

# Convergence Theorem

Mingwei Fei

January 22, 2023

## 1 Measurement Theorem and Integral

## 2 Monotone Convergence

## 3 Convergence Mode

It is very important to understand the definition and the notation for each definition.

### (i) Converge almost everywhere

A sequence  $X_n$  converges almost everywhere (a.e) to  $X$ , denoted  $X_n \xrightarrow{a.e.} X$ , if  $X_n(w) \rightarrow X(w)$  for all  $w \in \Omega - N$  where  $\mu(N) = 0$ . If  $\mu$  is a probability, we write a.e. as a.s. (almost surely).

$$\lim_{n \rightarrow \infty} X_n = X$$
$$P(\sup_{m \geq n} |X_m - X| > \epsilon) \rightarrow 0$$

Remarks: Pay attention to the notation, it says that among all the observations that after  $X_n$ , the biggest difference is less than a certain value. When the  $\sup_{m \geq n}$  come up, it has listed almost all the observations, which is the same as almost sure.

### (ii) Converges in probability A sequence $X_n$ converges in measure to a measurable function $X$ , denoted $X_n \xrightarrow{\mu} X$ , if $\mu(|X_n - X| \geq \epsilon) \rightarrow 0$ for all $\epsilon > 0$ . If $\mu$ is a probability measure, we say $X_n$ converges in probability to $X$ .

$$\lim_{n \rightarrow \infty} P(\|X_n - X\| > \epsilon) = 0$$

### (iii) Converges in $L_r$ -distance (rth moment)

Notation:  $c = (c_1, \dots, c_k) \in R^k$ ,  $\|c\|_r = \left( \sum_{j=1}^k |c_j|^r \right)^{1/r}$ ,  $r > 0$ . If  $r \geq 1$ , then  $\|c\|_r$  is the  $L_r$ -distance between 0 and  $c$ . When  $r = 2$ ,  $\|c\| = \|c\|_2 = \sqrt{c^t c}$ .

$$X_n \xrightarrow{L_r} X$$

$$\lim_{n \rightarrow \infty} E \|X_n - X\|_r^r = 0$$

- (iv) Converges in distribution Let  $F, F_n, n = 1, 2, \dots$ , be c.d.f.'s on  $R^k$  and  $P, P_n, n = 1, 2, \dots$  be their corresponding probability measures. We say that  $\{F_n\}$  converges to  $F$  weakly and write  $F_n \xrightarrow{w} F$  iff, for each continuity point  $x$  of  $F$ ,

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

We say that  $\{X_n\}$  converges to  $X$  in distribution and write  $X_n \xrightarrow{d} X$  iff  $F_{X_n} \xrightarrow{w} F_X$ .

Note: converges in distribution is the cumulative distribution is the same.

- (v) Remarks

$\xrightarrow{a.s.}, \xrightarrow{p}, \xrightarrow{L_r}$ : measures how close is between  $X_n$  and  $X$  as  $n \rightarrow \infty$ .

$F_{X_n} \xrightarrow{w} F_X$ :  $F_{X_n}$  is close to  $F_X$ . but  $X_n$  and  $X$  may not be close, they may be on different spaces.

Example: Let  $\theta_n = 1 + n^{-1}$  and  $X_n$  be an random variable having the exponential distribution  $E(0, \theta_n), n = 1, 2, \dots$ . Let  $X$  be a random variable having the exponential distribution  $E(0, 1)$ .

For any  $x > 0$ , as  $n \rightarrow \infty$ ,

$$F_{X_n}(x) = 1 - e^{-x/\theta_n} \rightarrow 1 - e^{-x} = F_X(x)$$

Since  $F_{X_n}(x) = 0 = F_X(x)$  for  $x \leq 0$ , we have shown that  $X_n \xrightarrow{d} X$ .

How about  $X_n \xrightarrow{p} X$ ?

We will need the distribution of  $X_n - X$  as we need to get the probability  $P(|X_n - X| > \epsilon)$ .

The distribution has two cases depends on whether  $X_n$  and  $X$  are independent or not.

- (i) Suppose that  $X_n$  and  $X$  are not independent, and  $X_n \equiv \theta_n X$  (then  $X_n$  has the given c.d.f.).

$X_n - X = (\theta_n - 1)X = n^{-1}X$ , which has the c.d.f.  $(1 - e^{-nx})I_{[0, \infty)}(x)$ .

Then  $X_n \xrightarrow{p} X$  because, for any  $\epsilon > 0$ ,

$$P(|X_n - X| \geq \epsilon) = e^{-n\epsilon} \rightarrow 0$$

Also,  $X_n \xrightarrow{L_p} X$  for any  $p > 0$ , because

$$E(|X_n - X|^p) = n^{-p} E X^p \rightarrow 0$$

- (ii) Suppose that  $X_n$  and  $X$  are independent random variables. Since p.d.f.'s for  $X_n$  and  $-X$  are  $\theta_n^{-1} e^{-x/\theta_n} I_{(0,\infty)}(x)$  and  $e^x I_{(-\infty,0)}(x)$ , respectively, we have let  $y = X_n - X, x = X_n$ , then  $-X = y - X_n < 0$ . In the below range,  $y \in (-\infty, x)$

$$P(|X_n - X| \leq \epsilon) = \int_{-\epsilon}^{\epsilon} \int_0^{\infty} \theta_n^{-1} e^{-x/\theta_n} e^{y-x} I_{(0,\infty)}(x) I_{(-\infty,x)}(y) dx dy$$

which converges to (by the dominated convergence theorem)

$$\begin{aligned} \int_{-\epsilon}^{\epsilon} \int_0^{\infty} e^{-x} e^{y-x} I_{(0,\infty)}(x) I_{(-\infty,-x)}(y) dx dy &= 1 - e^{-\epsilon} \\ &= \int_0^{\epsilon} e^{-2x} \int_{-\epsilon}^x e^y dy dx \\ &= \int_0^{\epsilon} e^{-x} dx \\ &= 1 - e^{-\epsilon} \end{aligned}$$

Thus,  $P(|X_n - X| \leq \epsilon) \rightarrow e^{-\epsilon} > 0$  for any  $\epsilon > 0$  and, therefore,  $X_n \xrightarrow{p} X$  does not hold.

