Mathematical Statistics

Notes for STAT3001, The University of Queensland

George Stuart

Semester 1, 2026

About these notes. These are my notes from STAT3001. They combine two sets of notes I took during the course, one on the course's background notes and one on its lecture notes, arranged by topic. They were written to go alongside the course material rather than to replace it: some sections only summarise the main points, a few stop part-way through, and many proofs are left to the course notes. Page references are to those notes.

Contents

I Background **2**
1 Probability Background . . . 2
2 The Multivariate Normal Distribution . . . 4

II Point Estimation **9**
3 Point Estimation . . . 9
4 Maximum Likelihood Estimation . . . 13
5 The Cramér–Rao Lower Bound . . . 15
6 The Exponential Family . . . 19
7 Properties of Estimators . . . 27
8 Estimator Theorems . . . 30
9 Asymptotic Properties . . . 34
10 The EM Algorithm . . . 36

III Hypothesis Testing **46**
11 Hypothesis Testing . . . 46
12 The Chi-Squared Goodness of Fit Test . . . 51
13 Multiple Testing . . . 53
14 Errors in Testing . . . 57
15 Tests for Normal Means and Variances . . . 57
16 Tukey's Range Test . . . 65

Part I

Background

1 Probability Background

Given a continuous random variable $\boldsymbol{X}$ with some density function $f(\boldsymbol{x};\, \boldsymbol{\theta})$ we define,

$$P(\boldsymbol{X} \in A) = \int_A f(\boldsymbol{x};\, \boldsymbol{\theta})\, d\boldsymbol{x}$$

We then define,

$$E[h(\boldsymbol{X})] = \int h(\boldsymbol{x}) f(\boldsymbol{x};\, \boldsymbol{\theta})\, d\boldsymbol{x}$$

Given two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ the joint density $f_{\boldsymbol{XY}}$ is a function that satisfies,

$$P(\boldsymbol{X} \in A,\ \boldsymbol{Y} \in B) = \int_A \int_B f_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y};\, \boldsymbol{\theta})\, d\boldsymbol{y}\, d\boldsymbol{x}$$

We can then define the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by,

$$\text{cov}(\boldsymbol{X}, \boldsymbol{Y}) = E\Big[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{Y} - E[\boldsymbol{Y}])^T\Big]$$

(In the univariate case $\text{cov}(X, X)$ is referred to as $\text{var}(X)$ and is a scalar value.)

1.1 Mean and Covariance of a Random Vector

Suppose we have some random vector $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{X} = (X_1, \ldots, X_n)^T$. Here,

$$\boldsymbol{\mu} = E[\boldsymbol{X}]$$

and,

$$\begin{aligned}
\text{cov}(\boldsymbol{X}) &= \boldsymbol{\Sigma} \\
&= E[(\boldsymbol{X} - \boldsymbol{\mu}) \otimes (\boldsymbol{X} - \boldsymbol{\mu})] \\
&= E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]
\end{aligned}$$

1.2 Covariance Decomposition

An important result to know is the decomposition of the covariance matrix,

$$\begin{aligned}
\text{cov}(\boldsymbol{X}, \boldsymbol{Y}) &= E\Big[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{Y} - E[\boldsymbol{Y}])^T\Big] \\
&= E\Big[\boldsymbol{X}\boldsymbol{Y}^T - \boldsymbol{X}E[\boldsymbol{Y}]^T - E[\boldsymbol{X}]\boldsymbol{Y}^T + E[\boldsymbol{X}]E[\boldsymbol{Y}]^T\Big] \\
&= E\Big[\boldsymbol{X}\boldsymbol{Y}^T\Big] - E[\boldsymbol{X}]E[\boldsymbol{Y}]^T - E[\boldsymbol{X}]E[\boldsymbol{Y}]^T + E[\boldsymbol{X}]E[\boldsymbol{Y}]^T \\
&= E\Big[\boldsymbol{X}\boldsymbol{Y}^T\Big] - E[\boldsymbol{X}]E[\boldsymbol{Y}]^T
\end{aligned}$$

Taking $\boldsymbol{Y} = \boldsymbol{X}$ gives $\text{cov}(\boldsymbol{X}) = E[\boldsymbol{X}\boldsymbol{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$, which is the multivariate version of the familiar result $\text{var}(X) = E[X^2] - E[X]^2$.

1.3 Properties of the Covariance Matrix

Properties of $\boldsymbol{\Sigma}$:

- positive semi-definite (meaning that $\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$)
- symmetric (since $\left((\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\right)^T = (\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T$)
- the diagonal entries are the variances of each variable in the vector
- the off-diagonal entries are the covariances between pairs of variables in the vector

Recall,

$$\begin{aligned}\operatorname{cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY] - \mu_X\mu_Y\end{aligned}$$

Also recall that,

- If $\operatorname{cov}(X,Y) = 0$ then X and Y are said to be "uncorrelated"
- If X and Y are independent then X and Y will also be uncorrelated
- If X and Y are *jointly* normally distributed *and* uncorrelated then they are independent (proved in Section 2.4)

1.4 Random Variables as Vector Spaces

The set of all random variables on a probability space $(\Omega, \mathcal{F}, P)$ with finite variance can be shown to form a vector space (sums and scalar multiples of random variables are again random variables).

There are many interesting things that can be said about this, but perhaps one of the most important is that given this interpretation of random variables we can define an inner product by,

$$\langle X, Y\rangle = E[XY]$$

(strictly speaking this is an inner product once random variables that are equal with probability 1 are treated as the same). You can then easily confirm that,

$$\operatorname{cov}(X,Y) = \langle X - E[X], Y - E[Y]\rangle$$

and so the covariance can be thought of as a centralised inner product. (For random vectors, $E[\boldsymbol{X}\boldsymbol{Y}^T]$ is the matrix of these inner products between the components of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and $\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})$ is the matrix of the centralised ones.)

This allows us to use results from linear algebra on random variables. For example from the Cauchy–Schwarz inequality we have that,

$$\begin{aligned}\operatorname{cov}(X,Y)^2 &\leq E[(X-E[X])^2]E[(Y-E[Y])^2] \\ &= \operatorname{var}(X)\operatorname{var}(Y)\end{aligned}$$

(this will be used later in the proof of the CRLB in the univariate case).

1.5 Linear Functions of Random Vectors

Let $\boldsymbol{b} = (b_1, \ldots, b_d)^T$ and $\boldsymbol{A} \in \mathbb{R}^{d\times n}$ be a matrix of constants. Then define the random vector $\boldsymbol{Y}$ by,

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$$

Then,

$$\begin{aligned}E[\boldsymbol{Y}] &= E(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) \\ &= \boldsymbol{A}E[\boldsymbol{X}] + \boldsymbol{b}\end{aligned}$$

and,

$$\begin{aligned}\operatorname{cov}(\boldsymbol{Y}) &= E[(\boldsymbol{Y}-E[\boldsymbol{Y}])(\boldsymbol{Y}-E[\boldsymbol{Y}])^T] \\ &= E[(\boldsymbol{AX}+\boldsymbol{b}-\boldsymbol{A}E[\boldsymbol{X}]-\boldsymbol{b})(\boldsymbol{AX}+\boldsymbol{b}-\boldsymbol{A}E[\boldsymbol{X}]-\boldsymbol{b})^T] \\ &= E[(\boldsymbol{A}(\boldsymbol{X}-E[\boldsymbol{X}]))(\boldsymbol{A}(\boldsymbol{X}-E[\boldsymbol{X}]))^T] \\ &= E[\boldsymbol{A}(\boldsymbol{X}-E[\boldsymbol{X}])(\boldsymbol{X}-E[\boldsymbol{X}])^T\boldsymbol{A}^T] \\ &= \boldsymbol{A}E[(\boldsymbol{X}-E[\boldsymbol{X}])(\boldsymbol{X}-E[\boldsymbol{X}])^T]\boldsymbol{A}^T \\ &= \boldsymbol{A}\operatorname{cov}(\boldsymbol{X})\boldsymbol{A}^T\end{aligned}$$

(recall the analogous results in univariate probability: $E[aX+b]=aE[X]+b$ and $\operatorname{var}(aX+b)=a^2\operatorname{var}(X)$).

2 The Multivariate Normal Distribution

2.1 Density Function

If $\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ then the p.d.f. of $\boldsymbol{X}$ is given by,

$$f(\boldsymbol{x}\ ;\ \boldsymbol{\theta}) = (2\pi)^{\frac{-n}{2}}\det(\boldsymbol{\Sigma})^{\frac{-1}{2}}\exp\left(\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Proof. Intuitively, looking at the formula it does seem like a reasonable generalisation of the univariate normal distribution. However here I will show where the formula comes from explicitly. To do this we will start with the simple case where,

$$\boldsymbol{Z}=(Z_1,\dots,Z_n)^T \quad \text{and} \quad Z_1,\dots,Z_n \overset{\text{i.i.d.}}{\sim} N(0,1)$$

In other words this means that,

$$\boldsymbol{Z}\sim N(\boldsymbol{0},\boldsymbol{I})$$

since each r.v. has variance 1 and there is no covariance (since independence $\rightarrow$ zero covariance).

Therefore the p.d.f of $\boldsymbol{Z}$ which I will denote $f_{\boldsymbol{Z}}(\boldsymbol{z})$ is given by,

$$\begin{aligned}f_{\boldsymbol{Z}}(\boldsymbol{z}) &= \prod_{j=1}^{n}(2\pi)^{\frac{-1}{2}}\exp\left(\frac{-1}{2}z_j^2\right) \\ &= (2\pi)^{\frac{-n}{2}}\exp\left(\frac{-1}{2}\sum_{j=1}^{n}z_j^2\right) \\ &= (2\pi)^{\frac{-n}{2}}\exp\left(\frac{-1}{2}\boldsymbol{z}^T\boldsymbol{z}\right)\end{aligned}$$

Now with this result we can generalise to an arbitrary multivariate normal vector $\boldsymbol{X}$. To do this we can apply a linear transformation to the standard normal $\boldsymbol{Z}$ as follows,

$$\text{let } \boldsymbol{X}=\boldsymbol{AZ}+\boldsymbol{\mu}$$

where the non-singular matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ and $\boldsymbol{\mu}\in\mathbb{R}^n$ are constants. Note that as we have shown in Section 1,

$$E[\boldsymbol{X}]=\boldsymbol{\mu} \qquad \text{and} \qquad \begin{aligned}\operatorname{cov}(\boldsymbol{X}) &= \boldsymbol{A}\operatorname{cov}(\boldsymbol{Z})\boldsymbol{A}^T \\ &= \boldsymbol{AA}^T\end{aligned}$$

Therefore the distribution of $\boldsymbol{X}$ is given by,

$$\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad \text{where } \boldsymbol{\Sigma}=\boldsymbol{AA}^T$$

To find the p.d.f of $\boldsymbol{X}$ we can then use the change of variables formula,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Z}}(\boldsymbol{z}) \left| \det \frac{d\boldsymbol{z}}{d\boldsymbol{x}} \right|$$

Recall that in this formula $\frac{d\boldsymbol{z}}{d\boldsymbol{x}}$ is the Jacobian matrix of the transformation function mapping the random variable $\boldsymbol{Z}$ to $\boldsymbol{X}$. This relationship can be found by,

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu} \quad \implies \quad \boldsymbol{z} = \boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

The derivative is then clearly,

$$\frac{d\boldsymbol{z}}{d\boldsymbol{x}} = \boldsymbol{A}^{-1}$$

Thus,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= f_{\boldsymbol{Z}}(\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})) \left| \det \boldsymbol{A}^{-1} \right| \\ &= (2\pi)^{\frac{-n}{2}} \left| \det \boldsymbol{A}^{-1} \right| \exp\left(\frac{-1}{2} (\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))^T (\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})) \right) \\ &= (2\pi)^{\frac{-n}{2}} \left| \det \boldsymbol{A}^{-1} \right| \exp\left(\frac{-1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{A}^{-T} \boldsymbol{A}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \end{aligned}$$

From here this is quite close to the form for the multivariate normal distribution however we are using this matrix $\boldsymbol{A}$ instead of the covariance matrix $\boldsymbol{\Sigma}$.

First looking at the determinant,

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{A}\boldsymbol{A}^T \\ \det(\boldsymbol{\Sigma}) &= \det(\boldsymbol{A}) \det(\boldsymbol{A}^T) \\ &= \det(\boldsymbol{A})^2 \\ \implies \sqrt{\det(\boldsymbol{\Sigma})} &= |\det(\boldsymbol{A})| \\ \frac{1}{\sqrt{\det(\boldsymbol{\Sigma})}} &= \frac{1}{|\det(\boldsymbol{A})|} \\ \frac{1}{\sqrt{\det(\boldsymbol{\Sigma})}} &= \left| \det(\boldsymbol{A}^{-1}) \right| \\ \det(\boldsymbol{\Sigma})^{\frac{-1}{2}} &= \left| \det(\boldsymbol{A}^{-1}) \right| \end{aligned}$$

Then looking at the inner $\boldsymbol{A}^{-T}\boldsymbol{A}^{-1}$ term,

$$\begin{aligned} \boldsymbol{A}^{-T}\boldsymbol{A}^{-1} &= (\boldsymbol{A}\boldsymbol{A}^T)^{-1} \\ &= \boldsymbol{\Sigma}^{-1} \end{aligned}$$

And so with this all together we get the expected result. □

2.2 Useful Results for Sub-Vectors

Here are some useful results when dealing with sub-vectors of a multivariate normal distribution.

Suppose that $\boldsymbol{W}$ is a multivariate normal random variable partitioned into two sub-vectors $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ such that,

$$\boldsymbol{W} = (\boldsymbol{W}_1^T, \boldsymbol{W}_2^T)^T$$

where $\boldsymbol{W}_i$ is of dimension p_i and $\boldsymbol{W}$ is of dimension $p = p_1 + p_2$.

Then let,

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

(just for clarity note that $\boldsymbol{\Sigma}_{11} = \text{cov}(\boldsymbol{W}_1, \boldsymbol{W}_1)$ and so it is also a matrix).

Then if $\boldsymbol{W} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the conditional distribution of $\boldsymbol{W}_1$ given $\boldsymbol{W}_2 = \boldsymbol{w}_2$ is p_1-variate normal with mean,

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{w}_2 - \boldsymbol{\mu}_2)$$

and covariance matrix,

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

These are not derived in the course, but they are useful to remember (and they are not too bad to memorise).

In general this formula applies whenever the two sub-vectors are jointly normal, as they are here since $(\boldsymbol{W}_1^T, \boldsymbol{W}_2^T)^T$ follows a joint normal distribution.

2.2.1 Example on Sub-Vectors

Suppose we have a random sample from a univariate normal distribution,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

We can then define the sample as a single instance of an n-dimensional normal random variable $\boldsymbol{X}$ where,

$$\boldsymbol{X} \sim N(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n)$$

Then the question is,

$$\text{What is the distribution of } X_i | \overline{X} = \overline{x}?$$

Here we can show that $\boldsymbol{W} = (X_i, \overline{X})^T$ follows a joint normal distribution since,

$$\boldsymbol{W} = \boldsymbol{A}\boldsymbol{X}$$

where $\boldsymbol{A}$ is a $2 \times n$ matrix where the first row is all zeros with a 1 in the ith position and the second row is all $\frac{1}{n}$. You can confirm that expanding this matrix product out would give $(X_i, \overline{X})$.

Given this we know that $\boldsymbol{W}$ will be jointly normal since,

$$\boldsymbol{W} \sim N(\boldsymbol{A}\mu \mathbf{1}_n, \boldsymbol{A}\sigma^2 \boldsymbol{I}_n \boldsymbol{A}^T)$$

With this we have established that $\boldsymbol{W}$ is a jointly normal random variable and so we can use the result shown above to determine the conditional distribution of its subvectors (i.e. $X_i | \overline{X}$). To use the formula however we need to compute $\boldsymbol{\mu}_{\boldsymbol{W}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{W}}$.

This can be done two ways:

- Calculate by looking at the expression for the distribution of $\boldsymbol{W}$
  - Here $\boldsymbol{\mu}_{\boldsymbol{W}} = \boldsymbol{A}\mu \mathbf{1}_n$ and $\boldsymbol{\Sigma}_{\boldsymbol{W}} = \boldsymbol{A}\sigma^2 \boldsymbol{I}_n \boldsymbol{A}^T$
  - You would then need to think about what this matrix product expands out to.
- Calculate intuitively using the fact that $\boldsymbol{W} = (X_i, \overline{X})^T$
  - Then $\boldsymbol{\mu}_{\boldsymbol{W}} = (E[X_i], E[\overline{X}])^T$ and $\Sigma_{\boldsymbol{W}} = \text{cov}(\boldsymbol{W})$
  - Then compute each value term by term.

The lecture notes use the second method. In that approach the matrix $\boldsymbol{A}$ isn't used for the calculation at all; its only purpose is to show that $\boldsymbol{W}$ is normal (which is required to use the conditional normality result above). Either way works.

Using method 1: Firstly looking at the mean vector, $\boldsymbol{\mu}_{\boldsymbol{W}} = \boldsymbol{A}\mu \mathbf{1}_n = \mu \boldsymbol{A}\mathbf{1}_n$. Here $\boldsymbol{A}\mathbf{1}_n$ is just going to expand to give you $\mathbf{1}_2$ and so $\boldsymbol{\mu}_{\boldsymbol{W}} = \mu \mathbf{1}_2 = (\mu, \mu)^T$.

Next looking at the covariance matrix, $\boldsymbol{\Sigma_W} = \boldsymbol{A}\sigma^2\boldsymbol{I}_n\boldsymbol{A}^T = \sigma^2\boldsymbol{A}\boldsymbol{A}^T$. Expanding this out you would get that $\boldsymbol{\Sigma_W} = \begin{pmatrix} \sigma^2 & \frac{\sigma^2}{n} \\ \frac{\sigma^2}{n} & \frac{\sigma^2}{n} \end{pmatrix}$.

Using method 2: Here $E[X_i] = \mu$ and $E[\overline{X}] = \mu$ therefore $\boldsymbol{\mu_W} = (\mu, \mu)^T$.

Then for the covariance matrix we need to think about it term by term,

$$\begin{aligned} \operatorname{cov}(X_i, X_i) &= \operatorname{var}(X_i) = \sigma^2 \\ \operatorname{cov}(\overline{X}, \overline{X}) &= \operatorname{var}(\overline{X}) = \frac{\sigma^2}{n} \\ \operatorname{cov}(X_i, \overline{X}) &= \frac{1}{n}\sum_{j=1}^{n} \operatorname{cov}(X_i, X_j) = \frac{1}{n}\operatorname{var}(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

(using that $\operatorname{cov}(X_i, X_j) = 0$ for $j \neq i$ by independence). Putting this all together gives the same covariance matrix $\boldsymbol{\Sigma_W}$ as was found using method 1. In this case method 1 turned out to be easier.

Thus given that $\boldsymbol{W} = (X_i, \overline{X})$ is distributed $N(\boldsymbol{\mu_W}, \boldsymbol{\Sigma_W})$ we can compute the conditional distribution $X_i|\overline{x}$ as,

$$X_i|\overline{X} = \overline{x} \sim N(\mu_V, \sigma_V^2)$$

where,

$$\begin{aligned} \mu_V &= \mu_{X_i} + \Sigma_{12}\Sigma_{22}^{-1}(\overline{x} - \mu_{\overline{X}}) \\ &= \mu + \frac{\sigma^2}{n}\left(\frac{\sigma^2}{n}\right)^{-1}(\overline{x} - \mu) \\ &= \mu + \overline{x} - \mu \\ &= \overline{x} \end{aligned}$$

and,

$$\begin{aligned} \sigma_V^2 &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \sigma^2 - \frac{\sigma^2}{n}\left(\frac{\sigma^2}{n}\right)^{-1}\frac{\sigma^2}{n} \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \sigma^2\left(1 - \frac{1}{n}\right) \end{aligned}$$

Thus,

$$X_i|\overline{X} = \overline{x} \sim N\left(\overline{x}, \left(1 - \frac{1}{n}\right)\sigma^2\right)$$

(Note that in this case the two constituent subvectors were just scalars X_i and $\overline{X}$, but in general if they were vectors then $\boldsymbol{\mu}_i$ would be vectors and $\boldsymbol{\Sigma}_{ij}$ would be matrices. Then in the formula $\boldsymbol{\Sigma}_{ij}^{-1}$ would represent a matrix inverse, not just the reciprocal as it did here).

2.3 The Sample Variance, s^2

Lecture note 1 also had some material on the distribution of s^2 (said to come from STAT2004). It doesn't match the multivariate theme of the rest of that lecture note, so it is kept separate here.

Recall that for,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

we define,

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad \text{and} \qquad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

Here we want to prove two important results:

1. Independence: that $\overline{X}$ and s^2 are independent
2. Distribution: that $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

To do this I am going to rely on the result that,

$$\text{uncorrelated} + \text{jointly normal} \implies \text{independence}$$

and so I will prove this first.

2.4 Independence of Uncorrelated Normal Random Variables

Suppose $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are jointly normal with $\text{cov}(X, Y) = 0$.

Given that they are jointly normal,

$$\text{let } \boldsymbol{W} = (X, Y) \quad \text{where} \quad \boldsymbol{W} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

In this case,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}$$

Note that the off diagonals on $\boldsymbol{\Sigma}$ are zero since covariance between X and Y is zero.

