theta}_0$ in the support of $\boldsymbol{\theta}|Y_{\text{obs}}$.
- Clearly, the sampling-wise IBF (4.30) provides a natural class of IS densities:
$$\left\{ f(z|Y_{\text{obs}}, \boldsymbol{\theta}): \ \ \boldsymbol{\theta} \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})} \right\},$$
which are available from the model specification.
- However, the efficiency of the IBF sampler depends on how well the IS density approximates the target function $f(z|Y_{\text{obs}})$.
- Since (4.30) holds for any given $\boldsymbol{\theta}_0 \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}$, it suffices to select one $\boldsymbol{\theta}_0$ such that $f(z|Y_{\text{obs}}, \boldsymbol{\theta}_0)$ best approximates $f(z|Y_{\text{obs}})$.
- Heuristically, if $\boldsymbol{\theta}_0$ is chosen to be the observed posterior mode $\tilde{\boldsymbol{\theta}}$, the overlap area under the two functions would be substantial since the approximation is accurate to the order of $O(1/n)$, as shown in the following theorem.

Theorem 4.1 (The relation of $f(z|Y_{\text{obs}})$ and conditional predictive density). Let the observed posterior density $p(\boldsymbol{\theta}|Y_{\text{obs}})$ be unimodal with mode $\tilde{\boldsymbol{\theta}}$ and n be the sample size of the observed data Y_{obs}. Then

$$f(z|Y_{\text{obs}}) = f(z|Y_{\text{obs}}, \tilde{\boldsymbol{\theta}})\{1 + O(1/n)\}, \tag{4.36}$$

see Tan *et al.* (2003). ‖

Proof: Let $g(\boldsymbol{\theta})$ be an arbitrarily smooth, positive function for $\boldsymbol{\theta} \in \mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})} \subseteq \mathbb{R}^d$, $L(\boldsymbol{\theta}|Y_{\text{obs}})$ be the likelihood function and $\pi(\boldsymbol{\theta})$ be the prior. The posterior mean of $g(\boldsymbol{\theta})$ is given by

$$\begin{aligned} E\{g(\boldsymbol{\theta})|Y_{\text{obs}}\} &= \int_{\mathcal{S}_{(\boldsymbol{\theta}|Y_{\text{obs}})}} g(\boldsymbol{\theta}) p(\boldsymbol{\theta}|Y_{\text{obs}}) \, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{\int g(\boldsymbol{\theta}) \exp\{n\, \ell(\boldsymbol{\theta})\} \, \mathrm{d}\boldsymbol{\theta}}{\int \exp\{n\, \ell(\boldsymbol{\theta})\} \, \mathrm{d}\boldsymbol{\theta}}, \end{aligned} \tag{4.37}$$

where $n\,\ell(\boldsymbol{\theta}) = \log\{L(\boldsymbol{\theta}|Y_{\text{obs}})\pi(\boldsymbol{\theta})\} \propto \log\{p(\boldsymbol{\theta}|Y_{\text{obs}})\}$. Thus $\ell(\boldsymbol{\theta})$ has the same mode as $p(\boldsymbol{\theta}|Y_{\text{obs}})$; i.e., $\nabla\ell(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$. Applying Laplace's method to the numerator in (4.37) gives

$$\int g(\boldsymbol{\theta}) \exp\{n\,\ell(\boldsymbol{\theta})\}\,\mathrm{d}\boldsymbol{\theta} \approx g(\tilde{\boldsymbol{\theta}}) \exp\{n\,\ell(\tilde{\boldsymbol{\theta}})\}\Big(\frac{2\pi}{n}\Big)^{d/2}|\boldsymbol{\Sigma}|^{1/2},$$

where

$$\boldsymbol{\Sigma}_{d\times d} = -\left(\frac{\partial^2\ell(\tilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}\right)^{-1}.$$

Similarly, for the denominator in (4.37), we have

$$\int \exp\{n\,\ell(\boldsymbol{\theta})\}\,\mathrm{d}\boldsymbol{\theta} \approx \exp\{n\,\ell(\tilde{\boldsymbol{\theta}})\}\Big(\frac{2\pi}{n}\Big)^{d/2}|\boldsymbol{\Sigma}|^{1/2}.$$

The resulting ratio is $g(\tilde{\boldsymbol{\theta}})$ up to error $O(1/n)$ as shown in Tierney & Kadane (1986). Thus,

$$E\{g(\boldsymbol{\theta})|Y_{\text{obs}}\} = g(\tilde{\boldsymbol{\theta}})\{1 + O(1/n)\}.$$

Since

$$f(z|Y_{\text{obs}}) = \int f(z|Y_{\text{obs}},\boldsymbol{\theta})p(\boldsymbol{\theta}|Y_{\text{obs}})\,\mathrm{d}\boldsymbol{\theta} = E\{f(z|Y_{\text{obs}},\boldsymbol{\theta})|Y_{\text{obs}}\},$$

(4.36) follows immediately. □

Example 4.12 (The genetic linkage model revisited). To illustrate the IBF sampler, we revisit the genetic linkage model with observed data $Y_{\text{obs}} = \{y_1, y_2, y_3, y_4\} = \{125, 18, 20, 34\}$.

Solution: We first use the EM algorithm to compute the posterior mode $\tilde{\theta}$. Based on (2.19) and (4.33), both E-step and M-step have following closed-form expressions:

$$z^{(t)} = \frac{y_1\theta^{(t)}}{\theta^{(t)}+2}, \quad \theta^{(t+1)} = \frac{a + y_4 + z^{(t)} - 1}{(a + y_4 + z^{(t)} - 1) + (b + y_2 + y_3 - 1)}.$$

Setting $\theta^{(0)} = 0.5$ and $a = b = 1$ corresponding to the uniform prior, the EM converged to $\tilde{\theta} = 0.6268$ after four iterations.

We implement the IBF sampler based on (4.30) by generating

$$z^{(j)} \overset{\text{iid}}{\sim} \text{Binomial}(y_1, \tilde{\theta}/(\tilde{\theta}+2)), \quad j = 1, \ldots, J,$$

with $J = 30{,}000$ and computing the weights $\{\omega_j\}_{j=1}^J$ based on (4.35). Then, resample without replacement from the discrete distribution on $\{z^{(j)}\}_{j=1}^J$ with probabilities $\{\omega_j\}_{j=1}^J$ to obtain an i.i.d. sample of size $I = 10{,}000$ approximately from $f(z|Y_{\text{obs}})$, denoted by $\{z^{(k_i)}\}_{i=1}^I$. Finally, we generate

$$\theta^{(i)} \overset{\text{ind}}{\sim} \text{Beta}(1 + y_4 + z^{(k_i)},\ 1 + y_2 + y_3), \quad i = 1, \ldots, I,$$

then $\{\theta^{(i)}\}_{i=1}^I$ are i.i.d. posterior samples from $p(\theta|Y_{\text{obs}})$.

The accuracy of the IBF sampler is remarkable as shown in Figure 4.2(a), where the solid curve is given exactly by (4.34) while the dotted curve is estimated by a kernel density smoother based on these i.i.d. IBF output. In addition, the histogram based on these samples is plotted in Figure 4.2(b), which shows that the IBF sampler has recovered the density completely.

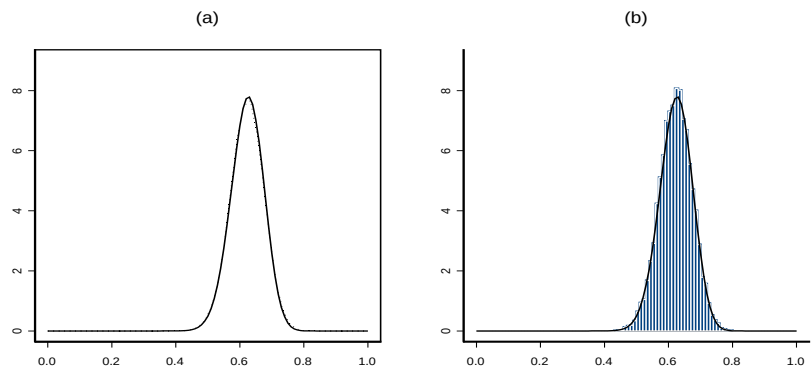

Figure 4.2 (a) The comparison between the observed posterior density of θ (solid curve) given exactly by (4.34) with the dotted curve estimated by a kernel density smoother based on i.i.d. samples obtained via the IBF sampler ($J = 30{,}000$, $I = 10{,}000$, without replacement). (b) The histogram of θ based on $I = 10{,}000$ i.i.d. samples generated via the IBF sampler. ‖

Exercise 4

4.1 Let $X \sim N(0, 1)$ and $Y \doteq X^2$.

(a) Find the conditional pdfs $f_{(X|Y)}(x|y)$ and $f_{(Y|X)}(y|x)$.

(b) Find the joint cdf of X and Y.

(c) Does the identity (4.1) hold? Why?

4.2 Let $X|(Y = y) \sim N(y, 1)$ and $Y|(X = x) \sim N(x, 1)$. Use the sampling-wise IBF to show that the marginal pdf of X does not exist.

4.3 Let two conditional distributions be truncated normal; i.e.,

$$\begin{aligned} X|(Y=y) &\sim \mathrm{TN}(\mu_1+\rho\sigma_1\sigma_2^{-1}(y-\mu_2),\, \sigma_1^2(1-\rho^2); a_1, b_1) \quad \text{and} \\ Y|(X=x) &\sim \mathrm{TN}(\mu_2+\rho\sigma_2\sigma_1^{-1}(x-\mu_1),\, \sigma_2^2(1-\rho^2); a_2, b_2). \end{aligned}$$

Use the point-wise IBF to find the marginal pdfs of X and Y.

4.4 Let two conditional pdfs be exponential restricted to $[0, b)$; that is,

$$\begin{aligned} f_{(X|Y)}(x|y) &= \frac{y\exp(-yx)}{1-\exp(-by)}, \quad 0 \leqslant x < b < +\infty, \\ f_{(Y|X)}(y|x) &= \frac{x\exp(-xy)}{1-\exp(-bx)}, \quad 0 \leqslant y < b < +\infty. \end{aligned}$$

(a) Find the marginal distribution of X.

(b) If $b = +\infty$, please discuss the existence of $f_X(x)$.

4.5 Let X be a discrete r.v. with pmf $p_i = \Pr(X = x_i)$ for $i = 1, 2$ and Y be a discrete r.v. with pmf $q_j = \Pr(Y = y_j)$ for $j = 1, 2$. Given two conditional distribution matrices

$$\boldsymbol{A} = \begin{pmatrix} 1/4 & 1/2 \\ 3/4 & 1/2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{pmatrix} 1/3 & 2/3 \\ 3/5 & 2/5 \end{pmatrix},$$

where the (i, j) element of $\boldsymbol{A}$ is $a_{ij} = \Pr(X = x_i|Y = y_j)$ and the (i, j) element of $\boldsymbol{B}$ is $b_{ij} = \Pr(Y = y_j|X = x_i)$.

(a) Find the marginal pmfs of X and Y.

(b) Find the joint distribution of (X, Y).

4.6 (Censored data from an exponential density). In survival or reliability analysis, one observes a random sample (McLachlan & Krishnan 1997)

$$Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Exponential}(\theta),$$

which is generally terminated before all of them are able to be observed. Suppose that the first r observations of $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$ are uncensored and the remaining $m - r$ are censored (c_i denotes a censored time) so that the observed data $Y_{\text{obs}} = \{y_1, \ldots, y_r; c_{r+1}, \ldots, c_m\}$. Augment the Y_{obs} with latent failure times $\mathbf{z} = (Z_{r+1}, \ldots, Z_m)^\top$, where $Z_i > c_i$ for $i = r + 1, \ldots, m$. Let $\boldsymbol{z} = (z_{r+1}, \ldots, z_m)^\top$ denote the realizations of $\mathbf{z}$.

(a) Show that the complete-data likelihood is given by

$$L(\theta|Y_{\text{obs}}, \boldsymbol{z}) \propto \theta^m \exp\{-\theta(\textstyle\sum_{i=1}^r y_i + \mathbf{1}^\top \boldsymbol{z})\}.$$

(b) If the Gamma(α_0, β_0) prior is put on θ, then

$$\begin{aligned} \theta|(Y_{\text{obs}}, \boldsymbol{z}) &\sim \text{Gamma}(m + \alpha_0,\ y^* + \mathbf{1}^\top \boldsymbol{z} + \beta_0), \\ \mathbf{1}^\top \mathbf{z} - c.|(Y_{\text{obs}}, \theta) &\sim \text{Gamma}(m - r, \theta), \end{aligned}$$

where $y^* \doteq \sum_{i=1}^r y_i$ and $c. \doteq \sum_{i=r+1}^m c_i$.

(c) Write down the Gibbs sampler for this example.

4.7 (Failures for nuclear pumps). Table 4.2 gives multiple failures of pumps in a nuclear plant (Gaver & O'Muircheartaigh 1987). Let the failures in time interval $(0, t_i]$ for the i-th pump follow a homogeneous Poisson process with rate λ_i $(i = 1, \ldots, m)$. Then, the number of failures

$$N_i \doteq N(t_i) \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_i t_i).$$

Let $Y_{\text{obs}} = \{t_i, N_i\}_1^m$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$.

(a) If prior distributions are specified by $\lambda_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_0, \beta)$ and $\beta \sim \text{Gamma}(a_0, b_0)$, then

$$\begin{aligned} f(\boldsymbol{\lambda}|Y_{\text{obs}}, \beta) &= \prod_{i=1}^m \text{Gamma}(\lambda_i|N_i + \alpha_0,\ t_i + \beta), \\ f(\beta|Y_{\text{obs}}, \boldsymbol{\lambda}) &= \text{Gamma}(\beta|a_0 + m\alpha_0,\ b_0 + \textstyle\sum_{i=1}^m \lambda_i). \end{aligned}$$

(b) For the data in Table 4.2, implement the Gibbs sampler.

Table 4.2 Numbers of failures and times for 10 pumps in a nuclear plant

Pump i	1	2	3	4	5	6	7	8	9	10
Failures N_i	5	1	5	14	3	19	1	1	4	22
Time t_i	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

4.8 In Example 4.11, calculate $\{q_k(\theta_0)\}_{k=1}^K$ with $\theta_0 = 0.8$, and then compute $\{p_k\}_{k=1}^K$. Compare your $\{p_k\}_{k=1}^K$ with the $\{p_k\}_{k=1}^K$ in Table 4.1.

4.9 In Exercise 2.21(d), let the prior distributions $\phi \sim \text{Beta}(a_0, b_0)$, $\lambda \sim \text{Gamma}(\alpha_0, \beta_0)$, and they are independent. What are the I-step and P-Step of the DA algorithm?

4.10 In Example 2.6, let $Y_{\text{obs}} = (y_1, y_2, y_3, y_4)^\top = (125, 18, 20, 34)^\top$ be the observed frequencies. Assume that the prior distribution of θ is $\text{Beta}(a_0, b_0)$ with $a_0 = b_0 = 1$.

(a) Use an EM algorithm to calculate the value of mode $\tilde{\theta}$ of the observed-data posterior density $p(\theta|Y_{\text{obs}})$ by writing an R code.

(b) Let the fixed point iteration of the EM algorithm in Exercise 4.10(a) can be represented by $\theta^{(t+1)} = h_1(\theta^{(t)})$. Derive the expression of $h_1(\cdot)$ and calculate the value of the convergence rate $r_1 \doteq |h_1'(\tilde{\theta})|$ of this EM algorithm. [Hint: see (2.9) and (2.2)]

(c) Use a corresponding DA algorithm to calculate the value of the posterior mean of θ by discarding the first half posterior samples of a total of 100,000 posterior samples $\{\theta^{(t)}\colon\ t = 1, \ldots, 100{,}000\}$.

(d) If we write $\theta + 2 = 3\theta + 2(1 - \theta)$, and augment the observed data Y_{obs} with another latent variable W by splitting $y_1 = W + (y_1 - W)$, please construct the second EM algorithm to calculate the value of the mode $\tilde{\theta}$ of the observed-data posterior density $p(\theta|Y_{\text{obs}})$ by writing an R code.

(e) Let the fixed point iteration of the second EM in Exercise 4.10(d) can be represented by $\theta^{(t+1)} = h_2(\theta^{(t)})$. Derive the expression of $h_2(\cdot)$ and calculate the value of the convergence rate $r_2 \doteq |h_2'(\tilde{\theta})|$ of the second EM. Which EM algorithm converges faster?

4.11 In Exercise 2.22(b), let the prior distribution of λ be $\text{Gamma}(\alpha_0, \beta_0)$. Based on the SR (2.65), derive the complete-data posterior distribution of λ. What are the I-step and P-Step of the DA algorithm?

4.12 (Clinical mastitis data). Table 4.3 displays numbers of cases of clinical mastitis (which is an inflammation usually caused by infection) in 127 dairy cattle herds over one year period (Robert & Casella 1999, p.334). Let $X_i \sim \text{Poisson}(\lambda_i)$, where X_i is the number of cases in herd i ($i = 1, \ldots, m$) and λ_i is the rate of infection in herd i. However, the data lack of independence because mastitis is infectious. To account for this over-dispersion, Schukken *et al.* (1991) presented a

complete hierarchical model by specifying $\lambda_i \sim \text{Gamma}(\alpha_0, \beta_i)$ and $\beta_i \sim \text{Gamma}(a_0, b_0)$. Let $Y_{\text{obs}} = \{x_i\}_{i=1}^m$, show that

$$\begin{aligned} \lambda_i|(Y_{\text{obs}}, \beta_i) &\sim \text{Gamma}(\lambda_i|x_i + \alpha_0,\ 1 + \beta_i), \\ \beta_i|(Y_{\text{obs}}, \lambda_i) &\sim \text{Gamma}(\beta|a_0 + \alpha_0,\ b_0 + \lambda_i). \end{aligned}$$

For this data set, implement the Gibbs sampler.

Table 4.3 Occurrences of clinical mastitis in 127 herds of dairy cattle

0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
1	1	1	1	2	2	2	2	2	2	2	2	3	3	3
3	3	3	3	3	3	4	4	4	4	4	4	4	4	5
5	5	5	5	5	5	5	6	6	6	6	6	6	6	6
6	7	7	7	7	7	7	8	8	8	8	8	9	9	9
10	10	10	10	11	11	11	11	11	11	11	12	12	12	12
13	13	13	13	13	14	14	15	16	16	16	16	17	17	17
18	18	18	19	19	19	19	20	20	21	21	22	22	22	22
23	25	25	25	25	25	25								

4.13 (Poisson change–point models). Carlin *et al.* (1992) analyzed British coal–mining disaster data (see Maguire *et al.* 1952 and corrected by Jarrett 1979) in Table 4.4 from 1851–1962 by using a hierarchical Poisson change–point model. At the first stage, they assumed that

$$\begin{aligned} Y_i &\sim \text{Poisson}(\theta_1 t_i), \quad i = 1, \ldots, k, \\ Y_i &\sim \text{Poisson}(\theta_2 t_i), \quad i = k+1, \ldots, n, \end{aligned}$$

where θ_1 and θ_2 are unknown parameters. At the second stage, they considered independent prior distributions: k is assumed to follow a discrete uniform distribution on $\{1, \ldots, n\}$,

$$\theta_j|b_j \sim \text{Gamma}(a_{j0}, b_j), \quad j = 1, 2.$$

At the third stage, they supposed that

$$b_j \sim \text{Gamma}(c_{j0}, d_{j0}), \quad j = 1, 2,$$

and b_1 is independent of b_2.

Table 4.4 British coal–mining disaster data (1851–1962)

Year	Count	Year	Count	Year	Count	Year	Count
1851	4	1879	3	1907	0	1935	4
1852	5	1880	4	1908	3	1936	1
1853	4	1881	2	1909	2	1937	1
1854	1	1982	5	1910	2	1938	1
1855	0	1883	2	1911	0	1939	1
1856	4	1884	2	1912	1	1940	2
1857	3	1885	3	1913	1	1941	4
1858	4	1886	4	1914	1	1942	2
1859	0	1887	2	1915	0	1943	0
1860	6	1888	1	1916	1	1944	0
1861	3	1889	3	1917	0	1945	0
1862	3	1890	2	1918	1	1946	1
1863	4	1891	2	1919	0	1947	4
1864	0	1892	1	1920	0	1948	0
1865	2	1893	1	1921	0	1949	0
1866	6	1894	1	1922	2	1950	0
1867	3	1895	1	1923	1	1951	1
1868	3	1896	3	1924	0	1952	0
1869	5	1897	0	1925	0	1953	0
1870	4	1898	0	1926	0	1954	0
1871	5	1899	1	1927	1	1955	0
1872	3	1900	0	1928	1	1956	0
1873	1	1901	1	1929	0	1957	1
1874	4	1902	1	1930	2	1958	0
1875	4	1903	0	1931	3	1959	0
1876	1	1904	0	1932	3	1960	1
1877	5	1905	3	1933	1	1961	0
1878	5	1906	1	1934	1	1962	1

Let $Y_{\text{obs}} = \{y_i\}_{i=1}^n$ and define

$$S_1 \hat{=} \sum_{i=1}^{k} y_i, \quad S_2 \hat{=} \sum_{i=k+1}^{n} y_i, \quad T_1 \hat{=} \sum_{i=1}^{k} t_i, \quad T_2 \hat{=} \sum_{i=k+1}^{n} t_i.$$

Prove that three full conditional distributions are given by

$$\begin{aligned} p(\theta_1, \theta_2 | Y_{\text{obs}}, k, b_1, b_2) &= \prod_{j=1}^{2} \text{Gamma}(\theta_j | a_{j0} + S_j,\ b_j + T_j), \\ p(b_1, b_2 | Y_{\text{obs}}, k, \theta_1, \theta_2) &= \prod_{j=1}^{2} \text{Gamma}(b_j | a_{j0} + c_{j0},\ \theta_j + d_{j0}), \\ p(k | Y_{\text{obs}}, \theta_1, \theta_2, b_1, b_2) &\propto \left(\frac{\theta_1}{\theta_2}\right)^{S_1} \exp\{(\theta_2 - \theta_1) T_1\}. \end{aligned}$$

For the data in Table 4.4, implement the Gibbs sampler.

Chapter 5

Bootstrap Methods

1• What is the bootstrap?

- The bootstrap is a data–based method for statistical inference.
- Its introduction into statistics is relatively recent because the method is computationally intensive.

2• The purpose of the bootstrap

- The bootstrap provides a general method for obtaining *estimated standard errors* of estimators and *confidence intervals* (CIs) of parameters.
- The bootstrap can be applied to *testing hypotheses* to calculate a bootstrap p-value or to provide an upper quantile point of the distribution of a test statistic when its density is not available in closed–form.

5.1 Bootstrap Confidence Intervals

5.1.1 Parametric bootstrap

3• Describing parametric bootstrap for computing ci

- Suppose that we have a method to calculate the point estimator $\hat{\theta}$, repeatedly computing the point estimator G times based on G bootstrap samples will result in the CI of θ.
- Thus, the key is how to generate a bootstrap sample.

4• Large–sample ci for one–sample problem

- Let $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, where $\theta = \Pr(X_1 = 1)$ is the unknown parameter of population mean. Let $\mathbf{x} = (X_1, \ldots, X_n)^\top$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.
- The likelihood function for θ is $L(\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$ for $\theta \in (0,1)$, so that the MLE of θ is the sample mean defined by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i = s(\mathbf{x}). \tag{5.1}$$

4.1• Central limit theorem

— Note that $n\hat{\theta} = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \theta)$, then $E(\hat{\theta}) = \theta$ and $\text{Var}(\hat{\theta}) = \theta(1-\theta)/n$.

— According to the central limit theorem, we have

$$\frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{\text{Var}(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/n}} \xrightarrow{\text{D}} Z \sim N(0,1).$$

— Namely, $\{\hat{\theta} - E(\hat{\theta})\}/\{\text{Var}(\hat{\theta})\}^{1/2}$ converges in distribution to a random variable following $N(0,1)$.

4.2• Limiting properties of MLE

— Based on limiting properties of MLE, we approximately have

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \dot{\sim} N(0,1) \quad \text{as} \quad n \to \infty.$$

— Let z_α be the upper α-th quantile of $N(0,1)$ satisfying $\Pr(Z \geqslant z_\alpha) = \alpha$.

— Therefore, an asymptotic $100(1-\alpha)\%$ CI of θ is given by

$$\begin{aligned} 1-\alpha &= \Pr\left(-z_{\alpha/2} \leqslant \frac{\hat{\theta} - \theta}{\hat{\sigma}} \leqslant z_{\alpha/2}\right) \\ &= \Pr\left(\hat{\theta} - z_{\alpha/2}\hat{\sigma} \leqslant \theta \leqslant \hat{\theta} + z_{\alpha/2}\hat{\sigma}\right); \end{aligned}$$

i.e.,

$$[\hat{\theta}_l,\ \hat{\theta}_u] = \left[\hat{\theta} - z_{\alpha/2}\hat{\sigma},\ \ \hat{\theta} + z_{\alpha/2}\hat{\sigma}\ \right], \tag{5.2}$$

where $\hat{\sigma} = \sqrt{\hat{\theta}(1-\hat{\theta})/n}$.

4.3• Two problems associated with the asymptotic CI (5.2)

— First, even though for large sample size n, the lower bound may be beyond zero when the true value of θ is close to zero; while the upper bound may be beyond 1 when the true value of θ is near to 1.

— Second, for small to moderate sample sizes, the asymptotic CI is not reliable.

5• Small-sample CI via bootstrap

- To overcome the two difficulties with the traditional large-sample method, we employ the bootstrap approach.
- The essential steps in the bootstrap approach for deriving the estimated standard error of $\hat{\theta}$ (hence, the *bootstrap CI* (BCI) of θ) are as follows.

Step 1: Calculate the point estimator $\hat{\theta} = s(\mathbf{x})$ like (5.1).

Step 2: Generate a bootstrap sample $\mathbf{x}^* = (X_1^*, \ldots, X_n^*)^\top$ with $\{X_i^*\}_{i=1}^n \overset{\text{iid}}{\sim} \text{Bernoulli}(\hat{\theta})$ and compute the corresponding bootstrap replication $\hat{\theta}^* = s(\mathbf{x}^*)$.

Step 3: Independently repeating this process (i.e., Step 2) G times, we obtain G bootstrap replications $\{\hat{\theta}^*(g)\}_{g=1}^G$.

Step 4: Consequently, the standard error, $\text{Se}(\hat{\theta})$, of $\hat{\theta}$ can be estimated by the sample standard deviation of the G replications; i.e.,

$$\widehat{\text{Se}}^*(\hat{\theta}) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}\{\hat{\theta}^*(g) - \bar{\theta}^*\}^2}\ , \tag{5.3}$$

where $\bar{\theta}^* = \{\hat{\theta}^*(1) + \cdots + \hat{\theta}^*(G)\}/G$.

Step 5: If $\{\hat{\theta}^*(g)\}_{g=1}^G$ are approximately normally distributed, a $100(1-\alpha)\%$ BCI for θ is

$$[\hat{\theta}_l^*,\ \hat{\theta}_u^*] = \left[\bar{\theta}^* - z_{\alpha/2}\cdot \widehat{\mathrm{Se}}^*(\hat{\theta}),\ \ \bar{\theta}^* + z_{\alpha/2}\cdot \widehat{\mathrm{Se}}^*(\hat{\theta})\right]. \quad (5.4)$$

Step 6: If the BCI (5.4) is beyond the unit interval $[0,1]$ or the bootstrap replications $\{\hat{\theta}^*(g)\}_{g=1}^G$ are non-normally distributed, a $100(1-\alpha)\%$ BCI for θ is

$$[\hat{\theta}_{\mathrm{L}}^*,\ \hat{\theta}_{\mathrm{U}}^*], \quad (5.5)$$

where $\hat{\theta}_{\mathrm{L}}^*$ and $\hat{\theta}_{\mathrm{U}}^*$ are the $(\alpha/2)G$-th and the $(1-\alpha/2)G$-th *order statistics* of $\{\hat{\theta}^*(g)\}_{g=1}^G$.

5.1• Sample order statistics

— For example, when $\alpha = 0.05$ and $G = 1000$, $\hat{\theta}_{\mathrm{L}}^*$ is the 25-th order statistic and $\hat{\theta}_{\mathrm{U}}^*$ is the 975-th order statistic of $\hat{\theta}^*(1),\ldots,\hat{\theta}^*(1000)$, respectively.

6• A REAL WORLD AND A PARAMETRIC BOOTSTRAP WORLD

Table 5.1 A comparison between a real world and a parametric bootstrap world for one–sample problem

	A real world	A parametric bootstrap world
1	Known distribution class with unknown parameter: $F(x,\theta)$	Estimated cdf: $F(x,\hat{\theta})$
2	One observed random sample: $\mathbf{x} = (X_1,\ldots,X_n)^\top \overset{\text{iid}}{\sim} F(x,\theta)$	G independent bootstrap samples: $\mathbf{x}^*(g) = (X_1^*(g),\ldots,X_n^*(g))^\top \overset{\text{iid}}{\sim} F(x,\hat{\theta})$ $g = 1,\ldots,G$
3	Estimator: $\hat{\theta} = s(\mathbf{x})$	G Bootstrap replications: $\hat{\theta}^*(g) = s(\mathbf{x}^*(g)),\ g = 1,\ldots,G$
4		Bootstrap sample mean: $\bar{\theta}^*$
5		Bootstrap sample std: $\widehat{\mathrm{Se}}^*(\hat{\theta})$
6		Normality-based BCI: $[\hat{\theta}_l^*,\ \hat{\theta}_u^*]$
7		non-normality-based BCI: $[\hat{\theta}_{\mathrm{L}}^*,\ \hat{\theta}_{\mathrm{U}}^*]$

NOTE: BCI = Bootstrap confidence interval.

