Expectation - Maximization Algorithm

Mingwei Fei

January 26, 2023

1 Jessen's Inequality

In a convex function, we have $E[f(x)] \geq f[E(X)]$. Associate this with the graph, the average of two points is high above the point in the middle.

This could be related to the likelihood function, as the optimal point is between two points, that we can approximate the likelihood of the optimal point by average the likelihood of the two points.

But log-likelihood function is concave, while not convex.

1.1 Application in EM Example

Consider an experiment with coin A that has a probability θ_A of heads, and a coin B that has a probability θ_B of tails. We draw m samples as follows - for each sample, pick one of the coins at random, flip it n times, and record the number of heads and tails (that sum to n). If we recorded which coin we used for each sample, we have complete information and can estimate θ_A and θ_B in closed form. To be very explicit, suppose we drew 5 samples with the number of heads and tails represented as a vector x, and the sequence of coins chosen was A,A,B,A,B.

(i) Complete Information with Maximum Likelihood Then the complete log likelihood is

$$logp(X, \theta) = logp(x_1, \theta_A) + logp(x_2, \theta_A) + logp(x_3, \theta_B) + logp(x_4, \theta_A) + logp(x_5, \theta_B)$$

We will use z_i to indicate the label of the ith coin, that is - whether we used coin A or B to gnerate the ith sample.

(ii) Incomplete information However, if we did not record the coin we used, we have missing data and the problem of estimating θ_B is harder to solve. One way to approach the problem is to ask - can we assign weights w_i to each sample according to how likely it is to be generated from coin A or coin B?

Remark: It is a common practice dealing with missing discrete variables, which use the posterior probability based on the data. That is how the Jessen's inequality come in the play. The missing information is an expectation in the log-likelihood function, so the expectation of the log-likelihood function is smaller than the log-likelihood function of the expected missing information.

With knowledge of w_i, we can maximize the likelihod to find θ . Similarly, given w_i, we can calculate what θ should be. So the basic idea behind Expectation Maximization (EM) is simply to start with a guess for θ, then calculate z, then update θ using this new value for z, and repeat till convergence. The derivation below shows why the EM algorithm using this "alternating" updates actually works.

A verbal outline of the derivtion - first consider the log likelihood function as a curve (surface) where the base is θ. Find another function Q of θ that is a lower bound of the log-likelihood but touches the log likelihodd function at some θ (E-step). Next find the value of θ that maximizes this function (M-step). Now find yet antoher function of θ that is a lower bound of the log-likelihood but touches the log likelihodd function at this new θ. Now repeat until convergence - at this point, the maxima of the lower bound and likelihood functions are the same and we have found the maximum log likelihood. See illustratioin below.

$$\phi^N(x) = \begin{cases} 1 & f_1(x) > kf_0(x) \\ 0 & f_1(x) < kf_0(x) \end{cases}$$

(iii) Proof

We can see ϕ^N has the following characteristics: (montone) If $\phi^N - \phi > 0$, then $\phi^N > 0$ and $f_1(x) \geq kf_0(x)$. If $\phi^N - \phi < 0$, then $\phi^N < 1$ and $f_1(x) \leq kf_0(x)$.

In any case,

$$[\phi^N - \phi][f_1(x) - kf_0(x)] \geq 0$$

and therefore,

$$\int [\phi^N - \phi][f_1(x) - kf_0(x)]dv \geq 0$$

$$\int [\phi^N - \phi]f_1(x)dv \geq \int [\phi^N - \phi]kf_0(x)dv$$

The left-hand side is $E_1(\phi^N) - E_1(\phi)$ and the right-hand side

$$k(E_0[\phi^N] - E_0[\phi]) = k(\alpha - E_0[\phi]) \geq 0$$

1.2 Example

The example just gives two distributions: Suppose that X is a sample of size 1, $P_0 = \{p_0\}$ and $P_1 = \{p_1\}$, where P_0 is $N(0,1)$ and P_1 is the double exponential distribution $DE(0,2)$ with the p.d.f. $4^{-1}e^{-|x|/2}$

Since $P\Big(f_1(x) = cf_0(x)\Big) = 0$, there is a unique nonrandomized UMP test.

$$\phi(x) = \begin{cases} 1 & \Big(4^{-1}e^{-|x|/2}\Big)^2 > k^2(2\pi)^{-\frac{1}{2}\times 2}\exp\Big(-x^2\Big) \\ 0 & f_1(x) < kf_0(x) \end{cases}$$

which is $|x| > t$ or $|x| < 1 - t$ for some $t > \frac{1}{2}$. Suppose that $\alpha < \frac{1}{3}$, to determine t, we use

$$\alpha = E_0[\phi] = P_0[|x| > t] + P_0[|x| < 1 - t]$$

(i) t should be larger than 1.

If $t \leq 1$, then $P_0(|x| > t) \geq P_0(|x| > 1) = 0.3374 > \alpha$

(ii) So the probability simplified

$$\alpha = P_0(|x| > t) = \Phi(-t) + 1 - \Phi(t)$$

thus, $t = \Phi^{-1}(1 - \alpha/2)$ and $\phi = I_{(t,\infty)}(|X|)$. Note, it is not necessary to find out what c is.

Another common example is in binomial distribution, which we will have randomized UMP test.

An interesting phenomenon is that ϕ is a test that does not depend on P_1. So it is the range of parameters in H_1.

2 Likelihood Ratio Test

Likelihood ratio test is associated with the Neyman-Pearson UMP test, it realizes on the monotone likelihood ratio and an extension of UMP test.

2.1 Definition

Let $l(\theta) = f_\theta(X)$ be the likelihood function. For testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, a likelihood ratio (LR) test is any test that rejects H_0 if and only if $\lambda(X) < c$, where $c \in [0,1]$ and $\lambda(X)$ is the likelihood ratio defined by

$$\lambda(X) = \frac{\sup_{\theta\in\Theta_0} l(\theta)}{\sup_{\theta\in\Theta} l(\theta)}$$

Question:

(i) Why we need to use the SUP here? Is it related to the monotone likelihood ratio? We could see that $\sup_{\theta\in\Theta} l(\theta)$ is a fixed number, comparing with $\sup_{\theta\in\Theta_1} l(\theta)$. While in the nominator, any likelihood would be less than $\sup_{\theta\in\Theta_0} l(\theta)$, so if we rejected the H_0 by the sup, any other θ_0 would be rejected as well.