The density function is thus given by,

$$f(\boldsymbol{w}) = (2\pi)^{\frac{-2}{2}} (\det \boldsymbol{\Sigma})^{\frac{-1}{2}} \exp\left(\frac{-1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w} - \boldsymbol{\mu})\right)$$

Looking at this notice that,

$$\det \boldsymbol{\Sigma} = \sigma_X^2 \sigma_Y^2$$

and,

$$\begin{aligned}
(\boldsymbol{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w} - \boldsymbol{\mu}) &= \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}^T \frac{1}{\sigma_X^2 \sigma_Y^2} \begin{pmatrix} \sigma_Y^2 & 0 \\ 0 & \sigma_X^2 \end{pmatrix} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix} \\
&= \frac{1}{\sigma_X^2 \sigma_Y^2} \begin{pmatrix} \sigma_Y^2 (x - \mu_X) \\ \sigma_X^2 (y - \mu_Y) \end{pmatrix}^T \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix} \\
&= \frac{1}{\sigma_X^2 \sigma_Y^2} (\sigma_Y^2 (x - \mu_X)^2 + \sigma_X^2 (y - \mu_Y)^2) \\
&= \frac{(x - \mu_X)^2}{\sigma_X^2} + \frac{(y - \mu_Y)^2}{\sigma_Y^2}
\end{aligned}$$

Putting this all together we can see that $f(\boldsymbol{w})$ decomposes into the product of its marginal distributions,

$$\begin{aligned}
f(\boldsymbol{w}) &= (2\pi)^{\frac{-2}{2}} (\sigma_X^2 \sigma_Y^2)^{\frac{-1}{2}} \exp\left(\frac{-1}{2}\left(\frac{(x - \mu_X)^2}{\sigma_X^2} + \frac{(y - \mu_Y)^2}{\sigma_Y^2}\right)\right) \\
&= (2\pi)^{\frac{-1}{2}} (\sigma_X^2)^{\frac{-1}{2}} \exp\left(\frac{-1}{2}\frac{(x - \mu_X)^2}{\sigma_X^2}\right) \times (2\pi)^{\frac{-1}{2}} (\sigma_Y^2)^{\frac{-1}{2}} \exp\left(\frac{-1}{2}\frac{(y - \mu_Y)^2}{\sigma_Y^2}\right) \\
&= f_X(x) \times f_Y(y)
\end{aligned}$$

2.5 Independence of $\overline{X}$ and s^2

The proof in the lecture notes is an old one, originally due to Fisher. There are cleaner and arguably "better" proofs of this independence (Basu's theorem in Section 8.5 gives one), but here I follow the idea of the proof in the lecture notes.

Given our random sample we can view it as a single observation of a multivariate normal distribution where,

$$\boldsymbol{X} = (X_1, \ldots, X_n)^T \quad \text{and} \quad \boldsymbol{X} \sim N(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n)$$

Now the idea of this proof comes from an interesting result that rotating a multivariate normal distribution preserves independence structure and Gaussianity. More concretely a rotation corresponds to an orthogonal matrix, and what we are claiming is that if we let $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$ for some orthogonal matrix $\boldsymbol{A}$ then $\boldsymbol{Y}$ will still be normally distributed and its constituent random variables will still be independent. The problem then becomes choosing a suitable rotation.

Take $\boldsymbol{A}$ to be any orthogonal matrix whose first row is $\frac{1}{\sqrt{n}}\mathbf{1}_n^T$ (one exists, since any unit vector can be extended to an orthonormal basis). Then,

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} \sim N(\mu \boldsymbol{A}\mathbf{1}_n,\ \sigma^2 \boldsymbol{A}\boldsymbol{A}^T) = N(\mu \boldsymbol{A}\mathbf{1}_n,\ \sigma^2 \boldsymbol{I}_n)$$

so $Y_1, \ldots, Y_n$ are independent normal random variables, each with variance σ^2 (the density of $N(\cdot, \sigma^2 \boldsymbol{I}_n)$ factorises into a product, exactly as in the previous subsection). Looking at the mean, the first entry of $\boldsymbol{A}\mathbf{1}_n$ is $\frac{1}{\sqrt{n}}\mathbf{1}_n^T\mathbf{1}_n = \sqrt{n}$, and every other entry is 0 since the other rows of $\boldsymbol{A}$ are orthogonal to the first row (and hence to $\mathbf{1}_n$). So,

$$Y_1 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j = \sqrt{n}\,\overline{X} \sim N(\sqrt{n}\,\mu, \sigma^2) \qquad \text{and} \qquad Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Next, since $\boldsymbol{A}$ is orthogonal it preserves lengths, so

$$\sum_{j=1}^{n} Y_j^2 = \boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{X} = \sum_{j=1}^{n} X_j^2$$

and therefore,

$$(n-1)s^2 = \sum_{j=1}^{n}(X_j - \overline{X})^2 = \sum_{j=1}^{n} X_j^2 - n\overline{X}^2 = \sum_{j=1}^{n} Y_j^2 - Y_1^2 = \sum_{j=2}^{n} Y_j^2$$

Both results now follow:

1. Independence: $\overline{X} = Y_1/\sqrt{n}$ is a function of Y_1 only, and s^2 is a function of $Y_2, \ldots, Y_n$ only. Since Y_1 is independent of $(Y_2, \ldots, Y_n)$, $\overline{X}$ and s^2 are independent.
2. Distribution: $\frac{(n-1)s^2}{\sigma^2} = \sum_{j=2}^{n}\left(\frac{Y_j}{\sigma}\right)^2$ is a sum of $n-1$ squared independent standard normals, so it is χ^2_{n-1}.

Part II

Point Estimation

3 Point Estimation

Suppose you have a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ that comes from some distribution $F_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta}$ belongs to a parameter space Ω.

Here $\boldsymbol{\theta}$ can be understood to contain all the unknown features of the probability distribution.

In point estimation we estimate $\boldsymbol{\theta}$ by a single point which is taken to be some function of the observed data. That is we estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$ where,

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

We refer to $\boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ as an estimate and the corresponding random quantity $\boldsymbol{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ as an estimator.

3.1 Risk of an Estimator

A natural question that may arise is to ask how good an estimator is at estimating $\boldsymbol{\theta}$. We can formalise this with the idea of the estimator's "risk".

Let ℓ denote a loss function where,

- $\forall \boldsymbol{\theta}, \boldsymbol{T} \quad \ell(\boldsymbol{\theta}, \boldsymbol{T}) \geq 0$
- $\forall \boldsymbol{\theta} \quad \ell(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$

The accuracy (or rather inaccuracy) of $\boldsymbol{T}$ can then be measured by the risk function,

$$R(\boldsymbol{\theta}, \boldsymbol{T}) = E\Big[\ell(\boldsymbol{\theta}, \boldsymbol{T})\Big]$$

A popular choice of loss function for example is one of the form:

$$\ell(\boldsymbol{\theta}, \boldsymbol{T}) = \|\boldsymbol{T} - \boldsymbol{\theta}\|^{\alpha}$$

for some constant $\alpha > 0$. Notice that:

- If $\alpha = 1$: then the risk corresponds to the "mean absolute error"
- If $\alpha = 2$: then the risk corresponds to the "mean squared error"

Ideally one would like to find an estimator that minimises the risk over all values of $\boldsymbol{\theta}$ in Ω. However no uniformly best estimator exists except in the trivial case where the parameter space Ω only contains a single point. This should be intuitive if you think about it. Suppose that the true value of $\boldsymbol{\theta}$ was $\boldsymbol{\theta}_0$. Then the estimator

$$\boldsymbol{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \boldsymbol{\theta}_0$$

would have zero risk. But this would be a terrible estimator if the true value was not $\boldsymbol{\theta}_0$ since this estimator just returns $\boldsymbol{\theta}_0$ regardless of the observed data.

3.2 Unbiased Estimators

One way of avoiding the difficulty mentioned above is to restrict one's attention to a specific class of estimators that may have some desirable properties. For example we may want the estimators to enforce a certain degree of impartiality.

One such condition could be to restrict attention to unbiased estimators. An estimator $\boldsymbol{T}$ is said to be unbiased if,

$$\forall \boldsymbol{\theta} \in \Omega, \quad E[\boldsymbol{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] = \boldsymbol{\theta}$$

More generally we define the bias of an estimator $\boldsymbol{T}$ in estimating $\boldsymbol{\theta}$ by,

$$\text{bias}(\boldsymbol{T}) = E\Big[\boldsymbol{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)\Big] - \boldsymbol{\theta}$$

And so $\boldsymbol{T}$ is said to be unbiased if $\text{bias}(\boldsymbol{T}) = \boldsymbol{0}$.

3.3 Mean Squared Error Decomposition

As was mentioned above, the effectiveness of an estimator can be measured using the risk function.

In particular when,

$$\ell(\boldsymbol{\theta}, \boldsymbol{T}) = \|\boldsymbol{\theta} - \boldsymbol{T}\|^2 \quad \Longrightarrow \quad R(\boldsymbol{\theta}, \boldsymbol{T}) = E\Big[\|\boldsymbol{T} - \boldsymbol{\theta}\|^2\Big]$$

The risk function is referred to as the mean squared error (MSE) and this is the most commonly used measure of risk.

A common result for the MSE in the univariate case is that,

$$\text{MSE}(\theta, T) = \text{var}(T) + \text{bias}^2(T)$$

Proof.

$$\begin{aligned}
\text{MSE}(\theta, T) &= E\Big[(T-\theta)^2\Big] \\
&= E\Big[(T - E[T] + E[T] - \theta)^2\Big] \\
&= E\Big[(T - E[T])^2 + 2(T - E[T])(E[T] - \theta) + (E[T] - \theta)^2\Big] \\
&= E\Big[(T - E[T])^2\Big] + 2E\Big[(T - E[T])\Big](E[T] - \theta) + E\Big[(E[T] - \theta)^2\Big] \\
&= \text{var}(T) + \text{bias}^2(T)
\end{aligned}$$

(the middle term is zero since $E[T - E[T]] = 0$). □

This also extends to the multivariate case where we have that,

$$\text{MSE}(\boldsymbol{\theta}, \boldsymbol{T}) = \text{tr}(\text{cov}(\boldsymbol{T})) + \|\text{bias}(\boldsymbol{T})\|^2$$

The proof of this follows along almost exactly the same lines as the univariate case, using the fact that,

$$\|\boldsymbol{a} + \boldsymbol{b}\|^2 = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + 2\boldsymbol{a}^T\boldsymbol{b}$$

which is easy to prove from the definition of the Euclidean norm as an inner product.

Proof. Take $\boldsymbol{a} = \boldsymbol{T} - E[\boldsymbol{T}]$ and $\boldsymbol{b} = E[\boldsymbol{T}] - \boldsymbol{\theta}$. Then,

$$\begin{aligned}
\text{MSE}(\boldsymbol{\theta}, \boldsymbol{T}) &= E\Big[\|\boldsymbol{T} - E[\boldsymbol{T}]\|^2\Big] + 2E\Big[\boldsymbol{T} - E[\boldsymbol{T}]\Big]^T\Big(E[\boldsymbol{T}] - \boldsymbol{\theta}\Big) + \|E[\boldsymbol{T}] - \boldsymbol{\theta}\|^2 \\
&= E\Big[\,\text{tr}\,\Big((\boldsymbol{T} - E[\boldsymbol{T}])(\boldsymbol{T} - E[\boldsymbol{T}])^T\Big)\Big] + 0 + \|\text{bias}(\boldsymbol{T})\|^2 \\
&= \text{tr}(\text{cov}(\boldsymbol{T})) + \|\text{bias}(\boldsymbol{T})\|^2
\end{aligned}$$

using that a scalar is equal to its own trace and $\boldsymbol{a}^T\boldsymbol{a} = \text{tr}(\boldsymbol{a}\boldsymbol{a}^T)$. □

3.4 Comparing Unbiased Estimators

Notice from the MSE decomposition that when we restrict our attention to the class of unbiased estimators then

$$\text{MSE}(\theta, T) = \text{var}(T)$$

(or $\text{tr}(\text{cov}(\boldsymbol{T}))$ in the multivariate case).

Therefore, the estimator with the lowest MSE (i.e. lowest risk) will be the estimator with the lowest variance.

This then leads one quite naturally to search for unbiased estimators that have the smallest variance. We will come back to this idea later when looking at UMVU estimators (uniform minimum variance unbiased estimators).

3.5 Other Properties of Estimators

As we will see later there are many other properties of estimators that may make them desirable.

For example we may want the estimator to be:

- Unbiased
- Median-Unbiased
- Sufficient
- Complete
- Minimal sufficient
- Consistent
- Asymptotically Efficient

3.6 Finding Estimators

There are several methods for finding suitable estimators. For example suppose $X_1, \ldots, X_n$ are distributed symmetrically around θ. That is,

- $E[X] = \theta$
- median $= \theta$

Then one could take the estimate $\widehat{\theta}$ to be the value of θ that minimises,

$$\left(\sum_{j=1}^{n} |x_j - \theta|^{\alpha} \right)^{\frac{1}{\alpha}}$$

For different values of α this gives different solutions,

- $\alpha = 1$ (absolute error) leads to $\widehat{\theta} = q_{0.5}$ (sample median)
- $\alpha = 2$ (squared error) leads to $\widehat{\theta} = \overline{X}$ (sample mean)
- $\alpha \to 0$ leads to the mode in the limit
- $\alpha \to \infty$ leads to the mid-range (average of max and min)

Other more general approaches to finding estimators include:

- Method of Moments (MOM)
- Maximum Likelihood Estimation (MLE)

The latter is the primary method discussed in this course, but the method of moments is described briefly here.

3.6.1 Method of Moments

For a parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$, the method of moments equates the first d population moments with the corresponding sample moments,

$$E_{\boldsymbol{\theta}}[X^k] = \frac{1}{n} \sum_{j=1}^{n} x_j^k \qquad k = 1, \ldots, d$$

and takes $\widehat{\boldsymbol{\theta}}$ to be the solution of these d equations (see page 5 of the background notes). For example, for $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, the equations $E[X] = \mu$ and $E[X^2] = \sigma^2 + \mu^2$ give $\widehat{\mu} = \overline{x}$ and $\widehat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \overline{x})^2$, which are the same as the MLEs (Section 6.4).

4 Maximum Likelihood Estimation

To explain Maximum Likelihood (ML) Estimation we first need to introduce the concept of the likelihood function.

Suppose you have a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from some distribution $F_{\boldsymbol{\theta}}$. Then the likelihood function $L(\boldsymbol{\theta})$ is equal to the joint density of the random variables. That is,

$$L(\boldsymbol{\theta}) = f_{\boldsymbol{X}_1 \cdots \boldsymbol{X}_n}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

Notice that $L(\boldsymbol{\theta})$ is a function of both the unknown parameter and the data but it is generally thought about from the perspective of it being a function of $\boldsymbol{\theta}$. Say for example that you observe some data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, now the data is fixed and we can think of $L(\boldsymbol{\theta})$ as a measure of the relative likelihood that the data came from a distribution with the given value $\boldsymbol{\theta}$.

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{MLE}$ is then simply,

$$\widehat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta})$$

In the case where $L(\boldsymbol{\theta})$ has multiple local maxima we may sometimes refer to each of these points as ML estimators since they will share many of the desirable properties of the global maximum.

Given that log is a monotone increasing function, one typically maximises $\log L(\boldsymbol{\theta})$ since it is generally easier.

4.1 Notes on the Likelihood Function

In the case where the random sample is independent and identically distributed (i.i.d.) from a distribution with density $f(\boldsymbol{x}; \boldsymbol{\theta})$, the likelihood is given by the product of the individual densities such that,

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\boldsymbol{x}_j; \boldsymbol{\theta})$$

Given the likelihood function define,

$$\text{Score statistic:} \quad S(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\text{Fisher's information:} \quad \mathscr{I}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right]$$

Equating the score statistic to zero gives what is known as the likelihood equation,

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

(solving this equation gives you the MLEs)

4.2 Regularity Conditions

There are some nice results that hold under certain regularity conditions. Typical regularity conditions are that:

- the support of $f(\boldsymbol{x}; \boldsymbol{\theta})$ (the set of $\boldsymbol{x}$ where it is positive) does not depend on $\boldsymbol{\theta}$
- the parameter space Ω is an open set and $\log L(\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$
- differentiation with respect to $\boldsymbol{\theta}$ and integration over the data can be interchanged (as is done in the proofs below)

The conditions above are intended to illustrate the type of restrictions that are needed. They can be modified or perhaps relaxed to some extent depending on the specific result that is needed.

4.3 Results Under Regularity Conditions

The expected value of the score statistic is $\mathbf{0}$,

$$E\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$$

Consequently the covariance of the score is the Fisher's information,

$$\operatorname{cov}\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = \mathscr{I}(\boldsymbol{\theta})$$

Furthermore we can show that,

$$\mathscr{I}(\boldsymbol{\theta}) = E[\boldsymbol{I}(\boldsymbol{\theta})]$$

where,

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

which is the negative Hessian of the log-likelihood function.

Given the MLE $\widehat{\boldsymbol{\theta}}$, we call $\boldsymbol{I}(\widehat{\boldsymbol{\theta}})$ the observed information, and we will show later that in certain cases (the regular exponential family, Section 6) the relationship $\mathscr{I}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{I}(\widehat{\boldsymbol{\theta}})$ holds. This is often an easier way to compute the Fisher's information at the MLE since it avoids needing to compute expectations.

Proof. Since $L(\boldsymbol{\theta})$ is a density function,

$$\begin{aligned}
\int \cdots \int L(\boldsymbol{\theta})\, d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_n &= 1 \\
\int \cdots \int \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_n &= 0 \\
\int \cdots \int \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})\, d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_n &= 0 \qquad (1) \\
E\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] &= \mathbf{0}
\end{aligned}$$

The second result then follows trivially from this since,

$$\begin{aligned}
\operatorname{cov}\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) &= E\left[\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] - E\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] E\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \\
&= E\left[\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] \\
&= \mathscr{I}(\boldsymbol{\theta})
\end{aligned}$$

To get the last result we continue from equation (1) taking partial derivatives again,

$$\begin{aligned}
\int\cdots\int \frac{\partial}{\partial \boldsymbol{\theta}^T}\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})\right) d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_n &= 0 \\
\int\cdots\int \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} L(\boldsymbol{\theta}) + \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\, d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_n &= 0 \\
\int\cdots\int \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} L(\boldsymbol{\theta}) + \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} L(\boldsymbol{\theta})\, d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_n &= 0 \\
E\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} + \left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] &= 0 \\
E\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right] + \mathscr{I}(\boldsymbol{\theta}) &= 0 \\
\mathscr{I}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right] & \\
= E[\boldsymbol{I}(\boldsymbol{\theta})] &
\end{aligned}$$

□

5 The Cramér–Rao Lower Bound

5.1 Univariate Parameter Case

Suppose you have some estimator T that is unbiased in estimating $g(\theta)$. Under the regularity conditions of Section 4, the Cramér–Rao lower bound (CRLB) imposes a lower bound on the variance of such an estimator T such that,

$$\operatorname{var}(T) \geq \frac{[g'(\theta)]^2}{\mathscr{I}(\theta)}$$

Proof. Suppose T is unbiased in estimating $g(\theta)$. This implies that,

$$\begin{aligned}
E[T] &= g(\theta) \\
\int\cdots\int T L(\theta)\, dx_1 \ldots dx_n &= g(\theta) \\
\int\cdots\int T \frac{\partial L(\theta)}{\partial \theta}\, dx_1 \ldots dx_n &= g'(\theta) \\
\int\cdots\int T \frac{\partial \log L(\theta)}{\partial \theta} L(\theta)\, dx_1 \ldots dx_n &= g'(\theta) \\
E\left[T \frac{\partial \log L(\theta)}{\partial \theta}\right] &= g'(\theta) \\
\operatorname{cov}\left(T,\ \frac{\partial \log L(\theta)}{\partial \theta}\right) &= g'(\theta) \\
\operatorname{var}(T) \operatorname{var}\left(\frac{\partial \log L(\theta)}{\partial \theta}\right) &\geq [g'(\theta)]^2 \\
\operatorname{var}(T) &\geq \frac{[g'(\theta)]^2}{\mathscr{I}(\theta)}
\end{aligned}$$

Here the expectation becomes a covariance because the score has mean zero (Section 4), the inequality is the Cauchy–Schwarz inequality for random variables (Section 1), and var of the score is $\mathscr{I}(\theta)$. □

5.2 Multivariate Parameter Case

Let $\boldsymbol{\theta}$ be a d-dimensional vector of parameters and let $\boldsymbol{T}$ be a d-dimensional statistic with,

$$E[\boldsymbol{T}] = \boldsymbol{g}(\boldsymbol{\theta})$$

The covariance matrix of $\boldsymbol{T}$ satisfies the inequality,

$$\operatorname{cov}(\boldsymbol{T}) \succeq \boldsymbol{g}'(\boldsymbol{\theta}) \mathscr{I}^{-1}(\boldsymbol{\theta}) \boldsymbol{g}'(\boldsymbol{\theta})^T$$

Notice here that:

- $\operatorname{cov}(\boldsymbol{T})$ is a d by d matrix
- $\boldsymbol{g}'(\boldsymbol{\theta})$ is a d by d Jacobian matrix
- $\mathscr{I}^{-1}(\boldsymbol{\theta})$ is a d by d matrix
- $\boldsymbol{A} \succeq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite. I.e. that $\boldsymbol{x}^T(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x} \geq 0$ for all vectors $\boldsymbol{x}$

The proof of this is not covered in this course.

