MAT7035: Computational Statistics

Midterm Test
(16:20–18:20, 12/03/2018)

1. **(15 marks)** Use the inversion method to generate a random variable from the following distribution, and write down the algorithm:

   (a) (Poisson distribution) The *probability mass function* (pmf) is
   $$p_i = \Pr(X = i) = \frac{\lambda^i \, \mathrm{e}^{-\lambda}}{i!}, \qquad i = 0, 1, \ldots, +\infty.$$

   (b) (Laplace distribution) The density function is
   $$\mathrm{Laplace}(x|\mu, \sigma^2) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right),$$
   where $x \in \mathbb{R} \doteq (-\infty, +\infty)$, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+ \doteq (0, \infty)$.

2. **(20 marks)** Suppose that we want to draw random samples from the target density $f(x)$ with support $\mathcal{S}_X$. Furthermore, we assume that there exist an envelope constant $c\,(\geqslant 1)$ and an envelope density $g(x)$ having the same support $\mathcal{S}_X$ such that $f(x) \leqslant cg(x)$ for all $x \in \mathcal{S}_X$.

   (a) State the rejection algorithm for generating one random sample X from $f(x)$. **[5 marks]**

   (b) Using the Laplace density $g_\sigma(x) = \mathrm{Laplace}(x|0, \sigma^2)$ for $x \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$ as the envelope function to generate one random sample from the standard normal density
   $$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R},$$
   by the rejection method. **[12 marks]**

(c) What is the acceptance probability for this rejection algorithm? **[3 marks]**

3. **(6 marks)** Write down the relationships of the *stochastic representation* (SR) of random variables between the following distributions.

   (a) Generate $\chi^2(n)$ from $N(0,1)$;

   (b) Generate $t(n)$ from $N(0,1)$ and $\chi^2(n)$;

   (c) Generate $F(m,n)$ from $\chi^2(m)$ and $\chi^2(n)$;

   (d) Generate Gamma(n,β) from Exponential(β);

   (e) Generate Beta(α_1,α_2) from Gamma(α_1,β) and Gamma(α_2,β);

   (f) Generate Log-normal(μ,σ^2) from $N(\mu,\sigma^2)$.

   HINT: (1) The density of $X \sim$ Gamma(α,β) is

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}, \quad x \in \mathbb{R}_+.$$

   (2) Gamma$(1,\beta) =$ Exponential(β).
   (3) The density of $X \sim$ Log-normal(μ,σ^2) is

$$\frac{1}{\sqrt{2\pi}\,\sigma x} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right], \quad x \in \mathbb{R}_+.$$

4. **(9 marks)** Let $X = (X_1, X_2, X_3)^\top \sim$ Multinomial$(n; p_1, p_2, p_3)$ with the joint pmf

$$\Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \binom{n}{x_1, x_2, x_3} \prod_{i=1}^{3} p_i^{x_i},$$

   where $x_i \geqslant 0$, $\sum_{i=1}^3 x_i = n$, $p_i \geqslant 0$ and $\sum_{i=1}^3 p_i = 1$.

   (a) Find the marginal distribution of X_1. **[3 marks]**

   (b) Derive the conditional distribution of $X_2|(X_1 = x_1)$. **[3 marks]**

(c) State the conditional sampling method for generating one random sample for $X = (X_1, X_2, X_2)^\top$. **[3 marks]**

5. **(20 marks)** Let $y_1, \ldots, y_n$ be the corresponding realizations of independent random variables $Y_1, \ldots, Y_n$, and

$$Y_i \sim \text{Poisson}(\lambda_i),$$

$$\log(\lambda_i) = \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}, \quad 1 \leqslant i \leqslant n,$$

where $\boldsymbol{x}_{(i)}$ is $q \times 1$ covariates vector, and $\boldsymbol{\theta}_{q\times 1}$ is unknown parameter vector.

