Review for the Midterm Test

Ch 1. Generation of Random Variables
Ch 2. Optimization

Midterm Test 2021

- **Time and Date**
  16:20 – 18:20,
  December 20 (Monday), 120 minutes
- **Venue**
  Lychee Hills Building No. 2, Room 101
- **Range**
  Chapters 1 and 2
- **Assessment**
  - **Each assignment (5%)**
  - **Midterm test (25%)**

Key Points in the Midterm Test 2021

1. **Use the inverse method to generate a discrete/continuous r.v. (20 marks);**
2. **Use the rejection algorithm to generate a continuous r.v. (20 marks);**
3. **State SIR algorithm; Use the SR, conditional sampling method to generate a multivariate r.v. (10 marks);**
4. **The Newton–Raphson and Fisher scoring algorithms (20 marks).**
5. **The EM, (ECM, MM) algorithms (30 marks)**

The Policy of Closed Book Midterm Test

- Please bring one calculator and check the battery.
- Please bring two pens/pencils in case one is not available.
- You can prepare anything on one side of an A4 paper and bring it with you to the test venue.
- You are not allowed to bring any other material (including iPhone/iPad) to the test venue.

Ch 1. Generation of Random Variables

- §1.1 The Inversion Method
- §1.2 The Grid Method
- §1.3 The Rejection Method
- §1.4 The SIR Method
- §1.5 The SR Method
- §1.6 The CS Method

§1.1 Three Key Points to The Inversion Method

1. **The algorithm for the continuous r.v.:**

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

   Step 2. Return $x = F^{-1}(u)$.

2. **How to generate samples from a discrete distribution using the inversion method (please review all examples in Subsection 1.1.2).**

3. **The built-in R function:**
   `sample`(x, N, **prob** $= p$, **replace** $=$ **T/F**)
   (please review Appendix A.1.1).

Remarks on the Inversion Method

1° **Applicability:** **If F^{-1} has an explicit expression, the inversion method is the most efficient method to generate a random sample of X from $F(\cdot)$.**

2° **Inapplicability:** **If F^{-1} is not available analytically, the inversion method may not be efficient.**

3° **The inversion method is a special case of the stochastic representation (SR) method:**

$$X \stackrel{\mathrm{d}}{=} F^{-1}(U) \stackrel{\mathrm{d}}{=} F^{-1}(1-U).$$

§1.2 The Grid Method

- **The essence of the grid method is to generate a finite discrete distribution:**

$$X \dot{\sim} \mathbf{FDiscrete}_d(\{x_i\}, \{p_i\}).$$

where $\{x_i\}_{i=1}^d$ **is a set of grid points, covering the support** $\mathcal{S}_X$**,**

- **and**

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

Remarks on the Grid Method

1° **Applicability:** **If (i) the normalizing constant of** $f_X(x)$ **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** $f_X(\cdot)$**.**

2° **Inapplicability:** **If** $\mathcal{S}_X$ **is an infinite interval, the grid method cannot be applied.**

§1.3 Four Key Points to The Rejection Method

1. **Several basic notions:** **Target density $f(x)$; Envelope constant c; Envelope density $g(x)$; Acceptance probability $1/c$.**

2. **The algorithm:**

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

   Step 2. **If $U \leqslant \frac{f(Y)}{cg(Y)}$, return $X = Y$; otherwise, go to Step 1.**

§1.3 Four Key Points to The Rejection Method (Cont'd)

3. Theoretical justification **(see page 17)**

4. How to find the optimal $c_{\rm opt}$ among

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

where $\{g_\theta(x) : \theta \in \Theta\}$ is a family of the candidate envelope densities indexed by a parameter θ.

(Please review all examples in Subsection 1.3.3)

§1.4 Two Points to the SIR Method

1. **Two basic notions:** **Target density** $f(x)$; **Important sampling density** $g(x)$.

2. **The algorithm:**

   Step 1. **Generate** $X^{(1)}, \ldots, X^{(J)} \overset{\text{iid}}{\sim} g(\cdot)$;

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

§1.5 A Key Point to The SR Method

1. **How to prove that a given multivariate random vector can be represented by other independent r.v.s.**

- **Prove that** $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim \text{Dirichlet}(a_1, \ldots, a_d)$ **has the following SR**

$$X_i \overset{\text{d}}{=} (1 - W_{i-1}) \prod_{j=i}^{d-1} W_j,\ i = 1, \ldots, d-1,$$

$$X_d \overset{\text{d}}{=} 1 - W_{d-1}$$

where $W_0 \equiv 0$, $\{W_i\}_{i=1}^{d-1} \overset{\text{ind}}{\sim} \text{Beta}(a_1 + \cdots + a_i,\ a_{i+1})$.