7• Three examples

Example 5.1 (CIs of parameter for Bernoulli distribution). Let

$$1, 1, 0, 1, 1, 1, 1, 1, 1, 1 \tag{5.6}$$

be an observed sample of size $n = 10$ from Bernoulli distribution with mean parameter θ. From (5.1) and (5.2), the MLE and an asymptotic 95% CI of θ are given by $\hat{\theta} = 0.9$ and

$$\begin{aligned}[\hat{\theta}_l,\ \hat{\theta}_u] &= \left[0.9 - 1.96\sqrt{0.9(0.1)/10},\ 0.9 + 1.96\sqrt{0.9(0.1)/10}\ \right] \\ &= [0.71406,\ 1.0859],\end{aligned}$$

respectively. We noted that the upper limit of this asymptotic CI is larger than 1, resulting in a useless CI. Use the bootstrap approach to estimate $\mathrm{Se}(\hat{\theta})$ and to obtain two BCIs.

Solution for small sample: Note that $n = 10$ is very small, we generated $G = 20{,}000$ bootstrap samples and computed 20,000 bootstrap replications $\{\hat{\theta}^*(g)\}_{g=1}^{G}$. Numerical results corresponding to (5.3)–(5.5) are as follows:

$$\begin{aligned}\bar{\theta}^* &= 0.90027, \quad \widehat{\mathrm{Se}}^*(\hat{\theta}) = 0.09516, \\ [\hat{\theta}_l^*,\ \hat{\theta}_u^*] &= [0.71376,\ 1.0868], \\ [\hat{\theta}_{\mathrm{L}}^*,\ \hat{\theta}_{\mathrm{U}}^*] &= [0.7,\ 1].\end{aligned}$$

It is noted that the upper bound of the normality-based BCI is also beyond 1. Hence, only the non-normality-based BCI is available.

Solution for moderate sample: If the observed data (5.6) is replaced by

$$1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1,$$

then the corresponding results are given by

$$\begin{aligned}\hat{\theta} &= 0.60, \\ [\hat{\theta}_l,\ \hat{\theta}_u] &= [0.38529,\ 0.81471], \\ \bar{\theta}^* &= 0.60041, \\ \widehat{\mathrm{Se}}^*(\hat{\theta}) &= 0.10947, \\ [\hat{\theta}_l^*,\ \hat{\theta}_u^*] &= [0.38585,\ 0.81497], \\ [\hat{\theta}_{\mathrm{L}}^*,\ \hat{\theta}_{\mathrm{U}}^*] &= [0.40,\ 0.80].\end{aligned}$$

We can see that the second BCI has the shortest width among the three CIs.

R codes:

```
CS.Example5.1 <- function(ind, G) {
     # ================ Aim ================================
     # Conduct Example 5.1
     # Call: M <- mean.std.CI(thsample)
     # ================ Input ==============================
     # ind = 1, if the data set (5.6) is used
     # ind = 2, if the other data set is used
     # G   = bootstrap sample size, 20000 in this example
     # ================ Output =============================
     # thhat, thmean, thstd, thl, thu, thL, thU
     # =====================================================
     if(ind == 1) { x <- c(1, 1, 0, 1, 1, 1, 1, 1, 1, 1) }
     if(ind == 2) { x <- c(1, 1, 0, 1, 0, 0, 1, 1, 0, 0,
                           0, 1, 0, 1, 0, 1, 1, 1, 1, 1)
     }
     n <- length(x)
     thhat <- mean(x)
     th.star.sample <- matrix(0, G, 1)
     for(g in 1:G) {
          xstar <- rbinom(n, 1, thhat)
          thstar <- mean(xstar)
          th.star.sample[g] <- thstar
     }
     M <- mean.std.CI(th.star.sample)
     results <- c(thhat, M)
     return(results)
}

mean.std.CI <- function(thsample) {
     # ================ Aim ================================
     # Calculating the mean, std, two 95% CIs based on column
     # ================ Input ==============================
     # thsample = a matrix of G x c
     # ================ Output =============================
```

```
    # thmean, thstd, thl, thu, thL, thU
    # ======================================================
    G <- dim(thsample)[1]
    thmean <- apply(thsample, 2, mean)
    thstd <- sqrt(apply(thsample, 2, var))
    thl <- thmean - 1.96 * thstd
    thu <- thmean + 1.96 * thstd
    thsort <- apply(thsample, 2, sort)
    indexx <- floor(c(0.025 * G, 0.975 * G))
    thL <- (thsort[indexx[1], ]+ thsort[indexx[1]+1, ])/2
    thU <- (thsort[indexx[2], ]+ thsort[indexx[2]+1, ])/2
    results <- c(thmean, thstd, thl, thu, thL, thU)
    return(results)
}
```

‖

Example 5.2 (CI for linear combination of means in independent Poisson distributions). Let $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} \text{Poisson}(\theta_i)$, construct a $100(1-\alpha)\%$ CI for $\phi \triangleq \sum_{i=1}^n c_i\theta_i$, where $c_i > 0$, $i = 1, \ldots, n$.

Solution for a special case: We first consider a special case that all $c_i = 1$, for which, we can construct an exact CI for $\lambda = \sum_{i=1}^n \theta_i$. The pmf of X_i is

$$\text{Poisson}(x_i|\theta_i) = \frac{\theta_i^{x_i}}{x_i!}e^{-\theta_i}, \quad x_i = 0, 1, \ldots, \infty.$$

Note that the *moment generating function* (mgf) of X_i is

$$M_{X_i}(t) = E(\,\mathrm{e}^{tX_i}) = \exp\{(\,\mathrm{e}^t - 1)\theta_i\},$$

so that the mgf of $Y = \sum_{i=1}^n X_i$ is given by

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = \exp\{(\,\mathrm{e}^t - 1)\lambda\},$$

which implies $Y \sim \text{Poisson}(\lambda)$. Let y_0 be an observed value of Y and $[\hat{\lambda}_l, \hat{\lambda}_u]$ denote a $100(1-\alpha)\%$ CI of λ, then

$$\frac{\alpha}{2} = \Pr(Y \geqslant y_0|\lambda = \hat{\lambda}_l) = \sum_{y=y_0}^{\infty} \frac{(\hat{\lambda}_l)^y}{y!}\,\mathrm{e}^{-\hat{\lambda}_l}, \tag{5.7}$$

$$\frac{\alpha}{2} = \Pr(Y \leqslant y_0|\lambda = \hat{\lambda}_u) = \sum_{y=0}^{y_0} \frac{(\hat{\lambda}_u)^y}{y!}\,\mathrm{e}^{-\hat{\lambda}_u}. \tag{5.8}$$

Note that Poisson pmf and gamma density have the following relationship

$$\sum_{y=k}^{\infty} \text{Poisson}(y|\lambda) = \int_0^{\lambda} \text{Gamma}(z|k,1)\,\mathrm{d}z = \int_0^{2\lambda} \chi^2(t|2k)\,\mathrm{d}t,$$

then (5.7) becomes

$$\frac{\alpha}{2} = \sum_{y=y_0}^{\infty} \text{Poisson}(y|\hat{\lambda}_l) = \int_0^{2\hat{\lambda}_l} \chi^2(t|2y_0)\,\mathrm{d}t,$$

resulting in

$$\hat{\lambda}_l = \frac{\chi^2(1-\alpha/2, 2y_0)}{2},$$

where $\chi^2(\alpha, m)$ denotes the upper α-th quantile of the $\chi^2(m)$ distribution satisfying $\Pr\{\chi^2(m) \geqslant \chi^2(\alpha, m)\} = \alpha$. Similarly, from (5.8), we obtain

$$\hat{\lambda}_u = \frac{\chi^2(\alpha/2, 2y_0+2)}{2}.$$

Solution for the general case: For the general case, it is not possible to find an exact CI for ϕ by using the frequentist approach. However, the bootstrap method can be employed easily for this purpose. First, we calculate the MLE of ϕ:

$$\hat{\phi} = \sum_{i=1}^{n} c_i \hat{\theta}_i,$$

where $\hat{\theta}_i = X_i$ is the MLE of θ_i for $i = 1, \ldots, n$. Second, generate a bootstrap sample $\{X_i^*\}_{i=1}^n \overset{\text{ind}}{\sim} \text{Poisson}(\hat{\theta}_i)$ and compute the bootstrap replication $\hat{\phi}^* = \sum_{i=1}^n c_i X_i^*$. Independently repeating this process G times, we obtain a $100(1-\alpha)\%$ BCI for ϕ similar to (5.5). ‖

Example 5.3 (Two–parameter multinomial model revisited). In Example 2.7, we use the EM algorithm to obtain the MLEs of the parameter vector $\boldsymbol{\theta} \in \mathbb{T}_3$ given by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)^\top = (0.585900,\ 0.0716178,\ 0.342482)^\top$. On the other hand, in Example 2.9, we utilize the missing information principle to obtain the estimated standard errors given by $\widehat{\text{Se}}(\hat{\theta}_1) = 0.1417$, $\widehat{\text{Se}}(\hat{\theta}_2) = 0.0696$ and $\widehat{\text{Se}}(\hat{\theta}_3) = 0.1332$. Clearly, the lower bound of the asymptotic 95% CI of θ_2 is $0.0716178 - 1.96 \times 0.0696 = -0.064798 < 0$, leading to a useless result. Use the bootstrap approach to calculate a CI for each component of the $\boldsymbol{\theta}$.

Solution: First, we generate a bootstrap vector $Y^*_{\text{obs}} = (y^*_1, \ldots, y^*_5)^\top$ from

$$\text{Multinomial}(22; a_1\hat{\theta}_1 + b_1, a_2\hat{\theta}_1 + b_2, a_3\hat{\theta}_2 + b_3, a_4\hat{\theta}_2 + b_4, 0.5\hat{\theta}_3),$$

where $\{a_i, b_i\}_{i=1}^4$ are given by (2.26). Second, we use the EM algorithm (2.23) and (2.25) to compute the bootstrap replication $\hat{\boldsymbol{\theta}}^* = (\hat{\theta}^*_1, \hat{\theta}^*_2, \hat{\theta}^*_3)^\top$. Independently repeating this process $G = 10{,}000$ times, we obtain three bootstrap sample means, standard deviations, and 95% CIs as follows:

$$\begin{aligned}
\bar{\boldsymbol{\theta}}^* &= (\bar{\theta}^*_1, \bar{\theta}^*_2, \bar{\theta}^*_3)^\top = (0.58606, 0.07492, 0.33902)^\top, \\
\widehat{\text{Se}}^*(\hat{\theta}_1) &= 0.16726, \quad \widehat{\text{Se}}^*(\hat{\theta}_2) = 0.11140, \quad \widehat{\text{Se}}^*(\hat{\theta}_3) = 0.15298, \\
\left[\hat{\theta}^*_{1\text{L}},\ \hat{\theta}^*_{1\text{U}}\right] &= [0.2484,\ 0.90143], \quad \left[\hat{\theta}^*_{2\text{L}},\ \hat{\theta}^*_{2\text{U}}\right] = [0,\ 0.36445], \\
\left[\hat{\theta}^*_{3\text{L}},\ \hat{\theta}^*_{3\text{U}}\right] &= [0.0827,\ 0.65902].
\end{aligned}$$

R codes:

```
CS.Example5.3 <- function(EMn, G) {
     # ================= Aim ================================
     # Conduct Example 5.3
     # Call: ystar <- rmultinom(1, n, p)
     # Call: thMLE <- CS.Example5.3.MLE(ystar, EMn)
     # Call: M     <- mean.std.CI(thsample)
     # ================= Input ==============================
     # EMn = 20 is the iterative number needed for the EM
     # G   = 10000 is the bootstrap sample size
     # ================= Output =============================
     # thmean, thstd, thL, thU
     # ======================================================
     n <- 22
     a <- rep(0.25, 4)
     b <- c(1/8, 0, 0, 3/8)
     hth <- c(0.5859, 0.0716178, 0.342482)
     p <- c(a[1] * hth[1] + b[1], a[2] * hth[1] + b[2],
          a[3] * hth[2] + b[3], a[4] * hth[2] + b[4],
          0.5 * hth[3])
     th.star.sample <- matrix(0, G, 3)
     for(g in 1:G) {
```

```
          ystar <- rmultinom(1, n, p)
          thstar <- Example5.3.MLE(ystar, EMn)
          th.star.sample[g,  ] <- thstar
     }
     M <- mean.std.CI(th.star.sample)
     thmean <- M[1]
     thstd <- M[2]
     thL <- M[5]
     thU <- M[6]
     results <- c(thmean, thstd, thL, thU)
     return(results)
}


CS.Example5.3.MLE <- function(y, EMn) {
     # ================ Aim ================================
     # Calculate MLEs of \theta in Example 5.3
     # ================ Input ==============================
     # y   = data vector in Example 2.7
     # EMn = 10 is the iterative number needed for the EM
     # ================ Output =============================
     # thMLE = (theta_1, theta_2, theta_3)
     # =====================================================
     a <- rep(0.25, 4)
     b <- c(1/8, 0, 0, 3/8)
     th1 <- th2 <- 1/3
     for(tt in 1:EMn) {
          p1 <- (a[1] * th1)/(a[1] * th1 + b[1])
          p2 <- (a[2] * th1)/(a[2] * th1 + b[2])
          p3 <- (a[3] * th2)/(a[3] * th2 + b[3])
          p4 <- (a[4] * th2)/(a[4] * th2 + b[4])
          z1 <- y[1] * p1
          z2 <- y[2] * p2
          z3 <- y[3] * p3
          z4 <- y[4] * p4
          De <- z1 + z2 + z3 + z4 + y[5]
          th1 <- (z1 + z2)/De
          th2 <- (z3 + z4)/De
```

```
    }
    thMLE <- c(th1, th2, 1 - th1 - th2)
    return(thMLE)
}
```

5.1.2 Non-parametric bootstrap

8• Why do we need the non-parametric bootstrap?

- In §5.1.1, the BCIs are based on known distributions.
- However, in many real applications, the true distribution is unknown.
- It is desirable to use a non-parametric bootstrap method to obtain the estimated standard error of an estimator (e.g., the *least square estimator* (LSE)) or the BCI for a population parameter (e.g., the mean of population distribution).

9• Describing non-parametric bootstrap for computing ci

- Suppose that we have a method to calculate the point estimator $\hat{\theta}$, repeatedly computing the point estimator G times based on G bootstrap samples will result in the CI of θ.
- The key is how to generate a bootstrap sample from the empirical cdf.

10• The point estimator for one–sample problem

- Assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x)$, where $F(\cdot)$ is an unknown distribution function, and the quantity of interest is $\theta = T(F)$, expressed as a function of F.
- For example,
$$\theta = T(F) = \int x \, \mathrm{d}F(x)$$
is the mean of population distribution.

10.1• Objective of the non-parametric bootstrap

— The objective of the non-parametric bootstrap is first to estimate the cdf F and then to obtain a BCI for θ.

— The empirical distribution function based on the observations $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ is defined by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leqslant x), \tag{5.9}$$

where we assume $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$.

— Based on the function $\hat{F}_n$, one can estimate θ by $\hat{\theta} = T(\hat{F}_n) = s(\mathbf{x})$.

— For example, when θ is a univariate population mean, the estimator is the sample mean,

$$\hat{\theta} = T(\hat{F}_n) = \int x \, \mathrm{d}\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i = s(\mathbf{x}). \tag{5.10}$$

11• Bootstrap confidence interval

- The essential steps in the non-parametric bootstrap approach for obtaining the estimated standard error of $\hat{\theta}$ (hence, the BCI of θ) are as follows.

  Step 1: Calculate the point estimator $\hat{\theta} = s(\mathbf{x})$ like (5.10).

  Step 2: Generate a bootstrap sample $\mathbf{x}^* = (X_1^*, \ldots, X_n^*)^\top$ with

  $$\{X_i^*\}_{i=1}^n \overset{\text{iid}}{\sim} \hat{F}_n(x)$$

  and compute the corresponding bootstrap replication $\hat{\theta}^* = s(\mathbf{x}^*)$.

  Step 3: Independently repeating this process (i.e., Step 2) G times, we obtain G bootstrap replications $\{\hat{\theta}^*(g)\}_{g=1}^G$.

  Step 4: Consequently, the standard error, $\mathrm{Se}(\hat{\theta})$, of $\hat{\theta}$ can be estimated by the sample standard deviation of the G replications, i.e.,

  $$\widehat{\mathrm{Se}}^*(\hat{\theta}) = \sqrt{\frac{1}{G-1} \sum_{g=1}^{G} \{\hat{\theta}^*(g) - \bar{\theta}^*\}^2},$$

  where $\bar{\theta}^* = \{\hat{\theta}^*(1) + \cdots + \hat{\theta}^*(G)\}/G$.

Step 5: If $\{\hat{\theta}^*(g)\}_{g=1}^G$ are approximately normally distributed, a $100(1-\alpha)\%$ BCI for θ is

$$[\hat{\theta}_l^*,\ \hat{\theta}_u^*] = [\bar{\theta}^* - z_{\alpha/2} \cdot \widehat{\mathrm{Se}}^*(\hat{\theta}),\ \bar{\theta}^* + z_{\alpha/2} \cdot \widehat{\mathrm{Se}}^*(\hat{\theta})].$$

Step 6: If the bootstrap replications $\{\hat{\theta}^*(g)\}_{g=1}^G$ are non-normally distributed, a $100(1-\alpha)\%$ BCI for θ is

$$[\hat{\theta}_{\mathrm{L}}^*,\ \hat{\theta}_{\mathrm{U}}^*],$$

where $\hat{\theta}_{\mathrm{L}}^*$ and $\hat{\theta}_{\mathrm{U}}^*$ are the $(\alpha/2)G$-th and the $(1-\alpha/2)G$-th order statistics of $\{\hat{\theta}^*(g)\}_{g=1}^G$.

12• A real world and a non-parametric bootstrap world

Table 5.2 Comparison between a real world and a non-parametric bootstrap world for one–sample problem

	A real world	A non-parametric bootstrap world
1	Unknown distribution: F	Empirical distribution: $\hat{F}_n$
2	One observed random sample: $\mathbf{x} = (X_1, \ldots, X_n)^\top \overset{\text{iid}}{\sim} F$	G independent bootstrap samples: $\mathbf{x}^*(g) = (X_1^*(g), \ldots, X_n^*(g))^\top \overset{\text{iid}}{\sim} \hat{F}_n$, $g = 1, \ldots, G$
3	Statistic of interest: $\hat{\theta} = s(\mathbf{x})$	G Bootstrap replications: $\hat{\theta}^*(g) = s(\mathbf{x}^*(g)),\ g = 1, \ldots, G$
4		Bootstrap sample mean: $\bar{\theta}^*$
5		Bootstrap sample std: $\widehat{\mathrm{Se}}^*(\hat{\theta})$
6		Normality-based BCI: $[\hat{\theta}_l^*,\ \hat{\theta}_u^*]$
7		non-normality-based BCI: $[\hat{\theta}_{\mathrm{L}}^*,\ \hat{\theta}_{\mathrm{U}}^*]$

NOTE: BCI = Bootstrap confidence interval.

13• Generating independent bootstrap samples from $\hat{F}_n$

- The key step for the non-parametric bootstrap is to generate independent bootstrap samples from $\hat{F}_n$ defined by (5.9).
- Let $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)^\top$ be an i.i.d. bootstrap sample from $\hat{F}_n$.
- In fact, $\boldsymbol{x}^*$ is a random sample drawn *with replacement* from $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.

- Thus, we might have $x_1^* = x_7$, $x_2^* = x_3$, $x_3^* = x_3$, $x_4^* = x_{22}, \ldots, x_n^* = x_7$.
- The bootstrap sample $x_1^*, \ldots, x_n^*$ consists of members of the original observations $x_1, \ldots, x_n$, some appearing zero times, some appearing once, some appearing twice, etc.

13.1• Theoretical illustration

— Theoretically, the inversion method in Chapter 1 can be used to sample from $\hat{F}_n$ for obtaining $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)^\top$ as follows:

Step 1: Sort $x_1, \ldots, x_n$ in ascending order.

Step 2: Construct the empirical cdf $\hat{F}_n(x)$ according to (5.9).

Step 3: At the i-th iteration, generate a uniform random number $u \sim U(0, 1)$.

Step 4: Accept $x_i^* = x_1$ if $u \leqslant \hat{F}_n(x_1) = 1/n$. Otherwise, $x_i^* = x_j$ if $\hat{F}_n(x_{j-1}) = u \leqslant \hat{F}_n(x_j)$ or $(j-1)/n < u \leqslant j/n$ simply.

Step 5: Repeat Steps 3 and 4 for $i \leftarrow i+1$ until $x_1^*, \ldots, x_n^*$ are obtained.

13.2• Practical illustration

— In practice, note that each x_i^* equals any one of the n values x_i with probability $1/n$, we can use the R function,

$$\mathsf{sample}(\boldsymbol{x},\ \mathsf{N},\ \mathsf{prob} = \mathsf{rep}(1/\mathsf{n},\ \mathsf{n}),\ \mathsf{replace} = \mathsf{T}),$$

to produce a vector of length N randomly chosen from $\{x_1, \ldots, x_n\}$ with equal probabilities $\{1/n, \ldots, 1/n\}$ with replacement.

— For example,

```
============================================================
> sample(c(2,1,6), 1, prob = rep(1/3, 3), replace = T)
[1] 2
> sample(c(2,1,6), 2, prob = rep(1/3, 3), replace = T)
[1] 2 6
```

```
> sample(c(2,1,6), 3, prob = rep(1/3, 3), replace = T)
[1] 2 1 6
> sample(c(2,1,6), 4, prob = rep(1/3, 3), replace = T)
[1] 2 1 6 1
> sample(c(2,1,6), 5, prob = rep(1/3, 3), replace = T)
[1] 2 2 2 2 6
> sample(c(2,1,6), 6, prob = rep(1/3, 3), replace = T)
[1] 1 1 1 2 2 1
> sample(c(2,1,6), 10, prob = rep(1/3, 3), replace = T)
[1] 1 1 2 2 1 6 2 1 1 6
> sample(c(2,1,6), 12, prob = rep(1/3, 3), replace = T)
[1] 6 6 2 1 2 1 1 1 6 1 1 1
> sample(c(2,1,6), 15, prob = rep(1/3, 3), replace = T)
[1] 2 2 1 1 6 6 6 6 1 2 1 6 6 6 1
************************************************************
```

14• Two examples

Example 5.4 (An illustration). Suppose $n = 3$ univariate data points, say $\boldsymbol{x} = (1, 2, 6)^\top$, are observed as an i.i.d. sample from a cdf F with mean θ. The point estimate from the observed data is $\hat{\theta} = 9/3$ according to (5.10). Use the non-parametric bootstrap approach to estimate $\mathrm{Se}(\hat{\theta})$ and to obtain two BCIs.

Solution: To obtain the BCI of θ, firstly, we generate a bootstrap sample $\boldsymbol{x}^* = (x_1^*, x_2^*, x_3^*)^\top$ from $\hat{F}_3(x)$ and compute the bootstrap replication $\hat{\theta}^* = (x_1^* + x_2^* + x_3^*)/3$. Independently repeating this process $G = 10{,}000$ times, we obtain the bootstrap sample mean, standard deviations, and two 95% CIs:

$$\bar{\theta}^* = 2.9852, \qquad \widehat{\mathrm{Se}}^*(\hat{\theta}) = 1.2414,$$
$$\left[\hat{\theta}_l^*,\ \hat{\theta}_u^*\right] = [0.5521,\ 5.4183], \quad \left[\hat{\theta}_{\mathrm{L}}^*,\ \hat{\theta}_{\mathrm{U}}^*\right] = [1,\ 6].$$

R codes:

```
CS.Example5.4 <- function(G) {
    # ================ Aim ================================
    # Conduct Example 5.4
    # Call: M <- mean.std.CI(thsample)
```

```
     # ================ Input ==============================
     # G = 10000 is the bootstrap sample size
     # ================ Output =============================
     # thmean, thstd, thl, thu, thL, thU
     # =====================================================
     x <- c(1, 2, 6)
     n <- length(x)
     hth <- mean(x)
     p <- rep(1/3, 3)
     th.star.sample <- matrix(0, G, 1)
     for(g in 1:G) {
          xstar <- sample(x, n, prob = p, replace = T)
          thstar <- mean(xstar)
          th.star.sample[g, 1] <- thstar
     }
     M <- mean.std.CI(th.star.sample)
     return(M)
}
```

‖

Example 5.5 (Estimated standard errors of LSE). Consider the following linear regression model

$$Y_i = \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where Y_i is the response variable for subject i, $\boldsymbol{x}_{(i)}^\top$ denotes the i-th row of the covariate matrix $\boldsymbol{X}_{n\times q}$, $\boldsymbol{\theta}_{q\times 1}$ is the unknown parameter vector, and the error terms $\{\varepsilon_i\}_{i=1}^n$ are assumed to be i.i.d. from an unknown distribution $F(\cdot)$ with mean zero. The LSE of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||^2 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ are realizations of $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$. Use the nonparametric bootstrap approach to estimate $\mathrm{Se}(\hat{\theta}_j)$ for $j = 1, \ldots, q$.

Solution: Since $\hat{\boldsymbol{\theta}}$ is available, we can calculate

$$\hat{\varepsilon}_i = y_i - \boldsymbol{x}_{(i)}^\top \hat{\boldsymbol{\theta}}$$

for each i. The obvious estimate of $F(\cdot)$ is the empirical distribution of $\{\hat{\varepsilon}_i\}_{i=1}^n$, denoted by $\hat{F}_n(\cdot)$ with probability $1/n$ on $\hat{\varepsilon}_i$. Thus, we can generate a random sample of bootstrap error terms, denoted by $\{\varepsilon_i^*\}_{i=1}^n$, where

each ε_i^* equals any one of the n values $\hat{\varepsilon}_i$ with probability $1/n$. Then, the bootstrap responses are generated by $y_i^* = \boldsymbol{x}_{(i)}^\top \hat{\boldsymbol{\theta}} + \varepsilon_i^*$, $i = 1, \ldots, n$, or equivalently,

$$\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*,$$

where $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^\top$ and $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)^\top$. Notice that the covariate matrix $\boldsymbol{X}$ are the same for the bootstrap data as for the actual data, and $\hat{\boldsymbol{\theta}}$ is a fixed quantity in (5.10).

Having obtained $\boldsymbol{y}^*$, the bootstrap replication is given by

$$\hat{\boldsymbol{\theta}}^* = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\theta}||^2 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}^*.$$

Independently repeating the above process G times, we obtain G bootstrap replications $\{\hat{\boldsymbol{\theta}}^*(g)\}_{g=1}^G$ with $\hat{\boldsymbol{\theta}}^*(g) = (\hat{\theta}_1^*(g), \ldots, \hat{\theta}_q^*(g))^\top$. Thus, the standard error $\mathrm{Se}(\hat{\theta}_j)$ of $\hat{\theta}_j$ can be estimated by the sample standard deviation of the G replications; i.e.,

$$\widehat{\mathrm{Se}}^*(\hat{\theta}_j) = \left[\sum_{g=1}^{G} \left\{ \hat{\theta}_j^*(g) - (1/G) \textstyle\sum_{g=1}^G \hat{\theta}_j^*(g) \right\}^2 \Big/ (G-1) \right]^{1/2},$$

for $j = 1, \ldots, q$. ‖

5.2 Hypothesis Testing with the Bootstrap

5.2.1 Testing equality of two unknown distributions

15• The two–sample problem

- Assume that we have observed two independent random samples $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$ drawn from possibly different cdfs $F(\cdot)$ and $G(\cdot)$.
- We wish to test the null hypothesis

$$H_0\colon\ F(\cdot) = G(\cdot) \quad \text{against} \quad H_1\colon\ F(\cdot) \neq G(\cdot). \tag{5.11}$$

15.1• Achieved significance level

— Let $\boldsymbol{x} = \{\boldsymbol{z}, \boldsymbol{y}\}$ denote the combined sample.

— A bootstrap hypothesis test is based on a test statistic, denoted by

$$t(\mathbf{x}) = \bar{Z} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{m}\sum_{j=1}^{m} Y_j,$$

where $\mathbf{x} = \{\mathbf{z}, \mathbf{y}\} = (Z_1, \ldots, Z_n, Y_1, \ldots, Y_m)^\top$. The observed value of $t(\mathbf{x})$ is denoted by

$$t_{\text{obs}} = t(\boldsymbol{x}) = \bar{z} - \bar{y} = \frac{1}{n}\sum_{i=1}^{n} z_i - \frac{1}{m}\sum_{j=1}^{m} y_j.$$

— We seek an *achieved significance level* (ASL) of the test. Sometimes, the ASL is also called the bootstrap p-value.

— The ASL is defined by

$$\text{ASL} = \Pr\{t(\mathbf{x}^*) \geqslant t_{\text{obs}} | H_0\},$$

where $\mathbf{x}^* = (X_1^*, \ldots, X_{n+m}^*)^\top$, $X_1^*, \ldots, X_{n+m}^*$ is a random sample from a population random variable X^* with cdf $F_0(\cdot)$.

— Now, the question is, what is $F_0(\cdot)$?

15.2• What is $F_0(\cdot)$?

— In bootstrap hypothesis testing, let the empirical distribution based on $\boldsymbol{x}$ be $\hat{F}_0(\cdot)$, putting probability $1/(n+m)$ on each component of $\boldsymbol{x}$.

— Under H_0, $\hat{F}_0(\cdot)$ provides a non-parametric estimate of the common population that gave rise to both $\boldsymbol{z}$ and $\boldsymbol{y}$.

15.3• Estimating the ASL

— Computing the bootstrap test statistic for testing H_0 in (5.11) is as follows:

Step 1: Generate G bootstrap samples $\{\boldsymbol{x}^*(g)\}_{g=1}^G$ of size $n+m$ *with replacement* from $\mathbf{x} = \{\mathbf{z}, \mathbf{y}\}$, where

$$\boldsymbol{x}^*(g) = \{\boldsymbol{z}^*(g), \boldsymbol{y}^*(g)\} = (z_1^*(g), \ldots, z_n^*(g), y_1^*(g), \ldots, y_m^*(g))^\top.$$

Step 2: For each bootstrap sample, calculate

$$t(\boldsymbol{x}^*(g)) = \bar{z}^*(g) - \bar{y}^*(g) = \frac{1}{n}\sum_{i=1}^{n} z_i^*(g) - \frac{1}{m}\sum_{j=1}^{m} y_j^*(g),$$

where $g = 1, \ldots, G$.

