1. Checking Normality

We will talk later on about Confidence Intervals, Hypothesis Testing etc. Often there lies the assumptions that the data follow a (multivariate) normal distribution. It is, therefore, important to develop criteria that will test this assumption.

1.1. **Univariate Data.** Probability plots are a useful graphical tool for qualitatively assesing the fit of empirical data to a theoretical probability distribution. Consider sample data $y_1, y_2, \ldots, y_n$ of size n from a uniform distribution on $[0, 1]$. Then *order* the sample data to a increasing order, as $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. The values of a sample data, ordered in the above fashion are called the **order statistics** of the sample. It can be shown (see exercises) that

$$E[y_{(i)}] = \frac{j}{n+1}$$

Therefore, if we plot the observation $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ against the points $1/(n+1), 2/(n+1), \ldots, n/(n+1)$ and assuming that the underlying distribution is uniform, we expect the plotted data to look roughly linear.

We could extend this technique to other continuous distributions. The theoretical approach is the following. Suppose that we have a continuous random variable with strictly increasing cumulative distribution function F_Y. Consider the random variable $X = F_Y(Y)$. Then it follows that the distribution of X is the uniform on $[0, 1]$ (check this !). We can therefore follow the previous procedure and plot the data $F(Y_{(k)})$ vs $k/(n+1)$, or equivallently the data $Y_{(k)}$ vs. $F^{-1(k/(n+1))}$. Again, if the sample data follow the cumulative distribution F_Y, we should observe an almost linear plot.

In Figure 1 we see the probability plot of a sample of 100 random data generated from a Normal distribution. The probability plot (often called in this case **normal plot**) is, indeed, almost linear.

In Figure 2 we drew the probability plot of a sample of 100 random data generated from an exponential distribution, i.e. a distribultion with cumulative distribution function $1 - e^{-x}$, versus a normal distribution. The (normal) probability plot that we obtain is cleary non linear. In particular we see that both the right and left tails of the plot are superlinear (notice that in the middle it is sort of linear). The reason for this is that the exponential distribution has *heavier* tails than the normal distribution. For example the probability to get a very large value is much larger if the data follow an exponential distribution, than a normal distribution. This is because for large x we have $e^{-x} >> e^{-x^2/2}$. There the frequency of large value in a sample that follows an exponential distribution would be much larger, than a sample of normal distribution. The same happens near zero. The sample data of the exponential distribution are superlinear, since the probaility density of an exponential near zero is almost 1, while for a normal it is almost $1/\sqrt{2\pi}$.

1.1.1. *Some more graphs.* We have collected the data from the stock value of the Barclays Bank, from 16 June 2006 until 10 October 2008. Below we have drawn the histograms of the returns $R(t) = (P(t) - P(t-1))/P(t-1)$, where $P(t)$ is the value of the stock at time t. In Figures 3,4 we exhibit histograms with different binning size. You see that as the size of the bin get small the histogram get finer and resembles more a continuous distribution. In Figure 5 we tried to fit a Normal distribution and in the Figure 6 we drew the normal probability plot.

1.2. **Bivariate Data.** Consider a vector $\underline{X} = (X_1, X_2)^T$, whose distribution is bivariate normal $\mathcal{N}_2(0, \mathbf{\Sigma})$. Then we know that the level curves of the bivariate normal are ellipses. Therefore any scatter plot of data chosen from this distribution should resemble an ellipses.

1.3. **Multivariate Date.** Consider a vector $\underline{X} = (X_1, X_2, \ldots, X_p)^T \sim \mathcal{N}_p(0, \mathbf{\Sigma})$, then as in the previous case, the scatter plot of the date chosen from this distribution should resemble an ellipsoid. Nevertheless, it is difficult to visualise scatter plots in high dimensions. To go around this problem we can check (i) the univariate marginals, using the Normal probability plots, (ii) the bivariate marginals using two dimensional scatter plots.

A remark should be in order: Namely, one can construct bivariate non-Normal distributions, whose one dimensional marginals are normal (see exercises) !

1.4. **Altenative to scatter plots.** These methods are based on the following proposition

Proposition 1. *Let* $\underline{X} \sim \mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$ *with* $det(\mathbf{\Sigma}) > 0$. *Then*

(a) $(\underline{X} - \underline{\mu})^T \mathbf{\Sigma}^{-1} (\underline{X} - \underline{\mu})$ *is distributed as a chi-square distribution with* p *degrees of freedom,* χ^2_p.

(b) The $\mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$ *assigns probability* $(1 - \alpha)$ *to the ellipsoid* $\{\underline{x} \in \mathbb{R}^p \colon (\underline{x} - \underline{\mu})^T \mathbf{\Sigma}^{-1} (\underline{x} - \underline{\mu}) \leq \chi^2_p(\alpha)\}$, *where* $\chi^2_p(\alpha)$ *denotes the upper* 100α *percentile of the* χ^2_p.

Proof. We refer to [JW] page 163.

A consequence of the above proposition is that if we have a sample $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ of a multivariate normal distribution then about half of the observations should lie within the set

$$\{\underline{x} \in \mathbb{R}^p \colon (\underline{x} - \overline{\underline{X}})^T \mathbf{S}^{-1} (\underline{x} - \overline{\underline{X}}) \leq \chi^2_p(0.5)\}.$$

Example. Consider the list of observations x_1 =sales, x_2=profits of the 10 largest companies in the world

	x_1(sales)	x_2(profits)	x_3(assets)
Citigroup	108.28	17.05	1,484.10
General Electric	152.36	16.59	750.33
American Group	95.04	10.91	766.42
Bank of America	65.45	14.14	1,110.46
HSBC	62.97	9.52	1,031.29
Exxon	263.99	25.33	195.26
Royal Dutch	265.19	18.54	193.83
BP	285.06	15.73	191.11
ING	92.01	8.10	1,175.16
Toyota	165.68	11.13	211.15

This sample is obviously not random, but we can pretend so. We can then compute the sample mean

$$\overline{\underline{X}} = \begin{pmatrix} 155.60 \\ 14.70 \end{pmatrix}$$

and the sample covariance matric

$$\mathbf{S} = \begin{pmatrix} 7476.45 & 303.62 \\ 303.63 & 26.19 \end{pmatrix}$$

computing the $\mathbf{S}^{-1}$ we get

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.000253 & -0.002930 \\ -0.002930 & 0.072148 \end{pmatrix}$$

From the table of the chi-square distribution we conclude that $\chi^2_2(0.5) = 1.39$ and therefore the 50% region is given by

$$\{\underline{x} \in \mathbb{R}^2 \colon (\underline{x} - \overline{\underline{X}})^T \mathbf{S}^{-1} (\underline{x} - \overline{\underline{X}}) \leq 1.39\}.$$

Entering the numbers from the above data we see that 6/10 do not fall into this region. This might provide evidence that the joint distribution of the sales and profits is not bivariate normal. Of course, one should also bear in mind that a sample of size ten is too small...

2. Confidence Intervals and Regions

2.1. **Univariate case.** Suppose that we have a Normal $N(\mu, \sigma^2)$ distribution with unkown mean and and standard deviation. We know that the sample mean and variance are estimators of the population mean and variance. These estimators tell what is *a most likely* value for these parameters. However it is very unlikely that these estimators will produce the *exact value* of the inknown parameter. It would be more desirable to produce a range of value, such that the unkown parameter will lie within these values with *high probability*. This is achieved by the construction of *confidence intervals*. We will defer a formal defintion for later on, after we see the philosophy behind the construction of the confidence interval.

2.1.1. *Confidence Interval for Mean with Known Standard Deviation.*

Suppose we have a distribution, not necessarily normal, with known variance σ^2, but unkown mean μ. Suppose that we form a sample $X_1, X_2, \ldots, X_n$ from this distribution. Then the sample mean

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

will have variance $\mathrm{Var}(\overline{X}) = \sigma^2/n$ and mean $E[\overline{X}] = \mu$, while the Central Limit Theorem will tell us that the distribution of

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal, that is

$$\begin{aligned} P\left(-z < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z\right) &\simeq \Phi(z) - \Phi(-z) \\ &\simeq 2\Phi(z) - 1 \end{aligned}$$

where $\Phi(z) = \int_{-\infty}^{z} e^{-x^2/2}/\sqrt{2\pi}$. By a simple manipulation in the above we get that

$$P\left(\overline{X} - z\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z\frac{\sigma}{\sqrt{n}}\right) \simeq 2\Phi(z) - 1.$$

Suppose now that we choose $z = z_{\alpha/2}$ such that $2\Phi(z_{\alpha/2}) - 1 = 1 - \alpha$ (this equation cannot be solved explicitly, but there are tables that give you the values $z_{\alpha/2}$ for different values of α, or you can use some statistical software) then

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \simeq 1 - \alpha.$$

So, for example, if $\alpha = 0.05$, we obtain from the tables that $z_{0.025} \simeq 1.96$ and therefore the interval $[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}]$ will contain the unkown value of the mean μ with probability 0.95.

2.1.2. *Confidence Interval for Mean with Unkown Standard Deviation.*

Suppose now that we want to construct a conidence interval for the mean of a distribution, but this thime we don't know its standard deviation. In this case the variance σ^2 in the $(1-\alpha)$-confidence interval $\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ should be replaced by an estimator, which we choose to be the sample variance $\hat{s}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$. In other words we would tend to say that the $(1-\alpha)-$confidence interval for the mean μ is

$$\left[\overline{X} - z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}\right].$$

This is not exactly correct, though. The reason is that when we replace σ by s i.e. when we consider the fraction

$$\frac{\overline{X} - \mu}{\hat{s}/\sqrt{n}}$$

the correct approximation of the distribution of this random variable is not the standard normal distribution, but rather the t_{n-1}, the t-distribution with $(n-1)$ degrees of freedom. Therefore the $z_{\alpha/2}$ normal quantiles should be replaced with the corresponding $t_{n-1,\alpha/2}$ quantiles of the t-distribution with $(n-1)$ degrees of freedom. The correct $(1-\alpha)$ -confidence interval in this case is

$$\left[\overline{X} - t_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}\right].$$

We will now attempt to give an explanation as to why the t-distribution is the correct distribution to consider, rather than the normal distribution.

Assume that the underlying distribution is normal $N(\mu, \sigma^2)$, with unkown mean and standard deviation. Suppose that a sample of size n, $X_1, X_2, \ldots, X_n$ is drawn from this distribution and consider the fraction

$$\frac{\overline{X} - \mu}{\hat{s}/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)/\sigma}{\hat{s}/\sigma}. \tag{1}$$

Then we claim that the distribution of the above random variable is *exaclty* the t_{n-1}-distribution with $(n-1)$ degrees of freedom.