5.3 Example: Normal Distribution with Known Mean

Suppose,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

where μ is known but σ^2 is unknown (i.e. $\theta = \sigma^2$). The kernel of the log likelihood is given by,

$$\log L(\theta) = \frac{-n}{2} \log(\theta) - \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\theta}$$

(kernel means the log likelihood without all of the parts that don't depend on θ).

The score statistic is thus,

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{-n}{2\theta} + \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\theta^2}$$

Solving the likelihood equation gives us the MLE for θ,

$$\begin{aligned}
\frac{n}{2\widehat{\theta}} &= \frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\widehat{\theta}^2} \\
\frac{n}{\widehat{\theta}} &= \frac{1}{\widehat{\theta}^2} \sum_{j=1}^{n} (x_j - \mu)^2 \\
\widehat{\theta} &= \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu)^2
\end{aligned}$$

We can further show that the variance of $\widehat{\theta}$ attains the CRLB. We can show this by computing its variance explicitly,

$$\begin{aligned}
\frac{n}{\sigma^2}\widehat{\theta} \sim \chi^2_n \quad \implies \quad \operatorname{var}\left(\frac{n}{\sigma^2}\widehat{\theta}\right) &= 2n \\
\frac{n^2}{\sigma^4} \operatorname{var}(\widehat{\theta}) &= 2n \\
\operatorname{var}(\widehat{\theta}) &= \frac{2\sigma^4}{n}
\end{aligned}$$

Now we also need to compute the Fisher information,

$$\begin{aligned} I(\theta) &= -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\ &= -\frac{\partial}{\partial \theta}\Big(\frac{-n}{2\theta} + \frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\theta^2}\Big) \\ &= \frac{-n}{2\theta^2} + \frac{1}{\theta^3}\sum_{j=1}^{n}(x_j-\mu)^2 \\ \mathscr{I}(\theta) &= E[I(\theta)] \\ &= E\left[\frac{-n}{2\theta^2} + \frac{1}{\theta^3}\sum_{j=1}^{n}(x_j-\mu)^2\right] \\ &= \frac{-n}{2\theta^2} + \frac{1}{\theta^3}\sum_{j=1}^{n}E\Big[(X_j-\mu)^2\Big] \\ &= \frac{-n}{2\theta^2} + \frac{n\theta}{\theta^3} \\ &= \frac{n}{2\theta^2} \\ &= \frac{n}{2\sigma^4} \end{aligned}$$

Notice that,

$$\mathrm{var}(\widehat{\theta}) = \frac{1}{\mathscr{I}(\theta)}$$

and so $\widehat{\theta}$ attains the CRLB.

5.4 Example: Normal Distribution

Suppose now that both μ and σ^2 are unknown. The parameter vector is now therefore 2-dimensional,

$$\boldsymbol{\theta} = (\mu, \sigma^2)^T$$

Now consider the estimator $\boldsymbol{T}$ given by,

$$\boldsymbol{T} = (\overline{X}, s^2)^T$$

This estimator is unbiased in estimating $\boldsymbol{\theta}$. I.e.

$$E[\boldsymbol{T}] = \boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$$

Therefore,

$$\boldsymbol{g}'(\boldsymbol{\theta}) = \boldsymbol{I}_2$$

Now to see if $\boldsymbol{T}$ attains the CRLB we need the Fisher information $\mathscr{I}(\boldsymbol{\theta})$ and $\mathrm{cov}(\boldsymbol{T})$. The Fisher information (computed in tutorial sheet 1) is,

$$\mathscr{I}(\boldsymbol{\theta}) = \mathrm{diag}\left(\frac{n}{\sigma^2}, \frac{n}{2\sigma^4}\right)$$

The covariance on the other hand is given by,

$$\begin{aligned} \mathrm{cov}(\boldsymbol{T}) &= \begin{pmatrix} \mathrm{var}(\overline{X}) & \mathrm{cov}(\overline{X}, s^2) \\ \mathrm{cov}(s^2, \overline{X}) & \mathrm{var}(s^2) \end{pmatrix} \\ &= \mathrm{diag}\Big(\mathrm{var}(\overline{X}),\ \mathrm{var}(s^2)\Big) \end{aligned}$$

(the off-diagonals are zero because $\overline{X}$ and s^2 are independent, as shown in Section 2.5).

Now consider the variance of $\overline{X}$,

$$\begin{aligned}\operatorname{var}(\overline{X}) &= \operatorname{var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\operatorname{var}(X_j) \qquad \text{(by independence)} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Now looking at s^2,

$$\begin{aligned}\frac{n-1}{\sigma^2}s^2 \sim \chi^2_{n-1} \quad \implies \quad \operatorname{var}\left(\frac{n-1}{\sigma^2}s^2\right) &= 2(n-1) \\ \left(\frac{n-1}{\sigma^2}\right)^2 \operatorname{var}(s^2) &= 2(n-1) \\ \operatorname{var}(s^2) &= \frac{2(n-1)\sigma^4}{(n-1)^2} \\ &= \frac{2\sigma^4}{n-1}\end{aligned}$$

And so,

$$\operatorname{cov}(\boldsymbol{T}) = \operatorname{diag}\left(\frac{\sigma^2}{n},\ \frac{2\sigma^4}{n-1}\right)$$

The CRLB here states that,

$$\begin{aligned}\operatorname{cov}(\boldsymbol{T}) &\succeq \boldsymbol{g}'(\boldsymbol{\theta})\mathscr{I}(\boldsymbol{\theta})^{-1}\boldsymbol{g}'(\boldsymbol{\theta})^T \\ \operatorname{diag}\left(\frac{\sigma^2}{n},\ \frac{2\sigma^4}{n-1}\right) &\succeq \operatorname{diag}\left(\frac{\sigma^2}{n},\ \frac{2\sigma^4}{n}\right)\end{aligned}$$

Notice here however that equality is not achieved. This is because the penalty incurred in not knowing the mean increases the variance of the unbiased estimator.

5.5 Example: Using the MLE Instead

Now let's try the exact same process as above but using the MLE, given by,

$$\boldsymbol{T} = (\overline{X}, \widehat{\sigma}^2)^T$$

This estimator is no longer unbiased for $\boldsymbol{\theta}$ since,

$$\begin{aligned}\frac{n}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_{n-1} \quad \implies \quad E\left[\frac{n}{\sigma^2}\widehat{\sigma}^2\right] &= n-1 \\ E[\widehat{\sigma}^2] &= \frac{n-1}{n}\sigma^2\end{aligned}$$

Therefore,

$$\begin{aligned}E[\boldsymbol{T}] &= \boldsymbol{g}(\boldsymbol{\theta}) \\ &= \left(\mu, \frac{n-1}{n}\sigma^2\right)^T\end{aligned}$$

So $\boldsymbol{T}$ is not unbiased for $\boldsymbol{\theta}$ but trivially it is unbiased for $\boldsymbol{g}(\boldsymbol{\theta})$.

Computing $\text{cov}(\boldsymbol{T})$ we again use the fact that $\overline{X}$ and $\widehat{\sigma}^2$ are independent to justify why the off-diagonal terms are zero. The first diagonal term is then $\frac{\sigma^2}{n}$ by the same logic as above and so the only term that needs to be computed is the second diagonal term which is $\text{var}(\widehat{\sigma}^2)$.

$$\frac{n}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_{n-1} \quad \implies \quad \text{var}\left(\frac{n}{\sigma^2}\widehat{\sigma}^2\right) = 2(n-1)$$
$$\text{var}(\widehat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2}$$

Thus the covariance is given by,

$$\text{cov}(\boldsymbol{T}) = \text{diag}\left(\frac{\sigma^2}{n}, \frac{2(n-1)\sigma^4}{n^2}\right)$$

We can then also compute the lower bound on $\text{cov}(\boldsymbol{T})$ as follows,

$$\begin{aligned}\boldsymbol{g}'(\boldsymbol{\theta})\mathscr{I}(\boldsymbol{\theta})^{-1}\boldsymbol{g}'(\boldsymbol{\theta})^T &= \begin{pmatrix}1 & 0\\ 0 & \frac{n-1}{n}\end{pmatrix}\begin{pmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{n}{2\sigma^4}\end{pmatrix}^{-1}\begin{pmatrix}1 & 0\\ 0 & \frac{n-1}{n}\end{pmatrix}^T \\ &= \text{diag}\left(\frac{\sigma^2}{n}, \frac{2(n-1)^2\sigma^4}{n^3}\right) \\ &= \text{diag}\left(\frac{\sigma^2}{n}, \frac{2(n-1)\sigma^4}{n^2}\frac{(n-1)}{n}\right)\end{aligned}$$

Which we can notice is close to the covariance of $\boldsymbol{T}$ but is off by the factor of $\frac{n-1}{n}$ in the second diagonal entry. Therefore this estimator also does not achieve the CRLB.

6 The Exponential Family

6.1 Regular Exponential Family

Suppose you have a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ where $\boldsymbol{x}_i$ is a p-dimensional random vector coming from a distribution parameterised by a d-dimensional parameter $\boldsymbol{\theta}$.

The likelihood function $L(\boldsymbol{\theta})$ is part of the exponential family if it can be written in the form,

$$L(\boldsymbol{\theta}) = b(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\exp\left(\boldsymbol{c}(\boldsymbol{\theta})^T\boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\right) \,/\, a(\boldsymbol{\theta})$$

Where:

- $\boldsymbol{T}$ is a q-dimensional sufficient statistic ($q \geq d$)
- $\boldsymbol{c}(\boldsymbol{\theta})$ is a q-dimensional vector

Under certain conditions we refer to the likelihood as being part of the "**regular**" exponential family. These conditions are that:

- $q = d$ (i.e. the dimension of $\boldsymbol{T}$ is equal to the dimension of $\boldsymbol{\theta}$)
- $\boldsymbol{J_c} = \boldsymbol{\nabla c}^T = \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{\theta}}$ is of full rank

In such cases we refer to $\boldsymbol{c}(\boldsymbol{\theta})$ as the natural or canonical parameter, and it is unique up to non-singular linear transformations. The same goes for $\boldsymbol{T}$: for example one could take,

$$\left[\boldsymbol{A}^{-T}\boldsymbol{c}(\boldsymbol{\theta})\right]^T\left[\boldsymbol{AT}\right] = \boldsymbol{c}(\boldsymbol{\theta})^T\boldsymbol{T}$$

for any non-singular $\boldsymbol{A}$. The statistic $\boldsymbol{T}$ is a complete sufficient statistic (these terms are defined in Section 7).

6.2 Likelihood Equation for the REF

In the case of the REF, the score statistic can be determined as follows,

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \log\Big(b(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\Big) + \boldsymbol{c}(\boldsymbol{\theta})^T\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - \log\Big(a(\boldsymbol{\theta})\Big)\\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - \frac{\partial \log a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \qquad (1)\end{aligned}$$

We can then take the expectation of both sides and recall that under regularity conditions the expected value of the score statistic is $\mathbf{0}$ (Section 4).

$$\begin{aligned}\mathbf{0} &= E\left[\frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - \frac{\partial \log a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\\ \frac{\partial \log a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}E[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]\end{aligned}$$

By combining this equation with equation (1) we get the following,

$$\begin{aligned}\frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}E\Big[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)\Big]\\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}E\Big[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)\Big]\end{aligned}$$

Setting the score statistic equal to $\mathbf{0}$ yields,

$$\begin{aligned}\mathbf{0} &= \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}E\Big[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)\Big]\\ &= \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big(\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - E[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]\Big)\end{aligned}$$

Given that $L(\boldsymbol{\theta})$ is in the **regular** exponential family, the $d \times d$ matrix $\frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is of full rank, which implies that the only vector that it maps to zero is the zero vector. I.e.,

$$\begin{aligned}\mathbf{0} &= \boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - E[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]\\ E[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)] &= \boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\end{aligned}$$

And so the likelihood equation can be solved by equating the expected value of the statistic $\boldsymbol{T}$ with its observed value (a similar idea to the method of moments).

6.3 Expected Fisher Information

Given the MLE $\widehat{\boldsymbol{\theta}}$, the expected Fisher information $\mathscr{I}(\widehat{\boldsymbol{\theta}})$ for data coming from a regular exponential family can be calculated by,

$$\mathscr{I}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{I}(\widehat{\boldsymbol{\theta}}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

To show this we have from equation (1) above that,

$$\begin{aligned}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - \frac{\partial \log a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\\ \boldsymbol{I}(\boldsymbol{\theta}) &= -\frac{\partial^2 \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}\boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) + \frac{\partial^2 \log a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}\end{aligned}$$

Taking the expected value of both sides then yields,

$$\mathscr{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \boldsymbol{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}E\Big[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)\Big] + \frac{\partial^2 \log a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}$$

(since under regularity conditions $\mathscr{I}(\boldsymbol{\theta}) = E[\boldsymbol{I}(\boldsymbol{\theta})]$, Section 4)

However notice that when $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$,

$$E\big[\boldsymbol{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)\big] = \boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

And so this implies that,

$$\mathscr{I}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{I}(\widehat{\boldsymbol{\theta}})$$

6.4 MLEs of the Normal Distribution via the REF

This subsection covers the MLE for:

- Univariate normal distribution
- Multivariate normal distribution

Here we use the fact that the normal distribution comes from the regular exponential family and so its MLEs can be calculated without the need to differentiate the likelihood function. This is especially useful in the multivariate case since otherwise you would need to differentiate with respect to vectors and matrices which can be very difficult. (since $\boldsymbol{\mu}$ is a vector and $\boldsymbol{\Sigma}$ is a matrix).

(The lecture notes derive the multivariate case slightly differently, but with the same idea and the same end result.)

6.4.1 Univariate Normal Distribution

First we start by looking at the univariate normal distribution.

Suppose we have a random sample,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

Let $\boldsymbol{\theta} = (\mu, \sigma^2)$ be the unknown parameter.

The likelihood function can then be determined as follows,

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{n} (2\pi)^{\frac{-1}{2}} (\sigma^2)^{\frac{-1}{2}} \exp\left(\frac{-1}{2\sigma^2}(x_j - \mu)^2\right) \\ &= (2\pi)^{\frac{-n}{2}} (\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2\right) \end{aligned}$$

To then find the MLE, $\widehat{\boldsymbol{\theta}}$, there are two approaches that we could take:

- Solve the likelihood equation directly (i.e. solve $\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \mathbf{0}$)
- Show that $L(\boldsymbol{\theta})$ belongs to the regular exponential family

Whilst both routes work taking the second approach using the regular exponential family is generally easier (especially when it comes to the multivariate case which we will do later).

Recall that to be in the regular exponential family the likelihood function must be in the form,

$$L(\boldsymbol{\theta}) = b(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \exp\Big(\boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\Big) / a(\boldsymbol{\theta})$$

together with the conditions given at the start of this section.

In this case we can manipulate $L(\boldsymbol{\theta})$ to be in this form.

$$\begin{aligned}
L(\boldsymbol{\theta}) &= (2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right) \\
&= (2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n}\left(x_j^2-2x_j\mu+\mu^2\right)\right) \\
&= (2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma^2}\left(\sum_{j=1}^{n}x_j^2-2\mu\sum_{j=1}^{n}x_j+n\mu^2\right)\right) \\
&= (2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n}x_j^2+\frac{\mu}{\sigma^2}\sum_{j=1}^{n}x_j-\frac{n\mu^2}{2\sigma^2}\right) \\
&= (2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{\mu}{\sigma^2}\sum_{j=1}^{n}x_j-\frac{1}{2\sigma^2}\sum_{j=1}^{n}x_j^2\right) / \exp\left(\frac{n\mu^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{\mu}{\sigma^2}\sum_{j=1}^{n}x_j-\frac{1}{2\sigma^2}\sum_{j=1}^{n}x_j^2\right) / \left(\exp\left(\frac{n\mu^2}{2\sigma^2}\right)(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}\right)
\end{aligned}$$

Which is now in the form of the regular exponential family where,

$$\begin{aligned}
b(x_1,\ldots,x_n) &= 1 \qquad\qquad \text{(and } q=d=2\text{, with } \partial\boldsymbol{c}/\partial\boldsymbol{\theta} \text{ of full rank)} \\
\boldsymbol{c}(\boldsymbol{\theta}) &= \left(\frac{\mu}{\sigma^2}, \quad \frac{-1}{2\sigma^2}\right)^T \\
\boldsymbol{T}(x_1,\ldots,x_n) &= \left(\sum_{j=1}^{n}x_j, \quad \sum_{j=1}^{n}x_j^2\right)^T \\
a(\boldsymbol{\theta}) &= \exp\left(\frac{n\mu^2}{2\sigma^2}\right)(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}
\end{aligned}$$

As shown above, this implies that the MLEs can be found by solving,

$$E_{\widehat{\boldsymbol{\theta}}}\Big(\boldsymbol{T}(X_1,\ldots,X_n)\Big) = \boldsymbol{T}(x_1,\ldots,x_n)$$

First equating $\boldsymbol{T}_1$ we get that,

$$\begin{aligned}
E_{\widehat{\boldsymbol{\theta}}}\left(\sum_{j=1}^{n}X_j\right) &= \sum_{j=1}^{n}x_j \\
\sum_{j=1}^{n}E_{\widehat{\boldsymbol{\theta}}}(X_j) &= \sum_{j=1}^{n}x_j \\
n\widehat{\mu} &= \sum_{j=1}^{n}x_j \\
\widehat{\mu} &= \frac{1}{n}\sum_{j=1}^{n}x_j \\
&= \overline{x}
\end{aligned}$$

Then from $\boldsymbol{T}_2$ we get,

$$\begin{aligned}
E_{\widehat{\boldsymbol{\theta}}}\left(\sum_{j=1}^{n} X_j^2\right) &= \sum_{j=1}^{n} x_j^2 \\
\sum_{j=1}^{n} E_{\widehat{\boldsymbol{\theta}}}(X_j^2) &= \sum_{j=1}^{n} x_j^2 \\
n(\widehat{\sigma}^2 + \widehat{\mu}^2) &= \sum_{j=1}^{n} x_j^2 \\
n\widehat{\sigma}^2 &= \sum_{j=1}^{n} (x_j^2) - n\widehat{\mu}^2 \\
&= \sum_{j=1}^{n} (x_j^2) - n\overline{x}^2 \\
&= \sum_{j=1}^{n} (x_j)^2 - 2n\overline{x}^2 + n\overline{x}^2 \\
&= \sum_{j=1}^{n} (x_j)^2 - 2\overline{x}\sum_{j=1}^{n}(x_j) + \sum_{j=1}^{n}\overline{x}^2 \\
&= \sum_{j=1}^{n} (x_j - \overline{x})^2 \\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{j=1}^{n} (x_j - \overline{x})^2
\end{aligned}$$

And so the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma}^2)$ for data coming from a univariate normal distribution is,

$$\widehat{\boldsymbol{\theta}} = \left(\overline{x}, \quad \frac{1}{n}\sum_{j=1}^{n} (x_j - \overline{x})^2\right)$$

6.4.2 Multivariate Normal Distribution

Now suppose that we instead have a random sample from a multivariate normal distribution.

That is,

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Let $\boldsymbol{\theta}$ be the vector of $d = p + \frac{1}{2}p(p+1)$ dimensions containing the p elements in the $\boldsymbol{\mu}$ vector and the $\frac{1}{2}p(p+1)$ distinct elements of the $\boldsymbol{\Sigma}$ matrix (you can show that this is the case by counting the number of distinct elements with an arithmetic sequence).

The likelihood function $L(\boldsymbol{\theta})$ is below,

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{j=1}^{n} (2\pi)^{\frac{-p}{2}} |\boldsymbol{\Sigma}|^{\frac{-1}{2}} \exp\left(\frac{-1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu})\right) \\
&= (2\pi)^{\frac{-np}{2}} |\boldsymbol{\Sigma}|^{\frac{-n}{2}} \exp\left(\frac{-1}{2}\sum_{j=1}^{n} (\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu})\right)
\end{aligned}$$

What we want to do now is expand out the quadratic form (like what we did in the univariate

case). This can be done as follows,

$$\begin{aligned}\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}) &= \sum_{j=1}^{n}(\boldsymbol{x}_j^T-\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})\\ &= \sum_{j=1}^{n}(\boldsymbol{x}_j^T\boldsymbol{\Sigma}^{-1}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1})(\boldsymbol{x}_j-\boldsymbol{\mu})\\ &= \sum_{j=1}^{n}(\boldsymbol{x}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j-\boldsymbol{x}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\end{aligned}$$

The two central terms can actually combine here since for any symmetric matrix $\boldsymbol{M}$ we have the property that,

$$\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{b}=\boldsymbol{b}^T\boldsymbol{M}\boldsymbol{a}$$

(since $\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{b}$ is a scalar and any scalar is equal to its own transpose, $\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{b}=\boldsymbol{b}^T\boldsymbol{M}^T\boldsymbol{a}=\boldsymbol{b}^T\boldsymbol{M}\boldsymbol{a}$).