Step 3: Estimate ASL by

$$\begin{aligned} \widehat{\text{ASL}} &= \frac{\#\{t(\boldsymbol{x}^*(g)) \geqslant t_{\text{obs}}\}}{G} \\ &= \frac{1}{G}\sum_{g=1}^{G} I\{t(\boldsymbol{x}^*(g)) \geqslant t_{\text{obs}}\}. \end{aligned}$$

15.4• Comments on the ASL

— The smaller the value of ASL, the stronger the evidence against H_0.

— The hypothesis test of H_0 consists of computing ASL, and seeing if it is too small according to certain conventional thresholds.

— Formally, we choose a small probability α, like 0.05 or 0.01, and reject H_0 if ASL $< \alpha$.

— Otherwise, we cannot reject H_0, which amounts to saying that the experimental data do not decisively reject H_0.

— Less formally, we observe ASL and rate the evidence against H_0 according to the following rough conventions:

ASL < 0.100	borderline evidence against H_0,
ASL < 0.050	reasonably strong evidence against H_0,
ASL < 0.025	strong evidence against H_0,
ASL < 0.010	very strong evidence against H_0.

16• One example

Example 5.6 (The mouse data). Table 5.3 shows the results of a small experiment, in which 7 out of the 16 mice were randomly selected to receive a new medical treatment, while the remaining 9 mice were assigned to the

control group. The treatment was intended to prolong survival after a test surgery. The table shows the survival time following surgery, in days, for all 16 mice. Did the treatment prolong survival?

Table 5.3 The mouse data

Group	Survival time in days	Sample size
Treatment	94, 197, 16, 38, 99, 141, 23	7
Control	52, 104, 146, 10, 51, 30, 40, 27, 46	9

Solution: In this example, $t(\boldsymbol{x}) = \bar{z} - \bar{y} = 86.857 - 56.222 = 30.635$. Twenty thousand bootstrap samples were generated, and 2544 had $t(\boldsymbol{x}^*) > 30.63$. The estimated value of ASL is $2544/20000 = 0.1272$ so that we cannot reject H_0: $F(\cdot) = G(\cdot)$. Figure 5.1 shows a histogram of bootstrap replications of $\bar{z} - \bar{y}$ for testing H_0: $F(\cdot) = G(\cdot)$.

R codes:

```
CS.Example5.6 <- function(G) {
     # ================= Aim ================================
     # Conduct Example 5.6
     # ================= Input ==============================
     # G = 20000 is the bootstrap sample size
     # ================= Output =============================
     # ASL and Figure 5.1
     # ======================================================
     z <- c(94, 197, 16, 38, 99, 141, 23)
     n <- length(z)
     y <- c(52, 104, 146, 10, 51, 30, 40, 27, 46)
     m <- length(y)
     x <- c(z, y)
     tx <- mean(z) - mean(y)
     N <- n + m
     p <- rep(1/N, N)
     t.star.sample <- rep(0, G)
     for(g in 1:G) {
          xstar <- sample(x, N, prob = p, replace = T)
          thstar <- mean(xstar[1:n]) - mean(xstar[(n+1):N])
          t.star.sample[g] <- thstar
```

```
    }
    ASL <- length(t.star.sample[t.star.sample >= tx])/G
    #--------------------- Figure 5.1 --------------------
    par(pty = "s")
    hist(t.star.sample, probability = T, plot = T,
         include.lowest = T, main = "H_0: F=G", xlab = " ",
         ylab = " ")
    return(ASL)
}
```

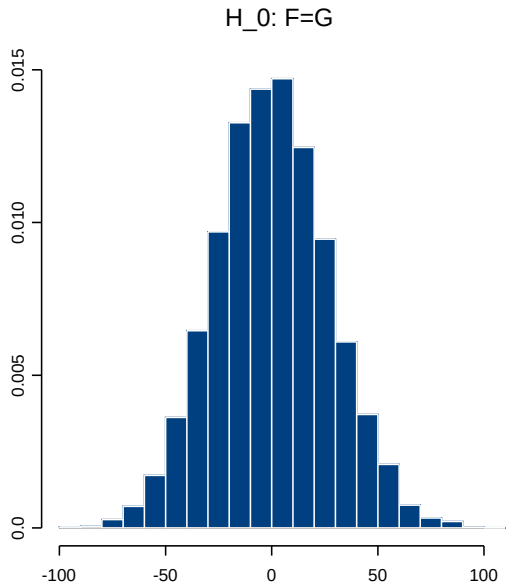

Figure 5.1 A histogram of bootstrap replications of $\bar{z} - \bar{y}$ for testing H_0: $F(\cdot) = G(\cdot)$ for the mouse data. ‖

5.2.2 Testing equality of two group means

17• A NEW ISSUE

- In (5.11), we tested that two cdfs are equal.
- What could happen if we would like to test only whether their means were equal?

18• Two–sample t test with equal variances

- Consider the following hypotheses H_0: $\mu_z = \mu_y$ against H_1: $\mu_z > \mu_y$.
- The two–sample t statistic and its realization are defined by

$$t(\mathbf{x}) = \frac{\bar{Z} - \bar{Y}}{S\sqrt{1/n + 1/m}} \quad \text{and} \quad t(\boldsymbol{x}) = \frac{\bar{z} - \bar{y}}{s\sqrt{1/n + 1/m}}, \tag{5.12}$$

  respectively, where

$$s^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})^2 + \sum_{j=1}^{m}(y_j - \bar{y})^2}{n + m - 2}.$$

- Assuming normal populations with equal variances, under H_0, we have $t(\mathbf{x}) \sim t(n + m - 2)$.

19• Motivation from two-sample t test with unequal variances

- When $\sigma_1^2 \neq \sigma_2^2$, the test statistic and its realization:

$$t(\mathbf{x}) = \frac{\bar{Z} - \bar{Y}}{\sqrt{S_1^2/n + S_2^2/m}} \quad \text{and} \quad t(\boldsymbol{x}) = \frac{\bar{z} - \bar{y}}{\sqrt{s_1^2/n + s_2^2/m}} \tag{5.13}$$

  can be used, where

$$s_1^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})^2}{n - 1} \quad \text{and} \quad s_2^2 = \frac{\sum_{j=1}^{m}(y_j - \bar{y})^2}{m - 1}.$$

- Even for the normal populations, the $t(\mathbf{x})$ in (5.13) no longer follows a t-distribution and a number of approximate solutions have therefore been proposed.
- In the literature, this is known as the Behrens–Fisher problem.

20• The bootstrap test

- The assumption of equal variance is a key for deriving the t-distribution for the t statistic $t(\mathbf{x})$ in (5.12).
- In the bootstrap test, it is not necessary to make this assumption.

- To proceed, we need estimates of $F(\cdot)$ and $G(\cdot)$ that use only the assumption of a common mean.
- Letting $\bar{x}$ be the mean of the combined sample, we can translate both samples so that they have mean $\bar{x}$, and then resample each population separately.

20.1• The bootstrap procedure

— Computation of the bootstrap test statistic for testing equality of two means is as follows:

Step 1: Let $\hat{F}(\cdot)$ put equal probability on the points $\tilde{z}_i = z_i - \bar{z} + \bar{x}$ for $i = 1, \ldots, n$, and $\hat{G}(\cdot)$ put equal probability on the points $\tilde{y}_j = y_j - \bar{y} + \bar{x}$ for $j = 1, \ldots, m$, where $\bar{z}$ and $\bar{y}$ are the group means and $\bar{x}$ is the mean of the combined sample.

Step 2: Form G bootstrap data sets $\boldsymbol{x}^*(g) = \{\boldsymbol{z}^*(g), \boldsymbol{y}^*(g)\}$, where

$$\boldsymbol{z}^*(g) = (z_1^*(g), \ldots, z_n^*(g))^\top$$

is sampled with replacement from $\tilde{z}_1, \ldots, \tilde{z}_n$ and

$$\boldsymbol{y}^*(g) = (y_1^*(g), \ldots, y_m^*(g))^\top$$

is sampled with replacement from $\tilde{y}_1, \ldots, \tilde{y}_m$.

Step 3: Evaluate $t(\cdot)$ defined by (5.13) on each data set,

$$t(\boldsymbol{x}^*(g)) = \frac{\bar{z}^*(g) - \bar{y}^*(g)}{\sqrt{s_1^{*2}(g)/n + s_2^{*2}(g)/m}},$$

where $g = 1, \ldots, G$,

$$\begin{aligned}
\bar{z}^*(g) &= \frac{\sum_{i=1}^n z_i^*(g)}{n}, \\
\bar{y}^*(g) &= \frac{\sum_{j=1}^m y_j^*(g)}{m}, \\
s_1^{*2}(g) &= \frac{\sum_{i=1}^n (z_i^*(g) - \bar{z}^*(g))^2}{n-1}, \quad \text{and} \\
s_2^{2*}(g) &= \frac{\sum_{j=1}^m (y_j^*(g) - \bar{y}^*(g))^2}{m-1}.
\end{aligned}$$

Step 4: Estimate ASL by

$$\widehat{\text{ASL}} = \frac{\#\{t(\mathbf{x}^*(g)) \geqslant t_{\text{obs}}\}}{G} = \frac{\sum_{g=1}^{G} I\{t(\mathbf{x}^*(g)) \geqslant t_{\text{obs}}\}}{G},$$

where $t_{\text{obs}} = t(\boldsymbol{x})$ is the observed value of the statistic $t(\mathbf{x})$.

21• One example

Example 5.7 (The mouse data revisited). Assume that we only consider to test H_0: $\mu_z = \mu_y$ against H_1: $\mu_z > \mu_y$, where μ_z denotes the average survival time after a test surgery in the treatment group, while μ_y denotes the average survival time in the control group.

Solution: From (5.13), $t(\boldsymbol{x}) = 1.0587$. Twenty thousand bootstrap samples were generated, and 2896 had $t(\boldsymbol{x}^*) > 1.0587$. The estimated value of ASL is $2896/20000 = 0.1448$ so that we cannot reject H_0. Figure 5.2 shows a histogram of bootstrap replications of $t(\mathbf{x})$ defined in (5.13) for the test of H_0: $\mu_z = \mu_y$.

R codes:

```
CS.Example5.7 <- function(G) {
    # ================= Aim ================================
    # Conduct Example 5.7
    # ================= Input ==============================
    # G = 20000 is the bootstrap sample size
    # ================= Output =============================
    # tx, ASL and Figure 5.2
    # ======================================================
    z <- c(94, 197, 16, 38, 99, 141, 23)
    n <- length(z)
    y <- c(52, 104, 146, 10, 51, 30, 40, 27, 46)
    m <- length(y)
    x <- c(z, y)
    zwave <- z - mean(z) + mean(x)
    ywave <- y - mean(y) + mean(x)
    tx <- (mean(z) - mean(y))/sqrt(var(z)/n + var(y)/m)
    p1 <- rep(1/n, n)
```

```
    p2 <- rep(1/m, m)
    t.star.sample <- rep(0, G)
    for(g in 1:G) {
        zstar <- sample(zwave, n, prob = p1, replace = T)
        ystar <- sample(ywave, m, prob = p2, replace = T)
        txstar <- (mean(zstar) - mean(ystar))/
                  sqrt(var(zstar)/n + var(ystar)/m)
        t.star.sample[g] <- txstar
    }
    ASL <- length(t.star.sample[t.star.sample >= tx])/G
    #---------------------- Figure 5.2 ----------------------
    par(pty = "s")
    hist(t.star.sample, probability = T, plot = T,
        include.lowest = T, main = "H_0: Equality of two
        means", xlab = " ", ylab = " ")
    return(tx, ASL)
}
```

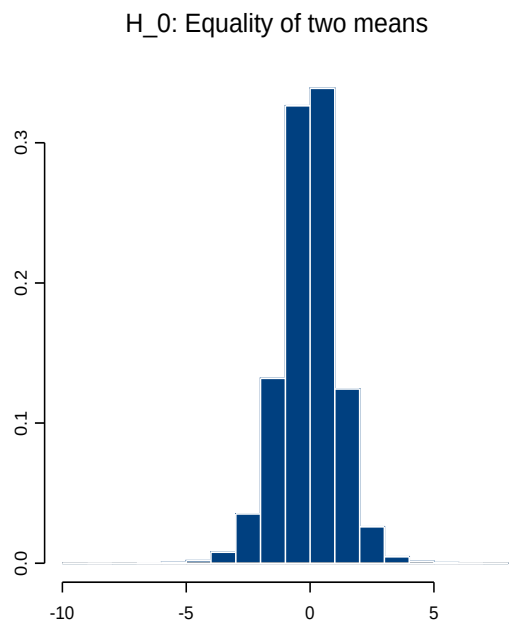

Figure 5.2 A histogram of bootstrap replications of $t(\boldsymbol{x})$ defined in (5.13) for testing H_0: $\mu_z = \mu_y$ against H_1: $\mu_z > \mu_y$ for the mouse data. ‖

5.2.3 One–sample problem

22• The normal test with known variance

- Consider the following hypotheses H_0: $\mu_z = \mu_0$ against H_1: $\mu_z < \mu_0$.
- Assuming a normal population and a known variance, the test statistic and its realization are

$$Z = \frac{\bar{Z} - \mu_0}{\sigma/\sqrt{n}} \quad \text{and} \quad z = \frac{\bar{z} - \mu_0}{\sigma/\sqrt{n}},$$

  respectively.
- Under H_0, $Z \sim N(0,1)$. Then the ASL or the p-value is given by

$$\text{ASL} = p\text{-value} = \Phi\left(\frac{\bar{z} - \mu_0}{\sigma/\sqrt{n}}\right). \tag{5.14}$$

23• The t-test with unknown variance

- When σ^2 is unknown, the t-test gives

$$\text{ASL} = p\text{-value} = \Pr\left\{t(n-1) < \frac{\bar{z} - \mu_0}{s/\sqrt{n}}\right\}, \tag{5.15}$$

  where

$$s^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})^2}{n-1}.$$

24• The bootstrap test

- In the bootstrap hypothesis test, it is not necessary to make the assumption of normality.
- We base the bootstrap hypothesis test on the distribution of the test statistic

$$t(\mathbf{z}) = \frac{\bar{Z} - \mu_0}{S/\sqrt{n}} \tag{5.16}$$

  under the null hypothesis H_0.
- But what is the appropriate null distribution?
- We need a distribution $F(\cdot)$ that estimates, e.g., the population of treatment times, *under* H_0.

- Note first that the empirical distribution $\hat{F}(\cdot)$ is not an appropriate estimate for $F(\cdot)$ because it *does not obey* H_0.
- That is, the mean of $\hat{F}(\cdot)$ is not equal to μ_0.
- Somehow we need to obtain an estimate of the population that has mean μ_0.
- A simple way is to translate the empirical distribution $\hat{F}(\cdot)$ so that it has the desired mean.
- In other words, we use the empirical distribution $\hat{F}(\cdot)$ on the values

$$\tilde{z}_i = z_i - \bar{z} + \mu_0, \quad i = 1, \ldots, n,$$

as the estimated null distribution.

24.1• The bootstrap procedure

— The bootstrap test approach can be summarized as follows:

Step 1: Let $\hat{F}(\cdot)$ put equal probability on the points $\tilde{z}_i = z_i - \bar{z} + \mu_0$, $i = 1, \ldots, n$.

Step 2: Form G bootstrap data sets $\boldsymbol{z}^*(g) = (z_1^*(g), \ldots, z_n^*(g))^\top$ by sampling with replacement from $\tilde{z}_1, \ldots, \tilde{z}_n$.

Step 3: Compute the statistic $t(\cdot)$ defined by (5.16) on each data set,

$$t(\boldsymbol{z}^*(g)) = \frac{\bar{z}^*(g) - \mu_0}{s^*(g)/\sqrt{n}},$$

where $g = 1, \ldots, G$,

$$\begin{aligned} \bar{z}^*(g) &= \frac{\sum_{i=1}^n z_i^*(g)}{n} \quad \text{and} \\ s^{*2}(g) &= \frac{\sum_{i=1}^n (z_i^*(g) - \bar{z}^*(g))^2}{n-1}. \end{aligned}$$

Step 4: Estimate ASL by

$$\widehat{\text{ASL}} = \frac{\#\{t(\mathbf{z}^*(g)) < t_{\text{obs}}\}}{G} = \frac{\sum_{g=1}^G I\{t(\mathbf{z}^*(g)) < t_{\text{obs}}\}}{G},$$

where $t_{\text{obs}} = t(\boldsymbol{z})$ is the observed value of the statistic $t(\mathbf{z})$.

25• One example

Example 5.8 (The mouse data revisited). Consider only the treated mice data in Table 5.3. Suppose that other investigators have run experiments similar to ours but with many more mice, and they observed a mean lifetime of 129.0 days for treated mice. We might want to test whether the mean of the treatment group in Table 5.3 was 129 as well

$$H_0\colon\ \mu_z = 129.0 \quad \text{against} \quad H_1\colon\ \mu_z < 129.0.$$

Solution: Use s to estimate σ, from (5.14), we have

$$\text{ASL} = \Phi\left(\frac{86.857 - 129.0}{66.767/\sqrt{7}}\right) = \Phi(-1.67) = 0.04746 < 0.05.$$

So there is marginal evidence that the treated mice in our study have a mean survival time of less than 129.0 days. However, from (5.15), we have

$$\text{ASL} = \Pr\{t(6) < -1.67\} = 0.072979 > 0.05.$$

From (5.16), $t(\boldsymbol{z}) = -1.67$. Twenty thousand bootstrap samples were generated, and 1938 had $t(\boldsymbol{z}^*) < -1.67$. The estimated value of ASL is $1938/20000 = 0.0969$ so that we cannot reject H_0. Figure 5.3 shows a histogram of bootstrap replications of $t(\mathbf{z})$ defined in (5.16) for the test of $H_0\colon\ \mu_z = 129.0$ against $H_1\colon\ \mu_z < 129.0$.

R codes:

```
CS.Example5.8 <- function(G) {
     # ================ Aim ================================
     # Conduct Example 5.7
     # ================ Input ==============================
     # G = 20000 is the bootstrap sample size
     # ================ Output =============================
     # ASL and Figure 5.3
     # =====================================================
     mu0 <- 129.
     z <- c(94, 197, 16, 38, 99, 141, 23)
     n <- length(z)
     zwave <- z - mean(z) + mu0
     tz <- (mean(z) - mu0)/sqrt(var(z)/n)
```

```
    p <- rep(1/n, n)
    t.star.sample <- rep(0, G)
    for(g in 1:G) {
        zstar <- sample(zwave, n, prob = p, replace = T)
        tzstar <- (mean(zstar) - mu0)/sqrt(var(zstar)/n)
        t.star.sample[g] <- tzstar
    }
    ASL <- length(t.star.sample[t.star.sample < tz])/G
    #---------------------- Figure 5.3 ------------------
    par(pty = "s")
    hist(t.star.sample, probability = T, plot = T,
        include.lowest = T, main = "H_0: mean = 129.0",
        xlab = " ", ylab = " ")
    return(ASL)
}
```

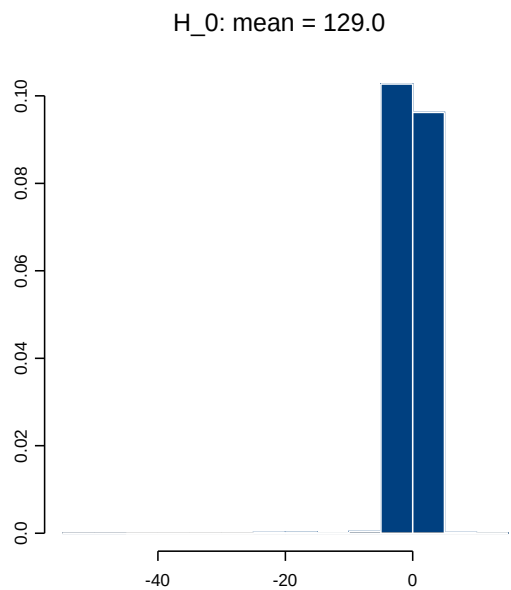

Figure 5.3 A histogram of bootstrap replications of $t(\mathbf{z})$ defined in (5.16) for the test of H_0: $\mu_z = 129.0$ against H_1: $\mu_z < 129.0$ for the treated mouse data. ‖

Exercise 5

5.1 Assume that the following observations

32.0,	46.4,	48.1,	27.7,	35.5,	52.6,	66.0,	41.3,
49.9,	36.1,	50.0,	44.7,	48.2,	36.9,	40.8,	35.1,
63.3,	42.5,	52.4,	40.9,	38.6,	43.2,	41.7,	35.6

are from a normal distribution with mean μ and variance σ^2.

(a) Calculate the MLEs of μ and σ^2.

(b) The *coefficient of variation* (CV) is defined as the ratio of the standard deviation to the mean; i.e., CV $= \sigma/\mu$. Use the parametric bootstrap method to find a 95% CI for CV.

(c) Let θ denote the population median. Use both the parametric bootstrap method and the non-parametric bootstrap method to find the 95% CIs for θ.

5.2 The definition of a ZIP distribution is given in Exercise 1.12. Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{ZIP}(\phi, \lambda)$ and $y_1, \ldots, y_n$ denote their realizations. Let $Y_{\text{obs}} = \{y_i\}_{i=1}^n$, $\mathbb{O} \hat{=} \{y_i\colon\ y_i = 0,\ i = 1, \ldots, n\}$, and m denote the number of elements in $\mathbb{O}$.

(a) What is the observed likelihood function for (ϕ, λ)?

(b) Derive the MLEs of ϕ and λ by using the EM algorithm.

(c) State the bootstrap method to obtain $100(1-\alpha)\%$ bootstrap confidence intervals for ϕ and λ.

5.3 A discrete random variable X is said to follow a *zero-truncated binomial* (ZTB) distribution or positive binomial distribution, denoted by $X \sim \text{ZTB}(m, \pi)$, if its pmf is

$$\Pr(X = x) = c \cdot \binom{m}{x} \pi^x (1-\pi)^{m-x}, \quad c \hat{=} \frac{1}{1-(1-\pi)^m},$$

for $x = 1, 2, \ldots, m$. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{ZTB}(m, \pi)$ and $x_1, \ldots, x_n$ denote their realizations. Let $Y_{\text{obs}} = \{x_i\}_{i=1}^n$.

(a) Derive the MLE of π by using an MM algorithm. [Hint: Use the inequality in Exercise 2.22(b3)]

(b) State the bootstrap method to obtain $100(1-\alpha)\%$ bootstrap confidence interval for π.

Appendix A

Some Statistical Distributions and Stochastic Processes

1• Introduction

- Some commonly used distributions including abbreviation, support, density, mean, variance, covariance, properties, relationship with other distributions and the r.v. generation method or some existing R functions, are listed.
- An exhaustive review of statistical distributions is provided by Johnson & Kotz (1969–1972) and Johnson *et al.* (1992, 1994, 1995).
- In addition, a summary of the homogeneous Poisson process and the nonhomogeneous Poisson process (Ross 1983) is also presented.
- All computer codes are written in R.

A.1 Discrete Distributions

A.1.1 Finite discrete distribution

2• Definition, moments and random variable generation

- A discrete r.v. X is said to follow a *finite discrete* distribution, denoted by $X \sim \text{FDiscrete}_d(\{x_i\}, \{p_i\})$, $x_i \in \mathbb{R}$, $p_i > 0$, $\sum_{i=1}^{d} p_i = 1$, if its pmf is $\Pr(X = x_i) = p_i$ for $i = 1, \ldots, d$, or equivalently

X	$x_1, \ldots, x_i, \ldots, x_d$
$\Pr(X = x_i)$	$p_1, \ldots, p_i, \ldots, p_d$

  where d is a known positive integer.

- $E(X) = \sum_{i=1}^{d} x_i p_i$ and $\text{Var}(X) = \sum_{i=1}^{d} x_i^2 p_i - (\sum_{i=1}^{d} x_i p_i)^2$.

- The built-in R function,

$$\mathsf{sample}(\boldsymbol{x},\ N,\ \mathsf{prob} = \boldsymbol{p},\ \mathsf{replace} = \mathsf{F/T})$$

produces random samples from this distribution; i.e., a vector of length N randomly chosen from $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ with corresponding probabilities $\boldsymbol{p} = (p_1, \ldots, p_d)^\top$ without/with replacement.

— For example, sample(0:1, 100, c(0.3, 0.7), T) will produce 100 i.i.d. Bernoulli(0.7) samples.

3• Some special cases

- The *uniform discrete* distribution is a special case of the finite discrete distribution with $p_i = 1/d$ for all $i = 1, \ldots, d$.
- A constant c can be viewed as a *degenerate* r.v. ξ with $\Pr(\xi = c) = 1$, denoted by $\xi \sim \text{Degenerate}(c)$, which is a special case of the finite discrete distribution with $d = 1$, $x_1 = c$ and $p_1 = 1$.
- $\text{FDiscrete}_2(\{x_1, x_2\}, \{1-p, p\})$ denotes the *two–point* distribution.
- Let $X \sim \text{FDiscrete}_2(\{x_1, x_2\}, \{1-p, p\})$, define $Y = (X - x_1)/(x_2 - x_1)$, then Y is said to follow the *Bernoulli* distribution with mean parameter p, denoted by $Y \sim \text{Bernoulli}(p)$.
  - The pmf of Y is $p^y(1-p)^{1-y}$ for $y = 0, 1$.
  - $E(Y) = p$ and $\text{Var}(Y) = p(1-p)$.
  - For any positive integer r, we have $Y^r \overset{\text{d}}{=} Y$.
  - Let r and s be two positive integers, we have $(1-Y)^r Y^s \sim \text{Degenerate}(0)$.

A.1.2 Hypergeometric distribution

4• Probability mass function and moments

- A discrete r.v. X is said to follow the *hypergeometric* distribution, denoted by $X \sim \text{Hgeometric}(m, n, k)$, if its pmf is

$$\text{Hgeometric}(x|m, n, k) = \binom{m}{x}\binom{n}{k-x} \Big/ \binom{m+n}{k},$$

for $x = \max(0, k-n), \ldots, \min(m, k)$.

- $E(X) = km/N_+$ and $\text{Var}(X) = kmn(N_+ - k)/\{N_+^2(N_+ - 1)\}$, where $N_+ \triangleq m + n$.
- The hypergeometric distribution can be described by an urn model with m red and n black balls. Any sequence of k drawings resulting in x red and $k - x$ black balls has the same probability.
- It is similar to the binomial distribution but sampled from a finite population without replacement.
- The built-in R function, rhyper(N, m, n, k), can be used to generate N i.i.d. samples of X.

A.1.3 Binomial and related distributions

5• Binomial distribution

- Let $\{X_j\}_{j=1}^n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, define $X = \sum_{j=1}^n X_j$, then X is said to follow the binomial distribution, denoted by $X \sim \text{Binomial}(n, p)$ with pmf

$$\text{Binomial}(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \tag{A.1}$$

  where n is a known positive integer and $p \in (0, 1)$.
- $E(X) = np \triangleq \mu$ and $\text{Var}(X) = np(1-p) = \mu(1 - \mu/n) < \mu$, so the binomial distribution is under-dispersed.
- The binomial generator in R is rbinom(N, n, p).

5.1• Three special distributions

— When $n = 1$, the binomial reduces to the Bernoulli(p) distribution.

— When $p = 0$, $\text{Binomial}(n, 0) = \text{Degenerate}(0)$.

— When $p = 1$, $\text{Binomial}(n, 1) = \text{Degenerate}(n)$.

5.2• Some properties

— Let $\{X_i\}_{i=1}^d \overset{\text{ind}}{\sim} \text{Binomial}(n_i, p)$, then $\sum_{i=1}^d X_i \sim \text{Binomial}(\sum_{i=1}^d n_i, p)$.

— Let $X \sim \text{Binomial}(n, p)$ and $Y|(X = x) \sim \text{Binomial}(x, \theta)$, then $Y \sim \text{Binomial}(n, p\theta)$.

— The binomial and beta distributions have the relationship:

$$\sum_{x=0}^{k} \text{Binomial}(x|n, p) = \int_0^{1-p} \text{Beta}(x|n-k, k+1)\,\mathrm{d}x, \tag{A.2}$$

where $0 \leqslant k \leqslant n$.

6• Zero-truncated binomial distribution

- Let $X \sim \text{Binomial}(n, p)$, define $Y = X|(X > 0)$, then Y follows the ZTB distribution, denoted by $Y \sim \text{ZTB}(n, p)$ with pmf

$$\text{ZTB}(y|n, p) = \frac{1}{1-(1-p)^n}\binom{n}{y}p^y(1-p)^{n-y}, \quad y = 1, \ldots, n, \tag{A.3}$$

  where n is a known positive integer and $p \in (0, 1)$.

- Y has the SR: $X \overset{\text{d}}{=} ZY$, where $Z \sim \text{Bernoulli}(1-(1-p)^n)$ and $Z \perp\!\!\!\perp Y$.

- $E(Y) = np/\{1-(1-p)^n\} \mathrel{\hat{=}} \mu$ and

$$\text{Var}(Y) = \mu\left\{1 - p - \frac{np(1-p)^n}{1-(1-p)^n}\right\} = \mu - \mu\{p + \mu(1-p)^n\} < \mu,$$

  so the ZTB distribution is under-dispersed.

- Let $\{Y_i\}_{i=1}^d \overset{\text{iid}}{\sim} \text{ZTB}(n, p)$, then the pmf of $Y_+ = \sum_{i=1}^d Y_i$ is given by

$$\Pr(Y_+ = y) = \frac{p^y(1-p)^{nd-y}}{\{1-(1-p)^n\}^d}\sum_{k=0}^{d}(-1)^k\binom{d}{k}\binom{nd-nk}{y},$$

  for $d \leqslant y \leqslant nd$, where $\binom{nd-nk}{y} \mathrel{\hat{=}} 0$ kf $y > n(d-k)$.

