STATISTICAL INFERENCE THEORY

STAT 560 LECTURE NOTES

SHIHAO (OWEN) TONG[*]

University of British Columbia

Lecture given by prof Ruben H. Zamar .Template from here. No guarantee for accuracy. For personal use only.

CONTENTS

1 Asymptotic Inference 2
- 1.1 Pointwise Convergence . . . 2
- 1.2 Almost Sure Convergence . . . 3
- 1.3 Convergence in Probability . . . 3
- 1.4 Convergence in Distribution . . . 4
- 1.5 Relationship among Modes of Convergence . . . 4

2 Multinomial Distribution 7

3 Introduction To Information Theory 7
- 3.1 Entropy . . . 7
- 3.2 Differential Entropy . . . 7
- 3.3 Mutual Information . . . 8
- 3.4 Proof of non-negative mutual information . . . 9

4 Multivariate Normal Distribution 11
- 4.1 Generalized Multivariate Normal . . . 11
- 4.2 Properties of Multivariate Normal Distribution . . . 13

5 linear model and least square 16
- 5.1 Linear Regression Model . . . 16

[*]tongshihaoowne@outlook.com

6 Distribution of Normal Quadratic Form 21
6.1 The Fisher-Cochran Theorem 22
6.2 Application of Fisher-Cochran Theorem: The Normal Location-Scale Model 24

7 Comparison of Nested Model 26
7.1 Example: Constant signal vs Pure error 26
7.2 Example 2: Linear regression model vs Local scale Model . . . 27

8 Extended Fisher - cochran Theorem 29

9 Maximum Likelihood 29
9.1 The Information Inequality 31
9.2 Score function and Fisher Information matrix 32
9.3 Regularity Conditions 33
9.4 Properties of Score function 33
9.5 Confidence Interval Construction MLE 34

10 Expectation maximization (EM) Algorithm 35
10.1 Mixture Model 35

Notations: We use upper case to represent randomness, lower case to be represent a single realization. We use bold to represent a vector (i.e dimension ≥ 2) and non-bold for a one-dimensional variable.

1 ASYMPTOTIC INFERENCE

Several definition & concepts before moving on are

- Probability Space $(\Omega, \mathcal{F}, \mathrm{P})$: Where Ω is the **Sample space**, $\mathcal{F}$ is the σ-field, and $(\Omega, \mathcal{F})$ is a measurable space. P is the probability measure (Refer to Mathematical Statistics, Jun Shao for detail).
- Random vector: The random vector is a map

$$\mathbb{X} : \Omega \to \mathbb{R}^p \tag{1}$$

  If $p = 1$, then X is a random variable.

Let $\mathrm{X}_n \sim \mathrm{F}_n$, n = 1,2,..,n be a sequence of r.var or vect. We wish to define convergence of X_n to X_0 (i.e the convergence of sequence of r.v).

1.1 *Pointwise Convergence*

1.1 DEFINITION. Let X_n ne a sequence of random variable. If

$$\lim_{n\to\infty} X_n(\omega) = X_0(\omega), \quad \forall \omega \in \Omega \tag{2}$$

then we say X_n converges to X_0 pointwisely.

- This type of convergence is very strong. At every point of $\omega \in \Omega$ it converges so difficult to obtain.
- Not necessary to have in order to get a good approximation for the probability behavior of X_n.

1.2 *Almost Sure Convergence*

A.s convergence also called

- "Convergence with probability 1"
- "Convergence almost everyshere"

1.2 DEFINITION. (A.S convergence) Let X_n, $n = 0, 1, 2, ...$ are defined on the **same probability space** $(\Omega, \mathcal{F}, \mathcal{P})$. We say X_n converges almost surely to X_0 if

$$\mathcal{P}(\lim_{n\to\infty} X_n = X_0) = 1 \tag{3}$$

Equivalently, $\exists$ a set A, with $\mathcal{P}(A) = 1$?, such that

$$\lim_{n\to\infty} X_n(\omega) = X_0(\omega), \quad \forall \omega \in A$$

Equivalently, $X_n \to X_0$ a.s if

$$\lim_{n\to\infty} \mathcal{P}(\left|X_m - X_0\right| < \epsilon, \forall m \geq n) = 1, \ \forall \epsilon > 0 \tag{4}$$

Proof. (Equivalence of definition (3) and (4)) Make up □

1.3 EXAMPLE. (a.s convergence) Make up later

1.3 *Convergence in Probability*

1.4 DEFINITION. (Converges in probability) Let X_n, $n = 0, 1, 2, ...$ be defined on the same probability space $(\Omega, \mathcal{F}, \mathcal{P})$. If

$$\lim_{n\to\infty} \mathcal{P}(\left|X_n - X_0\right| < \epsilon) = 1 \tag{5}$$

notation $X_n \to_p X_0$, then we say X_n converges in probability to X_0.

1.4 *Convergence in Distribution*

1.5 DEFINITION. (Convergence in Distribution) Let X_n be sequence of r.v. which are NOT necessarily defined on the same probability space. If

$$\lim_{n\to\infty} F_n(x) = F_0(x) \tag{6}$$

for all points x at which F_0 is **continuous**, then we say $X_n \to_d X_0$

The "Different probability space" means the Ω can be different.

1.5 *Relationship among Modes of Convergence*

1.6 PROPOSITION. *Convergence is preserved by continuous functions. Let $g(x)$ be a continuous function. Then*

$$X_n \to_{a.s} X_0 \implies g(X_n) \to_{a.s} g(X_0)$$

$$X_n \to_p X_0 \implies g(X_n) \to_p g(X_0)$$

$$X_n \to_d X_0 \implies g(X_n) \to_d g(X_0)$$

Type	$\Longrightarrow$ A.S	In Probability	In Distribution
Almost Surely	Yes	Yes	Yes
In Probability	No	Yes	Yes
In Distribution	No	No	Yes
In Distribution to C (constant)	No	Yes	Yes

Table 1: Relationship between Modes of Convergence

1.7 PROPOSITION. *Let* $\mathbf{X}_n = \begin{pmatrix} X_{1,n} \\ X_{2,n} \\ \dots \\ X_{m,n} \end{pmatrix}$ *and* $\mathbf{X}_0 = \begin{pmatrix} X_{1,0} \\ X_{2,0} \\ \dots \\ X_{m,0} \end{pmatrix}$. *Then*

$$X_{i,n} \to_{a.s} X_{i,0}\ (i = 1, \dots, m) \implies \mathbf{X_n} \to_{a.s} \mathbf{X}_0$$

$$X_{i,n} \to_p X_{i,0}\ (i = 1, \dots, m) \implies \mathbf{X_n} \to_p \mathbf{X}_0$$

$$X_{i,n} \to_d X_{i,0}\ (i = 1, \dots, m) \not\implies \mathbf{X_n} \to_d \mathbf{X}_0$$

Exmples make up

1.8 THEOREM. *(Slutzky's Theorem) If $X_n \to_d X_0$ and $Y_n \to_p c$, then*

$$X_n + Y_n \to X_0 + c$$

$$X_n Y_n \to_d cX_0$$

Proof. Makeup □

1.9 THEOREM. *(Weak Law of Large Number) Suppose that* $X_1, X_2, ...$ *are i.i.d with* $\mathbb{E}(|X_i|) < \infty$ *and* $\mathbb{E}(X_i) = \mu$. *Then*

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \to_p \mu \tag{7}$$

1.10 THEOREM. *(Strong Law of Large Number) Suppose that* $X_1, X_2, ...$ *are i.i.d with* $\mathbb{E}(|X_i|) < \infty$ *and* $\mathbb{E}(X_i) = \mu$. *Then*

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \to_{a.s} \mu \tag{8}$$

1.11 THEOREM. *(Central Limit Theorem)(Standard Version) Suppose* $X_1, ..., X_n, ...$ *are i.i.d with* $\mathbb{E}(X_i) = \mu$ *and* $Var(X_i) = \sigma^2 < \infty$, *Then*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to_d N(0, 1) \tag{9}$$

1.12 THEOREM. *(Central Limit Theorem for random Vectors)(Standard Version) Suppose* $\mathbf{X}_1, ..., \mathbf{X}_n, ...$ *are i.i.d with* $\mu = \begin{pmatrix} \mathbb{E}(X_1) \\ ... \\ \mathbb{E}(X_p) \end{pmatrix}$ *and* $Var(X_i) = \sigma_i^2 < \infty$, $i = 1, 2, ..., p$, *Then*

$$\sqrt{n}(\overline{\mathbf{X}}_n - \mu) \to_d N(\mathbf{0}, \Sigma) \tag{10}$$

Equivalently

$$\sqrt{n}\Sigma^{-1/2}(\overline{\mathbf{X}}_n - \mu) \to_d N(\mathbf{0}, \mathbf{I}) \tag{11}$$

where

$$\Sigma^{-1/2} = \sum_{j=1}^{p} \lambda_j^{-1/2} \mathbf{a}_j \mathbf{a}_j'$$

here λ_j *are the eigenvalues of* Σ *and* $\mathbf{a}_j$ *are those corresponding eigenvectors.*

Property of the covariance matrix Σ:

1. Σ is symmetric, positive definite (i.e one of equivalent statement is that for real symmetric matrix $\mathbf{A}$, $\exists \mathbf{C}$ which is real and invertible s.t $A = CC'$). Due to the existence of $\mathbf{C}$, we define the $\Sigma^{-1/2}$ to be

$$\Sigma = \mathbf{C}\mathbf{C}' \to \Sigma^{-1/2} = \mathbf{C}$$

2. Further more, $\Sigma^{-1/2} = \sum_{j=1}^{p} \lambda_j^{-1/2} \mathbf{a}_j \mathbf{a}_j'$ is called the *Spectral Decomposition* of Σ which is not unique since let $\mathbf{Q}$ be any orthogonal matrix

$$\Sigma = \mathbf{C}\mathbf{C}' = \underbrace{\mathbf{C}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{C}' = \mathbf{C}\mathbf{Q}\mathbf{Q}'\mathbf{C}'}_{\text{orthogonal so } \mathbf{Q}^{\mathrm{T}}=\mathbf{Q}'} = (\mathbf{C}\mathbf{Q})(\mathbf{C}\mathbf{Q})'$$

so $\mathbf{CQ}$ is also a solution. Also it can be verified taht $\Sigma^{-1/2}$ is also symmetric Makeup.

3. From (10) to (11) involving the matrix operation. Prove it in a general way. The general conclusion is if vector $\mathbf{X} \sim \mathrm{F}_{\mathrm{X}}(\mu, \Sigma)$, then let $\mathrm{A} \in \mathrm{M}(\mathbb{R}, n \times n)$, then

$$\mathbf{AX} \sim \mathrm{F}_{\mathrm{AX}}(\mathbf{AX}, \mathbf{A}\Sigma\mathbf{A}')$$

the operation here involves $\mathrm{V}ar(\mathbf{AX}) = \mathbf{A}\mathrm{V}ar(\mathrm{X})\mathbf{A}'$. Simple proof here is

$$\mathrm{V}ar(\mathbf{AX}) = \mathbb{E}[(\mathbf{A}(\mathbf{X}-\mu))(\mathbf{A}(\mathbf{X}-\mu))']$$

$$= \mathbb{E}[(\mathbf{A}(\mathbf{X}-\mu))((\mathbf{X}-\mu)'\mathbf{A}')] = \mathbf{A}\mathbb{E}[((\mathbf{X}-\mu))((\mathbf{X}-\mu)')]\mathbf{A}'$$

$$= \mathbf{A}\Sigma\mathbf{A}'$$

Notice the CLT requires the second moment constrains (i.e the finite variance). In practice, usually when we have $n \geq 15$ we can confidently imply CLT.