### 3.1 Relationship between coverage modes

- (i) If  $X_n \xrightarrow{a.s.} X$ , then  $X_n \xrightarrow{p} X$ .

Proof:

$$P(|X_i - X| > \epsilon) \leq P(\sup_{m \geq n} |X_m - X| > \epsilon) \rightarrow 0$$

- (ii) If  $X_n \xrightarrow{L_r} X$  for an  $r > 0$ , then  $X_n \xrightarrow{p} X$ . Consider the definition of moment convergence and probability convergence, the link that connect Expectation and Probability with inequality is Markov Inequality.

For any positive and increasing function  $g(\cdot)$  and random variable  $Y$ ,

$$P(|Y| > \epsilon) \leq E\left[\frac{g(|Y|)}{g(\epsilon)}\right]$$

In particular, we choose  $Y = |X_n - X|$  and  $g(y) = |y|^r$ . It gives that

$$P(|X_n - X| > \epsilon) \leq E\left[\frac{|X_n - X|^r}{\epsilon^r}\right] \rightarrow 0$$

(iii) If  $X_n \xrightarrow{p} X$ , then  $X_n \xrightarrow{d} X$ .

Prove: need to use the definition of convergence in probability, and construct the cumulative probability  $F_X(x)$ .

The purpose is to induce  $F_{X_n}(x)$ , so that we can compare  $F_{X_n}(x)$  and  $F(x)$ . So the  $F(x)$  will be rewritten as  $F_{X_n}(x)$  and a probability involves  $X_n - X$ .

Assume  $k = 1$ , let  $x$  be a continuity point of  $F_X$  and  $\epsilon > 0$  be given. Then

$$\begin{aligned} F_X(x - \epsilon) &= P(X \leq x - \epsilon, X_n \leq x) + P(X \leq x - \epsilon, X_n > x) \\ &\leq P(X_n \leq x) + P(X \leq x - \epsilon, X_n > x), \quad P(X_n \leq x) > P(X \leq x - \epsilon, X_n \leq x) \\ &\leq F_{X_n}(x) + P(|X_n - X| > \epsilon), \quad X_n - X > x - (x - \epsilon) = \epsilon \end{aligned}$$

Letting  $n \rightarrow \infty$ , we obtain that

$$F_X(x - \epsilon) \leq \liminf_n F_{X_n}(x)$$

Switching  $X_n$  and  $X$  in the previous argument,

$$\begin{aligned} F_X(x + \epsilon) &= P(X \leq x + \epsilon, X_n \leq x) + P(X \leq x + \epsilon, X_n > x) \\ &\geq P(X_n \leq x) + P(X \leq x + \epsilon, X_n > x) \\ &\geq F_{X_n}(x) + P(|X_n - X| > \epsilon) \end{aligned}$$

Letting  $n \rightarrow \infty$ , we obtain that

$$\begin{aligned} F_X(x - \epsilon) &\leq \liminf_n F_{X_n}(x) \\ F_X(x + \epsilon) &\geq \limsup_n F_{X_n}(x) \end{aligned}$$

Since  $\epsilon$  is arbitrary and  $F_X$  is continuous at  $x$ ,

$$F_X(x) = \lim_{n \rightarrow \infty} F_{X_n}(x)$$

- (iv) Skorohod's theorem: a conditional converse of (i)-(iii). If  $X_n \xrightarrow{d} X$ , then there are random vectors  $Y_n, Y_n \xrightarrow{a.s.} Y$ .
- (v) If, for every  $\epsilon > 0$ ,  $\sum_{n=1}^{\infty} P(\|X_n - X\| \geq \epsilon) < \infty$ , then  $X_n \xrightarrow{a.s.} X$ .
- (vi) If  $X_n \xrightarrow{p} X$ , then there is a subsequence  $\{X_{n_j}, j = 1, 2, \dots\}$  such that  $X_{n_j} \xrightarrow{a.s.} X$  as  $j \rightarrow \infty$ .

We need to show that such a sequence exists, and prove by the almost surely definition. Such a sequence generally use the  $2^{-k}$ . Because  $2^{-k}$  is almost surely convergence, so any sequence that is smaller than this sequence, will definitely be almost surely convergence as well.

For any  $\epsilon > 0$ ,  $P(|X_n - X| > \epsilon) \rightarrow 0$ , we choose  $\epsilon = 2^{-m}$  then there exists a  $X_{n_m}$  such that

$$P(|X_{n_m} - X| > 2^{-m}) < 2^{-m}$$

Particularly, we can choose  $n_m$  to be increasing. For the sequence  $\{X_{n_m}\}$ , we note that for any  $\epsilon > 0$ , when  $n_m$  is large,

$$P(\sup_{k \geq m} |X_{n_k} - X| > \epsilon) \leq \sum_{k \geq m} P(|X_{n_k} - X| > 2^{-k}) \leq \sum_{k \geq m} 2^{-k} \rightarrow 0$$

Thus,  $X_{n_m} \xrightarrow{a.s.} X$ .

Remarks: Need to pay attention to the SUP and sum of probability, it is similar to the max of the sequence. So we need to think about the all sequence observations probability.

- (vii) If  $X_n \xrightarrow{d} X$ , and  $P(X \equiv c) \equiv 1$ , where  $c \in R^k$  is a constant vector, then  $X_n \xrightarrow{p} c$ . Let  $X \equiv c$ .