7• Beta–binomial distribution

- Let $p \sim \text{Beta}(a, b)$ and $X|p \sim \text{Binomial}(n, p)$, then X is said to follow the beta–binomial distribution, denoted by $X \sim \text{BBinomial}(n, a, b)$ with pmf

$$\text{BBinomial}(x|n, a, b) = \binom{n}{x}\frac{B(x+a, n-x+b)}{B(a, b)}, \tag{A.4}$$

  for $x = 0, 1, \ldots, n$, where $n > 0$ is an integer and $a, b > 0$.

- $E(X) = na/(a+b) \hat{=} \mu$ and

$$\begin{aligned} \mathrm{Var}(X) &= \frac{nab(a+b+n)}{(a+b)^2(a+b+1)} = \mu\left(1-\frac{\mu}{n}\right)\frac{a+b+n}{a+b+1} \\ &> \mu\frac{a+b+n}{a+b+1} \geqslant \mu, \end{aligned}$$

  so the beta–binomial distribution is over-dispersed.

- It reduces to the Bernoulli$(a/(a+b))$ distribution when $n = 1$.
- When $a = b = 1$, it becomes the discrete uniform distribution from 0 to n, i.e., $\mathrm{FDiscrete}_{n+1}(0:n, \mathbf{1}_{n+1}/(n+1))$.
- The beta–binomial is a robust alternative to the binomial distribution.

A.1.4 Poisson and related distributions

8• Poisson distribution

- A discrete r.v. X is said to follow the *Poisson* distribution, denoted by $X \sim \mathrm{Poisson}(\lambda)$, $\lambda > 0$, if its pmf is

$$\mathrm{Poisson}(x|\lambda) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \quad x = 0, 1, \ldots, \infty. \tag{A.5}$$

- $E(X) = \mathrm{Var}(X) = \lambda$, so the Poisson distribution is equi-dispersed.
- It is well known that when $n \to \infty$, $p \to 0$ while $np \to \lambda$, the pmf of Binomial(n, p) tends to the pmf of the Poisson(λ).
- The Poisson generator in R is rpois(N, λ).

8.1• Some properties

— When $\lambda = 0$, Poisson(0) = Degenerate(0).

— If $\{X_i\}_{i=1}^d \overset{\text{ind}}{\sim} \mathrm{Poisson}(\lambda_i)$, then

$$\textstyle\sum_{i=1}^d X_i \sim \mathrm{Poisson}(\sum_{i=1}^d \lambda_i), \tag{A.6}$$

$$(X_1, \ldots, X_d)|(\textstyle\sum_{i=1}^d X_i = n) \sim \mathrm{Multinomial}_d(n, \boldsymbol{p}), \tag{A.7}$$

where $\boldsymbol{p} = (\lambda_1, \ldots, \lambda_d)^\top / \sum_{i=1}^d \lambda_i$.

— The Poisson and gamma distributions have the relationship:

$$\sum_{x=k}^{\infty} \text{Poisson}(x|\lambda) = \int_0^{\lambda} \text{Gamma}(y|k,1)\,\mathrm{d}y. \tag{A.8}$$

— Let $X \sim \text{Poisson}(\lambda)$ and $Y|(X = x) \sim \text{Binomial}(x, p)$, then $Y \sim \text{Poisson}(\lambda p)$.

9• Zero-truncated poisson distribution

- Let $X \sim \text{Poisson}(\lambda)$ with $\lambda > 0$, define $Y = X|(X > 0)$, then Y follows the ZTP distribution, denoted by $Y \sim \text{ZTP}(\lambda)$ with pmf

$$\text{ZTP}(y|\lambda) = \frac{1}{1 - \mathrm{e}^{-\lambda}} \cdot \frac{\lambda^y \mathrm{e}^{-\lambda}}{y!}, \quad y = 1, 2, \ldots, \infty. \tag{A.9}$$

- Y has the SR: $X \overset{\mathrm{d}}{=} ZY$, where $Z \sim \text{Bernoulli}(1 - \mathrm{e}^{-\lambda})$ and $Z \perp\!\!\!\perp Y$.
- $E(Y) = \lambda/(1 - \mathrm{e}^{-\lambda}) \,\hat{=}\, \mu$ and

$$\text{Var}(Y) = \frac{\lambda}{1 - \mathrm{e}^{-\lambda}} \left(1 - \frac{\lambda \mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}}\right) = \mu - \mu^2 \mathrm{e}^{-\lambda} < \mu,$$

  so the ZTP distribution is under-dispersed.

- Let $\{Y_i\}_{i=1}^d \overset{\text{iid}}{\sim} \text{ZTP}(\lambda)$, then the pmf of $Y_+ = \sum_{i=1}^d Y_i$ is given by

$$\Pr(Y_+ = y) = \frac{\lambda^y \mathrm{e}^{-\lambda d}}{(1 - \mathrm{e}^{-\lambda})^d y!} \sum_{k=0}^{d} (-1)^k \binom{d}{k} (d-k)^y, \quad y \geqslant d.$$

10• Charlier series distribution

- Let $X_1 \sim \text{Binomial}(n, p)$, $X_2 \sim \text{Poisson}(\lambda)$, and $X_1 \perp\!\!\!\perp X_2$, define $X = X_1 + X_2$, then X is said to follow the *Charlier series* (CS) distribution (Ong 1988), denoted by $X \sim \text{CS}(n, p, \lambda)$ with pmf

$$\Pr(X = x) = \sum_{k=0}^{\min(n,x)} \binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{\lambda^{x-k} \mathrm{e}^{-\lambda}}{(x-k)!}, \tag{A.10}$$

  for $x = 0, 1, \ldots, \infty$, where n is a known positive integer, $p \in (0, 1)$ and $\lambda > 0$.

- $E(X) = np + \lambda \hat{=} \mu$ and $\text{Var}(X) = np(1-p) + \lambda = \mu - np^2 < \mu$, so the CS distribution is under-dispersed.
- When $p = 0$, $\text{CS}(n, 0, \lambda) = \text{Poisson}(\lambda)$.
- When $\lambda = 0$, $\text{CS}(n, p, 0) = \text{Binomial}(n, p)$.

11• Gamma–poisson (mixture) distribution

- Let $\lambda \sim \text{Gamma}(\alpha, \beta)$ and $X|\lambda \sim \text{Poisson}(\lambda)$, then X is said to follow the gamma–Poisson (mixture) distribution, denoted by $X \sim \text{GPoisson}(\alpha, \beta)$ with pmf

$$\text{GPoisson}(x|\alpha, \beta) = \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{x}, \qquad \text{(A.11)}$$

  for $x = 0, 1, \ldots, \infty$, where the shape parameter $\alpha > 0$ and the rate parameter $\beta > 0$.
- $E(X) = \alpha/\beta \hat{=} \mu$ and $\text{Var}(X) = \alpha(\beta+1)/\beta^2 = \mu(1 + 1/\beta) > \mu$, so the gamma–Poisson distribution is over-dispersed.
- The gamma–Poisson is a robust alternative to the Poisson distribution.

11.1• Type I negative–binomial distribution

— Reparametrizing (α, β) by (μ, β) through $\mu = \alpha/\beta$, Cameron & Trivedi (1986) called (A.11) as Type I negative–binomial distribution, denoted by $X \sim \text{NB1}(\mu, \beta)$ with pmf

$$\text{NB1}(x|\mu, \beta) = \frac{\Gamma(x+\mu\beta)}{x!\Gamma(\mu\beta)} \left(\frac{\beta}{\beta+1}\right)^{\mu\beta} \left(\frac{1}{\beta+1}\right)^{x}, \qquad \text{(A.12)}$$

for $x = 0, 1, \ldots, \infty$, where the mean parameter $\mu > 0$ and the rate parameter $\beta > 0$.

— The pmf (A.12) can be used for constructing a regression model to link the mean parameter with a set of covariates.

— $E(X) = \mu$ and $\text{Var}(X) = \mu(1 + 1/\beta)$ is a linear function of the mean μ.

11.2• Type II negative–binomial distribution

— Reparametrizing (α, β) by (α, μ) through $\mu = \alpha/\beta$, Cameron & Trivedi (1986) called (A.11) as Type II negative–binomial distribution, denoted by $X \sim \text{NB2}(\alpha, \mu)$ with pmf

$$\text{NB2}(x|\alpha, \mu) = \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \left(\frac{\alpha}{\alpha+\mu}\right)^{\alpha} \left(\frac{\mu}{\alpha+\mu}\right)^{x}, \tag{A.13}$$

for $x = 0, 1, \ldots, \infty$, where the shape parameter $\alpha > 0$ and the mean parameter $\mu > 0$.

— $E(X) = \mu$ and $\text{Var}(X) = \mu + \mu^2/\alpha$ is a quadratic function of μ.

A.1.5 Negative–binomial and related distributions

12• Negative–binomial distribution

- In (A.11), let $\beta/(\beta+1) = p$ and $\alpha = r$, the gamma–Poisson reduces to the negative–binomial distribution or the Polya distribution (after George Pólya), denoted by $X \sim \text{NBinomial}(r, p)$ with pmf

$$\text{NBinomial}(x|r, p) = \frac{\Gamma(x+r)}{x!\Gamma(r)} p^r (1-p)^x, \tag{A.14}$$

for $x = 0, 1, \ldots, \infty$, where $r > 0$ is a real number and $p \in (0, 1)$.

- $E(X) = r(1-p)/p$ and $\text{Var}(X) = r(1-p)/p^2$.
- The negative–binomial in R is `rnbinom(N, r, p)`.

12.1• Pascal distribution

— When r is a positive integer, the negative-binomial distribution is also called the *Pascal* distribution (after Blaise Pascal) with pmf

$$\text{NBinomial}(x|r, p) = \binom{x+r-1}{x} p^r (1-p)^x, \tag{A.15}$$

for $x = 0, 1, \ldots, \infty$.

— The Pascal r.v. X can be viewed as the number of failures in a sequence of i.i.d. Bernoulli trials before r successes occur, where p is the probability of success.

— Define $\pi = 1 - p$, we have

$$p^{-r} = (1-\pi)^{-r} = \sum_{x=0}^{\infty} \binom{-r}{x} (-\pi)^x = \sum_{x=0}^{\infty} \binom{x+r-1}{x} (1-p)^x.$$

Thus $\sum_{x=0}^{\infty} \text{NBinomial}(x|r,p) = 1$.

12.2• Geometric distribution

— The *geometric* distribution is a special case of the Pascal distribution with $r = 1$.

13• Zero-truncated negative–binomial distribution

- Let $X \sim \text{NBinomial}(r,p)$, define $Y = X|(X > 0)$, then Y is said to follow the ZTNB distribution, denoted by $Y \sim \text{ZTNB}(r,p)$ with pmf

$$\text{ZTNB}(y|r,p) = \frac{1}{1-p^r} \frac{\Gamma(y+r)}{y!\Gamma(r)} p^r (1-p)^y, \tag{A.16}$$

for $y = 1, 2, \ldots, \infty$, where $r > 0$ is a real number and $p \in (0,1)$.

- Y has the SR: $X \overset{\text{d}}{=} ZY$, where $Z \sim \text{Bernoulli}(1-p^r)$ and $Z \perp\!\!\!\perp Y$.

- $E(Y) = r(1-p)/\{p(1-p^r)\}$ and

$$\text{Var}(Y) = \frac{r(1-p)\{1-p^r - r(1-p)p^r\}}{p^2(1-p^r)^2}.$$

- Let $\{Y_i\}_{i=1}^d \overset{\text{iid}}{\sim} \text{ZTNB}(r,p)$, where $r > 0$ is assumed to be a known real number, then the pmf of $Y_+ = \sum_{i=1}^d Y_i$ is given by

$$\Pr(Y_+ = y) = \frac{(1-p)^y p^{rd}}{(1-p^r)^d} \sum_{k=0}^{d} (-1)^k \binom{d}{k} \frac{\Gamma(y+rd-rk)}{y!\Gamma(rd-rk)}, \quad y \geqslant d.$$

14• Logarithmic series distribution

- The pmf of the logarithmic series (or log-series) distribution is

$$\text{Logseries}(y|p) = \frac{-1}{\log p} \cdot \frac{(1-p)^y}{y}, \quad y = 1, 2, \ldots, \infty, \tag{A.17}$$

where $p \in (0,1)$, denoted by $Y \sim \text{Logseries}(p)$.

- From the following Maclaurin's expansion:

$$-\log p = (1-p) + \frac{(1-p)^2}{2} + \frac{(1-p)^3}{3} + \cdots,$$

we can obtain the identity $\sum_{y=1}^{\infty} \text{Logseries}(y|p) = 1$.

- The mean and variance are given by

$$E(Y) = \frac{-1}{\log p} \cdot \frac{1-p}{p} \quad \text{and} \quad \text{Var}(Y) = -\frac{(1-p)(1-p+\log p)}{p^2(\log p)^2}.$$

14.1• Limiting distribution of the ZTNB distribution

— The mgf is $E(\mathrm{e}^{tY}) = \log\{1-(1-p)\,\mathrm{e}^t\}/\log p$.

— The log-series distribution can be obtained as a limiting distribution of the zero-truncated negative–binomial distribution as $r \to 0$; i.e.,

$$\lim_{r\to 0} \text{ZTNB}(y|r,p) = \text{Logseries}(y|p).$$

— Let $\{Y_k\}_{k=1}^{\infty} \overset{\text{iid}}{\sim} \text{Logseries}(p)$, $N \sim \text{Poisson}(\lambda)$ and $N \perp\!\!\!\perp \{Y_k\}_{k=1}^{\infty}$, then $\sum_{k=1}^{N} Y_k \sim \text{NBinomial}(r,p)$, where $r = \lambda/(-\log p)$.

A.1.6 Generalized Poisson and related distributions

15• Generalized poisson distribution

- A discrete r.v. X is said to have a *generalized Poisson* (GP) distribution, denoted by $X \sim \text{GP}(\lambda,\theta)$ with pmf (Consul & Jain 1973; Consul 1989, p.4)

$$\text{GP}(x|\lambda,\theta) = \begin{cases} \dfrac{\lambda(\lambda+\theta x)^{x-1}\,\mathrm{e}^{-\lambda-\theta x}}{x!}, & x = 0,1,\ldots,\infty, \\ 0, & \text{for } x > q, \text{ when } \theta < 0, \end{cases} \tag{A.18}$$

where $\lambda > 0$, $\max(-1, -\lambda/q) < \theta \leqslant 1$ and $q\,(\geqslant 4)$ is the largest positive integer for which $\lambda + \theta q > 0$ when $\theta < 0$.

- $E(X) = \lambda/(1-\theta)$ and $\text{Var}(X) = \lambda/(1-\theta)^3$, where $\theta < 1$.

- When $\theta = 0$, $\text{GP}(\lambda, 0) = \text{Poisson}(\lambda)$.

- The GP(λ, θ) distribution is over-dispersed when $\theta > 0$ and under-dispersed when $\theta < 0$.
- The most frequently used version of the GP distribution assumes that $\lambda > 0$ and $\theta \in [0, 1)$.
- Let $\{X_i\}_{i=1}^2 \overset{\text{ind}}{\sim} \text{GP}(\lambda_i, \theta)$, then $X_1 + X_2 \sim \text{GP}(\lambda_1 + \lambda_2, \theta)$.

16• Zero-truncated generalized poisson distribution

- Let $X \sim \text{GP}(\lambda, \theta)$, define $Y = X|(X > 0)$, then Y is said to follow the ZTGP distribution, denoted by $Y \sim \text{ZTGP}(\lambda, \theta)$ with pmf

$$\text{ZTGP}(y|\lambda, \theta) = \frac{1}{1 - \mathrm{e}^{-\lambda}} \cdot \frac{\lambda(\lambda + \theta y)^{y-1} \mathrm{e}^{-\lambda-\theta y}}{y!}, \tag{A.19}$$

  for $y = 1, 2, \ldots, \infty$, where $\lambda > 0$ and $\theta \in [0, 1)$.
- Y has the SR: $X \overset{\text{d}}{=} ZY$, where $Z \sim \text{Bernoulli}(1 - \mathrm{e}^{-\lambda})$ and $Z \perp\!\!\!\perp Y$.
- $E(Y) = \lambda/\{(1 - \theta)(1 - \mathrm{e}^{-\lambda})\} \hat{=} \mu$ and

$$\text{Var}(Y) = \mu \frac{1 + \lambda(1 - \theta)}{(1 - \theta)^2} - \mu^2.$$

- Let $\{Y_i\}_{i=1}^d \overset{\text{iid}}{\sim} \text{ZTGP}(\lambda, \theta)$, then the pmf of $Y_+ = \sum_{i=1}^d Y_i$ is

$$\Pr(Y_+ = y) = \frac{\lambda \mathrm{e}^{-\lambda d - \theta y}}{(1 - \mathrm{e}^{-\lambda})^d y!} \sum_{k=0}^{d} (-1)^k \binom{d}{k} (d - k)\{(d - k)\lambda + \theta y\}^{y-1},$$

  for $y \geqslant d$.

A.1.7 Multinomial and related distributions

17• Multinomial distribution

- A d-dimensional random vector $\mathbf{x} = (X_1, \ldots, X_d)^\top$ is said to follow the multinomial distribution, denoted by $\mathbf{x} \sim \text{Multinomial}(n; p_1, \ldots, p_d)$ or $\mathbf{x} \sim \text{Multinomial}_d(n, \boldsymbol{p})$, if its joint pmf is

$$\text{Multinomial}_d(\boldsymbol{x}|n, \boldsymbol{p}) = \binom{n}{x_1, \ldots, x_d} \prod_{i=1}^{d} p_i^{x_i}, \quad \boldsymbol{x} \in \mathbb{T}_d(n), \tag{A.20}$$

  where n is a positive integer and $\boldsymbol{p} = (p_1, \ldots, p_d)^\top \in \mathbb{T}_d$.

- $E(X_i) = np_i$, $\mathrm{Var}(X_i) = np_i(1-p_i)$, $\mathrm{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$.
- For $i = 1, \ldots, d$, we have $X_i \sim \mathrm{Binomial}(n, p_i)$, and for $i = 2, \ldots, d$,

$$X_i | (X_1 = x_1, \ldots, X_{i-1} = x_{i-1}) \sim \mathrm{Binomial}\left(n - \sum_{j=1}^{i-1} x_j, \ \frac{p_i}{\sum_{j=i}^{d} p_j} \right).$$

- The binomial distribution is a special case of the multinomial with $d = 2$.
- The multinomial generator in R is rmultinom(N, n, $\boldsymbol{p}$).

18• Dirichlet–multinomial distribution

- Let $\boldsymbol{a} = (a_1, \ldots, a_d)^\top$ with $a_i > 0$ for $i = 1, \ldots, d$,

$$\boldsymbol{p} \sim \mathrm{Dirichlet}_d(\boldsymbol{a}) \quad \text{and} \quad \mathbf{x} | \boldsymbol{p} \sim \mathrm{Multinomial}_d(n, \boldsymbol{p}),$$

then $\mathbf{x} = (X_1, \ldots, X_d)^\top$ is said to follow the Dirichlet–multinomial distribution, denoted by $\mathbf{x} \sim \mathrm{DMultinomial}_d(n, \boldsymbol{a})$ with pmf

$$\mathrm{DMultinomial}_d(\boldsymbol{x} | n, \boldsymbol{a}) = \binom{n}{x_1, \ldots, x_d} \frac{B_d(\boldsymbol{x} + \boldsymbol{a})}{B_d(\boldsymbol{a})}, \tag{A.21}$$

for $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{T}_d(n)$, where $n > 0$ is an integer, and

$$B_d(\boldsymbol{a}) = \frac{\prod_{i=1}^{d} \Gamma(a_i)}{\Gamma(\sum_{i=1}^{d} a_i)} \tag{A.22}$$

is the multivariate beta function.

- The moments are given by

$$\begin{aligned} E(X_i) &= na_i/a_+, \qquad a_+ \hat{=} \textstyle\sum_{i=1}^{d} a_i, \\ \mathrm{Var}(X_i) &= n(a_+ + n)a_i(a_+ - a_i)/\{a_+^2(a_+ + 1)\}, \\ \mathrm{Cov}(X_i, X_j) &= -n(a_+ + n)a_ia_j/\{a_+^2(a_+ + 1)\}, \quad i \neq j. \end{aligned}$$

- When $d = 2$, the Dirichlet–multinomial reduces to the beta–binomial.
- The Dirichlet–multinomial is a robust alternative to the multinomial distribution.

A.2 Continuous Distributions

A.2.1 Uniform, beta and Dirichlet distributions

19• Uniform distribution

- A continuous r.v. X follows the uniform distribution on the interval (a, b) for $a < b$, denoted by $X \sim U(a, b)$, if its pdf is

$$U(x|a, b) = \frac{1}{b-a}, \quad x \in (a, b).$$

- $E(X) = (a + b)/2$ and $\text{Var}(X) = (b - a)^2/12$.
- A non-informative distribution is obtained by letting $a \to -\infty$ and $b \to \infty$.
- If $Y \sim U(0, 1)$, then $X = a + (b - a)Y \sim U(a, b)$.
- The uniform random number generator in R is runif(N, a, b).

20• Beta distribution

- A r.v. X follows the beta distribution, denoted by $X \sim \text{Beta}(a, b)$ with $a > 0$ and $b > 0$, if its pdf is

$$\text{Beta}(x|a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}, \quad x \in (0, 1), \tag{A.23}$$

  where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ denotes the beta function.

- The moments of $X \sim \text{Beta}(a, b)$ are given by

$$\begin{aligned} E(X) &= \frac{a}{a+b}, \quad E(X^2) = \frac{a(a+1)}{(a+b)(a+b+1)}, \quad \text{and} \\ \text{Var}(X) &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

20.1• Some properties

— When $a = b = 1$, $\text{Beta}(1, 1) = U(0, 1)$.

— The k-th order statistic of a random sample of N i.i.d. $U(0, 1)$ follows $\text{Beta}(k, N - k + 1)$.

— The beta and gamma distributions have the following relationship:

$$X \overset{\mathrm{d}}{=} \frac{Y}{Y+Z},$$

where $Y \sim \text{Gamma}(a,1)$, $Z \sim \text{Gamma}(b,1)$ and $Y \perp\!\!\!\perp Z$.

— The beta distribution is the conjugate prior for the parameter of success probability in a binomial (or negative–binomial) distribution.

— A non-informative distribution is obtained as $a, b \to 0$.

— The beta generator is `rbeta(N, a, b)`.

21• Dirichlet distribution

- A random vector $\mathbf{x} = (X_1, \ldots, X_d)^\top$ follows the Dirichlet distribution, denoted by $\mathbf{x} \sim \text{Dirichlet}(a_1, \ldots, a_d)$ or $\mathbf{x} \sim \text{Dirichlet}_d(\boldsymbol{a})$ with $\boldsymbol{a} = (a_1, \ldots, a_d)^\top$, $a_i > 0$, $i = 1, \ldots, d$, if its joint pdf is

$$\text{Dirichlet}_d(\boldsymbol{x}|\boldsymbol{a}) = \frac{\prod_{i=1}^d x_i^{a_i-1}}{B_d(\boldsymbol{a})}, \quad \boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{T}_d, \tag{A.24}$$

where $B_d(\boldsymbol{a})$ is the multivariate beta function defined by (A.22).

- Let $a_+ = \sum_{i=1}^d a_i$, then

$$\begin{aligned} E(X_i) &= \frac{a_i}{a_+}, \quad \text{Var}(X_i) = \frac{a_i(a_+ - a_i)}{a_+^2(a_+ + 1)}, \quad \text{and} \\ \text{Cov}(X_i, X_j) &= -\frac{a_i a_j}{a_+^2(a_+ + 1)}, \quad i \neq j. \end{aligned}$$

21.1• Basic properties

— When $d = 2$, we have $\text{Dirichlet}(a_1, a_2) = \text{Beta}(a_1, a_2)$.

— The Dirichlet and gamma distributions have the following relationship:

$$X_i \overset{\mathrm{d}}{=} \frac{Y_i}{\sum_{j=1}^d Y_j}, \quad \text{where} \quad \{Y_i\}_{i=1}^d \overset{\text{ind}}{\sim} \text{Gamma}(a_i, 1). \tag{A.25}$$

— The Dirichlet is the conjugate prior for the parameter vector in a multinomial distribution.

— A non-informative distribution is obtained as $a_i \to 0$ for all i.

21.2• Generation of Dirichlet random vector

— Let $\{W_i\}_{i=1}^{d-1} \overset{\text{ind}}{\sim} \text{Beta}(a_1 + \cdots + a_i,\ a_{i+1})$, then, the following SR (Fang *et al.* 1990, p.146)

$$\begin{cases} X_i \overset{\text{d}}{=} (1 - W_{i-1}) \prod_{j=i}^{d-1} W_j,\ i = 1, \ldots, d-1,\ W_0 \equiv 0, \\ X_d \overset{\text{d}}{=} 1 - W_{d-1} \end{cases} \tag{A.26}$$

can be used to generate a random sample of $\mathbf{x}$ from Dirichlet$(a_1, \ldots, a_d)$.

— The corresponding R codes are given by

```
rDirichlet <- function(a) {
    # ================= Aim ==============================
    # Generate one random vector (x_1, ..., x_d) from
    # Dirichlet(a_1, ..., a_d) with d >= 3
    # ================= Input ============================
    # a = (a_1, ..., a_d)
    # ================= Output ===========================
    # x = (x_1, ..., x_d)
    # ====================================================
    d <- length(a)
    w <- rep(0, d - 1)
    for(i in 1:(d - 1)) {
        w[i] <- rbeta(1, sum(a[1:i]), a[i + 1])
    }
    x <- rep(0, d)
    x[1] <- prod(w)
    for(i in 2:(d - 1)) {
        x[i] <- (1 - w[i - 1]) * prod(w[i:(d - 1)])
    }
    x[d] <- 1 - sum(x[1:(d - 1)])
    return(x)
}
```

A.2.2 Logistic and Laplace distributions

22• Logistic distribution

- A r.v. X is said to follow the logistic distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$, denoted by $X \sim \text{Logistic}(\mu, \sigma^2)$, if its pdf is

$$\text{Logistic}(x|\mu, \sigma^2) = \frac{\exp(-\frac{x-\mu}{\sigma})}{\sigma\{1 + \exp(-\frac{x-\mu}{\sigma})\}^2}, \quad x \in \mathbb{R}. \tag{A.27}$$

- $E(X) = \mu$ and $\text{Var}(X) = \pi^2\sigma^2/3$.
- The logistic is another symmetric and unimodal distribution, more similar to the normal in appearance than the Laplace, but with even heavier tails.
- Both the cdf and its inverse function have closed–form expressions:

$$\begin{aligned} F(x|\mu, \sigma^2) &= \left(1 + \mathrm{e}^{-\frac{x-\mu}{\sigma}}\right)^{-1}, \quad x \in \mathbb{R}, \\ F^{-1}(x|\mu, \sigma^2) &= \mu + \sigma \log\left(\frac{x}{1-x}\right), \quad x \in (0, 1). \end{aligned}$$

- The logistic generator in R is rlogis(N, μ, σ).

23• Laplace (or double exponential) distribution

- A r.v. X is said to follow the Laplace distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$, denoted by $X \sim \text{Laplace}(\mu, \sigma^2)$, if its pdf is

$$\text{Laplace}(x|\mu, \sigma^2) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right), \quad x \in \mathbb{R}. \tag{A.28}$$

- $E(X) = \mu$ and $\text{Var}(X) = 2\sigma^2$.
- Like the Gaussian distribution, the Laplace is symmetric and unimodal, but has heavier tail.
- $|X - \mu| \sim \text{Exponential}(1/\sigma)$.

- Let $\{Y_i\}_{i=1}^4 \overset{\text{iid}}{\sim} N(0,1)$, then $Y_1Y_4 - Y_2Y_3 \sim \text{Laplace}(0, 2^2)$, see Nyquist *et al.* (1954).

23.1• Generation of Laplace random variable

— The inversion method (see §1.1) can be used to draw N i.i.d. samples from the standard Laplace$(0, 1)$ distribution.

— The R codes are given by

```
rLaplace <- function(N) {
    # ================= Aim ==============================
    # Generate N i.i.d. samples from the standard
    # Laplace(0, 1) distribution
    # ================= Input ============================
    # N = the sample size
    # ================= Output ===========================
    # x = (x_1, ..., x_N)
    # ====================================================
    u <- runif(N)
    x <- rep(0, N)
    for(i in 1:N) {
       if(u[i] < 0.5) { x[i] <- log(2 * u[i]) }
       else { x[i] <-  - log(2 * (1 - u[i]))  }
    }
    return(x)
}
```

A.2.3 Exponential, gamma and inverse gamma distributions

24• Exponential distribution

- $X \sim \text{Exponential}(\beta)$, rate $\beta > 0$,

$$\text{Exponential}(x|\beta) = \beta\, \mathrm{e}^{-\beta x}, \quad x \geqslant 0.$$

- $E(X) = 1/\beta$ and $\text{Var}(X) = 1/\beta^2$.
- If $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} \text{Exponential}(\beta)$, then $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \beta)$.
- The exponential distribution generator in R is rexp(N, β).

25• Gamma distribution

- $X \sim \text{Gamma}(\alpha, \beta)$, shape $\alpha > 0$, rate $\beta > 0$,

$$\text{Gamma}(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\,\mathrm{e}^{-\beta x}, \quad x \in \mathbb{R}_+,$$

  where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}\,\mathrm{e}^{-y}\,\mathrm{d}y$ denotes the gamma function.
- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.
- $E(X) = \alpha/\beta$ and $\text{Var}(X) = \alpha/\beta^2$.
- $\text{Gamma}(1,\beta) \equiv \text{Exponential}(\beta)$.
- $\text{Gamma}(\nu/2, 1/2) \equiv \chi^2(\nu)$.