1.13 THEOREM. *(Chevychev's Inequality) Let* $\mathbb{E}(\mathrm{X}) = \mu$. *Then*

$$\mathcal{P}(\left|\mathrm{X}-\mu\right| < \epsilon) \geq 1 - \frac{\mathrm{V}ar(\mathrm{X})}{\epsilon^2}$$

There are other versions of this inequality and the one stated here is not the standard one. The WLLN is proved by this inequality. However the SLLN is harder to prove.

1.14 THEOREM. *(Delta Method for Univariate) Suppose that* $\sqrt{n}(\mathrm{X}_n - \mu) \to_d \mathrm{N}(0, \sigma^2)$ *where* $\mathbb{E}(\mathrm{X}_i) = \mu$ *and* $\mathrm{V}ar(\mathrm{X}_i) = \sigma^2$ *(not necessarily i.i.d) Let* $g(t)$ *be a continuously differentiable function at* μ. *Then*

$$\sqrt{n}[g(\mathrm{X}_n) - g(\mu)] \to_d \mathrm{N}(0, [g'(\mu)]^2\sigma^2) \tag{12}$$

Notice $\mathbb{E}(g(\mathrm{X})) \neq g(\mathbb{E}(\mathrm{X}))$.

1.15 THEOREM. *(Delta Method for Multivariate) Suppose that* $\sqrt{n}(\mathbf{X}_n - \mu) \to_d \mathrm{N}(\mathbf{0}, \Sigma)$. *Let* $g(\mathbf{t})$ *be a continuously differentiable function at* μ. *Then*

$$\sqrt{n}[g(\mathbf{X_n}) - g(\mu)] \to_d \mathrm{N}(\mathbf{0}, \nabla_g(\mu)'\Sigma\nabla_g(\mu)) \tag{13}$$

where $\nabla_g(\mathbf{t})$ *is the gradient of* g *which is*

$$\nabla_g(\mathbf{t}) = \left(\frac{\partial g(\mathbf{t})}{\partial t_j}\right) = \begin{pmatrix} \frac{\partial g(\mathbf{t})}{\partial t_1} \\ \frac{\partial g(\mathbf{t})}{\partial t_2} \\ \dots \\ \frac{\partial g(\mathbf{t})}{\partial t_p} \end{pmatrix}$$

2 MULTINOMIAL DISTRIBUTION

3 INTRODUCTION TO INFORMATION THEORY

3.1 *Entropy*

3.1 DEFINITION. (Entropy) Let $\mathbf{X}$ be a random vector with pmf $f(\mathbf{x}_i) = \mathbb{P}(\mathbf{X} = \mathbf{x}_i)$. Then the entropy of $\mathbf{X}$ is defined as

$$\mathrm{H}(\mathbf{X}) = -\sum f(\mathbf{x}_i)\log(f(\mathbf{x}_i)) = -\mathbb{E}\{\log(f(\mathbf{X}))\} \tag{14}$$

The negative sign before the summation is to make the entropy positive, since $f(\mathbf{x}_i)$ is always between 0 and 1. Entropy is interpreted as measure of randomness or uncertainty. Some comments:

- Entropy does not depend on the value of random variable. It can be seen that the value of $\mathbf{X}$, which is $\mathbf{x}_i$ only appear in the density function. So what matters is the probability. Compare with other characteristics of r.v like covariance, mean etc. all to do with the value of random variable.
- Let the r.v can only take two possible outcomes and one with probability p and one with $1-p$. Then we take derivative w.r.t p and find at $p = 1/2$ the entropy is maximized. That is, when two cases are equally likely, entropy is maximized. This can be extended to more than two possible outcomes.
- Bigger difference between those probability, bigger entropy.

Exercises: Suppose that $\mathbf{X}$ *can take* n *possible values with probabilities* $p_1, p_2, ..., p_n$. *Show that in this case* $\mathrm{H}(\mathbf{X})$ *is maximized when* $p_i = 1/n$ *for all* i

3.2 *Differential Entropy*

3.2 DEFINITION. Let $\mathbf{X}$ be a continuous random vector with density $f(x)$. THe differential ecntropy of $\mathbf{X}$ is defined as

$$\mathrm{H}(\mathbf{X}) = -\mathbb{E}\{\log(f(\mathbf{X}))\} = -\int \dots \int f(x)\log(f(x))\mathbf{dx} \tag{15}$$

NOTES:

- The definition of entropy and differential entropy in terms of expected values are identical.
- But the behavior of H in the discrete and continuous cases are rather different.
  - **Differential entropy can be negative** This is cause by the pdf which is no longer required to be within $[0, 1]$. For example the uniform distribution of $X \in [1, a]$, then log of $f(x)$ will all be greater than 0 and so the entropy will be negative. More over, the limit

$$\lim_{a \to 0} \mathrm{H}(\mathrm{U}ni([0, a])) = -\infty$$

  - **H(X) is invariant under one-to-one transformation in the discrete case but not in the continuous case**. For example, let

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) = \begin{pmatrix} g_1(\mathbf{X}) \\ g_2(\mathbf{X}) \\ \dots \\ g_n(\mathbf{X}) \end{pmatrix}, \quad \mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y}) = \mathbf{h}(\mathbf{Y}) = \begin{pmatrix} h_1(\mathbf{Y}) \\ h_2(\mathbf{Y}) \\ \dots \\ h_n(\mathbf{Y}) \end{pmatrix}$$

    First Let $\mathbf{X}$ be discrete r.v then we have

$$\mathrm{H}(\mathbf{Y}) = -\sum \mathbb{P}(\mathbf{Y} = \mathbf{y}_i) \log(\mathbb{P}(\mathbf{Y} = \mathbf{y}_i))$$

$$= -\sum \mathbb{P}(\mathbf{g}(\mathbf{X}) = \mathbf{g}(\mathbf{x}_i)) \log(\mathbf{g}(\mathbf{X}) = \mathbf{g}(\mathbf{x}_i))$$

$$= -\sum \mathbb{P}(\mathbf{X} = \mathbf{x}_i) \log(\mathbb{P}(\mathbf{X} = \mathbf{x}_i)) = \mathrm{H}(\mathbf{X})$$

    Now let $\mathbf{X}$ be a continuous random vector. For example let $\mathbf{X} \sim \mathrm{U}nif orm(0, 1)$ and make the transformation to be $\mathbf{Y} = a\mathbf{X}$. Then we can check that

$$\mathrm{H}(\mathbf{X}) = \log(1) = 0, \quad \mathrm{H}(a\mathbf{X}) = \log(a)$$

    which is not equivalent.

Exerciese: Let $\mathbf{X}$ *be a continuous random vector and* $\mathbf{Y} - \mathbf{MX}$ *where* $\mathbf{M}$ *is invertiable constant matrix then*

$$\mathrm{H}(\mathbf{Y}) = \mathrm{H}(\mathbf{X}) + \log(\det\mathbf{M})$$

Exerciese: Derive and analyze the entropy of the following variables: Binomial(n,p), Begative binomial(m,p), Possion(λ*), Normal(*μ*,* σ^2*) and Gamma(k,*λ*)*

3.3 *Mutual Information*

3.3 DEFINITION. Let $\mathbf{X}$ be a random vector. The mutual information of r.v $\mathbf{X}$ is defined as

$$D(\mathbf{X}) = \sum_{i=1}^{d} H(X_i) - H(\mathbf{X}) \tag{16}$$

Mutual information can be thought as an extension of the correlation coefficient which measure the ability if one variable can be used to predict the other in a linear sense. The mutual information is more general. It does not assume the linear relationship. It is a general measure of the information that one variable contains which is useful to predict the others.

3.4 FACT. If those components of the random vectors are independent from each other, then $H(\mathbf{X}) = \sum_{i=1}^{d} H(X_i)$.

Proof. Assume all X_i components are continuous random variables. Then

$$H(\mathbf{X}) = -\int ... \int f_{\mathbf{X}}(X_1 = x_1, ...X_d = x_d) \log(f_{\mathbf{X}}(X_1 = x_1, ..., X_d = x_d))\mathbf{dx}$$

$$= -\int ... \int \prod_{i=1}^{d} f_{X_i}(X_i = x_i) * \sum_{i=1}^{d} \log_{X_i}(f(X_i = x_i))\mathbf{dx}$$

$$= -\sum_{i=1}^{d} \int f_{X_i}(X_i = x_i) \log(f_{X_i}(X_i = x_i))$$

$$= \sum_{i=1}^{d} H(X_i)$$

□

So it can be seen that the mutual information is the difference between the entropy that we would have when X_is are independent and the actual entropy of the random vector. Also notice intuition suggests that $D(\mathbf{X})$ should be greater than 0 (i.e more messy when X_i are independent). This will be proved.

Exercises: let $\mathbf{X}$ *be the bivariate normal with mean* μ *and covariance matrix*

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

calculate the mutual information of $\mathbf{X}$.

Exercises: Consider the linear model and what the mutual information suggests.

3.4 *Proof of non-negative mutual information*

3.5 DEFINITION. (Janson inequality) Suppose that $g(x)$ is a convex function (second derivative is positive). If X is a random variable with mean μ, then

$$g(\mu) \leq \mathbb{E}(g(X)) \tag{17}$$

that is

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X))$$

3.6 DEFINITION. (Kullback-Leibler Divergence/ Distance, not completely correct) The Kullback-Leibler divergence between two (multivariate) density distribution $f_1(x)$ and $f_2(x)$ is defined as

$$\delta(f_1, f_2) = \mathbb{E}_{f_1}\left\{\log\left(\frac{f_1(\mathbf{X})}{f_2(\mathbf{X})}\right)\right\} = \int \ldots \int f_1(\mathbf{x}) \log\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) \tag{18}$$

Notice the δ is not a symmetric operator. Also if $f_2(\mathbf{x}) = 0$ on a set A with $\mathbb{P} > 0$ then the distance becomes $\delta(f_1, f_2) = \infty$. The divergence can be thought to be a measure of difference or distance of two distribution.

3.7 FACT. $\delta(f_1, f_2) \geq 0$

Proof.

$$\mathbb{E}_{f_1}\left\{\log\left(\frac{f_1(\mathbf{X})}{f_2(\mathbf{X})}\right)\right\} = \mathbb{E}_{f_1}\left\{-\log\left(\frac{f_2(\mathbf{X})}{f_1(\mathbf{X})}\right)\right\}$$

$$\underbrace{\geq}_{\text{Jensen Inequality}} \quad -\log\left\{\mathbb{E}_{f_1}\left(\frac{f_2(\mathbf{X})}{f_1(\mathbf{X})}\right)\right\}$$

notice $\log(\cdot)$ is a convex function

$$= -\log\left(\int \ldots \int f_2(\mathbf{x}) d\mathbf{x}\right) = -\log(1) = 0$$

□

Finally we observed that **mutual information is simply the Kullback-Leibler divergence between**

$$f_1(\mathbf{x}) = f(\mathbf{x}), \quad f_2(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i) \tag{19}$$

which is

$$\delta(f_1, f_2) = \int \ldots \int f(\mathbf{x}) \log\left[\frac{f(\mathbf{x})}{\prod_{i=1}^{d} f_i(x_i)}\right] \mathbf{dx}$$

$$= \int \ldots \int f(\mathbf{x}) \log(f(\mathbf{x})) \mathbf{dx} - \int \ldots \int f(\mathbf{x}) \log\left(\prod_{i=1}^{d} f_i(x_i)\right) \mathbf{dx}$$

$$= -\mathrm{H}(\mathbf{X}) - \sum_{i=1}^{d} \underbrace{\int f_i(x_i) \log\Big(f_i(x_i)\Big) dx_i}_{-\mathrm{H}(\mathrm{X}_i)}$$

$$= \sum_{i=1}^{d} \mathrm{H}(\mathrm{X}_i) - \mathrm{H}(\mathbf{X}) \geq 0$$

the non-negativity of mutual information is by the property of Kullback-Leibler Divergence.