Because of this result we have that $\boldsymbol{x}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}=\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j$ and so,

$$\begin{aligned}\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}) &= \sum_{j=1}^{n}(\boldsymbol{x}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j-2\boldsymbol{x}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\\ &= \sum_{j=1}^{n}(\boldsymbol{x}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j)-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j)+n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\end{aligned}$$

We have now made good progress in separating out the quadratic form however notice that the first term in the sum above is still a joint function of the data and the unknown parameters. To split it up though we can use another useful matrix property which is that,

$$a^TMa=\mathrm{tr}(Maa^T)$$

(this follows a similar idea to the previous property: a scalar is equal to its own trace, so $\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{a}=\mathrm{tr}(\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{a})=\mathrm{tr}(\boldsymbol{M}\boldsymbol{a}\boldsymbol{a}^T)$ by the cyclic property of the trace).

Using this and the fact that the trace is a linear operator we can do the following,

$$\begin{aligned}\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}) &= \sum_{j=1}^{n}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^T)-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j)+n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ &= \mathrm{tr}\left(\sum_{j=1}^{n}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j)+n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ &= \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j)+n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\end{aligned}$$

By doing this we have now separated the parameters from the data. Whilst they don't look fully separated in this form with the trace wrapped around both of them if you expand out the operation that is being performed you will see that each individual term in the sum can be decomposed as a function of the parameters multiplied by a function of the data.

In fact you may recall that for two matrices A and B the matrix inner product is typically defined as,

$$\begin{aligned}\langle A,B\rangle &= \mathrm{tr}(A^TB)\\ &= \mathrm{vec}(A)^T\,\mathrm{vec}(B)\\ &= \mathrm{vec}(A)\cdot\mathrm{vec}(B)\end{aligned}$$

Where here vec$(\cdot)$ just means to flatten out a matrix into a column vector.

Rewriting the expression in this way makes it clearer how the term factorises into a linear combination of parameter and data components.

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right) &= \left\langle (\boldsymbol{\Sigma}^{-1})^T, \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right\rangle \\ &= \left\langle \boldsymbol{\Sigma}^{-1}, \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right\rangle \\ &= \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)^T \operatorname{vec}\left(\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)\end{aligned}$$

And so putting this all back into the likelihood function we get the following,

$$\begin{aligned}L(\boldsymbol{\theta}) &= (2\pi)^{\frac{-np}{2}}|\boldsymbol{\Sigma}|^{\frac{-n}{2}}\exp\left(\frac{-1}{2}\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)^T\operatorname{vec}\left(\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j)-\frac{n}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\ &= \exp\left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j)-\frac{1}{2}\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)^T\operatorname{vec}\left(\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)\right) \\ &\quad /\left((2\pi)^{\frac{np}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}\exp\left(\frac{n}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)\end{aligned}$$

From here it is quite clear how this fits in the regular exponential family. However I am going to revert back to using a matrix inner product as opposed to flattening out the matrices into vectors. Mathematically they are identical, so it is more of a stylistic difference, but in general you lose structure when flattening out matrices, which is why keeping it as a matrix is preferred.

Using this the likelihood function has the following form,

$$L(\boldsymbol{\theta}) = \exp\left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j)+\left\langle\frac{-1}{2}\boldsymbol{\Sigma}^{-1},\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right\rangle\right)/\left((2\pi)^{\frac{np}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}\exp\left(\frac{n}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)$$

Which matches the form of the regular exponential family with,

$$\begin{aligned}b(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) &= 1 \\ \boldsymbol{c}(\boldsymbol{\theta}) &= \left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1},\quad \frac{-1}{2}\boldsymbol{\Sigma}^{-1}\right) \\ \boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) &= \left(\sum_{j=1}^{n}\boldsymbol{x}_j,\quad \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)^T \\ a(\boldsymbol{\theta}) &= (2\pi)^{\frac{np}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}\exp\left(\frac{n}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\end{aligned}$$

One small pedantic note. Typically one takes the first term in the canonical parameter to be $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ instead of $\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}$ as I did. These have the same entries (one is the transpose of the other). This follows from the fact that $\boldsymbol{\Sigma}$ is symmetric and the inner product is commutative (so that $\boldsymbol{a}^T\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{a}$).

Now that this has been established we can solve for the MLE, $\widehat{\boldsymbol{\theta}}$, using again the fact that,

$$E_{\widehat{\boldsymbol{\theta}}}[\boldsymbol{T}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)] = \boldsymbol{T}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$$

First equating $\boldsymbol{T}_1$ we get,

$$\begin{aligned} E_{\widehat{\boldsymbol{\theta}}}\left(\sum_{j=1}^{n}\boldsymbol{X}_j\right) &= \sum_{j=1}^{n}\boldsymbol{x}_j \\ n\widehat{\boldsymbol{\mu}} &= \sum_{j=1}^{n}\boldsymbol{x}_j \\ \widehat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j \\ &= \overline{\boldsymbol{x}} \end{aligned}$$

And then equating $\boldsymbol{T}_2$,

$$\begin{aligned} E_{\widehat{\boldsymbol{\theta}}}\left(\sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{X}_j^T\right) &= \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T \\ \sum_{j=1}^{n}E_{\widehat{\boldsymbol{\theta}}}\left(\boldsymbol{X}_j\boldsymbol{X}_j^T\right) &= \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T \end{aligned}$$

Here we can use the fact that $\boldsymbol{\Sigma} = E(\boldsymbol{X}_j\boldsymbol{X}_j^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$ (Section 1) to rewrite the above expression as,

$$\begin{aligned} n(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T) &= \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T \\ n\widehat{\boldsymbol{\Sigma}} &= \sum_{j=1}^{n}(\boldsymbol{x}_j\boldsymbol{x}_j^T) - n\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{x}_j\boldsymbol{x}_j^T - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\right) \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{x}_j\boldsymbol{x}_j^T - \overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T\right) \end{aligned}$$

This is equivalent to the more standard form for the sample covariance matrix as I show below,

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})^T \\ n\widehat{\boldsymbol{\Sigma}} &= \sum_{j=1}^{n}\left(\boldsymbol{x}_j\boldsymbol{x}_j^T - \boldsymbol{x}_j\overline{\boldsymbol{x}}^T - \overline{\boldsymbol{x}}\boldsymbol{x}_j^T + \overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T\right) \\ &= \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T - \sum_{j=1}^{n}\boldsymbol{x}_j\overline{\boldsymbol{x}}^T - \sum_{j=1}^{n}\overline{\boldsymbol{x}}\boldsymbol{x}_j^T + \sum_{j=1}^{n}\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T \\ &= \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T - \left(\sum_{j=1}^{n}\boldsymbol{x}_j\right)\overline{\boldsymbol{x}}^T - \overline{\boldsymbol{x}}\left(\sum_{j=1}^{n}\boldsymbol{x}_j^T\right) + n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T \\ &= \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T - n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T - n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T + n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T \\ &= \sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T - n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T \\ &= \sum_{j=1}^{n}\left(\boldsymbol{x}_j\boldsymbol{x}_j^T - \overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T\right) \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{x}_j\boldsymbol{x}_j^T - \overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T\right) \end{aligned}$$

And so the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ for data coming from a multivariate normal distribution is,

$$\widehat{\boldsymbol{\theta}} = \left(\overline{\boldsymbol{x}}, \quad \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})^T \right)$$

7 Properties of Estimators

7.1 Minimum Variance Bound (MVB) Unbiased Estimators

An unbiased estimator $\boldsymbol{T}$ of $\boldsymbol{g}(\boldsymbol{\theta})$ is said to be MVB if it attains the CRLB. I.e.,

$$\operatorname{cov}(\boldsymbol{T}) = \boldsymbol{g}'(\boldsymbol{\theta}) \mathscr{I}(\boldsymbol{\theta})^{-1} \boldsymbol{g}'(\boldsymbol{\theta})^T$$

For the univariate case this implies that,

$$\operatorname{var}(T) = \frac{[g'(\theta)]^2}{\mathscr{I}(\theta)}$$

7.1.1 Necessary and sufficient condition for MVB

Recall that in the proof of the CRLB for the univariate case (Section 5) we applied the Cauchy–Schwarz inequality to,

$$\operatorname{cov}\left(T, \ \frac{\partial \log L(\theta)}{\partial \theta} \right)^2 \leq \operatorname{var}(T) \operatorname{var}\left(\frac{\partial \log L(\theta)}{\partial \theta} \right)$$

To make this clearer I am going to rewrite the expression above in terms of random variable inner products. Also for simplicity let $s(\theta)$ be the score statistic. Then the expression can be written as,

$$\Big\langle T - E[T], \ s - E[s] \Big\rangle^2 \leq \Big\langle T - E[T], \ T - E[T] \Big\rangle \Big\langle s - E[s], \ s - E[s] \Big\rangle$$

This comes from the definition of the covariance as a "centralised" inner product. Now recall that equality is achieved in the Cauchy–Schwarz inequality if and only if one variable is a scalar multiple of the other. I.e.,

$$s - E[s] = k(\theta) \Big(T - E[T] \Big)$$

where here this constant is $k(\theta)$.

We can write this in a nicer form though by realising that:

- $E[T] = g(\theta)$
- $E[s] = 0$ (under regularity conditions)

Thus,

$$\frac{\partial \log L(\theta)}{\partial \theta} = k(\theta)(T - g(\theta)) \qquad (1)$$

Comments:

- This equation above is a necessary and sufficient condition for an estimator to be MVB
- The course notes justify this formula differently, but I find the equality case of Cauchy–Schwarz the most natural way to see where it comes from.

7.1.2 Relation to the Exponential Family

From (1) above we have a necessary and sufficient condition for attaining the CRLB. Integrating both sides with respect to θ leads to a likelihood of the form,

$$\log L(\theta) = T \int k(\theta)\, d\theta - \int k(\theta) g(\theta)\, d\theta + \underbrace{b(x_1, \ldots, x_n)}_{\text{constant of integration}}$$

$$L(\theta) = \exp(b) \exp\left(T \int k(\theta)\, d\theta\right) / \exp\left(\int k(\theta) g(\theta)\, d\theta\right)$$

This fits the form of the 1-parameter exponential family exactly. And in fact the reverse works as well (i.e. starting with the exponential family and differentiating to form the score statistic).

Therefore we have an important result,

$$T \text{ attains the CRLB for } g(\theta) \iff L(\theta) \text{ is a 1-parameter exponential family with statistic } T$$

Comments:

- This is one of the reasons why the form of the regular exponential family is so important
- This is presumably also why the course writes the exponential family with the division by $a(\theta)$
- The results also extend to the multivariate case.

7.2 Uniform Minimum Variance Unbiased Estimators (UMVUE)

A UMVU estimator is an unbiased estimator that has the lowest variance of all unbiased estimators, whatever the true parameter value θ is.

I.e. U is a UMVUE if:

- $E[U] = g(\theta)$
- For all other estimators T with $E[T] = g(\theta)$, $\text{var}(U) \leq \text{var}(T)$

We will show later with the Rao–Blackwell and Lehmann–Scheffé theorems how one can form UMVU estimators.

Also, if U is MVB then U is the UMVUE (since no unbiased estimator can have a variance below the CRLB).

7.3 Sufficiency

Suppose we have a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ coming from a distribution with a joint density function $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta})$.

A statistic $\boldsymbol{T}$ is said to be sufficient for $\boldsymbol{\theta}$ if:

- It captures all the information about $\boldsymbol{\theta}$ that exists in the data
- Formally, the distribution of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \,|\, \boldsymbol{T}$ must not depend on $\boldsymbol{\theta}$
- Or in other words the distribution of the data given the statistic does not depend on the unknown parameter.

7.3.1 Fisher–Neyman Factorisation Theorem

The Fisher–Neyman factorisation theorem gives us a convenient way to check if a statistic $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$. It says that this will be the case if and only if the joint density can be factorised

as follows,

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) = h_1(\boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n); \boldsymbol{\theta}) h_2(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

I.e. the joint density (likelihood function) factorises into:

- A function of $\boldsymbol{T}$ and $\boldsymbol{\theta}$
- A function of the data

7.3.2 Relation to Exponential Family

Looking at the likelihood function in the regular exponential family notice that,

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) = \underbrace{b(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}_{h_2} \underbrace{\exp\left(\boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{T}\right) / a(\boldsymbol{\theta})}_{h_1}$$

And so by the Fisher–Neyman factorisation theorem this implies that,

$$\boldsymbol{T} \text{ is a sufficient statistic}$$

Notice however that belonging to the exponential family is not a strictly necessary condition for a statistic to be sufficient.

7.3.3 Minimal Sufficiency

One "problem" with sufficiency is that it is not unique. For example suppose that $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ and that $\boldsymbol{T} = \boldsymbol{H}(\boldsymbol{U})$. Then $\boldsymbol{U}$ is also going to be sufficient for $\boldsymbol{\theta}$. However we can think of $\boldsymbol{T}$ as providing a greater reduction of the data than $\boldsymbol{U}$. This is because given the statistic $\boldsymbol{U}$ we can still reconstruct $\boldsymbol{T}$ exactly (but not necessarily the other way round since $\boldsymbol{T}$ can be a greater reduction of the data).

This idea then lends itself to the concept of minimal sufficiency. A sufficient statistic $\boldsymbol{T}$ is said to be minimal sufficient if:

- Of all the sufficient statistics it provides the greatest possible reduction of the data
- That is it is a function of every other sufficient statistic

(In other words, $\boldsymbol{T}$ is minimal sufficient if for any other sufficient statistic $\boldsymbol{U}$ you can find a function $\boldsymbol{H}$ such that $\boldsymbol{T} = \boldsymbol{H}(\boldsymbol{U})$. So $\boldsymbol{T}$ is a reduction of every other sufficient statistic.)

7.4 Completeness

A statistic $\boldsymbol{T}$ is said to be complete if,

$$E_{\boldsymbol{\theta}}(\omega(\boldsymbol{T})) = 0 \quad \forall \boldsymbol{\theta} \in \Omega \qquad \Longrightarrow \qquad \omega(\boldsymbol{T}) \equiv 0$$

(where $\omega(\boldsymbol{T}) \equiv 0$ means $P_{\boldsymbol{\theta}}(\omega(\boldsymbol{T}) = 0) = 1$ for all $\boldsymbol{\theta}$).

Intuitively the best way to think about completeness is that if $\boldsymbol{T}$ is complete then all parts of $\boldsymbol{T}$ are important in estimating $\boldsymbol{\theta}$. That is because assuming that $\omega(\boldsymbol{T}) \neq 0$, the expected value of $\omega(\boldsymbol{T})$ is going to be non-zero for at least some value of $\boldsymbol{\theta}$. I.e. $\omega(\boldsymbol{T})$ carries with it some information of $\boldsymbol{\theta}$. And completeness says that the only way that a function $\omega(\boldsymbol{T})$ can carry no information about $\boldsymbol{\theta}$ at all is if $\omega(\boldsymbol{T})$ is zero. This sounds very similar to the definition of minimal sufficiency and in fact they are related but completeness is a stronger requirement.

An important result:

$$\boldsymbol{T} \text{ is complete sufficient} \implies \boldsymbol{T} \text{ is minimal sufficient}$$

Furthermore, it can be shown that:

- the statistic $\boldsymbol{T}$ from the regular exponential family is a complete sufficient statistic (and thus also minimal sufficient).

7.5 Ancillary Statistic

An ancillary statistic is a statistic whose distribution does not depend on $\boldsymbol{\theta}$.

Therefore an ancillary statistic can be thought of as providing no information about the parameter $\boldsymbol{\theta}$. In this sense it is the opposite of a complete statistic which contains no ancillary information.

8 Estimator Theorems

Here I include a collection of useful theorems involving estimators.

8.1 Jensen's Inequality

Suppose $h(x)$ is a convex function and X is a random variable with mean μ. Then Jensen's inequality states that,

$$E[h(X)] \geq h(\mu)$$

If $h(x)$ is strictly convex then the inequality is strict (unless X is constant with probability 1).

Sometimes you may see the inequality written as:

$$E[h(X)] \geq h(E[X])$$

Here I am going to give two proofs:

- A weak proof that was only mentioned very quickly in the lectures. It is worth including since it follows through very simply. It is a "weak" proof because it assumes that $h(x)$ is twice differentiable and generally "well-behaved", but it is still useful to bear in mind.
- A full proof that is mathematically rigorous

8.1.1 Weak Proof

Suppose $h(x)$ is twice differentiable and so it can be approximated by its Taylor series,

$$h(x) \approx h(\mu) + h'(\mu)(x-\mu) + \frac{1}{2}h''(\mu)(x-\mu)^2$$

Taking the expected value of both sides (with $\sigma^2 = \text{var}(X)$),

$$\begin{aligned} E[h(X)] &\approx h(\mu) + h'(\mu)E[X-\mu] + \frac{1}{2}h''(\mu)E[(X-\mu)^2] \\ &= h(\mu) + \frac{1}{2}h''(\mu)\sigma^2 \\ &\geq h(\mu) \end{aligned}$$

This is using the fact that for a convex function $h''(x) \geq 0$ for all x.

8.1.2 Full Proof

It can be shown that if $h(x)$ is convex then at any point $(t, h(t))$ there exists a line through that point such that $h(x)$ is always on or above that line.

Let this line be denoted $L(x)$ where,

$$L(x) = k(x-t) + h(t)$$

for some constant k.

(notice that this is just the point gradient form of the tangent line and if h was known to be differentiable then $k = h'(t)$).

We therefore have that,

$$h(x) \geq L(x) \quad \forall x$$

On taking $t = \mu$ we get Jensen's inequality,

$$\begin{aligned} E[h(X)] &\geq E[L(X)] \\ &= E[k(X - \mu) + h(\mu)] \\ &= kE[X - \mu] + h(\mu) \\ &= h(\mu) \end{aligned}$$

8.2 Kullback–Leibler Distance

Let F_1 and F_2 be two distributions with densities f_1 and f_2 respectively. Then the Kullback–Leibler distance between F_1 and F_2 is defined by,

$$\delta_{KL}(F_1, F_2) \coloneqq \int f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx$$

Note that it is directional since generally,

$$\delta_{KL}(F_1, F_2) \neq \delta_{KL}(F_2, F_1)$$

Jeffreys' measure is a symmetric combination of the Kullback–Leibler distances such that,

$$\delta_J(F_1, F_2) = \delta_{KL}(F_1, F_2) + \delta_{KL}(F_2, F_1)$$

We can show that the KL distance is non-negative,

$$\begin{aligned} \delta_{KL} &= \int f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx \\ &= \int -f_1(x) \log\left(\frac{f_2(x)}{f_1(x)}\right) dx \\ &= E_{F_1}\left[-\log\left(\frac{f_2(X)}{f_1(X)}\right)\right] \end{aligned}$$

Here notice that $-\log(x)$ is a convex function and so we can apply Jensen's inequality,

$$\begin{aligned} \delta_{KL} &\geq -\log\left(E_{F_1}\left[\frac{f_2(X)}{f_1(X)}\right]\right) \\ &= -\log\left(\int \frac{f_2(x)}{f_1(x)} f_1(x)\, dx\right) \\ &= -\log\left(\int f_2(x)\, dx\right) \\ &= -\log(1) \\ &= 0 \end{aligned}$$

8.3 Rao–Blackwell Theorem

Suppose you have a random sample,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_\theta$$

with sufficient statistic T and some estimator U of θ (possibly biased). The Rao–Blackwell theorem then tells us that,

$$W(T) = E[U|T]$$

is a new estimator that:

- Is a function of T alone
- $\text{var}(W) \leq \text{var}(U)$

(note that in the background notes we additionally consider U to be unbiased but this isn't actually needed for Rao–Blackwell. If U were to be unbiased it would then let us conclude that $W(T)$ is also unbiased but it is not actually necessary for the core argument of Rao–Blackwell and that is of the reduction in variance).

Recall that the accuracy (or rather inaccuracy) of an estimator W can be formulated in terms of a risk function R where,

$$R(W;\theta) = E[\ell(W,\theta)]$$

Or in other words the risk of W in estimating θ is equal to its expected loss that is given by the loss function $\ell(W,\theta)$. Typically we take,

$$\ell(W,\theta) = (W-\theta)^2$$

to be the loss function (measures squared deviation from what it needs to be estimating). Using this definition the risk function is simply the MSE,

$$R(W;\theta) = E[(W-\theta)^2] = MSE(W) = \text{bias}(W)^2 + \text{var}(W)$$

In the context of the Rao–Blackwell theorem notice that,

$$\text{bias}(W) = E[W] - \theta = E[E[U|T]] - \theta = E[U] - \theta = \text{bias}(U) + \theta - \theta = \text{bias}(U)$$

Therefore $\text{var}(W) \leq \text{var}(U)$ implies that,

$$MSE(W) \leq MSE(U)$$

Or in other words the Rao–Blackwell theorem can be thought of as a reduction in the risk function.

8.3.1 Generalised Rao–Blackwell

As above, suppose $U(X_1, \ldots, X_n)$ is an estimator of θ and $T(X_1, \ldots, X_n)$ is a sufficient statistic. Then take,

$$W(T) = E[U|T]$$

Then the generalised Rao–Blackwell theorem states that for any strictly convex **loss function**,

$$R(W;\theta) \leq R(U;\theta) \quad \forall\theta$$

The "standard" Rao–Blackwell is just the special case where the loss is the convex function $\ell = (W-\theta)^2$.