Prove by Polya's theorem:

$$P(|X - n - c| > \epsilon) \leq 1 - F_n(c + \epsilon) + F_n(c - \epsilon) \rightarrow 1 - F_X(c + \epsilon) + F_X(c - \epsilon) = 0$$

Remarks: Polya's theorem is very useful when dealing with the  $F_n$  change to  $F$ .

- (viii) Moment convergence: Suppose that  $X_n \xrightarrow{d} X$ , then for any  $r > 0$ ,

$$\lim_{n \rightarrow \infty} E\|X_n\|_r^r = E\|X\|_r^r < \infty$$

iff  $\{\|X_n\|_r^r\}$  is uniformly integrable (UI) in the sense that

$$\limsup_{t \rightarrow \infty} E(\|X_n\|_r^r I_{\|X_n\|_r > t}) = 0$$

In particular,  $X_n \xrightarrow{L_r} X$  if and only if  $\{\|X_n - X\|_r^r\}$  is UI

(viii) If  $X_n \xrightarrow{p} X$  and  $|X_n|^r$  is uniformly integrable, then  $X_n \xrightarrow{r} X$ .

## 4 Polya's theorem

If  $F_n \xrightarrow{w} F$  and  $F$  is continuous on  $R^k$ , then

$$\limsup_{n \rightarrow \infty} \sup_{x \in R^k} |F_n(x) - F(x)| = 0.$$

This proposition implies the following useful result: If  $c_n \in R^k$  with  $C_n \rightarrow C$ , then

$$F_n(C_n) \rightarrow F(C)$$

## 5 Fatou's lemma

Given a measure space  $(\Omega, \mathbf{F}, \mu)$ , and a set  $X \in \mathbf{F}$ , let  $\{f_n\}$  be a sequence of  $(F, B_{R \geq 0})$  - measurable non-negative functions:  $f_n : X \rightarrow [0, +\infty]$ . Define the function  $f : X \rightarrow [0, +\infty]$  by setting  $f(x) = \liminf_{n \rightarrow \infty} f_n(x)$ , for every  $x \in X$ . Then  $f$  is  $(F, B_{R \geq 0})$  - measurable, and also

$$\int_X f d\mu \leq \liminf_{n \rightarrow \infty} \int_X f_n d\mu, \text{ where the integral may be infinite.}$$

## 6 Big O and small o

In calculus, two sequences of real numbers,  $\{a_n\}$  and  $\{b_n\}$ , satisfy 1.  $a_n = O(b_n)$  iff  $|a_n| \leq c|b_n|$  for all  $n$  and a constant  $c$ . 2.  $a_n = o(b_n)$  iff  $a_n/b_n \rightarrow 0$  as  $n \rightarrow \infty$ .

### 6.1 Definition

Let  $X_1, X_2, \dots$  be random vectors and  $Y_1, Y_2, \dots$  be random variables defined on a common probability space.

(i)  $X_n = O(Y_n)$  a.s. iff  $P(\|X_n\| = O(|Y_n|)) = 1$

Since  $a_n = O(1)$  means that  $\{a_n\}$  is bounded,  $\{X_n\}$  is said to be bounded in probability if  $X_n = O_p(1)$ .

$X_n = O(Y_n)$  and  $Y_n = O(Z_n)$  implies  $X_n = O(Z_n)$ .

$X_n = O(Y_n)$  does not imply  $Y_n = O_p(X_n)$ .

If  $X_n = O(Z_n)$ , then  $X_n Y_n = O_p(Y_n Z_n)$ .

If  $X_n = O(Z_n)$  and  $Y_n = O(Z_n)$ , then  $X_n + Y_n = O_p(Z_n)$ .

If  $X_n \xrightarrow{d} X$  for a random variable  $X$ , then  $X_n = O_p(1)$ .

If  $E(|X_n|) = O(a_n)$ , then  $X_n = O_p(a_n)$ , where  $a_n \in (0, \infty)$ .

If  $X_n \xrightarrow{a.s.} X$ , then  $\sup_n |X_n| = O_p(1)$ .

(ii)  $X_n = o(Y_n)$  a.s. iff  $X_n/Y_n \xrightarrow{a.s.} 0$

$X_n = o(Y_n)$  implies  $X_n = O_p(Y_n)$ .

(iii)  $X_n = O_p(Y_n)$  iff, for any  $\epsilon > 0$ , there is a constant  $C_\epsilon > 0$  such that

$$\sup_n P(\|X_n\| \geq C_\epsilon(|Y_n|)) < \epsilon$$

(iv)  $X_n = o_p(Y_n)$  iff  $X_n/Y_n \xrightarrow{p} 0$ .

Let

$$U = Y_1 + \dots + Y_i, \quad V = Y_{i+1} + \dots + Y_{n+1}$$

Then  $U \sim \text{Gamma}(i, \theta)$ ,  $V \sim \text{Gamma}(n+1-i, \theta)$ , Let

$$Z_i = U/(U+V), \quad W = U+V$$

Consider the transformation  $(U, V)^T \rightarrow (Z_i, W)^T$ , note that the transform is one-to-one with the Jacobian

$$\left| \frac{\partial(U, V)}{\partial(Z_i, W)} \right| = |W|$$

For joint distribution of  $(U, V)^T$ ,

$$\frac{1}{\Gamma(i)} \theta \exp(-\theta u) (\theta u)^{i-1} I(u > 0) \times \frac{1}{\Gamma(n+1-i)} \theta \exp(-\theta v) (\theta v)^{n-i} I(v > 0)$$