(a) Derive the score vector and the observed information matrix. **[15 marks]**

(b) Using the Newton-Raphson algorithm to find the MLE $\hat{\boldsymbol{\theta}}$ and the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. **[5 marks]**

6. **(30 marks)** Let $Y_{\text{obs}} = \{n_1, n_2, n_3, n_4; n_{12}, n_{34}\}$ denote the observed frequencies and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top$ be the cell probability vector satisfying $\theta_i > 0, \theta_1 + \cdots + \theta_4 = 1$. Suppose that the observed-data likelihood function of $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) \propto \left(\prod_{i=1}^{4} \theta_i^{n_i}\right) (\theta_1 + \theta_2)^{n_{12}} (\theta_3 + \theta_4)^{n_{34}}.$$

(a) Use the EM algorithm to find the MLEs $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. **[18 marks]**

(b) Combine the E-step with the M-step of the EM algorithm in (a), please show that

$$\hat{\theta}_1 + \hat{\theta}_2 = \theta_1^{(t+1)} + \theta_2^{(t+1)} = \frac{n_1 + n_2 + n_{12}}{n_1 + n_2 + n_3 + n_4 + n_{12} + n_{34}},$$

which does not depend on t, where $\theta_i^{(t+1)}$ denotes the $(t+1)$-th approximation of $\hat{\theta}_i$. Then derive the closed-form expressions for the MLEs $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. **[4 marks]**

(c) Derive the closed-form expressions for the restricted MLEs $\hat{\boldsymbol{\theta}}^R$ of $\boldsymbol{\theta}$ subject to the equality constraint $\theta_1\theta_4/(\theta_2\theta_3) = 1$. [**4 marks**]

(d) Let $n_1 = 164$, $n_2 = 164$, $n_3 = 103$, $n_4 = 221$, $n_{12} = 43$ and $n_{34} = 59$. Based on the results obtained in (b) and (c), calculate the values of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^R$. [**4 marks**]

=== END OF THE PAPER ===

1. **Solution**. (a) See **Example T1.2** in Tutorial 1.

   (b) See **Example T2.2** in Tutorial 2.

2. **Solution**. See **Example 1.4** in page 6 of Lecture Notes Chapter 1.

3. **Solution**. See **Example T3.1** in Tutorial 3.

4. **Solution**. See **Example T3.4** in Tutorial 3.

5. **Solution**. See **Example T4.3** in Tutorial 4.

6. **Solution**. (a) The observed-data likelihood function of $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) \propto \left(\prod_{i=1}^{4} \theta_i^{n_i}\right) \times (\theta_1+\theta_2)^{n_{12}}(\theta_3+\theta_4)^{n_{34}}.$$

By writing $n_{12} = Z_1 + Z_2$ with $Z_2 \equiv n_{12} - Z_1$ and $n_{34} = Z_3 + Z_4$ with $Z_4 \equiv n_{34} - Z_3$, a natural latent vector $Z = (Z_1, Z_3)^\top$ can be introduced so that the likelihood function for the complete-data $\{Y_{\text{obs}}, Z\}$ is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}, Z) \propto \prod_{i=1}^{4} \theta_i^{n_i+Z_i}.$$

Thus, the MLEs of $\boldsymbol{\theta}$ based on the complete data are given by

$$\hat{\theta}_i = \frac{n_i + Z_i}{N}, \quad i = 1, 2, 3, 4, \tag{1.1}$$

where $N = n_1 + n_2 + n_3 + n_4 + n_{12} + n_{34}$.

On the other hand, note that when Y_{obs} and $\boldsymbol{\theta}$ are given, Z_1 and Z_3 are independent binomially distributed. Thus, the conditional predictive distribution is

$$\begin{aligned} f(Z|Y_{\text{obs}}, \boldsymbol{\theta}) &= \text{Binomial}(Z_1|n_{12},\ \theta_1/(\theta_1+\theta_2)) \\ &\quad \times \text{Binomial}(Z_3|n_{34},\ \theta_3/(\theta_3+\theta_4)). \end{aligned}$$

Thus, the E-step of the EM algorithm is to compute the conditional expectations

$$E(Z_1|Y_{\rm obs}, \boldsymbol{\theta}) = \frac{n_{12}\theta_1}{\theta_1 + \theta_2} \quad \text{and} \quad E(Z_3|Y_{\rm obs}, \boldsymbol{\theta}) = \frac{n_{34}\theta_3}{\theta_3 + \theta_4}, \tag{1.2}$$

and the M-step is to update (1.1) by replacing Z_1 and Z_3 with $E(Z_1|Y_{\rm obs}, \boldsymbol{\theta})$ and $E(Z_3|Y_{\rm obs}, \boldsymbol{\theta})$, respectively.