(ii)

2.2 Relationship between LR test and UMP test

Why do we need to study this?

(i) Why do we use LR test while not UMP test?

One requirement for UMP test is to have monotone likelihood ratio, however ie. when $H_0 : a < \theta < b$, the likelihood increases and decreases, so the monotone LR doesn't hold here. However, it won't stop us using the LR test. As the LR test only uses the Sup of likelihood, which does not depend on the monotone ratio.

(ii) Here we compare the rejection region that the two tests give. UMP test won't give a rejection region that go to different direction, as it requires the power uniformly increase to max.

Suppose that X has a p.d.f. in a one-parameter exponential family:

$$f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x)$$

w.r.t. a $\sigma-$ finite measure v, where η is a strictly increasing and differentiable function of θ.

(i) For testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ there is an LR test whose rejection region is the same as that of the UMP test given in Theorem 6.2.

proof: monotone increasing of $l(\theta_0)/l(\hat{\theta})$, and when $\lambda = l(\theta_0)/l(\hat{\theta}) < c$, is equal to $\theta_0 < \hat{\theta}$

Let $\hat{\theta}$ be the MLE of θ. Note that $l(\theta)$ is increasing when $\theta < \hat{\theta}$ and decreasing when $\theta > \hat{\theta}$. Thus

$$\lambda = \begin{cases} 1 & \theta \geq \hat{\theta} \\ \frac{l(\theta)}{l(\hat{\theta})} & \theta < \hat{\theta} \end{cases}$$

Then $\lambda(X) < c$ is the same as $\hat{\theta} > \theta_0$ and $\frac{l(\theta)}{l(\hat{\theta})} < c$.

From the property of exponential families, $\hat{\theta}$ is a solution of the likelihood equation. For any $\theta_0 \in \Theta$, $log(l(\hat{\theta})) - logl(\theta_0)$ is strictly increasing in Y when $\hat{\theta} > \theta_0$, and strictly decreasing in Y when $\hat{\theta} < \theta_0$.

Hence, for any $d \in R, \hat{\theta} > \theta$ and $l(\theta_0)/l(\hat{\theta}) < c$ is equivalent to $Y > d$ for some $c \in (0, 1)$.

Here, we need to pay attention to the monotone increasing or decreasing is respective to Y. When Y increases, the LR increases or decreases.

(ii) For testing $H_0 : \theta \leq \theta_1, \theta \geq \theta_2$ versus $H_1 : \theta_1 < \theta < \theta_2$, there is an LR test whose rejection region is the same as that of the UMP test T given in Theorem 6.3.

(iii) For testing the other two-sided hypotheses, there is an LR test whose rejection region is equivalent to $Y(X) < c_1$ or $Y(X) > c_2$ for some constants c_1 and c_2.

2.3 Example

We would like to show the connection between LR test and UMP test.

(i) Consider the testing problem $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on i.i.d. $X_1, \ldots, X_n$ from the uniform distribution $U(0, \theta)$.

This problem we can do in both UMP and LR test. The lebesgue p.d.f. used in UMP test is the same as the likelihood function Let

(ii)

3 Generalized NP lemma

Let $f_1, \ldots, f_{m+1}$ be real-valued, μ- integrable functions defined on a Euclidean space X. Suppose that for given constants $c_1, \ldots, c_m$ there exists a critical function ϕ satisfying

$$\int \phi f_i d\mu = c_i, \qquad i = 1, ..m \tag{1}$$

Let C be the class of critical functions ϕ for which (**??**) holds

(i) Among all members of C there exists one that maximizes $\int \phi f_{m+1} d\mu$

Note that, $\int \phi f_{m+1} d\mu$ is the power of the test, as the critical function ϕ is to $f_1, ..f_m$.

(ii) A sufficient condition for a member ϕ_0 of C to maximize $\int \phi f_{m+1} d\mu$ (over C) is the existence of constants $k_1, \ldots, k_m$ such that

$$\phi_0(x) = \begin{cases} 1 & f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x) \\ 0 & f_{m+1}(x) < \sum_{i=1}^m k_i f_i(x) \end{cases} \tag{2}$$

Proof: it is always a good practice to transform the complicated and general case into simple case, in this problem, we will try to transform into two parameter case.

Take $\phi \in C$. Note that $\int(\phi_0 - \phi)(f_{m+1} - \sum_{i=1}^m k_i f_i)d\mu \geq 0$ since the integrand is ≥ 0. [this is the same as in the UMP proof].

$$\int(\phi_0 - \phi)f_{m+1}d\mu \geq \sum_{i=1}^m k_i \int(\phi_0 - \phi)f_i d\mu \geq 0,$$
$$\int \phi_0 f_{m+1}d\mu \geq \int \phi f_{m+1}d\mu$$

(iii) If a member of C satisfies (**??**) with $k_1, \ldots k_m \geq 0$, then it maximizes $\int \phi f_{m+1}d\mu$ among all critical functions satisfying $\int \phi f_i d\mu \leq c_i$, for $i = 1, ..m$.

(iii) The set

$$M := \Big(\int \phi f_1 d\mu, .., \int \phi f_m d\mu \Big)$$

where ϕ is a critical function. It is convex and closed. If $(c_1, ..c_m)$ is an interior point of M, then there exists constants $k_1, .., k_m$ and a test ϕ_0 satisfying (**??**) and (**??**). And a necessary condition for a member of C to maximize $\int \phi f_{m+1}d\mu$ is that (**??**) holds a.e. μ.