To prove this we need the following very interesting lemma

Lemma 1. *Consider a sequence $X_1, X_2, \ldots, X_n$ of i.i.d. standard normal variables. Then the sample average $\overline{X}$ is independent of the random vector $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$.*

The proof of this fact is not difficult, but we will skip it since it is a bit lengthy. The detailed proof can be found in the book of Rice, Section 6.3. Let us just say

that to prove the statement it is enough to prove the for any numbers $u, u_1, \ldots, u_n$ it holds that

$$E\left[e^{u\overline{X}+\sum_{i=1}^n u_i(X_i-\overline{X})}\right] = E\left[e^{u\overline{X}}\right]\ E\left[e^{\sum_{i=1}^n u_i(X_i-\overline{X})}\right].$$

Then, clearly $\sqrt{n}(\overline{X}-\mu)/\sigma$ is a standard normal distribution. On the other hand we have

Lemma 2. *If $X_1, X_2, \ldots, X_n$ are i.i.d. normal $N(\mu, \sigma^2)$ then the distribution of*

$$(n-1)\frac{\hat{s}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2$$

is a χ^2_{n-1} distribution, with $(n-1)$ degrees of freedom.

Proof. Note that

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 = \sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$$

as a sum of the squares of n i.i.d. standard normals. Moreover,

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 &= \frac{1}{\sigma^2}\sum_{i=1}^n ((X_i-\overline{X})+(\overline{X}-\mu))^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^n ((X_i-\overline{X}))^2 + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2,\end{aligned}$$

where we also used the fact the $\sum_{i=1}^n (X_i - \overline{X}) = 0$. The above equation is of the form $W = U + V$, where U,V are independent by the previous lemma. Also, W, V have distributions χ^2_n, χ^2_1, respectively. If $M_W(t)$ denotes the moment generating function of W, and similarly for U, V we have by independence that

$$M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2},$$

where we used the fact that the moment generating function of a χ^2_n with n degrees of freedom is $(1-2t)^{-n/2}$. □

From the above Lemma, as well as the definition of a t-distribution, it follows that the distribution of (1) is exactly the t_{n-1} -distribution.

2.1.3. *Confidence Intervals for the Variance.*

Let us consider the particular case of an i.i.d. **Normal sample** $X_1, X_2, \ldots, X_n$. Let $\hat{\sigma}^2$ the maximum likelihood estimator of the variance, i.e.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Then by Lemma 2 we have that

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

Let us denote by $\chi^2_{m,\alpha}$, the chi square quantile, i.e. the point beyond which the chi square distribution with m degrees of freedom has probability α. Then we have

$$P\left(\chi^2_{n-1,1-\alpha/2} < \frac{n\hat{\sigma}^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right) = 1-\alpha$$

and solving for σ^2 we get that

$$P\left(\frac{n\hat{\sigma}^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{n\hat{\sigma}^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha$$

Therefore the $(1-\alpha)$-confidence interval for the variance is

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{n-1,\alpha/2}}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

2.1.4. *What is the Confidence Interval ?*

We have so far several way of constructing confidence intervals. Let us now discuss how we should interpret a confidence interval.

The confidence interval should be interpreted itself as a random object. It is a random inerval e.g., in the case of the mean, of the form

$$\left[\overline{X} - t_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}\right]. \tag{2}$$

but $\overline{X}$ and $\hat{s}$ are functions of the sample and so they should be considered are random variables.

The interpretation of an interval like the above one, should be a *realization* of a *random* interval, which with probability $(1-\alpha)$ contains the unkown parameter (in this case the mean).

As an example we do the following experiment. We generate 20 independent samples of size 11 each, from a normal distribution with mean $\mu = 10$ and variance $\sigma^2 = 9$. For each one of these samples we form the resulting 0.9-confidence intervals...

2.2. **Multivariate case.** Let us start with the following multivariate generalisation of Lemma 1

Proposition 2. *Let $\underline{X}_1, \ldots, \underline{X}_N$ an i.i.d collection of $\mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$. Then (i) the sample mean $\overline{\underline{X}} = \frac{1}{N}\sum_{i=1}^N \underline{X}_i$ has distribution $\mathcal{N}_p(\underline{\mu}, \frac{1}{N}\mathbf{\Sigma})$.*

(ii) The sample variance matrix

$$\begin{aligned}\mathbf{S} &:= \frac{1}{N-1}\sum_{i=1}^{N}(\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})^T \\ &:= \frac{1}{N-1}\mathbf{C}\end{aligned}$$

and $\mathbf{C}$ *has the Wishart distribution* $W_p(N-1, \mathbf{\Sigma})$ *defined below.*
(iii) $\mathbf{C}$ *and* $\overline{\underline{X}}$ *are independent.*

Proof. *The proof is very interesting but not trivial and we will not make use of it. So we'll just refer to Anderson, T.W.: "An Introduction to Multivariate Statistical Analysis . Let us just mention that it generalises the the derivation of the t distribution we presented when we dealt with the univariate case.*

The p-variate Wishart distribution, with m degrees of freedom with respect to a covariance matrix $\mathbf{\Sigma}$, is defined as the distribution of

$$\sum_{i=1}^{m}\underline{Z}_i\underline{Z}_i^T$$

where $\underline{Z}_i$, $i = 1, 2, ..., m$ are independent with distribution $\mathcal{N}_p(\underline{0}, \mathbf{\Sigma})$. Let us mention some properties of the Wishart distribution

(i) If $\mathbf{A_1} \sim^d W_p(n_1, \mathbf{\Sigma})$ and $\mathbf{A_2} \sim^d W_p(n_1, \mathbf{\Sigma})$ and they are independent, then $\mathbf{A_1} + \mathbf{A_1} \sim^d W_p(n_1 + n_2, \mathbf{\Sigma})$, ie. the degrees of freedom add up.

(ii) If $\mathbf{A} \sim^d W_p(n, \mathbf{\Sigma})$ and $\mathbf{B}$ is a matrix (of appropriate dimensions), then $\mathbf{BAB^T} \sim^d W_p(n, \mathbf{B\Sigma B}^t)$.

(iii) When $n > p$ then the density of the Wishart distribution, i.e. the probability to have a Wishart matrix equal to a matrix $\mathbf{A}$ is

$$\frac{\det(\mathbf{A})^{(n-p-2)/2)}e^{-\mathbf{Tr}(\mathbf{A\Sigma}^{-1}/2)}}{2^{p(n-1)/2}\pi^{p(p-1)/4}\det(\mathbf{\Sigma})^{(n-1)/2}\prod_{i=1}^{p}\Gamma(\frac{1}{2}(n-i))}.$$

REMARK: Let us recall that the estimators $\overline{X}$ and $\mathbf{S}$ are unbiased estimators for the mean and the variance of the distribution.

Given data $\underline{x}_1, \ldots, \underline{x}_N$ observed from a p-variate distribution we want to construct a region $\mathcal{R} \subset \mathbb{R}^p$, that is a function of the observed data, i.e. $\mathcal{R} = \mathcal{R}(\underline{x}_1, \ldots, \underline{x}_n)$, that will determine, i.e. contain with certain probability, the investigated parameter of the distribution, e.g. $\underline{\mu}$.

Definition 1. *Given a sample* $\underline{X}_1, \ldots, \underline{X}_N$, $\mathcal{R} = \mathcal{R}(\underline{X}_1, \ldots, \underline{X}_n)$, *is a* $100(1-\alpha)\%$ *confidence region for the population mean if*

$$\mathbb{P}\left(\underline{\mu} \in \mathcal{R}(\underline{X}_1, \ldots, \underline{X}_n)\right) = 1 - \alpha$$

2.2.1. *Confidence Interval for Mean with Known Covariance Matrix.*

In the case that the population distribution is p-variate normal with known covariance matrix $\mathbf{\Sigma}$, then we can follow Proposition 2 and the definition of the confidence region and deduce that the $100(1-\alpha)\%$ confidence region is

$$\mathcal{R}(\underline{X}_1, \ldots, \underline{X}_n) = \{\underline{x} \in \mathbb{R}^p \colon (\overline{\underline{X}} - \underline{x})^T \mathbf{\Sigma}^{-1} (\overline{\underline{X}} - \underline{x}) \leq \frac{1}{N} \chi_p^2(\alpha)\}. \tag{3}$$

One should also compare this with Proposition 1.

In the case that the population distribution is not multivariate normal, then Proposition 2 is not exactly correct, but it is *approximately* correct, as a consequence of the Central Limit Theorem. Then still (3) can define the $100(1-\alpha)\%$ confidence interval. The only difference now is that the probability of this region will not be exactly α, but approximately α, when the sample size is large enough.

2.2.2. *Confidence Interval for Mean with Unknown Covariance Matrix.* The procedure here is analogous to the univariate case. We would have to replace the unknown population covariance matrix with the sample covariance matrix $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})^T$. It is crucial in the study of this case that the population distribution is multivariate normal. To proceed we will need the analogue of the distribution of relation (1).

Proposition 3. *Let* $\underline{X}_1, \ldots, \underline{X}_n$ *i.i.d with distribution* $\mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$. *Then*

$$\frac{n(n-p)}{p(n-1)} (\overline{\underline{X}} - \underline{\mu})^T \mathbf{S}^{-1} (\overline{\underline{X}} - \underline{\mu}) \sim^d F_{p,n-p},$$

where $F_{p,n-p}$ *is the* F*-distribution with* p *and* $n-p$ *degrees of freedom. Therefore the* $100(1-\alpha)\%$ *confidence interval for the population mean is*

$$\mathcal{R}(\underline{X}_1, \ldots, \underline{X}_n) = \{\underline{x} \in \mathbb{R}^p \colon (\overline{\underline{X}} - \underline{x})^T \mathbf{S}^{-1} (\overline{\underline{X}} - \underline{x}) \leq \frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)\},$$

where $F_{p,n-p}(\alpha)$ *is the* $100(1-\alpha)\%$ *percentile of the* $F_{p,n-p}$ *distribution.*

Proof. We will not prove it but we will describe how this can be reduced to a calculus problem. Let's first write

$$\begin{aligned} T^2_{p,n-1} &= n(\overline{\underline{X}} - \underline{\mu})^T \mathbf{S}^{-1} (\overline{\underline{X}} - \underline{\mu}) \\ &= \sqrt{n}(\overline{\underline{X}} - \underline{\mu})^T \left(\frac{\sum_{i=1}^n (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})^T}{n-1} \right)^{-1} \sqrt{n}(\overline{\underline{X}} - \underline{\mu}) \\ &= \sqrt{n}(\overline{\underline{X}} - \underline{\mu})^T \left(\frac{\mathbf{C}}{n-1} \right)^{-1} \sqrt{n}(\overline{\underline{X}} - \underline{\mu}) \end{aligned}$$

(Notice that $T^2_{p,n-1}$ is a generalisation of the t-distribution in the case $p > 1$.) It only remains to use Proposition 2 that says that $\mathbf{C}$ and $\overline{\underline{X}}$ are independent, calculus and lots of work to complete the proof. □

