Review and Summary

Dr. Qiuzhuang Sun

STAT3023

Moment generating functions (MGF)

The MGF $M_X(t) = E[\exp(tX)]$ encodes the sequence of moments $E[X^r]$, $r = 1, 2, \ldots$

- The MGF is defined provided the above expectation exists for t in some open interval containing zero
- $M_X^{(r)}(0) = E[X^r]$
- If the MGFs for X and Y exist, and $M_X(t) = M_Y(t)$, then X and Y have the same distribution
- Let $Z = X_1 + \cdots + X_n$ with $X_1, \ldots, X_n$ independent. We have $M_Z(t) = \prod_{i=1}^n M_{X_i}(t)$
- Assume the MGF of X is $M_X(t)$. Then $Z = aX + b$ has MGF $M_Z(t) = e^{tb} M_X(at)$

Multivariate distributions

- Two continuous random variables X, Y are independent iff there exist functions $g(x), h(y)$ so that $f_{X,Y}(x,y) = g(x)h(y)$
- Computing mean and variance by conditioning:
    - $E[Y] = E[E[Y|X]]$
    - $\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}(E[X|Y])$
- How to find the marginal distribution of a hierarchical model
    - The above formulas are useful
    - Use MGF or compute the marginal distribution directly
- If (X, Y) follows a bivariate normal distribution
    - Marginal distribution is normal
    - Conditional distribution $(Y|X)$ is normal
    - Zero correlation implies independence for bivariate normal distributions

Transformation of random variables

- For monotone function g, let $Y = g(X)$. Then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

- Let $U = h_1(X,Y)$ and $V = h_2(X,Y)$. If the transformation between (X,Y) and (U,V) is one-to-one, then

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v)) |\det(J)|,$$

where

$$J = \begin{bmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{bmatrix}$$

Transformation of random variables

- Samples from normal random variables: let $X_1, \ldots, X_n$ be iid random variables from $N(\mu, \sigma^2)$
  - The sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$
  - $\bar{X}_n$ and sample variance $s_n^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ are independent
  - $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$
- Let $Z \sim N(0,1)$ and $V \sim \chi^2_d$ be independent random variables. Then

  $$T = \frac{Z}{\sqrt{V/d}} \sim t_d,$$

  the t-distribution with d degrees of freedom

Exponential family

The exponential family has PDF or PMF

$$f(x|\theta) = h(x) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x) - A(\theta)\right), \quad x \in \mathbb{R}$$

- Let d be the number of elements in θ. If $d = k$, the PDF or PMF belongs to a full exponential family; if $d < k$, it belongs to the curved exponential family.
- Examples: Poisson, binomial with known n, normal...

The canonical form is

$$f(x|\theta) = h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x) - A^*(\boldsymbol{\eta})\right), \ \boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$$

Exponential family

Let $T(x) = [t_1(x), \ldots, t_k(x)]^\top$. Assume $X \sim f(x|\theta)$. Then we have

$$E[T(X)] = \frac{\partial A^*}{\partial \boldsymbol{\eta}^\top};$$

$$\mathrm{Var}(T(X)) = \frac{\partial^2 A^*}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top};$$

$$\log M_{T(X)}(\boldsymbol{s}) = A^*(\boldsymbol{s} + \boldsymbol{\eta}) - A^*(\boldsymbol{\eta})$$

Sufficiency

- A statistic $T = T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if the conditional distribution of $\mathbf{X}$ given the value of T does not depend on θ
- Factorization theorem: T is a sufficient statistic for θ iff the likelihood function is written in the following form:

$$L(\theta; \mathbf{X}) = g(T(\mathbf{X}; \theta))h(\mathbf{X})$$

- For exponential family with PDF or PMF

$$f(x|\theta) = h(x) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x) - A(\theta)\right),$$

a sufficient statistic for θ is $[\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)]$

Unbiasedness, consistency, and efficiency

- $\hat{\theta}$ is consistent for θ if $\lim_{n\to\infty} P(|\hat{\theta}-\theta| \geq \varepsilon) = 0$
- $\hat{\theta}$ is unbiased for θ if $\mathrm{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta = 0$
- The mean squared error of $\hat{\theta}$, $\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta}-\theta)^2]$, can be decomposed into

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}))^2$$

- Let $W(\mathbf{X})$ be an estimator and $f(x;\theta)$ be the PDF of each iid random variable. Under some regularity conditions,

$$\mathrm{Var}(W(X)) \geq \frac{\left\{\frac{\partial}{\partial\theta} E_\theta[W(X)]\right\}^2}{\mathrm{Var}\left(\frac{\partial}{\partial\theta}\log(f(X;\theta))\right)},$$

which is known to be the Cramer-Rao Lower Bound (CRLB)

Unbiasedness, consistency, and efficiency

- The CRLB is attained iff

$$\frac{\partial}{\partial\theta}\log(f(\mathbf{X};\theta)) = C_\theta(W(\mathbf{X}) - E[W(\mathbf{X})])$$