3.1 Example

How do we use the general NP lemma? The probability could be all different

Suppose that $X_1, \ldots X_n$ are i.i.d. from the Cauchy location family $p_\theta(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$, for $x \in R$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Can we find a test ϕ of size α such that ϕ maximizes

$$\frac{d}{d\theta}\beta_\phi(\theta_0) = \frac{d}{d\theta}E_\theta[\phi(X)]\Big|_{\theta=\theta_0}$$

For any test ϕ the power is given by

$$\beta_\phi(\theta) = E_\phi[\phi(X)] = \int \phi(X)p(x,\theta)dx$$

where $p(x, \theta)$ is the joint density of the model. So, if the interchange of differentiation and integration is justifiable, then

$$\beta_{\phi}^{'}(\theta) = \int \phi(X) \frac{\partial}{\partial \theta} p(x, \theta) dx$$

Thus, by the generalized N-P lemma, a test of the form

$$\phi_{)}(x) = \begin{cases} 1 & \frac{\partial}{\partial \theta} p(X, \theta_0) > k p(X, \theta_0) \\ 0 & \frac{\partial}{\partial \theta} p(X, \theta_0) < k p(X, \theta_0) \end{cases}$$

maximizes $\beta_{\phi}^{'}(\theta_0)$ among all ϕ with $E_{\theta_0}\phi(X) = \alpha$. This test is said to be locally most powerful of size α.

Also the likelihood function and density function are correlated.

$$\frac{\partial}{\partial \theta} p(X, \theta_0) > k p(X, \theta_0) \rightarrow \frac{\partial}{\partial \theta} log p(X, \theta_0) > k$$

$$S_n(\theta_0) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) > k'$$

The point of doing that way is to simplify the equation, so that we be able to get the statistics involving X.

4 Unbiased tests

4.1 Definition

(i) Unbiased test (Constrain by enforcing unbiasedness):

A test ϕ for $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ with level α is unbiased if its power $\beta_\phi(\theta) := E_\theta[\phi(X)]$ satisfies

$$\begin{aligned} \beta_\phi(\theta) &\leq \alpha, \qquad \theta \in \Theta_0 \\ \beta_\phi(\theta) &\geq \alpha, \qquad \theta \in \Theta_1 \end{aligned}$$

If there is a UMP test ϕ^*, then it is automatically unbiased because $\beta_{\phi^*} \geq \beta(\theta)$, for all $\theta \in \Theta_1$, where ϕ is the degenerate test, which equals α regardless of the observed data.

Unbiasedness enforces the appealing property that the probability of rejection is greater under any alternative distribution than it is under any null distribution. A uniformly most powerful test is always unbiased if it exists.

While this notion of unbiasedness differs from the definition we encountered when discussing point estimation, we can check that this is actually a special case of risk unbiasedness when the loss function L is such that $L(\theta_0, reject) = 1 - \alpha, L(\theta_1, accept) = \alpha$

(ii) UMP Unbiased test:

we will focus on finding uniformly most powerful unbiased (UMPU) tests in settings in which UMP tests do not exist. These tests often exist for testing $\theta_1 \leq \tilde{\theta}$ vs $\theta_1 > \tilde{\theta}$ in the presence of nuisance parameters $(\theta_2, \ldots \theta_n)$ and for testing $\theta = \tilde{\theta} vs \theta \neq \tilde{\theta}$.

A UMP unbiased test (UMPU) level α is a test ϕ_0 for which

$$\beta_{\phi_0}(\theta) \geq \beta_{\phi}(\theta), \qquad \theta \in \Theta_1$$

for all unbiased level α tests ϕ.

This definition says that, among all the unbiased test, the UMPU test has higher power.

Generally, if the power function $\theta \to \beta_\phi(\theta)$ is continuous in θ (as is the case for any canonical form exponential family on the natural parameter space), then ϕ unbiased and of level α implies that $\beta_\phi(\theta) = \alpha$ for all $\theta \in \omega$. We have a name for tests that match the level on the boundary.

Proof. Firstly, because ϕ_0 is UMP α-similar tests, it is at least as powerful as $\phi_\alpha(X) \equiv \alpha$, and the power of ϕ_0 on Ω_1 is therefore $\geq \alpha$. Hence, ϕ_0 is unbiased.

Secondly, an unbiased level-α test must, by definition, have expectation value $\leq \alpha$ for $\theta \in \Omega_0$ and $\geq \alpha$ for $\theta \in \Omega_1$. By continuity such a test must have expectation α on the common boundary. Therefore, the set of unbiased level-α tests is a subset of α -similar level-α tests, amongst which ϕ_0 is most powerful. Hence, ϕ_0 is also as powerful as any unbiased level-α test. ϕ_0 is UMPU.

4.2 Two sided testing without Nuisance parameters

It is based on the hypothesis that, the null hypothesis has only one θ_0, while the H_1 has two different direction of range. So based on the UMPU test definition, we look at the power function as a function of θ. And it reaches the minimum at θ_0.

Remember that UMP test is just based on the type I error rate and power, it is not important to find out what the likelihood ratio is. And the same applies to UMPU test. We only need to focus on the power function, and reaches the minimum at θ_0. Furthermore, we can find out the critical function ϕ.

Let us test $H_0 : \theta = \theta_0 vs. H_1 : \theta \neq \theta_0$, when X is distributed according some member of the one-dimensional exponential family

$$p_\theta(x) = h(x) \exp\Big(\theta T(x) - A(\theta)\Big)$$

We have seen that no UMP test exists in the normal case. Our goal here is to find a UMPU test.

Since we are working with an exponential family, the power function is continuous, and, by Lemma 1, it suffices to find a UMP level α test amongst α-similar tests. Since

$\omega = \Omega_0$, any UMP α-similar test ϕ has

$$\beta_\phi(\theta_0) = E_{\theta_0}\phi(X) = \alpha$$
$$\beta_\phi(\theta_0) \leq \beta_\phi(\theta), \qquad \theta \in R$$

since $\phi_\alpha(x) \equiv \alpha$ is also α-similar.