Example. Consider a sample data for radiation from microwave ovens. The sample size is $n = 42$ and the two quantities measured are

$$x_1 = (\text{measured radiation with door closed})^{1/4}$$

$$x_2 = (\text{measured radiation with door open})^{1/4}.$$

The sample mean is

$$\overline{\underline{x}} = \begin{pmatrix} 0.564 \\ 0.603 \end{pmatrix}$$

and the sample covariance matrix is

$$\mathbf{S} = \begin{pmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{pmatrix}$$

and we compute the inverse of the sample covariance matrix

$$\mathbf{S}^{-1} = \begin{pmatrix} 203.018 & -163.391 \\ -163.391 & 200.228 \end{pmatrix}$$

Therefore the 95% confidence region for the population mean is

$$(0.564-x_1,\, 0.603-x_2) \begin{pmatrix} 203.018 & -163.391 \\ -163.391 & 200.228 \end{pmatrix} \begin{pmatrix} 0.564 - x_1 \\ 0.603 - x_2 \end{pmatrix} \leq \frac{2 \cdot 41}{42 \cdot 40} F_{2,40}(0.05).$$

From the tables we find that $F_{2,40}(0.05) = 3.23$. If you were asked to draw the confidence region (or describe it) you would say that it is an ellipsoid given by the above equation, but if you were asked to describe it in more detail, then you should say what is the center, the major and minor axis of it and their lengths. The center is obviously $(0.564, 0.603)$. The major and minor axis are given by the eigenvectors of the sample covariance $\mathbf{S}$, which are computed to be $\mathbf{e_1}^T = (0.704, 0.710)$ and $\mathbf{e_2}^T = (-.710, 0.704)$. To find the half lengths you need to know the eigenvalues,

which are computed to be $\lambda_1 = 0.026$ and $\lambda_2 = 0.002$. Then the half lengths of the axis of the ellipsoid are given by

$$\sqrt{\lambda_1}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)} = ... = 0.064$$

and

$$\sqrt{\lambda_2}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)} = ... = 0.018$$

2.2.3. *Simultaneous Confidence Intervals.* This is related to the problem of constructing confidence region that will *simultaneously* and with *the same probability* cover the means of the random variables $\underline{a}^T\underline{X}$, for every vector $\underline{a} \in \mathbb{R}^p$. We will assume that the covariance matrix is uknown and therefore we will be force to restrict ourselves to the case that the population distribution is multivariate normal. Then for any $\underline{a} \in \mathbb{R}^p$ the random variable $\underline{a}^T\underline{X}$ is normal $\mathcal{N}_1(\underline{a}^T\underline{\mu}, \underline{a}^T\mathbf{\Sigma}\underline{a})$. Suppose now that we have a random sample $\underline{X}_j$ from this distribution, which will produce a random sample $\underline{Z}_j = \underline{a}^T\underline{X}_j$ for the random variable $\underline{Z}$. We denote its sample mean by $\overline{z} = \underline{a}^T\overline{\underline{X}}$ and the sample covariance matrix $s_z^2 = \underline{a}^T\mathbf{\Sigma}\underline{a}$

The first attempt would be to consider the t-fraction

$$t_{\underline{a}} = \frac{\overline{Z} - \mu_z}{s_z/\sqrt{n}} = \frac{\sqrt{n}(\underline{a}^T\overline{\underline{X}} - \underline{a}^T\underline{\mu})}{\sqrt{\underline{a}^t\mathbf{S}\underline{a}}}$$

and follow the one-dimensional theory of confidence intervals which will give the following confidence interval with probability $(1-\alpha)$:

$$\overline{Z} - t_{n-1}(\alpha)\frac{s_z}{\sqrt{n}} \leq \mu_z \leq \overline{Z} + t_{n-1}(\alpha)\frac{s_z}{\sqrt{n}}.$$

The problem with this approach is that it will give a $(1-\alpha)$ confidence interval *only* for fixed $\underline{a}$ and *not* simultaneously for *all* $\underline{a}$. To achieve this we need to consider the optimised fraction

$$\begin{aligned} \max_{\underline{a}} t_{\underline{a}}^2 &= \max_{\underline{a}} \frac{n(\underline{a}^T\overline{\underline{X}} - \underline{a}^T\underline{\mu})^2}{\underline{a}^t\mathbf{S}\underline{a}} \\ &= n(\overline{\underline{X}} - \underline{\mu})^T\mathbf{S}^{-1}(\overline{\underline{X}} - \underline{\mu}). \end{aligned} \tag{4}$$

Check this as an exercise. We have already encountered this quantity and we know that its distribution, when multiplied by $n(n-p)/p(n-1)$ is $F_{p,n-p}$. There we can easily conclude that

Proposition 4. *Let $\underline{X}_1, ..., \underline{X}_n$ a random sample from a $\mathcal{N}_p(\underline{\mu}, \mathbf{S})$) population with* $\mathbf{S}$ *positive definite. Then simultaneously for all $\underline{a}$ the interval*

$$\left(\underline{a}^T \overline{\underline{X}} - \sqrt{\frac{p(n-1)}{n(n-p)} F_{n,n-p}(\alpha) \cdot \underline{a}^T \mathbf{S} \underline{a}}, \quad \underline{a}^T \overline{\underline{X}} + \sqrt{\frac{p(n-1)}{n(n-p)} F_{n,n-p}(\alpha) \cdot \underline{a}^T \mathbf{S} \underline{a}} \right),$$

will contain $\underline{a}^T \underline{\mu}$ with probability $1 - \alpha$.

Proof. Exercise. □

One nice thing about this proposition is that it allows us to find confidence regions for a number of other quantities of potential importance. For example, confidence intervals for the means of the $i-th$ one dimensional marginals follow from the above proposition by setting $\underline{a} = \mathbf{e}_i$, the $i - th$ unit vector of $\mathbb{R}^p$.

If we want a confidence interval for the difference $\mu_i - \mu_j$, then we just need to use the vector $\underline{a} = \mathbf{e}_i - \mathbf{e}_j$.

3. Hypothesis Testing

3.1. **Univariate Case.** Let us start with an example. Suppose that $X_1, X_2, \ldots, X_n$ is a sample drawn from a normal distribution with unkown mean μ and variance σ^2. Consider testing the following hypotheses:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0 \end{aligned}$$

The hypothesis H_0 is called **null hypothesis**, while the hypothesis H_A is called **alternative hypothesis**. The idea is that one starts *assuming* that the mean of the normal distribution is μ_0 and then proceeds in checking whether this assumption should be true and therefore be *accepted* or whether it should *rejected* in favor of the alternative hypothesis H_A, which claims that the mean $\mu \neq \mu_0$.

We would like to construct a **test**, based on which we will be rejecting or accepting the null hypothesis. Of course, since we deal with random events, there will always be a probability of false decision, that is, to accept the null hypothesis as correct, when it is not, or to accept the alternative hypothesis as correct, while it is not. The former type of error is called **Type II error**, while the latter error is called **Type I error**. We will come back to this point in a minute. First, we need to construct the test. Again, as in many occasions so far, there are several ways to construct an appropriate test. Here we will present the thest that is **dual** to confidence intervals.

We start with the assumption that the null hypothesis is correct, i.e. that the mean of the distribution is μ_0. Then, as before, the random variable

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

is standard normal. The random variable Z is called the **test statistic**, that we use. Suppose that the actual value of the random variable Z, as this emerges from the sampling is such that

$$\left| \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \tag{5}$$

where $z_{\alpha/2}$ is the $\alpha/2$ standard normal quantile. The porbability that something like this happens is

$$P(|Z| > z_{\alpha/2}) = \alpha.$$

Therefore, if α is sufficiently small, the probability of obtaining sample data that result to a test statistic satisfying (5) is very small (and equal to α). It is, therefore, unlikely that we got "strange data" and we prefer to say that our null hypothesis was wrong and reject it in favor of the alternative hypothesis. Of course there is always the possibility that we really got "strange data" and we falsely rejected the null hypothesis. In this case we fell into a type I error. The probability of this happening is α and it is called the **significance level** of our test.

We finally say that the region

$$\left\{x\colon \left| \frac{x - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}\right\}$$

is the **rejection region** for the test statistic (5) at significance level α. In other words we will reject the null hypothesis, if the data form a sample mean that falls into the rejection region.

The above type of hypothesis testing is called two-side. We could also have a one-sided hypothesis testing, which would consist of

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}$$

In this case it is easy to see that the rejection region at significance level α should be

$$\left\{x\colon \frac{x - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha}\right\}.$$

One proceeds similarly in the case where $>$ is replaced by $<$.

Since we were dealing with normal distributions the computations of the above probabilities were exact. In the case that we want to test the mean of a **general** distribution, we to make use of the Central Limit Theorem and then proceed similarly. The only thing that will change is that the equation $P(|Z| > z_{\alpha/2}) = \alpha$ will be replaced by $P(|Z| > z_{\alpha/2}) \simeq \alpha$.

As in the case of confidence intervals with unknown variance, when the variance of the distribution is unknown, we will have to replace it with the sample variance

$\hat{s}^2$. Then we also need to make use of the t-distribution, instead of the normal. In this case the test statistic that we will be using is

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}}.$$

The rejection region at significance level α (in the case of o two-sided hypothesis testing) will be

$$\left\{x\colon \left|\frac{x-\mu_0}{\sigma/\sqrt{n}}\right| > t_{n-1,\alpha/2}\right\}$$

where $t_{n-1,\alpha/2}$ is quantile for the t-distribution with $(n-1)$ degrees of freedom (if our sample size is n).

3.2. **Multivariate Case.** The multivariate case proceeds analogously to the univariate case. Let us go straight to the case the the population covariance is unkown and restrict ourselves to a multivariate normal distribution. So let's assume that we want to test if a vector $\underline{\mu}_0 \in \mathbb{R}^p$ is a plausible value for the mean of the multivariate normal.

By analogy to the univariate case we need to construct an appropriate test statistic. The analogous test statistic based on a random sample $\underline{X}_1, ...\underline{X}_n$ is

$$T^2 = n(\overline{\underline{X}} - \underline{\mu}_0)^T \mathbf{S}^{-1}(\overline{\underline{X}} - \underline{\mu}_0). \tag{6}$$

We have already encoutered this distribution, so T^2 is distributed according to $\frac{(n-1)p}{(n-p)}F_{p,n-p}$.

Suppose that we have the null hypothesis $H_0 : \underline{\mu} = \underline{\mu}_0$ versus the alternative hypothesis $H_1 : \underline{\mu} \neq \underline{\mu}_0$. Then at level of significance α we will reject the null hypothesis in favor of the alternative one if the observed sample fall into the region

$$\left\{\underline{x} \in \mathbb{R}^p\colon n(\overline{\underline{X}} - \underline{\mu}_0)^T \mathbf{S}^{-1}(\overline{\underline{X}} - \underline{\mu}_0) > \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\right\},$$

where $F_{p,n-p}(\alpha)$ is the $(100\alpha)th$ upper percentile of the $F_{p,n-p}$ distribution.