Proof.

$$\begin{aligned}
\ell(W,\theta) &= \ell(E[U|T],\theta) \\
&\le E[\ell(U,\theta)\,|\,T] \qquad \text{(Jensen's Inequality)} \\
E[\ell(W,\theta)] &\le E\big[E[\ell(U,\theta)\,|\,T]\big] \\
&= E[\ell(U,\theta)]
\end{aligned}$$

Therefore,

$$R(W;\theta) \le R(U;\theta) \quad \forall\theta$$

Note on the application of Jensen's inequality: recall that $P(\,\cdot\,|\,X)$ is a valid probability measure (it satisfies the probability axioms) and so you can treat $E[\,\cdot\,|\,X]$ as an expectation. Therefore Jensen's inequality can be written as,

$$h(E[A\,|\,X]) \le E[h(A)\,|\,X]$$

which is what I used above. □

(The proof in the lecture notes is slightly different but gets the same result: it essentially uses the law of total probability backwards in the first step.)

8.4 Lehmann–Scheffé Theorem

If T is a complete sufficient statistic and $W(T)$ is a function of T that is unbiased for $g(\theta)$, then $W(T)$ is the unique UMVUE of $g(\theta)$. Combined with the Rao–Blackwell theorem, this means that for any unbiased estimator U of $g(\theta)$, $E[U\,|\,T]$ is the UMVUE. (The proof isn't written up here; see the background notes.)

8.5 Basu's Theorem

Let $\boldsymbol{T}$ be a complete sufficient statistic for $\boldsymbol{\theta}$ and suppose that $\boldsymbol{V}$ is an ancillary statistic. Then $\boldsymbol{V}$ is independent of $\boldsymbol{T}$.

Recall that:

- $\boldsymbol{T}$ is sufficient if the conditional distribution of $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ given $\boldsymbol{T}$ does not depend on $\boldsymbol{\theta}$.
- $\boldsymbol{T}$ is complete if for any function ω, $E_{\boldsymbol{\theta}}[\omega(\boldsymbol{T})] = 0$ for all $\boldsymbol{\theta} \implies \omega(\boldsymbol{T}) \equiv 0$
- $\boldsymbol{V}$ is an ancillary statistic if the distribution of $\boldsymbol{V}$ does not depend on $\boldsymbol{\theta}$

Proof. Fix a value v and let $\omega(T) = f(v\,|\,T) - f(v)$. Notice that ω is a function of T alone (and not $\boldsymbol{\theta}$): $f(v\,|\,T)$ doesn't depend on $\boldsymbol{\theta}$ because T is sufficient, and $f(v)$ doesn't depend on $\boldsymbol{\theta}$ because V is ancillary.

$$\begin{aligned}
E[\omega(T)] &= E[f(v\,|\,T) - f(v)] \\
&= E[f(v\,|\,T)] - f(v) \\
&= E\left[\frac{f(v,T)}{f(T)}\right] - f(v) \\
&= \sum_t \left(\frac{f(v,t)}{f(t)} f(t)\right) - f(v) \\
&= \sum_t f(v,t) - f(v) \\
&= f(v) - f(v) \\
&= 0
\end{aligned}$$

(for continuous T the sums are integrals, which is how the background notes write it). This holds for every $\boldsymbol{\theta}$, so as T is complete, $\omega(T) \equiv 0$. That is, $f(v\,|\,T) = f(v)$ for every v, which implies independence. □

8.5.1 Proving Independence of $\overline{X}$ and s^2 using Basu's Theorem

Suppose,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

Treat σ^2 as known for the moment, so that μ is the only parameter. Recall that the likelihood function then belongs to the regular exponential family, and that $\overline{X}$ is a complete sufficient statistic for μ.

Now take,

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

We can show that s^2 is an ancillary statistic (with respect to μ). Intuitively this makes sense since the distribution of squared deviations from the mean should not depend on μ but mathematically recall that,

$$\frac{n-1}{\sigma^2}s^2 \sim \chi^2_{n-1}$$

which clearly shows that its distribution does not depend on μ.

Therefore by Basu's theorem we can conclude that $\overline{X}$ and s^2 are independent. This holds for every value of σ^2, so it holds in general.

9 Asymptotic Properties

Large sample theory considers the observed sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ as n grows and assesses the performance of the estimator as $n \to \infty$. Let,

$$\boldsymbol{X}_s = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$$

be a sample of the first n observations. We can then define $\boldsymbol{T}_n$ as the estimator $\boldsymbol{T}$ applied to the first n samples.

9.1 Consistency

A sequence of estimators T_n of $g(\theta)$ is said to be consistent if for every $\theta \in \Omega$,

$$T_n \xrightarrow{p} g(\theta) \quad \text{as} \quad n \to \infty$$

That is, it converges in probability to $g(\theta)$.

Mathematically this means that,

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P(|T_n - g(\theta)| \geq \varepsilon) = 0$$

9.1.1 Relation to the MSE

An interesting result is that,

$$\begin{pmatrix} \text{bias}(T_n) \to 0 \\ \text{var}(T_n) \to 0 \end{pmatrix} \iff \text{MSE}(T_n) \to 0 \implies T_n \text{ is consistent}$$

(notice that the other direction does not hold).

Proof. The result follows easily from the Markov inequality,

$$\begin{aligned} P(|T_n - g(\theta)| \geq \varepsilon) &= P\big((T_n - g(\theta))^2 \geq \varepsilon^2\big) \\ &\leq \frac{E[(T_n - g(\theta))^2]}{\varepsilon^2} \qquad \text{(Markov inequality)} \\ &= \frac{\mathrm{MSE}(T_n)}{\varepsilon^2} \end{aligned}$$

which tends to 0 as $n \to \infty$ for every $\varepsilon > 0$. □

(The background notes also prove this directly, without using the Markov inequality, but that proof is more involved; see page 31.)

9.2 Large-Sample Properties of the MLE

The following theorems describe how the MLE behaves for large samples. Their proofs aren't covered here.

9.2.1 Theorem 1

Suppose,

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{\theta}}$$

and that the following regularity conditions hold:

- the densities $f(\boldsymbol{x}; \boldsymbol{\theta})$ are distinct for different values of $\boldsymbol{\theta}$
- the densities $f(\boldsymbol{x}; \boldsymbol{\theta})$ have common support
- the true value of $\boldsymbol{\theta}$ (denoted $\boldsymbol{\theta}_o$) does not lie on the boundary of Ω

Then, for any fixed $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$,

$$P_{\boldsymbol{\theta}_o}(L(\boldsymbol{\theta}_o) > L(\boldsymbol{\theta})) \to 1$$

as the sample size n tends to ∞.

Interpretation: This shows us that the density of $\boldsymbol{X}$ at the true value $\boldsymbol{\theta}_o$ exceeds that at any other value $\boldsymbol{\theta}$ with a high probability for large sample sizes n. If $L(\boldsymbol{\theta})$ is smooth then we can be quite confident that for large values of n the MLE will be close to $\boldsymbol{\theta}_o$.

9.2.2 Theorem 2

Suppose the same setup as above with the same regularity conditions, but with the extra condition that:

- for almost all $\boldsymbol{x}$, $f(\boldsymbol{x}; \boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$

Then as the sample size $n \to \infty$, the likelihood equation will have a sequence of roots that tend to the true value $\boldsymbol{\theta}_o$ in probability.

(This was only stated in words in the lecture notes, without proof. The important thing is to understand that such a theorem exists and what it means.)

9.2.3 Theorem 3

As before suppose,

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{\theta}}$$

and further suppose that certain regularity conditions hold (see page 6 of the lecture notes for details).

Then any consistent sequence of roots $\widehat{\theta}_n$ of the likelihood equation is asymptotically normal and efficient. I.e.,

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{\mathscr{L}} N\left(0, \frac{1}{i(\theta_o)}\right)$$

where $i(\theta)$ is the Fisher information of a single observation (so that $\mathscr{I}(\theta) = n\, i(\theta)$). Or in other words, for large n we have that,

$$\widehat{\theta}_n \overset{\text{approx.}}{\sim} N\left(\theta_o, \frac{1}{\mathscr{I}(\theta_o)}\right)$$

9.2.4 Further Results

The lecture notes contain some further theorems, which aren't covered here. (Theorem 7 on page 9 is what is generally known as the delta method, although the lecture notes don't call it that.)

10 The EM Algorithm

10.1 Setup

Suppose we observe some data,

$$\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$$

from some distribution $F_{\boldsymbol{\Psi}}$ and we want to find the MLE $\widehat{\boldsymbol{\Psi}}$. As we have shown previously this is typically achieved by solving the likelihood equation,

$$\frac{\partial \log L(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = \mathbf{0}$$

In some situations, solving this system can be quite challenging (even when using numerical techniques such as the Newton–Raphson method). However, the problem can often be made easier if we assume that there is some additional "missing" data.

Let $\boldsymbol{z}$ denote this missing data where,

$$\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$$

The complete data (denoted $\boldsymbol{x}$) is then given by,

$$\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$$

Where I am using this as a sort of "loose" notation to denote that,

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} (\boldsymbol{y}_1, \boldsymbol{z}_1) \\ \vdots \\ (\boldsymbol{y}_n, \boldsymbol{z}_n) \end{pmatrix}$$

Where $\boldsymbol{x}$ is an n-dimensional data vector with each element $\boldsymbol{x}_i$ being a column vector of dimensions $\dim(\boldsymbol{y}_i) + \dim(\boldsymbol{z}_i)$.

We then define the likelihood function of the data by,

- Let $L(\boldsymbol{\Psi}) = g(\boldsymbol{y};\, \boldsymbol{\Psi})$
- Let $L_c(\boldsymbol{\Psi}) = g_c(\boldsymbol{x};\, \boldsymbol{\Psi})$

The notes then use both L and g depending on whether you should be thinking of it as a likelihood function or a joint density respectively. Just bear in mind that these are the same underlying mathematical function, just thought about from different points of view.

10.2 The Algorithm

The EM algorithm is an iterative algorithm starting with some initial guess $\boldsymbol{\Psi}^{(0)}$ and iteratively updating this value until it converges onto the MLE $\widehat{\boldsymbol{\Psi}}$ (not necessarily the global maximising MLE but at least some local maximiser).

Given some iterate $\boldsymbol{\Psi}^{(k)}$ the next iterate is computed in two steps:

1. E-step (expectation)
2. M-step (maximisation)

10.2.1 E-Step

Here we compute what is known as the Q function that is given by,

$$Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\Big[\log L_c(\boldsymbol{\Psi}) \,|\, \boldsymbol{y}\Big]$$

You can think of the Q function as representing what our "best guess" of the complete log likelihood function is given the data $\boldsymbol{y}$ that we have observed and assuming the true parameter value was $\boldsymbol{\Psi}^{(k)}$.

10.2.2 M-step

Here we maximise the Q function over all $\boldsymbol{\Psi} \in \Omega$ and take this to be the new parameter $\boldsymbol{\Psi}^{(k+1)}$. That is,

$$\boldsymbol{\Psi}^{(k+1)} = \arg\max_{\boldsymbol{\Psi}\in\Omega} Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$$

This step is then just finding the "MLEs" of the function Q which is our "best guess" of the complete likelihood L_c.

10.3 Idea Behind the Algorithm

Recall here that our goal is to find the MLE $\widehat{\boldsymbol{\Psi}}$ which is the value that maximises $L(\boldsymbol{\Psi})$, or equivalently $\log L(\boldsymbol{\Psi})$. The key idea is that it may be easier to do this if we had access to the complete data $\boldsymbol{x}$.

Let $k(\boldsymbol{x} \,|\, \boldsymbol{y};\, \boldsymbol{\Psi})$ be the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$. I.e.,

$$k(\boldsymbol{x} \,|\, \boldsymbol{y};\, \boldsymbol{\Psi}) = \frac{g_c(\boldsymbol{x};\, \boldsymbol{\Psi})}{g(\boldsymbol{y};\, \boldsymbol{\Psi})}$$

And thus,

$$\log g(\boldsymbol{y};\, \boldsymbol{\Psi}) = \log g_c(\boldsymbol{x};\, \boldsymbol{\Psi}) - \log k(\boldsymbol{x} \,|\, \boldsymbol{y};\, \boldsymbol{\Psi})$$

Which can be equivalently written in terms of the likelihood function as,

$$\log L(\boldsymbol{\Psi}) = \log L_c(\boldsymbol{\Psi}) - \log k(\boldsymbol{x} \,|\, \boldsymbol{y};\, \boldsymbol{\Psi})$$

We can now take the conditional expectation of both sides given the observed data $\boldsymbol{y}$,

$$E_{\boldsymbol{\Psi}^{(k)}}\Big[\log L(\boldsymbol{\Psi}) \,|\, \boldsymbol{y}\Big] = E_{\boldsymbol{\Psi}^{(k)}}\Big[\log L_c(\boldsymbol{\Psi}) \,|\, \boldsymbol{y}\Big] - E_{\boldsymbol{\Psi}^{(k)}}\Big[\log k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}) \,|\, \boldsymbol{y}\Big]$$

$$\log L(\boldsymbol{\Psi}) = Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) - H(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$$

where $H(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\big[\log k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}) \,|\, \boldsymbol{y}\big]$. Notice that the expectation did not affect the LHS since $L(\boldsymbol{\Psi})$ is a function of $\boldsymbol{y}$ only and not $\boldsymbol{x}$ and so taking its expected value given that $\boldsymbol{y}$ is fixed just leaves it unchanged. It essentially becomes a constant when $\boldsymbol{y}$ is taken to be fixed.

We can then show that by iteratively maximising the Q function the likelihood will monotonically increase.

10.4 Monotonicity of the EM Algorithm

It can be shown that for iterates $\boldsymbol{\Psi}^{(k)}$ and $\boldsymbol{\Psi}^{(k+1)}$ of the EM algorithm,

$$\log L(\boldsymbol{\Psi}^{(k+1)}) \geq \log L(\boldsymbol{\Psi}^{(k)})$$

Proof.

$$\log L(\boldsymbol{\Psi}^{(k+1)}) - \log L(\boldsymbol{\Psi}^{(k)}) = \underbrace{\left[Q(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) - Q(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)})\right]}_{\geq 0 \text{ by the M-step}} - \underbrace{\left[H(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) - H(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)})\right]}_{\leq 0 \text{ by Jensen's inequality}}$$

Just to clarify further:

- $Q(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) - Q(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)})$ is greater than or equal to zero by the definition of the M-step, since we take $\boldsymbol{\Psi}^{(k+1)}$ to be the argument that maximises the Q function.
- The second term is non-positive due to applying Jensen's inequality to the concave log function as I show below,

$$\begin{aligned}
H(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) - H(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)}) &= E_{\boldsymbol{\Psi}^{(k)}}\Big[\log k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k+1)}) \,|\, \boldsymbol{y}\Big] \\
&\quad - E_{\boldsymbol{\Psi}^{(k)}}\Big[\log k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k)}) \,|\, \boldsymbol{y}\Big] \\
&= E_{\boldsymbol{\Psi}^{(k)}}\left[\log \frac{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k+1)})}{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k)})} \,\middle|\, \boldsymbol{y}\right] \\
&\leq \log\left(E_{\boldsymbol{\Psi}^{(k)}}\left[\frac{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k+1)})}{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k)})} \,\middle|\, \boldsymbol{y}\right]\right) \\
&= \log(1) \\
&= 0
\end{aligned}$$

The reason that the expectation term was just equal to 1 is because,

$$\begin{aligned}
E_{\boldsymbol{\Psi}^{(k)}}\left[\frac{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k+1)})}{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k)})} \,\middle|\, \boldsymbol{y}\right] &= \int \cdots \int \frac{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k+1)})}{k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k)})} k(\boldsymbol{x} \,|\, \boldsymbol{y};\, \boldsymbol{\Psi}^{(k)})\, d\boldsymbol{x} \\
&= \int \cdots \int k(\boldsymbol{x} \,|\, \boldsymbol{y}; \boldsymbol{\Psi}^{(k+1)})\, d\boldsymbol{x} \\
&= 1 \qquad \text{(since } k \text{ is a density)}
\end{aligned}$$

□

10.5 Generalised EM (GEM) Algorithm

Notice that the monotonicity argument from above relies on the fact that in the M-step,

$$Q(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)})$$

This is clearly the case when performing the EM algorithm since $\boldsymbol{\Psi}^{(k+1)}$ is taken to be,

$$\boldsymbol{\Psi}^{(k+1)} = \underset{\boldsymbol{\Psi} \in \Omega}{\arg\max}\, Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$$

In some cases finding the argmax may be challenging or even impossible however in practice we don't actually have to take $\boldsymbol{\Psi}^{(k+1)}$ to be the global maximum to ensure monotonicity. All that we require is that,

$$Q(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)})$$

And so this leads to the notion of the GEM Algorithm that still preserves monotonicity but provides a weaker requirement on the new iterate.

Important note:

- Just because the likelihood is non-decreasing with every new iterate does not automatically guarantee that the iterates converge to a stationary point. To prove convergence you need additional regularity conditions.

10.6 EM on the Regular Exponential Family

Suppose that the density of the complete data comes from the regular exponential family. That is,

$$g_c(\boldsymbol{x};\, \boldsymbol{\Psi}) = b(\boldsymbol{x}) \exp(\boldsymbol{c}(\boldsymbol{\Psi})^T \boldsymbol{T}(\boldsymbol{x})) \,/\, a(\boldsymbol{\Psi})$$

And:

- The dimension of $\boldsymbol{\Psi}$ matches the dimension of $\boldsymbol{T}(\boldsymbol{x})$
- The Jacobian of $\boldsymbol{c}(\boldsymbol{\Psi})$ is of full rank

Given this the Q function can be reduced to the following,

$$\begin{aligned}
Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) &= E_{\boldsymbol{\Psi}^{(k)}}\Big[\log L_c(\boldsymbol{\Psi}) \,|\, \boldsymbol{y}\Big] \\
&= E_{\boldsymbol{\Psi}^{(k)}}\Big[\log\Big(b(\boldsymbol{x}) \exp(\boldsymbol{c}(\boldsymbol{\Psi})^T \boldsymbol{T}(\boldsymbol{x})) \,/\, a(\boldsymbol{\Psi})\Big)\,\Big|\,\boldsymbol{y}\Big] \\
&= E_{\boldsymbol{\Psi}^{(k)}}\Big[\log(b(\boldsymbol{x}))\,\Big|\,\boldsymbol{y}\Big] + E_{\boldsymbol{\Psi}^{(k)}}\Big[\boldsymbol{c}(\boldsymbol{\Psi})^T \boldsymbol{T}(\boldsymbol{x})\,\Big|\,\boldsymbol{y}\Big] - E_{\boldsymbol{\Psi}^{(k)}}\Big[\log a(\boldsymbol{\Psi})\,\Big|\,\boldsymbol{y}\Big] \\
&= \boldsymbol{c}(\boldsymbol{\Psi})^T E_{\boldsymbol{\Psi}^{(k)}}\Big[\boldsymbol{T}(\boldsymbol{x})\,\Big|\,\boldsymbol{y}\Big] - \log a(\boldsymbol{\Psi})
\end{aligned}$$

Notice that the reason that I dropped the $E_{\boldsymbol{\Psi}^{(k)}}\Big[\log(b(\boldsymbol{x}))\,\Big|\,\boldsymbol{y}\Big]$ term from above is not because it is equal to zero but because the quantity is just a number (i.e. it doesn't depend on $\boldsymbol{\Psi}$) at all and so in the M-step when we maximise over $\boldsymbol{\Psi}$ it is irrelevant.

Furthermore we can use the structure of the regular exponential family to reduce the M-step to finding the value $\boldsymbol{\Psi}^{(k+1)}$ that satisfies,

$$E_{\boldsymbol{\Psi}^{(k+1)}}\Big[\boldsymbol{T}(\boldsymbol{X})\Big] = E_{\boldsymbol{\Psi}^{(k)}}\Big[\boldsymbol{T}(\boldsymbol{X})\,\Big|\,\boldsymbol{y}\Big]$$

(this result just follows from the property for finding the MLEs of data in the REF, Section 6).

10.7 Example: Mixtures of Normal Distributions

Here I am going to run through an example of the EM algorithm applied to finding the maximum likelihood estimators of data coming from a mixture model of normal distributions.

10.7.1 Problem Setup

Suppose you observe data,

$$\boldsymbol{y} = (y_1, \ldots, y_n)^T$$

that is independently sampled from a mixture of G normal distributions (i.e. G different classes).

That is the density $f(y; \boldsymbol{\Psi})$ is given by,

$$f(\boldsymbol{y};\, \boldsymbol{\Psi}) = \prod_{j=1}^{n} \sum_{i=1}^{G} \pi_i \phi(y_j;\, \mu_i, \sigma_i^2)$$

where $\phi(\,\cdot\,;\, \mu, \sigma^2)$ is the normal density, $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_G, \mu_1, \ldots, \mu_G, \sigma_1^2, \ldots, \sigma_G^2)^T$ and,

$$\sum_{i=1}^{G} \pi_i = 1$$

Given this setup, one thing you may want to do is to compute the ML estimate $\widehat{\boldsymbol{\Psi}}$ however in practice it turns out that doing this is very hard. In fact, in the case of a mixture of normal distributions there is no closed-form solution for the MLEs.

However by treating the problem as a missing data problem we can use the EM algorithm to numerically find the MLEs.

10.7.2 EM Setup

Suppose we had access to some missing data: namely, which class each observation came from. This would make the problem much easier to solve.