We obtain the joint density of  $(Z_i, W)$  as

$$\begin{aligned} & \frac{1}{\Gamma(i)} \theta \exp(-\theta z_i w) (\theta z_i w)^{i-1} \\ & \times \frac{1}{\Gamma(n+1-i)} \theta \exp(-\theta(1-z_i)w) (\theta(1-z_i)w)^{n-i} w \times I(0 < z_i < 1) I(w > 0) \end{aligned}$$

Thus, the marginal density of  $Z_i = X/(X+Y)$  is equal to

$$\begin{aligned} & (1-z_i)^{n-i} z_i^{i-1} I(0 < z_i < 1) \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} \int_w \theta \exp(-\theta w) (\theta w)^n I(w > 0) dw \\ & = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} (1-z_i)^{n-i} z_i^{i-1} I(0 < z_i < 1) \end{aligned}$$

That is  $Z_i \sim \text{Beta}(i, n+1-i)$

### 6.1.1 Prove Uniform distribution

We need to note that, we always construct a joint distribution as we are using the  $Y_1, \dots, Y_{n+1}$  in joint distribution. So the transformed variables are also joint distribution.

Then we need to find the relationship between  $Y_1, \dots, Y_{n+1}$  and  $(Z_1, \dots, Z_n)$ . We can set the total sum as  $S_{n+1} = Y_1 + \dots + Y_{n+1}$ .

$$\begin{aligned} Y_1 &= Z_1 \times S_{n+1} \\ Y_2 &= Z_2 \times S_{n+1} - Y_1 = (Z_2 - Z_1) \times S_{n+1} \\ Y_3 &= Z_3 \times S_{n+1} - Y_1 - Y_2 = (Z_3 - Z_2) \times S_{n+1} \\ &\dots \\ Y_n &= Z_n \times S_{n+1} - Y_1 - \dots - Y_{n-1} = (Z_n - Z_{n-1}) \times S_{n+1} \\ Y_{n+1} &= S_{n+1} - Y_1 - \dots - Y_n = (1 - Z_n) \times S_{n+1} \end{aligned}$$

Then we have the Jacobian transform distribution

$$J = \begin{pmatrix} \frac{\partial Y_1}{\partial Z_1} & \frac{\partial Y_1}{\partial Z_2} \dots & \frac{\partial Y_1}{\partial Z_n} & \frac{\partial Y_1}{\partial S_{n+1}} \\ \frac{\partial Y_2}{\partial Z_1} & \frac{\partial Y_2}{\partial Z_2} \dots & \frac{\partial Y_2}{\partial Z_n} & \frac{\partial Y_2}{\partial S_{n+1}} \\ \frac{\partial Y_3}{\partial Z_1} & \frac{\partial Y_3}{\partial Z_2} \dots & \frac{\partial Y_3}{\partial Z_n} & \frac{\partial Y_3}{\partial S_{n+1}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial Y_n}{\partial Z_1} & \frac{\partial Y_n}{\partial Z_2} \dots & \frac{\partial Y_n}{\partial Z_n} & \frac{\partial Y_n}{\partial S_{n+1}} \\ \frac{\partial Y_{n+1}}{\partial Z_1} & \frac{\partial Y_{n+1}}{\partial Z_2} \dots & \frac{\partial Y_{n+1}}{\partial Z_n} & \frac{\partial Y_{n+1}}{\partial S_{n+1}} \end{pmatrix} = \begin{pmatrix} S_{n+1} & 0 & 0 & \dots & \dots & Z_1 \\ -S_{n+1} & S_{n+1} & 0 & \dots & \dots & Z_2 - Z_1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots - S_{n+1} & S_{n+1} & 0 & Z_n - Z_{n-1} \\ 0 & 0 & \dots & 0 & S_{n+1} & 1 - Z_n \end{pmatrix}$$

$$|J| = S_{n+1}^n$$

The joint distribution of  $Y_1, \dots, Y_{n+1}$

$$\begin{aligned} f(Y_1, Y_2, \dots, Y_{n+1}) &= \prod_{i=1}^{n+1} \frac{1}{\theta} \exp\left(-\frac{y_i}{\theta}\right) I(0 < y_i < \infty) \\ &= \frac{1}{\theta^{n+1}} \exp\left(-\frac{\sum_{i=1}^{n+1} y_i}{\theta}\right) I(0 < y_i < \infty) \\ &= \frac{1}{\theta^{n+1}} \exp\left(-\frac{S_{n+1}}{\theta}\right) I(0 < y_i < \infty) \end{aligned}$$

Then the transformed variables

$$f(Z_1, Z_2, \dots, Z_n, S_{n+1}) = \frac{1}{\theta^{n+1}} \exp\left(-\frac{S_{n+1}}{\theta}\right) S_{n+1}^n I(0 < z_1 < z_2 < \dots < z_n < 1) I(S_{n+1} > 0)$$

We obtain the joint distribution of  $(Z_1, \dots, Z_n)$

$$\begin{aligned} f(Z_1, Z_2, \dots, Z_n) &= \int_s f(Z_1, Z_2, \dots, Z_n, S_{n+1}) dS \\ &= \int_s \frac{1}{\theta^{n+1}} \exp\left(-\frac{S}{\theta}\right) S^n I(0 < z_1 < z_2 < \dots < z_n < 1) I(S > 0) dS \\ &= n! I(0 < z_1 < z_2 < \dots < z_n < 1), \quad \text{Gamma integral} \end{aligned}$$



---

(b) Let

$$W_1 = Y_1, W_2 = Y_1 + Y_2, \dots, W_n = Y_1 + \dots + Y_n, S = Y_1 + \dots + Y_{n+1}.$$

Consider the transformation  $(Y_1, \dots, Y_{n+1})' \mapsto (W_1, \dots, W_n, S)'$ . Note that the transformation is one-to-one with the Jacobian

$$\left| \det \left( \frac{\partial(Y_1, \dots, Y_{n+1})}{\partial(W_1, \dots, W_n, S)} \right) \right| = \left| \det \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix} \right| = 1$$

Figure 1: Alt method

which is the joint density of order statistics of  $n$  uniform  $(0,1)$  random variables.