Example. Perspiration from 20 healthy females was analyzed. Three components X_1=sweat rate, X_2=sodium content, X_3=potassium content were measured and the results were recorded. We want to test the Hypothesis $H_0 : \underline{\mu}^T = (4, 50, 10)$ against the alternative hypothesis $H_1 :: \underline{\mu}^T \neq (4, 50, 10)$. To do this we first need to compute the sample mean and cavariance matrix

$$\overline{\underline{X}} = \begin{pmatrix} 4.640 \\ 45.400 \\ 9.965 \end{pmatrix}$$

and

$$\mathbf{S} = \begin{pmatrix} 2.879 & 10.010 & -1.810 \\ 10.010 & 199.788 & -5.640 \\ -1.810 & -5.640 & 3.628 \end{pmatrix}$$

Also we have

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.586 & -0.022 & 0.258 \\ 0.258 & 0.006 & -0.002 \\ 0.258 & -0.002 & 0.402 \end{pmatrix}.$$

Plugging the values into the T^2 statistic we get that $T^2 = 9.74$. While the critical value of the rejection region is

$$\frac{(n-1)p}{n-p} F_{p,n-p}(0.10) = 8.18.$$

Therefore, we will reject the null hypothesis at the 10% level of significance.

4. Maximum Likelihood and Hypothesis Tests.

The maximum likelihood estimation is a very powerful method to obtain estimators and confidence intervals for various parameters of distributions. We will start by considering the mean and the covariance of a multivariate normal distribution. So, let us consider $\underline{X}_1, ..., \underline{X}_n$ an i.i.d. sample of $\mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$. The joint density of the sample is

$$\begin{aligned} f_{\underline{X}_1,...,\underline{X}_n}(\underline{x}_1, ..., \underline{x}_n \,|\, \underline{\mu}, \mathbf{\Sigma}) &= \prod_{j=1}^{n} \frac{1}{(2\pi)^{p/2} \det(\mathbf{\Sigma})^{1/2}} e^{-\frac{1}{2}(\underline{x}_j - \underline{\mu})^T \mathbf{\Sigma}^{-1} (\underline{x}_j - \underline{\mu})} \\ &= \frac{1}{(2\pi)^{np/2} \det(\mathbf{\Sigma})^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n} (\underline{x}_j - \underline{\mu})^T \mathbf{\Sigma}^{-1} (\underline{x}_j - \underline{\mu})}. \end{aligned}$$

The Maximum Likelihood Estimation (MLE) is based on the idea that we are seeking the parameters $\underline{\mu}, \mathbf{\Sigma}$, which will maximise the likelihood of the observed sample $\underline{x}, ..., \underline{x}_n$. This is formalised as seeking a solution to the following optimisation problem

$$\begin{aligned} \max_{\underline{\mu}, \mathbf{\Sigma}} f_{\underline{X}_1,...,\underline{X}_n}(\underline{x}_1, ..., \underline{x}_n \,|\, \underline{\mu}, \mathbf{\Sigma}) &= \max_{\underline{\mu}, \mathbf{\Sigma}} \frac{1}{(2\pi)^{np/2} \det(\mathbf{\Sigma})^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n} (\underline{x}_j - \underline{\mu})^T \mathbf{\Sigma}^{-1} (\underline{x}_j - \underline{\mu})} \\ &= \max_{\underline{\mu}, \mathbf{\Sigma}} \frac{1}{(2\pi)^{np/2} \det(\mathbf{\Sigma})^{n/2}} e^{-\frac{1}{2}\mathrm{Tr}\left(\sum_{j=1}^{n} (\underline{x}_j - \underline{\mu})^T \mathbf{\Sigma}^{-1} (\underline{x}_j - \underline{\mu})\right)} \\ &= \max_{\underline{\mu}, \mathbf{\Sigma}} \frac{1}{(2\pi)^{np/2} \det(\mathbf{\Sigma})^{n/2}} e^{-\frac{1}{2}\mathrm{Tr}\left(\sum_{j=1}^{n} \mathbf{\Sigma}^{-1} (\underline{x}_j - \underline{\mu})(\underline{x}_j - \underline{\mu})^T\right)} \\ &= \max_{\underline{\mu}, \mathbf{\Sigma}} \frac{1}{(2\pi)^{np/2} \det(\mathbf{\Sigma})^{n/2}} e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1} \sum_{j=1}^{n} (\underline{x}_j - \underline{\mu})(\underline{x}_j - \underline{\mu})^T\right)}, \end{aligned}$$

where in the second equality we used the fact that the trace of a number equals the number, while in the second that $\mathrm{Tr}(\underline{x}^T\mathbf{A}\underline{x}) = \mathrm{Tr}(\mathbf{A}\underline{x}\underline{x}^T)$, for a vector $\underline{x}$ and a matrix $\mathbf{A}$. A simple algebra by adding and subtracting the sample mean, shows that

$$\sum_{j=1}^{n}(\underline{x}_j - \underline{\mu})(\underline{x}_j - \underline{\mu})^T = \sum_{j=1}^{n}(\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T + n(\overline{\underline{x}} - \underline{\mu})^T(\overline{\underline{x}} - \underline{\mu})$$

and therefore the above optimisation problem writes us

$$\max_{\underline{\mu},\mathbf{\Sigma}} \frac{1}{(2\pi)^{np/2}\det(\mathbf{\Sigma})^{n/2}} e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\sum_{j=1}^{n}(\underline{x}_j-\overline{\underline{x}})(\underline{x}_j-\overline{\underline{x}})^T\right)-\frac{n}{2}n(\overline{\underline{x}}-\underline{\mu})^T(\overline{\underline{x}}-\underline{\mu})},$$

Since the covariance matrix is positive definite we have the

$$(\overline{\underline{x}} - \underline{\mu})^T(\overline{\underline{x}} - \underline{\mu}) \geq 0,$$

for any $\overline{\underline{x}}$ and $\underline{\mu}$. Therefore this term is minimised (hence maximised when multiplied by a minus) for

$$\underline{\mu} = \overline{\underline{x}}. \tag{7}$$

We need to see when the term

$$\max_{\mathbf{\Sigma}} \frac{1}{(2\pi)^{np/2}\det(\mathbf{\Sigma})^{n/2}} e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\sum_{j=1}^{n}(\underline{x}_j-\overline{\underline{x}})(\underline{x}_j-\overline{\underline{x}})^T\right)}$$

is achieved. This is taken care by the following Lemma

Lemma 3. *Given a $p \times p$ symmetric, positive definite matrix $\mathbf{B}$ and $b > 0$ we have that*

$$\frac{1}{\det(\mathbf{\Sigma})^b} e^{-\frac{1}{2}Tr(\mathbf{\Sigma}^{-1}B)} \leq \frac{1}{\det(\mathbf{B})^b}(2b)^{pb}e^{-bp},$$

with the equality when $\mathbf{\Sigma} = \frac{1}{2b}\mathbf{B}$.

The proof of this lemma is interesting, but we will skip it and refer to [JW], page 170. We just remark that it reduces to an eigenvalue optimisation problem.

Applying this lemma we have that the desired optimization problem is solved for

$$\mathbf{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T \tag{8}$$

Equation (7), (8) provide the Maximum Likelihood Estimators for the mean and the covariance matrix.

Remark: The MLE is a very stable estimator. If $\hat{\theta}_{MLE}$ is the MLE estimator of the parameter θ and $h(\cdot)$ a function, then the MLE of $h\theta$ is $h(\hat{\theta}_{MLE})$. So, for example, the MLE estimator of the variance of the i^{th} coordinate of the multivariate normal is readily given by $n^{-1}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2$.

4.1. **MLE and Hypothesis Testing.** Let us now see how we can use the MLE into Hypothesis Testing (still working with multivariate normal). Suppose that we have to test the null hypothesis $H_0 : \underline{\mu} = \underline{\mu}_0$. Assuming H_0 we have that the MLE functional is given by

$$f_{\underline{X}_1,...,\underline{X}_n}(\underline{x}_1, ..., \underline{x}_n \,|\, \underline{\mu}_0, \mathbf{\Sigma}) = \prod_{j=1}^{n} \frac{1}{(2\pi)^{p/2} \det(\mathbf{\Sigma})^{1/2}} e^{-\frac{1}{2}(\underline{x}_j - \underline{\mu}_0)^T \mathbf{\Sigma}^{-1} (\underline{x}_j - \underline{\mu}_0)}$$

and we will therefore need to solve the optimisation problem

$$\max_{\mathbf{\Sigma}} f_{\underline{X}_1,...,\underline{X}_n}(\underline{x}_1, ..., \underline{x}_n \,|\, \underline{\mu}_0, \mathbf{\Sigma}).$$

Following the same steps as before we have the maximum is achieved for

$$\hat{\mathbf{\Sigma}}_0 = \frac{1}{n} \sum_{j=1}^{n} (\underline{x}_j - \underline{\mu}_0)(\underline{x}_j - \underline{\mu}_0)^T$$

To determine whether H_0 is a good assumption we look at the ratio

$$\Lambda = \frac{\max_{\mathbf{\Sigma}} f_{\underline{X}_1,...,\underline{X}_n}(\underline{x}_1, ..., \underline{x}_n \,|\, \underline{\mu}_0, \mathbf{\Sigma})}{\max_{\underline{\mu}, \mathbf{\Sigma}} f_{\underline{X}_1,...,\underline{X}_n}(\underline{x}_1, ..., \underline{x}_n \,|\, \underline{\mu}, \mathbf{\Sigma})}$$

which turns out to be equal to

$$\Lambda = \left(\frac{\det(\hat{\mathbf{\Sigma}}_{MLE})}{\det(\hat{\mathbf{\Sigma}}_0)} \right)^{n/2}.$$

The statistic $\Lambda^{2/n}$ is called the Wilks lambda. If the observed value of Λ is too small, then we will reject the null hypothesis in favor of the alternative. We can write the rejection region explicitly as

$$\left\{ \underline{x} \in \mathbb{R}^p : \frac{\det\left(\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T \right)}{\det\left(\sum_{j=1}^{n} (\underline{x}_j - \underline{\mu}_0)(\underline{x}_j - \underline{\mu}_0)^T \right)} < c_\alpha^{2/n} \right\}, \tag{9}$$

where c_α is the 100α percentile of the distribution of Λ. Fortunately, the Wilks Lambda is related directly to the T^2 statistic

Proposition 5. *Let $\underline{X}_1, ..., \underline{X}_n$ be a random sample from $\mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$. Then*

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1} \right)^{-1},$$

where T^2 is given by

$$T^2 = n(\overline{\underline{X}} - \underline{\mu}_0)^T \mathbf{S}^{-1} (\overline{\underline{X}} - \underline{\mu}_0). \tag{10}$$

Proof. The proof is based on the very interesting fact from linear algebra that if a matrix $\mathbf{A}$ can be partitioned as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{pmatrix},$$

where $\mathbf{A_{ij}}, \mathbf{i}, \mathbf{i} = \mathbf{1}, \mathbf{2}$ are matrices of appropriate dimensions, then

$$\begin{aligned} \det(\mathbf{A}) &= \det(\mathbf{A_{11}}) \det(\mathbf{A_{22}} - \mathbf{A_{21}}\mathbf{A_{11}^{-1}}\mathbf{A_{12}}) \\ &= \det(\mathbf{A_{22}}) \det(\mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{A_{21}}). \end{aligned}$$