Exercises: Let $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *be multivariate normal random vectors with mean* $\boldsymbol{m}_1$ *and* $\boldsymbol{m}_2$ *and covariance* $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *respectively. Calculate the Kullback-Leibler distane between* $\mathbf{X}_1$ *and* $\mathbf{X}_2$.

4 MULTIVARIATE NORMAL DISTRIBUTION

Notation introduction first. Let

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{1}) \iff \mathrm{Z}_1, \mathrm{Z}_2, ..., \mathrm{Z}_p \text{ are independent } \mathrm{N}(0, 1)$$

Notice the covariance matrix is the unit diagonal matrix. since we are assuming the covariance of those components to be independent. Recall normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

then joint density of $\mathbf{Z}$ is

$$f(z) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|\mathbf{Z}\|^2} = (2\pi)^{-\frac{d}{2}} e^{\xi'\xi}$$

last one use the inner products of two random vector. The level curve of the joint density of standard multinomormal are all circles.

4.1 *Generalized Multivariate Normal*

4.1 DEFINITION. Let $\mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_p)'$ and the mean for each components is $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)$ and the covariance matrix is

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1p} \\ a_{21} & a_{22} & ... & a_{2p} \\ ... & ... & ... & ... \\ a_{p1} & a_{p2} & ... & a_{pp} \end{pmatrix}$$

where the matrix is assumed to be of full rank. We call the vector $\mathbf{X}$ the multivariate normal random vector which is

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff \mathbf{X} = \mathrm{A}\mathbf{Z} + \boldsymbol{\mu}$$

We will assume that the $\Sigma = \mathbf{A}\mathbf{A}' > 0$ (positive definite) and $rank(\mathrm{A}_{p\times p}) = \mathrm{p}$

4.2 FACT. $\Sigma = \mathbf{A}\mathbf{A}' > 0$

Proof.

$$\Sigma = Cov(\mathbf{X}) = Cov(\mathbf{A}\mathbf{Z} + \boldsymbol{\mu}) = Cov(\mathbf{A}\mathbf{Z}) = \mathbf{A}\mathbf{A}'$$

□

where we say **A**is a square root of Σ. Fact 4.2 shows that given **A** we are able to find the covariance matrix of **X**. The following fact shows that if we are given Σ then how can we find **A**.

4.3 DEFINITION. (Positive Definite) We list several equivalent statement here

- Given $\Sigma \in \mathrm{M}(\mathrm{N}\times\mathrm{N},\mathbb{R})$. If $\forall x \neq 0$ which is N dimensional, we have $x'\Sigma x > 0$, then Σ is positive definite.
- Σ is positive definite if it is symmetric and all its eigenvalues are positive.

4.4 FACT. (Spectral Decomposition) Let Σ be $\mathrm{N}\times\mathrm{N}$ matrix with full rank N. Then we have

$$\Sigma = \sum_{i=1}^{\mathrm{N}} \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i'$$

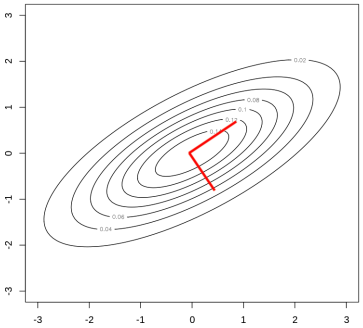
where we have λ_i to be the eigenvalues in ascending order with corresponding eigenvectors $\boldsymbol{a}_i$. Notices that those $\boldsymbol{a}_i$ are all orthonormal.

4.5 FACT. Given Σ which is of full rank with same notation in fact 4.4, we have

$$\mathbf{A} = \sum_{i=1}^{\mathrm{N}} \sqrt{\lambda_i}\, \boldsymbol{a}_i \boldsymbol{a}_i'$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{\mathrm{N}} > 0$ and **A** is the **unique symmetric square root of** Σ**.**

It is easy to check that fact(4.4) matches fact(4.3)



The plot above shows the contour (level) plot for bi variate normal random variables. The vector of Σ is the red arrows (which are unit vectors and orthonormal) and the eigenvalues determines the size of the ellipses. The multinational distribution is used as a modeling tool. We usually chose the eignvectors and eignvalue to build this distribution.

4.6 DEFINITION. (Multivariate Normal Density) let $\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}$ where $\mathbf{Z}$ s the standard multivariate normal. Then the PDF of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\boldsymbol{x}) = |\det(\mathbf{A})^{-1}|(2\pi)^{-p/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'(\mathbf{A}^{-1})'\mathbf{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{20}$$

$$= \det(\Sigma)^{-1/2}(2\pi)^{-p/2} \exp\left(-\frac{1}{2}\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}_{\text{Mahalanobis Distance}}\right)$$

Proof. make up later □

4.7 DEFINITION. (Mahalanobis Distance) The commonly used distance in statistics is not the Euclidean distance. The Mahalanobis distance is

$$d_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{\mu}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})'\,\Sigma\,(\boldsymbol{x}-\boldsymbol{\mu})} \tag{21}$$

The Mahalanobis distance can be thought of the version of euclidean distance after changing the basis of the random vector by using the new basis of eigenvector of Σ. How to see this in the expression (21)?

4.2 *Properties of Multivariate Normal Distribution*

Summary of the important properties:

- **Linear combination of X are normal (Jointly Normal random variables)**
- **Normal Marginals (subsets of X are normal)**
- **Cov(X, Y) = 0 iff X and Y are independent**
- **Normal Conditional Distribution. More importantly, $\mathbb{E}(\mathbf{Y}\,|\,\mathbf{X})$ is linear in X**

4.8 FACT. The MGF of a normal random variable is

$$\mathrm{M}_{\mathrm{X}}(t) = e^{t\mu + (t^2/2\sigma^2)}$$

so for standard normal we have

$$\mathrm{M}_{\mathrm{Z}}(t) = e^{t^2/2}$$

4.9 THEOREM. *(MGF of standard Normal random vector) Let* $\mathbf{Z}$ *ne the standard normal random vector. Then*

$$M_{\mathbf{Z}}(\boldsymbol{t}) = \exp(\frac{1}{2}\boldsymbol{t}'\boldsymbol{t}) \tag{22}$$

Notice the way we define the MGF of a random vector is by

$$M_{\mathbf{X}}(\boldsymbol{t}) = \mathbb{E}\Big(\exp(\mathbf{X}\boldsymbol{t})\Big)$$

where the product is the inner product.

4.10 THEOREM. *(MGF of Multivariate Noraml) Still let* $\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$. *Then*

$$M_{\mathbf{X}} = \exp\left(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}\right) \tag{23}$$

Proof. Exercises. Make up later. □

4.11 FACT. Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff \boldsymbol{a}'\mathbf{X} \sim \mathrm{N}(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\Sigma\boldsymbol{a})$$

for all $\boldsymbol{a} \neq 0$.

Proof. Using the moment generation function.

" $\Longrightarrow$ ": Assume the RHS. Let $\boldsymbol{a} \in \mathbb{R}^p$. Then

$$M_{\boldsymbol{a}'\mathbf{X}}(t) = \mathbb{E}\Big(\exp\Big\{\boldsymbol{a}'\mathbf{X}t\Big\}\Big) = \mathbb{E}\Big(t(\boldsymbol{a}'\mathbf{X})\Big) = M_{\mathbf{X}}(t\boldsymbol{a}')$$

Then by out assumption we have multinormal distribution of $\mathbf{X}$. Then by theorem 4.10 we have

$$= \mathbb{E}\Big(\exp\Big\{t\boldsymbol{a}'\boldsymbol{\mu} + \frac{1}{2}t^2\boldsymbol{a}'\boldsymbol{\mu}\boldsymbol{a}\Big\}\Big)$$

this is the moment generation function the r.v $\boldsymbol{a}'\mathbf{X}$.

" $\Longleftarrow$ ": Make up □

Fact 4.11 says that the linear combination of random variables is sitll normal. The conclusion is heard before while this time using the covariance matrix. The next fact extend this one which indicates the joint normality.

4.12 THEOREM. *If* $\forall t \in (-\epsilon, \epsilon)$ *for some* $\epsilon > 0$ *s.t* $M_X(t) = M_Y(t)$, *then* X *and* Y *have the same distribution. Or we can say this is the uniqueness of MGF.*

4.13 FACT. Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma), \mathbf{X} \in \mathbb{R}^p$. Let $\mathbf{A} \in \mathrm{M}(q \times p, \mathbb{R})$ with full rank, $q \leq p$. Then let $\mathbf{Y} = \mathbf{A}\mathbf{X}$ we have

$$\mathbf{Y} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\Sigma\mathbf{A}')$$

Proof. make up □

4.14 FACT. (Marginal Distribution) Let $\mathbf{X} \in \mathbb{R}^p$, a normal random vector. Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{1,q\times 1} \\ \mathbf{X}_{2,(p-q)\times 1} \end{pmatrix} \sim \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

then we have the fact that

$$\mathbf{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \Sigma_{11})$$

Proof. Make up □

4.15 FACT. Let the notation be the same as fact(4.13) which is let

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{1,q\times 1} \\ \mathbf{X}_{2,(p-q)\times 1} \end{pmatrix} \sim \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

Then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if

$$\Sigma_{12} = Cov(\mathbf{X}_1, \mathbf{X}_2) = 0$$

Proof. Make up □

4.16 DEFINITION. (Conditional distribution) Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{1,q\times 1} \\ \mathbf{X}_{2,(p-q)\times 1} \end{pmatrix} \sim \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

as before. Define

$$\widehat{\mathbf{X}}_1 = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \mu_2)$$

and

$$\boldsymbol{e} = \mathbf{X}_1 - \widehat{\mathbf{X}}_1 = (\mathbf{X}_1 - \boldsymbol{\mu}_1) - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \mu_2)$$

then we define the conditional distribution of the first partition by given the second partition.

4.17 FACT. By definition of 4.16 we can derive the folloing results

- $\boldsymbol{e}$ and $\mathbf{X}_2$ are independent
- $\boldsymbol{e}$ and $\widehat{\mathbf{X}}_1$ are independent
- $\boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$
- $\widehat{\mathbf{X}}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$
- $\mathbf{X}_1 \,|\, (\mathbf{X}_2 = \boldsymbol{x_2}) \sim \mathrm{N}(\widehat{\boldsymbol{x}}_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$
- $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ where $\Sigma^{1/2}$ is the unique symmetric square root of Σ (i.e $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$). Divided by standard error for standardiza-

tion. Also notice the sign of 1/2 indicates the sign of power of eigenvalues in the spectral decomposition.

Notice the last result $\widehat{\boldsymbol{x}}_1$ is the version with a given $\boldsymbol{x}_2$ defined in 4.16.

Proof. make up □

5 LINEAR MODEL AND LEAST SQUARE

5.1 *Linear Regression Model*

Notations: Let training data from $(y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$ where $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^p$. The linear model becomes

$$y = \underbrace{g(\mathbf{x}, \boldsymbol{\theta})}_{\text{Signal}} + \underbrace{\sigma\epsilon}_{\text{Noise}} \tag{24}$$

where $\boldsymbol{\theta}$ is the vector of parameters. $\sigma > 0$ to be the dispersion parameter. $\epsilon \sim N(0, 1)$ in the theory. Notice the ϵ is random variable collecting two part of noise: one is the **model noise** (i.e the signal that we model is not exactly the signal) and the other is the **observation noise**. Goals for the model is

- **To describe/understand the probabilistic mechanism that generates observations similar to those in the training data.**
- **To predict future values of** y.