Since θ_0 minimizes β_ϕ, and β_ϕ is differentiable with derivative

$$\beta_\phi^{'}(\theta) = \int \phi(x)\frac{d}{d\theta}p_\theta(x)d\mu(x)$$

We have the constraint

$$0 = \beta_\phi^{'}(\theta_0) = \int \phi(x)\frac{d}{d\theta}p_{\theta_0}(x)d\mu(x)$$

5 Multiple Constraints - Method of Undetermined Multipliers (MoUM)

In this setting, $H_0 : \theta = \theta_0 vs. H_1 : \theta = \theta^{'}$. We will fix a simple alternative $\theta = \theta^{'}$ and hope that our best test has no $\theta^{'}$ dependence. We would like to maximize power subject to

$$\int \phi p_{\theta_0}(x)d\mu(x) = \alpha$$
$$\int \phi\frac{d}{d\theta}p_{\theta_0}(x)d\mu(x) = 0$$

For a 1-parameter exponential family, we have

$$p_\theta(x) = h(x)e^{\theta T(x) - A(\theta)}$$
$$\frac{d}{d\theta}p_\theta(x) = h(x)e^{\theta T(x)-A(\theta)}(T(x) - A^{'}(\theta)) = p_\theta(x)\Big[T(x) - E_\theta(T(X))\Big]$$

Applying the reasoning from the previous section, we find that a most powerful test has rejection region defined by

$$p_{\theta'}(x) > k_1 p(\theta_0)(x) + k_2\frac{d}{d\theta}p(\theta_0)(x)$$

for some values of k_1 and k_2, which is equivalent to

$$\frac{e^{(\theta^{'}-\theta_0)T(x)}}{k_1^{'} + k_2^{'}T(x)} > const$$

Now consider the set of values of T(x) satisfying this constraint. Because the constraint is that an exponential function exceeds a linear function, the set of values of T(x) satisfying this constraint is either a one-sided interval

6 Weighted Statistics

Let $X_1, ..X_n$ be i.i.d $N(0, \sigma^2)$. $w_1, ..., w_n$ is a constant vector such that $w_1, .., w_n > 0$ and $w_1 + .. + w_n = 1$. Define $\bar{X}_{nw} = \sqrt{w_1}X_1 + .. + \sqrt{w_n}X_n$. Show that $Y_n = \bar{X}_{nw}/\sigma \sim N(0, 1)$.

6.1 Question

Note that $\bar{X}_{nw}$ is a linear combination of $X_1, ..X_n$, we need to use the vector/matrix to show the distribution, while not single one variable.

If $X_i \sim N(\mu_i, \sigma_i^2)$, which we can have a MVN distribution, which each X_i has its own normal distribution. Then the transformation matrix, orthogonal matrix, etc could be applied. Here all the X_i follows the same distribution, and we also can use the similar concept by applying orthogonal matrix.

Also we have the slutsky's theorem, delta method for the asymptotic distribution, however that is under the $n \to \infty$. In this problem, we can't use that.

So this is the exact distribution using the transformation (just the transform is by orthogonal matrix). The MGF or characteristic distribution is always the method when doing transformation.

6.2 MGF

$$\begin{aligned} M(t) &= exp(\mu t + \sigma^2 t^2/2), \qquad \text{MGF for } N(\mu, \sigma^2) \\ M_{\sqrt{w_i}X_i}(t) &= E[exp(\sqrt{w_i}tX_i)] = exp(\mu\sqrt{w_i}t + \sigma^2[\sqrt{w_i}t]^2/2), \qquad \mu = 0 \\ &= exp(\sigma^2 w_i t^2/2) \end{aligned}$$

Then the linear combination y_n

$$\begin{aligned} M_{Y_n}(t) &= E[exp\left((\sqrt{w_1}X_1 + \sqrt{w_2}X_2 + .. + \sqrt{w_n}X_n)t\right)] \\ &= E[exp(\sqrt{w_1}X_1 t)]E[exp(\sqrt{w_2}X_2 t)]E[exp(\sqrt{w_3}X_3 t)]..E[exp(\sqrt{w_n}X_n t)] \\ &= exp(\sigma^2 w_1 t^2/2)exp(\sigma^2 w_2 t^2/2)exp(\sigma^2 w_3 t^2/2)..exp(\sigma^2 w_n t^2/2) \\ &= exp(\sigma^2[w_1 + w_2 + ..w_n]t^2/2) = exp(\sigma^2 t^2/2) \end{aligned}$$

So $Y_n \sim N(0, \sigma^2)$.

6.3 Orthogonal Matrix

Consider an orthogonal matrix Σ such that the first row is $(\sqrt{w_1}, \sqrt{w_2}, ..\sqrt{w_n})$. Let

$$(Z_1, Z_2, ..Z_n)^T = \Sigma(X_1, X_2, ..X_n)^T$$

We have $Z^T Z = (\Sigma X)^T(\Sigma X) = X^T \Sigma^T \Sigma X = X^T X$. The characteristic function of Z is

$$\phi_Z(t) = E[exp(it'Z)] = E[exp(i(\Sigma' t)'X)] = exp(-\sigma^2 t't/2)$$

Need to get familiar with the vector form in MGF/characteristic function
Therefore, we have $Z_1, ..Z_n \sim N(0, \sigma^2)$

$$\begin{aligned} Y_n &= \bar{X}_{nw}/\sigma = (\sqrt{w_1}X_1 + \sqrt{w_2}X_2 + .. + \sqrt{w_n}X_n)/\sigma \\ &= Z_1/\sigma \sim N(0,1) \end{aligned}$$

Also,

$$\begin{aligned} (n-1)S_n^2/\sigma^2 &= \sum_{i=1}^{n}(X_i^2 - \bar{X}_{nw}^2)/\sigma^2 \\ &= (X^T X - Z_1^2)/\sigma^2 = (Z_2^2 + ... + Z_n^2)/\sigma^2 \sim \chi_{n-1}^2 \end{aligned}$$

Since Y_n and S_n^2 are functions of Z_1 and $(Z_2, ..Z_n)$ respectively, and from the independence of $Z_i, (i = 1, ..n)$, we have Y_n and S_n^2 are independent. It follows that, by the definition of t-distribution, $T_n \sim t_{n-1}/\sigma$. When $w_1 = w_2 = .. = w_n = 1/n, Y_n = \sum_{i=1}^{n} X_i/(\sigma\sqrt{n})$, which is the standardized sample mean. Also,

$$\begin{aligned} S_n^2/\sigma^2 &= \frac{\sum_{i=1}^{n} X_i^2 - \bar{X}_{nw}^2}{n-1} \\ &= \frac{\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} {X_i}^2/n}{n-1} \\ &= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}_i^2}{n-1} \end{aligned}$$

which is the sample variance.