25.1• Some properties

— If $X \sim \text{Gamma}(\alpha,\beta)$ and $c > 0$, then $Y = cX \sim \text{Gamma}(\alpha, \beta/c)$.

— If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} \text{Gamma}(\alpha_i,\beta)$, then $\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.

— The gamma is the conjugate prior for the Poisson mean and for the inverse of the normal variance.

— A noninformative distribution is obtained as $\alpha, \beta \to 0$.

— The gamma generator in R is rgamma(N, α, β).

26• Inverse gamma distribution

- $X \sim \text{IGamma}(\alpha,\beta)$, shape $\alpha > 0$, scale $\beta > 0$,

$$\text{IGamma}(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)}\,\mathrm{e}^{-\beta/x}, \quad x > 0.$$

- $E(X) = \beta/(\alpha-1)$ (if $\alpha > 1$) and $\text{Var}(X) = \beta^2/\{(\alpha-1)^2(\alpha-2)\}$ (if $\alpha > 2$).
- $\text{IGamma}(x|\alpha,\beta) = \text{Gamma}(x^{-1}|\alpha,\beta)/x^2$.
- If $X^{-1} \sim \text{Gamma}(\alpha,\beta)$, then $X \sim \text{IGamma}(\alpha,\beta)$.
- The inverse gamma is the conjugate prior for the normal variance σ^2.
- A noninformative distribution is obtained as $\alpha, \beta \to 0$.

A.2.4 Chi-square, F and inverse chi-square distributions

27• Chi-square distribution

- $X \sim \chi^2(\nu) \equiv \text{Gamma}(\nu/2, 1/2)$, degree of freedom $\nu > 0$,

$$\chi^2(x|\nu) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} \, \mathrm{e}^{-x/2}, \quad x \in \mathbb{R}_+.$$

- $E(X) = \nu$ and $\text{Var}(X) = 2\nu$.
- If $Y \sim N(0,1)$, then $X = Y^2 \sim \chi^2(1)$.
- If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} \chi^2(\nu_i)$, then $\sum_{i=1}^n X_i \sim \chi^2(\sum_{i=1}^n \nu_i)$.
- The chi-square generator is $\mathsf{rchisq}(N, \nu)$.

28• F or fisher's f distribution

- $X \sim F(\nu_1, \nu_2)$, ν_1, ν_2 are positive integers,

$$F(x|\nu_1, \nu_2) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2, \nu_2/2)} x^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1 x}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}}, \quad x \in \mathbb{R}_+.$$

- $E(X) = \nu_2/(\nu_2 - 2)$ (if $\nu_2 > 2$) and

$$\text{Var}(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)(\nu_2 - 2)^2} \quad (\text{if } \nu_2 > 4).$$

- Let $\{Y_i\}_{i=1}^2 \overset{\text{ind}}{\sim} \chi^2(\nu_i)$, then $X \overset{\text{d}}{=} (Y_1/\nu_1)/(Y_2/\nu_2) \sim F(\nu_1, \nu_2)$.
- The F generator is $\mathsf{rf}(N, \nu_1, \nu_2)$.

29• Inverse chi-square distribution

- $X \sim \text{I}\chi^2(\nu) \equiv \text{IGamma}(\frac{\nu}{2}, \frac{1}{2})$, $\nu > 0$,

$$\text{I}\chi^2(x|\nu) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{-(\nu/2+1)} \, \mathrm{e}^{-1/(2x)}, \quad x > 0.$$

- $E(X) = 1/(\nu - 2)$ (if $\nu > 2$) and

$$\text{Var}(X) = \frac{2}{(\nu - 2)^2(\nu - 4)} \quad (\text{if } \nu > 4).$$

A.2.5 Normal, Lognormal and Inverse Gaussian distributions

30• Normal or gaussian distribution

- $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$,

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R}.$$

- $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.
- If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} N(\mu_i, \sigma_i^2)$, then $\sum_{i=1}^n a_i X_i \sim N(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2)$.
- If $X_1|X_2 \sim N(X_2, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 \sim N(\mu_2, \sigma_1^2 + \sigma_2^2)$.
- A noninformative or flat distribution is obtained as $\sigma^2 \to \infty$.
- The normal generator is rnorm(N, μ, σ).

31• Lognormal distribution

- $X \sim \text{Lognormal}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$,

$$\text{Lognormal}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{ -\frac{(\log x - \mu)^2}{2\sigma^2} \right\}, \quad x > 0.$$

- $E(X) = \exp(\mu + 0.5\sigma^2)$ and $\text{Var}(X) = \{E(X)\}^2\{\exp(\sigma^2) - 1\}$.
- If $\log(X) \sim N(\mu, \sigma^2)$, then $X \sim \text{Lognormal}(\mu, \sigma^2)$.
- The lognormal generator is rlnorm(N, μ, σ).

32• Inverse gaussian or wald distribution

- $X \sim \text{IGaussian}(\mu, \lambda)$, $\mu > 0$, $\lambda > 0$,

$$\text{IGaussian}(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{ -\frac{\lambda(x-\mu)^2}{2\mu^2 x} \right\}, \quad x > 0. \quad \text{(A.29)}$$

- $E(X) = \mu$ and $\text{Var}(X) = \mu^3/\lambda$.
- If $X \sim \text{IGaussian}(\mu, \lambda)$ and $c > 0$, then $cX \sim \text{IGaussian}(c\mu, c\lambda)$.

- If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} \text{IGaussian}(\mu_i, c\mu_i^2)$, then

$$\sum_{i=1}^n X_i \sim \text{IGaussian}(\mu_+, c\mu_+^2) \quad \text{with} \quad \mu_+ = \sum_{i=1}^n \mu_i.$$

32.1• Generation of inverse Gaussian random variable

— Let $X \sim \text{IGaussian}(\mu, \lambda)$, then $\lambda(X-\mu)^2 \,/(\mu^2 X) \sim \chi^2(1)$.

— This result first obtained by Shuster (1968) can be used to simulate N i.i.d. samples from the inverse Gaussian distribution.

— The corresponding R codes are given by

```
rigaussian <- function(N, mu, lambda) {
    # ================ Aim ============================
    # Generate N i.i.d. samples from IGaussian(mu, lambda)
    # with pdf (A.29)
    # ================ Input ==========================
    # N      = the sample size
    # mu     = location parameter
    # lambda = scale parameter
    # ================ Output =========================
    # x = (x_1, ..., x_N)
    # =================================================
    y <- (rnorm(N, 0, 1))^2
    a <- (mu^2/(2 * lambda)) * y
    b <- 4 * mu * lambda * y + mu^2 * y^2
    x1 <- mu + a - (mu/(2 * lambda)) * sqrt(b)
    u <- runif(N)
    x <- rep(0, N)
    for(i in 1:N) {
       if(u[i] < mu/(mu + x1[i])) { x[i] <- x1[i] }
       else { x[i] <- mu^2/x1[i] }
    }
    return(x)
}
```

A.2.6 Multivariate normal distribution

33• Multivariate normal or gaussian distribution

- $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} > 0$,

$$N_d(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi}\,)^d |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},$$

  where $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$.

- $E(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$.

- Assume that

$$\mathbf{x}|(\mathbf{y} = \boldsymbol{y}) \sim N_d(\boldsymbol{A}\boldsymbol{y},\ \boldsymbol{\Sigma}_{x|y}) \quad \text{and} \quad \mathbf{y} \sim N_q(\boldsymbol{\mu}_y,\ \boldsymbol{\Sigma}_y), \tag{A.30}$$

  then the joint pdf of $(\mathbf{x}, \mathbf{y})$ is also multinormal.

  Proof: Note that $E(\mathbf{x}|\mathbf{y} = \boldsymbol{y}) = \boldsymbol{A}\boldsymbol{y}$ is a linear function of $\boldsymbol{y}$. By using the following identities:

$$\begin{aligned} E(\mathbf{x}) &= E\{E(\mathbf{x}|\mathbf{y})\} \quad \text{and} \\ \mathrm{Var}(\mathbf{x}) &= E\{\mathrm{Var}(\mathbf{x}|\mathbf{y})\} + \mathrm{Var}\{E(\mathbf{x}|\mathbf{y})\}, \end{aligned}$$

  we have $E(\mathbf{x}) = \boldsymbol{A}\boldsymbol{\mu}_y$, $\mathrm{Var}(\mathbf{x}) = \boldsymbol{\Sigma}_{x|y} + \boldsymbol{A}\boldsymbol{\Sigma}_y\boldsymbol{A}^\top$ and

$$\begin{aligned} \mathrm{Cov}(\mathbf{x}, \mathbf{y}) &= E(\mathbf{x}\mathbf{y}^\top) - E(\mathbf{x})E(\mathbf{y}^\top) \\ &= E\{E(\mathbf{x}\mathbf{y}^\top|\mathbf{y})\} - \boldsymbol{A}\mu_y E(\mathbf{y}^\top) \\ &= \boldsymbol{A}\{E(\mathbf{y}\mathbf{y}^\top) - E(\mathbf{y})E(\mathbf{y}^\top)\} \\ &= \boldsymbol{A}\boldsymbol{\Sigma}_y, \end{aligned}$$

  which results in

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \sim N_{d+q}\left(\begin{pmatrix}\boldsymbol{A}\boldsymbol{\mu}_y\\ \boldsymbol{\mu}_y\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{x|y} + \boldsymbol{A}\boldsymbol{\Sigma}_y\boldsymbol{A}^\top & \boldsymbol{A}\boldsymbol{\Sigma}_y\\ \boldsymbol{\Sigma}_y\boldsymbol{A}^\top & \boldsymbol{\Sigma}_y\end{pmatrix}\right). \tag{A.31}$$

  □

- The multinormal generator is rmvnorm(N, mean = $\boldsymbol{\mu}$, cov = $\boldsymbol{\Sigma}$).

A.2.7 Student's t and multivariate t distributions

34• Student's t or t-distribution

- $X \sim t(\mu, \sigma^2, \nu)$, location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$, degree of freedom $\nu > 0$,

$$t(x|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma}\left\{1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}. \tag{A.32}$$

- $E(X) = \mu$ (if $\nu > 1$) and $\mathrm{Var}(X) = \sigma^2\nu/(\nu-2)$ (if $\nu > 2$).
- The case of $\nu = 1$ is called the *Cauchy* distribution.
- When $\mu = 0$ and $\sigma^2 = 1$, it is called the standard t-distribution, denoted by $t(\nu)$.
- The t-distribution is a common heavy–tail alternative to the normal distribution in a robust analysis.

34.1• Generation of t random variable

— If

$$\tau \sim \mathrm{Gamma}(\nu/2, \nu/2) \quad \text{and} \quad X|\tau \sim N(\mu, \tau^{-1}\sigma^2), \tag{A.33}$$

then $X \sim t(\mu, \sigma^2, \nu)$.

— The mixture (A.33) gives an algorithm to generate the t-distribution.

— Let $Z \sim N(0,1)$ and $X \sim t(\mu, \sigma^2, \nu)$, we have the following SR:

$$\frac{X-\mu}{\sqrt{\tau^{-1}\sigma^2}}\Big|\tau \overset{\mathrm{d}}{=} Z \quad \text{or}$$

$$X \overset{\mathrm{d}}{=} \mu + \frac{\sigma Z}{\sqrt{\nu\tau/\nu}} \overset{\mathrm{d}}{=} \mu + \frac{N(0,\sigma^2)}{\sqrt{\chi^2(\nu)/\nu}}. \tag{A.34}$$

— The SR (A.34) results in an alternative method to generate the t-variate.

— The $t(\nu)$ generator is $\mathsf{rt}(N, \nu)$.

35• Multivariate t distribution

- $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, location parameter vector $\boldsymbol{\mu} \in \mathbb{R}^d$, dispersion matrix $\boldsymbol{\Sigma} > 0$, degree of freedom $\nu > 0$,

$$t_d(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})(\sqrt{\nu\pi})^d |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left\{1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right\}^{-\frac{\nu+d}{2}}, \quad \text{(A.35)}$$

  where $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$.

- $E(\mathbf{x}) = \boldsymbol{\mu}$ (if $\nu > 1$) and $\mathrm{Var}(\mathbf{x}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$ (if $\nu > 2$).

- The multivariate t provides a heavy–tail alternative to the multinormal while accounting for correlation among the components of $\mathbf{x}$.

35.1• Generation of multivariate t random vector

— If

$$\tau \sim \mathrm{Gamma}(\nu/2, \nu/2) \quad \text{and} \quad \mathbf{x}|\tau \sim N_d(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Sigma}), \quad \text{(A.36)}$$

then $\mathbf{x} \sim t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$.

— The mixture (A.36) gives a sampling method to generate the multivariate t distribution.

— Equivalently, we have the following SR:

$$\mathbf{x} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \frac{N_d(\mathbf{0}, \boldsymbol{\Sigma})}{\sqrt{\chi^2(\nu)/\nu}},$$

which itself can be served as the definition of the multivariate t-variate.

A.2.8 Wishart and inverse Wishart distributions

36• Wishart distribution

- $\mathbf{X}_{d\times d} = (X_{ij}) \sim \mathrm{Wishart}_d(\boldsymbol{A}, \nu)$, scale matrix $\boldsymbol{A}_{d\times d} > 0$, degree of freedom $\nu > 0$,

$$\mathrm{Wishart}_d(\boldsymbol{X}|\boldsymbol{A}, \nu) = \frac{\{\det(\boldsymbol{X})\}^{\frac{\nu-d-1}{2}} \mathrm{e}^{-0.5\,\mathrm{tr}\,(\boldsymbol{A}^{-1}\boldsymbol{X})}}{2^{\frac{\nu d}{2}} \pi^{\frac{d(d-1)}{4}} \{\det(\boldsymbol{A})\}^{\frac{\nu}{2}} \prod_{i=1}^d \Gamma(\frac{\nu+1-i}{2})},$$

  where $\boldsymbol{X}_{d\times d} = (x_{ij}) > 0$.

- $E(\mathbf{X}) = \nu \boldsymbol{A}$.
- When $d = 1$ and $\boldsymbol{A} = 1$, the Wishart reduces to the chi-square distribution: $\text{Wishart}_1(1, \nu) = \chi^2(\nu)$.
- The Wishart is the conjugate prior for the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ of a multinormal distribution.
- A non-informative distribution is proportional to $\{\det(\boldsymbol{X})\}^{-(d+1)/2}$ obtained as $\nu \to 0$ and $\det(\boldsymbol{A}^{-1}) \to 0$.

36.1• Generation of Wishart random matrix

— When ν is an integer and $\nu \geqslant d$, the multinormal generator can be used to simulate a draw from the Wishart based on the following idea:
$***$ Draw $\mathbf{y}_1, \ldots, \mathbf{y}_\nu \overset{\text{iid}}{\sim} N_d(\mathbf{0}, \boldsymbol{A})$;
$***$ Set $\mathbf{X} = \sum_{i=1}^{\nu} \mathbf{y}_i \mathbf{y}_i^\top$, then $\mathbf{X} \sim \text{Wishart}_d(\boldsymbol{A}, \nu)$.

— The corresponding R codes are given by

```
rwishart <- function(d, A, v) {
    # ================ Aim ============================
    # Generate one random matrix X from Wishart_d(A, v)
    # ================ Input ==========================
    # d = dimension >= 2,
    # v = degree of freedom >= d,
    # A = positive define matrix of d by d
    # ================ Output =========================
    # X = positive define random matrix of d by d
    # ================================================
    D <- diag(eigen(A)$values)
    G <- eigen(A)$vectors      # A= G D G'
    Asqrt <- G %*% sqrt(D) %*% t(G)
                               # Asqrt= G D^{1/2} G'
    Z <- matrix(rnorm(d * v), ncol = v)
    X <- Asqrt %*% Z %*% t(Z) %*% t(Asqrt)
                               # X = A^{1/2} ZZ' A^{1/2}
    return(X)
}
```

— Non-integral ν requires the general algorithm originally proposed by Odell & Feiveson (1966).

37• Inverse wishart distribution

- $\mathbf{X}_{d\times d} = (X_{ij}) \sim \text{IWishart}_d(\boldsymbol{A}^{-1}, \nu)$, scale matrix $\boldsymbol{A}_{d\times d} > 0$, degree of freedom $\nu > 0$,

$$\text{IWishart}_d(\boldsymbol{X}|\boldsymbol{A}^{-1}, \nu) = \frac{\{\det(\boldsymbol{X})\}^{-\frac{\nu+d+1}{2}} \, \mathrm{e}^{-0.5\,\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{X}^{-1})}}{2^{\frac{\nu d}{2}} \pi^{\frac{d(d-1)}{4}} \{\det(\boldsymbol{A}^{-1})\}^{\frac{\nu}{2}} \prod_{i=1}^{d} \Gamma(\frac{\nu+1-i}{2})},$$

  where $\boldsymbol{X}_{d\times d} = (x_{ij}) > 0$.

- $E(\mathbf{X}) = (\nu - d - 1)^{-1}\boldsymbol{A}$.

- When $d = 1$ and $\boldsymbol{A}^{-1} = 1$, the inverse Wishart reduces to the inverse chi-square distribution: $\text{IWishart}_1(1, \nu) = \text{I}\chi^2(\nu)$.

- If $\mathbf{X}^{-1} \sim \text{Wishart}_d(\boldsymbol{A}^{-1}, \nu)$, then $\mathbf{X} \sim \text{IWishart}_d(\boldsymbol{A}^{-1}, \nu)$.

  Proof: This can be verified by using the following fact:

$$\text{If } \boldsymbol{X}_{d\times d} = \boldsymbol{X}^{\top}, \quad \text{then } J(\boldsymbol{X}^{-1} \to \boldsymbol{X}) = \{\det(\boldsymbol{X})\}^{-(d+1)},$$

  where $J(\boldsymbol{Y} \to \boldsymbol{X}) = \partial\boldsymbol{Y}/\partial\boldsymbol{X}$ is the Jacobian determinant. □

- The inverse Wishart is the conjugate prior for the covariance matrix $\boldsymbol{\Sigma}$ in a multinormal distribution.

- A non-informative distribution is proportional to $\{\det(\boldsymbol{X})\}^{-(d+1)/2}$ obtained as $\nu \to 0$ and $\det(\boldsymbol{A}) \to 0$.

A.3 Stochastic Processes

A.3.1 Homogeneous Poisson process

38• Definition and property

- $\{X_i, i \geqslant 1\} \sim \text{HPP}(\lambda)$, where $\lambda > 0$ denotes the rate. A counting process $\{N(t), t \geqslant 0\}$ with successive event times $0 < X_1 < X_2 < \cdots$ is said to be a *homogeneous Poisson process* (HPP) with rate λ, if

  (a) $N(0) = 0$;

(b) The process has stationary and independent increments;

(c) $\Pr\{N(h) \geqslant 2\} = o(h)$;

(d) $\Pr\{N(h) = 1\} = \lambda h + o(h)$.

38.1• Property

— Let $\{X_i, i \geqslant 1\} \sim \text{HPP}(\lambda)$, then

(1) For all $s, t \geqslant 0$, $N(t+s) - N(s) = N(t) \sim \text{Poisson}(\lambda t)$;

(2) $\{X_i - X_{i-1}\} \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$, $X_0 \equiv 0$;

(3) The joint pdf of $X_1, \ldots, X_n$ is $\lambda^n \, \mathrm{e}^{-\lambda x_n}$;

(4) Given that $N(t) = n$, X_i $(i = 1, \ldots, n)$ has the same distribution as the i-th order statistic of n i.i.d. samples from $U(0, t)$.

— The second property can be used to generate all the event times occurring in $(0, t]$ of a Poisson process with rate λ.

A.3.2 Nonhomogeneous Poisson process

39• Definition and property

- $\{X_i, i \geqslant 1\} \sim \text{NHPP}(\lambda(t), t \geqslant 0)$ or $\{X_i, i \geqslant 1\} \sim \text{NHPP}(m(t), t \geqslant 0)$, where $\lambda(t)$ and $m(t) \triangleq \int_0^t \lambda(s)\,\mathrm{d}s$ denote the *intensity* function and the *mean* function, respectively. A counting process $\{N(t), t \geqslant 0\}$ is said to be a *nonhomogeneous Poisson process* (NHPP) with intensity $\lambda(t)$, if

  (a) $N(0) = 0$;

  (b) The process has independent increments;

  (c) $\Pr\{N(t+h) - N(t) \geqslant 2\} = o(h)$;

  (d) $\Pr\{N(t+h) - N(t) = 1\} = \lambda(t)h + o(h)$.

39.1• Property

— Let $\{X_i, i \geqslant 1\} \sim \text{NHPP}(m(t), t \geqslant 0)$, then

(1) $N(t+s) - N(s) \sim \text{Poisson}(m(t+s) - m(t))$;

(2) The joint pdf of $X_1, \ldots, X_n$ is

$$\left\{\prod_{i=1}^{n} \lambda(x_i)\right\} \mathrm{e}^{-m(T)},$$

where $T = x_n$ if it is failure truncated and $T = t$ if it is time truncated;

(3) Conditional on $X_n = x_n$, $X_1 < \cdots < X_{n-1}$ are distributed as $n-1$ order statistics from the following cdf

$$G(x) = 0 \cdot I(x \leqslant 0) + \frac{m(x)}{m(x_n)} \cdot I(0 < x \leqslant x_n) + 1 \cdot I(x > x_n).$$

(4) Conditional on $N(t) = n$, $X_1 < \cdots < X_n$ are distributed as n order statistics from the following cdf

$$G(x) = 0 \cdot I(x \leqslant 0) + \frac{m(x)}{m(t)} \cdot I(0 < x \leqslant t) + 1 \cdot I(x > t).$$

Appendix B

R Programming

1• What is R?

- R is a statistical computer program, made available through the Internet under the general public license.
- R provides an environment in which you can perform statistical analysis and produce graphics.
- It is designed in such a way that it is always possible to do further computations on the results of a statistical procedure.

1.1• R is free

— R is supplied with a license that allows you to use it freely, distribute it, or even sell it, as long as the receiver has the same rights and the source code is freely available.

— It is actually a complete programming language.

— Here we only learn the elementary concepts and see a number of cookbook examples.

1.2• From language C, language S, S-plus to R

— R owes its name to typical internet humor.

— You may be familiar with the programming language C (whose name is a story in itself).

— Inspired by this, Becker and Chambers chose in the early 1980s to call their newly developed statistical programming language S.

— Language S was further developed into the commercial product S-plus, which by the end of the decade was in widespread use among statisticians of all kinds.

— Ross Ihaka and Robert Gentleman from the University of Auckland, New Zealand, chose to write a reduced version of S for teaching purpose.

— In 1995, Martin Maechler persuaded Ross and Robert to release the source codes for R under the general public license.

1.3• Difference between R and S-plus

— R implements a dialect of the S language.

— There are some differences, but in everyday use the two are very similar.

— However, some functions do differ, often because the R version tries to simplify things for the user. The differences are not all that big.

2• How to obtain R?

- The way to obtain R is to download it from one of the CRAN (Comprehensive R Archive Network) sites. The main site is

  http://cran.r.-project.org/

- It has a number of mirror sites worldwide, which may be closer to you and give faster download times.

- Installation details tend to vary over time, so you should read the accompanying documents and any other information offered on CRAN.

- Information and further internet resources for R can be obtained from the R home-page at www.r-project.org

B.1 Basic Commands

B.1.1 Expressions

3• Numeric expressions

- When R is ready for input, it prints out its prompt, a '>'.

- One of the simplest possible tasks in R is to enter an expression and receive a result.

Table B.1 Arithmetic operators

Operator	Meaning	Expression	Result
+	plus, addition	$4+3$	7
−	minus, subtraction, sign	$9-5$	4
*	times, multiplication	$3*5$	15
/	division	7/3	2.3333
		8/3	2.6667
`% / %`	integer division	7 %/% 3	2
		8 %/% 3	3
`^`	power	2^3	8

Table B.2 Commonly used functions

R function	Meaning
`sqrt()`	square root
`log()`	natural logarithm
`log10()`	logarithm base 10
`exp()`	exponential
`abs()`	absolute value
`round()`	round to nearest integer
`ceiling()`	round up
`floor()`	round down
`sin(), cos(), tan()`	sine, cosine, tangent
`asin(), acos(), atan()`	arc-sine, arc-cosine, arc-tangent
`min(x)`	smallest value in vector `x`
`min(x1, x2, ...)`	minimum over several vectors (one number)
`pmin(x1, x2, ...)`	parallel (element-wise) minimum over multiple equally long vectors
`max(x)`	largest value in vector `x`
`max(x1, x2, ...)`	maximum over several vectors (one number)
`pmax(x1, x2, ...)`	parallel (element-wise) maximum
`range(x)`	like `c(min(x), max(x))`
`length(x)`	number of elements in vector `x`

- For example, the computer can perform 2 plus 2 making 4 (the second line is the answer from the computer):

```
> 2 + 2
 [1] 4
```

- Of course, it also knows how to do other standard calculations.
- For example, the following is how to compute e^{-2}:

```
> exp(-2)
[1] 0.13533528
```

3.1• Arithmetic operators and commonly used functions

— Arithmetic operators are presented in Table B.1.

— Some commonly used functions are listed in Table B.2.

3.2• The `options()` and `help()` functions

— The options() function can be used to control the appearance of the output, e.g.,

```
============================================================
> options(width=68, digits=8)
> pi
[1] 3.1415927
> options(width=68, digits=4)
> pi
[1] 3.142
> -5/3
[1] -1.667
> 1/3
[1] 0.3333
> -1/3
[1] -0.3333
************************************************************
```

— The [1] in front of the result is part of R's way of printing numbers and vectors.

— It is useless here, but it becomes useful when the result is a longer vector.

— Consider the case of generating 10 random numbers from uniform distribution on $(0, 1)$:

```
> runif(10)
 [1] 0.050808 0.195130 0.391954 0.300020 0.143770 0.895648
 [7] 0.031605 0.723146 0.528792 0.887409
```

— Here the `[7]` indicates that 0.031605 is the seventh element in the vector.

— More information about any R functions can be found using the help() function.

```
> help(t.test)
> ?t.test
> ??t.test
```

4• Logical expressions

- So far, we have mentioned values of type numeric.
- When a numeric value is missing, it is of type `NA`, i.e., not available.
- Another type in R is logical with three values, `TRUE` (or its abbreviation `T`), `FALSE` (or `F`), and `NA`.
- Logical operations are extremely useful when making comparisons and choosing particular elements from vectors and matrices.

Table B.3 Logical operators

Operator	Meaning
`<`	less than
`>`	greater than
`<=`	less than or equal to
`>=`	greater than or equal to
`==`	equal to
`!=`	not equal to
`&`	and
`|`	or
`!`	not
`is.na(x)`	missing?

- The symbols used for logical operations are listed in Table B.3.
- We can use a logical expression to assign a logical value.

```
============================================================
> x <- 3 == 4
> x
[1] FALSE
> x <- 3 < 4
> x
[1] TRUE
> 3 == 4 & 3 < 4
[1] FALSE
> 3 == 4 | 3 < 4
[1] TRUE
> 1/0
[1] Inf
> is.numeric(3)
[1] TRUE
> is.character("3")
[1] TRUE
> is.infinite(1/0)
[1] TRUE
------------------------------------------------------------
> x <- 1:15
> x[x < 10]
[1] 1 2 3 4 5 6 7 8 9
> y <- c(rep(0, 10), rep(1, 5))
> x[y == 0]
 [1]  1  2  3  4  5  6  7  8  9 10
************************************************************
```

B.1.2 Assignment operator

5• Assignment statement

- To assign the value 2 to the variable x, you can input

```
> x <- 2
```

- The two characters <- should be read as a single symbol: an arrow pointing to the variable to which the value is assigned.
- There is no immediately visible result, but from now on, x has the value 2 and can be used in subsequent calculations. For instance,

```
> x*x
[1] 4
```

- Assignment can also be made using the function assign().
- Assignments can also be made in the other direction.

```
============================================================
> assign("x", c(10.4, 5.6, 3.1, 6.4, 21.7))
> x
[1] 10.4  5.6  3.1  6.4 21.7
------------------------------------------------------------
>  1:4 -> y
> y
[1] 1 2 3 4
------------------------------------------------------------
> x1 <- 1; x2 <- -2; x3 <- 4
> c(x1, x2, x3)
[1]  1 -2  4
------------------------------------------------------------
> a <- b <- c <- 2
> c(a, b, c)
[1] 2 2 2
************************************************************
```

6• Names of variables

- Names of variables in R can be built from letters, digits and the period (dot) symbol.
- However, names must not start with a digit and avoid starting with period.

- For example, `height.2yr` may be used to describe the height of a child at the age of 2 years.
- Names are case-sensitive: `WT` and `wt` do not refer to the same variable.
- Some names, e.g., `c, q, t, C, D, F, I, T, diff, df, pt` are already used/defined by the system.

B.2 Vectors and Matrices

B.2.1 Vectors

7• Numeric vectors

7.1• The colon operator ":"

— Many methods can be used to generate vectors in R.

— The simplest way is to use the colon operator.

— The colon operator has high priority within an expression.

```
============================================================
> x <- 1:5
> x
[1] 1 2 3 4 5
> 5:1
[1] 5 4 3 2 1
> 2*1:5
[1]  2  4  6  8 10
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> is.vector(5:1)
[1] TRUE
> is.vector(5:1, mode="integer")
[1] TRUE
> is.vector(5:1, mode="numeric")
[1] TRUE
> is.vector(5:1, mode="character")
[1] FALSE
> is.matrix(5:1)
[1] FALSE
```

```
------------------------------------------------------------
> n <- 10
> 1:n-1                                        # = (1:n) - 1
 [1] 0 1 2 3 4 5 6 7 8 9
> 1:(n-1)
[1] 1 2 3 4 5 6 7 8 9
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> is.numeric(n)
[1] TRUE
> is.character(n)
[1] FALSE
************************************************************
```

7.2• Concatenate function `c()`

— `c()` function is the second way to generate a (column) vector.