1. **Model Assumptions:** Basic assumptions:

   - The model is
   $$y_i = g(\boldsymbol{x}_i, \boldsymbol{\theta}) + \sigma\epsilon_i \tag{25}$$

   The compact matrix form is $\boldsymbol{y} = \mathbf{Z}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$. Here the ϵ is the **model error** which is random and **non-observable**. It is **not the same as residuals**.

   - $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are random variables satisfying
   $$\begin{aligned} \mathbb{E}(\epsilon_i \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x_n}) &= 0 \\ Var(\epsilon_i \,|\, \boldsymbol{x}_1, \boldsymbol{x_2}, ..., \boldsymbol{x}_n) &= 1 \end{aligned} \tag{26}$$
   for all $i = 1, 2, ..., n$.
   - The ϵ_i are uncorrelated
   $$Cov(\epsilon_i, \epsilon_j \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x_n}) = 0 \tag{27}$$

   This can be turned in to more compact matrix form

$$\mathbb{E}(\boldsymbol{\epsilon} \mid \mathrm{X}) = 0 \quad Cov(\boldsymbol{\epsilon} \mid \mathrm{X}) = \mathbf{I} \tag{28}$$

and the stronger assumption below can be $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$

So the basic assumptions consists of the **first and second moments assumptions.** The stronger assumption is that $\epsilon_1, ..., \epsilon_n$ are all iid and follows $\mathrm{N}(0, 1)$, and $(\epsilon_1, \epsilon_2, ..., \epsilon_n)$, $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ are independent.

2. **Data Construction**: The data, information matrix, consists each row as a case of observation which is

$$\mathbf{D} = \begin{pmatrix} y_1 & x_{11} & x_{12} & x_{13} & ... & x_{1p} \\ y_2 & x_{21} & x_{22} & x_{23} & ... & x_{2p} \\ ... & ... & ... & ... & ... & ... \\ y_n & x_{n1} & x_{n2} & x_{n3} & ... & x_{np} \end{pmatrix}$$

Then we introduce the design matrix which is

$$\mathbf{Z} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & ... & x_{2p} \\ ... & ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & x_{n3} & ... & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_1' \\ \mathbf{Z}_2' \\ \mathbf{Z}_3' \\ ... \\ \mathbf{Z}_n' \end{pmatrix} = (\mathbf{Z}^0, \mathbf{Z}^1, ..., \mathbf{Z}^p)$$

where we use the subscript to represent the row and superscript for column. The covariates matrix is similar but without the first column of 1. The design matrix is simply designed to match up with the regression model with the constant term. We assume the matrix $\mathbf{Z}$ always has full rank and $rank(\mathbf{Z}) = p + 1 < n$. Link this with econometric of identification problem in a similar way that we need at least k equation to specify k parameter.

3. **Cases for the choice of** $g(\boldsymbol{x}_i, \boldsymbol{\theta})$:

   - **Location Model:** This is by chosen the $\boldsymbol{\theta} \in \mathbb{R}$. This leads to the location model $g(\boldsymbol{x}_i, \mu) = \mu$ for all $i = 1, 2, ..., n$.
   - **Linear regression model**: Let the $\boldsymbol{\beta}$ to be

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}$$

   where $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta}_1 \in \mathbb{R}^p$. Then our linear model becomes

$$g(\boldsymbol{x}_i, \beta_0, \boldsymbol{\beta}_1) = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}_i \quad i = 1, 2, ..., n \tag{29}$$

   where the $\boldsymbol{\theta}$ is interpreted as regression coefficient. Notice linear model means **linear in regression parameters** but NOT necessarily in explanatory variables.

4. **Regression Residuals**: No more random and it's observable. They are computed given the training data which is

$$e_i(\boldsymbol{t}) = y_i - g(\boldsymbol{x}_i, \boldsymbol{t}) \tag{30}$$

where $\boldsymbol{t}$ is the given parameters (not the true one). Then the sum of square residuals becomes

$$\mathrm{S}(\boldsymbol{t}) = \sum_{i=1}^{n} \Big(y_i - g(\boldsymbol{x}_i, \boldsymbol{t})\Big)^2 = \sum_{i=1}^{n} e_i^2(\boldsymbol{t}) \tag{31}$$

Notice the above equation is convex (i.e sum of convex function is convex) so taking derivatives get the minimum. In case of the linear regression model, we can show that

5.1 FACT. The OLS estimate for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{y}$$

The prove is by showing the first derivative of $\mathrm{S}(\boldsymbol{t}) = 0$ generate the estimate. Lecture notes do this by checking the given estimator satisfies

$$\mathrm{S}(\boldsymbol{b}) \geq \mathrm{S}(\widehat{\boldsymbol{\beta}}) \quad \forall \boldsymbol{b} \in \mathbb{R}^{p+1} \tag{32}$$

with equality **if and only if** $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}$. The detailed proof see documents Least Squares page 15.

5. **Property of LS Estimate**:

5.2 FACT. (Unbiasness) Suppose that

$$\mathbb{E}(\boldsymbol{\epsilon} \,|\, \mathrm{X}) = \mathbf{0}$$

then the $\widehat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$:

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} \tag{33}$$

for all $\boldsymbol{\beta}$

Proof. The notes in lecture confusing estimate and estimator. Assume

the first moment of $\boldsymbol{\epsilon}$ is zero. Then

$$\begin{aligned}
\mathbb{E}(\widehat{\boldsymbol{\beta}}) &= \mathbb{E}(\mathbb{E}[\widehat{\boldsymbol{\beta}} \mid X]) = \mathbb{E}\left(\mathbb{E}\left[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \mid X\right]\right) \\
&= \mathbb{E}\left((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathrm{E}[\mathbf{y} \mid X]\right) \\
&= \mathbb{E}\left((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbb{E}[\mathbf{Z}\beta + \sigma\boldsymbol{\epsilon} \mid X]\right) \\
&= \mathbb{E}\left((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' * (\mathbf{Z}\beta + \sigma\mathbb{E}[\boldsymbol{\epsilon} \mid X])\right) \\
&= \mathbb{E}\left((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}\boldsymbol{\beta}\right) = \mathbb{E}(\boldsymbol{\beta}) = \boldsymbol{\beta}
\end{aligned}$$

□

5.3 FACT. (Covariance of LS estimate) Suppose that $\mathbb{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$ and $Cov(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{I}$. Then

$$Cov(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1} \tag{34}$$

Notice that the conclusion is w.r.p to $\widehat{\boldsymbol{\beta}}$ **conditioning on X**. The $Cov(\widehat{\boldsymbol{\beta}})$ is too tricky so not going to explore.

Proof. Makeup. Page 27 & 28. The proof contains some valuable manipulation of matrix operation when dealing with covariance.

$$\begin{aligned}
\mathrm{Cov}[\widehat{\beta} \mid X] &= \mathrm{Cov}\left[(Z'Z)^{-1}Z'y \mid X\right] \\
&= (Z'Z)^{-1}Z'\,\mathrm{Cov}[y \mid X]Z\,(Z'Z)^{-1} \\
&= (Z'Z)^{-1}Z'\,\mathrm{Cov}[Z\beta + \sigma\varepsilon \mid X]Z\,(Z'Z)^{-1} \\
&= (Z'Z)^{-1}Z'\sigma^2\overbrace{\mathrm{Cov}[\varepsilon \mid X]}^{=1}Z\,(Z'Z)^{-1} \\
&= \sigma^2\,(Z'Z)^{-1}Z'Z\,(Z'Z)^{-1} \\
&= \sigma^2\,(Z'Z)^{-1}
\end{aligned}$$

The font not adjusted yet □

6. Gauss-Markov Theorem

5.4 THEOREM. *Suppose the* $\mathbb{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$ *and* $Cov(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{I}$ *(no distribution is assumed). Suppose the* $\widetilde{\boldsymbol{\beta}} = \mathbf{L}\boldsymbol{y}$*, where* $\mathbf{L} \in Matrix((p+1) * n, \mathbb{R})$*, to be another linear unbiased estimator of* $\boldsymbol{\beta}$ *(i.e that is* $\mathbb{E}(\widetilde{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta}$*). Then*

$$\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \tag{35}$$

is ***non-negative definite (i.e*** ≥ 0*). Therefore,* $\widehat{\boldsymbol{\beta}}$ *is the BLUE (Best Linear Unbiased Estimate for* $\boldsymbol{\beta}$*).*

The theorem is saying that comparing the linear unbiased estimator is worse than the LS estimate in terms of variability.

Proof. Page 30 and 31. Make up □

7. **Fitted Values**

After obtaining the LS estimate of the regression coefficient, next step is to make prediction. $\boldsymbol{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and with given assumption we have $\mathbb{E}(\boldsymbol{y} \mid \mathbf{X}) = \mathbf{Z}\boldsymbol{\beta}$. We want to use the expectation as an estimator of y given x. So

$$\begin{aligned}\widehat{\boldsymbol{y}} = \widehat{\mathbb{E}}(\boldsymbol{y} \mid \mathbf{X}) &= \mathbf{Z}\widehat{\boldsymbol{\beta}} \\ &= \underbrace{\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'}_{\text{H, Hat matrix}}\boldsymbol{y} = \mathrm{H}\boldsymbol{y} \text{ (Fitted Values)}\end{aligned}$$

The hat matrix has very nice properties

- **Symmetric** very easy to check
- **Idempotent** This means HH = H. So H to any power is still H. Easy to check.
- **H is a projection matrix** H projects $\boldsymbol{y}$ on space V which is the column space of Z. In particular, HZ = Z. The detailed discussion is at 47:13, 10-06

8. **Residuals** Residual comes from the difference between fitted values and the observed real values which is

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \mathrm{H}\boldsymbol{y} = (\mathrm{I} - \mathrm{H})\boldsymbol{y}$$

Then we have

$$\mathbb{E}(\boldsymbol{e}) = \mathbb{E}\Big((\mathrm{I} - \mathrm{H})\boldsymbol{y}\Big) = \Big(\mathrm{I} - \mathrm{H}\Big)\mathbb{E}(\boldsymbol{y}) = \Big(\mathrm{I} - \mathrm{H}\Big)\mathbf{Z}\boldsymbol{\beta} = \mathbf{0}$$

and $Cov(\boldsymbol{e}) = \sigma^2(\mathrm{I} - \mathrm{H})$ by a similar means. Here are some remaining properites to be checked by yourself

- $\widehat{\boldsymbol{y}}' \neq= 0$ Fitted values and residuals are orthogonal
- $\mathbb{E}(\boldsymbol{e}'\boldsymbol{e}) = (n - p - 1)\sigma^2$ there fore

$$\widehat{\sigma^2} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n - p - 1}$$

is an unbiased estimator of σ^2.exercises

Notice the total sum of square (residuals) in matrix form is

$$\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = \mathbf{Y}'\mathbf{Y}$$

6 DISTRIBUTION OF NORMAL QUADRATIC FORM

Review of joint normal and Chi-square distribution. All the detail about multi normal is in the previous section. Some notations

- Σ^{-1} is called the *precision matrix* and the $\Sigma^{-1/2}$ is the square root of the precision matrix. This is important for doing sparse estimation of Σ^{-1} in geographical model. This is

$$\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2} = \Big(\sum_{i=1}^{p} \lambda_i^{-1/2}\boldsymbol{a}_i\boldsymbol{a}_i'\Big)\Big(\sum_{j=1}^{p} \lambda_j^{-1/2}\boldsymbol{a}_j\boldsymbol{a}_j'\Big)$$

$$= \sum_{m=1}^{p} \lambda_m^{-1}\boldsymbol{a}_m\boldsymbol{a}_m'$$

  notice usually eigenvalues of a matrix are NOT orthogonal while for symmetric matrix the eigenvectors are orthogonal (i.e the covaiance matrix here). Also check $\Sigma\Sigma^{-1} = \mathbf{I}$. So this is a unique symmetric square root for Σ. Square root of a matrix is never unique while if further more we want a symmetric root this becomes unique.

