5 Multivariate Sampling Distributions

In this course we focus on data that are assumed to be a random sample from a multivariate population. That is, we observe n independent copies, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ of a p-dimensional random variable.

1. The sample mean $\bar{\mathbf{X}}$ is an unbiased estimate of $\boldsymbol{\mu}$, i.e. $E\bar{\mathbf{X}} = \boldsymbol{\mu}$.
2. $Var(\bar{\mathbf{X}}) = \frac{1}{n}\Sigma$
3. The sample variance matrix

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T$$

is an unbiased estimate of Σ, i.e. $ES = \Sigma$.

Proof as an exercise.

Definition The term $|S|$ is called a *generalized variance*.

5.1 Large Sample Results for a General Case

1. $\bar{\mathbf{X}} \to \boldsymbol{\mu}$ and $S \to \Sigma$ in probability
2. $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \Sigma)$
3. Combining (1) and (2), we get

$$\sqrt{n}S^{-1/2}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \Sigma)$$

4. This implies that

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T S^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} \chi^2(p)$$

Recall that $a_n \xrightarrow{p} a$, convergence in probability, means as the sample size gets large, a_n converge to a in the sense that the probability that a_n is "close" to b goes to 1.

Convergence in distribution, *statistic* $\xrightarrow{d} F$ means that as the sample size gets large the distribution of the statistic becomes approximately equal to the distribution F.

5.2 Large Sample Results for a Multivariate Normal Case

Now consider the special case where we assume that the common distribution of our random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is $N_p(\boldsymbol{\mu}, \Sigma)$. Our additional assumption of normality provides us with some stronger results:

1. Unlike the general case, these results are exact and hold for any sample size n:

$$\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, n^{-1}\Sigma)$$
$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T S^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi^2(p)$$

2. The random $p \times p$ matrix nS follows the **Wishart distribution**.

   The Wishart distribution has two parameters, namely the degrees of freedom (i.e. n in this case) and the *scale matrix* (Σ in this case). We denote it by $nS \sim W_p(\Sigma, n)$. In other words, the Wishart distribution $W_p(\Sigma, m)$ is a probability distribution of random nonnegative-definite $p \times p$ matrices that is used to model random covariance matrices.

3. $\bar{\mathbf{X}}$ and S are independent.

Assessment of Normality We can use the following strategies:

- We have seen that $\mathbf{X} = (X_1, \ldots, X_p)^T$ is multivariate normal if and only if every linear combination of $X_1, \ldots, X_p$ is normal. (Hard!)
- As a subcase of the method above, we can check whether each component of $\mathbf{X}$ is univariate normal. (Easier!)
- Alternatively, we can also analyze linear combinations of $\hat{\mathbf{e}}_1^T\mathbf{X}$ and $\hat{\mathbf{e}}_p^T\mathbf{X}$ where $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_p$ are the eigenvectors associated with the largest and smallest eigenvalues of S respectively.

To assess univariate normality, we can use QQ plots, Shapiro-Wilk test etc. Under certain particular cases of nonnormality, we can apply power transformations such as $\log(x)$ and x^κ for $\kappa \neq 0$. More on multivariate normality in the next few sections.

6 Maximum Likelihood Estimation

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent p-dimensional vectors that follow $N(\boldsymbol{\mu}, \Sigma)$. The joint probability density function is

$$\begin{aligned} f(\mathbf{X}_1, \ldots, \mathbf{X}_n) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{X}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu})\right] \\ &= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu})\right] \end{aligned}$$

Now, if we treat $\mathbf{X}_1, \ldots, \mathbf{X}_n$ as given, $f(\mathbf{X}_1, \ldots, \mathbf{X}_n|\boldsymbol{\mu}, \Sigma)$ becomes a function of $\boldsymbol{\mu}$ and Σ and is called the *likelihood* function of the random sample, denoted by $L(\cdot, \cdot)$. (Exactly as in the univariate case!) Then, the values of of $\boldsymbol{\mu}$ and Σ are called the maximum likelihood estimates (MLE) of mean and covariance, respectively.