## 7 Construct Distribution

### 7.1 Order Statistics

The joint distribution of minimum and maximum:

Let's go for the joint cdf of the minimum and maximum

$$F_{X_1, X_n}(x, y) = P(X_{(1)} \leq x, X_{(n)} \leq y)$$

Why do we start from cdf? It is much easier to get cdf than pdf, as the pdf need to take derivative, while cdf only needs to get the integral.

And if the observations are independent, the pdf and cdf also follows the same rule.

We will need to write this in terms of the individual  $X_i$  as the minimum and maximum are statistics of the individuals. Consider instead the relationship

$$P(X_n) \leq y = P(X_{(1)} \leq x, X_{(n)} \leq y) + P(X_{(1)} > x, X_{(n)} \leq y)$$

This is the integral of  $x$ , which is a common in getting the marginal distribution from joint distribution.

We know how to write out the term on the left-hand side. The first term on the right-hand side is what we want to compute. As for the final term,  $P(X_{(1)} > x, X_{(n)} \leq y)$ ,

note that this is 0 if  $x > y$ . So, we consider  $x < y$

$$\begin{aligned} P(X_{(1)} > x, X_{(n)} \leq y) &= P(x < X_1 \leq y, x < X_2 \leq y, \dots, x < X_n \leq y) \\ &= [P(x < X_1 \leq y)]^n, \quad \text{i.i.d} \\ &= [F(y) - F(x)]^n \end{aligned}$$

From the joint density of  $(Y_1, \dots, Y_{n+1})'$ ,

$$\prod_{i=1}^{n+1} \frac{1}{\theta} \exp\{-\frac{1}{\theta} y_i\} I(0 < y_i < \infty) = (\frac{1}{\theta})^{n+1} \exp\{-\frac{\sum_{i=1}^{n+1} y_i}{\theta}\} I(0 < y_i < \infty)$$

we obtain the joint density of  $(W_1, \dots, W_n, S)$  as

$$(\frac{1}{\theta})^{n+1} \exp\{-\frac{s}{\theta}\} I(0 < w_1 < w_2 < \dots < w_n < s < \infty),$$

Let

$$Z_1 = W_1/S, Z_2 = W_2/S, \dots, Z_n = W_n/S, S.$$

Consider the transformation  $(W_1, \dots, W_n, S)' \mapsto (Z_1, \dots, Z_n, S)'$ . Note that the transformation is one-to-one with the Jacobian

$$|det \left( \frac{\partial(W_1, \dots, W_n, S)}{\partial(Z_1, \dots, Z_n, S)} \right) | = |det \begin{pmatrix} s & 0 & \dots & 0 & z_1 \\ 0 & s & \dots & 0 & z_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & s & z_n \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} | = s^n$$

From the joint density of  $(Z_1, \dots, Z_n, S)'$ ,

$$= (\frac{1}{\theta})^{n+1} \exp\{-\frac{s}{\theta}\} s^n I(0 < z_1 < z_2 < \dots < z_n < 1) I(s > 0)$$

Figure 2: Alt method 2

So, we have

$$\begin{aligned}
F_{(X_{(1)}, X_{(n)})}(x, y) &= P(X_{(1)} \leq x, X_{(n)} \leq y) \\
&= P(X_n \leq y) - P(X_1 > x, X_n \leq y) \\
&= [F(y)]^n - [F(y) - F(x)]^n
\end{aligned}$$

Now the joint pdf is

$$\begin{aligned}
f_{(X_{(1)}, X_{(n)})}(x, y) &= \frac{d}{dx} \frac{d}{dy} \{[F(y)]^n - [F(y) - F(x)]^n\} \\
&= \frac{d}{dx} nF(y)^{n-1} f(y) - n(F(y) - F(x))^{n-1} f(y) \\
&= n(n-1)(F(y) - F(x))^{n-2} f(x) f(y)
\end{aligned}$$

This hold for  $x < y$  and for  $x$  and  $y$  both in the support of the original distribution.

## 8 Moment Generating Function

### 8.1 Chi-square MGF

We can get MGF for chi-square from  $E[x^2 t]$  and  $E[(\mu + Z)^2 t]$ , where  $Z \sim N(0, 1)$ . Let's prove it in two methods:

(i) Method 1:

$$\begin{aligned}
M_i(t) &= E[x^2 t] = \frac{1}{\sqrt{2\pi}} \int \exp(x^2 t) \exp\left(-\frac{(x - \mu)^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left((t - \frac{1}{2})x^2 + \mu x - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{1}{2}(1 - 2t)\left\{x^2 - \frac{2\mu x}{(1 - 2t)} + \frac{\mu^2}{(1 - 2t)^2}\right\} + \frac{\mu^2}{2(1 - 2t)} - \frac{\mu^2}{2}\right) dx \\
&= \frac{1}{\sqrt{(1 - 2t)}} \int \frac{(1 - 2t)}{\sqrt{2\pi}} \exp\left(-\frac{(x - \frac{\mu}{1 - 2t})^2}{2(1 - 2t)^{-1}}\right) dx \left[\exp\left(\frac{\mu^2 t}{1 - 2t}\right)\right] \\
&= \frac{1}{\sqrt{(1 - 2t)}} \exp\left(\frac{\mu^2 t}{1 - 2t}\right), \quad \lambda = \mu^2 \\
&= \frac{1}{\sqrt{(1 - 2t)}} \exp\left(\frac{\lambda t}{1 - 2t}\right)
\end{aligned}$$