6.1 FACT. Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ defined as before as a p dimensional random vector. Then we obtain that

$$\mathbf{Q} = (\mathbf{X} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \underbrace{(\mathbf{X} - \boldsymbol{\mu})'\Sigma^{-1/2}}_{\mathbf{Z}'}\underbrace{\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})}_{\mathbf{Z}} \tag{36}$$

where $\mathbf{Z}$ is the standard joint normal random variables. Then

$$\mathbf{Q} \sim \chi^2_{(p)}$$

which means $\sum_{i=1}^{p} Z_i^2$ follows a Chi-square distribution.

6.2 FACT. Follow the definition in 6.1, now consider if $\mathbf{X}$ is not centered, then the case becomes

$$\mathbf{W} = \mathbf{X}'\Sigma^{-1}\mathbf{X} \sim \chi^2_{(p)}(\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu})$$

where we define

$$\gamma = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$$

to be the non-centrality parameter.

6.3 FACT. (Reproductive Property) If $W_1 \sim \chi^2_{(p_1)}(\lambda_1)$ and $W_2 \sim \chi^2_{(p_2)}(\lambda_2)$ are independent then

$$W_1 + W_2 \sim \chi^2_{(p_1+p_2)}(\lambda_1 + \lambda_2)$$

this can be verified by MFG.

6.1 *The Fisher-Cochran Theorem*

In short this theorem study the partition of sum of square of quadratic form of normal real distribution.

6.4 LEMMA. *Given the quadratic form* $\mathbf{Q} = \boldsymbol{x}'\mathrm{A}\boldsymbol{x}$*, it can also be written as*

$$\mathbf{Q} = \boldsymbol{x}' \underbrace{\left(\frac{1}{2}\Big(\mathrm{A} + \mathrm{A}'\Big)\right)}_{\textbf{\textit{Symmetric matrix}}} \boldsymbol{x}$$

that is, we can assume, wlg, that the matrix A is symmetric.

Proof. Since $\mathbf{Q} \in \mathrm{R}$ we have that $\mathbf{Q}' = \mathbf{Q}$. Therefore

$$\begin{aligned} \mathbf{Q} &= \mathbf{Q}' = x'\mathrm{A}'x \\ \mathbf{Q} &= x'\mathrm{A}x \end{aligned}$$

So

$$\begin{aligned} \mathbf{Q} &= \frac{1}{2}\left(\mathbf{Q}' + \mathbf{Q}\right) = \frac{1}{2}\left(x'\mathrm{A}'x + x'\mathrm{A}x\right) \\ &= x'\left(\frac{1}{2}\left(\mathrm{A} + \mathrm{A}'\right)\right)x \end{aligned}$$

□

6.5 LEMMA. *Suppose that*

$$\mathbf{Q} = x'\mathrm{A}'x \quad \textbf{\textit{rank}}(\mathrm{A}) = q$$

where $\boldsymbol{x} \in \mathbb{R}^p$*, then there are q linearly independent linear combinations*

$$y_i = \boldsymbol{b}_i'\boldsymbol{x} = \sum_{j=1}^{p} b_{ij}x_j, \quad i = 1, 2, ..., q$$

and

$$\mathbf{Q} = \sum_{i=1}^{q} \delta_i y_i^2, \quad \delta_i^2 = 1$$

that is $\delta_i = \pm 1$*. Also notice if* $\mathrm{A} \geq 0$ *then*

$$\mathbf{Q} = \sum_{i=1}^{q} y_i^2$$

A more compact notation for lemma6.5 is , there exists a matrix $\mathbf{B}$ which is

$$\mathrm{B}_{q\times p} = \begin{pmatrix} \boldsymbol{b}_1' \\ \boldsymbol{b}_2' \\ . \\ . \\ \boldsymbol{b}_q' \end{pmatrix}, rank(\mathbf{B}) = q$$

such that

$$\begin{aligned} \boldsymbol{y} &= \mathbf{B}\boldsymbol{x} \\ \mathbf{Q} &= \boldsymbol{x}'\mathrm{A}\boldsymbol{x} = \boldsymbol{y}'\Delta\boldsymbol{y} \end{aligned} \tag{37}$$

with $\Delta = diag(\delta_1, ..., \delta_q)$ and $\delta_i^2 = 1$, $i = 1, 2, ..., q$. Moreover by (37) we can write

$$\boldsymbol{x}'\mathrm{A}\boldsymbol{x} = \boldsymbol{x}'\mathrm{B}'\Delta\mathrm{B}\boldsymbol{x} \quad \forall \boldsymbol{x}$$

therefore

$$\mathrm{A} = \mathrm{B}'\Delta\mathrm{B}$$

Proof is omitted. See page 18 of ppt

6.6 THEOREM. *(Fisher-Cochran Theorem) Suppose* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$ *in* $\mathbb{R}^p$*, and that*

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{k} \mathbf{X}'\mathrm{A}_i\mathbf{X}, \quad \operatorname{rank}(\mathrm{A}_i) = q_i, \quad i = 1, \ldots, k$$

Set

$$\mathrm{Q}_i = \mathbf{X}'\mathrm{A}_i\mathbf{X}$$

then the following two statement are equivalent:

- $\mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{Q}_k$ *are independent with*

$$\mathrm{Q}_i \sim \chi^2_{(q_i)}(\boldsymbol{\mu}'\mathrm{A}_i\boldsymbol{\mu}), \quad i = 1, \ldots, k$$

- $\sum_{i=1}^{k} q_i = p.$

Recall the definition of non-centered chi-square distribution.

6.7 DEFINITION. Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I}) \in \mathbb{R}^p$ and $\mathrm{A}_{p\times p}$. Then

$$\mathbf{X}'\mathrm{A}\mathbf{X} \sim \chi^2_p(\boldsymbol{\mu}'\mathrm{A}\boldsymbol{\mu})$$

which is the non centered chi squared distribution. Proof on page 22 pdf.

6.2 *Application of Fisher-Cochran Theorem: The Normal Location-Scale Model*

This mode is useful for independent measurement with errors. The model is

$$Y_i = \mu + \sigma\epsilon_i$$

where $\epsilon_.$ are iid $N(0,1)$. Turn into matrix form we have

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \\ \dots \\ \mu \end{pmatrix}, \sigma^2\mathbf{I} \right) = N(\mu\mathbf{1}, \sigma^2\mathbf{I}) \tag{38}$$

and standardize form

$$\frac{1}{\sigma}\mathbf{Y} \sim N\left(\frac{\mu}{\sigma}\mathbf{1}, \mathbf{I}\right), \mathbb{E}(\mathbf{Y}) = \frac{\mu}{\sigma}\mathbf{1}$$

The linear model becomes

$$\mathbf{Y} = \mathbf{1}\mu + \sigma\boldsymbol{\epsilon}$$

where the design matrix in this case is just $\mathbf{1}$. The recall the LS estimator

$$\widehat{\boldsymbol{\beta}} = \widehat{\mu} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{Y} = \frac{1}{n}\sum Y_i = \overline{Y}$$

and the hat matrix here is

$$H = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \frac{1}{n}\mathbf{1}\mathbf{1}'$$

so the signal + noise decomposition is

$$\mathbf{Y} = H\mathbf{Y} + (1 - H)\mathbf{Y} = \widehat{\mathbf{Y}} + \boldsymbol{e}$$

The hypothesis usually of interested is $\mu =, \neq \mu_0$. The decomposition of sum of square residual is

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'H\mathbf{Y} + \mathbf{Y}'(I - H)\mathbf{Y}$$

which is a general result which is a sum of two quadratic form. Here we apply the Fisher-Cochran's theorem with

$$A_1 = H = \frac{1}{n}\mathbf{1}\mathbf{1}' \qquad A_2 = I_n - H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

In summary

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'A_1\mathbf{Y} + \mathbf{Y}'(I_n - A_1)\mathbf{Y}$$

Notice we need to check $\text{Cov}(\mathbf{Y}) \neq \mathbf{I}$ and $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

6.8 FACT. if B is a projection matrix then rank(B) = trace(B)

In order to be a projection matrix, the matrix need to satisfy **Symmetry** and **Idenpotence** (i.e $B^n = B$). Prove page28 slides Then

$$\text{rank}(A_1) = \text{tr}(A_1) = 1$$

$$\text{rank}(A_2) = \text{tr}(A_2) = \text{tr}(I_n)\text{tr}(A_1)$$

Therefore by F – C theorem

$$Q_1 = \frac{\mathbf{Y}'H\mathbf{Y}}{\sigma^2} \quad Q_2 = \frac{\mathbf{Y}'(I_n - H)\mathbf{Y}}{\sigma^2}$$

are independent and

$$Q_1 \sim \chi^2_{(1)}(\frac{\mu^2}{\sigma^2}\mathbf{1}'H\mathbf{1})$$

$$Q_2 \sim \chi^2_{(n-1)}(\frac{\mu^2}{\sigma^2}\mathbf{1}'(I_n - H)\mathbf{1})$$

Notice

$$\frac{\mu^2}{\sigma^2}1'H1 = \frac{\mu^2}{\sigma^2}1'\left(\frac{1}{n}11'\right)1 = \underbrace{n\frac{\mu^2}{\sigma^2}}_{\text{Noncenter Parameter}}$$

$$\frac{\mu^2}{\sigma^2}1'(I_n - H)1 = \frac{\mu^2}{\sigma^2}(1'1 + 1'H1)$$

$$= \frac{\mu^2}{\sigma^2}\left(1'1 + 1'\left(\frac{1}{n}11'\right)1\right) = 0$$

The signal to noise ratio is zero for all μ all parameter to be centered. The ratio of two chi square distribution is F distribution. So in our case, Q_2 is always central while Q_1 is central under the null hypothesis. The quadratics form here is

$$Q_1 = \frac{Y'A_1Y}{\sigma^2} = \frac{1}{n}\frac{Y'11'Y}{\sigma^2} = \frac{n\bar{Y}^2}{\sigma^2} \sim \chi^2_1\left(n\frac{\mu^2}{\sigma^2}\right)$$

$$Q_2 = \frac{Y'(I_n - A_1)Y}{\sigma^2} = \frac{Y'(I_n - A_1)Y}{\sigma^2}$$

$$= \frac{\left(\sum_{i=1}^n Y_i^2\right) - n\bar{Y}^2}{\sigma^2} = \frac{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}{\sigma^2}$$

$$= \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}(0)$$

Test Statistics

Under null, $H_0 : \mu = 0$ the ratio is centered F distribution. So we use

$$\text{Signal} = \overline{Y}^2, \quad \text{Noise} = \frac{S^2}{n}$$

and its ratio as test statistics

$$F = \frac{n\overline{Y}^2}{S^2}$$

Actually the non-center F distribution is defined with a non-centred numerator chi-square

7 COMPARISON OF NESTED MODEL

We wish to compare two models:

- A full model, called Model 1 versus
- A parsimonious model, called Model 0
- We assume that Model 0 is a restriction (particular case) of Model 1

Compassion based on the ratio

$$F = \frac{\text{Signal}}{\text{Noise}}$$

where

- signal = RSS_0 - RSS_1 (difference of residual sum of squares)
- noise = RSS_1 (minimum residual sum of squares)

where $RRS_0 = \sum_{i=1}^{n} (e_i^0)^2$ similar for RSS_1 using model 1. Favour model 1 over model 0 if the F value is large and we need a reference range to decide whether F is in fact large, facilitated by F-C theorem.