We will now apply this to the matrix

$$\mathbf{A} = \begin{pmatrix} \sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T & \sqrt{n}(\overline{\underline{x}} - \underline{\mu}_0) \\ \sqrt{n}(\overline{\underline{x}} - \underline{\mu}_0)^T & -1 \end{pmatrix}$$

Then, on the one hand we have that

$$\begin{aligned} \det(\mathbf{A}) &= \det\left(\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T\right) \\ &\quad \times \left(-1 - n(\overline{\underline{x}} - \underline{\mu}_0)^T \left(\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T\right)^{-1} (\overline{\underline{x}} - \underline{\mu}_0)\right) \\ &= (-1) \det\left(\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T\right) \\ &\quad \times \left(1 + \frac{n}{n-1}(\overline{\underline{x}} - \underline{\mu}_0)^T \left(\frac{\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T}{n-1}\right)^{-1} (\overline{\underline{x}} - \underline{\mu}_0)\right) \\ &= (-1) \det\left(\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T\right) \left(1 + \frac{n(\overline{\underline{x}} - \underline{\mu}_0)^T \mathbf{S}^{-1} (\overline{\underline{x}} - \underline{\mu}_0)}{n-1}\right) \\ &= (-1) \det\left(\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T\right) \left(1 + \frac{T^2}{n-1}\right). \end{aligned}$$

On the other hand, using the second form of the $\det(\mathbf{A})$ we get that Then, on the one hand we have that

$$\begin{aligned} \det(\mathbf{A}) &= (-1) \det\left(\sum_{j=1}^{n} (\underline{x}_j - \overline{\underline{x}})(\underline{x}_j - \overline{\underline{x}})^T + n(\overline{\underline{x}} - \underline{\mu}_0)(\overline{\underline{x}} - \underline{\mu}_0)^T\right) \\ &= (-1) \det\left(\sum_{j=1}^{n} (\underline{x}_j - \underline{\mu}_0)(\underline{x}_j - \underline{\mu}_0)^T\right). \end{aligned}$$

Equating the two expressions of $\det(\mathbf{A})$ we have that

$$\frac{\det\left(\sum_{j=1}^{n}(\underline{x}_j-\overline{\underline{x}})(\underline{x}_j-\overline{\underline{x}})^T\right)}{\det\left(\sum_{j=1}^{n}(\underline{x}_j-\underline{\mu}_0)(\underline{x}_j-\underline{\mu}_0)^T\right)}=\left(1+\frac{T^2}{n-1}\right)^{-1},$$

which is the desired equality. □

Remark: Having the above proposition we can rewrite the rejection region in (9) in terms of the T^2 statistic as

$$\left\{\underline{x}\in\mathbb{R}^p\colon T^2>(n-1)(c_\alpha^{-2/n}-1).\right\}$$

This follows from just solving the inequality in (9) in terms of T^2 using the above proposition. Therefore it follows that the parameter c_α should be chosen such that

$$(n-1)(c_\alpha^{-2/n}-1)=F_{p,n-p}(\alpha),$$

with $F_{p,n-p}(\alpha)$ the upper 100α percentile of the F distribution.

4.2. **Generalised Likelihood Ratio Method.** We would now like to address the question of how to do Hypothesis Testing for general distributions (not necessarily normal) and their corresponding parameters. So let us suppose that we have a distribution whose parameters are captured by the vector $\underline{\theta}\in\mathbf{\Theta}\subset\mathbb{R}^\nu$. and suppose that we have a hypothesis testing $H_0\colon\underline{\theta}=\underline{\theta}_0$, which in effect restrict the parameter $\underline{\theta}$ to a subspace $\Theta_0\subset\mathbb{R}^{\nu_0}$ with $\nu_0<\nu$. Then we wrill reject the null hypothesis (at some significance level α) if the likelihood ration is too small ie

$$\Lambda=\frac{\max_{\underline{\theta}\in\mathbf{\Theta_0}} f_{\underline{X}_1,\ldots,\underline{X}_n}(\underline{x}_1,\ldots,\underline{x}_n\,|\,\underline{\theta})}{\max_{\underline{\theta}\in\mathbf{\Theta}} f_{\underline{X}_1,\ldots,\underline{X}_n}(\underline{x}_1,\ldots,\underline{x}_n\,|\,\underline{\theta})}<c_\alpha.$$

Of course, to quantify things we will need to determine the distribution of Λ and this can be difficult. Luckily, there is a very powerful result that says that, when the sample size n is large, the Λ approximates a chi-square distribution.

Proposition 6.

$$-2\log\Lambda=-2\log\frac{\max_{\underline{\theta}\in\mathbf{\Theta_0}} f_{\underline{X}_1,\ldots,\underline{X}_n}(\underline{X}_1,\ldots,\underline{X}_n\,|\,\underline{\theta})}{\max_{\underline{\theta}\in\mathbf{\Theta}} f_{\underline{X}_1,\ldots,\underline{X}_n}(\underline{X}_1,\ldots,\underline{X}_n\,|\,\underline{\theta})},$$

is approximately, for large n, a $\chi^2_{\nu-\nu_0}$ variable, where ν =dimension of $\mathbf{\Theta}$ and ν_0 =dimension of $\mathbf{\Theta_0}$.

5. Comparisons of Several Multivariate Means.

We will be working under assumptions of normality. If this fails then it will have to be replaced by a large sample size.

Suppose we want to test the comparative effect of two treatments, say Treatement 1 and Treatment 2. The treatment can be thought as an application of a drug or the effect of an advertising campaign (in the latter case we should think of Treatement 1 as effect on the population of the advertising and Treatment 2 as the effect that a no advertising would have). We do this by applying the two treatments to the same sample (so to reduce other exogenous factors) and then measuring the responses.

Let us start with the case that we only measure on parameter. We denote X_{j1} the response of the j^{th} member of the sample to Treatment 1 and X_{j2} the response of the j^{th} member of the sample to Treatment 2. What interests us is the difference $D = X_1 - X_2$, or in terms of sample $D_j = X_{j1} - X_{j2}$. This quantity should measure the differential effect of the two treatments, but on top of it there will be some noise sitting. Let us assume that this noise is modelled as a normal $\mathcal{N}(\delta, \sigma_\delta^2)$ variable. Our goal is to estimate the mean δ. The relevant test-statistic should

$$t = \frac{\overline{D} - \delta}{s_d/\sqrt{n}}$$

where $\overline{D}$ is the sample mean of the sample $D_1, ..., D_n$ and s_d the sample standard deviation. Then, as we have already seen, the interval

$$\left[\overline{D} - t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}}, \overline{D} - t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}}\right]$$

would be a $100(1-\alpha)$ confidence interval for δ. In the same way we can thest the hypothesis

$$\begin{aligned} H_0 &: \quad \delta = 0 \\ H_A &: \quad \delta \neq 0, \end{aligned}$$

which, at significance level α will have a rejection region

$$\left\{\left|\frac{\overline{D}}{s_d/\sqrt{n}}\right| > t_{n-1}(\alpha/2)\right\}.$$

Let us now pass to the multivariate case. That is, we are interested in measuring $p \geq 1$ parameters. Let us denote X_{jk1} the response of the j^{th} member of the sample to the measurement of the k^{th} parameter, with $1 \leq k \leq p$ under Treatment 1 and X_{jk2} the response of the j^{th} member of the sample to the measurement of the k^{th} parameter, with $1 \leq k \leq p$ under Treatment 2. We record the difference in the vectors $\underline{D} = (D_{j1}, ..., D_{jp})^T$ with $D_{jk} = X_{jk1} - X_{jk2}$ for $k = 1, ..., p$. Our assumption again is that

the components of $\underline{D}$ are i.i.d. multivariate normals $\mathcal{N}_p(\underline{\delta}, \mathbf{\Sigma}_d)$. Inferences about $\underline{\delta}$ can be made from the T^2 statistics

$$T^2 = n(\overline{\underline{D}} - \underline{\delta})^T \mathbf{S}_d^{-1} (\overline{\underline{D}} - \underline{\delta}),$$

where $\overline{\underline{D}}$ is the sample mean and $\mathbf{S}_d$ the sample covariance. The null hypothesis $H_0 : \underline{\delta} = 0$ is rejected in favor of the alternative hypothesis $H_A : \underline{\delta} \neq 0$ at significance level α if the sample mean $\overline{\underline{D}}$ falls into the region

$$\left\{ \underline{d}^T \in \mathbb{R}^p \colon n\, \underline{d}\, \mathbf{S}_d^{-1}\, \underline{d} > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \right\},$$

As an exercise write the confidence intervals for $\underline{\delta}$.

5.1. **Comparison of means of different population.** Suppose we apply a treatment to two different populations and we want to measure the relative effect, ie the difference of the mean responses. We denote by $\underline{X}_{11}, ..., \underline{X}_{1n_1}$ the sample from the one population of size n_1 and $\underline{X}_{21}, ..., \underline{X}_{2n_2}$ the sample from the second population of size n_2. We assume that the population distribution is multivariate normal. We naturally assume that the two samples are independent from each other.

5.1.1. *CASE 1: Equal Covariances.* The first case we treat makes the crucial assumption that the covaiances of the two populations are equal ie $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. The population means $\underline{\mu}_1, \underline{\mu}_2$ of the two populations are naturally estimated by the saple means $\overline{\underline{X}}_1, \overline{\underline{X}}_2$. Therefore the estimator for the difference $\underline{\mu}_1, \underline{\mu}_2$ is $\overline{\underline{X}}_1 - \overline{\underline{X}}_2$. Let's now look at the covariance

$$\mathrm{Cov}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2) = \mathrm{Cov}\overline{\underline{X}}_1 + \mathrm{Cov}(\overline{\underline{X}}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{\Sigma},$$

where $\mathbf{\Sigma}$ is the common covariance. We now come to the need of estimating $\mathbf{\Sigma}$. To do this we should utilise both sample. Let

$$\mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n} \left(\underline{X}_{ij} - \overline{\underline{X}}_{ij} \right) \left(\underline{X}_{ij} - \overline{\underline{X}}_{ij} \right)^T, \qquad i = 1, 2$$

the sample covariances of the two populations. A natural guess for the estimator of the covariance is the linear combination of the two sample coariances

$$\mathbf{S}_{pool} := \frac{n_1 - 1}{n_1 + n_2 - 2} \mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} \mathbf{S}_2.$$

As the sample size get larger the pooled covariance converges to the true, common population covariance.