If there are quadratic forms, we can consider the orthogonal matrix that transform to standard normal distribution.

7 Sufficient and Complete Statistics

7.1 Minimum Sufficient Statistics

7.2 Complete Statistics

7.3 Ancillary Statistics

8 Bivariate Normal Distribution / Partition Matrix

The Bivariate Normal Distribution is always connected with partitioned covariance matrix. Assume vector (X, Y) is Gaussian.

Definition: Two random variables X and Y are said to be bivariate normal, or jointly normal, if $aX + bY$ has a normal distribution for all $a, b \in R$.

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are jointly normal, then $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho(X,Y)\sigma_X\sigma_Y)$.

We consider $X + Y$ is a also normal distribution, then the covariance

$$Cov(X+Y) = Cov(X) + Cov(Y) + 2Cov(XY) = \sigma_X^2 + \sigma_Y^2 + 2\rho(X,Y)\sigma_X\sigma_Y$$

How to provide a simple way to generate jointly normal random variables? The basic idea is that we can start from several independent random variables and by considering their linear combinations, we can obtain bivariate normal random variables.

Let Z_1 and Z_2 be two independent N(0,1) random variables. Define

$$X = Z_1, \qquad Y = \rho Z_1 + \sqrt{1-\rho^2} Z_2$$

where ρ is a real number in (-1, 1). Show that X and Y are bivariate normal.

First, note that since Z_1 and Z_2 are normal and independent, they are jointly normal, with the joint PDF

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &= f_{Z_1}(z_1) f_{Z_2}(z_2) \\ &= \frac{1}{2\pi} exp\left(-\frac{1}{2}[z_1^2 + z_2^2]\right) \end{aligned}$$

We need to show $aX + bY$ is normal for all $a, b \in R$. We have

$$\begin{aligned} aX + bY &= aZ_1 + b(\rho Z_1 + \sqrt{1-\rho^2} Z_2) \\ &= (a + b\rho) Z_1 + b\sqrt{1-\rho^2} Z_2 \end{aligned}$$

which is a linear combination of Z_1 and Z_2, and thus it is normal.

We can use the method of transformation to find the joint PDF of X and Y. The inverse transformation is given by

$$\begin{aligned} Z_1 &= X = h_1(X,Y) \\ Z_2 &= -\frac{\rho}{\sqrt{1-\rho^2}} X + \frac{1}{\sqrt{1-\rho^2}} Y = h_2(X,Y) \end{aligned}$$

We have

$$\begin{aligned} f_{XY}(z_1, z_2) &= f_{Z_1,Z_2}(h_1(X,Y), h_2(X,Y))|J| \\ &= f_{Z_1,Z_2}(x, -\frac{\rho}{\sqrt{1-\rho^2}} x + \frac{1}{\sqrt{1-\rho^2}} y)|J| \end{aligned}$$

where

$$J = det \begin{bmatrix} \frac{\partial h_1}{\partial x} & \frac{\partial h_1}{\partial y} \\ \frac{\partial h_2}{\partial x} & \frac{\partial h_2}{\partial y} \end{bmatrix} = det \begin{bmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{bmatrix} = \frac{1}{\sqrt{1-\rho^2}}$$

Thus, we conclude that,

$$f_{XY}(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(-\frac{1}{2(1-\rho^2)}[x^2 - 2\rho xy + y^2]\right)$$

To find the ρ

$$\begin{aligned} Var(X) &= Var(Z_1) = 1 \\ Var(Y) &= \rho^2 Var(Z_1) + (1-\rho^2)Var(Z_2) = 1 \\ \rho(X,Y) &= Cov(X,Y) = Cov(Z_1, \rho Z_1 + \sqrt{1-\rho^2}Z_2) \\ &= \rho Cov(Z_1, Z_2) + \sqrt{1-\rho^2}Cov(Z_1, Z_2) \\ &= \rho \end{aligned}$$

Now, if you want two jointly normal random variables X and Y such that $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$, and $\rho(X,Y) = \rho$, you can start with two independent N(0,1) random variables, Z_1 and Z_2, and define

$$\begin{aligned} X &= \sigma_X Z_1 + \mu_X \\ Y &= \sigma_Y \left(\rho Z_1 + \sqrt{1-\rho^2}Z_2\right) + \mu_Y \end{aligned}$$

construction using Z1 and Z2 can be used to solve problems regarding bivariate normal distributions. Third, this method gives us a way to generate samples from the bivariate normal distribution using a computer program.

8.1 Conditional Distribution

Suppose X and Y are jointly normal random variables with parameters $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$, and ρ. Then, given $X = x, Y$ is normally distributed with

$$\begin{aligned} E[Y|X=x] &= \mu_Y + \rho\sigma_Y \frac{x - \mu_X}{\sigma_X} \\ Var(Y|X=x) &= (1-\rho^2)\sigma_Y^2 \end{aligned}$$

One way to solve this problem is by using the joint PDF formula. In particular, since $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$, we can use

$$f_{Y|X=x}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

or we use

$$\begin{aligned} X &= \sigma_X Z_1 + \mu_X \\ Y &= \sigma_Y \left(\rho Z_1 + \sqrt{1-\rho^2} Z_2\right) + \mu_Y \end{aligned}$$

Thus, given $X = x$,

$$\begin{aligned} Z_1 &= \frac{x - \mu_X}{\sigma_X} \\ Y &= \sigma_Y \rho \frac{x - \mu_X}{\sigma_X} + \sigma_Y \sqrt{1-\rho^2} Z_2 + \mu_Y \end{aligned}$$

Since Z_1 and Z_2 are independent, knowing Z_1 does not provide any information on Z_2. We have shown that given $X = x, Y$ is a linear function of Z_2, thus it is normal. In particular

$$\begin{aligned} E[Y|X=x] &= \sigma_Y \rho \frac{x - \mu_X}{\sigma_X} + \sigma_Y \sqrt{1-\rho^2} E[Z_2] + \mu_Y \\ &= \mu_Y + \rho \sigma_Y \frac{x - \mu_X}{\sigma_X} \\ Var[Y|X=x] &= \sigma_Y^2 (1-\rho^2) Var(Z_2) = (1-\rho^2)\sigma_Y^2 \end{aligned}$$

8.1.1 Marginal and conditional distributions of a multivariate normal vector

A $K \times 1$ random vector X is multivariate normal if its joint probability density function is

$$f_X(x) = (2\pi)^{-K/2} |det(V)|^{-1/2} exp(-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu))$$

where μ is a $K \times 1$ mean vector, V is a $K \times K$ covariance matrix.
Partition of the vector:
We partition X into two sub-vectors X_a and X_b such that

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}$$

The sub-vectors X_a and X_b have dimensions $K_a \times 1$ and $K_b \times 1$ respectively. Moreover, $K_a + K_b = K$.