7.1 *Example: Constant signal vs Pure error*

Lets consider two model:

$$\text{Model 0: Pure noise} \quad Y_i = \sigma\epsilon_i$$

$$\text{Model 1: Constant Signal} \quad Y_i = \mu + \sigma\epsilon_i$$

we assume ϵ_i are iid $N(0, 1)$. Use linear model here for estimation (least square). The model is

$$\mathbf{Y} = \mathbf{1}\mu + \sigma\boldsymbol{\epsilon}$$

and the design matrix $\mathbf{Z} = \mathbf{1}_{n\times 1}$ then we can work out the

$$\widehat{\mu} = \overline{Y}$$

and the hat matrix H_1 is a $(n \times n)$ with all entries $1/n$. The model residual

$$\boldsymbol{e}^1 = \mathbf{Y} - \widehat{\mathbf{Y}}^1 = (1 - H_1)\mathbf{Y}$$

and RRS become

$$\text{RRS}_1 = (\boldsymbol{e}^1)'\boldsymbol{e}^1 = \mathbf{Y}'(\mathbf{I} - \text{H}_1)\mathbf{Y}$$

using the idempotent of $\mathbf{Y}'(\mathbf{I} - \text{H}_1)$ For model 0, the $\text{RSS}_0 = \mathbf{Y}'\mathbf{Y}$. So our test becomes

- Signal = $\mathbf{Y}'\text{H}\mathbf{Y}$=$n\overline{\text{Y}}^2$
- Noise = $\mathbf{Y}'(\mathbf{I} - \text{H})\mathbf{Y}$=(n-1)$\text{S}^2$, S^2 is the sample variance
- Signal + Noise = $\mathbf{Y}'\mathbf{Y}$

So we clearly see the decomposition

$$\frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2} = \frac{\mathbf{Y}'(\text{I} - \text{H})\mathbf{Y}}{\sigma^2} + \frac{\mathbf{Y}'\text{H}\mathbf{Y}}{\sigma^2} = \text{Q}_1 + \text{Q}_2$$

Since tr((I – H)) = $n - 1$ and tr((H)) = 1, by the F-C Theorem we have

$$\text{Q}_1 \sim \chi^2_{(n-1)}\left(\frac{\mu^2}{\sigma^2}1'(\text{I} - \text{H})1\right) = \chi^2_{(n-1)}(0), \quad forall\mu$$

$$\text{Q}_2 \sim \chi^2_{(1)}\left(\frac{\mu^2}{\sigma^2}1'\text{H}1\right) = \chi^2_{(1)}\left(\frac{\mu^2}{\sigma^2}1'1\right) = \chi^2_{(1)}\left(\frac{\mu^2}{\sigma^2}n\right), \quad forall\mu$$

Q_1 and Q_2 are independent. Therefore

$$\text{F} = \frac{\text{Q}_2}{\text{Q}_1/(n-1)} = \frac{\mathbf{Y}'\text{H}\mathbf{Y}}{\mathbf{Y}'(\text{I} - \text{H})\mathbf{Y}/(n-1)} = \frac{n\overline{\text{Y}}^2}{\text{S}_n^2} \sim \text{F}_{(1,n-1)}\left(\frac{\mu^2}{\sigma^2}n\right) \text{ for all}\mu$$

which has a centered F distribution at null. This is why we say we need a reference range: So range fro comparison for F is range of F when pans. model holds. So null distribution is a central F distribution

$$\text{F}_{(1,n-1)}(0) = \text{F}_{(1,n-1)}$$

with rejection region

$$\frac{n\overline{\text{Y}}^2}{\text{S}_n^2} \geq \text{F}^{-1}_{1,n-1}(1-\alpha)$$

for some small α, usually 0.05 or 0.01.

7.2 *Example 2: Linear regression model vs Local scale Model*

Our model one is linear model

$$\text{Y}_i = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + \sigma\epsilon_i, \quad \mathbf{X}_i \in \mathbb{R}^p$$

with i.i.d N(0,1) ϵ_i. We assume the design matrix is of rank $p + 1 < n$. Recall the previous results, the $\text{RRS}_1 = \mathbf{Y}'(\text{I} - \text{H}_1)\mathbf{Y}$. It is of the same form as the constant signal model.

$$\text{Y}_i = \beta + \sigma\epsilon_i$$

where $\widehat{\beta_0} = \overline{Y}$ and $H_0 = (1/n)\mathbf{1}\mathbf{1}'$. Then

$$\text{Signal} = \text{RSS}_0 - \text{RSS}_1 = \mathbf{Y}'(I - H_0)\mathbf{Y} - \mathbf{Y}'(I - H_1)\mathbf{Y} = \mathbf{Y}'(H_1 - H_0)\mathbf{Y}$$

$$\text{Noise} = \text{RSS}_1 = \mathbf{Y}'(I - H_1)\mathbf{Y}$$

Then the decomposition can be

$$\text{Signal} + \text{Noise} = \mathbf{Y}'(H_1 - H_0)\mathbf{Y} + \mathbf{Y}'(I - H_1)\mathbf{Y} = \mathbf{Y}'\mathbf{Y} + \mathbf{Y}'H_0\mathbf{Y}$$

so the beauty becomes

$$\mathbf{Y}'\mathbf{Y} = \text{Signal} + \text{Noise} + \mathbf{Y}'H_0\mathbf{Y}$$

then we can apply F-C theorem to construct F test. This becomes

$$\frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2} = \frac{\mathbf{Y}'\overbrace{(H_1 - H_0)}^{A_1}\mathbf{Y}}{\sigma^2} + \frac{\mathbf{Y}'\overbrace{(I - H_1)}^{A_2}\mathbf{Y}}{\sigma^2} + \frac{\mathbf{Y}'\overbrace{H_0}^{A_3}\mathbf{Y}}{\sigma^2}$$

Notice the reason why divided by σ^2 is that the F-C theorem requires the covariance matrix of $\mathbf{Y}$ must by identity. All A_i are projection matrix. Easy to check the two properties. Then check the rank of As' by checking their trace (tr. is a linear operator). Therefore, by the F-C Theorem we have that

$$Q_1 = \frac{Y'(H_1 - H_0)Y}{\sigma^2}, \quad Q_2 = \frac{Y'(I - H_1)Y}{\sigma^2} \text{ and } \quad Q_3 = \frac{Y'H_0Y}{\sigma^2}$$

are independent

$$Q_1 \sim \chi^2_{(p)}\left[\beta' Z_1'(H_1 - H_0)Z_1\beta\right]$$

$$Q_2 \sim \chi^2_{[n-(p+1)]}\left[\beta' Z_1'(I - H_1)Z_1\beta\right] = \chi^2_{(n-p-1)}(0)$$

$$Q_3 \sim \chi^2_{(1)}(\beta' Z_1' H_0 Z_1 \beta)$$

so our test statistics need a non central chi-square and a ALWAYS centered numerator chi-square. That is

$$\frac{Q_1/p}{Q2/(n-p-1)} \sim F_{(p,n-p-1)}(\beta' Z_1'(H_1 - H_0)Z_1\beta)$$

and under null where $\beta_1 = 0$ F is central distributed.

8 EXTENDED FISHER - COCHRAN THEOREM

We star with a lemma.

8.1 LEMMA. *Suppose the rank(A) = m and let* $Q = (\mathbf{Y}'A\mathbf{Y})/\sigma^2$. *Then*

$$Q \sim \chi^2_m(\lambda)$$

if and only if $A^2 = A$.

Proof in screen shot Proof1.jpeg and Proof2.jpeg

8.2 LEMMA. *Let*

$$Q_1 = \frac{\mathbf{Y}'A_1\mathbf{Y}}{\sigma^2} \sim \chi^2_{(m_1)}(\lambda_1)$$

$$Q_2 = \frac{\mathbf{Y}'A_2\mathbf{Y}}{\sigma^2} \sim \chi^2_{(m_2)}(\lambda_2)$$

then Q_1 *and* Q_2 *are independent* ***iff*** $A_1A_2 = 0$.

Proof in slides pages 58 and 57

8.3 THEOREM. *(Further Characterization of F-C Theorem) Suppose that* $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ *and*

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{k} \mathbf{Y}'A_i\mathbf{Y}, \quad rank(A_i) = q_i, \quad i = 1, ..., k$$

then the following statements are equivalent:

- *The* $(\mathbf{Y}'A_i\mathbf{Y})/\sigma^2$ *are independent* $\chi^2_{(q_i)}(\lambda_i)$
- *The matrices* A_i *are idempotent*
- $A_iA_j = 0$ *for all* $i \neq j$

Proof in slides

9 MAXIMUM LIKELIHOOD

Similar idea as before. Notations here we will use: The population that a series of sample comes from has a common density

$$f(\boldsymbol{y}; \boldsymbol{\theta}) \equiv f(\boldsymbol{y}, \boldsymbol{\theta}) \equiv f(\boldsymbol{y} \mid \boldsymbol{\theta})$$

We assume the $\boldsymbol{\theta} \in \mathbb{R}^p$ and unknown. The range of possible values of $\boldsymbol{\theta}$ is the *parameter space* denoted by Θ. Notice the $\boldsymbol{y}_i$ can be either r.v or random vectors (i.e f is there joint density). Important: MLE is a frequenstist approach, that is, the parameters are unknown but not random.

Likelihood Function: It is the joint density function of the data at the observed values $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n$, but viewed as a function of parameters. This is

$$\mathrm{L}_n(\theta) = f_\theta(\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\theta})$$

and the **Maximum Likelihood Estimate**:

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\theta \in \Theta} \ \mathrm{L}_n(\boldsymbol{\theta})$$

this is usually approached by least square or sometimes non-trivial optimization. The log-likelihood function

$$l_n(\theta) = \frac{1}{n} \log\left[\mathrm{L}_n(\theta)\right] = \frac{1}{n} \sum_{i=1}^{n} \log\left(f\left(y_i; \theta\right)\right)$$

we divided by n for some average reason or we want to use LLN or CLT. Clearly the estimation result are the same.

9.1 EXAMPLE. Let $\mathrm{Y}_i (i = 1, .., n)$ be i.i.d Unif$(0, \theta)$, $\theta > 0$. Suppose

$$\max\{y_i\} = 1.5$$

If we simply apply the same idea before we will get the likelihood which is $1/\theta^n$ which does not utilize the sample information at all. So we should revise the method. So we use

$$\begin{aligned} \mathrm{L}_n(\theta) = \prod_{i=1}^{n} f\left(\mathrm{T}_i \mid \theta\right) &= \frac{1}{\theta^n} \mathrm{I}\left(\mathrm{Y}_1, \ldots, \mathrm{Y}_n \in [0, \theta]\right) \\ &= \frac{1}{\theta^n} \mathrm{I}\left(\max\left(\mathrm{Y}_1, \ldots, \mathrm{Y}_n\right) \leq \theta\right) \end{aligned}$$

where $\mathrm{I}(\cdot)$ is an indicator function of value 1 if sth inside the bracket happens or 0 otherwise provided that $\min\{y_i\} \geq 0$. Another way to say is this

$$\mathrm{L}_n(\theta) = 0 \, if \, \theta < \max\left(\mathrm{Y}_1, \ldots, \mathrm{Y}_n\right)$$

$$\mathrm{L}_n(\theta) = \frac{1}{\theta^n} if \, \theta \geq \max\left(\mathrm{Y}_1, \ldots, \mathrm{Y}_n\right)$$

The results will be $\widehat{\theta} = \max\{\mathrm{Y}_1, \ldots, \mathrm{Y}_n\}$.