(b) From (1.1) and (1.2), we have

$$\theta_1 = \frac{n_1 + n_{12}\theta_1/(\theta_1 + \theta_2)}{N}, \tag{1.3}$$

$$\theta_2 = \frac{n_2 + n_{12}\theta_2/(\theta_1 + \theta_2)}{N}, \tag{1.4}$$

$$\theta_3 = \frac{n_3 + n_{34}\theta_3/(\theta_3 + \theta_4)}{N}, \tag{1.5}$$

$$\theta_4 = \frac{n_4 + n_{34}\theta_4/(\theta_3 + \theta_4)}{N}, \tag{1.6}$$

By adding both sides of (1.3) with those of (1.4), we obtain

$$\theta_1 + \theta_2 = \frac{n_1 + n_2 + n_{12}}{N}. \tag{1.7}$$

Replacing $\theta_1 + \theta_2$ in the right-hand side of (1.3) with (1.7), we obtain

$$\theta_1 = \frac{n_1}{N} \cdot \frac{n_1 + n_2 + n_{12}}{n_1 + n_2}.$$

Replacing $\theta_1 + \theta_2$ in the right-hand side of (1.4) with (1.7), we obtain

$$\theta_2 = \frac{n_2}{N} \cdot \frac{n_1 + n_2 + n_{12}}{n_1 + n_2}.$$

Similarly, we have

$$\begin{aligned} \theta_3 + \theta_4 &= \frac{n_3 + n_4 + n_{34}}{N}, \\ \theta_3 &= \frac{n_3}{N} \cdot \frac{n_3 + n_4 + n_{34}}{n_3 + n_4}, \\ \theta_4 &= \frac{n_4}{N} \cdot \frac{n_3 + n_4 + n_{34}}{n_3 + n_4}. \end{aligned}$$

(c) From $\theta_4 = 1 - \theta_1 - \theta_2 - \theta_3$ and

$$\theta_3 = \frac{\theta_1\theta_4}{\theta_2} = \frac{\theta_1(1-\theta_1-\theta_2-\theta_3)}{\theta_2},$$

we have

$$\theta_3 = \theta_1\left(\frac{1}{\theta_1+\theta_2} - 1\right) \quad \text{and} \quad \theta_4 = \theta_2\left(\frac{1}{\theta_1+\theta_2} - 1\right). \tag{1.8}$$

Then, the likelihood function is

$$\begin{aligned} L(\theta_1,\theta_2) &= \theta_1^{n_1}\theta_2^{n_2}\cdot\theta_1^{n_3}\left(\frac{1}{\theta_1+\theta_2}-1\right)^{n_3}\cdot\theta_2^{n_4}\left(\frac{1}{\theta_1+\theta_2}-1\right)^{n_4} \\ &\quad \cdot(\theta_1+\theta_2)^{n_{12}}\cdot(1-\theta_1-\theta_2)^{n_{34}} \\ &= \theta_1^{n_1+n_3}\theta_2^{n_2+n_4}\cdot(\theta_1+\theta_2)^{n_{12}-n_3-n_4}\cdot(1-\theta_1-\theta_2)^{N_1}, \end{aligned}$$

where $N_1 \hat{=} n_3 + n_4 + n_{34}$. The log-likelihood function is given by

$$\begin{aligned} \ell(\theta_1,\theta_2) &= (n_1+n_3)\log(\theta_1) + (n_2+n_4)\log(\theta_2) \\ &\quad + N_1\log(1-\theta_1-\theta_2) + (n_{12}-n_3-n_4)\log(\theta_1+\theta_2). \end{aligned}$$