- **You only need to prove that**

$$f(\boldsymbol{x}_{-d}) = f(x_1, \ldots, x_{d-1}) = \frac{\Gamma(\sum_{i=1}^{d} a_i)}{\prod_{i=1}^{d} \Gamma(a_i)} \prod_{i=1}^{d} x_i^{a_i - 1}.$$

- **Note that**

$$f(\boldsymbol{x}_{-d}) = \left[\prod_{i=1}^{d-1} g_i(w_i)\right] \cdot |J(\boldsymbol{w}_{-d} \to \boldsymbol{x}_{-d})|,$$

where the Jacobian determinant is

$$J(\boldsymbol{w}_{-d} \to \boldsymbol{x}_{-d}) = \frac{\partial(w_1, \ldots, w_{d-1})}{\partial(x_1, \ldots, x_{d-1})}.$$

§1.6 The Conditional Sampling Method

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

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

- **To generate x from** $f_{\mathbf{x}}(\boldsymbol{x})$**, we only need to generate** x_1 **from the marginal density** $f_1(x_1)$**, then to generate** x_i **sequentially from the conditional density** $f_i(x_i|x_1, x_2, \ldots, x_{i-1})$**. (Review all examples in Section 1.6)**

Ch 2. Optimization

- §2.1 Rate of Convergence
- §2.2 The NR Algorithm
- §2.3 The EM Algorithm
- §2.4 The ECM Algorithm
- §2.5 MM Algorithms

§2.1 Rate of convergence

1. **Let an EM/MM algorithm can be represented by $\theta^{(t+1)} = h(\theta^{(t)})$, then, the rate of convergence of the EM/MM algorithm is defined by (2.9), i.e.,**

$$c = \lim_{t\to\infty} \frac{|\theta^{(t+1)} - \hat{\theta}|}{|\theta^{(t)} - \hat{\theta}|} = |h'(\hat{\theta})|,$$

   where $\hat{\theta}$ is the MLE of θ.

2. **Calculate the value of the rate of convergence of the EM algorithm in Example 2.6**

3. **Let an NR/FS algorithm can be represented by $\theta^{(t+1)} = f(\theta^{(t)})$, then, the rate of convergence of the NR/FS algorithm is defined by**

$$c = \lim_{t\to\infty} \frac{|\theta^{(t+1)} - \hat{\theta}|}{|\theta^{(t)} - \hat{\theta}|^2} = \frac{1}{2}|f''(\hat{\theta})|,$$

where $\hat{\theta}$ is the MLE of θ.

4. **Calculate the value of the rate of convergence of the NR/FS algorithm in Example 2.6**

§2.2 The NR Algorithm

1. **Several important notions:** **Score vector** $\nabla \ell(\boldsymbol{\theta}|Y_{\rm obs})$**; Observed information matrix** $\boldsymbol{I}(\boldsymbol{\theta}|Y_{\rm obs})$**; Fisher/expected information matrix** $\boldsymbol{J}(\boldsymbol{\theta})$**; Estimated asymptotic covariance matrix** $\widehat{\rm Cov}(\hat{\boldsymbol{\theta}})$**.**

2. **The NR algorithm**

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

3. **Application to logistic regression. (Please review §2.2.5)**

4. How to apply the NR/FS algorithms to the Poisson regression. [See Ex. T4.3]

Let $Y_{\text{obs}} = \{y_i\}_{i=1}^n$ and consider the following Poisson regression

$$\begin{aligned} Y_i &\overset{\text{ind}}{\sim} \textbf{Poisson}(\lambda_i), \\ \log(\lambda_i) &= \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}, \qquad 1 \leqslant i \leqslant n, \end{aligned}$$

then the log-likelihood function for $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = c + \sum_{i=1}^n \left\{ y_i(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) - \exp(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}) \right\}.$$