- Rao-Blackwell Theorem: Let $\hat{\theta}_1$ be an unbiased estimator for θ, and T be a sufficient statistic for θ. Then $\hat{\theta}_2 = E[\hat{\theta}_1|T]$ is unbiased for θ and uniformly more efficient than $\hat{\theta}_1$
- For all the unbiased estimators, the one having the minimum variance is called the minimum variance unbiased estimator (MVUE)
- For the full exponential family, any function of a sufficient statistic T is the MVUE for its expected value

(i) Simple vs. simple

- The NP lemma guarantees the existence of the most power test, where the test statistic is the likelihood ratio $\Upsilon = f_1(\mathbf{X})/f_0(\mathbf{X})$ with critical region $\{\mathbf{X} : f_1(\mathbf{X})/f_0(\mathbf{X}) \geq y_\alpha\}$, where y_α is chosen such that $P_0(\Upsilon \geq y_\alpha) = \alpha$
- Generalising to discrete distribution
  - Test function $\delta(\mathbf{X}) \in [0, 1]$
  - Randomised test

$$\delta(\mathbf{X}) = \begin{cases} 1, & \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > y_\alpha \\ \gamma, & \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} = y_\alpha \\ 0, & \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} < y_\alpha \end{cases}$$

  - γ, y_α are chosen such that $E_{\theta_0}[\delta(\mathbf{X})] = \alpha$

Hypothesis testing

(ii) Simple vs. composite

- Conditions for the existence of the UMP test for one-sided alternatives
  - Definition of the UMP test
  - Monotone likelihood ratio: (a) $f_{\theta_0} \neq f_{\theta_1}$ for any $\theta_0 < \theta_1$; (b) for any $\theta_0 < \theta_1$, $\frac{f_{\theta_1}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})}$ is an increasing function of $T(\mathbf{X})$
  - For $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$,

$$\delta(\mathbf{X}) = \begin{cases} 1, & T(\mathbf{X}) > C \\ \gamma, & T(\mathbf{X}) = C \\ 0, & T(\mathbf{X}) < C \end{cases}$$

  - Special case: 1-parameter exponential family with sufficient statistic $T(\mathbf{X})$

Hypothesis testing

(ii) Simple vs. composite

- Condition for the existence of a UMPU test for two-sided alternatives
  - Definition of a UMPU test
  - Special case: 1-parameter exponential family with sufficient statistic $T(\mathbf{X})$
  - GLRT for $H_0 : \theta = \theta_0$ vs $H_1 : \theta \in \Theta \backslash \{\theta_0\}$ with test statistic

$$\frac{\prod_{i=1}^{n} f_{\hat{\theta}}(X_i)}{\prod_{i=1}^{n} f_{\theta_0}(X_i)} \quad \text{with } \hat{\theta} = \underset{\theta \in \Theta}{\arg\max} \prod_{i=1}^{n} f_{\theta}(X_i) \text{ being the MLE}$$

  - Take L_n to be the logarithm of the test statistic, then $2L_n$ approximately follows χ_1^2 for large n (under "regular" models)

Hypothesis testing

(iii) Composite vs. composite

- Consider $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. If the family of distribution has monotone likelihood ratio, then the UMP test exists and is the same as testing $H_1 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$
- Similar for $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$
- Consider $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ (for some $\theta_1 < \theta_2$) vs $H_1 : \theta_1 < \theta < \theta_2$. If the distribution of interest belongs to a 1-parameter exponential family, then the UMPU test exists
- GLRT has test statistic $\frac{\prod_{i=1}^n f_{\hat{\theta}}(X_i)}{\prod_{i=1}^n f_{\hat{\theta}_0}(X_i)}$, where $\hat{\theta}$ is the MLE and $\hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} \ell(\theta; \mathbf{X})$ is the "restricted" MLE under the null

Decision theory

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be iid from $f_\theta(\cdot)$ where θ is unknown to be estimated and f_θ belongs to a family of distribution $\mathcal{F}$

- Key definitions: decision, decision space, loss function $L(d(\mathbf{X})|\theta)$, risk $R(\theta|d) = E_\theta[L(d(\mathbf{X})|\theta)]$
- Overall risk: Bayes risk (w.r.t. weight function or prior $w(\theta)$) and maximum risk (over a subset of parameter space)
- Optimal decisions
  - Bayes risk: the optimal procedure is called the Bayes procedure, under
    - squared error loss
    - absolute error loss
    - 0-1 loss
    - discrete selection

    (We need to first work out the posterior of θ)

Decision theory

- Optimal decisions
  - Maximum risk: the optimal procedure is called a minimax procedure
  - Finding minimax procedures:
    - Week 10 Theorem 1: by taking the limiting Bayes risk of a sequence of Bayes procedures
    - Week 10 Theorem 2: Bayes procedures with constant risk
- Asymptotic minimax procedures
  - AMLB theorem
    - Find the limiting (rescaled) risk of the Bayes procedure under uniform prior. In many cases, it is the same as that for Bayes procedure under flat prior
    - Check if the limiting (rescaled) risk is a continuous function of θ; if so, we can use the AMLB theorem to find a lower bound of the limiting maximum risk for any sequence of procedures
  - Show the procedures of interest exactly attains the lower bound