Partition of the parameters

We partition the mean vector and covariance matrix as follows:

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

and

$$V = \begin{pmatrix} V_a & V_{ab}^T \\ V_{ab} & V_b \end{pmatrix}$$

Normality of the sub-vectors

The marginal distributions of the two sub-vectors are also multivariate normal.

8.1.2 Proof

The random vector X_a can be written as a linear transformation of X:

$$X_a = AX$$

Where A is a $K_a \times K$ matrix whose entries are either zero or one. Thus, X_a has a multivariate normal distribution because it is a linear transformation of the multivariate normal random vector X and multivariate normality is preserved by linear transformations. Same as $X_b = BX$ where B is a $K_b \times K$ matrix whose entries are either zero or one.

Independence of the sub-vectors

X_a and X_b are independent if and only if $V_{ab} = 0$.

X_a and X_b are independent if and only if their joint moment generating function is equal to the product of their individual moment generating functions. Since X_a is multivariate normal, its joint moment generating function is

$$M_{X_a}(t_a) = exp(t_a^T \mu_a + \frac{1}{2} t_a^T V_a t_a)$$
$$M_{X_b}(t_b) = exp(t_b^T \mu_b + \frac{1}{2} t_b^T V_b t_b)$$

The joint moment generating function of X_a and X_b, which is just the joint moment generating function of X, is

$$
\begin{aligned}
M_{X_a,X_b}(t_a,t_b) &= M_X(t) \\
&= exp(t^T\mu + \frac{1}{2}t^TVt) \\
&= exp\left([t_a^T t_b^T]\begin{bmatrix}\mu_a \\ \mu_b\end{bmatrix} + [t_a^T t_b^T]\begin{bmatrix}V_a & V_{ab}^T \\ V_{ab} & V_b\end{bmatrix}[t_a t_b]\right) \\
&= exp\left(t_a^T\mu_a + t_b^T\mu_b + \frac{1}{2}t_a^TV_at_a + \frac{1}{2}t_b^TV_bt_b + \frac{1}{2}t_b^TV_{ab}t_a + \frac{1}{2}t_a^TV_{ab}^Tt_b\right) \\
&= exp\left(t_a^T\mu_a + t_b^T\mu_b + \frac{1}{2}t_a^TV_at_a + \frac{1}{2}t_b^TV_bt_b + t_b^TV_{ab}t_a\right) \\
&= exp\left(t_a^T\mu_a + \frac{1}{2}t_a^TV_at_a\right)exp\left(t_b^T\mu_b + \frac{1}{2}t_b^TV_bt_b\right)exp(t_b^TV_{ab}t_a)
\end{aligned}
$$

from which it is obvious that $M_{X_a,X_b}(t_a,t_b) = M_{X_a}(t_a)M_{X_b}(t_b)$ if and only if $V_{ab} = 0$.

8.2 Schur Complement

In order to derive the conditional distributions, we are going to reply on Schur complements.

In linear algebra and the theory of matrices, the Schur complement of a block matrix is defined as follows.

Suppose p, q are nonnegative integers, and suppose A, B, C, D are respectively $p \times p, p \times q, q \times p$, and $q \times q$ matrices of complex numbers. Let

$$
M = \begin{bmatrix}A & B \\ C & D\end{bmatrix}
$$

So that M is a $(p+q) \times (q+p)$ matrix. If D is invertible, then the **Schuler complement** of the block D of the matrix M is the $p \times p$ matrix defined by

$$
M/D := A - BD^{-1}C
$$

If A is invertible, the Schur complement of the block A of the matrix M is the q × q matrix defined by

$$
M/A := D - CA^{-1}B
$$

In the case that A or D is singular, substituting a generalized inverse for the inverses on M/A and M/D yields the generalized Schur complement.

8.2.1 Background

The Schur complement arises when performing a block Gaussian elimination on the matrix M. In order to eliminate the elements below the block diagonal, one multiplies the matrix M by a block lower triangular matrix on the right as follows:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \rightarrow \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & B \\ 0 & D \end{bmatrix}$$

where Ip denotes a $p \times p$ identity matrix. As a result, the Schur complement $M/D = A - BD^{-1}C$ appears in the upper-left $p \times p$ block.

Continuing the elimination process beyond this point (i.e., performing a block Gauss–Jordan elimination),

$$\begin{bmatrix} A - BD^{-1}C & B \\ 0 & D \end{bmatrix} \rightarrow \begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & B \\ 0 & D \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix}$$

leads to an LDU decomposition of M, which reads

$$M = \begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix}$$

Thus, the inverse of M may be expressed involving D^{-1} and the inverse of Schur's complement, assuming it exists, as

$$M^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \left(\begin{bmatrix} I_p & BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I_p & 0 \\ D^{-1}C & I_q \end{bmatrix} \right)^{-1}$$

Here I need to separate the inverse of 2×2 matrix and this partitioned matrix.

$$\begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} I_p & BD^{-1} \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix}$$

$$\begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix}^{-1} = \begin{bmatrix} I_p & BD^{-1} \\ 0 & I_q \end{bmatrix}$$

$$\begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix}^{-1} = \begin{bmatrix} I_p & 0 \\ D^{-1}C & I_q \end{bmatrix}$$

So, we have

$$M^{-1} = \begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix} \begin{bmatrix} [A - BD^{-1}C]^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix}$$

If p and q are both 1 (i.e., A, B, C and D are all scalars), we get the familiar formula for the inverse of a 2-by-2 matrix:

$$M^{-1} = \frac{1}{AD - BC} \begin{bmatrix} D & -B \\ -C & A \end{bmatrix}$$

8.2.2 Applications to probability theory and statistics

Suppose the random column vectors X, Y live in R_n and R_m respectively, and the vector (X, Y) in R_{n+m} has a multivariate normal distribution whose covariance is the symmetric positive-definite matrix

$$\Sigma = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

where $\mathbf{R}^{n \times n}$ is the covariance matrix of X, $C \in \mathbf{R}^{m \times m}$ is the covariance matrix of Y and $B \in \mathbf{R}^{n \times m}$ is the covariance matrix between X and Y.