We now want to test the hypothesis $H_0 \colon \underline{\mu}_1 - \underline{\mu}_2 = \underline{\delta}$. To proceed we will need to device a statistic whose distribution we can also compute. This turns out to be the

T^2 statistic

$$T^2 = (\overline{X}_1 - \overline{X}_2 - \underline{\mu}_1 + \underline{\mu}_2)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pool} \right]^{-1} (\overline{X}_1 - \overline{X}_2 - \underline{\mu}_1 + \underline{\mu}_2).$$

We have essentially seen how to compute the distribution of T^2. We just need to notice that

$$\overline{X}_1 - \overline{X}_2 - \underline{\mu}_1 + \underline{\mu}_2$$

is a multivariate normal with mean zero, while $\mathbf{S}_{pool}$ is a Wishart distribution with $n_1 + n_2 - 2$ degrees of freedom. Therefore, T^2 has a $\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$ distribution. We immediately deduce then that the rejection region for H_0 is when the sample data result to a statistic such that

$$(\overline{X}_1 - \overline{X}_2 - \underline{x})^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pool} \right]^{-1} (\overline{X}_1 - \overline{X}_2 - \underline{x}) > \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p,n_1+n_2-p-1}(\alpha)$$

6. Factor Analysis.

The purpose of Factor Analysis is to describe the covariance relationships among many variables in terms of a few underlying (BUT unobservables) random quantiles called *factors*.

IDEA: Suppose that variables can be grouped together by their correlations. This means that we can split the variables (of our functional) to groups, such that the variables within a particular group are highly correlated, but variables within different groups are essentially uncorrelated. Then each group of variable represents a single factor that is responsible fro the observed correlations.

Let's discussed the following example (taken from Wikipedia !). The following example is a fictionalized simplification for expository purposes, and should not be taken as being realistic. Suppose a psychologist proposes a theory that there are two kinds of intelligence, "verbal intelligence" and "mathematical intelligence", neither of which is directly observed. Evidence for the theory is sought in the examination scores from each of 10 different academic fields of 1000 students. If each student is chosen randomly from a large population, then each student's 10 scores are random variables. The psychologist's theory may say that for each of the 10 academic fields, the score averaged over the group of all students who share some common pair of values for verbal and mathematical "intelligences" is some constant times their level of verbal intelligence plus another constant times their level of mathematical intelligence, i.e., it is a linear combination of those two "factors". The numbers for a particular subject, by which the two kinds of intelligence are multiplied to obtain the expected score, are posited by the theory to be the same for all intelligence level pairs, and are called "factor loadings" for this subject. For example, the theory may

hold that the average student's aptitude in the field of amphibiology is

{10×the student's verbal intelligence}+{6×the student's mathematical intelligence}.

The numbers 10 and 6 are the factor loadings associated with amphibiology. Other academic subjects may have different factor loadings. Two students having identical degrees of verbal intelligence and identical degrees of mathematical intelligence may have different aptitudes in amphibiology because individual aptitudes differ from average aptitudes. That difference is called the "error" a statistical term that means the amount by which an individual differs from what is average for his or her levels of intelligence (see errors and residuals in statistics). The observable data that go into factor analysis would be 10 scores of each of the 1000 students, a total of $10,000$ numbers. The factor loadings and levels of the two kinds of intelligence of each student must be inferred from the data.

Let us formalise the above. Assume that we have the observable random vector $\underline{X} \in \mathbb{R}^p$, with mean $\underline{\mu}$ and covariance matrix $\mathbf{\Sigma}$. The factor model postulates that $\underline{X}$ is linearly dependent upon a *few unobservable* variables $F_1, ..., F_m$ and p additional sources of randomness $\varepsilon_1, ..., \varepsilon_p$, that is

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \cdots \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{11}F_2 + \cdots \ell_{2m}F_m + \varepsilon_2 \\ &\cdot \\ &\cdot \\ &\cdot \\ X_p - \mu_p &= \ell_{p1}F_1 + \cdots \ell_{pm}F_m + \varepsilon_p, \end{aligned}$$

which in matrix notation writes as

$$\underline{X} - \underline{\mu} = \mathbf{L}\underline{F} + \underline{\varepsilon}.$$

The coefficients ℓ_{ij} are called the loadings of the i^{th} variable and the matrix $\mathbf{L}$ is the matrix of loadings.

Remark. Let us remark that the deviations $X_i - \mu_i$ are expressed in terms of $p + m$ random variables $F_1, ..., F_m, \varepsilon_1, ..., \varepsilon_p$, which are unobservables. This also constitutes the difference with the linear regression, where the variables $F_1, ..., F_m$ are observed.

ASSUMPTIONS: Let us make the following assumptions:

$$\begin{aligned} E[\underline{F}] &= 0 \\ \text{Cov}(F) &= E[\underline{F}\underline{F}^T] = \mathbf{I} \\ E[\underline{\varepsilon}] &= 0 \\ \text{Cov}(\underline{\varepsilon}) &= E[\underline{\varepsilon}\underline{\varepsilon}^T] = \mathbf{\Psi} \\ &= diag(\psi_1, ..., \psi_p) \end{aligned}$$

We also assume that $\underline{\varepsilon}$ and $\underline{F}$ are independent and therefore $\text{Cov}(\underline{\varepsilon}, \underline{F}) = 0$. These assumptions constitute the *Orthogonal Factor Model*.

We can now compute the covariance matrix, $\mathbf{\Sigma}$ of $\underline{X}$. This is a simple calculation and leads to

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}.$$

We can spell out this matrix relation as

$$\begin{aligned} Var(X_i) &= \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i \\ Cov(X_i, X_k) &= \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}. \end{aligned}$$

Finally we can also easily compute

$$Cov(\underline{X}, \underline{F}) = \mathbf{L},$$

and which is also spelled out as

$$Cov(X_i, F_j) = \ell_{ij}.$$

The above covariance relations show that the covariance of $\underline{X}$ (which involves $p(p+1)/2$) can be recovered from the pm variables ℓ_{ij} and the p variables ε_i. The method is, therefore, useful if we can do so in the case that $m << p$.

Remark. It is not always (in fact it usually not) possible to write the covariance matrix in the Orthogonal Factor Model form. For example consider the covariance matrix

$$\begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix}$$

You can check (by trying to solve the equations) that this covariance cannot be written in the form

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \varepsilon_2 \\ X_3 - \mu_3 &= \ell_{31}F_1 + \varepsilon_3 \end{aligned}$$

6.1. **Principal Component Method.** This method uses the spectral decomposition of symmetric matrices. We know that the covariance matrix $\mathbf{\Sigma}$ is symmetric and positive definite. Therefore it can be diagonalised is the form $\mathbf{\Sigma} = \mathbf{U}\, diag(\lambda_1, ..., \lambda_n)\mathbf{U}^T$, where $\mathbf{U}$ is the matrix consisting of eigenvectors. We can rewrite this diagonalization in the form

$$\mathbf{\Sigma} = \lambda_1 \underline{e}_1 \underline{e}_1^T + \cdots + \lambda_p \underline{e}_p \underline{e}_p^T = (\sqrt{\lambda_1}\underline{e}_1, ..., \sqrt{\lambda_p}\underline{e}_p) \begin{pmatrix} \sqrt{\lambda_1}\underline{e}_1 \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{\lambda_p}\underline{e}_p \end{pmatrix} \tag{11}$$

and therefore $\mathbf{\Sigma}$ can be written in the form $\mathbf{\Sigma} = \mathbf{LL^T}$. This representation is not efficient since the dimension of the $\mathbf{L}$ matrix is $p \times p$, but we would like it to have dimensions $m \times m$ with $m << p$. One way out would be if, for some m, the last $p-m$ eigenvalues are negligible and so we could approximately represent the covariance matrix as

$$\mathbf{\Sigma} \simeq \lambda_1 \underline{e}_1 \underline{e}_1^T + \cdots + \lambda_m \underline{e}_m \underline{e}_m^T = (\sqrt{\lambda_1}\underline{e}_1, ..., \sqrt{\lambda_m}\underline{e}_m) \begin{pmatrix} \sqrt{\lambda_1}\underline{e}_1 \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{\lambda_m}\underline{e}_m \end{pmatrix} \tag{12}$$

In the derivation of (11) and (12) we haven't included any specific factors $\underline{\varepsilon}$. If we do this, then their variance may be taken to be the diagonal elements of $\mathbf{\Sigma} - \mathbf{LL^T}$. Therefore we arrive at

$$\mathbf{\Sigma} \simeq \mathbf{LL^T} + \mathbf{\Psi} := (\sqrt{\lambda_1}\underline{e}_1, ..., \sqrt{\lambda_m}\underline{e}_m) \begin{pmatrix} \sqrt{\lambda_1}\underline{e}_1 \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{\lambda_m}\underline{e}_m \end{pmatrix} + diag(\psi_1, ..., \psi_p)$$

where $\psi_i = \sigma_{ii} - \sum_{j=1}^m \ell_{ij}^2$, for $i = 1, 2, ..., p$, where ℓ_{ij} are the entries of the matrix $\mathbf{L}$.

Proposition 7. *The principal component factor analysis of the sample covariance matrix* $\mathbf{S}$ *is specified in terms of its eigenvalues and eigenvectors* $(\hat{\lambda}_i, \underline{\hat{e}}_i)_{i=1,...,p}$. *Let* $m < p$ *be the number of common factors. Then the matrix estimated factor loadings* $\{\tilde{\ell}_{ij}\}$ *is given by*

$$\tilde{\mathbf{L}} = \left(\sqrt{\hat{\lambda}_1}\underline{\hat{e}}_1, ..., \sqrt{\hat{\lambda}_m}\underline{\hat{e}}_m \right).$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \mathbf{LL^T}$, *and they are* $\tilde{\mathbf{\Psi}} := diag(\tilde{\psi}_1, ..., \tilde{\psi}_p)$, *with* $\tilde{\psi}_i = s_{ij} - \sum_{j=1}^m \tilde{\ell}_{ij}^2$.

Often it is better to work with centered and scaled variables, so that to avoid the case where we have a fictitious influence on the loadings due to observation with large variances or means. This means, that once the data $\underline{x}_1, ..., \underline{x}_n$ are collected we should transform them to $\underline{z}_1, ..., \underline{z}_n$, with $z_{ji} := (x_{ji} - \overline{x}_i)/\sqrt{s_{ii}}$, $j = 1, ..., n$ and $i = 1, ..., p$. The sample covariance matrix for these centered and rescaled data is the sample correlation matrix $\mathbf{R}$. Proposition 7 could now be stated for the sample correlation matrix, by simply replacing $\mathbf{R}$ with $\mathbf{S}$.

HOW TO DECIDE HOW MANY FACTORS TO CHOOSE ? To answer this question we consider the residual matrix $\mathbf{S} - (\mathbf{L}\mathbf{L}^{\mathbf{T}} + \mathbf{\Psi})$. The diagonal elements of this matrix are zero (why ?). It is not difficult to see, using the spectral decomposition, that

$$\sum_{ij} (\mathbf{S} - (\mathbf{L}\mathbf{L}^{\mathbf{T}} + \mathbf{\Psi}))_{ij}^2 \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2. \tag{13}$$

We should therefore use m such that the remainder sum of the squares of the eigenvalues is small. In some computer packages m is set to equal the number of positive eigenvalues of $\mathbf{S}$, if the sample covariance is factored, or the number of eigenvalues of $\mathbf{R}$, which larger than one, if the sample correlation is factored.