This example although is a bit tricky but still has a closed from results. The following one is not even closed.

9.2 EXAMPLE. (Gamma distribution) See pdf of MLE Example 3. The final result involves numerical methods.

9.1 *The Information Inequality*

The Kulback-Leibler divergence between $f(\boldsymbol{y};\boldsymbol{\theta}_0)$, $f(\boldsymbol{y};\boldsymbol{\theta}_1)$ is defined as

$$\begin{aligned}\mathrm{K}\left(f\left(y;\theta_0\right),f\left(y;\theta_1\right)\right) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\log\left(\frac{f\left(y;\theta_0\right)}{f\left(y;\theta_1\right)}\right)f\left(y;\theta_0\right)dy \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\log\left(f\left(y;\theta_0\right)\right)-\log\left(fy;\theta_1\right)\right]f\left(y;\theta_0\right)dy\end{aligned}$$

If $f\left(y;\theta_1\right)=0$ on a set A with $\mathrm{P}_{\theta_0}(\mathrm{A})>0$, then $\mathrm{K}\left(f\left(y;\theta_0\right),f\left(y;\theta_1\right)\right)=\infty$. Also notice the $\mathrm{K}(\cdot,\cdot)$ is not a symmetric operator.

9.3 THEOREM. *(Information Inequality) Let*

$$f_i(y)=f(y;\theta_i),\ i=0,1$$

then

$$\mathrm{K}(f_0,f_1)\geq 0$$

with = iff $f_0=f_1$. proof on pdf page8

The connection between K-L distance and MLE is this. We restate theorem 9.3

$$\mathrm{E}_{\theta_0}\left\{\log\left(\frac{f\left(\mathrm{Y};\theta_0\right)}{f(\mathrm{Y};\theta)}\right)\right\}\geq 0,\qquad \text{for all }\theta$$

as follows

$$\mathrm{E}_{\theta_0}\left\{\log\left(f\left(\mathrm{Y};\theta_0\right)\right)-\log(f(\mathrm{Y};\theta))\right\}\geq 0,\qquad \text{for all }\theta$$

the following inequality is important

$$\mathrm{E}_{\theta_0}\left\{\log\left(f\left(\mathrm{Y};\theta_0\right)\right)\right\}\geq \mathrm{E}_{\theta_0}\{\log(f(\mathrm{Y};\theta))\},\qquad \text{for all }\theta\neq\theta_0$$

What we want is that the estimator to be consistent.

9.4 DEFINITION. A parametric model $\{\mathrm{P}_\theta\}$ is identifiable if $\theta_1\neq\theta_2 \implies \mathrm{P}_{\theta_1}\neq \mathrm{P}_{\theta_2}$

If the model is identifiable, the expected log-likelihood function

$$\mathrm{L}(\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{\theta}_0}\{\log(f(\mathbf{Y};\boldsymbol{\theta}))\} \tag{39}$$

This is exactly the population counterpart of the log likelihood and this is the reason that we divide the likelihood by 1/n. That is

$$l_n(\boldsymbol{\theta})\longrightarrow_{a.s} l(\boldsymbol{\theta})\ \text{ as }\ n\longrightarrow\infty$$

Page 10 in pdf model 7 should be added here.

9.5 THEOREM. *Theorem 3 on page 10*

9.2 *Score function and Fisher Information matrix*

The score function $\psi(y;\theta)$ is defined as the gradient of the log density function $\log[f(\boldsymbol{y};\theta)]$, for all $\theta \in \Theta$. That is,

$$\begin{aligned}\psi(\boldsymbol{y};\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \log[f(\boldsymbol{y};\boldsymbol{\theta})] \\ &= \begin{pmatrix} \frac{\partial}{\partial \theta_1}\log[f(y;\theta)] \\ \frac{\partial}{\partial \theta_2}\log[f(y;\boldsymbol{\theta})] \\ \vdots \\ \frac{\partial}{\partial \theta_p}\log[f(y;\boldsymbol{\theta})] \end{pmatrix} = \begin{pmatrix} \psi_1(y;\boldsymbol{\theta}) \\ \psi_2(y;\boldsymbol{\theta}) \\ \vdots \\ \psi_p(y;\boldsymbol{\theta}) \end{pmatrix} \quad \boldsymbol{\theta} \in \Theta\end{aligned}$$

The expected score function is an expectation based on the true parameter $\boldsymbol{\theta}_0$

$$\Psi(\boldsymbol{t}) = \mathbb{E}_{\theta_0}\{\Psi(\boldsymbol{y};\boldsymbol{t})\} = \int_{-\infty}^{\infty} \Psi(\boldsymbol{y};\boldsymbol{t}) f(\boldsymbol{y};\theta_0) d\boldsymbol{y}$$

where $\boldsymbol{t}$ is a particular possible values of $\boldsymbol{\theta}$. Finally under regualrity conditions including diffrentiability of $f(\boldsymbol{y};\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$ and the interchangeable of the order of differentiation we have

$$\Psi(\boldsymbol{\theta}_0) = 0$$

if the true value of parameter is $\boldsymbol{\theta}_0$. Watch the video 11-03 and proof in pdf.

MLE Equation

Given data $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$, the log-likelihood function is

$$\mathcal{L}_n(\boldsymbol{y};\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(f(\boldsymbol{y}_i;\boldsymbol{t})\right)$$

Then we take differentiation w.r.t $\boldsymbol{t}$ we get

$$\Psi_n(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{n} \Psi(\boldsymbol{y}_i;\boldsymbol{t}) = \boldsymbol{0}$$

and we set it to be $\boldsymbol{0}$. Under regularity conditions the MLE solves the equation

$$\Psi_n(\widehat{\boldsymbol{\theta}_0}) = \frac{1}{n}\sum_{i=1}^{n} \Psi(\boldsymbol{y}_i;\boldsymbol{\theta}_0) = \boldsymbol{0}$$

Unfortunately there may be no solution or multiple solution to the equation.

9.6 DEFINITION. The Hessian Matrix is defined by differentiating the score function

$$\begin{pmatrix} \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) & \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) \\ \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) & \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2}{\partial^2\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) & \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta}\Psi(\boldsymbol{y};\boldsymbol{\theta}) \end{pmatrix}$$

which contains the differentiation w.r.t $\boldsymbol{\theta}$ of score function.

Notice the score function is first direvative of log pdf/pmy w.r.t $\boldsymbol{\theta}$ which is the gradient while Hessian is the second derivative of log pdf/pmf w.r.t $\boldsymbol{\theta}$.

9.7 DEFINITION. (Fisher Information matrix) The matrix is given as the covariance of score function under **true** parameter which is

$$\begin{aligned} \mathrm{I}(\boldsymbol{\theta}) &= \mathrm{Cov}(\Psi(\mathbf{Y};\boldsymbol{\theta}_0)) = \mathbb{E}_{\boldsymbol{\theta}_0}\{\Psi(\mathbf{Y};\boldsymbol{\theta}_0)\Psi(\mathbf{Y};\boldsymbol{\theta}_0)'\} \\ &= \mathbb{E}_{\boldsymbol{\theta}_0}\left\{\Big(\frac{\partial}{\partial\boldsymbol{\theta}}\log[f(\mathbf{Y};\boldsymbol{\theta})]\Big)\Big(\frac{\partial}{\partial\boldsymbol{\theta}}\log[f(\mathbf{Y};\boldsymbol{\theta})]\Big)'\right\} \end{aligned}$$

It also can be shown that under regularity conditions:

$$\mathrm{I}(\boldsymbol{\theta}) = -\mathbb{E}\Big\{\mathrm{H}(\boldsymbol{\theta}_0)\Big\}$$

Usually the second way is easier to compute.

9.3 *Regularity Conditions*

The standard regularity conditions in the context of MLE theory are

- The parameter $\boldsymbol{\theta}$ is identifiable ($\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \implies \mathrm{F}_{\boldsymbol{\theta}_1} \neq \mathrm{F}_{\boldsymbol{\theta}_2}$)
- The support of $f(\boldsymbol{y};\boldsymbol{\theta})$ dose NOT depends on $\boldsymbol{\theta}$.
- THe parameter space Θ contains an open set of which the true parameter is an interior point, that is the true parameter is not on the boundary of Θ
- The order of differentiation and expected values can be interchanged

These conditions will guarantee most of the asymptotic behaviour in MLE.

9.4 *Properties of Score function*

9.8 PROPOSITION. *Let $\boldsymbol{\theta}_0$ be the true parameter. Then*

$$\Psi(\boldsymbol{\theta}_0) = 0$$

Notice the difference of the two notation

- $\Psi(\cdot) = \mathbb{E}_{\boldsymbol{\theta}_0}(\boldsymbol{y};\cdot)$ w.r.t $\boldsymbol{y}$

- $\Psi(\cdot;\cdot)$ is the score function.

9.9 PROPOSITION. *Under regularity conditions*

$$\mathrm{I}(\boldsymbol{\theta}_0) = \mathbb{E}_{\theta_0}\left\{\Psi(\mathbf{Y};\boldsymbol{\theta}_0)\Psi(\mathbf{Y};\boldsymbol{\theta}_0)'\right\} = -\mathrm{H}(\boldsymbol{\theta}_0)$$

We have seen the consistency of $\widehat{\boldsymbol{\theta}}$, the MLE. It actually also has a asymptotic normal behaviour

9.10 THEOREM. *(Asymptotic Normality of $\widehat{\boldsymbol{\theta}}$) Let θ be the true parameter and $\widehat{\boldsymbol{\theta}}$ be the MLE. Then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \longrightarrow_d \mathrm{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$$

where $\mathbf{I}$ is the Fisher information matrix.

Prove of 1-dimensional case is in the note. Thus for n large enough we have

$$\widehat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{\theta}, \frac{1}{n}\mathbf{I}^{-1}(\widehat{\boldsymbol{\theta}})) \tag{40}$$

usually, $n/p \approx 15$ or 20 works well.

9.5 *Confidence Interval Construction MLE*

Assume the MLE $\widehat{\boldsymbol{\theta}}$ is already obtained and the true value of parameter is $\boldsymbol{\theta}$.

Case 1: $\boldsymbol{\theta} \in \mathbb{R}^1$

This is an algorithm that we should use as engineering. Our ultimate goal is to find the $\mathrm{SE}(\widehat{\theta})$. With the asymptotic normality we get (40) which seems, but not necessarily and not guaranteed that

$$\frac{1}{n}\mathbf{I}^{-1}(\widehat{\boldsymbol{\theta}}) \longrightarrow \mathrm{Var}(\widehat{\theta})$$

However, we will still consider to use this (i.e the LHS) as an estimation of the asymptotic variance.

- Compute

  $$\log(f(y;\theta))$$

  at a singe sample
- Compute the score function at the point

  $$\Psi(y;\theta) = \frac{\partial}{\partial\theta}\log\left\{f(y;\theta)\right\}$$
- Hessian matrix (a number in this case)

  $$\mathrm{H}(y;\theta) = \frac{\partial}{\partial\theta}\Psi(y;\theta)$$

- Fisher information matrix

$$\mathrm{I}(\theta) = -\mathbb{E}\Big(\mathrm{H}(y;\theta)\Big)$$

  Then finally

$$\frac{1}{n}\mathbf{I}^{-1}(\widehat{\theta}) \approx \mathrm{V}ar(\widehat{\theta})$$

  and thus

$$\mathrm{SE}(\widehat{\theta}) = \sqrt{\frac{1}{n}\mathbf{I}^{-1}(\widehat{\theta})}$$

- The (1-α) 100% C.I for θ

$$\widehat{\theta} \pm \Phi^{-1/2}\Big(1 - \frac{\alpha}{2}\Big) \times \mathrm{SE}(\widehat{\theta})$$

10 EXPECTATION MAXIMIZATION (EM) ALGORITHM

The EM algorithm is used to compute MLE estimators when some information is missing. For example, some entries in a data table are missing. Then the maximization of the likelihood function can be difficult. The EM strategy is to replace a single difficult problem by a sequence of easy optimization steps. The main application if the estimation of the parameter of a mixture models. So we will introduce the model first

10.1 *Mixture Model*

10.1 DEFINITION. A random vector $\mathbf{Y}$ has a mixture distribution if the joint density of $\mathbf{Y}$ is of the form

$$f(\boldsymbol{y}) = \alpha_1 f_1(\boldsymbol{y}) + \alpha_2 f_2(\boldsymbol{y}) + \cdots + \alpha_m f_m(\boldsymbol{y})$$

where

- α_j are positive numbers s.t

$$\sum_{j=1}^{m} \alpha_j = 1$$

- $f_j(\boldsymbol{y})$ are density functions of KNOWN shape (i.e normal with mean μ_j and σ_j^2)

So the $f(\boldsymbol{y})$ specify a density function that a series of sample comes from.