Then the conditional covariance of X given Y is the Schur complement of C in $\Sigma\Sigma$

$$Cov(X|Y) = A - BC^{-1}B^T$$
$$E(X|Y) = E(X) + BC^{-1}(Y - E(Y))$$

Let V_a be invertible. Let V/V_a be the Schur complement of V_a in V, defined as

$$V/V_a = V_b - V_{ab}V_a^{-1}V_{ab}^T$$

If V/V_a is invertible, then V is invertible

8.3 b

Consider the following

(a) For an arbitrary model, consider the conditional score statistic

$$U_\psi(\xi) = \frac{\partial l_c(\xi, \psi_0)}{\partial \psi}|_{\psi_0 = \psi}$$

Show that the conditional score statistic for any model can be written as

$$U_\psi(\xi) = \partial_\psi logp(Y|\xi) - E[\partial_\psi logp(Y|\xi)|s_\lambda(\psi_0)]|_{\psi_0 = \psi}$$

The conditional score statistic is the derivative of the conditional distribution

$$U_\psi(\xi) = \frac{\partial l_c(\xi, \psi_0)}{\partial \psi}|_{\psi_0 = \psi}$$
$$p(\mathbf{Y}|\xi) = p(\mathbf{Y}|s_\lambda(\psi_0), \xi)p(s_\lambda(\psi_0)|\xi), \qquad p(\mathbf{Y}|s_\lambda(\psi_0), \xi) = \frac{p(\mathbf{Y}|\xi)}{p(s_\lambda(\psi_0)|\xi)}$$
$$l_c(\xi, \psi_0) = logp(\mathbf{Y}|s_\lambda(\psi_0), \xi) = logp(\mathbf{Y}|\xi) - logp(s_\lambda(\psi_0)|\xi)$$

Then we need to prove

$$U_\psi(\xi) = \frac{\partial l_c(\xi, \psi_0)}{\partial \psi}|_{\psi_0=\psi} = \partial_\psi logp(\mathbf{Y}|\xi) - \partial_\psi logp(s_\lambda(\psi_0)|\xi)$$
$$\partial_\psi logp(s_\lambda(\psi_0)|\xi) = E[\partial_\psi logp(Y|\xi)|s_\lambda(\psi_0)]|_{\psi_0=\psi}$$

We can write

$$logp(\mathbf{Y}|\xi) = logp(\mathbf{Y}|s_\lambda(\psi_0), \xi) + logp(s_\lambda(\psi_0)|\xi)$$
$$E\left(\partial_\psi[logp(\mathbf{Y}|\xi)|s_\lambda]\right) = E\left(\partial_\psi[logp(\mathbf{Y}|s_\lambda(\psi_0), \xi)|s_\lambda]\right) + E\left(\partial_\psi[logp(s_\lambda(\psi_0), \xi)|s_\lambda]\right)$$

in which, the integral and expectation can switch, then we have

$$E\left(\partial_\psi[logp(\mathbf{Y}|s_\lambda(\psi_0), \xi)|s_\lambda]\right) = \partial_\psi E\left([logp(\mathbf{Y}|s_\lambda(\psi_0), \xi)|s_\lambda]\right) = \partial_\psi E\left([logp(\mathbf{Y}|\xi)]\right) = 0$$

So,

$$E\left(\partial_\psi[logp(\mathbf{Y}|\xi)|s_\lambda]\right) = \partial_\psi logp(s_\lambda(\psi_0), \xi)$$

Then we show

$$U_\psi(\xi) = \partial_\psi logp(Y|\xi) - E[\partial_\psi logp(Y|\xi)|s_\lambda(\psi_0)]|_{\psi_0=\psi}$$

(b) Suppose that $y_1; ..y_n$ are independent and y_i follows a Poisson distribution with mean $exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2})$, where $(x_{i1}; x_{i2})$ are covariates, $\lambda = (\lambda_0; \lambda_1)$ is the nuisance parameter vector and ψ is the parameter of interest. Derive the conditional likelihood of ψ and show that this conditional likelihood is free of λ.
The joint distribution of $(y_1, , y_n)$ is given by

$$P(Y|\lambda, \psi) = exp\left(\sum_{i=1}^{n} y_i(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^{n} exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i!\right)$$

Thus, $S_0 = \sum_{i=1}^n y_i$ is the sufficient and complete statistics for λ_0, and $S_1 = \sum_{i=1}^n y_i x_{i1}$ is the sufficient and complete statistics for λ_1.
The conditional distribution of ψ given S_0, S_1 is given by

$$\begin{aligned} p(\mathbf{Y}, \psi|S = (S_0, S_1)) &= \frac{exp\left(\sum_{i=1}^n y_i(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^n exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i!\right)}{\sum_{y' \in S} exp\left(\sum_{i=1}^n y_i'(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^n exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy\right.} \\ &= \frac{exp\left(S_1\lambda_0 + S_2\lambda_1 + S_3\psi) - \sum_{i=1}^n exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i!\right)}{\sum_{y' \in S} exp\left(S_1'\lambda_0 + S_2'\lambda_1 + S_3'\psi) - \sum_{i=1}^n exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - logy_i'!\right)} \\ &= \frac{exp\left(S_3\psi - logy_i!\right)}{\sum_{y' \in S} exp\left(S_3'\psi - logy_i'!\right)}, \qquad S_3 = \sum_{i=1}^n y_i x_{i2}, S_3' = \sum_{i=1}^n y_i' x_{i2} \end{aligned}$$

which is independent of λ.