Let,

$$\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$$

where,

$$\boldsymbol{z}_j = (z_{1j}, \ldots, z_{Gj})^T \quad \text{and} \quad z_{ij} = \begin{cases} 1 & \text{if } y_j \text{ is from class } i \\ 0 & \text{otherwise} \end{cases}$$

Essentially $\boldsymbol{z}$ can be thought of as a vector of one-hot encodings, where $z_{ij} = (\boldsymbol{z}_j)_i$ is 1 if the jth observation is from class i.

We now form the complete data as,

$$\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$$

Given this we can now define the density of complete data f_c by,

$$f_c(\boldsymbol{x};\, \boldsymbol{\Psi}) = \prod_{j=1}^{n} \prod_{i=1}^{G} \Big[\pi_i \phi(y_j;\, \mu_i, \sigma_i^2)\Big]^{z_{ij}}$$

(This makes sense because raising each term to the power of z_{ij} selects only the term with $z_{ij} = 1$, which is the class that the particular observation y_j belongs to.)

10.7.3 Constructing the Q-Function

By definition, the Q-function is given by,

$$\begin{aligned} Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) &= E_{\boldsymbol{\Psi}^{(k)}}\Big[\log L_c(\boldsymbol{\Psi}) \,|\, \boldsymbol{y}\Big] \\ &= E_{\boldsymbol{\Psi}^{(k)}}\left[\log \prod_{j=1}^{n} \prod_{i=1}^{G} \Big[\pi_i \phi(y_j;\, \mu_i, \sigma_i^2)\Big]^{z_{ij}} \,\middle|\, \boldsymbol{y}\right] \\ &= E_{\boldsymbol{\Psi}^{(k)}}\left[\sum_{j=1}^{n} \sum_{i=1}^{G} z_{ij} \log \Big[\pi_i \phi(y_j;\, \mu_i, \sigma_i^2)\Big] \,\middle|\, \boldsymbol{y}\right] \end{aligned}$$

It is worth noting now what in the above expression is actually random.

- π_i, μ_i and σ_i^2 are all unknown constants (at least in frequentist statistics) and so they are not random
- y_j is also not "random" here since we are conditioning on $\boldsymbol{y}$ in the expectation
- the only term remaining that is actually random is z_{ij}

Therefore,

$$Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{G} E_{\boldsymbol{\Psi}^{(k)}}\Big[z_{ij} \,|\, \boldsymbol{y}\Big] \log\Big[\pi_i \phi(y_j;\, \mu_i, \sigma_i^2)\Big]$$

For ease of notation let,

$$\begin{aligned}\tau_{ij}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}}\Big[z_{ij} \,|\, \boldsymbol{y}\Big] \\ &= E_{\boldsymbol{\Psi}^{(k)}}\Big[z_{ij} \,|\, y_j\Big]\end{aligned}$$

The reason that I only need to condition on y_j is because all of the samples are independent and so the class of a particular observation should only depend on y_j and $\boldsymbol{\Psi}^{(k)}$ (not on all of the other data points in $\boldsymbol{y}$).

The Q-function can then be written as,

$$\begin{aligned}Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) &= \sum_{j=1}^{n}\sum_{i=1}^{G} \tau_{ij}^{(k)} \log\Big[\pi_i \phi(y_j;\, \mu_i, \sigma_i^2)\Big] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{G} \tau_{ij}^{(k)} \Big(\log(\pi_i) + \log(\phi(y_j;\, \mu_i, \sigma_i^2))\Big) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{G} \tau_{ij}^{(k)} \left(\log(\pi_i) - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_j - \mu_i)^2\right)\end{aligned}$$

We still need to compute what $\tau_{ij}^{(k)}$ is but since it is a constant with respect to $\boldsymbol{\Psi}$ I am going to leave it as it is for now and continue on to the maximisation step of the algorithm.

10.7.4 Maximising the Q-Function

One thing that you may notice is that in this mixture model situation the Q function is actually just equal to $\log L_c(\boldsymbol{\Psi})$ but with z_{ij} replaced with $\tau_{ij}^{(k)}$. Therefore maximising the Q-function is effectively the same as finding the MLEs of $\log L_c$.

Let $\boldsymbol{\Psi}^{(k+1)}$ be the value that maximises the Q function such that,

$$Q(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) \quad \forall \boldsymbol{\Psi}$$

To find $\boldsymbol{\Psi}^{(k+1)}$ we need to solve,

$$\left.\frac{\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})}{\partial \boldsymbol{\Psi}}\right|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k+1)}} = \mathbf{0}$$

Subject to the constraint that,

$$\sum_{i=1}^{G} \pi_i = 1$$

This constraint is important since it means that there is a dependence on the values of π_i. To maximise with respect to π_i we thus have to use Lagrange multipliers.

$$\mathcal{L}(\boldsymbol{\Psi}, \lambda) = Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) + \lambda\left(\sum_{i=1}^{G} \pi_i - 1\right)$$

Set partial derivative with respect to π_i to zero (for some arbitrary $i \in \{1, \ldots, G\}$),

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\Psi}, \lambda)}{\partial \pi_i} &= 0 \\
\sum_{j=1}^{n} \frac{\tau_{ij}^{(k)}}{\pi_i} + \lambda &= 0 \\
\pi_i &= \frac{-1}{\lambda} \sum_{j=1}^{n} \tau_{ij}^{(k)} \qquad (1) \\
\sum_{i=1}^{G} \pi_i &= \frac{-1}{\lambda} \sum_{i=1}^{G} \sum_{j=1}^{n} \tau_{ij}^{(k)} \\
1 &= \frac{-1}{\lambda} \sum_{i=1}^{G} \sum_{j=1}^{n} \tau_{ij}^{(k)} \\
\lambda &= -\sum_{i=1}^{G} \sum_{j=1}^{n} \tau_{ij}^{(k)} \\
&= -\sum_{j=1}^{n} E_{\boldsymbol{\Psi}^{(k)}} \left[\sum_{i=1}^{G} z_{ij} \,\middle|\, \boldsymbol{y} \right] \\
&= -\sum_{j=1}^{n} E_{\boldsymbol{\Psi}^{(k)}} \left[1 \,\middle|\, \boldsymbol{y} \right] \\
&= -n \qquad (2)
\end{aligned}$$

(The last few steps use that $\sum_{i=1}^{G} z_{ij} = 1$, since each observation belongs to exactly one class.) The key result now is that we can combine expressions (1) and (2) to find the maximiser π_i (which I will denote $\pi_i^{(k+1)}$)

$$\pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \tau_{ij}^{(k)}$$

Now to find $\mu_i^{(k+1)}$ we can go back to maximising the Q function directly since there are no more constraints,

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})}{\partial \mu_i} &= 0 \\
\sum_{j=1}^{n} \tau_{ij}^{(k)} \left(\frac{1}{\sigma_i^2} (y_j - \mu_i) \right) &= 0 \\
\sum_{j=1}^{n} \tau_{ij}^{(k)} y_j &= \mu_i \sum_{j=1}^{n} \tau_{ij}^{(k)} \\
\mu_i^{(k+1)} &= \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} y_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}
\end{aligned}$$

And lastly for $\sigma_i^{2(k+1)}$,

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})}{\partial \sigma_i^2} &= 0 \\
\sum_{j=1}^{n} \tau_{ij}^{(k)} \left(\frac{-1}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}(y_j - \mu_i)^2 \right) &= 0 \\
\frac{1}{2\sigma_i^4} \sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i)^2 &= \frac{1}{2\sigma_i^2} \sum_{j=1}^{n} \tau_{ij}^{(k)} \\
\sigma_i^2 &= \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i)^2}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}
\end{aligned}$$

Thus,

$$\sigma_i^{2(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)})^2}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}$$

And so the maximisers are,

$$\pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \tau_{ij}^{(k)} \qquad \mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} y_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \qquad \sigma_i^{2(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)})^2}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}$$

for $i \in \{1, \ldots, G\}$.

10.7.5 Computing $\tau_{ij}^{(k)}$

Given this the only thing that is left is to compute $\tau_{ij}^{(k)}$ explicitly.

Notice that z_{ij} is really just a Bernoulli random variable since,

$$z_{ij} = \begin{cases} 1 & \text{if observation } j \text{ belongs to class } i \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$\begin{aligned}
\tau_{ij}^{(k)} &= E_{\boldsymbol{\Psi}^{(k)}} \left[z_{ij} \,|\, y_j \right] \\
&= P_{\boldsymbol{\Psi}^{(k)}} \left(z_{ij} = 1 \,|\, y_j \right) \\
&= \frac{f\left(y_j \,|\, z_{ij} = 1;\, \boldsymbol{\Psi}^{(k)} \right) P_{\boldsymbol{\Psi}^{(k)}}(z_{ij} = 1)}{f\left(y_j;\, \boldsymbol{\Psi}^{(k)} \right)}
\end{aligned}$$

What I applied above is essentially Bayes' rule but for the case where you have a discrete state (z_{ij}) and continuous observation (y_j). The result is the same as the standard Bayes' rule but the probabilities involving y_j are turned into the corresponding density functions (since otherwise the probabilities would just be zero).

Looking at each probability term on the RHS,

- $P_{\boldsymbol{\Psi}^{(k)}}(z_{ij} = 1)$ is just the probability that observation j came from class i and thus,

$$P_{\boldsymbol{\Psi}^{(k)}}(z_{ij} = 1) = \pi_i^{(k)}$$

- $f(y_j \,|\, z_{ij} = 1;\, \boldsymbol{\Psi}^{(k)})$ is equal to the density function of y_j given that it comes from class i,

$$f(y_j \,|\, z_{ij} = 1;\, \boldsymbol{\Psi}^{(k)}) = \phi\left(y_j;\, \mu_i^{(k)}, \sigma_i^{2(k)} \right)$$

- $f\left(y_j;\, \boldsymbol{\Psi}^{(k)}\right)$ is the density of y_j given no information of what class it comes from,

$$f\left(y_j;\, \boldsymbol{\Psi}^{(k)}\right) = \sum_{l=1}^{G} \pi_l^{(k)} \phi\left(y_j;\, \mu_l^{(k)}, \sigma_l^{2(k)}\right)$$

(with the last one you could also think of it as the marginal density of $f(\boldsymbol{x}_j;\, \boldsymbol{\Psi}^{(k)})$ which can be found by summing out over all values of $z_{\bullet j}$ and would lead to the same result).

Therefore putting this together we have an expression for $\tau_{ij}^{(k)}$,

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} \phi\left(y_j;\, \mu_i^{(k)}, \sigma_i^{2(k)}\right)}{\sum_{l=1}^{G} \pi_l^{(k)} \phi\left(y_j;\, \mu_l^{(k)}, \sigma_l^{2(k)}\right)}$$

10.7.6 Bonus: Intuition Behind the Maximisers and τ_{ij}

If we go back to the complete log likelihood $\log L_c(\boldsymbol{\Psi})$ notice that it takes the form,

$$\log L_c(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \sum_{i=1}^{G} z_{ij} \left(\log(\pi_i) - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_j - \mu_i)^2 \right)$$

And I already mentioned this at the top of the last section but this is exactly the same as the Q function with the only difference being that in the complete likelihood the z_{ij} terms are known whereas in the Q function they are unknown and are thus estimated by $\tau_{ij}^{(k)}$.

In the same way, the MLEs (i.e. the maximisers) of the complete likelihood function are,

$$\widehat{\pi}_i = \frac{1}{n}\sum_{j=1}^{n} z_{ij} \qquad \widehat{\mu}_i = \frac{\sum_{j=1}^{n} z_{ij}\, y_j}{\sum_{j=1}^{n} z_{ij}} \qquad \widehat{\sigma}_i^2 = \frac{\sum_{j=1}^{n} z_{ij}(y_j - \widehat{\mu}_i)^2}{\sum_{j=1}^{n} z_{ij}}$$

These now lend themselves nicely to interpretation:

- $\widehat{\pi}_i$ says that the best estimate for the proportion of the population in class i is to take it to be the proportion of the sample that is in class i
- $\widehat{\mu}_i$ says that the best estimate of the mean of class i is to sum up all of the data-points that are from class i and divide by the total number of data-points in class i
- $\widehat{\sigma}_i^2$ says that the best estimate of the variance of class i is to take the sample variance of the data-points that are from that class (the biased version of the sample variance)

Now in the case of the Q function each z_{ij} data-point is not known and so our best estimate of it is given by τ_{ij} where recall that,

$$\tau_{ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}\left[z_{ij} \,|\, y_j\right]$$

Essentially you can think of $\tau_{ij}^{(k)}$ as a posterior probability that y_j belongs to class i.

Above we calculated $\tau_{ij}^{(k)}$ explicitly to be given by,

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} \phi\left(y_j;\, \mu_i^{(k)}, \sigma_i^{2(k)}\right)}{\sum_{l=1}^{G} \pi_l^{(k)} \phi\left(y_j;\, \mu_l^{(k)}, \sigma_l^{2(k)}\right)}$$

And all this is saying is that given your current estimates $\boldsymbol{\Psi}^{(k)}$ the probability that observation y_j belongs to class i is equal to:

- The likelihood of class i evaluated at y_j, weighted by how common class i is
- Normalised by the sum of the weighted likelihoods from all classes

In fact for a given observation y_j, the values $\tau_{1j}^{(k)}, \ldots, \tau_{Gj}^{(k)}$ define a discrete probability distribution where,

$$P_{\boldsymbol{\Psi}^{(k)}}(y_j \in \text{class } i) = \tau_{ij}^{(k)}$$

10.7.7 Stating the Full EM Algorithm

The full algorithm is therefore as follows:

1. Pick a starting value $\boldsymbol{\Psi}^{(0)}$ and set $k = 0$
2. Compute $\tau_{ij}^{(k)}$
$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} \phi\left(y_j;\, \mu_i^{(k)}, \sigma_i^{2(k)}\right)}{\sum_{l=1}^{G} \pi_l^{(k)} \phi\left(y_j;\, \mu_l^{(k)}, \sigma_l^{2(k)}\right)}$$
3. Use this to construct the Q function,
$$Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = \sum_{j=1}^{n} \sum_{i=1}^{G} \tau_{ij}^{(k)} \left(\log(\pi_i) - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_j - \mu_i)^2 \right)$$
4. Take $\boldsymbol{\Psi}^{(k+1)}$ to be the values that maximise the Q function, which can be computed as follows,
$$\pi_i^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n} \tau_{ij}^{(k)} \qquad \mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} y_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}$$
$$\sigma_i^{2(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)})^2}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}$$
5. Set $k \leftarrow k + 1$ and repeat from step 2 until the estimates converge

10.8 Example: Grouped and Truncated Data

This example, from lecture note 12, applies EM to grouped and truncated data.

10.8.1 Setup

Suppose W is a random variable with density function $f(w; \boldsymbol{\theta})$. Further suppose that the support $\mathcal{W}$ of W is partitioned into v mutually exclusive intervals such that,

$$\mathcal{W} = \bigcup_{j=1,\ldots,v} \mathcal{W}_j$$

Independent observations are made on W but only the number n_j falling in $\mathcal{W}_j$ is recorded. That is individual observations are not recorded but only the class intervals in which they fall are recorded.

Furthermore, the observations are only made if the W value falls in one of the intervals $j = 1, \ldots, r$ where $r \leq v$ (i.e. we have a truncated problem).

The observed data,

$$\boldsymbol{y} = (n_1, \ldots, n_r)^T$$

follows a multinomial distribution consisting of $n = \sum_{j=1}^{r} n_j$ draws on r categories with probabilities,

$$p_j = \frac{P_j(\boldsymbol{\theta})}{P(\boldsymbol{\theta})}$$

where,

$$P_j(\boldsymbol{\theta}) = \int_{\mathcal{W}_j} f(w; \boldsymbol{\theta})\, dw$$

$$P(\boldsymbol{\theta}) = \sum_{j=1}^{r} P_j(\boldsymbol{\theta})$$

10.8.2 The EM Framework

We can solve this problem within the EM framework by introducing the vectors,

$$\boldsymbol{u} = (n_{r+1}, \ldots, n_v)^T$$
$$\boldsymbol{w}_j = (w_{j,1}, \ldots, w_{j,n_j})^T \quad \text{for } j = 1, \ldots, v$$

Here:

- The vector $\boldsymbol{u}$ represents the missing frequencies from the truncation
- The vector $\boldsymbol{w}_j$ contains the n_j observations of W that were in the jth interval (i.e. the true values of the grouped data).

The complete data vector $\boldsymbol{x}$ is thus,

$$\boldsymbol{x} = (\boldsymbol{y}^T, \boldsymbol{u}^T, \boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_v^T)^T$$

10.8.3 The Complete-Data Likelihood

$$L_c(\boldsymbol{\theta}) = \prod_{j=1}^{v} \prod_{l=1}^{n_j} f(w_{jl}; \boldsymbol{\theta})$$

The rest of this example (the E- and M-steps) isn't written up here; see lecture note 12.

Part III

Hypothesis Testing

11 Hypothesis Testing

11.1 P-Values

The p-value is the probability under the null hypothesis of observing a value as extreme or more extreme than the observed value of the test-statistic.

It is a measure of "how surprising" under the null hypothesis the result of the experiment is.

11.2 Basic Hypothesis Testing Setup

Let,

$$\boldsymbol{x}_S = (x_1, \ldots, x_n)^T$$

be an observed random sample from the pdf $f(x; \boldsymbol{\theta})$.

Suppose we wish to test the hypothesis:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

versus the alternative hypothesis,

$$H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$$

(this is perhaps the simplest setup in hypothesis testing)

11.3 Types of Errors in a Binary Test

There are two types of errors that one can make in any binary decision problem:

- reject when it should accept (type I)
- accept when it should reject (type II)

Specifically in the framework of hypothesis testing:

- A type I error is the error of rejecting the null hypothesis when it should have been "accepted"
- A type II error is the error of "accepting" the null hypothesis when it should have been rejected

The significance level α is the probability of making a type I error:

$$\begin{aligned}\alpha &= P(\text{reject } H_0 \,|\, H_0) \\ &= \int_R L(\boldsymbol{\theta}_0;\, \boldsymbol{x}_S)\, d\boldsymbol{x}_S\end{aligned}$$

Where R is known as the critical region (CR) of the test. It is the region in the support of $\boldsymbol{x}_S$ where the test rejects H_0.

Likewise the probability of a type II error is denoted β,

$$\begin{aligned}\beta &= P(\text{reject } H_1 \,|\, H_1) \qquad \text{(i.e. not rejecting } H_0 \text{ when } H_1 \text{ is true)} \\ &= 1 - \int_R L(\boldsymbol{\theta}_1; \boldsymbol{x}_S)\, d\boldsymbol{x}_S\end{aligned}$$

The power of the test π is given by,

$$\pi = 1 - \beta$$

11.4 Neyman–Pearson Lemma

In practice we want a test with high power π for a specified low value of the significance level α.

The Neyman–Pearson lemma yields the critical region of the most powerful (MP) test of a simple H_0 vs simple H_1 at a fixed significance level. Specifically the critical region of the most powerful test is given by,

$$R = \left\{ \boldsymbol{x}_S : \frac{L(\boldsymbol{\theta}_1; \boldsymbol{x}_S)}{L(\boldsymbol{\theta}_0; \boldsymbol{x}_S)} > k \right\}$$

where the constant k determines α.

(An important note: in the Neyman–Pearson lemma you divide the likelihood at θ_1 by the likelihood at θ_0. However in the GLRT, which is directly inspired by this, it is the other way around: you divide the supremum restricted to the null hypothesis by the unrestricted supremum.)

Proof. Let R^* be the critical region of any other test of H_0 vs H_1 having significance level α. Then we shall show that its power, $\pi(R^*)$, is at most the power of the test specified above, $\pi(R)$.

$$\pi(R) - \pi(R^*) = \int_R L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S - \int_{R^*} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S$$

Now we use the fact that any two sets R and R^* can be decomposed as,

$$\begin{aligned} R &= (R \cap R^*) \cup (R \cap \overline{R^*}) \\ R^* &= (R \cap R^*) \cup (\overline{R} \cap R^*) \end{aligned}$$

Using this we can decompose the integral and simplify,

$$\begin{aligned} \pi(R) - \pi(R^*) &= \left(\int_{R \cap R^*} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S + \int_{R \cap \overline{R^*}} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S \right) \\ &\quad - \left(\int_{R \cap R^*} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S + \int_{\overline{R} \cap R^*} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S \right) \\ &= \int_{R \cap \overline{R^*}} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S - \int_{\overline{R} \cap R^*} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S \end{aligned}$$

On the region $R \cap \overline{R^*}$ we have that $\boldsymbol{x}_S \in R$ and so,

$$L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) > k L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) \quad \Longrightarrow \quad \int_{R \cap \overline{R^*}} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S \geq \int_{R \cap \overline{R^*}} k L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) d\boldsymbol{x}_S$$

On the region $\overline{R} \cap R^*$ we have that $\boldsymbol{x}_S \notin R$ and so,

$$\begin{aligned} \frac{L(\boldsymbol{\theta}_1; \boldsymbol{x}_S)}{L(\boldsymbol{\theta}_0; \boldsymbol{x}_S)} \leq k \quad &\Longrightarrow \quad L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) \leq k L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) \\ &\Longrightarrow \quad - \int_{\overline{R} \cap R^*} L(\boldsymbol{\theta}_1; \boldsymbol{x}_S) d\boldsymbol{x}_S \geq - \int_{\overline{R} \cap R^*} k L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) d\boldsymbol{x}_S \end{aligned}$$

Therefore,

$$\begin{aligned} \pi(R) - \pi(R^*) &\geq k \int_{R \cap \overline{R^*}} L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) d\boldsymbol{x}_S - k \int_{\overline{R} \cap R^*} L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) d\boldsymbol{x}_S \\ &= k \int_R L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) d\boldsymbol{x}_S - k \int_{R^*} L(\boldsymbol{\theta}_0; \boldsymbol{x}_S) d\boldsymbol{x}_S \\ &= k\alpha - k\alpha \\ &= 0 \end{aligned}$$

(the second line just uses the decomposition backwards, after adding and subtracting the integral of $kL(\boldsymbol{\theta}_0; \boldsymbol{x}_S)$ over $R \cap R^*$). □

11.5 Generalise the Alternative Hypothesis

Now suppose that the alternative hypothesis is that,

$$H_1 : \boldsymbol{\theta} \in \Omega_{H_1} \quad \text{where} \quad \Omega_{H_1} \subset \Omega$$

A test is said to be uniformly most powerful (UMP) if it is most powerful (MP) in testing,

$$H_1(\boldsymbol{\theta}_1) : \boldsymbol{\theta} = \boldsymbol{\theta}_1$$

for each $\boldsymbol{\theta}_1 \in \Omega_{H_1}$.