— The number of elements in a vector can be determined using the `length()` function.

— The `t()` function transposes an n-dimensional column vector into a $1 \times n$ matrix.

— In R, there is no row vector.

```
============================================================
> x <- c(1,2,3,4)
> x
[1] 1 2 3 4
> length(x)
[1] 4
> tx <- t(x)
> tx
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
> is.vector(tx)
[1] FALSE
> is.matrix(tx)
[1] TRUE
************************************************************
```

7.3• Sequence function `seq()`

— The general syntax of `seq()` is

```
============================================================
seq(from= 1, to = 1, by = (to - from)/(length.out - 1),
                       length.out = NULL)

# Typical usages are

seq(from, to)
seq(from, to, by= )
seq(from, to, length.out= )
************************************************************
```

— If `by` is 1, the `seq()` function can be replaced by `from:to`.

```
============================================================
> seq(from= 1, to= 4)
[1] 1 2 3 4
> seq(1, 4)
[1] 1 2 3 4
> seq(from= 1, to= 4, by= 1)
[1] 1 2 3 4
> 1:4
[1] 1 2 3 4
> seq(-10, 0, 1)
 [1] -10  -9  -8  -7  -6  -5  -4  -3  -2  -1   0
------------------------------------------------------------
> seq(-pi, pi, length= 10)  # length = length.out
 [1] -3.14159 -2.44346 -1.74533 -1.04720 -0.34907
 [6]  0.34907  1.04720  1.74533  2.44346  3.14159
------------------------------------------------------------
> seq(0, 1, by= 0.1)
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> seq(0, 1, length.out = 11)
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
************************************************************
```

7.4• Repeat function `rep()`

— The general syntax of `rep()` is

```
rep(x, times= 1, length.out= NA, each= 1)
```

```
============================================================
> rep(10, times= 5)
[1] 10 10 10 10 10
> rep(1:3, 3)
[1] 1 2 3 1 2 3 1 2 3
> rep(1:3, c(3,3,3))
[1] 1 1 1 2 2 2 3 3 3
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> rep(1:3, c(1,2,3))
[1] 1 2 2 3 3 3
> rep(c(2,3,4,5), 1:4)
 [1] 2 3 3 4 4 4 5 5 5 5
------------------------------------------------------------
> rep(1:4, 2)
[1] 1 2 3 4 1 2 3 4
> rep(1:4, each= 2)
[1] 1 1 2 2 3 3 4 4
> rep(1:4, c(2,2,2,2))
[1] 1 1 2 2 3 3 4 4
> rep(1:4, c(2,1,2,1))
[1] 1 1 2 3 3 4
> rep(1:4, each= 2, len= 4)         # first 4 only
[1] 1 1 2 2
> rep(1:4, each= 2, length.out= 4)  # length.out = len
[1] 1 1 2 2
> rep(1:4, each = 2, len = 10)      # 8 integers plus two
 [1] 1 1 2 2 3 3 4 4 1 1            # recycled 1's
> rep(1:4, each= 2, times= 3)
 [1] 1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4
************************************************************
```

7.5• Obtaining a sub-vector by square brackets

```
================================================================
> x <- 1:6
> x[3]            # [1] 3
> x[c(1,3)]
[1] 1 3
> x[1:3]
[1] 1 2 3
> x[-2]
[1] 1 3 4 5 6
> x[-c(2,4)]
[1] 1 3 5 6
----------------------------------------------------------------
> y <- c(3, 3, 3, 3, 3, 3)
> x[x>y]
[1] 4 5 6
> x[x==y]
[1] 3
> x[x!=y]
[1] 1 2 4 5 6
****************************************************************
```

7.6• Missing values

— Missing data are frequently encountered in practice (e.g., some patient withdrew from a study; an experiment failed).

— In R, missing value is denoted by `NA`.

— Operations on `NA` yield `NA` as the result.

```
============================================================
> x <- c(1, 2, 3, NA, 5)
> x
[1]  1  2  3 NA  5
> x+1
[1]  2  3  4 NA  6
> x*4
[1]  4  8 12 NA 20
```

```
> x*0
[1]  0  0  0 NA  0
------------------------------------------------------------
> y <- c(1, NA, 3, 4, 5)
> x+y
[1]  2 NA  6 NA 10
> x*y
[1]  1 NA  9 NA 25
------------------------------------------------------------
> is.na(x)
[1] FALSE FALSE FALSE  TRUE FALSE
> is.vector(x)
[1] TRUE
> is.vector(x, mode="integer")
[1] FALSE
> is.vector(x, mode="character")
[1] FALSE
> is.vector(x, mode="numeric")
[1] TRUE
************************************************************
```

8• Character vectors

8.1• Concatenate function `c()`

— A character vector is a vector of (text) strings, whose elements are specified and printed in double quotes.

— It also works with a mixture of numeric and string values, but in this case, all elements will be converted to strings

```
============================================================
> BMN <- c("Brain", "Mouth", "Nose")
> BMN
[1] "Brain" "Mouth" "Nose"
> > is.character(BMN)
[1] TRUE
> is.vector(BMN)
[1] TRUE
```

```
> is.vector(BMN, mode="character")
[1] TRUE
------------------------------------------------------------
> mix <- c(BMN, 45, -20)
> mix
[1] "Brain" "Mouth" "Nose"  "45"    "-20"
************************************************************
```

8.2• The `paste()` function

— The `paste()` function takes an arbitrary number of arguments and concatenates them one by one into character strings.

— The general syntax of `paste()` is

```
paste (..., sep = " ", collapse = NULL)
```

```
============================================================
>  paste(c("X","Y"), 1:4)
[1] "X 1" "Y 2" "X 3" "Y 4"

> paste(c("X","Y"), 1:4, sep="")
[1] "X1" "Y2" "X3" "Y4"

> paste(c("X","Y"), 1:4, sep="_")
[1] "X_1" "Y_2" "X_3" "Y_4"
------------------------------------------------------------
> x <- c("st", "nd", "rd", "th", "th")
> paste(1:5, x, sep="-")
[1] "1-st" "2-nd" "3-rd" "4-th" "5-th"

> paste(1:5, x, sep= "-", collapse =" ")
[1] "1-st 2-nd 3-rd 4-th 5-th"

> paste(1:5, x, sep= "-", collapse =", ")
[1] "1-st, 2-nd, 3-rd, 4-th, 5-th"

> paste(1:5, x, sep= "-", collapse =" | ")
[1] "1-st | 2-nd | 3-rd | 4-th | 5-th"
```

```
------------------------------------------------------------
> p <- 0.03
> paste("The p-value = ", p)
[1] "The p-value =  0.03"
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> tv <- 2.14; pv <- 0.03
> paste(c("The t-value is ", "The p-value = "), c(tv, pv))
[1] "The t-value is  2.14" "The p-value =  0.03"

> paste(c("The t-value is ", "the p-value = "), c(tv, pv),
                                          collapse =" and ")
[1] "The t-value is  2.14 and the p-value =  0.03"
************************************************************
```

9• Logical vectors

- The elements of a logical vector can have the values `TRUE` (or its abbreviation `T`), `FALSE` (or `F`), and `NA`.

```
============================================================
> c(T, T, F, T)
[1]  TRUE  TRUE FALSE  TRUE
------------------------------------------------------------
> x <- c(1, 2, 3, NA, 5)
> is.na(x)
[1] FALSE FALSE FALSE  TRUE FALSE
> is.vector(is.na(x), mode="logical")
[1] TRUE
> !is.na(x)
[1]  TRUE  TRUE  TRUE FALSE  TRUE
> x[!is.na(x)]      # A vector containing the non-missing
[1] 1 2 3 5         # values of x
------------------------------------------------------------
> x <- c(-1, 2, 3, NA, -5, 6)
> x[x>0]
[1]  2  3 NA  6
> x[(!is.na(x)) & x>0]
[1] 2 3 6
```

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> x+1
[1]  0  3  4 NA -4  7
> (x+1)[(!is.na(x)) & x>0]
[1] 3 4 7
# A sub-vector of x+1 with those elements, where the
# corresponding elements in x are non-missing & positive
**********************************************************
```

B.2.2 Matrices

10• Dimension function dim()

- A matrix in mathematics is just a two-dimensional array of *numbers*.
- In R, the matrix notation is extended to elements of any type, e.g., a matrix of *character strings*.
- A matrix is represented as a vector with dimensions:

```
==========================================================
> M1 <- 1:15
> dim(M1) <- c(3, 5)       # The storage is column-wise
> M1
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> M1[, c(5, 4, 3, 1, 2)]       # no change in rows
     [,1] [,2] [,3] [,4] [,5]  # change order of columns
[1,]   13   10    7    1    4
[2,]   14   11    8    2    5
[3,]   15   12    9    3    6
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> M1[c(3, 2, 1), ]             # exchange row 3 with
     [,1] [,2] [,3] [,4] [,5]  # row 1
[1,]    3    6    9   12   15
[2,]    2    5    8   11   14
[3,]    1    4    7   10   13
```

```
-----------------------------------------------------------
> M2 <- c("a1", "a2", "a3", "b1", "b2", "b3",
                            "c1", "c2", "c3")
> dim(M2) <- c(3, 3)
> M2
     [,1] [,2] [,3]
[1,] "a1" "b1" "c1"
[2,] "a2" "b2" "c2"
[3,] "a3" "b3" "c3"
-----------------------------------------------------------
> M3 <- c(1:3, "a1", "a2", "a3")   # This why we need
> M3                               # data.frame()
[1] "1"  "2"  "3"  "a1" "a2" "a3"
> dim(M3) <- c(3, 2)
> M3
     [,1] [,2]
[1,] "1"  "a1"
[2,] "2"  "a2"
[3,] "3"  "a3"
***********************************************************
```

11• Matrix function `matrix()`

- The general syntax of `matrix()` is

  `matrix(data, nrow, ncol, byrow= F, dimnames= NULL).`

- `byrow = T` specifies that the matrix is to be filled row by row and `byrow = F` is the default.

```
===========================================================
> matrix(1:6, nrow = 2, byrow = T)        #list by rows
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
> matrix(1:6, nrow = 2)
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```
------------------------------------------------------------
> X <- matrix(1:6, 2, 3)      #list by columns (default)
> X
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> rownames(X)<- LETTERS[1:2]
> colnames(X)<- letters[1:3]
> X
  a b c
A 1 3 5
B 2 4 6
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> rownames(X) <- month.name[1:2]
> colnames(X) <- month.abb[1:3]
> X
         Jan Feb Mar
January    1   3   5
February   2   4   6
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> rownames(X) <- c("R1", "R2")
> colnames(X) <- c("C1", "C2", "C3")
> X
   C1 C2 C3
R1  1  3  5
R2  2  4  6
------------------------------------------------------------
> Y <- matrix(1:6, nrow= 2, dimnames= list(c("R1", "R2"),
                                    c("C1", "C2", "C3")))
> Y
   C1 C2 C3
R1  1  3  5
R2  2  4  6
************************************************************
```

12• Merging vectors/matrices by rbind() and cbind()

```
================================================================
> M1 <- rbind(A= 1:4, B= 5:8, C= 9:12)
> M1
  [,1] [,2] [,3] [,4]
A    1    2    3    4
B    5    6    7    8
C    9   10   11   12
> M2 <- cbind(a= -(1:3), b= -(4:6), c= -(7:9))
> M2
      a  b  c
[1,] -1 -4 -7
[2,] -2 -5 -8
[3,] -3 -6 -9
> rownames(M2) <- LETTERS[1:3]
> M2
   a  b  c
A -1 -4 -7
B -2 -5 -8
C -3 -6 -9
----------------------------------------------------------------
> M <- cbind(M1, M2)
> M
               a  b  c
A 1  2  3  4 -1 -4 -7
B 5  6  7  8 -2 -5 -8
C 9 10 11 12 -3 -6 -9
> colnames(M)[1:4] <- month.abb[1:4]
> M
  Jan Feb Mar Apr  a  b  c
A   1   2   3   4 -1 -4 -7
B   5   6   7   8 -2 -5 -8
C   9  10  11  12 -3 -6 -9
----------------------------------------------------------------
> cbind(1, 1:3)
     [,1] [,2]
[1,]    1    1
[2,]    1    2
```

```
[3,]    1    3
> cbind(1:3, diag(3))
     [,1] [,2] [,3] [,4]
[1,]    1    1    0    0
[2,]    2    0    1    0
[3,]    3    0    0    1
**************************************************************
```

13• Obtaining a sub-matrix by square brackets

```
==============================================================
> X
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
--------------------------------------------------------------
> X[2, 3]      # X[i,j] is the (i,j)-th element of X
[1] 6
> X[1, ]       # X[i, ] is the vector of the i-th row of X
[1] 1 3 5
> X[, 2]       # X[, j] is the vector of the j-th column of X
[1] 3 4
--------------------------------------------------------------
> X * X               # multiply element by element
     [,1] [,2] [,3]
[1,]    1    9   25
[2,]    4   16   36
> X %*% t(X)          # matrix multiplication
     [,1] [,2]
[1,]   35   44
[2,]   44   56
--------------------------------------------------------------
> dim(X)              # dimension function
[1] 2 3
> nrow(X)             # number of rows of X
[1] 2
> ncol(X)             # number of columns of X
[1] 3
**************************************************************
```

14• The diag() function

- If `n` is a single numeric value, then `diag(n)` is the `n` by `n` identity matrix.
- If `v` is a vector, then `diag(v)` gives a diagonal matrix.
- If `M` is a matrix, then `diag(M)` gives the vector of main diagonal entries of `M`.

```
============================================================
> diag(3)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
------------------------------------------------------------
> diag(1:3)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    2    0
[3,]    0    0    3
------------------------------------------------------------
> M <- matrix(1:9, 3, 3)
> M
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> diag(M)
[1] 1 5 9
------------------------------------------------------------
> diag(diag(M))
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    5    0
[3,]    0    0    9
************************************************************
```

15• The crossprod() function

- `crossprod(X, y)` is the same as `t(X) %*% y` but the operation is more efficient.
- If the second argument to `crossprod()` is omitted, it is taken to be the same as the first.

```
============================================================
> x <- 1:3
> crossprod(x)          # = crossprod(x, x) = t(x) %*% x
     [,1]
[1,]   14
> t(x) %*% x
     [,1]
[1,]   14
> x %*% t(x)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    4    6
[3,]    3    6    9
************************************************************
```

16• Linear equations and matrix inversion

- Let $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, then $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ is the solution of the linear equations.
- In R, we use `x <- solve(A, b)`.
- Although, we can compute `x <- solve(A) %*% b`, it is inefficient and potentially unstable.
- The quadratic form $\boldsymbol{x}^\top \boldsymbol{A}^{-1}\boldsymbol{x}$ should be computed as `x %*% solve(A,x)`.

17• Eigenvalues and eigenvectors for a symmetric matrix

- Let $\boldsymbol{A} = (a_{ij})$ be an $n \times n$ symmetric matrix, then $\boldsymbol{A} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top$ or $\boldsymbol{A}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}$, where
  - $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$,
  - $\{\lambda_1, \ldots, \lambda_n\}$ are eigenvalues of $\boldsymbol{A}$,

  - $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$ is orthogonal satisfying $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \boldsymbol{I}_n$, and
  - $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n\}$ are corresponding eigenvectors of $\boldsymbol{A}$.

- The function `eigen(A)` calculates the eigenvalues and eigenvectors of a symmetric matrix `A`.
  - The result of this function is a list of two components named `values` and `vectors`.
- The function `det(A)` computes the determinant of an arbitrary square matrix $\boldsymbol{A}$.
  - However, for a symmetric matrix $\boldsymbol{A}$, we have $\det(\boldsymbol{A}) = \prod_{i=1}^n \lambda_i$.
- The trace of $\boldsymbol{A}$ is defined as $\mathrm{tr}\,(\boldsymbol{A}) = \sum_{i=1}^n a_{ii}$ for any square matrix $\boldsymbol{A}$.
  - Note that, there is no `tr(A)` function in R because it is simply computed as `sum(diag(A))`.
  - However, for the symmetric matrix $\boldsymbol{A}$, we have $\mathrm{tr}\,(\boldsymbol{A}) = \sum_{i=1}^n \lambda_i$.

```
============================================================
> X <- matrix(rnorm(9), 3, 3)
> A <- t(X) %*% X
> ev <- eigen(A)
> ev
$values
[1] 2.2833 0.8654 0.4781

$vectors
         [,1]     [,2]     [,3]
[1,] -0.08887  0.92143  0.37825
[2,]  0.02771  0.38190 -0.92379
[3,] -0.99566 -0.07161 -0.05947
------------------------------------------------------------
> L <- diag(ev$values); G <- ev$vectors;
> A %*% G                          #  Checking: A G = G L
         [,1]     [,2]     [,3]
[1,] -0.20291  0.79736  0.18085
[2,]  0.06326  0.33048 -0.44169
```

```
[3,] -2.27335 -0.06197 -0.02843
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> G %*% L
         [,1]     [,2]     [,3]
[1,] -0.20291  0.79736  0.18085
[2,]  0.06326  0.33048 -0.44169
[3,] -2.27335 -0.06197 -0.02843
------------------------------------------------------------
> G %*% t(G)                              # Checking: G G'
          [,1]        [,2]       [,3]  #  = G' G = I_n
[1,] 1.000e+00  1.110e-16  5.551e-17
[2,] 1.110e-16  1.000e+00 -6.939e-18
[3,] 5.551e-17 -6.939e-18  1.000e+00
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> t(G) %*% G
           [,1]       [,2]      [,3]
[1,]  1.000e+00 -4.163e-17 1.388e-17
[2,] -4.163e-17  1.000e+00 5.985e-17
[3,]  1.388e-17  5.985e-17 1.000e+00
------------------------------------------------------------
> det(A)                        # = 2.2833 *0.8654 *0.4781
[1] 0.9447
> prod(ev$values)               # = 2.2833 *0.8654 *0.4781
[1] 0.9447
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> sum(diag(A))                  # tr(A)
[1] 3.627
> sum(ev$values)
[1] 3.627
************************************************************
```

B.3 Lists, Data Frames and Arrays

B.3.1 Lists

18• Why need we `list()` besides vectors and matrices?

- An R `list()` is an object consisting of a collection of objects known as its *components*.

- The components could consist of a numeric vector, a logical value, a matrix, a complex vector, a character array, a function, and so on.
- Vectors and matrices are not enough to store such data. For example, the outcome of `eigen()` is a list of two components: a vector of eigenvalues and a matrix of eigenvectors.

```
============================================================
> L <- list(husband= "Fred", wife= "Mary",
              number.children= 3, child.ages= c(4,7,9) )
> L                           # is the name of this list
$husband                      # with four components
[1] "Fred"

$wife
[1] "Mary"

$number.children
[1] 3

$child.ages
[1] 4 7 9
------------------------------------------------------------
> length(L)     # gives the number of components
[1] 4
------------------------------------------------------------
> L$husband     # = L[[1]] = L[["husband"]]
[1] "Fred"      # name of the 1-st component of the list
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> L$wife        # = L[[2]] = L[["wife"]]
[1] "Mary"      #    [[ ]] : double square brackets
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> L$child.ages  # = L[[4]]
[1] 4 7 9
> L[[4]][1]     # the 1-st element of the vector L[[4]]
[1] 4
------------------------------------------------------------
> L[1]             # a sub-list with the first component
$husband
```

```
[1] "Fred"
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> L[4]
$child.ages
[1] 4 7 9
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> L[c(1, 2)]       # a sub-list with two components
$husband
[1] "Fred"

$wife
[1] "Mary"
**********************************************************
```

19• Forming a new list from existing lists via c()

```
==============================================================
> La <- list(score.child= c(80, 90, 100), university.child=
                                        c("U1", "U2", "U3"))
> La
$score.child
[1]  80  90 100

$university.child
[1] "U1" "U2" "U3"
--------------------------------------------------------------
> L.new <- c(L, La)
> L.new
$husband
[1] "Fred"

$wife
[1] "Mary"

$number.children
[1] 3

$child.ages
```

```
[1] 4 7 9

$score.child
[1]  80  90 100

$university.child
[1] "U1" "U2" "U3"
**************************************************************
```

B.3.2 Data frames

20• Why need we `data.frame()` besides matrices?

- We have only three kinds of matrix: numeric matrix, character matrix, and logical matrix.
- A data frame, a matrix-like structure whose columns may be of differing types (numeric, character, logical, factor and so on).

21• Data entry

21.1• Creating data frame from pre-existing variables

```
==============================================================
> sex <- c(rep("F", 3), rep("M", 3))
> sex
[1] "F" "F" "F" "M" "M" "M"
> y <- 1:6
--------------------------------------------------------------
> d <- data.frame(y= y, sex= sex)
> d
  y sex
1 1   F
2 2   F
3 3   F
4 4   M
5 5   M
6 6   M
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> d$y
```

```
[1] 1 2 3 4 5 6
> d$sex          # character vectors is coerced to be factors
[1] F F F M M M
Levels: F M
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
> is.data.frame(d)
[1] TRUE
> is.vector(d$y, mode="numeric")
[1] TRUE
> is.vector(d$sex, mode="character")
[1] FALSE
> is.factor(d$sex)
[1] TRUE
----------------------------------------------------------------
> d$z <- 6:1               # add a new variable
> d
  y sex z
1 1   F 6
2 2   F 5
3 3   F 4
4 4   M 3
5 5   M 2
6 6   M 1
> d <- d[c(1, 3, 2)]      # insert a new variable
> d
  y z sex
1 1 6   F
2 2 5   F
3 3 4   F
4 4 3   M
5 5 2   M
6 6 1   M
> d <- d[-2]              # delete the 2-nd column
> d
  y sex
1 1   F
2 2   F
3 3   F
```

```
4 4   M
5 5   M
6 6   M
**************************************************************
```

21.2• The data-frame editor for small data sets

— To enter data into a blank data frame, use

```
==============================================================
> dd <- data.frame()
> fix(dd)
**************************************************************
```

— This brings up a spreadsheet-like editor.

— An alternative would be `dd <- edit(data.frame())`.

22• Indexing of data frames

```
==============================================================
> d
     y sex
1  1.2   F
2  3.0   F
3  2.5   F
4 -2.6   M
5 10.0   M
6  7.0   M
--------------------------------------------------------------
> d[5, 1]
[1] 10
> d[5, 2]
[1] M
Levels: F M
> d[5, ]
   y sex
5 10   M
> d[, 2]
```

```
[1] F F F M M M
Levels: F M
-----------------------------------------------------------------
> d[d$y>2, ]
     y sex
2  3.0   F
3  2.5   F
5 10.0   M
6  7.0   M
*****************************************************************
```

23• subset AND transform

```
=================================================================
> d
     y sex
1  1.2   F
2  3.0   F
3  2.5   F
4 -2.6   M
5 10.0   M
6  7.0   M
-----------------------------------------------------------------
> d2 <- subset(d, d$y>2)        # delete some rows
> d2
     y sex
2  3.0   F
3  2.5   F
5 10.0   M
6  7.0   M
-----------------------------------------------------------------
> d3 <- transform(d, z= y*y)    # add a row named as z
> d3
     y sex      z
1  1.2   F   1.44
2  3.0   F   9.00
3  2.5   F   6.25
4 -2.6   M   6.76
```

```
5 10.0   M 100.00
6  7.0   M  49.00
*************************************************************
```

B.3.3 Arrays

24• Dimension function `dim()`

- A vector is a 1-dimensional array.
- A matrix is a 2-dimensional array.
- The following is an example of 3–dimensional array.

```
=============================================================
> z <- 1:24
> dim(z) <- c(2, 4, 3)
> z
, , 1

     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8

, , 2

     [,1] [,2] [,3] [,4]
[1,]    9   11   13   15
[2,]   10   12   14   16

, , 3

     [,1] [,2] [,3] [,4]
[1,]   17   19   21   23
[2,]   18   20   22   24
*************************************************************
```

25• Array function array()

```
============================================================
> Z <- array(1:24, dim= c(3, 4, 2))
> Z
, , 1

     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12

, , 2

     [,1] [,2] [,3] [,4]
[1,]   13   16   19   22
[2,]   14   17   20   23
[3,]   15   18   21   24
************************************************************
```

B.4 Flow Control

26• while statement

- The R allows conditional execution and looping constructs.
- Note that `while (condition) expression` construction, which says that the expression should be evaluated as long as the condition is `TRUE`.
- For example, suppose that we want to use a version of Newton's method for calculating the square root of y.

```
============================================================
> y <- 12345
> x <- y/2
> while (abs(x^2 - y) > 1e-10)    x <- (x + y/x)/2
> x
[1] 111.11
```

```
> x^2
[1] 12345
************************************************************
```

- The test occurs at the top of the loop so that the expression might never be evaluated.

27• `repeat` STATEMENT

- A variation of the same algorithm with test at the bottom of the loop can be written with a `repeat` construction:

```
============================================================
> x <- y/2
> repeat {
+     x <- (x + y/x)/2
+     if (abs(x^2 - y) < 1e-10)  break
+}
> x
[1] 111.11
************************************************************
```

28• OTHER LOOPS

- A *compound expression*: several expressions held together between curly braces.
- An `if` construction for conditional execution.
- A `break` expression, which causes the enclosing loop to exit.
- Table B.4 lists other loops.

Table B.4 Flow controls

Function	Meaning
`if(p<1) print('good')`	conditional execution
`if(p<1) print('good')` `else print('bad')`	conditional execution with alternative
`for(i in 1:9) print(i)`	loop over list

B.5 User Functions

29• Creating a user function

- The R provides an extremely powerful method of writing functions for specific tasks of interest.
- First, we use `ls()` to list all objects in the workspace, use `rm()` to remove objects from the working directory, use `q()` to terminate the current R session.
- Then, we use `fix()` to edit your new function.
- For example, when you type `fix(mysum)` and press the key "Enter", a window will jump out so that you can edit your function with name `mysum`.

```
============================================================
function(a, b)
{
  # Function name: mysum(a, b)
  x <- a^2 + b^2
  return(x)
}
************************************************************
```

- Here `a` and `b` are two arguments.

```
============================================================
> mysum(3, 4)
[1] 25
************************************************************
```

30• Inserting a sub-function into a main function

- When a main function requires to *repeatedly* call a sub-function, we can insert the sub-function into the main function.
- For example,

```
============================================================
function(k)
{
  # Function name: main.mysum(k)
  nest.fun <- function(x, y, p)
  {
         (x + y)^p
  }
  x <- y <- 1:4
  z <- nest.fun(x,y,1) + nest.fun(x,y,2)*nest.fun(x,y,3)
  w <- sum(z)
  result <- list(z= z, w= w)
  return(result)
}
************************************************************
```

B.6 Some Commonly Used R Functions for Data Analysis

31• apply() FUNCTION

- Table B.5 lists some statistical functions.
- For large data sets, we can use the `apply()` function.
- The syntax is

```
apply(object, dim, function)
```

where `object` is the name of a matrix, `dim` can take the value 1 (row) or 2 (column), and `function` is the name of an R function (already available or created by the user).