(c) Derive the conditional score statistic for part (b) and write out a Newton-Raphson algorithm for obtaining the conditional maximum likelihood estimate of ψ based on $U_\psi(\xi)$.
The log likelihood of the conditional distribution is

$$l_c(\psi) = S_3\psi - logy_i! - log\left[\sum_{y'\in S} exp\left(S_3'\psi - logy_i'!\right)\right], \qquad S_3 = \sum_{i=1}^{n} y_i x_{i2}, S_3' = \sum_{i=1}^{n} y_i' x_{i2}$$

The score function and observed fisher information is

$$\begin{aligned} U_\psi(\xi) &= \frac{\partial l_c(\xi,\psi_0)}{\partial\psi}|_{\psi_0=\psi} \\ &= \psi - \frac{\sum_{y'\in S} S_3' exp\left(S_3'\psi - logy_i'!\right)}{\sum_{y'\in S} exp\left(S_3'\psi - logy_i'!\right)} \\ \frac{\partial^2 l_c(\xi,\psi_0)}{\partial\psi^2} &= \left[\frac{\sum_{y'\in S} S_3' exp\left(S_3'\psi - logy_i'!\right)}{\sum_{y'\in S} exp\left(S_3'\psi - logy_i'!\right)}\right]^2 - \frac{\sum_{y'\in S} S_3'^2 exp\left(S_3'\psi - logy_i'!\right)}{\sum_{y'\in S} exp\left(S_3'\psi - logy_i'!\right)} \end{aligned}$$

The newton-Raphson algorithm

$$\psi^{k+1} = \psi^k - \left[\frac{\partial^2 l_c(\psi^k)}{\partial\psi^2}\right]^{-1} U_\psi(\psi^k)$$

where $\frac{\partial^2 l_c(\psi^k)}{\partial\psi^2}, U_\psi(\psi^k)$ are from above equations.

(d) Now suppose that we only have two random variables $y_1 \sim Poisson(\mu_1)$ and $y_2 \sim Poisson(\mu_2)$, where y_1 and y_2 are independent. We are interested in making inferences on the ratio $\psi = \mu_1/\mu_2$. Let $\xi = (\psi, \lambda)$, where λ represents the nuisance parameter.

(i) Show that the log-likelihood function of ξ can be written as

$$l(\xi) = (y_1 + y_2)\lambda + y_1 log(\psi) - exp(\lambda)(1+\psi)$$

where λ is a function of μ_2. Explicitly state what λ is.
Write the joint distribution of y_1, y_2

$$\begin{aligned} P(y_1,y_2) &= \frac{\mu_1^{y_1} e^{-\mu_1}}{y_1!}\frac{\mu_2^{y_2} e^{-\mu_2}}{y_2!} \\ logP(y_1,y_2) &= y_1 log\mu_1 - \mu_1 + y_2 \log\mu_2 - \mu_2 - logy_1! - logy_2! \\ &= y_1 log\frac{\mu_1}{\mu_2} + y_1 log\mu_2 + y_2 log\mu_2 - \mu_1 - \mu_2 - logy_1! - logy_2! \\ &= y_1 log\frac{\mu_1}{\mu_2} + (y_1+y_2)log\mu_2 - \mu_2(\mu_1/\mu_2 + 1) - logy_1! - logy_2! \end{aligned}$$

where

$$\psi = log\frac{\mu_1}{\mu_2}$$
$$\lambda = log\mu_2$$

(ii) Derive the conditional likelihood of ψ and write out a Newton-Raphson algorithm for obtaining the conditional maximum likelihood estimate of ψ .
From part (a), we see $y_1 + y_2$ is the sufficient statistics for λ, while $y_1 + y_2 \sim Poission(\mu_1 + \mu_2)$ then we have conditional distribution of ψ condition on $S = y_1 + y_2$.

$$\begin{aligned} Y(\psi|S = y_1 + y_2, \lambda) &= \frac{exp\left[y_1\psi + (y_1 + y_2)\lambda - exp(\lambda)(\psi + 1) - logy_1! - logy_2!\right]}{exp\left[(y_1 + y_2)log(\mu_1 + \mu_2) - (\mu_1 + \mu_2) - log(y_1 + y_2)!\right]} \\ &= \frac{exp\left[y_1\psi + S\lambda - exp(\lambda)(\psi + 1) - logy_1! - logy_2!\right]}{exp\left[S(\lambda + log(\psi + 1)) - exp(\lambda)(\psi + 1) - logS!\right]} \\ &= \frac{exp\left[y_1\psi - logy_1! - logy_2!\right]}{exp\left[(y_1 + S - y_1)log(\psi + 1)) - logS!\right]} \\ &= \binom{S}{y_1}\left(\frac{\psi}{1+\psi}\right)^{y_1}\left(\frac{1}{1+\psi}\right)^{S-y_1} \end{aligned}$$

The conditional distribution is a binomial, $B(S, \psi/(1 + \psi))$.
The score function and observed fisher information

$$logY(\psi|S, \lambda) = y_1 log\psi - Slog(1 + \psi) + log\binom{S}{y_1}$$
$$\partial_\psi logY(\psi|S, \lambda) = \frac{y_1}{\psi} - \frac{S}{1+\psi} = 0, \qquad \hat{\psi} = y_1/(S - y_1)$$
$$\partial^2_\psi logY(\psi|S, \lambda) = -\frac{y_1}{\psi^2} + \frac{S}{(1+\psi)^2}$$

The $CMLE = \hat{\psi} = y_1/(S - y_1)$. And the newton-Raphson equation

$$\begin{aligned} \psi^{k+1} &= \psi^k - \left[\frac{\partial^2 l_c(\psi^k)}{\partial\psi^2}\right]^{-1} U_\psi(\psi^k) \\ &= \psi^k - \left[-\frac{y_1}{\psi^2} + \frac{S}{(1+\psi)^2}\right]^{-1}\left[\frac{y_1}{\psi} - \frac{S}{1+\psi}\right]|_{\psi=\psi^k} \\ &= \psi^k + \frac{y_1/\psi^k - S/(1+\psi^k)}{y_1/{\psi^k}^2 - S/(1+\psi^k)^2} \end{aligned}$$