Then the MGF for  $Q_i \sim \chi_{k_i}^2(\lambda_i)$

$$\begin{aligned}
M(t) &= E\left[\sum_{i=1}^k x_i^2 t\right] = \prod_{i=1}^k M_i(t) \\
&= \left(\frac{1}{\sqrt{1-2t}}\right)^k \exp\left(\frac{\sum_{i=1}^k \lambda_i t}{1-2t}\right) \\
&= \left(\frac{1}{\sqrt{1-2t}}\right)^k \exp\left(\frac{\lambda t}{1-2t}\right) \\
&= (1-2t)^{-k/2} \exp\left(\frac{\lambda t}{1-2t}\right), \quad \text{i.i.d}
\end{aligned}$$

(ii) Method 2:

$$\begin{aligned}
M(t) &= E[(\mu + Z)^2 t] = \frac{1}{\sqrt{2\pi}} \int \exp((\mu + Z)^2 t) \exp\left(-\frac{Z^2}{2}\right) dz \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left((t - \frac{1}{2})z^2 + 2\mu t z + \mu^2 t\right) dz \\
&= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{(1-2t)}{2}\left\{z^2 - \frac{4\mu t z}{(1-2t)} + \frac{2\mu^2 t^2}{(1-2t)^2}\right\} + \frac{2\mu^2 t^2}{(1-2t)} + \mu^2 t\right) dz \\
&= \frac{1}{\sqrt{(1-2t)}} \left[\exp\left(\frac{\mu^2 t}{1-2t}\right)\right] \int \frac{1}{\sqrt{2\pi/(1-2t)}} \exp\left(-\frac{(z - \frac{2\mu t}{1-2t})^2}{2(1-2t)^{-1}}\right) dz \\
&= \frac{1}{\sqrt{(1-2t)}} \exp\left(\frac{\mu^2 t}{1-2t}\right), \quad t < 1/2
\end{aligned}$$

The general case of a linear combination of independent  $\chi_{k_i}^2(\lambda_i)$

$$Q = \sum_{i=1}^k a_i Q_i$$

We also can prove using MGF.

## 8.2 Linear Combination of Chi-Square Distribution

The linear combination of chi-square distribution  $Y_j$ . Let us denote by  $X \sim \Gamma(r, \lambda)$  the fact that the r.v.  $X$  has a Gamma distribution with shape parameter  $r$  and rate parameter  $\lambda$

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} \exp(-\lambda x) x^{r-1}, \quad (r, \lambda > 0, x > 0)$$

Then we have, for  $j = 1, \dots, p$ ,

$$Y_j \sim \Gamma\left(\frac{k_j}{2}, \frac{1}{2}\right) \rightarrow Z_j = w_j Y_j \sim \Gamma\left(\frac{k_j}{2}, \frac{1}{2w_j}\right)$$

The MGF for linear combinations  $Z_j = w_j Y_j$

$$\begin{aligned} M(t) &= E[\exp(Y_j t)] = (1 - 2t)^{-k/2} \exp\left(\frac{\lambda t}{1 - 2t}\right) \\ M_{Z_j}(t) &= E[\exp(w_j Y_j t)] = E[\exp(Y_j (w_j t))] \\ &= (1 - 2w_j t)^{-1/2} \exp\left(\frac{\lambda w_j t}{1 - 2w_j t}\right) \end{aligned}$$

$$\begin{aligned} M_Y(t) &= E[\exp(Y t)] = E[\exp(t[w_1 Y_1 + w_2 Y_2 + w_3 Y_3 + \dots w_n Y_n])] \\ &= E[\exp(w_1 t Y_1)] E[\exp(w_2 t Y_2)] \dots E[\exp(w_n t Y_n)] \\ &= M_{X_1}(w_1 t) M_{X_2}(w_2 t) M_{X_3}(w_3 t) \dots M_{X_n}(w_n t) \\ &= \prod_{i=1}^n M_{X_i}(w_i t) \end{aligned}$$

The third equation comes from the properties of exponents, as well as from the expectation of the product of functions of independent random variables.

I need to pay attention that, only under independent and identical situation, we can write

$$M_Y(t) = M_X(t)^n$$

Other than that, we can not further simplify that. So back to the non-central chi-square distribution, we have the MGF of Y

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^n M_{X_i}(w_i t) \\ &= \prod_{i=1}^n (1 - 2w_i t)^{-1/2} \exp\left(\frac{\lambda w_i t}{1 - 2w_i t}\right) \end{aligned}$$

Then we can see that the shape parameter is  $\frac{1}{2w_i}$ . If we want to have a non-central chi-square distribution for Y, then all  $w_j$  need to be the same.

### 8.3 Exponential Distribution Family

KGF could be used to directly get the expectation and variance, more common than MGF. To get KGF, we will need to write distribution in exponential distribution.

Suppose the exponential distribution family is

$$f(Y, \theta) = \exp(\phi(y\theta - b(\theta) - c(y)) - 0.5s(y, \phi))$$

The MGF of exponential family

$$\begin{aligned} M_Y(t) &= E[\exp(yt)] = \int \exp(yt) \exp(\phi(y\theta - b(\theta) - c(y)) - 0.5s(y, \phi)) dy \\ &= \int \exp(\phi(y(\theta + t/\phi) - b(\theta) - c(y)) - 0.5s(y, \phi)) dy \\ &= \exp(\phi[b(\theta + t/\phi) - b(\theta)]) \int \exp(\phi(y(\theta + t/\phi) - b(\theta + t/\phi) - c(y)) - 0.5s(y, \phi)) dy \\ &= \exp(\phi[b(\theta + t/\phi) - b(\theta)]) \end{aligned}$$

The KGF is  $\phi[b(\theta + t/\phi) - b(\theta)]$ , then we can get the expectation and variance

$$\begin{aligned} E(y) &= \left. \frac{\partial K(t)}{\partial t} \right|_{t=0} = \dot{b}(\theta) \\ Var(y) &= \left. \frac{\partial^2 K(t)}{\partial t \partial t} \right|_{t=0} = \phi^{-1} \ddot{b}(\theta) \end{aligned}$$

The MGF/KGF has shown that we can use the derivative function to get expectation or variance other than using the integral. The efficiency of computation also could be shown in the getting the covariance matrix using Fisher Information.