6.2. **Maximum Likelihood Solution.** We are now going to demonstrate how to produce a solution to the Factor problem via maximum likelihood, assuming that the common factors $\underline{F}$ and the specific factors $\underline{\varepsilon}$ are jointly normal . The maximum likelihood method is based on the observations $\underline{X}_j, j = 1, 2..., n$. The normality assumption implies that also the $\underline{X}_j$ which equals $\underline{\mu} + \mathbf{L}\underline{F} + \underline{\varepsilon}$ is normal, with mean $\underline{\mu}$ and an unkown covariance matrix, denote it by $\mathbf{\Sigma}$. The likelihood function, which we have already computed earlier is

$$f_{\underline{X}_1,...,\underline{X}_n}(\underline{x}_1, ..., \underline{x}_n \,|\underline{\mu}, \mathbf{L}, \mathbf{\Psi}) = \prod_{j=1}^{n} \frac{1}{(2\pi)^{p/2} \det(\mathbf{\Sigma})^{1/2}} e^{-\frac{1}{2}(\underline{x}_j - \underline{\mu})^T \mathbf{\Sigma}^{-1} (\underline{x}_j - \underline{\mu})} \tag{14}$$

Notice that we did not use the parameters $\mathbf{\Sigma}$ in the likelihood function, as our intention is to compute $\mathbf{L}, \mathbf{\Psi}$, and the right hand side of the likelihood function depends on these variables via the relation $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathbf{T}} + \mathbf{\Psi}$. Recall that the Factor Model does not have a unique solution. Any matrix of the form $\mathbf{LT}$ where $\mathbf{T}$ is orthogonal matrix is a solution if $\mathbf{L}$ is. To solve the Factor Problem via the Maximum Likelihood Method we will impose the workable assumption that

$$\mathbf{L}^{\mathbf{T}} \mathbf{\Psi}^{-\mathbf{1}} \mathbf{L} = \mathbf{\Delta},$$

where $\mathbf{\Delta}$ is a diagonal matrix.

Remark: There is no analytic solution of the maximum likelihood problem and therefore the solution is found numerically. A computational procedure to obtain the maximum likelihood procedure is described in Johnson-Wichern, pg 527.

We will now describe the solution to the factor problem, assuming a numerical solution of the maximum likelihood equations are obtained

Proposition 8. *Let $\underline{X}_1, ..., \underline{X}_n$ a random sample from $\mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathbf{T}} + \mathbf{\Psi}$ is the covariance matrix for the m common factor model. The maximum likelihood estimators $\hat{\mathbf{L}}, \hat{\mathbf{\Psi}}$ and $\underline{\hat{\mu}} = \overline{\underline{x}}$ maximise (14) under the constrain $\mathbf{L}^{\mathbf{T}}\mathbf{\Psi}^{-\mathbf{1}}\mathbf{L} = \mathbf{\Delta}$.*

The proportion of the total sample variance due to the j^{th} factor is

$$\frac{\hat{\ell}_{1j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + \cdots + s_{pp}}.$$

7. Discrimination and Classification.

Discrimination and calssification are multivariate techniques concerned with separating distinct sets of objects and allocating new objects to already existing groups. To start with, let's consider two classes π_1, π_2. We want to classify *new* objects as belonging to either π_1 or π_2. This will be done based on measurements on p random variables $\underline{X}^T = (X_1, ..., X_p)$. These populations will be characterised by different pdfs $f_1(\underline{x})$ and $f_2(\underline{x})$ for the measurable vector $\underline{X}$. These pdfs will be for example obtained after long observations, taking large samples. These samples are called learning samples. Essentially, we measure characteristics, ie the measurable vector $\underline{X}$, when the samples come from known populations. Then we determine two regions (nonintersecting), R_1, R_2. Now, let's say that a new observation arises and we want to classify as belonging to either π_1 or π_2. If the measurement falls into the R_1 region, then we will assign the new item to population π_1 and otherwise to π_2.

We are now interested in the probability of making a mistake and misclassifying the new observation. This would amount to computing the probability that the observation falls into, say, region R_2, while it actually comes from π_1. Let's denote this probability by $P(2|1)$. We then have

$$P(2|1) = P(\underline{X} \in R_2|\pi_1) = \int_{R_2} f_1(\underline{x})d\underline{x},$$

and analogously

$$P(1|2) = P(\underline{X} \in R_1|\pi_2) = \int_{R_1} f_2(\underline{x})d\underline{x}.$$

Let p_1 be the prior probability of π_1 and p_2 the prior probability of π_2. Then the overall probabilities of correctly or incorrectly clasifying objects are

$$\begin{aligned} P(\text{observation is correctly classified as}\,\pi_1) &= P(\underline{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1 \\ P(\text{observation misclassified as}\ \pi_1) &= P(\underline{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2, \end{aligned}$$

and analogously we get the probabilities of correctly or incorrectly classifying an observation as π_2.

Suppose, now, that we have costs associated to misclassifying an observation, namely $c(1|2)$ and $c(2|1)$. The total *expected misclassification cost* (ECM) is easily computed as

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2.$$

Our first goal is to minimise the ECM.

Proposition 9. *The regions R_1, R_2 that minimise the ECM are defined by the values $\underline{x}$, such that*

$$R_1 = \left\{ \underline{x} \colon \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right\},$$

and $R_2 = R_1^c$.

Minimising the ECM gives us one criterion for allocating new observations. Another criterion could be built via the posterior probability: Suppose we have an observation $\underline{x}_0$. Let us compute

$$\begin{aligned} P(\pi_1|\underline{x}_0) &= \frac{P(\underline{x}_0|\pi_1)P(\pi_1)}{P(\underline{x}_0|\pi_1)P(\pi_1) + P(\underline{x}_0|\pi_2)P(\pi_2)} \\ &= \frac{p_1 f_1(\underline{x}_0)}{p_1 f_1(\underline{x}_0) + p_2 f_2(\underline{x}_0)} \end{aligned}$$

We would then assign the observation $\underline{x}_0$ to π_1 if $P(\pi_1|\underline{x}_0) > P(\pi_2|\underline{x}_0)$.

7.1. **Classification among two multivariate normal populations.** Let's treat the case when the pdfs of the two populations are normal with the same covariance matrix $\mathbf{\Sigma}$. Then the criterion of minimal ECM in Proposition writes as

$$R_1 \quad \left\{ \underline{x} \colon -\frac{1}{2}(\underline{x} - \underline{\mu}_1)^T \mathbf{\Sigma}^{-1} (\underline{x} - \underline{\mu}_1) + \frac{1}{2}(\underline{x} - \underline{\mu}_2)^T \mathbf{\Sigma}^{-1} (\underline{x} - \underline{\mu}_2) \geq \log \left[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right] \right\}$$

$$\left\{ \underline{x} \colon (\underline{\mu}_1 - \underline{\mu}_2)^T \mathbf{\Sigma}^{-1} \underline{x} - \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)^T \mathbf{\Sigma}^{-1} (\underline{\mu}_1 + \underline{\mu}_2) \geq \log \left[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right] \right\},$$

where the first equality follows by taking logarithms on the criterion in Proposition 7.1 and the second by simple algebra.

As usual the parameters $\underline{\mu}_1, \underline{\mu}_2, \mathbf{\Sigma}$ are often not know and therefore have to be estimated. This is done by taking two samples $\underline{X}_{i1}, ..., \underline{X}_{in_i}, i = 1, 2$. Let $\overline{\underline{x}}_i, i = 1, 2$ the sample average of the population $i = 1, 2$ and $\mathbf{S}_i, i = 1, 2$ the sample covariance matrices of the two samples. As we have already seen, the common population

covariance $\mathbf{\Sigma}$ is estimated by an average of the two sample covariances, namely the pool sample covariance

$$\mathbf{S}_{pool} = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2$$

The region R_1 is then written as

(17)

$$R_1 \quad = \quad \left\{ \underline{x} \colon (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \mathbf{S}_{pool}^{-1} \underline{x} - \frac{1}{2}(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \mathbf{S}_{pool}^{-1} (\overline{\underline{x}}_1 + \overline{\underline{x}}_2) \geq \log \left[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right] \right\},$$

When the right hand side, ie the log, is zero, then the above criterion amounts to comparing

$$\hat{y} := (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \mathbf{S}_{pool}^{-1} \underline{x} = \hat{a}^T \underline{x} \tag{18}$$

where we define $\hat{a} := (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)\mathbf{S}_{pool}^{-1}$, the number

$$\hat{m} \quad := \quad \frac{1}{2}(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \mathbf{S}_{pool} (\overline{\underline{x}}_1 + \overline{\underline{x}}_2) \tag{19}$$

$$=: \quad \frac{1}{2}(\overline{y}_1 + \overline{y}_2), \tag{20}$$

where we define

$$\overline{y}_i := (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \mathbf{S}_{pool}\ \overline{\underline{x}}_i, \qquad i = 1, 2.$$

When formulated in terms of $\hat{y}$ and $\hat{m}$, ie assign observation $\underline{x}$ to π_1 if $\hat{y} \geq \hat{m}$ and to π_2 otherwise, is known as Fischer's Rule of Allocation. Fischer's Rule is actually more general and applies to nonormal populations, as well. It nevertheless, assumes implicitly that the two populations have the same covariance matrix since a pool estimate for the covariance matrix is used. The method is based on the idea of taking *projections* of the observation $\underline{x}_0$ on vectors $\hat{a} \in \mathbb{R}^p$. Suppose we take samples $\underline{x}_{i1}, ..., \underline{x}_{in}, i = 1, 2$ from the population $i = 1, 2$ and consider the projections $\underline{y}_{ij} := \hat{a}\underline{x}_{ij}, i = 1, 2,\ j = 1, 2..., n$. We then consider the normalised, squared separation distance

$$\frac{(\overline{\underline{y_1}} - \overline{\underline{y_2}})}{s_y^2} = \frac{(\hat{a}^T(\overline{\underline{x}}_i - \overline{\underline{x}}_2))^2)}{\hat{a}^T \mathbf{S}_{pool} \hat{a}}.$$

We now want to find a vector $\hat{a}$, such that the above distance is maximised. If then the projection of the observation $\underline{x}_0$ to the vector, which achieves the maximal separation distance, is greater than half this maximal distance, then we will assign the observation to population π_1.