Therefore, let

$$\frac{\partial\ell}{\partial\theta_1} = \frac{n_1+n_3}{\theta_1} - \frac{N_1}{1-\theta_1-\theta_2} + \frac{n_{12}-n_3-n_4}{\theta_1+\theta_2} = 0, \tag{1.9}$$

$$\frac{\partial\ell}{\partial\theta_2} = \frac{n_2+n_4}{\theta_2} - \frac{N_1}{1-\theta_1-\theta_2} + \frac{n_{12}-n_3-n_4}{\theta_1+\theta_2} = 0, \tag{1.10}$$

we obtain

$$\frac{n_1+n_3}{\theta_1} = \frac{n_2+n_4}{\theta_2} = \frac{n}{\theta_1+\theta_2}, \tag{1.11}$$

where $n = \sum_{i=1}^4 n_i$. So, replacing $(n_1+n_3)/\theta_1$ in (1.9) by $n/(\theta_1+\theta_2)$, we have

$$\begin{aligned} & \frac{n}{\theta_1+\theta_2} - \frac{N_1}{1-\theta_1-\theta_2} + \frac{n_{12}-n_3-n_4}{\theta_1+\theta_2} = 0 \\ \Rightarrow\ & \frac{n_1+n_2+n_{12}}{\theta_1+\theta_2} = \frac{N_1}{1-\theta_1-\theta_2} = \frac{N_0+N_1}{1} \\ \Rightarrow\ & \theta_1+\theta_2 = \frac{N_0}{N_0+N_1}, \end{aligned} \tag{1.12}$$

where $N_0 \hat{=} n_1 + n_2 + n_{12}$. From (1.12), (1.11) and (1.8), we obtain

$$\theta_1 = \frac{n_1 + n_3}{n} \cdot \frac{N_0}{N_0 + N_1}, \quad \theta_2 = \frac{n_2 + n_4}{n} \cdot \frac{N_0}{N_0 + N_1},$$
$$\theta_3 = \frac{n_1 + n_3}{n} \cdot \frac{N_1}{N_0 + N_1}, \quad \theta_4 = \frac{n_2 + n_4}{n} \cdot \frac{N_1}{N_0 + N_1}.$$

Hence, the restricted MLEs are given by

$$\hat{\theta}_1^R = \frac{n_1 + n_3}{\sum_{i=1}^4 n_i} \cdot \frac{n_1 + n_2 + n_{12}}{N},$$
$$\hat{\theta}_2^R = \frac{n_2 + n_4}{\sum_{i=1}^4 n_i} \cdot \frac{n_1 + n_2 + n_{12}}{N},$$
$$\hat{\theta}_3^R = \frac{n_1 + n_3}{\sum_{i=1}^4 n_i} \cdot \frac{n_3 + n_4 + n_{34}}{N},$$
$$\hat{\theta}_4^R = \frac{n_2 + n_4}{\sum_{i=1}^4 n_i} \cdot \frac{n_3 + n_4 + n_{34}}{N}.$$

where $N = \sum_{i=1}^4 n_i + n_{12} + n_{34}$.

(d) $\hat{\theta}_1 = 0.2460$, $\hat{\theta}_2 = 0.2460$, $\hat{\theta}_3 = 0.1615$, $\hat{\theta}_4 = 0.3465$.

$\hat{\theta}_1^R = 0.2014959$, $\hat{\theta}_2^R = 0.2905465$, $\hat{\theta}_3^R = 0.2080133$, $\hat{\theta}_4^R = 0.2999443$.

```
function(ind)
{   # Function name: MT3317.Q5(ind)
    # -------- Input --------------------------------
    # ind = 1: computae the MLEs of \theta_i
    # ind = 2: compute the restricted MLEs of \theta_i
    # -------- Output -------------------------------
    # MLEs or Restricted MLEs
    nv <- c(164, 164, 103, 221)
```

```
    n12 <- 43
    n34 <- 59
    N0 <- nv[1] + nv[2] + n12
    N1 <- nv[3] + nv[4] + n34
    N <- N0 + N1
    th <- rep(0, 4)
    if (ind==1) {
       m12 <- nv[1] + nv[2]
       m34 <- nv[3] + nv[4]
       th[1] <- nv[1]*N0/(N*m12)
       th[2] <- nv[2]*N0/(N*m12)
       th[3] <- nv[3]*N1/(N*m34)
       th[4] <- nv[4]*N1/(N*m34)
    }
    else {
       m13 <- nv[1] + nv[3]
       m24 <- nv[2] + nv[4]
       a <- sum(nv)*N
       th[1] <- m13*N0/a
       th[2] <- m24*N0/a
       th[3] <- m13*N1/a
       th[4] <- m24*N1/a
    }
    return(th)
}
```