§2.3 The EM Algorithm

1. Summary of the EM algorithm

- **Augment the observed data Y_{obs} with latent variables Z.**
- **M-Step: Find the complete-data log-likelihood function $\ell(\boldsymbol{\theta}|Y_{\text{obs}}, z)$ and derive the complete-data MLE $\hat{\boldsymbol{\theta}}$.**
- **E-Step: Find the conditional predictive distribution $f(z|Y_{\text{obs}}, \boldsymbol{\theta})$ and compute the conditional expectation $E(Z_i|Y_{\text{obs}}, \boldsymbol{\theta})$ or/and $E(Z_i^2|Y_{\text{obs}}, \boldsymbol{\theta})$.**

2. The ascent property of the EM

- **The definition of the Q function**

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\{\ell(\boldsymbol{\theta}|Y_{\rm obs}, Z)|Y_{\rm obs}, \boldsymbol{\theta}^{(t)}\} \\ &= \int_{\mathbb{Z}} \ell(\boldsymbol{\theta}|Y_{\rm obs}, z) \times f(z|Y_{\rm obs}, \boldsymbol{\theta}^{(t)})\,{\rm d}z, \end{aligned}$$

- **The original definition of the EM**

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{2.18}$$

- **Using the Kullback–Leibler divergence to prove that**

$$\ell(\boldsymbol{\theta}^{(t+1)}|Y_{\rm obs}) - \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) \geqslant 0.$$

3. Missing information principle

- **The observed information, the complete information, and the missing information:**

$$\boldsymbol{I}_{\text{obs}} = \boldsymbol{I}_{\text{com}} - \boldsymbol{I}_{\text{mis}}. \tag{2.38}$$

§2.5 MM Algorithms

1. The MM idea

- **Minorization.** **The function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is said to minorize $\ell(\boldsymbol{\theta}|Y_{\rm obs})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ if**

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \leqslant \ell(\boldsymbol{\theta}|Y_{\rm obs}) \quad \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}, \tag{2.43}$$

$$Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}). \tag{2.44}$$

- **Definition of an MM algorithm.**

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \tag{2.45}$$

- **The ascent property of an MM.**

2. The QLB algorithm

- **A key condition.** **There exists a positive definite matrix $\boldsymbol{B} > 0$ such that**

$$\nabla^2 \ell(\boldsymbol{\theta}|Y_{\rm obs}) + \boldsymbol{B} \geqslant 0 \qquad \forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta}. \quad (2.47)$$

- **Definition of the Q function.**

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^{\top} \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^{\top} \boldsymbol{B} (\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}. \quad (2.48)$$

- **The Definition of the QLB algorithm.**

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{B}^{-1} \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\rm obs}). \quad (2.49)$$

3. De Pierro's algorithm

- **A key condition.** **The log-likelihood function be of the form**

$$\ell(\boldsymbol{\theta}|Y_{\rm obs}) = \textstyle\sum_{i=1}^m f_i(\boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}),$$

where $\{f_i\}_{i=1}^m$ **are twice continuously differentiable and strictly concave functions defined in** $\mathbb{R}$**.**

- **Definition of the** Q **function.**

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^m \sum_{j\in\mathbb{J}_i} \lambda_{ij} f_i(\lambda_{ij}^{-1} x_{ij}(\theta_j - \theta_j^{(t)}) + \boldsymbol{x}_{(i)}^\top \boldsymbol{\theta}^{(t)}), \tag{2.52}$$

4. Jensen's Inequality

- **Let $\varphi(\cdot)$ be concave. If X is a r.v. taking values in the domain of $\varphi(\cdot)$, then**

$$\varphi[E(X)] \geqslant E[\varphi(X)],$$

provided that $E(X)$ and $E[\varphi(X)]$ exist.

- **Discrete version. For any concave function $f(\cdot)$,**

$$f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \geqslant \sum_{i=1}^{n} \alpha_i f(x_i),$$

where $\alpha_i \geqslant 0$ and $\sum_{i=1}^{n} \alpha_i = 1$.

5. The supporting hyperplane inequality

- **If $g(x)$ is convex; i.e., $g''(x) \geqslant 0$, we have**

$$g(x) \geqslant g(x_0) + (x - x_0)g'(x_0).$$

- **Please review all questions in Assignments 1-3.**

- **Please review all questions in Tutorials 1–7.**

End of the Review