## 9 Fisher Information

The Fisher Information always comes with the asymptotic normal distribution of the estimator, and hypothesis testing.

### 9.1 Multinomial Distribution

Multinomial distribution is a very typical distribution to demonstrate the relationship between the covariance matrix and Fisher Information.

If the observations from multinomial distribution are independent, so we can construct the variance and covariance between two observations, and then get the covariance matrix. This step we can use the MGF or by definition.

But the Fisher Information don't use the inverse of Covariance, use the definition of Fisher Information, which is the variance of score function.

The log-likelihood function of Multinomial distribution

$$p(x, \theta) = \binom{n}{x_1, x_2, \dots, x_k} \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3} \dots \theta_k^{x_k}$$

$$l_n(\theta) = \text{const} + \sum_{i=1}^k x_i \log(\theta_i)$$

Then the score function

$$S(x, \theta) = \left( \frac{x_1}{\theta_1}, \frac{x_2}{\theta_2}, \dots, \frac{x_k}{\theta_k} \right) = \text{diag}\left(\frac{1}{\theta}\right)x$$

Consequently, the Fisher Information

$$\begin{aligned} I_n(\theta) &= \text{Var}(S(\theta)) = \text{diag}\left(\frac{1}{\theta}\right) \text{Var}(x) \text{diag}\left(\frac{1}{\theta}\right) \\ &= n \text{diag}(1/\theta) [\text{diag}(\theta) - \theta \theta^T] \text{diag}(1/\theta) \\ &= n [\text{diag}(1/\theta) - (\text{diag}(1/\theta)\theta)(\text{diag}(1/\theta)\theta)^T] \\ &= n [\text{diag}(1/\theta) - 11^T] \\ &= n \begin{pmatrix} \frac{1-p_1}{p_1} & -1 & -1 & \dots & -1 \\ -1 & \frac{1-p_2}{p_2} & -1 & \dots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ -1 & -1 & -1 & \dots & \frac{1-p_k}{p_k} \end{pmatrix} \end{aligned}$$

## 10 Weighted Statistics

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

### 10.1 Question

Note that  $\bar{X}_{nw}$  is a linear combination of  $X_1, \dots, X_n$ , 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  $n \rightarrow \infty$ . 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.

## 10.2 MGF

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

Then the linear combination  $y_n$

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

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

## 10.3 Orthogonal Matrix

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

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

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  $Z_1, \dots, Z_n \sim N(0, \sigma^2)$

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

Also,

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

Since  $Y_n$  and  $S_n^2$  are functions of  $Z_1$  and  $(Z_2, \dots, Z_n)$  respectively, and from the independence of  $Z_i, (i = 1, \dots, n)$ , 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  $w_1 = w_2 = \dots = w_n = 1/n$ ,  $Y_n = \sum_{i=1}^n X_i/(\sigma\sqrt{n})$ , 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.

## 11 Sufficient and Complete Statistics

### 11.1 Minimum Sufficient Statistics

### 11.2 Complete Statistics

### 11.3 Ancillary Statistics

## 12 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

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

where  $\rho$  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

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

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  $\rho$

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

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  $Z_1$  and  $Z_2$  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.

## 12.1 Conditional Distribution

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

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

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  $X = x$ ,

$$Z_1 = \frac{x - \mu_X}{\sigma_X}$$

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

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

$$E[Y|X = x] = \sigma_Y \rho \frac{x - \mu_X}{\sigma_X} + \sigma_Y \sqrt{1 - \rho^2} E[Z_2] + \mu_Y$$

$$= \mu_Y + \rho \sigma_Y \frac{x - \mu_X}{\sigma_X}$$

$$Var[Y|X = x] = \sigma_Y^2 (1 - \rho^2) Var(Z_2) = (1 - \rho^2) \sigma_Y^2$$

### 12.1.1 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

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

where  $\mu$  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.

### 12.1.2 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

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

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

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

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  $V_{ab} = 0$ .

## 12.2 Schur Complement

In order to derive the conditional distributions, we are going to rely 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  $(p + q) \times (p + q)$  matrix. If  $D$  is invertible, then the **Schur 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  $q \times q$  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.

### 12.2.1 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  $I_p$  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  $2 \times 2$  matrix and this partitioned matrix.

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

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:

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

### 12.2.2 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  $\Sigma$

$$\begin{aligned} \text{Cov}(X|Y) &= A - BC^{-1}B^T \\ E(X|Y) &= E(X) + BC^{-1}(Y - E(Y)) \end{aligned}$$

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

### 12.3 b

Consider the following

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

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

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

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

The conditional score statistic is the derivative of the conditional distribution

$$\begin{aligned} U_\psi(\xi) &= \frac{\partial l_c(\xi, \psi_0)}{\partial \psi} \Big|_{\psi_0=\psi} \\ p(\mathbf{Y}|\xi) &= p(\mathbf{Y}|s_\lambda(\psi_0), \xi) p(s_\lambda(\psi_0)|\xi), \quad p(\mathbf{Y}|s_\lambda(\psi_0), \xi) = \frac{p(\mathbf{Y}|\xi)}{p(s_\lambda(\psi_0)|\xi)} \\ l_c(\xi, \psi_0) &= \log p(\mathbf{Y}|s_\lambda(\psi_0), \xi) = \log p(\mathbf{Y}|\xi) - \log p(s_\lambda(\psi_0)|\xi) \end{aligned}$$