11.6 Example

Suppose we observe a random sample $\boldsymbol{x}_S$ from the normal distribution where σ^2 is known (and so the unknown parameter $\theta = \mu$). Here we want to test the simple hypotheses,

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

The likelihood function $L(\theta)$ is given by,

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n} (2\pi)^{\frac{-1}{2}} (\sigma^2)^{\frac{-1}{2}} \exp\left(\frac{-1}{2\sigma^2}(x_j - \mu)^2\right) \\ &= (2\pi)^{\frac{-n}{2}} (\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2\right) \end{aligned}$$

The critical region (CR) of the MP test is given by,

$$R = \left\{ \boldsymbol{x}_S : \frac{L(\theta_1)}{L(\theta_0)} > k \right\}$$

The key idea now is to look at the ratio and try to simplify it.

- At each step you need to think, large value of $\frac{L(\theta_1)}{L(\theta_0)}$ corresponds to large values of what function of $\boldsymbol{x}_S$?
- What you really need to figure out is what function of $\boldsymbol{x}_S$ makes the ratio large?
- You can always apply a monotone increasing function to both sides. For example large values of $\frac{L(\theta_1)}{L(\theta_0)}$ corresponds to large values of $\log \frac{L(\theta_1)}{L(\theta_0)}$ or even large values of $\left(\frac{L(\theta_1)}{L(\theta_0)}\right)^2$ since the likelihood is always positive.

Looking at the ratio in this case,

$$\begin{aligned} \frac{L(\theta_1)}{L(\theta_0)} &= \exp\left(\frac{1}{2\sigma^2}\left[\sum_{j=1}^{n}(x_j - \theta_0)^2 - \sum_{j=1}^{n}(x_j - \theta_1)^2\right]\right) \\ &= \exp\left(\frac{n}{2\sigma^2}\left[2\overline{x}(\theta_1 - \theta_0) - (\theta_1^2 - \theta_0^2)\right]\right) \end{aligned}$$

Suppose $\theta_1 > \theta_0$,

- Then this ratio is going to be large if $\overline{x}$ is large
- $\{\boldsymbol{x}_S : \overline{x} > k'\}$

Suppose $\theta_0 > \theta_1$,

- Then this ratio is going to be large if $\overline{x}$ is small
- $\{\boldsymbol{x}_S : \overline{x} < k''\}$

Here k' and k'' are just constants that are determined by the given significance level α. For the case where $\theta_1 > \theta_0$,

$$\begin{aligned} \alpha &= P(\overline{X} > k' \,|\, H_0) \\ &= P\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > \frac{k' - \theta_0}{\sigma/\sqrt{n}} \,\Big|\, H_0\right) \\ &= P\left(Z > \frac{k' - \theta_0}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z < -\frac{k' - \theta_0}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(-\frac{k' - \theta_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

Hence we can rearrange this relationship to get that k' is given by,

$$k' = \theta_0 - \frac{\Phi^{-1}(\alpha)\sigma}{\sqrt{n}}$$

11.6.1 Extending to a one-sided test

Suppose now that the alternative hypothesis H_1 is of the form,

$$H_1 : \theta > \theta_0$$

(so this corresponds to $\Omega_{H_1} = \{\theta : \theta > \theta_0\}$).

Consider an arbitrary value θ_1 in Ω_{H_1}. We know from above that the MP test has CR given by,

$$\{\boldsymbol{x}_S : \overline{x} > k'\}$$

This critical region is the same regardless of what θ_1 is picked. Therefore it defines a Uniformly Most Powerful (UMP) test for H_0 versus H_1.

11.6.2 Extending to a two-sided test

Suppose now that the alternative hypothesis H_1 is of the form,

$$H_1 : \theta \neq \theta_0$$

It follows from the work above that there is not a UMP test. That is because for a given $\theta_1 \in \Omega_{H_1}$ the most powerful test depends on whether $\theta_1 > \theta_0$ or $\theta_1 < \theta_0$.

11.7 Generalised Likelihood Ratio Test (GLRT)

Suppose the most general case,

$$H_0 : \boldsymbol{\theta} \in \Omega_{H_0} \qquad\qquad H_1 : \boldsymbol{\theta} \notin \Omega_{H_0}$$

Then the likelihood ratio λ is given by,

$$\lambda = \frac{\sup_{\boldsymbol{\theta} \in \Omega_{H_0}} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta})}$$

Using λ as the test statistic the critical region R is

$$R = \Big\{\boldsymbol{x}_S : \lambda(\boldsymbol{x}_S) < k\Big\}$$

for some k that depends on the significance level α.

Notice a few key points:

- the denominator is just going to be $L(\boldsymbol{\theta})$ evaluated at the MLE. I.e. $L(\widehat{\boldsymbol{\theta}})$.
- λ is a statistic and should only be a function of $\boldsymbol{x}_S$
- $\lambda \in [0, 1]$ since the denominator is always at least as large as the numerator

11.8 Wilks' Theorem

One very nice property of the GLRT is that we have Wilks' theorem, which states that under appropriate regularity conditions, as $n \to \infty$ we have that,

$$-2\log(\lambda) \xrightarrow{d} \chi^2_d$$

where,

$$d = \dim(\Omega) - \dim(\Omega_{H_0})$$

Using this we can always construct a test having approximate significance level α by rejecting H_0 if,

$$-2\log(\lambda) > \chi^2_{d;1-\alpha}$$

(The proof isn't covered here. The lecture notes work through an example of the GLRT midway through page 7.)

12 The Chi-Squared Goodness of Fit Test

Suppose we observe some data that is categorised into r classes where o_j denotes the number of observations in the jth class. The null hypothesis is that,

$$H_0 : \theta_j = \theta_{j0} \quad j = 1, \ldots, r$$

where θ_{j0} is the probability of an observation falling in the jth class. The total number of observations is,

$$n = o_1 + \cdots + o_r$$

The chi-squared goodness of fit test is defined to be,

$$T = \sum_{j=1}^{r} \frac{(o_j - e_j)^2}{e_j} \quad \text{where} \quad e_j = n\theta_{j0}$$

Under H_0, the statistic T is approximately distributed (for large n) as

$$T \overset{\text{approx.}}{\sim} \chi^2_{r-1}$$

12.1 Derivation from the GLRT

The data comes from a multinomial distribution where the likelihood is given by

$$L(\boldsymbol{\theta}) = \frac{n!}{o_1! \ldots o_r!} \prod_{j=1}^{r} \theta_j^{o_j}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{r-1})^T$ since all θ sum to 1.

The null and alternative hypotheses can then be formulated as,

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

where $\boldsymbol{\theta}_0 = (\theta_{10}, \ldots, \theta_{r-1,0})^T$.

The likelihood ratio λ is given by,

$$\begin{aligned} \lambda &= \sup_{\Omega_{H_0}} L(\boldsymbol{\theta}) / \sup_{\Omega} L(\boldsymbol{\theta}) \\ &= L(\boldsymbol{\theta}_0) / L(\widehat{\boldsymbol{\theta}}) \end{aligned}$$

Solving for the MLE $\widehat{\boldsymbol{\theta}}$,

$$\begin{aligned}
\log L(\boldsymbol{\theta}) &= \log\left(\frac{n!}{o_1!\dots o_r!}\right) + \sum_{j=1}^{r} \log\left(\theta_j^{o_j}\right) \\
\log L(\boldsymbol{\theta}) &= K + \sum_{j=1}^{r} o_j \log\left(\theta_j\right) \\
&= K + \sum_{j=1}^{r-1} o_j \log\left(\theta_j\right) + o_r \log(\theta_r) \\
&= K + \sum_{j=1}^{r-1} o_j \log\left(\theta_j\right) + o_r \log\left(1 - \sum_{j=1}^{r-1} \theta_j\right) \\
\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_j} &= \frac{o_j}{\theta_j} - \frac{o_r}{\theta_r} \quad (j = 1, \dots, r-1)
\end{aligned}$$

Setting the score equal to zero,

$$\begin{aligned}
0 &= \frac{o_j}{\widehat{\theta}_j} - \frac{o_r}{\widehat{\theta}_r} \\
\widehat{\theta}_j &= \frac{o_j \widehat{\theta}_r}{o_r}
\end{aligned}$$

Summing both sides over $j = 1, \dots, r$ (the equation holds trivially for $j = r$),

$$\begin{aligned}
1 &= \frac{n\widehat{\theta}_r}{o_r} \\
\widehat{\theta}_r &= \frac{o_r}{n}
\end{aligned}$$

Substituting this result into the above we get that,

$$\widehat{\theta}_j = \frac{o_j}{n} \quad (j = 1, \dots, r)$$

Therefore the likelihood ratio is given by,

$$\begin{aligned}
\lambda &= L(\boldsymbol{\theta}_0)/L(\widehat{\boldsymbol{\theta}}) \\
&= \prod_{j=1}^{r} \theta_{j0}^{o_j} / \prod_{j=1}^{r} \left(\frac{o_j}{n}\right)^{o_j} \\
\log(\lambda) &= \sum_{j=1}^{r} o_j \log(\theta_{j0}) - \sum_{j=1}^{r} o_j \log\left(\frac{o_j}{n}\right) \\
&= \sum_{j=1}^{r} o_j \log\left(\frac{n\theta_{j0}}{o_j}\right) \\
-2\log(\lambda) &= -2\sum_{j=1}^{r} o_j \log\left(\frac{n\theta_{j0}}{o_j}\right) \\
&= -2\sum_{j=1}^{r} o_j \log\left(\frac{e_j}{o_j}\right)
\end{aligned}$$

By Wilks' theorem (Section 11), $-2\log\lambda$ is approximately χ^2_{r-1} under H_0, since $\dim(\Omega) = r - 1$ and $\dim(\Omega_{H_0}) = 0$. Now the idea here is that the chi-squared GOF test approximates this statistic

by looking at the Taylor series expansion of $\log(1+x)$.

$$\begin{aligned}
-2\log(\lambda) &= -2\sum_{j=1}^{r} o_j \log\left(1+\frac{e_j-o_j}{o_j}\right) \\
&= -2\sum_{j=1}^{r} o_j \left[\frac{e_j-o_j}{o_j} - \frac{1}{2}\frac{(e_j-o_j)^2}{o_j^2} + \dots\right] \\
&\approx -2\sum_{j=1}^{r}\left(e_j - o_j - \frac{1}{2}\frac{(e_j-o_j)^2}{o_j}\right) \\
&= -2\sum_{j=1}^{r}(e_j-o_j) + \sum_{j=1}^{r}\frac{(e_j-o_j)^2}{o_j}
\end{aligned}$$

Notice that the first term is just zero since $\sum e_j = \sum o_j = n$. Therefore,

$$\begin{aligned}
-2\log\lambda &\approx \sum_{j=1}^{r}\frac{(e_j-o_j)^2}{o_j} \\
&= \sum_{j=1}^{r}\frac{(o_j-e_j)^2}{o_j}
\end{aligned}$$

The usual chi-squared statistic has e_j in the denominator rather than o_j. Under H_0, $o_j/e_j \to 1$ as $n \to \infty$, so the two versions are asymptotically equivalent (the lecture notes make this swap without comment). Part of the reason for using e_j is that if a category had no observations, the version with o_j in the denominator would divide by zero.

13 Multiple Testing

13.1 The Family-Wise Error Rate

Suppose we wish to carry out N simultaneous tests $i = 1, \dots, N$. Let A_i be the event that a type I error occurs in test i. Let,

$$A = \bigcup_{i=1}^{N} A_i$$

be the event that at least 1 type I error occurs in the N tests. The probability of A is known as the Family-Wise Error Rate (FWER).

$$\begin{aligned}
\text{FWER} &= P(A) \\
&= P\left(\bigcup_{i=1}^{N} A_i\right) \\
&\leq \sum_{i=1}^{N} P(A_i)
\end{aligned}$$

which follows from the subadditivity of the probability measure.

If the N tests are statistically independent and each is carried out at significance level α then,

$$\begin{aligned}
\text{FWER} &= P\left(\bigcup_{i=1}^{N} A_i\right) \\
&= 1 - P\left(\bigcap_{i=1}^{N} \overline{A_i}\right) \\
&= 1 - \prod_{i=1}^{N} P(\overline{A_i}) \\
&= 1 - \prod_{i=1}^{N} (1 - P(A_i)) \\
&= 1 - \prod_{i=1}^{N} (1 - \alpha) \\
&= 1 - (1-\alpha)^N
\end{aligned}$$

Thus as the number of tests N grows, the FWER is inflated.

13.2 The Bonferroni Method

The idea is that we perform the ith test such that,

$$P(A_i) \leq \frac{\alpha}{N}$$

Using this it follows that,

$$\text{FWER} \leq \alpha \qquad (1)$$

However this does come at a significant reduction in the power of the test. It only really works well if N is very small.

(The proof of (1) follows directly from the inequality at the top of this section: FWER $\leq \sum_{i=1}^{N} \alpha/N = \alpha$.)

13.3 Limitations of Bonferroni

As described above, the Bonferroni method controls the FWER. However it is often too strict for common datasets since it tries to make it very unlikely for there to be any false rejections consequently leading to potentially missed findings.

13.4 The False Discovery Rate (FDR)

An alternative approach is the FDR. It emphasises the proportion of false positives among the rejected null hypotheses. It is a relatively new method with the initial paper published in 1995 and further foundational work up to 2001.

Table 1 summarises the possible outcomes of the N tests.

A few important points:

- Firstly a, b, c and d are all random variables. Even after performing the experiment we have no way of knowing how many false positives or true positives we made, for example.
- N_r is the total number of rejected tests (i.e. $N_r = b + d$)
- FDR stands for the False Discovery Rate and it is approximately $\frac{b}{N_r} = \frac{b}{b+d}$
- FNDR stands for the False Non-Discovery Rate and it is approximately $\frac{c}{a+c}$

	Predicted	
True	Retain null	Reject null
Null	a	b (false positive)
Non-null	c (false negative)	d
		$N_r = b + d$

Table 1: The outcomes of N tests. The FDR is approximately b/N_r and the FNDR is approximately $c/(a + c)$.

- Recall that the FWER is nothing but $P(b \geq 1)$

The False Discovery Proportion (FDP) is defined to be,

$$\frac{b}{N_r}$$

In the case where no null hypotheses were rejected (i.e. $N_r = 0$) we define the fraction to be zero. Notice that after an experiment N_r is known but b is not. We then define the FDR as,

$$\text{FDR} = E[\text{FDP}]$$

And there are some important properties of the FDR:

1. If all the null hypotheses are true then,
$$\text{FDR} = \text{FWER}$$
2. If the number of true null hypotheses N_0 is less than the total number of hypotheses N then,
$$\text{FDR} \leq \text{FWER}$$

13.5 Proofs

To show (1) first notice that if all the null hypotheses are true then any rejection is going to be a false positive (i.e. $d = 0$). Consequently this implies that if $b > 0$ then FDP $= 1$. Also notice that if $b = 0$ then the FDP is just $0/0$ which we define to be 0. Keeping this in mind, and writing $p_i = P(b = i)$,

$$\begin{aligned}
\text{FDR} &= E[\text{FDP}] \\
&= E\Big[E[\text{FDP} \,|\, b]\Big] \\
&= p_0 E[\text{FDP} \,|\, b = 0] + \sum_{i=1}^{N} p_i E[\text{FDP} \,|\, b = i] \\
&= p_0 \times 0 + \sum_{i=1}^{N} p_i \times 1 \\
&= \sum_{i=1}^{N} p_i \\
&= P(b \geq 1) \\
&= \text{FWER}
\end{aligned}$$

For (2) we have that $N_0 < N$. One thing to note here is that if $b = 0$ then, as before, the FDP is 0. If $b > 0$ however we don't necessarily have that the FDP $= 1$ but we do know that it is ≤ 1.

(since the denominator is always at least as big as the numerator). The logic then follows almost identically as previously (only one line changed to an inequality),

$$\begin{aligned}
\text{FDR} &= E[\text{FDP}] \\
&= E\Big[E[\text{FDP} \mid b]\Big] \\
&= p_0 E[\text{FDP} \mid b = 0] + \sum_{i=1}^{N} p_i E[\text{FDP} \mid b = i] \\
&\leq p_0 \times 0 + \sum_{i=1}^{N} p_i \times 1 \\
&= \sum_{i=1}^{N} p_i \\
&= P(b \geq 1) \\
&= \text{FWER}
\end{aligned}$$

13.6 Notes on the FDR

From the second key property, any procedure that controls the FWER also controls the FDR. However a procedure that only controls the FDR may be less stringent and may gain in power over the more restrictive requirements needed to control the FWER.

13.7 The Benjamini–Hochberg (BH) Procedure

This procedure controls the FDR at a level α when the P-values of the true null hypotheses are independent and uniformly distributed.

Note that here α denotes the level the FDR is controlled at, which is not the same as the significance level of each test (the lecture notes use α for this).

The procedure:

1. Let $P_{(1)}, \ldots, P_{(N)}$ be the ordered P-values.
2. Calculate
$$i_1 = \max\left\{i \in \{1, \ldots, N\} : P_{(i)} \leq \frac{i\alpha}{N}\right\}$$
3. If i_1 exists then reject the null hypotheses corresponding to,
$$P_{(1)}, \ldots, P_{(i_1)}$$
(i.e. all hypotheses with p-values less than or equal to $P_{(i_1)}$).

Some notes:

- Notice how $P_{(i)}$ only needs to be smaller than $\frac{i\alpha}{N}$ to be significant. This is less strict than Bonferroni, where all p-values need to be less than $\frac{\alpha}{N}$.
- In 2001 Benjamini & Yekutieli amended the procedure for situations where the tests are not necessarily independent. The only difference is that $\frac{i\alpha}{N}$ becomes $\frac{i\alpha}{C(N)N}$ where $C(N) = \sum_{i=1}^{N} \frac{1}{i}$. Generally here though we are just going to work with the simpler case without this $C(N)$ factor.

13.8 FDR Adjustment

The lecture notes also cover FDR-adjusted p-values, with a worked example on the last page (where it is worth understanding what each column of the table represents). This isn't written up here.

14 Errors in Testing

Suppose you have some test T for testing class A vs class B (where A and B are taken to be disjoint partitions of the sample space).

- If T is positive then we predict the data is from class A
- If T is negative then we predict the data is from class B

For example T could be a test for whether someone has a particular disease. Or more generally this could describe a typical hypothesis testing scenario where A and B correspond to H_0 and H_1.

Here is some useful terminology with regards to the test,

- Sensitivity: $P(T \mid A)$
- Specificity: $P(\neg T \mid B)$

In the framework of hypothesis testing, the sensitivity is equal to $1 - \alpha$ and the specificity is $1 - \beta$ (better known as the power).

You also have:

- Positive Predictive Value (PPV): $P(A \mid T)$
- Negative Predictive Value (NPV): $P(B \mid \neg T)$

(The rest of this lecture isn't written up here.)

15 Tests for Normal Means and Variances

15.1 Two-Sample t-Test

A random sample of size n_i is taken from class i with a $N(\mu_i, \sigma^2)$ distribution for $i \in \{1, 2\}$. I.e. the same variance but different means. Let $x_{i1}, \ldots, x_{in_i}$ denote a random sample from class i. For ease of notation let $N = n_1 + n_2$.

We wish to test,

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

We shall show that the LRT leads to the classical two-sample t-test where,

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and s^2 is the pooled sample variance. This follows a t_{N-2} distribution under H_0.

The test is then,

$$\text{reject } H_0 \text{ if } \quad P_{H_0}(|T| \geq |t|) < \alpha$$

(where t is the observed value of T).

15.2 Using the LRT

The whole parameter space Ω is,

$$\Omega = \{(\mu_1, \mu_2, \sigma^2)^T : \mu_i \in \mathbb{R},\, \sigma^2 > 0\}$$

On the other hand under H_0 we have,

$$\Omega_{H_0} = \{(\mu_1, \mu_2, \sigma^2)^T : \mu_1 = \mu_2 = \mu,\, \sigma^2 > 0\}$$

The likelihood ratio statistic λ is,

$$\begin{aligned}\lambda &= \frac{\sup\limits_{\boldsymbol{\theta}\in\Omega_{H_0}} L(\boldsymbol{\theta})}{\sup\limits_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})} \\ &= \frac{L(\widehat{\widehat{\boldsymbol{\theta}}})}{L(\widehat{\boldsymbol{\theta}})}\end{aligned}$$

(here $\widehat{\widehat{\boldsymbol{\theta}}}$ denotes the maximiser over Ω_{H_0}, following the notation of the lecture notes).