```
============================================================
> X <- matrix(1:9, 3, 3)
> X
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
```

```
[3,]    3    6    9
> apply(X, 2, sum)
[1]  6 15 24
> apply(X, 1, mean)
[1] 4 5 6
> apply(X, 2, median)
[1] 2 5 8
> apply(X, 1, var)
[1] 9 9 9
************************************************************
```

Table B.5 Some statistical functions

Function	Meaning
`sum()`	summation
`prod()`	multiplication
`mean()`	average
`var()`	variance
`sd()`	standard deviation
`median()`	median
`quantile(x, p)`	quantiles
`cor(x, y)`	correlation

Table B.6 Data manipulation functions

Function	Meaning
`sort(x)`	returns a vector which is a sorted version of x
`order(x)`	returns an integer vector containing the permutation that will sort x into ascending order
`sort.list(x)`	= `order(x)`
`rank(x)`	returns a vector of the ranks of x
`rev(x)`	returns an object with the same length as x but with the elements or components in the reverse order

Table B.7 Normal distribution

Function	Meaning
`dnorm(x, mean, sd)`	density
`pnorm(x, mean, sd)`	cumulative distribution function
`qnorm(p, mean, sd)`	lower p-quantile, $x, \Pr(X \leqslant x) = p$
`rnorm(n, mean, sd)`	n random numbers

Table B.8 Cumulative distribution functions

Function	Meaning
`pnorm(x, mean, sd)`	normal
`plnorm(x, mean, sd)`	lognormal
`pt(x, df)`	Student-t
`pf(x, n1, n2)`	F
`pgamma(x, shape, scale)`	gamma
`pchisq(x, df)`	χ^2
`pexp(x, rate)`	exponential
`punif(x, min, max)`	uniform
`pbeta(x, a, b)`	beta
`pbinom(x, n, p)`	binomial
`ppois(x, lambda)`	Poisson

Table B.9 Parametric methods for discrete data

Function	Meaning
`binom.test`	binomial test (including sign test)
`prop.test`	comparison of proportions
`prop.trend.test`	test for trend in relative proportions
`fisher.test`	Fisher's exact test in small tables
`chisq.test`	chi-square test
`glm(y~x1+x2+x3, binomial)`	logistic regression

Table B.10 Parametric and non-parametric methods

Function	Meaning
`t.test`	one– and two–sample t test
`pairwise.t.test`	pairwise comparisons
`var.test`	comparison of two variances (F test)
`bartlett.test`	Bartlett's test (k variances)
`cor.test`	correlation
`cor.test` variants:	
`method=''kendall''`	Kendall's τ
`method=''spearman''`	Spearman's ρ
`lm(y ~ x)`	regression analysis
`lm(y ~ f)`	one–way analysis of variance
`lm(y ~ f1 + f2)`	two–way analysis of variance
`lm(y ~ f + x)`	analysis of covariance
`lm(y ~ x1 + x2 + x3)`	multiple regression analysis
`wilcox.test`	one– and two–sample Wilcoxon test
`kruskal.test`	Kruskal–Wallis test
`friedman.test`	Friedman's two–way analysis of variance

Table B.11 Model formulas

Function	Meaning
`~`	distributed by
`+`	additive effects
`:`	interaction
`*`	main effects + interaction (`a*b = a + b + a:b`)
`-1`	remove intercept

Table B.12 Linear and generalized linear models

Function	Meaning
`lm.out <- lm(y ~ x)`	fit models and save results
`summary(lm.out)`	coefficients and so on
`anova(lm.out)`	analysis of variance table
`fitted(lm.out)`	fitted values
`resid(lm.out)`	residuals
`predict(lm.out, newdata)`	predictions for new data frame
`glm(y ~ x, binomial)`	logistic regression

Table B.13 Survival analysis

Function	Meaning
`S <- Surv(time, ev)`	create survival object
`survfit(S)`	Kaplan–Meier estimate
`plot(survfit(S))`	survival curve
`survdiff(S ~ g)`	log-rank test for equal survival curves
`coxph(S ~ x1 + x2)`	Cox's proportional hazards model

Table B.14 Graphics

Function	Meaning
`plot()`	scatter plot and more
`hist()`	histogram
`boxplot()`	box-and-whiskers plot
`stripplot()`	strip plot
`barplot()`	bar diagram
`dotplot()`	dot diagram
`piechart()`	cakes
`interaction.plot()`	interaction plot
`lines()`	lines
`abline()`	line given by intercept and slope
`points()`	points
`segments()`	line segments
`arrows()`	arrows
`axis()`	axis
`box()`	frame around plot
`title()`	title above plot
`text()`	text in plot
`mtext()`	text in margin
`legend()`	list of symbols
`pch`	symbol (*p*lotting *ch*aracter)
`mfrow, mfcol`	several plots on one (*m*ulti*f*rame)
`xlim, ylim`	plot limits
`lty, lwd`	line type/width
`col`	colour
`cex, mex`	character size and line spacing in margins

Appendix C

Introduction of Latent Variables

C.1 MLEs of Parameters in t Distribution

1• The issue and aim

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} t(\mu, \sigma^2, \nu)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are two unknown parameters and $\nu > 0$ is known.
- The aim is to find the MLEs of μ and σ^2.

1.1• The density of t distribution

— From (A.32) in Appendix A.2.7, the pdf of $X \sim t(\mu, \sigma^2, \nu)$ is

$$t(x|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma}\left\{1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}.$$

— When $\mu = 0$ and $\sigma^2 = 1$, it is called the standard t distribution, denoted by $t(\nu)$.

— When ν is known, we denotes the density by $t(x|\mu, \sigma^2)$.

2• Newton–raphson algorithm

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} t(\mu, \sigma^2)$.
- Let $Y_{\text{obs}} = \{x_1, \ldots, x_n\}$ and $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, the likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}|Y_{\text{obs}}) &= \prod_{i=1}^n t(x_i|\mu, \sigma^2) = \prod_{i=1}^n \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma}\left\{1 + \frac{(x_i-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}} \\ &\propto (\sigma^2)^{-n/2}\prod_{i=1}^n \left\{1 + \frac{(x_i-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}} \end{aligned}$$

so that the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = -\frac{n}{2}\log(\sigma^2) - \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left\{1+\frac{(x_i-\mu)^2}{\nu\sigma^2}\right\}.$$

- The score vector is

$$\begin{aligned}\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \begin{pmatrix} \dfrac{\partial\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu} \\ \dfrac{\partial\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\sigma^2} \end{pmatrix} \\ &= \begin{pmatrix} (\nu+1)\displaystyle\sum_{i=1}^{n}\frac{x_i-\mu}{\nu\sigma^2+(x_i-\mu)^2} \\ -\dfrac{n}{2\sigma^2}+\dfrac{\nu+1}{2\sigma^2}\displaystyle\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\nu\sigma^2+(x_i-\mu)^2} \end{pmatrix}.\end{aligned}$$

- Clearly, we cannot obtain a closed-form solution to the score equation $\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \mathbf{0}$.

- To apply the Newton–Raphson algorithm

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{I}^{-1}(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}),$$

we need to calculate the inverse of the observed information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}|Y_{\text{obs}}) = -\begin{pmatrix} \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu^2} & \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu\partial\sigma^2} \\ \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu\partial\sigma^2} & \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial(\sigma^2)^2} \end{pmatrix},$$

where

$$\begin{cases} \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu^2} &= -(\nu+1)\displaystyle\sum_{i=1}^{n}\frac{\nu\sigma^2-(x_i-\mu)^2}{\{\nu\sigma^2+(x_i-\mu)^2\}^2}, \\ \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu\partial\sigma^2} &= -\nu(\nu+1)\displaystyle\sum_{i=1}^{n}\frac{x_i-\mu}{\{\nu\sigma^2+(x_i-\mu)^2\}^2}, \\ \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial(\sigma^2)^2} &= \dfrac{n}{2\sigma^4}-\dfrac{\nu+1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}\frac{(x_i-\mu)^2\{2\nu\sigma^2+(x_i-\mu)^2\}}{\{\nu\sigma^2+(x_i-\mu)^2\}^2}. \end{cases}$$

- We can see that the Fisher scoring algorithm is not available for the current case.

3• EM algorithm

- To apply the EM algorithm, we first show that the univariate t distribution is a mixture distribution of a normal distribution.
- **Theorem C.1** (Gamma mixture of a normal distribution). Let $\tau \sim \text{Gamma}(\nu/2, \nu/2)$ and $X|\tau \sim N(\mu, \tau^{-1}\sigma^2)$, then $X \sim t(\mu, \sigma^2, \nu)$. ‖

Proof: The density of X is given by

$$\begin{aligned} f_X(x) &= \int f_{X,\tau}(x,\tau)\,\mathrm{d}\tau = \int f_\tau(\tau)\cdot f_{X|\tau}(x|\tau)\,\mathrm{d}\tau \\ &= \int \text{Gamma}(\tau|\nu/2,\nu/2)\cdot N(x|\mu,\tau^{-1}\sigma^2)\,\mathrm{d}\tau \\ &= \int \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\tau^{\frac{\nu}{2}-1}\,\mathrm{e}^{-\frac{\tau\nu}{2}}\cdot\frac{\sqrt{\tau}}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-\frac{\tau(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}\tau \\ &= \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})\sqrt{2\pi}\sigma}\int \tau^{\frac{\nu+1}{2}-1}\,\mathrm{e}^{-\left\{\frac{\nu}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right\}\tau}\,\mathrm{d}\tau \\ &= \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})\sqrt{2\pi}\sigma}\cdot\frac{\Gamma(\frac{\nu+1}{2})}{\left\{\frac{\nu}{2}+\frac{(x-\mu)^2}{2\sigma^2}\right\}^{\frac{\nu+1}{2}}} = t(x|\mu,\sigma^2,\nu), \end{aligned}$$

which completes the proof. □

3.1• Complete-data log-likelihood function

— Based on Theorem C.1, we introduce latent variables $Y_{\text{mis}} = \{\tau_1, \ldots, \tau_n\}$, where $\tau_1, \ldots, \tau_n \overset{\text{iid}}{\sim} \text{Gamma}(\nu/2, \nu/2)$.

— Thus, the complete data are $Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\}$.

— The complete-data likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}|Y_{\text{com}}) &= \prod_{i=1}^n f(x_i,\tau_i;\boldsymbol{\theta}) = \prod_{i=1}^n f(\tau_i)\cdot f(x_i|\tau_i,\boldsymbol{\theta}) \\ &= \prod_{i=1}^n \text{Gamma}(\tau_i|\nu/2,\nu/2)\cdot N(x_i|\mu,\tau_i^{-1}\sigma^2) \end{aligned}$$

$$= \prod_{i=1}^{n} \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \tau_i^{\nu/2-1} \mathrm{e}^{-\tau_i \nu/2} \cdot \frac{1}{\sqrt{2\pi\tau_i^{-1}\sigma^2}} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\tau_i^{-1}\sigma^2}}$$

$$\propto \prod_{i=1}^{n} \tau_i^{(\nu+1)/2-1} \mathrm{e}^{-\tau_i \nu/2} \cdot \sigma^{-1} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\tau_i^{-1}\sigma^2}}, \tag{C.1}$$

and the complete-data log-likelihood function is given by

$$\ell(\boldsymbol{\theta}|Y_{\text{com}}) \propto -\frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\tau_i^{-1}\sigma^2}.$$

3.2• M-step

— Let

$$0 = \frac{\partial \ell(\boldsymbol{\theta}|Y_{\text{com}})}{\partial \mu} = -\sum_{i=1}^{n} \frac{2(x_i-\mu)(-1)}{2\tau_i^{-1}\sigma^2},$$

then, the complete-data MLE of μ is

$$\mu = \frac{\sum_{i=1}^{n} \tau_i x_i}{\sum_{i=1}^{n} \tau_i}. \tag{C.2}$$

— Let

$$0 = \frac{\partial \ell(\boldsymbol{\theta}|Y_{\text{com}})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} - \left(\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\tau_i^{-1}}\right)(-1)(\sigma^2)^{-2},$$

then, the complete-data MLE of σ^2 is

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \tau_i (x_i-\mu)^2. \tag{C.3}$$

3.3• E-step

— To implement the E-step, i.e., to calculate $E(\tau_i|Y_{\text{obs}}, \boldsymbol{\theta})$, we need to find the conditional predictive distribution.

— From (C.1), we obtain

$$\begin{aligned} f(\tau_i|X_i = x_i, \boldsymbol{\theta}) &\propto f(x_i, \tau_i; \boldsymbol{\theta}) \\ &\propto \tau_i^{(\nu+1)/2-1} \mathrm{e}^{-\tau_i \nu/2} \cdot \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\tau_i^{-1}\sigma^2}} \\ &\propto \tau_i^{(\nu+1)/2-1} \mathrm{e}^{-\tau_i \frac{\nu+(x_i-\mu)^2/\sigma^2}{2}}, \end{aligned}$$

that is,

$$\tau_i|(X_i = x_i, \boldsymbol{\theta}) \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + (x_i - \mu)^2/\sigma^2}{2}\right), \quad i = 1, \ldots, n.$$

— Hence,

$$E(\tau_i|X_i = x_i, \boldsymbol{\theta}) = \frac{\nu+1}{\nu + (x_i - \mu)^2/\sigma^2}, \quad i = 1, \ldots, n. \tag{C.4}$$

— Therefore, the E-step is to calculate (C.4), and the M-step updates (C.2) and (C.3) by replacing τ_i with $E(\tau_i|X_i = x_i, \boldsymbol{\theta})$.

C.2 MLEs of Parameters in the Poisson Additive Model

4• The issue and aim

- Let

$$Y_i \overset{\text{ind}}{\sim} \text{Poisson}\,(\boldsymbol{a}_i^\top \boldsymbol{\theta}), \quad 1 \leqslant i \leqslant n, \tag{C.5}$$

  where $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{ip})^\top$ is a known vector and each $a_{ij} \geqslant 0$.

- The aim is to estimate the unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ based on $Y_{\text{obs}} = \{y_i\}_{i=1}^n$, where y_i is the realization of Y_i.

5• Newton–Raphson algorithm

- The observed-data likelihood function of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) = \prod_{i=1}^n \text{Poisson}\,(y_i|\boldsymbol{a}_i^\top\boldsymbol{\theta}) = \prod_{i=1}^n \frac{(\boldsymbol{a}_i^\top\boldsymbol{\theta})^{y_i} \exp(-\boldsymbol{a}_i^\top\boldsymbol{\theta})}{y_i!}$$

  so that the log-likelihood function is

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \text{constant} + \sum_{i=1}^n \Big\{ y_i \log(\boldsymbol{a}_i^\top\boldsymbol{\theta}) - \boldsymbol{a}_i^\top\boldsymbol{\theta} \Big\}. \tag{C.6}$$

- The score vector, the Hessian matrix and the observed information matrix are given by

$$\begin{aligned}
\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \sum_{i=1}^{n}\left(y_i\frac{\boldsymbol{a}_i}{\boldsymbol{a}_i^\top\boldsymbol{\theta}}-\boldsymbol{a}_i\right)=\sum_{i=1}^{n}\boldsymbol{a}_i(y_i/\boldsymbol{a}_i^\top\boldsymbol{\theta}-1),\\
\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \frac{\partial\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\boldsymbol{\theta}^\top}=-\sum_{i=1}^{n}\frac{y_i}{(\boldsymbol{a}_i^\top\boldsymbol{\theta})^2}\cdot\boldsymbol{a}_i\boldsymbol{a}_i^\top,\\
-\nabla^2\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \sum_{i=1}^{n}\frac{y_i}{(\boldsymbol{a}_i^\top\boldsymbol{\theta})^2}\cdot\boldsymbol{a}_i\boldsymbol{a}_i^\top,
\end{aligned}$$

  respectively.

- Thus, the Newton–Raphson algorithm is defined by

$$\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t)}+\{-\nabla^2\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})\}^{-1}\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}).$$

6• EM algorithm

- For a fixed i, we introduce a latent vector $\mathbf{z}_i=(Z_{i1},\ldots,Z_{ip})^\top$ by splitting Y_i and $\boldsymbol{a}_i^\top\boldsymbol{\theta}$ as follows:

$$\begin{array}{ccccccccc}
Y_i & = & Z_{i1} & +\cdots+ & Z_{ij} & +\cdots+ & Z_{ip}, & & \\
\updownarrow & & \updownarrow & & \updownarrow & & \updownarrow & & \\
\boldsymbol{a}_i^\top\boldsymbol{\theta} & = & a_{i1}\theta_1 & +\cdots+ & a_{ij}\theta_j & +\cdots+ & a_{ip}\theta_p, & &
\end{array} \tag{C.7}$$

  where

$$Z_{ij}\overset{\text{ind}}{\sim}\text{Poisson}\,(a_{ij}\theta_j),\quad 1\leqslant j\leqslant p. \tag{C.8}$$

6.1• M-step

— Let $Y_{\text{mis}}=\{\boldsymbol{z}_i\}_{i=1}^n=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\}$ and $Y_{\text{com}}=\{Y_{\text{obs}},Y_{\text{mis}}\}=Y_{\text{mis}}$.

— The complete-data likelihood function is

$$\begin{aligned}
L(\boldsymbol{\theta}|Y_{\text{com}}) &= \prod_{i=1}^{n}\prod_{j=1}^{p}\text{Poisson}\,(z_{ij}|a_{ij}\theta_j)\\
&\propto \prod_{j=1}^{p}\theta_j^{\sum_{i=1}^n z_{ij}}\exp\left(-\theta_j\sum_{i=1}^{n}a_{ij}\right),
\end{aligned}$$

so that complete-data MLEs of $\{\theta_j\}_{j=1}^p$ are

$$\hat{\theta}_j = \frac{\sum_{i=1}^n z_{ij}}{\sum_{i=1}^n a_{ij}}, \quad 1 \leqslant j \leqslant p.$$

6.2• E-step

— To derive the conditional predictive distribution of $\mathbf{z}_i|(Y_{\text{obs}}, \boldsymbol{\theta})$, we require the following result.

— **Theorem C.2** (Independent Poisson r.v.'s conditional on their sum). Let $X_j \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_j)$, then

$$(X_1, \ldots, X_p)|(\textstyle\sum_{j=1}^p X_j = N) \sim \text{Multinomial}_p(N, \boldsymbol{\pi}), \tag{C.9}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)^\top$ and $\pi_j = \lambda_j / \sum_{k=1}^p \lambda_k$, $j = 1, \ldots, p$. ‖

Proof: Let $X_+ = \sum_{j=1}^p X_j$ and $\lambda_+ = \sum_{j=1}^p \lambda_j$. Since all X_j are independent, we have $X_+ \sim \text{Poisson}(\lambda_+)$. The conditional distribution of $(X_1, \ldots, X_p)|(X_+ = N)$ is

$$\begin{aligned}
& \Pr(X_1 = x_1, \ldots, X_p = x_p | X_+ = N) \\
= \; & \frac{\Pr(X_1 = x_1, \ldots, X_p = x_p)}{\Pr(X_+ = N)} \qquad [\text{where } N = \textstyle\sum_{j=1}^p x_j] \\
= \; & \frac{\prod_{j=1}^p \mathrm{e}^{-\lambda_j} \lambda_j^{x_j} / x_j!}{\mathrm{e}^{-\lambda_+} \lambda_+^N / N!} = \binom{N}{x_1, \ldots, x_p} \prod_{j=1}^p \left(\frac{\lambda_j}{\lambda_+}\right)^{x_j},
\end{aligned}$$

which implies (C.9). □

— By applying Theorem C.2 to (C.7) and (C.8), we know that the conditional predictive distribution

$$\mathbf{z}_i|(Y_i = y_i, \boldsymbol{\theta}) \sim \text{Multinomial}_p\left(y_i;\ \frac{a_{i1}\theta_1}{\boldsymbol{a}_i^\top \boldsymbol{\theta}}, \ldots, \frac{a_{ip}\theta_p}{\boldsymbol{a}_i^\top \boldsymbol{\theta}}\right),$$

for $i = 1, \ldots, n$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent.

— The E-step is to calculate

$$E(Z_{ij}|Y_i = y_i, \boldsymbol{\theta}) = \frac{y_i \cdot a_{ij}\theta_j}{\boldsymbol{a}_i^\top \boldsymbol{\theta}}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant p.$$

— Finally, the iteration is given by

$$\theta_j^{(t+1)} = \theta_j^{(t)} \frac{\sum_{i=1}^n \{y_i a_{ij} / \boldsymbol{a}_i^\top \boldsymbol{\theta}^{(t)}\}}{\sum_{i=1}^n a_{ij}}, \quad 1 \leqslant j \leqslant p.$$

C.3 MLEs of Parameters in Constrained Normal Models

7• The issue and aim

- Let $\{Y_{ij}\}_{j=1}^{n_i} \overset{\text{iid}}{\sim} N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, m$, where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)^\top$ is a known parameter vector.
- The aim is to estimate the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ subject to a linear constraint

$$\boldsymbol{\mu} = \boldsymbol{B}\boldsymbol{\theta} \quad \text{or} \quad \mu_i = \sum_{k=1}^{q} b_{ik}\theta_k, \quad i = 1, \ldots, m,$$

where $\boldsymbol{B}_{m\times q} = (b_{ik})$ is a known scalar matrix and

$$\begin{aligned} \boldsymbol{\theta} \in \mathbb{R}^r \times \mathbb{R}_+^{q-r} \quad &= \quad \{(\theta_1, \ldots, \theta_q)^\top \colon \theta_k \in \mathbb{R},\ k = 1, \ldots, r; \\ &\qquad \theta_k \in \mathbb{R}_+,\ k = r+1, \ldots, q\}. \end{aligned} \tag{C.10}$$

8• Data augmentation

- Since $\boldsymbol{\sigma}^2$ is known, for each i, the sample mean

$$\overline{Y_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \sim N(\mu_i, \sigma_i^2/n_i)$$

is a sufficient statistic of μ_i.

- Thus, we denote the observed data by $Y_{\text{obs}} = \{\overline{Y_i}\}_{i=1}^m$.
- Noting the *linear invariant property* of the normal distribution, we augment the Y_{obs} by latent variables

$$Y_{\text{mis}} = \left\{Z_{ik} \colon 1 \leqslant i \leqslant m,\ 1 \leqslant k \leqslant q-1\right\}$$

to obtain the complete-data

$$Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\} = \{Z_{ik} \colon 1 \leqslant i \leqslant m,\ 1 \leqslant k \leqslant q\},$$

where

$$\begin{aligned} Z_{ik} \quad &\overset{\text{ind}}{\sim} \quad N(b_{ik}\theta_k,\ \sigma_i^2/(qn_i)), \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant k \leqslant q, \\ \overline{Y_i} \quad &= \quad \sum_{k=1}^{q} Z_{ik}, \quad 1 \leqslant i \leqslant m. \end{aligned} \tag{C.11}$$

- By mapping Y_{com} to Y_{obs}, the transformation (C.11) preserves the observed-data likelihood

$$\prod_{i=1}^{m} N(\overline{Y_i}|\mu_i, \sigma_i^2/n_i).$$

8.1• Linear invariant property

— The linear invariant property indicates that linear combinations of independent r.v.'s from a particular family of distributions still have distribution in that family.

8.2• A result on independent normal variables

— To derive the conditional predictive distribution of the latent data Y_{mis} given Y_{obs} and $\boldsymbol{\theta}$, we first prove the following result.

— **Theorem C.3** (Independent normal r.v.'s conditional on their sum). Let $\{W_k\}_{k=1}^q \overset{\text{ind}}{\sim} N(\beta_k, \delta_k^2)$, then

$$\begin{aligned} &(W_1, \ldots, W_{q-1})^\top | (\textstyle\sum_{k=1}^q W_k = w) \qquad \text{(C.12)} \\ \sim\ & N_{q-1}\left(\boldsymbol{\beta}_{-q} + \frac{w - \sum_{k=1}^q \beta_k}{\sum_{k=1}^q \delta_k^2} \cdot \boldsymbol{\delta}_{-q}^2,\ \operatorname{diag}(\boldsymbol{\delta}_{-q}^2) - \frac{\boldsymbol{\delta}_{-q}^2 \boldsymbol{\delta}_{-q}^{2\top}}{\sum_{k=1}^q \delta_k^2}\right), \end{aligned}$$

where $\boldsymbol{\beta}_{-q} = (\beta_1, \ldots, \beta_{q-1})^\top$ and $\boldsymbol{\delta}_{-q}^2 = (\delta_1^2, \ldots, \delta_{q-1}^2)^\top$. Especially, if $\delta_1^2 = \cdots = \delta_q^2 = \delta^2$, then

$$\begin{aligned} &(W_1, \ldots, W_{q-1})^\top | (\textstyle\sum_{k=1}^q W_k = w) \qquad \text{(C.13)} \\ \sim\ & N_{q-1}\left(\boldsymbol{\beta}_{-q} + \frac{w - \sum_{k=1}^q \beta_k}{q} \mathbf{1}_{q-1},\ \delta^2\left(\boldsymbol{I}_{q-1} - \frac{\mathbf{1}_{q-1}\mathbf{1}_{q-1}^\top}{q}\right)\right). \end{aligned}$$

‖

Proof: Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^\top$ and $\boldsymbol{\delta}^2 = (\delta_1^2, \ldots, \delta_q^2)^\top$. Note that

$$\mathbf{w} = (W_1, \ldots, W_q)^\top \sim N_q(\boldsymbol{\beta},\ \operatorname{diag}(\boldsymbol{\delta}^2)),$$

we have

$$\begin{pmatrix} \mathbf{w} \\ \sum_{k=1}^q W_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_q \\ \mathbf{1}_q^\top \end{pmatrix} \mathbf{w} \sim N_{q+1}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{1}_q^\top \boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \operatorname{diag}(\boldsymbol{\delta}^2) & \boldsymbol{\delta}^2 \\ \boldsymbol{\delta}^{2\top} & \mathbf{1}_q^\top \boldsymbol{\delta}^2 \end{pmatrix}\right).$$

From the property of multivariate normal distribution, we have

$$\begin{aligned}&(W_1,\ldots,W_q)^\top|(\textstyle\sum_{k=1}^q W_k = w)\\ \sim\ & N_q\left(\boldsymbol{\beta} + \frac{w-\sum_{k=1}^q \beta_k}{\mathbf{1}_q^\top\boldsymbol{\delta}^2}\cdot\boldsymbol{\delta}^2,\ \mathrm{diag}(\boldsymbol{\delta}^2) - \frac{\boldsymbol{\delta}^2\boldsymbol{\delta}^{2\top}}{\mathbf{1}_q^\top\boldsymbol{\delta}^2}\right). \qquad \text{(C.14)}\end{aligned}$$

Let $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\delta}^2) - \boldsymbol{\delta}^2\boldsymbol{\delta}^{2\top}/\mathbf{1}_q^\top\boldsymbol{\delta}^2$, then $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}(\boldsymbol{I}_q - \boldsymbol{\Sigma}_1)\boldsymbol{D}^{1/2}$, where

$$\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\delta}^2) \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \boldsymbol{D}^{1/2}\mathbf{1}_q(\mathbf{1}_q^\top\boldsymbol{D}\mathbf{1}_q)^{-1}\mathbf{1}_q^\top D^{1/2}.$$

Note that $\boldsymbol{I}_q - \boldsymbol{\Sigma}_1$ is a projection matrix and the rank of a projection matrix equals its trace. So $\mathrm{rank}\,(\boldsymbol{I}_q - \boldsymbol{\Sigma}_1) = q-1$, implying $\mathrm{rank}\,(\boldsymbol{\Sigma}) = q-1 < q$. That is, the distribution in (C.14) is a degenerate q-dimensional normal distribution. From (C.14), we obtain (C.12) and (C.13). □

9• MLE via the EM algorithm

- The likelihood function of $\boldsymbol{\theta}$ for the complete-data $Y_{\mathrm{com}}(\boldsymbol{\sigma}^2)$ is

$$\begin{aligned}L(\boldsymbol{\theta}|Y_{\mathrm{com}}) \ &\propto\ \prod_{i=1}^m (\sigma_i^2)^{-q/2}\\ &\quad\times\ \exp\left[-\frac{1}{2}\sum_{i=1}^m\left\{\frac{qn_i}{\sigma_i^2}\sum_{k=1}^q (Z_{ik}-b_{ik}\theta_k)^2\right\}\right],\end{aligned}$$

  where $\boldsymbol{\theta}\in\mathbb{R}^r\times\mathbb{R}_+^{q-r}$ defined by (C.10).

- Therefore, the sufficient statistics for θ_k are

$$S_k = \sum_{i=1}^m\left(\frac{n_i b_{ik}}{\sigma_i^2}\right)Z_{ik}, \quad k=1,\ldots,q.$$

- The complete-data MLEs of θ_k are given by

$$\begin{aligned}\hat\theta_k &= \frac{S_k}{\sum_{i=1}^m n_i b_{ik}^2/\sigma_i^2}, && k=1,\ldots,r, \quad \text{and}\\ \hat\theta_k &= \max\left(0,\ \frac{S_k}{\sum_{i=1}^m n_i b_{ik}^2/\sigma_i^2}\right), && k=r+1,\ldots,q.\end{aligned}$$

9.1• E-step

— Let $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{i,q-1})^\top$ and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{i,q-1})^\top$, then from (C.11) and (C.13), the conditional predictive distribution of $Y_{\text{mis}}|(Y_{\text{obs}}, \boldsymbol{\theta})$ is

$$\begin{aligned} f(Y_{\text{mis}}|Y_{\text{obs}}, \boldsymbol{\theta}) &= \prod_{i=1}^{m} f\Big(\boldsymbol{z}_i \Big| \textstyle\sum_{k=1}^{q} Z_{ik} = \overline{Y_i}, \boldsymbol{\theta}\Big) \\ &= \prod_{i=1}^{m} N_{q-1}(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{V}_i), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{a}_i &= (b_{i1}\theta_1, \ldots, b_{i,q-1}\theta_{q-1})^\top + \mathbf{1}_{q-1}\frac{\overline{Y_i} - \sum_{k=1}^{q} b_{ik}\theta_k}{q}, \\ \boldsymbol{V}_i &= \frac{\sigma_i^2}{qn_i}\Big(\boldsymbol{I}_{q-1} - \frac{\mathbf{1}_{q-1}\mathbf{1}_{q-1}^\top}{q}\Big). \end{aligned}$$

— Given Y_{obs} and the current estimate $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_q^{(t)})^\top$, the E-step is to calculate the conditional expectation of the complete-data sufficient statistic S_k,

$$\begin{aligned} S_k^{(t)} &= E\Big(S_k|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\Big) \\ &= \sum_{i=1}^{m}\left(\frac{n_i b_{ik}}{\sigma_i^2}\right)\cdot\left(b_{ik}\theta_k^{(t)} + \frac{\overline{Y_i} - \sum_{\ell=1}^{q} b_{i\ell}\theta_\ell^{(t)}}{q}\right) \end{aligned}$$

for $k = 1, \ldots, q$.

9.2• M-step

— The M-step is to update

$$\begin{aligned} \theta_k^{(t+1)} &= \frac{S_k^{(t)}}{\sum_{i=1}^{m} n_i b_{ik}^2/\sigma_i^2}, \qquad k = 1, \ldots, r, \quad \text{and} \\ \theta_k^{(t+1)} &= \max\left(0, \frac{S_k^{(t)}}{\sum_{i=1}^{m} n_i b_{ik}^2/\sigma_i^2}\right), \; k = r+1, \ldots, q. \end{aligned}$$

C.4 Binormal Model with Missing Data

10• A motivation example

- A bivariate normal sample with missing values on both variables is a classical missing data problem (Wilks 1932).
- However, a closed–form solution cannot be obtained by either the EM (Dempster *et al.* 1977; McLachlan & Krishnan 1997, p.45–49 & p.91–94), the DA (Tanner & Wong 1987), or the Gibbs sampling (Gelfand & Smith 1990, p.404–405).