Then we need to prove

$$\begin{aligned} U_\psi(\xi) &= \frac{\partial l_c(\xi, \psi_0)}{\partial \psi} \Big|_{\psi_0=\psi} = \partial_\psi \log p(\mathbf{Y}|\xi) - \partial_\psi \log p(s_\lambda(\psi_0)|\xi) \\ \partial_\psi \log p(s_\lambda(\psi_0)|\xi) &= E[\partial_\psi \log p(Y|\xi) | s_\lambda(\psi_0)] \Big|_{\psi_0=\psi} \end{aligned}$$

We can write

$$\begin{aligned} \log p(\mathbf{Y}|\xi) &= \log p(\mathbf{Y}|s_\lambda(\psi_0), \xi) + \log p(s_\lambda(\psi_0)|\xi) \\ E(\partial_\psi [\log p(\mathbf{Y}|\xi) | s_\lambda]) &= E(\partial_\psi [\log p(\mathbf{Y}|s_\lambda(\psi_0), \xi) | s_\lambda]) + E(\partial_\psi [\log p(s_\lambda(\psi_0), \xi) | s_\lambda]) \end{aligned}$$

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

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

So,

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

Then we show

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

- (b) Suppose that  $y_1; \dots; y_n$  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  $\psi$  is the parameter of interest. Derive the conditional likelihood of  $\psi$  and show that this conditional likelihood is free of  $\lambda$ .

The joint distribution of  $(y_1, \dots, y_n)$  is given by

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

Thus,  $S_0 = \sum_{i=1}^n y_i$  is the sufficient and complete statistics for  $\lambda_0$ , and  $S_1 = \sum_{i=1}^n y_i x_{i1}$  is the sufficient and complete statistics for  $\lambda_1$ .

The conditional distribution of  $\psi$  given  $S_0, S_1$  is given by

$$\begin{aligned} p(\mathbf{Y}, \psi | S = (S_0, S_1)) &= \frac{\exp(\sum_{i=1}^n y_i(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^n \exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \log y_i!)}{\sum_{y' \in S} \exp(\sum_{i=1}^n y'_i(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \sum_{i=1}^n \exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \log y'_i!)} \\ &= \frac{\exp(S_1 \lambda_0 + S_2 \lambda_1 + S_3 \psi) - \sum_{i=1}^n \exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \log y_i!}{\sum_{y' \in S} \exp(S'_1 \lambda_0 + S'_2 \lambda_1 + S'_3 \psi) - \sum_{i=1}^n \exp(\lambda_0 + \lambda_1 x_{i1} + \psi x_{i2}) - \log y'_i!} \\ &= \frac{\exp(S_3 \psi - \log y_i!)}{\sum_{y' \in S} \exp(S'_3 \psi - \log y'_i!)}, \quad S_3 = \sum_{i=1}^n y_i x_{i2}, S'_3 = \sum_{i=1}^n y'_i x_{i2} \end{aligned}$$

which is independent of  $\lambda$ .

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

The log likelihood of the conditional distribution is

$$l_c(\psi) = S_3 \psi - \log y_i! - \log \left[ \sum_{y' \in S} \exp(S'_3 \psi - \log y'_i!) \right], \quad S_3 = \sum_{i=1}^n y_i x_{i2}, S'_3 = \sum_{i=1}^n y'_i x_{i2}$$

The score function and observed fisher information is

$$\begin{aligned} U_\psi(\xi) &= \frac{\partial l_c(\xi, \psi_0)}{\partial \psi} \Big|_{\psi_0=\psi} \\ &= \psi - \frac{\sum_{y' \in S} S'_3 \exp(S'_3 \psi - \log y'_i!)}{\sum_{y' \in S} \exp(S'_3 \psi - \log y'_i!)} \\ \frac{\partial^2 l_c(\xi, \psi_0)}{\partial \psi^2} &= \left[ \frac{\sum_{y' \in S} S'_3 \exp(S'_3 \psi - \log y'_i!)}{\sum_{y' \in S} \exp(S'_3 \psi - \log y'_i!)} \right]^2 - \frac{\sum_{y' \in S} S'^2_3 \exp(S'_3 \psi - \log y'_i!)}{\sum_{y' \in S} \exp(S'_3 \psi - \log y'_i!)} \end{aligned}$$

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  $y_1 \sim \text{Poisson}(\mu_1)$  and  $y_2 \sim \text{Poisson}(\mu_2)$ , 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  $\lambda$  represents the nuisance parameter.

- (i) Show that the log-likelihood function of  $\xi$  can be written as

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

where  $\lambda$  is a function of  $\mu_2$ . Explicitly state what  $\lambda$  is.

Write the joint distribution of  $y_1, y_2$

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

where

$$\begin{aligned} \psi &= \log \frac{\mu_1}{\mu_2} \\ \lambda &= \log \mu_2 \end{aligned}$$

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

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

The conditional distribution is a binomial,  $B(S, \psi/(1 + \psi))$ .

The score function and observed fisher information

$$\begin{aligned}
\log Y(\psi|S, \lambda) &= y_1 \log \psi - S \log(1 + \psi) + \log \binom{S}{y_1} \\
\partial_\psi \log Y(\psi|S, \lambda) &= \frac{y_1}{\psi} - \frac{S}{1 + \psi} = 0, \quad \hat{\psi} = y_1/(S - y_1) \\
\partial_\psi^2 \log Y(\psi|S, \lambda) &= -\frac{y_1}{\psi^2} + \frac{S}{(1 + \psi)^2}
\end{aligned}$$

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

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