15.2.1 Finding $\widehat{\theta}$ (i.e. the MLE)

$$\begin{aligned}L(\boldsymbol{\theta}) &= (2\pi)^{\frac{-n_1}{2}}(\sigma^2)^{\frac{-n_1}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n_1}(x_{1j}-\mu_1)^2\right) \\ &\quad \times (2\pi)^{\frac{-n_2}{2}}(\sigma^2)^{\frac{-n_2}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2\right) \\ &= (2\pi)^{-\frac{N}{2}}(\sigma^2)^{-\frac{N}{2}} \exp\left(\frac{-1}{2\sigma^2}\left[\sum_{j=1}^{n_1}(x_{1j}-\mu_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2\right]\right) \\ \log L(\boldsymbol{\theta}) &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{j=1}^{n_1}(x_{1j}-\mu_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2\right]\end{aligned}$$

For μ_1,

$$\begin{aligned}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \mu_1} &= \frac{1}{\sigma^2}\sum_{j=1}^{n_1}(x_{1j}-\mu_1) \\ 0 &= \frac{1}{\widehat{\sigma}^2}\sum_{j=1}^{n_1}(x_{1j}-\widehat{\mu}_1) \\ \widehat{\mu}_1 &= \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} \\ &= \overline{x_1}\end{aligned}$$

(by symmetry the MLE for μ_2 is $\widehat{\mu}_2 = \overline{x_2}$)

For σ^2,

$$\begin{aligned}
\frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma^2} &= \frac{-N}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{j=1}^{n_1}(x_{1j}-\mu_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2\right] \\
0 &= \frac{-N}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\left[\sum_{j=1}^{n_1}(x_{1j}-\widehat{\mu}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\widehat{\mu}_2)^2\right] \\
\widehat{\sigma}^2 &= \frac{1}{N}\left[\sum_{j=1}^{n_1}(x_{1j}-\overline{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\overline{x}_2)^2\right] \\
&= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{N}
\end{aligned}$$

which can alternatively be written as,

$$\widehat{\sigma}^2 = \frac{N-2}{N}s^2$$

where s^2 is the unbiased pooled sample variance.

So the MLE is,

$$\widehat{\mu}_1 = \overline{x}_1 \qquad \widehat{\mu}_2 = \overline{x}_2 \qquad \widehat{\sigma}^2 = \frac{N-2}{N}s^2$$

15.2.2 Finding $\widehat{\widehat{\boldsymbol{\theta}}}$ (the maximiser over Ω_{H_0})

$$L(\boldsymbol{\theta}) = (2\pi)^{\frac{-n_1}{2}}(\sigma^2)^{\frac{-n_1}{2}}\exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n_1}(x_{1j}-\mu)^2\right)(2\pi)^{\frac{-n_2}{2}}(\sigma^2)^{\frac{-n_2}{2}}\exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{n_2}(x_{2j}-\mu)^2\right)$$

$$\log L(\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{j=1}^{n_1}(x_{1j}-\mu)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu)^2\right]$$

For μ,

$$\begin{aligned}
\frac{\partial \log L(\boldsymbol{\theta})}{\partial \mu} &= \frac{1}{\sigma^2}\left[\sum_{j=1}^{n_1}(x_{1j}-\mu) + \sum_{j=1}^{n_2}(x_{2j}-\mu)\right] \\
0 &= \frac{1}{\widehat{\sigma}^2}\left[\sum_{j=1}^{n_1}(x_{1j}-\widehat{\mu}) + \sum_{j=1}^{n_2}(x_{2j}-\widehat{\mu})\right] \\
N\widehat{\mu} &= \sum_{j=1}^{n_1}x_{1j} + \sum_{j=1}^{n_2}x_{2j} \\
\widehat{\mu} &= \frac{x_{1\bullet} + x_{2\bullet}}{N} \\
&= \overline{x_{\bullet\bullet}} \\
&= \overline{x} \quad \text{(as it is written in the lecture notes)}
\end{aligned}$$

For σ^2,

$$\begin{aligned}
\frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{j=1}^{n_1}(x_{1j}-\mu)^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu)^2\right] \\
0 &= -\frac{N}{2\widehat{\widehat{\sigma}}^2} + \frac{1}{2\widehat{\widehat{\sigma}}^4}\left[\sum_{j=1}^{n_1}(x_{1j}-\widehat{\mu})^2 + \sum_{j=1}^{n_2}(x_{2j}-\widehat{\mu})^2\right] \\
\widehat{\widehat{\sigma}}^2 &= \frac{1}{N}\left[\sum_{j=1}^{n_1}(x_{1j}-\overline{x})^2 + \sum_{j=1}^{n_2}(x_{2j}-\overline{x})^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2
\end{aligned}$$

so,

$$\widehat{\mu} = \overline{x} \qquad\qquad \widehat{\widehat{\sigma}}^2 = \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2$$

15.2.3 Constructing the test

$$\begin{aligned}
\lambda &= \frac{L(\widehat{\widehat{\boldsymbol{\theta}}})}{L(\widehat{\boldsymbol{\theta}})} \\
&= \frac{(2\pi)^{\frac{-N}{2}}(\widehat{\widehat{\sigma}}^2)^{\frac{-N}{2}}\exp\left(\frac{-1}{2\widehat{\widehat{\sigma}}^2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2\right)}{(2\pi)^{\frac{-N}{2}}(\widehat{\sigma}^2)^{\frac{-N}{2}}\exp\left(\frac{-1}{2\widehat{\sigma}^2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2\right)} \\
&= \left(\frac{\widehat{\widehat{\sigma}}^2}{\widehat{\sigma}^2}\right)^{-\frac{N}{2}}\exp\left(\frac{-1}{2\widehat{\widehat{\sigma}}^2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2 + \frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2\right) \\
&= \left(\frac{\widehat{\widehat{\sigma}}^2}{\widehat{\sigma}^2}\right)^{-\frac{N}{2}}\exp\left(\frac{-N\widehat{\widehat{\sigma}}^2}{2\widehat{\widehat{\sigma}}^2} + \frac{N\widehat{\sigma}^2}{2\widehat{\sigma}^2}\right) \\
&= \left(\frac{\widehat{\widehat{\sigma}}^2}{\widehat{\sigma}^2}\right)^{-\frac{N}{2}} \\
&= \left(\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2}{(N-2)s^2}\right)^{\frac{-N}{2}}
\end{aligned}$$

With the numerator here we can use a very handy result,

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2 = (N-2)s^2 + \frac{(\overline{x}_1-\overline{x}_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}}$$

(proved at the end of this subsection).

Then,

$$\lambda = \left(\frac{(N-2)s^2 + \frac{(\overline{x}_1 - \overline{x}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}}{(N-2)s^2} \right)^{\frac{-N}{2}}$$
$$= \left(1 + \frac{1}{N-2} \frac{(\overline{x}_1 - \overline{x}_2)^2}{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right)^{\frac{-N}{2}}$$
$$= \left(1 + \frac{t^2}{N-2} \right)^{\frac{-N}{2}}$$

where,

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Given that λ is monotone decreasing in $|t|$, it will be small when $|t|$ is large. So reject the null hypothesis if $|t|$ is large.

(where this t matches the observed t-test statistic for the two-sided t-test).

15.2.4 The Handy Result

The proof in the lecture notes starts at the very end of page 3. One way to see it is to split each deviation as $x_{ij} - \overline{x} = (x_{ij} - \overline{x}_i) + (\overline{x}_i - \overline{x})$. The cross terms vanish because $\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i) = 0$, so

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 = \underbrace{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2}_{(N-2)s^2} + \sum_{i=1}^{2} n_i(\overline{x}_i - \overline{x})^2$$

Since $\overline{x} = (n_1\overline{x}_1 + n_2\overline{x}_2)/N$, we have $\overline{x}_1 - \overline{x} = \frac{n_2}{N}(\overline{x}_1 - \overline{x}_2)$ and $\overline{x}_2 - \overline{x} = -\frac{n_1}{N}(\overline{x}_1 - \overline{x}_2)$. Therefore,

$$\sum_{i=1}^{2} n_i(\overline{x}_i - \overline{x})^2 = \frac{n_1 n_2^2 + n_2 n_1^2}{N^2}(\overline{x}_1 - \overline{x}_2)^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{x}_1 - \overline{x}_2)^2 = \frac{(\overline{x}_1 - \overline{x}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

which gives the result.

15.3 Example: Tests on the Variance

Consider a random sample,

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

where σ^2 is unknown but μ is known (until part (v)).

15.3.1 (i) The Most Powerful Test

Here we want to find the MP test of,

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \text{vs} \qquad H_1 : \sigma^2 = \sigma_1^2 \quad (\text{where } \sigma_1^2 > \sigma_0^2)$$

By the Neyman–Pearson lemma the CR of the MP test is given by,

$$\left\{ \lambda = \frac{L(\sigma_0^2)}{L(\sigma_1^2)} < k \right\}$$

Looking at the ratio,

$$\begin{aligned}\lambda &= \frac{(2\pi)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{n}{2}}\exp\left(\frac{-1}{2\sigma_0^2}\sum_{j=1}^n(x_j-\mu)^2\right)}{(2\pi)^{-\frac{n}{2}}(\sigma_1^2)^{-\frac{n}{2}}\exp\left(\frac{-1}{2\sigma_1^2}\sum_{j=1}^n(x_j-\mu)^2\right)} \\ &= \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{\frac{-n}{2}}\exp\left(\frac{-1}{2\sigma_0^2}\sum_{j=1}^n(x_j-\mu)^2+\frac{1}{2\sigma_1^2}\sum_{j=1}^n(x_j-\mu)^2\right) \\ &= \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{\frac{-n}{2}}\exp\left(\left[\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_0^2}\right]\sum_{j=1}^n(x_j-\mu)^2\right)\end{aligned}$$

Given that $\sigma_1^2 > \sigma_0^2$ then $\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2} < 0$. Therefore λ is going to be small when,

$$T = \sum_{j=1}^n (x_j - \mu)^2$$

is large.

So, the critical region R is given by,

$$R = \left\{\sum_{j=1}^n (x_j-\mu)^2 > k^*\right\}$$

The value k^* is chosen such that the type I error of the test is α. Under H_0 the test statistic T is distributed as,

$$\frac{1}{\sigma_0^2}T \sim \chi_n^2$$

Therefore,

$$\begin{aligned}\alpha &= P_{H_0}(T > k^*) \\ &= P\left(\chi_n^2 > \frac{k^*}{\sigma_0^2}\right) \\ \frac{k^*}{\sigma_0^2} &= \chi_{n;1-\alpha}^2 \\ k^* &= \sigma_0^2\chi_{n;\,1-\alpha}^2\end{aligned}$$

so,

$$R = \left\{\sum_{j=1}^n (x_j-\mu)^2 > \sigma_0^2\chi_{n;\,1-\alpha}^2\right\}$$

15.3.2 (ii) A UMP Test

Given that the critical region of the test is independent of the alternative hypothesis value σ_1^2 the test above will already be UMP for testing $H_1 : \sigma^2 > \sigma_0^2$.

15.3.3 (iii) A Confidence Interval

Here we need to construct a $100(1-\alpha)\%$ confidence interval for σ^2.

We can use,

$$\frac{T}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^n (x_j-\mu)^2$$

as a pivot variable, since its distribution, χ_n^2, does not depend on σ^2.

Then, with probability $1-\alpha$,

$$\chi^2_{n;\,\alpha/2} \le \frac{T}{\sigma^2} \le \chi^2_{n;\,1-\alpha/2}$$
$$\sigma^2\chi^2_{n;\,\alpha/2} \le T \le \sigma^2\chi^2_{n;\,1-\alpha/2}$$
$$\frac{T}{\chi^2_{n;\,1-\alpha/2}} \le \sigma^2 \le \frac{T}{\chi^2_{n;\,\alpha/2}}$$

and so the last line gives the confidence interval for σ^2.

15.3.4 (iv)

This can be constructed from the two one-sided tests above, adjusting the constants k so that the overall significance level is α. Not much new is needed.

15.3.5 (v) The GLRT with Unknown Mean

Now suppose μ is also unknown. First we need the MLE for σ^2 and μ (unrestricted)

$$\widehat{\mu} = \overline{x} \qquad \text{and} \qquad \widehat{\sigma}^2 = \frac{n-1}{n}s^2$$

Now we also need the MLEs subject to the constraint that $\sigma^2 = \sigma_0^2$. Then,

$$L(\mu) = (2\pi)^{\frac{-n}{2}}(\sigma_0^2)^{\frac{-n}{2}}\exp\left(\frac{-1}{2\sigma_0^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$
$$\log L(\mu) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{j=1}^{n}(x_j-\mu)^2$$
$$\frac{\partial \log L(\mu)}{\partial \mu} = \frac{1}{\sigma_0^2}\sum_{j=1}^{n}(x_j-\mu)$$
$$\implies \widehat{\mu} = \overline{x}$$

(it makes sense that "knowing" that $\sigma^2 = \sigma_0^2$ doesn't affect your best estimate of μ).

Then,

$$\lambda = \frac{(2\pi)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{n}{2}}\exp\left(\frac{-1}{2\sigma_0^2}\sum_{j=1}^{n}(x_j-\overline{x})^2\right)}{(2\pi)^{-\frac{n}{2}}(\frac{n-1}{n}s^2)^{-\frac{n}{2}}\exp\left(\frac{-1}{2\frac{n-1}{n}s^2}\sum_{j=1}^{n}(x_j-\overline{x})^2\right)}$$
$$= \left(\frac{n\sigma_0^2}{(n-1)s^2}\right)^{-\frac{n}{2}}\exp\left(\left[\frac{n}{2(n-1)s^2} - \frac{1}{2\sigma_0^2}\right]\sum_{j=1}^{n}(x_j-\overline{x})^2\right)$$

We reject when λ is small. (The rest of this part isn't worked through here. It is easier not to substitute the MLEs too early: work with σ_0^2 and $\widehat{\sigma}^2$, and only use $\widehat{\sigma}^2 = \frac{n-1}{n}s^2$ at the end.)

15.4 Satterthwaite's Approximation

Also known as the "Welch–Satterthwaite approximation".

15.4.1 Setup

Suppose you have k different populations (each with potentially different means and standard deviations). You then collect k independent samples where for sample i you have:

- n_i observations
- s_i^2 your calculated sample variance
- σ_i^2 the unknown variance of that population

You may want to estimate a combined variance by taking a linear combination of the sample variances,

$$s^2 = \sum_{i=1}^{k} \alpha_i s_i^2$$

A typical choice for α_i in this situation for example may be $\alpha_i = \frac{1}{n_i}$.

But the question is, what is the distribution of s^2?

15.4.2 The Satterthwaite Approximation

Firstly notice that each individual s_i^2 is distributed,

$$\frac{n_i - 1}{\sigma_i^2} s_i^2 \sim \chi^2_{n_i - 1}$$

However the distribution of s^2 as defined above does not have a simple closed form. Instead, the distribution can be approximated by a scaled chi-squared distribution where,

$$\frac{\nu}{\sigma^2} s^2 \overset{\text{approx.}}{\sim} \chi^2_\nu \qquad \text{with} \qquad \sigma^2 = E[s^2] = \sum_{i=1}^{k} \alpha_i \sigma_i^2$$

where,

$$\nu = \frac{(\sum_{i=1}^{k} \alpha_i s_i^2)^2}{\sum_{i=1}^{k} \frac{(\alpha_i s_i^2)^2}{\nu_i}} \qquad \text{where } \nu_i = n_i - 1$$

This is known as Satterthwaite's approximation.

15.4.3 Application to the Two-Sample t-Test

Suppose you want to test the difference in the means of two populations with different variances, estimated by s_1^2 and s_2^2. A natural test statistic would be,

$$T = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

To use this statistic though we need to know its distribution.

By Satterthwaite's approximation we know that,

$$\nu \frac{s_1^2/n_1 + s_2^2/n_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2} \overset{\text{approx}}{\sim} \chi^2_\nu$$

where,

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{1}{n_1 - 1}(s_1^2/n_1)^2 + \frac{1}{n_2 - 1}(s_2^2/n_2)^2}$$

Given this we know that, under $H_0 : \mu_1 = \mu_2$, T will approximately follow a t distribution where,

$$T \overset{\text{approx}}{\sim} t_\nu$$

16 Tukey's Range Test

Suppose we observe a random sample,

$$X_1, \ldots, X_k \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

Now consider the sample range R which is,

$$R = X_{(k)} - X_{(1)}$$

A natural question then may be: what is the distribution of R?

16.1 The Studentised Range Distribution

Tukey's studentised range statistic is defined as,

$$T = \frac{R}{s}$$

where s is an estimator of the standard deviation σ, independent of $X_1, \ldots, X_k$, with

$$\frac{\nu}{\sigma^2} s^2 \sim \chi^2_\nu$$

The distribution of Tukey's studentised range statistic T depends on both k (the number of groups) and ν. We shall let $T_{k,\nu;\alpha}$ denote the quantile of order α of the distribution.

16.2 Application: Tukey's Simultaneous Confidence Intervals

Suppose we observe k random samples,

$$X_{i1}, \ldots, X_{in_i} \overset{\text{i.i.d.}}{\sim} N(\mu_i, \sigma^2) \quad i = 1, \ldots, k$$

For all $i = 1, \ldots, k$, let $\widehat{\mu}_i$ be an estimator for μ_i such that,

$$\widehat{\mu}_i \sim N(\mu_i, c^2\sigma^2)$$

for some constant c that is the same for every i (for example $c^2 = \frac{1}{n}$ when every group has n observations), with $\widehat{\mu}_1, \ldots, \widehat{\mu}_k$ independent.

Furthermore let s^2 be an estimator of σ^2 distributed as,

$$\frac{\nu}{\sigma^2} s^2 \sim \chi^2_\nu$$

(which is assumed to be distributed independently of each $\widehat{\mu}_i$).

Then, Tukey's simultaneous confidence interval gives us the following $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu_j$ for all pairs $i \neq j$,

$$(\widehat{\mu}_i - \widehat{\mu}_j) - cT_{k,\nu;1-\alpha}s \leq \mu_i - \mu_j \leq (\widehat{\mu}_i - \widehat{\mu}_j) + cT_{k,\nu;1-\alpha}s$$

Such a procedure is valuable because it controls the FWER at α across all the pairwise comparisons. This makes Tukey's approach favourable in this situation over repeated t-tests with Bonferroni correction.

Proof. Assume the setup from above and let,

$$U_i = \frac{1}{c}(\widehat{\mu}_i - \mu_i) \quad \text{for } i = 1, \ldots, k$$

Then notice that,

$$U_i \sim N(0, \sigma^2)$$

It then follows that,

$$\frac{U_{(k)} - U_{(1)}}{s} \sim T_{k,\nu}$$

Given this we know that,

$$P\left(\frac{U_{(k)} - U_{(1)}}{s} \leq T_{k,\nu;1-\alpha}\right) = 1 - \alpha$$

The range is the largest of the differences $|U_i - U_j|$, so the range being at most $T_{k,\nu;1-\alpha}\, s$ is the same event as every difference being at most $T_{k,\nu;1-\alpha}\, s$. Therefore,

$$P\left(\frac{|U_i - U_j|}{s} \leq T_{k,\nu;1-\alpha} \quad \text{for all } i \neq j\right) = 1 - \alpha$$

On substituting the definition of U_i and U_j and rearranging we get the desired confidence intervals, which hold simultaneously with probability $1 - \alpha$. □

16.3 Extension to Contrasts

Now we consider a more general case. Suppose you have a contrast $\{c_i\}$ which by definition must satisfy,

$$\sum_{i=1}^{k} c_i = 0$$

Let ψ be the contrast in the parameter μ_i and let $\widehat{\psi}$ be an estimator for ψ,

$$\psi = \sum_{i=1}^{k} c_i \mu_i \quad \text{and} \quad \widehat{\psi} = \sum_{i=1}^{k} c_i \widehat{\mu}_i$$

It can then be shown that a $100(1-\alpha)\%$ confidence interval for ψ can be constructed as,

$$\widehat{\psi} - cT_{k,\nu;1-\alpha} sh \leq \psi \leq \widehat{\psi} + cT_{k,\nu;1-\alpha} sh$$

where,

$$h = \frac{1}{2}\sum_{i=1}^{k} |c_i|$$

Important points:

- This confidence interval holds simultaneously for *all* possible contrasts $\{c_i\}$ with probability exactly $1 - \alpha$.
- The FWER is controlled at α across the infinitely many possible contrasts
- The example above on Tukey's simultaneous confidence intervals was just a special case of this where we looked at the pair-wise comparison contrasts. The confidence interval for $\psi = \mu_i - \mu_j$ just corresponds to the contrast c with a 1 in the ith position, a -1 in the jth position and 0 elsewhere.

(The proof isn't written up here. In the lecture notes it starts at the very end of page 2, on the line "To establish (17)".)

16.4 Applications to ANOVA

Tukey's range test is often applied to ANOVA since it can control the FWER across all comparisons between the groups. See page 4 of the lecture notes for more details.