To put the above into context, we first need to solve the optimisation problem

$$\max_{\hat{a} \in \mathbb{R}^p} \frac{(\hat{a}^T(\overline{\underline{x}}_i - \overline{\underline{x}}_2))^2)}{\hat{a}^T \mathbf{S}_{pool} \hat{a}}$$

this is the same maximisation problem as in (4) and the solution equals $(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \mathbf{S}_{pool}^{-1}(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)$, and is achieved when $\hat{a}_{max} = (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \mathbf{S}_{pool}^{-1}$. Notice then that the $\hat{m}$ that appears in (19) satisfies

$$\hat{m} = \frac{1}{2} \max_{\hat{a} \in \mathbb{R}^p} \frac{(\hat{a}^T(\overline{\underline{x}}_i - \overline{\underline{x}}_2))^2)}{\hat{a}^T \mathbf{S}_{pool} \hat{a}},$$

while $\hat{a}_{max}^T \underline{x}_0 = \hat{y}$, which is what appears in (18).

7.2. **Classification of Normal Populations when $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$.** This follows again from Proposition 7.1. Spelling it out for the case of normal populations with different covariances we get that the region R_1 equals

$$R_1 = \left\{ \underline{x} \colon \ -\frac{1}{2} \underline{x}^T (\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}) \underline{x} + (\underline{\mu}_1^T \mathbf{\Sigma}_1^{-1} - \underline{\mu}_2^T \mathbf{\Sigma}_2^{-1}) \underline{x} - k \geq \log \left[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right] \right\},$$

where $k = \frac{1}{2} \log \frac{\det \mathbf{\Sigma}_1}{\det \mathbf{\Sigma}_2} + \frac{1}{2} (\underline{\mu}_1^T \mathbf{\Sigma}_1^{-1} \underline{\mu}_1 - \underline{\mu}_2^T \mathbf{\Sigma}_2^{-1} \underline{\mu}_2)$.

Again, when $\underline{\mu}_1, \underline{\mu}_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ are not known, then they are replaced by the sample estimators.

8. Error Rates.

We need to develop a criterion to judge the performance of the classification. One natural candidate is the Total Probability of Misclassification (TPM), which we have already computed as

$$TPM = p_1 \int_{R_2} f_2(\underline{x}) d\underline{x} + p_2 \int_{R_1} f_2(\underline{x}) d\underline{x}.$$

The regions R_1, R_2 have to be chosen so that this measure is minimal and this is done by the choices of Proposition 7.1.

Example. Let's conider the case of two normal populations with the same covariance matrix $\mathbf{\Sigma}$, when $p_1 = p_2 = 1/2$ and $c(2|1) = c(1|2)$. Then the regions R_1, R_2 are given by (15). We then have that

$$\int_{R_2} f_1(\underline{x}) d\underline{x} = P\left(\underline{a}^T \underline{X} < \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_2)^T \mathbf{\Sigma}^{-1} (\underline{\mu}_1 + \underline{\mu}_2) \right), \tag{21}$$

where $\underline{a} = (\underline{\mu}_1 - \underline{\mu}_2)^T \mathbf{\Sigma}^{-1}$ and the probability P is meant to be with respect to the normal $\mathcal{N}_p(\underline{\mu}_1, \mathbf{\Sigma})$. Since $\underline{X}$ is assumed to be normal $\mathcal{N}_p(\underline{\mu}_1, \mathbf{\Sigma})$, the projection $\underline{a}^T \underline{X}$ is also (univariate) normal, with mean $\mu_{1Y} = \underline{a}^T \underline{\mu}_1$ and variance $\sigma_Y^2 = \underline{a}^T \mathbf{\Sigma}^{-1} \underline{a}$. Therefore if we denote $Y = \underline{a}^T \underline{X}$ then (21) write as

$$P\left(Y < \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_2)^T \mathbf{\Sigma}^{-1} (\underline{\mu}_1 + \underline{\mu}_2) \right)$$

where now the probability P is meant to be with respect to $\mathcal{N}(\mu_{1Y}, \sigma_Y^2)$. Subtracting the mean and dividing by σ_Y we get that this probabiiity equals

$$P(Z < \frac{-\frac{1}{2}\sigma_Y^2}{\sigma_Y}) = \Phi\left(-\frac{\sigma_Y}{2}\right)$$

where $Z = (Y - \mu_{1Y})/\sigma_Y$ is a standard normal variable. Φ is given by the tables of the $z-$ scores.

Finally, we compute in the same way the $\int_{R_1} f_2(\underline{x})d\underline{x}$ and the TPM would be the sum of the two quantities.

The problem with this approach is that the mean and the covariance matrix would be in practice unkown and so the region will be determined by the corresponding formulae when these parameters are replace by the sample mean and covariance, as given by (17). But even in this case there is another problem, namely, that it is often the case that even the distributions $f_1.f_2$ are not known. So the question comes to whether we can device a criterion based only on the available sample. This is achieved by the *confusion matrix*. This works as follows: Consider n_1 observations from the population π_1 and n_2 observations from π_2. Based on this sample we can construct the regions R_1, R_2 based on (17). Once we have these region we can look and see how many of the available observations fall into the right and how many into the worng region. In other words, let's say that n_{1C} of the n_1 observations from π_1 fall into region R_1 and n_{1M} into R_2. The former would had been classified correctly and the latter ones would be misclassified. We do the same thing with the n_2 observations and therefore split them into n_{2C} correctly classified and n_{2M} misclassified. These can now be put into a matrix form

$$\begin{pmatrix} n_{1C} & n_{1M} \\ n_{2M} & n_{2C} \end{pmatrix}$$

The criterion we are looking for is the Apparent Error Rate (APER) given by

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}.$$

We should remark that the APER tends to underestimate the actual error rate (AER), the reason being that we use the data that we used to build the criterion in order to test its efficiency. So often other criteria are seeked. For example we can split the sample into two parts, the one being the training sample (ie the one we use to build our criterion) and the other the validation sample (ie the one we use to test the criterion). The drawback of this is that we would prefer to use almost all available data to build our criterion.

An alternative approach is called the Lachenbruch approach, where

1. Start with the π_1 group. Omit one observation from this group and develop a classification function (criterion) based on the remaining $n_1 - 1$ observations.

2. Classify the holdout observation using the function built in step 1.

3. Repeat steps 1,2 until all observation are classified. Let n_{1M}^H the number of holdout observation which were misclassified.

4. Repeat the same procedure for the population π_2 and record the number of misclassified holdout observations n_{2M}^H.

We then ook at the proportion of misclassified observations

$$\frac{n_{1M}^H + n_{2M}^H}{n_1 + n_2}.$$

This estimator offers a nearly unbiased estimators for the Expected Actual Error (AER). This algorithm is often built in statistical packages.

Example. We refer to page 600 of Johnson-Wichern.

9. Exercises

1. Let $X_1 \sim^d \mathcal{N}(0,1)$ and let $X_2 = -X_1 1_{|X_1|\leq 1} + X_1 1_{|X_1|>1}$. Show that X_2 is also normal but (X_1, X_2) is not bivariate normal.

2. Check relation (4).

3. Use Minitab to generate 100 random numbers following a uniform distribution on $[0,1]$. Draw the normal probability plot and explain it. Does it deviate from linear ? Why is this ? Are there any negative values in the plot ? Why is this ?

4. Consider a sample $Y_1, Y_2, \ldots, Y_n$ of a uniform distribution and consider the order statistics $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$. Compute $E[Y_{(k)}]$.

5. Consider a random variable with strictly increasing cumulative distribution function F_Y.Consider, now, the random variable $X = F_Y(Y)$ and compute its distribution.

6. Prove Proposition 4.

7. Evaluate the T^2, for testing the hypothesis $H_0 \colon \underline{\mu}^T = (7, 11)$ using the data

$$\underline{X} = \begin{pmatrix} 2 & 12 \\ 8 & 9 \\ 6 & 9 \\ 8 & 10 \end{pmatrix} \tag{22}$$

Specify the distributio of T^2 in this case. Test H_0 at significance level $\alpha = 0.5$.

8. Consider the data table

Individual	X_1	X_2	X_3
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Determine the axes of the 90% confidence ellipsoid for $\underline{\mu}$. Determine the lengths of these axes.

Construct probability plots for each observation. Construct the pairwise scatter plots. Does the assumption of multivariate normality seem to be justified ?

9 In the frame of the example of page 10, test the null hypothesis $H_0 \colon \underline{\mu} = (0.55, 0.60)$ at $\alpha = 0.05$ significance level.

10 For the data of exercise 8. construct 95% simultaneous confidence intervals for μ_1, μ_2, μ_2.

11. A researcher considered two indices measuring the severity of heart attacks. The values of these indices for $n = 40$ heart-attack patients arriving at the hospital produced the summary statistics

$$\underline{\overline{x}} = \begin{pmatrix} 46.1 \\ 57.3 \end{pmatrix}$$

and

$$\mathbf{S} = \begin{pmatrix} 101.3 & 63.0 \\ 63.0 & 80.2 \end{pmatrix}$$

Test for equality of the mean indices, at significance level $\alpha = 0.5$.

12. Consider the correlation matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix}$$

Show that the factor problem $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathbf{T}} + \mathbf{\Psi}$ does not have a proper solution, for $m = 1$.

13. Consider the covariance matrix(concerning turtle measurements)

$$\mathbf{S} = 10^{-3} \begin{pmatrix} 11.072 & & \\ 8.019 & 6.417 & \\ 8.160 & 6.005 & 6.773 \end{pmatrix}$$

Compute the principal component solution for $m = 1$. Compute the residual matrix.

14. Verify the estimate (13).

15. Consider two data sets

$$\underline{X}_1 = \begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{pmatrix}$$

and

$$\underline{X}_1 = \begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}$$

Calculate the linear discriminant function cf. (18).

Classify the observation $\underline{x}_0^T = (2, 7)$. You may consider that the prior probabilities and the costs are equal.

16. Consider the data of example 11.1 of Johnson-Wichern. Develop a linear classification function for the data set. Using the above classification function construct the confusion matrix by classifying the given observations. Calculate the apparent error rate (APER) State any assumption made to justify the use of the method.

17. Let $f_1(x) = (1 - |x|)$, for $|x| \leq 1$ and $f_2(x) = (1 - |x - 5|)$, for $-0.5 \leq x \leq 1.5$. Sketch the two densities. Identify the classification regions when $p_1 = p_2$ and $c(1|2) = c(2|1)$. Identify the classification regions when $p_1 = 0.2$ and $c(1|2) = c(2|1)$.

18. Suppose that $n_1 = 11$ and $n_2 = 12$ observations are made on two random variables X_1, X_2 where X_1 and X_2 are assumed to have bivariate normal distributions with a common covariance matrix $\mathbf{\Sigma}$, but possibly different means $\underline{\mu}_1$,$\underline{\mu}_2$, for the two samples. The sample mean vectors and pooled covariance matrix are $\overline{\underline{x}}_1^T = (-1, -1)$, $\overline{\underline{x}}_1^T = (2, 1)$ and

$$\mathbf{S}_{pool} = \begin{pmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{pmatrix}$$

Test for the difference in the population mean vectors using the T^2 tests. Construct Fischer's linear discriminant function. Assign the observation $\underline{x}_0^T = (0, 1)$ to either population π_1, π_2.