- In this section, we obtain a closed–form solution to this problem by means of the IBF.

Table C.1 Data from a bivariate normal distribution

Variate 1:	1	1	–1	–1	2	2	–2	–2	∗	∗	∗	∗
Variate 2:	1	–1	1	–1	∗	∗	∗	∗	2	2	–2	–2

NOTE: The ∗ denotes a missing value.

11• The model

- Motivated by the 12 observations in Table C.1 (Murray 1977), let $\boldsymbol{x}_i = (x_{1i}, x_{2i})^\top$, $i = 1, \ldots, n$, be a random sample of size n from $N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ with known mean $\boldsymbol{\mu}_0$ and unknown covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

- Without loss of generality, let $\boldsymbol{\mu}_0 = \mathbf{0}_2$.
- Denote the observed data and the missing data by

$$\begin{aligned} Y_{\text{obs}} &= \{x_i\}_{i=1}^{n_1} \cup \{x_{1i}\}_{i=n_1+1}^{n_1+n_2} \cup \{x_{2i}\}_{i=n_1+n_2+1}^{n} \quad \text{and} \\ Z &= \{x_{2i}\}_{i=n_1+1}^{n_1+n_2} \cup \{x_{1i}\}_{i=n_1+n_2+1}^{n}, \end{aligned}$$

respectively.

11.1• Complete-data posterior distribution

— A standard prior of $\boldsymbol{\Sigma}$ originated in Box & Tiao (1973, p.426) is

$$\pi(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2} = (\sigma_1\sigma_2\sqrt{1-\rho^2}\,)^{-(p+1)}$$

where p denotes the dimension of the multivariate normal distribution (for the current case, $p = 2$).

— Hence, the complete-data posterior $f(\boldsymbol{\Sigma}|Y_{\text{obs}}, Z)$ is proportional to

$$\frac{1}{(\sigma_1\sigma_2\sqrt{1-\rho^2})^{n+p+1}} \exp\left\{-\frac{\sigma_2^2 s_1^2(n) - 2\rho\sigma_1\sigma_2 s_{12}(n) + \sigma_1^2 s_2^2(n)}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right\},$$

where $s_k^2(n) \doteq \sum_{i=1}^n x_{ki}^2$ for $k = 1, 2$, and $s_{12}(n) = \sum_{i=1}^n x_{1i}x_{2i}$.

11.2• Conditional prediction distribution

— On the other hand, the conditional prediction distribution is

$$f(Z|Y_{\text{obs}}, \boldsymbol{\Sigma}) = \prod_{i=n_1+1}^{n_1+n_2} \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{(x_{2i} - \rho\sigma_2 x_{1i}/\sigma_1)^2}{2\sigma_2^2(1-\rho^2)}\right\}$$
$$\times \prod_{i=n_1+n_2+1}^{n} \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{(x_{1i} - \rho\sigma_1 x_{2i}/\sigma_2)^2}{2\sigma_1^2(1-\rho^2)}\right\}.$$

11.3• Using the point-wise IBF

— By the point-wise IBF, we obtain

$$\begin{aligned} f(\boldsymbol{\Sigma}|Y_{\text{obs}}) \;\propto\; & \frac{1}{\sigma_1^{n-n_3+1+p}\sigma_2^{n-n_2+1+p}(1-\rho^2)^{(n_1+1+p)/2}} \\ & \times \exp\left\{-\frac{\sigma_2^2 s_1^2(n_1) - 2\rho\sigma_1\sigma_2 s_{12}(n_1) + \sigma_1^2 s_2^2(n_1)}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right\} \\ & \times \exp\left\{-\frac{\sum_{i=n_1+1}^{n_1+n_2} x_{1i}^2}{2\sigma_1^2} - \frac{\sum_{i=n_1+n_2+1}^{n} x_{2i}^2}{2\sigma_2^2}\right\}. \end{aligned} \tag{C.15}$$

— Substituting the data in Table C.1 into (C.15), we obtain

$$\begin{aligned} f(\sigma_1^2, \sigma_2^2, \rho|Y_{\text{obs}}) \;\propto\; & (\sigma_1\sigma_2)^{-(9+p)}(1-\rho^2)^{-(5+p)/2} \\ & \times \exp\left\{-\frac{2\sigma_1^2 + 2\sigma_2^2}{\sigma_1^2\sigma_2^2(1-\rho^2)} - \frac{8}{\sigma_1^2} - \frac{8}{\sigma_2^2}\right\}. \end{aligned} \tag{C.16}$$

— Integrating (C.16) with respect to σ_1^2 and σ_2^2, we obtain

$$f(\rho|Y_{\text{obs}}) \propto \frac{(1-\rho^2)^{4.5+0.5p}}{(1.25-\rho^2)^{7+p}}.$$

— Let $p = 2$, we have

$$f(\rho|Y_{\text{obs}}) \propto \frac{(1-\rho^2)^{5.5}}{(1.25-\rho^2)^{9}},$$

which is slightly different from the result in Tanner (1996, p.96).

— By numerical integration, the normalizing constant is

$$\int_{-1}^{1} \frac{(1-\rho^2)^{5.5}}{(1.25-\rho^2)^{9}}\, d\rho = 0.3798.$$

— Thus, the posterior distribution of ρ is (see Figure C.1)

$$f(\rho|Y_{\text{obs}}) = \frac{(1-\rho^2)^{5.5}}{0.3798(1.25-\rho^2)^{9}}. \tag{C.17}$$

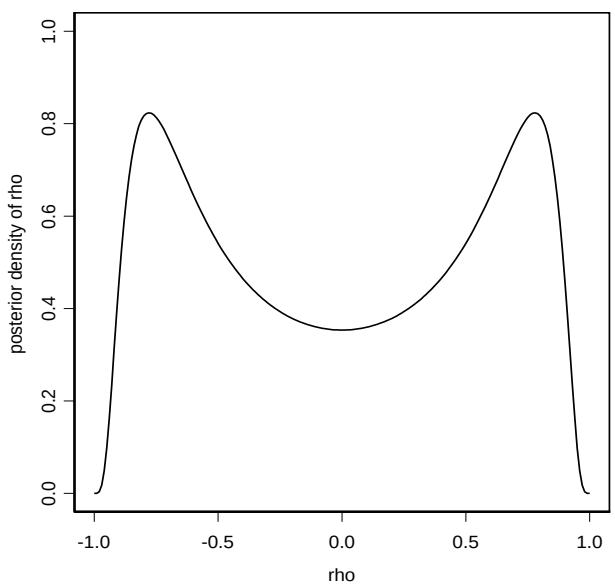

Figure C.1 Posterior distribution of ρ defined in (C.17).

List of Figures

List of Tables

List of Acronyms

AR	acceptance–rejection
arg	argument
a.s.	almost surely
ASL	achieved significance level
BCI	bootstrap confidence interval
cdf	cumulative distribution function
CI	confidence interval
CS	Charlier series
CV	coefficient of variation
DA	data augmentation
DP	De Pierro
ECM	expectation/conditional maximization
e.g.	for example
EM	expectation–maximization
GEM	generalized EM
GLM	generalized linear model
GP	generalized Poisson
HPP	homogeneous Poisson process
IBF	inverse Bayes formulae
i.e.	that is
iff	if and only if
i.i.d.	independent and identically distributed
inf	infimum
IS	importance sampling
KL	Kullback–Leibler
LSE	least square estimator
MAR	missing at random
MC	Monte Carlo
MCAR	missing completely at random
MCMC	Markov Chain Monte Carlo
mgf	moment generating function
MLE	maximum likelihood estimate
MM	minorization–maximization
MSE	mean square error
NHPP	nonhomogeneous Poisson process

NR	Newton–Raphson
pdf	probability density function
pmf	probability mass function
QLB	quadratic lower–bound
r.v.	random variable
r.v.'s	random variables
SIR	sampling/importance resampling
SR	stochastic representation
Std	standard deviation
sup	supremum
ZIP	zero-inflated Poisson
ZTB	zero-truncated binomial
ZTGP	zero-truncated generalized Poisson
ZTNB	zero-truncated negative–binomial
ZTP	zero-truncated Poisson

List of Symbols

Mathematics

$\propto$	proportional to
$\|$	end of a comment/example/solution/theorem
$\gg$	much greater than
$\square$	end of the proof
$\hat{=}$	equal by definition
$\equiv$	always equal to
$\doteq$, $\approx$	approximately equal to
$\neq$	not equal to
$\forall$	for all
$\|a\|$	absolute value of a
$\bar{x}$	average of the elements in the vector $\boldsymbol{x}_{n\times 1}$
$\tilde{x}$	mode of some function
$\|\boldsymbol{x}\|$	ℓ_2-norm of $\boldsymbol{x}$, $(\sum_{i=1}^{n} x_i^2)^{1/2}$
$\boldsymbol{x}^\top$	transpose of $\boldsymbol{x}$
$\boldsymbol{x}^{\otimes 2}$	$\boldsymbol{x}\boldsymbol{x}^\top$
$\mathrm{diag}(\boldsymbol{x})$	diagonal matrix with main diagonal elements $x_1, \ldots, x_n$
$\mathbf{0}_n$	n-dimensional vector of zeros
$\mathbf{1}_n$	n-dimensional vector of ones
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\det(\boldsymbol{A})$	determinant of the matrix $\boldsymbol{A}$
$\mathrm{tr}\,(\boldsymbol{A})$	trace of $\boldsymbol{A}$
$\boldsymbol{A}^\top$	transpose of $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	inverse of $\boldsymbol{A}$
$\boldsymbol{\Sigma} > 0$	positive definite matrix
$g'(x), g''(x), g'''(x)$	first, second, third derivatives of $g(x)$
$g^{(n)}(x), \dfrac{\mathrm{d}^n g(x)}{\mathrm{d}x^n}$	n-th derivative of $g(x)$
$\nabla f(\boldsymbol{x})$	gradient vector of $f(\boldsymbol{x})$
$\nabla^2 f(\boldsymbol{x})$	Hessian matrix of $f(\boldsymbol{x})$

$\Gamma(a)$	gamma function ($a > 0$)
$B(a, b)$	beta function ($a, b > 0$)
$B(a_1, \ldots, a_n)$	multivariate beta function
$\binom{n}{k}$	binomial coefficient
$\binom{n}{n_1, \ldots, n_k}$	multinomial coefficient
$I_{\mathbb{A}}(x), I(x \in \mathbb{A})$	indicator function
$\operatorname{sgn}(x)$	sign function
$\mathbb{A}, \mathbb{B}, \mathbb{C}$	events
$\mathbb{R}$	real line, $(-\infty, \infty)$
$\mathbb{R}^n$	n-dimensional Euclidean space, $\{\boldsymbol{x} = (x_1, \ldots, x_n)^\top \colon x_i \in \mathbb{R},\ i = 1, \ldots, n\}$
$\mathbb{R}_+$	positive real line $(0, \infty)$ or $[0, \infty)$
$\mathbb{R}_+^n$	n-dimensional positive orthant $\{\boldsymbol{x} \colon x_i \geqslant 0,\ 1 \leqslant i \leqslant n\}$
$\mathbb{B}_n(r)$	n-dimensional ball with radius r, $\{\boldsymbol{x} \colon \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{x}^\top \boldsymbol{x} \leqslant r^2\}$
$\mathbb{B}_n$	n-dimensional unit ball, $\mathbb{B}_n(1)$
$\mathbb{V}_n(c)$	n-dimensional simplex, $\{\boldsymbol{x} \colon \boldsymbol{x} \in \mathbb{R}_+^n,\ \boldsymbol{x}^\top \mathbf{1} \leqslant c\}$
$\mathbb{V}_n$	$\mathbb{V}_n(1)$
$\mathbb{T}_n(c)$	n-dimensional hyperplane, $\{\boldsymbol{x} \colon \boldsymbol{x} \in \mathbb{R}_+^n,\ \boldsymbol{x}^\top \mathbf{1} = c\}$
$\mathbb{T}_n$	$\mathbb{T}_n(1)$

Probability

$\sim$	distributed as
$\dot{\sim}$	approximately distributed as
$\overset{\text{iid}}{\sim}$	independent and identically distributed as
$\overset{\text{ind}}{\sim}$	independently distributed as
$\overset{\text{d}}{=}$	having the same distribution on both sides
X, Y, Z	random variables
x, y, z	realizations of X, Y and Z
$\mathbb{X}, \mathcal{X}$	sample space
$X \perp\!\!\!\perp Y$	X and Y are independent
$\mathbf{x} = (X_1, \ldots, X_n)^\top$	random vector
$\boldsymbol{x} = (x_1, \ldots, x_n)^\top$	realizations of $\mathbf{x} = (X_1, \ldots, X_n)^\top$
$\mathbf{X} = (X_{ij})$	random matrix
$\boldsymbol{X} = (x_{ij})$	realizations of random matrix $\mathbf{X} = (X_{ij})$
$X_{(1)}, \ldots, X_{(n)}$	order statistics of $X_1, \ldots, X_n$

$p(x)$	pmf of X
$p(x, y)$	joint pmf of (X, Y)
$f(x), f_X(x)$	pdf of X
$F(x), F_X(x)$	cdf of X
$f(x, y), f_{(X,Y)}(x, y)$	joint pdf of (X, Y)
$F(x, y), F_{(X,Y)}(x, y)$	joint cdf of (X, Y)
$\mathcal{S}_X, \mathcal{S}_{(X,Y)}$	supports of X and (X, Y)
$\mathcal{S}_{(X\|Y)}(y)$	support of $X\|(Y = y)$
$\phi(x)$	pdf of standard normal distribution
$\Phi(x)$	cdf of standard normal distribution
$\Psi(x)$	$\phi(x)/\{1 - \Phi(x)\}$
Pr	probability measure
$E(X)$	expectation of the r.v. X
$\mathrm{Var}(X)$	variance of X
$\mathrm{Cov}(X, Y)$	covariance of X and Y
$M_X(t)$	mgf of X
$\bar{\mu}_m \xrightarrow{\mathrm{D}} N(0, 1)$	convergence in distribution
$\bar{\mu}_m \xrightarrow{\mathrm{P}} \mu$	weak convergence in probability
$\bar{\mu}_m \xrightarrow{\mathrm{a.s.}} \mu$	strong convergence almost surely
$K(\boldsymbol{\theta}, \boldsymbol{\phi})$	kernel function
$\hat{F}_n(x)$	empirical distribution function

Statistics

Y_{com}	complete or augmented data
Y_{obs}	observed data
$Y_{\mathrm{mis}}, Z, \mathbf{z}$	missing data, latent variable, latent vector
$\boldsymbol{\theta}, \boldsymbol{\Theta}$	parameter vector, parameter space
$f(x; \boldsymbol{\theta})$	pmf or pdf of the population r.v. X
$L(\boldsymbol{\theta}\|Y_{\mathrm{obs}})$	observed-data likelihood function of $\boldsymbol{\theta}$
$\ell(\boldsymbol{\theta}\|Y_{\mathrm{obs}})$	observed-data log-likelihood function of $\boldsymbol{\theta}$
$\ell(\boldsymbol{\theta}\|Y_{\mathrm{obs}}, z)$	complete-data log-likelihood function of $\boldsymbol{\theta}$
$f(z\|Y_{\mathrm{obs}}, \boldsymbol{\theta})$	conditional predictive distribution
$E(Z\|Y_{\mathrm{obs}}, \boldsymbol{\theta})$	conditional expectation
$\nabla\ell(\boldsymbol{\theta}\|Y_{\mathrm{obs}})$	score vector $\partial\ell(\boldsymbol{\theta}\|Y_{\mathrm{obs}})/\partial\boldsymbol{\theta}$
$\boldsymbol{I}(\boldsymbol{\theta}\|Y_{\mathrm{obs}})$	observed information matrix $-\nabla^2\ell(\boldsymbol{\theta}\|Y_{\mathrm{obs}})$
$\boldsymbol{J}(\boldsymbol{\theta})$	Fisher/expected information matrix $E\{\boldsymbol{I}(\boldsymbol{\theta}\|Y_{\mathrm{obs}})\}$

$\hat{\boldsymbol{\theta}}$	MLE of $\boldsymbol{\theta}$
$\boldsymbol{\theta}^{(t)}$	the t-th approximation of $\hat{\boldsymbol{\theta}}$
$\mathrm{Se}(\hat{\boldsymbol{\theta}})$	standard error of $\hat{\boldsymbol{\theta}}$
$\mathrm{Cov}(\hat{\boldsymbol{\theta}})$	covariance matrix of $\hat{\boldsymbol{\theta}}$
$\boldsymbol{I}_{\mathrm{com}}$	complete information $I(\hat{\boldsymbol{\theta}}\|Y_{\mathrm{obs}})$
$\boldsymbol{I}_{\mathrm{obs}}$	observed information
$\boldsymbol{I}_{\mathrm{mis}}$	missing information
$\pi(\boldsymbol{\theta})$	prior density of $\boldsymbol{\theta}$
$p(\boldsymbol{\theta}\|Y_{\mathrm{com}})$	complete-data posterior density of $\boldsymbol{\theta}$
$p(\boldsymbol{\theta}\|Y_{\mathrm{obs}})$	observed-data posterior density of $\boldsymbol{\theta}$
$\tilde{\boldsymbol{\theta}}$	mode of the posterior density $p(\boldsymbol{\theta}\|Y_{\mathrm{obs}})$
$\mathrm{logit}(p)$	logit transformation $\log\{p/(1-p)\}$
$\mathrm{KL}(g\|h)$	Kullback–Leibler divergence between pdfs $g(\cdot)$ and $h(\cdot)$
$J(\boldsymbol{x} \to \boldsymbol{y})$	Jacobian determinant from $\boldsymbol{x}$ to $\boldsymbol{y}$
z_α	upper α-th quantile of $N(0,1)$
$t(\alpha, \nu)$	upper α-th quantile of $t(\nu)$
$\chi^2(\alpha, \nu)$	upper α-th quantile of $\chi^2(\nu)$
$f(\alpha, \nu_1, \nu_2)$	upper α-th quantile of $F(\nu_1, \nu_2)$

References

Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis.* Cambridge University Press, London.

Albert, J.H. (1993). Teaching Bayesian statistics using sampling methods and MINITAB. *Am. Statist.* **47**, 182–191.

Albert, J.H. and Gupta, A.K. (1983). Estimation in contingency tables using prior information. *J. Roy. Statist. Soc., B* **45**, 60–69.

Albert, J.H. and Gupta, A.K. (1985). Bayesian methods for binomial data with applications to a non-response problem. *J. Am. Statist. Assoc.* **80**, 167–174.

Arnold, S.F. (1993). Gibbs sampling. In *Handbook of Statistics: Computational Statistics* (C.R. Rao, ed.), Vol.9, 599–625. Elsevier Science Publishers B.V., New York.

Arnold, B.C. and Strauss, D. (1988). Bivariate distributions with exponential conditionals. *J. Am. Statist. Assoc.* **83**, 522–527.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 330–418. Reprinted with biographical note by G.A. Barnard, in *Biometrika* **45**, 293–315, 1958.

Becker, M.P., Yang, I. and Lange, K. (1997). EM algorithms without missing data. *Statistical Methods in Medical Research* **6**, 38–54.

Besag, J.E. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika* **76**, 183.

Besag, J.E. and Green, P.J. (1993). Spatial statistics and Bayesian computation (with discussions). *J. Roy. Statist. Soc., B* **55**, 25–37.

Bijleveld, C.C.J.H. and De Leeuw, J. (1991). Fitting longitudinal reduced-rank regression models by alternating least squares. *Psychometrika* **56**, 433–447.

Böhning, D. (1992). Multinomial logistic regression algorithm. *Ann. Institute Statist. Math.* **44**, 197–200.

Böhning, D. and Lindsay, B.G. (1988). Monotonicity of quadratic approximation algorithms. *Ann. Institute Statist. Math.* **40**, 641–663.

Box, G.E.P. and Muller, M.E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.* **29**, 610–611.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Addison-Wesley, Massachusetts.

Bulter, J.W. (1958). Machine sampling from given probability distributions. In *Symposium on Monte Carlo Methods* (M.A. Meyer, ed.). Wiley, New York.

Cameron, A.C. and Trivedi, P.K. (1986). Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics* **1**, 29–53.

Carlin, B.P., Gelfand, A.E. & Smith, A.F.M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.* **41**, 389–405.

Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *Am. Statist.* **46**, 167–174.

Casella, G. and Robert, C.P. (1998). Post-processing accept-reject samples: Recycling and rescaling. *J. Comput. Graph. Statist.* **7**(2), 139–157.

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *J. of Econometrics* **51**, 79–99.

Consul, P.C. (1989). *Generalized Poisson Distributions.* Marcel Dekker, New York.

Consul, P.C. and Jain, G.C. (1973). A generalization of the Poisson distribution. *Technometrics* **15**(4), 791–799.

Cox, D.R. (1972). Regression models and life tables (with discussions). *J. Roy. Statist. Soc., B* **74**, 187–220.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman & Hall, London.

De Leeuw, A.R. (1994). Block relaxation algorithms in statistics. In *Information Systems and Data Analysis* (H.H. Bock, W. Lenski & M.M. Richter, ed.), 308–325. Springer, Berlin.

De Leeuw, J. and Heiser, W.J. (1977). Convergence of correction matrix algorithms for multidimensional scaling. In *Geometric Representations of Relational Data* (J. C. Lingoes, E. Roskam & I. Borg, ed.), 735–752. Mathesis Press, Ann Arbor.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussions). *J. Roy. Statist. Soc., B* **39**, 1–38.

De Pierro, A.R. (1995). A modified EM algorithm for penalized likelihood estimation in emission tomography. *IEEE Transactions on Medical Imaging* **14**, 132–137.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* Springer, New York.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall, London.

Gaver, D.P. and O'Muircheartaigh, I.G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics* **29**, 1–15.

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc., B* **56**, 501–514.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

Gelfand, A.E., Smith, A.F.M. and Lee, T.M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Am. Statist. Assoc.* **87**, 523–532.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

Geman, S. and Geman D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337–348.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. and Kirby, A.J. (1993). Modeling complexity: Applications of Gibbs sampling in medicine (with discussions). *J. Roy. Statist. Soc., B* **55**, 39–52.

Givens, G.H. and Raftery, A.E. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *J. Am. Statist. Assoc.* **91**, 132–141.

Gnedenko, B.V. (1962). *The Theory of Probability*. Chelsea, New York.

Groer, P.G. and Pereira, C.A.DeB. (1987). Calibration of a radiation detector: Chromosome dosimetry for neturons. In *Probability and Bayesian Statistics* (R. Viertl, ed.), 225–252. Plenum Press, New York.

Hasselblad, V., Stead, A.G. and Crenson, J.P. (1980). Multiple probit analysis with a non-zero background. *Biometrics* **36**, 650–663.

Heiser, W.J. (1987). Correspondence analysis with least absolute residuals. *Comput. Statist. and Data Analysis* **5**, 337–356.

Heiser, W.J. (1995). Convergent computing by iterative majorization: Theory and applications in multidimensional data analysis. In *Recent Advances in Descriptive Multi. Analysis* (W.J. Krzanowski, ed.), 157–189. Clarendon Press, Oxford.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Hunter, D.R. and Lange, K. (2000). Rejoinder to discussions of "Optimization transfer using surrogate objective functions". *J. Comput. Graph. Statist.* **9**, 52–59.

Hunter, D.R. and Lange, K. (2004). A tutorial on MM algorithms. *Am. Statist.* **58**, 30–37.

Jarrett, R.G. (1979). A note on the intervals between coal–mining disasters. *Biometrika* **66**, 191–193.

Jeffreys, H. (1961). *Theory of Probability* (3rd Ed.). Oxford University Press, Oxford.

Johnson, N.L. and Kotz, S. (1969–1972). *Distributions in Statistics* (4 Vols.). Wiley, New York.

Johnson, N.L., Kotz, S. and Kemp, A.W. (1992). *Univariate Discrete Distributions* (2nd Ed.). Wiley, New York.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Vol.1* (2nd Ed.). Wiley, New York.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol.2* (2nd Ed.). Wiley, New York.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–795.

Khintchine, A.Y. (1938). On unimodal distributions. *Tomsk. Universitet. Nauchno-issledovatel' skii institut matematiki i mek-haniki IZVESITIIA* **2**, 1–7.

Kiers, H.A.L. and Ten Berge, J.M.F. (1992). Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika* **57**, 371–382.

Lange, K. (1999). *Numerical Analysis for Statisticians.* Springer, New York.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis* (2nd Ed.). Springer, New York.

Lange, K. and Fessler, J.A. (1995). Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Transactions on Image Processing* **4**, 1430–1438.

Lange, K., Hunter, D.R. and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussions). *J. Comput. Graph. Statist.* **9**, 1–20.

Larsen, R.I., Gardner, D.E. and Coffin, D.L. (1979). An air quality data analysis system for interrelating effects, standards and needed source reductions. Part 5: No.2, Mortality of mice. *J. Air. Pollut. Control Assoc.* **39**, 113–117.

Lee, E.T. (1974). A computer program for linear logistic regression analysis. *Computer Program in Biomedicine* **4**, 80–92.

Leonard, T. (2000). *A Course in Categorical Data Analysis.* Chapman & Hall/CRC, Boca Raton.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd Ed.). Wiley, New York.

Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computation with application to a gene-regulation problem. *J. Am. Statist. Assoc.* **89**(427), 958–966.

Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer, New York.

Liu, J.S., Wong, W.H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

Liu, J.S., Wong, W.H. and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc., B* **57**, 157–169.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc., B* **44**, 226–233.

Maguire, B.A., Pearson, E.S. and Wynn, A.H.A. (1952). The time intervals between industrial accidents. *Biometrika* **38**, 168–180.

Marshall, A. (1956). The use of multi-stage sampling schemes in Monte Carlo computations. In *Symposium on Monte Carlo Methods* (M. Mayer, ed.), 123–140. Wiley, New York.

Marshall, A.W. and Olkin, I. (1979). *Inequality: Theory of Majorization and Its Applications.* Academic, San Diego.

McAllister, M.K. and Ianelli, J.N. (1997). Bayesian stock assessment using catchage data and the sampling/importance resampling algorithm. *Canad. J. Fisheries and Aquatic Sci.* **54**, 284–300.

McAllister, M.K., Pikitch, E.K., Punt, A.E. and Hilborn, R. (1994). A Bayesian approach to stock assessment and harvest decisions using the sampling importance resampling algorithm. *Canad. J. Fisheries and Aquatic Sci.* **51**, 2673–2687.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd Ed.). Chapman & Hall/CRC, Boca Raton.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley, New York.

Meng, X.L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–278.

Meng, X.L. and van Dyk, D. (1997). The EM algorithm — an old folk-song sung to a fast new tune (with discussions). *J. Roy. Statist. Soc., B* **59**, 511–567.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *J. Am. Statist. Assoc.* **44**, 335–341.

Michael, J.R., Schucany, W.R. and Hass, R.W. (1976). Generating random variates using transformations with multiple roots. *Am. Statist.* **30**, 88–90.

Murray, G.D. (1977). Contribution to the discussion of paper by A.P. Dempster, N.M. Laird & D.B. Rubin. *J. Roy. Statist. Soc., B* **39**, 27-28.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussions). *J. Roy. Statist. Soc., B* **56**, 3–48.

Nyquist, H., Rice, S.O. and Riordan, J. (1954). The distribution of random determinants. *Quarterly of Appl. Math.* **42**, 97–104.

Odell, P.L. and Feiveson, A.H. (1966). A numerical procedure to generate a sample covariance matrix. *J. Am. Statist. Assoc.* **61**, 199–203.

Ong, S.H. (1988). A discrete Charlier series distribution. *Biometrical Journal* **30**(8), 1003–1009.

Ortega, J.M. and Rheinboldt, W.C. (1970). *Iterative Solutions of Nonlinear Equations in Several Variables.* Academic, New York.

Philippe, A. (1997a). Simulation output by Riemann sums. *J. Statist. Comput. Simul.* **59**(4), 295–314.

Philippe, A. (1997b). Importance sampling and Riemann sums. *Prepub. IRMA* **43**, VI, Université de Lille.

Raftery, A.E., Givens, G.H. and Zeh, J.E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *J. Am. Statist. Assoc.* **90**, 402–416.

Rao, C.R. (1973). *Linear Statistical Inferences and Its Applications* (2nd Ed.). Wiley, New York.

Robert, C.P. (1995). Simulation of truncated normal variables. *Statist. Comput.* **5**, 121–125.

Robert, C.P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.

Roberts, G.O. and Polson, N.G. (1994). On the geometric convergence of the Gibbs sampler. *J. Roy. Statist. Soc., B* **56**, 377–384.

Ross, S.M. (1983). *Stochastic Processes*. Wiley, New York.

Rubin, D.B. (1987a). *Multiple Imputation for Non-response in Surveys*. Wiley, New York.

Rubin, D.B. (1987b). Comments on "The calculation of posterior distributions by data augmentation" M.A. Tanner & W.H. Wong. *J. Am. Statist. Assoc.* **82**, 543–546.

Rubin, D.B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussions). In *Bayesian Statistics*, Vol.3 (J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds.), 395–402. Oxford University Press, Oxford.

Rubinstein, R.Y. and Kroese, D.P. (2004). *The Cross–Entropy Method*. Springer, New York.

Schervish, M.J. and Carlin, B.P. (1992). On the convergence of successive substitution sampling. *J. Comput. Graph. Statist.* **1**, 111–127.

Schmee, J. and Hahn, G.J. (1979). A simple method for regression analysis with censored data. *Technometrics* **21**, 417–434.

Schukken, Y.H., Casella, G. and van den Broek, J. (1991). Overdispersion in clinical mastitis data from dairy herds: A negative binomial approach. *Preventive Veterinary Medicine* **10**, 239–245.

Shuster, J. (1968). On the inverse Gaussian distribution function. *J. Am. Statist. Assoc.* **63**, 1514–1516.

Skare, Ø., Bølviken, E. and Holden, L. (2003). Improved sampling importance resampling and reduced bias importance sampling. *Scand. J. Statist.* **30**, 719–737.

Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling–resampling perspective. *Am. Statist.* **46**, 84–88.

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussions). *J. Roy. Statist. Soc., B* **55**, 3–23.

Tan, M., Tian, G.L. and Ng, K.W. (2003). A non-iterative sampling method for computing posteriors in the structure of EM-type algorithms. *Statistica Sinica* **13**(3), 625–639.

Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (3rd Ed.). Springer-Verlag, New York.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussions). *J. Am. Statist. Assoc.* **82**, 528–550.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82–86.

von Neumann, J. (1951). Various techniques used in connection with random digits. *National Bureau of Standards Appl. Math. Series* **12**, 36–38.

Wei, G.C.G. and Tanner, M.A. (1990). Posterior computations for censored regression data. *J. Am. Statist. Assoc.* **85**, 829–839.

Wilks, S.S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Ann. Math. Statist.* **2**, 163–195.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.

Yakowitz, S., Krimmel, J.E. and Szidarovszky, F. (1978). Weighted Monte Carlo integration. *SIAM J. Numer. Anal.* **15**(6), 1289–1300.

Subject Index