The generation of the mixture model can be thought as a two steps experiments.

- **Step one:** One of the mixture component is randomly selected. Notice each time only one component can be choosen, that is, we use a r.vector

$\mathbf{X}$ to model this step

$$\mathbf{X} = (\mathrm{X}_1, \mathrm{X} = 2, \cdots, \mathrm{X}_m) \sim \text{Multinomial}(n = 1, \alpha_1, \cdots, \alpha_m) \tag{41}$$

where α_j are exactly the probability been selected. So the joint density here, with $n = 1$ is simply

$$h(\mathbf{X}) = \prod_{j=1}^{m} \alpha_j^{\mathrm{X}_j}$$

If we look at the marginal density of $\mathbf{X}$ at $\boldsymbol{x}_j$ is

$$h(\boldsymbol{x}_j) = \alpha_j$$

where $\boldsymbol{x}_j$ is the vector that are all zero except the position j.

- **Step 2:** The vector Y is randomly obtained from the sub-population selected in the first step. Then

$$f(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{j=1}^{m} \left[f_j(\boldsymbol{y}) \right]^{x_j}$$

That is

$$f\left(\boldsymbol{y} \mid \boldsymbol{x}_j\right) = f_j(\boldsymbol{y})$$

Then we can put them together. The joint density of $\mathbf{X}$ and $\mathbf{Y}$ is

$$f(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x}) f(\boldsymbol{y} \mid \boldsymbol{x}) = \left(\prod_{j=1}^{m} \alpha_j^{\mathrm{X}_j} \right) \prod_{j=1}^{m} \left[f_j(\boldsymbol{y}) \right]^{x_j}$$

The result of the first step is not available in practice. So that is we only know $\boldsymbol{y}$ and the marginal density of $\mathbf{Y}$

$$f(\boldsymbol{y}) = \sum_{j=1}^{m} f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{m} \alpha_j f_j(\boldsymbol{y})$$

we don't know joint density but marginal density.

The density f_j usually includes unknown parameters. For example

$$f_j(\boldsymbol{y}) = \mathrm{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_j, \Sigma_j\right), \quad j = 1, 2, .., m$$

which is the Gaussian mixture model. Those α_j are unknown weights with sum 1 constrain. So the mixture model have plenty parameters. So the missing part of the model can be

- Unknown number of mixture components
- Unknown number of mixture components

- Missing entries in the data table

Application of EM algorithm

- Estimation of parameters of mixture mode, with application to cluster analysis
- Estimation of multivariate location and scatter matrix in the presence of missing data
- Estimation of model parameters in the presence of latent variables

Notations:

- $\mathbf{Y}$ as the incomplete data (only data we can use)
- $\mathbf{X}$ as the Augmented data (Artificial part may want to eliminate)
- $(\boldsymbol{y}, \mathbf{X})$ as the Complete data
- $\boldsymbol{\theta}$ as unknown parameters

The algorithm has two main steps: The **Expectation** step and the **Maximization step**

E-Step

The incomplete data log-likelihood is

$$\mathrm{I}(\boldsymbol{\theta}; \boldsymbol{y}) = \log[f(\boldsymbol{y}; \boldsymbol{\theta})]$$

The complete data log-likelihood is

$$\mathrm{I}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) = \log[f(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})]$$

Then the expected log-likelihood is

$$\begin{aligned} \widetilde{\mathrm{I}}\left(\boldsymbol{\theta} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(k)}\right) &= \mathbb{E}_{\mathrm{X}|\boldsymbol{y};\boldsymbol{\theta}^{(k)}}\{\mathrm{I}(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{X})\} \\ &= \int \cdots \int \log[f(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})] h\left(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(k)}\right) d\boldsymbol{x} \end{aligned} \tag{42}$$

So the expectation log-likelihood, by conditioning on $\boldsymbol{y}$, eliminate the unknown $\boldsymbol{x}$ where $\boldsymbol{\theta}^{(k)}$ is the current value of θ in the iteration and the $\boldsymbol{\theta}$ is the parameter we want to estimate. The density $h(\boldsymbol{x} \mid \boldsymbol{y})$ has no problem. We take expectation w.r.t $\boldsymbol{x}$ and with given $\boldsymbol{y}$. The specified version of it is

$$h\left(x \mid y; \theta^{(k)}\right) = \frac{f\left(y, x; \theta^{(k)}\right)}{f\left(y; \theta^{(k)}\right)}$$

M-Step

$$\theta^{(k+1)} = \arg\max_{\theta} \widetilde{\mathrm{I}}\left(\theta \mid y; \theta^{(k)}\right)$$

and it is guranteed

$$\mathrm{I}\left(\theta^{(k+1)}; y\right) \geq \mathrm{I}\left(\theta^{(k)}; y\right)$$

which is the ascending property of EM-algorithm.

Overview of the Steps in the EM algorithm

- Write the density for each SINGLE observation
- Write the likelihood for the (entire) incomplete data
- Construct the complete-data log-likelihood function
- Take expectation of the complete-data log-likelihood function using the conditional distribution of the augmented data given the incomplete data, and the current values of the parameters. This is called the E-step.
- Maximize the resulting expected log-likelihood function. This is called the M-step.
- Repeat steps 4 and 5 until convergence.

Notice the "complete data construction" is not int a real way while it is in a theoretical way.

10.2 EXAMPLE. Let the mixture density to be simply

$$f(y, p) = (1 - p) * f_0(y) + p * f_1(y)$$

and the only unknown parameter is the p (i.e both f_0 and f_1 are fully specified with no unknown parameters). The sample data observable are

$$y_1, y_2, ..., y_n$$

then the incomplete data likelihood is

$$\mathrm{I}(p; \boldsymbol{y}) = \sum_{i=1}^{n} \log\left[(1 - p) * f_0(y) + p * f_1(y)\right]$$

The complete data in this case should be

$$(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$$

where each X_i is from a binomial(1,p) (or Bernoulli(p) which is

$$\mathrm{X}_i = \begin{cases} 0 & \text{with prob} \quad 1 - p \\ 1 & \text{with prob} \quad p \end{cases}$$

Therefore

$$h(\mathrm{X}_i; p) = (1-p)^{1-\mathrm{X}_i} p^{\mathrm{X}_i}$$

then for each pair (y_i, x_i) iid, the bivariate density is

$$f(x, y) = h(x; p) f(y \mid x)$$

We assume that

$$f(y \mid 0) = f_0(y), \quad f(y \mid 1) = f_1(y)$$

Therefore

$$f(x, y) = h(x; p) f_x(y) = \begin{cases} (1-p) f_0(y) & \text{if} \quad x = 0 \\ p f_1(y) & \text{if} \quad x = 1 \end{cases}$$

Since X_i can only take values 0 and 1 ,

$$\begin{aligned} \mathrm{E}(\mathrm{X}_i \mid \mathrm{Y}_i = y_i) &= \mathrm{P}(\mathrm{X}_i = 1 \mid \mathrm{Y}_i = y_i) \\ &= f(1 \mid y_i) = \frac{f(1, y_i)}{f(y_i)} \\ &= \frac{h(1; p) f(y_i \mid 1)}{h(1; p) f(y_i \mid 1) + h(0; p) f(y_i \mid 0)} \\ &= \frac{p f_1(y_i)}{p f_1(y_i) + (1-p) f_0(y_i)} \\ &= \tilde{p}_i \end{aligned}$$

This is considered to be an estimate of the probabitlity that y_i comes from population of density $f_1(t)$. Then we compute the complete data log-liklihood

$$\mathrm{I}(p; \boldsymbol{y}, \mathbf{X}) = \sum_{i=1}^{n} \mathrm{X}_i \log\left[p f_1(y_i)\right] + \sum_{i=1}^{n} (1 - \mathrm{X}_i) \log\left[(1-p) f_0(y_i)\right]$$

a useful trick is

$$f(x, y; p) = \left[p f_1(y)\right]^x \left[(1-p) f_0(y)\right]^{1-x}$$

then

$$\begin{aligned} \tilde{\mathrm{I}}(p; \boldsymbol{y}) &= \mathrm{E}\left\{\mathrm{I}(p; y, \mathrm{X}) \mid y; p^{(k)}\right\} \\ &= \mathrm{E}\left\{\sum_{i=1}^{n} \mathrm{X}_i \log\left[p f_1(y_i)\right] + \sum_{i=1}^{n} (1 - \mathrm{X}_i) \log\left[(1-p) f_0(y_i)\right] \middle| y; p^{(k)}\right\} \\ &= \sum_{i=1}^{n} \mathrm{E}\left\{\mathrm{X}_i \mid y_i; p^{(k)}\right\} \log\left[p f_1(y_i)\right] + \sum_{i=1}^{n} \mathrm{E}\left\{(1 - \mathrm{X}_i) \mid y_i; p^{(k)}\right\} \log\left[(1-p) f_0(y_i)\right] \end{aligned}$$

where the conditional expectation part is done already, so

$$= \sum_{i=1}^{n} [\tilde{p}_i \log(p) + (1 - \tilde{p}_i) \log(1 - p)] + C$$

Then the M-Step is to find the maximizer $p^{(k+1)}$ of $\tilde{I}(p; \boldsymbol{y})$ srt p. So set the derivative to 0

$$\frac{d}{dp} \widetilde{I}(p; y) = \sum_{i=1}^{n} \left(\frac{\tilde{p}_i}{p} - \frac{1 - \tilde{p}_i}{1 - p} \right) = 0$$

and solve for p. Result is

$$p^{(k+1)} = \frac{\sum_{i-1}^{n} \tilde{p}_i}{n}$$

Initialization of $p^{(0)}$

Use K-means (for instance) to get an initial partition of the data into two sets like class 0 and class with size n_0 and n_1 respectively. We set

$$p^{(0)} = \frac{n_1}{n}$$

Iteration Step

Once initialization is done, we do

$$\tilde{p}_i = \frac{p^{(k)} f_1(y_i)}{p^{(k) f_1}(y_i) + \left(1 - p^{(k)}\right) f_0(y_i)'}$$

and set

$$p^{(k+1)} = \frac{\sum_{i-1}^{n} \tilde{p}_i}{n}$$

Stopping Rule

Given some small $\delta > 0$, stop when

$$\left| p^{(k+1)} - p^{(k)} \right| < \delta$$

and return

$$\widehat{p} = p^{(k+1)}$$

Possible extension

- There are $m > 2$ sub-populations
- The densities $f_j(y)$ have unknown parameters
- The observations are multivariate, $y \in R^p$
- The observed data table has missing data

All these situation can be expressed by the EM algorithm.

Case for more than 2 sub-population

See notes of several componets. Make up later.