By an analogy with the univariate case, it is not hard to guess that the ML estimates in the multivariate case will be

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i, \tag{6.1}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \tag{6.2}$$

Recall that the term $|S|$ is called a *generalized variance* where S is the unbiased estimate of the covariance matrix. Hence, since $|\hat{\Sigma}| = (n-1)^p/n^p|S|$, then

$$L(\boldsymbol{\mu}, \Sigma) = \text{constant} \times (\text{generalized variance})^{-\text{n}/2}.$$

Since the likelihood function of a random sample of normal distribution only depends on the sample mean $\bar{\mathbf{X}}$ and the sum-of-squares-and-cross-product matrix $\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}_i - \bar{\mathbf{X}}) = (n-1)S$, we can state that $\bar{\mathbf{X}}$ and S are the sufficient statistics of the multivariate normal distribution.

7 Inference about Sample Mean

7.1 Hotelling T^2

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a p-dimensional normal population with mean $\boldsymbol{\mu}$ and positive variance-covariance matrix Σ.

We are interested in the hypothesis testing problem:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \tag{7.3}$$

$$H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{7.4}$$

where $\boldsymbol{\mu}_0$ is a known vector. Note that this is a generalization of the one-sample t-test of the univariate case. When p is 1, the test statistic is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

and statistic t follows a Student distribution with $n-1$ degrees of freedom.

In a multivariate case, we can naturally generalize this test statistic

$$t^2 = (\bar{X} - \mu_0)\frac{n}{s^2}(\bar{X} - \mu_0)$$

to

$$T^2 = (\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' n S^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0),$$

where $S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}_i - \bar{\mathbf{X}})$ is our sample covariance matrix.

Definition Statistic T^2 is called the Hotelling's T^2 statistic. Statistic T^2 is distributed as $\frac{(n-1)p}{n-p}F_{p,n-p}$, where $F_{u,v}$ is the F-distribution with degrees of freedom u and v.

We know that $[t_{n-1}(\alpha/2)]^2 = F_{1,n-1}(\alpha)$, where $F_{u,v}(\alpha)$ is the upper 100α percentile of the F-distribution with u and v degrees of freedom. Hence, for a univariate case T^2 reduces to the usual one-sample t-statistic.

Example. Consider the daily closing prices of the two major European stock indices: Germany DAX and Switzerland SMI, 1991–1998. The data are sampled in business time, i.e., weekends and holidays are omitted. We look at the log returns of DAX and SMI and test the hypothesis that the log returns are 0.

```
> eu_data<-cbind(dax,smi)
```

```
> colMeans(eu_data)
         dax          smi
0.0006520417 0.0008178997
```

We start from testing multivariate normality. Notice that the true squared Mahalanobis distance $(\mathbf{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})$ has a χ^2-distribution with p degrees of freedom. When the sample is large, the estimated Mahalanobis distance $(\mathbf{X}-\hat{\boldsymbol{\mu}})^T\hat{\Sigma}^{-1}(\mathbf{X}-\hat{\boldsymbol{\mu}})$ will approximately follow χ^2-distribution with p-degrees of freedom. Then we can use this result to evaluate a multivariate normal distribution and outliers using quantile-quantile (QQ) plot.

There are many formal tests for multivariate normality that are essentially extensions from the univariate setting. The Mardia test (see the example below) is based on multivariate extensions of skewness and kurtosis measures and then using the fact that under H_0 they will follow χ^2-distribution (Mardia, 1970). The Mardia sample statistic for multivariate skewness is

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\mathbf{X}_i-\bar{\mathbf{X}})^T S^{-1}(\mathbf{X}_j-\bar{\mathbf{X}})\right]^3$$

and the corresponding sample statistic for kurtosis is

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{X}_i-\bar{\mathbf{X}})^T S^{-1}(\mathbf{X}_i-\bar{\mathbf{X}})\right]^2.$$

Note that both Mardia's skewness and kurtosis are functions of the squared Mahalanobis distances. A large value of multivariate kurtosis, in comparison to the expected value under normality, indicates that one or more observations have a large Mahalanobis distance and are thus located far from the centroid of the data set.

Mardia (1970, 1974) determined the asymptotic distributions of the multivariate skewness and kurtosis statistics. If we consider $A = nb_{1,p}/6$, it can be shown that A is asymptotically

distributed as a χ^2-random variable with $p(p+1)(p+2)/6$ degrees of freedom. Similarly, the statistic $B = [b_{2,p} - p(p+2)]/\sqrt{8p(p+2)/n}$ is s asymptotically distributed as a standard normal random variable.

```
> library(MVN)
```

```
> mardiaTest(eu_data, qqplot=TRUE)
   Mardia's Multivariate Normality Test
---------------------------------------
   data : eu_data

   g1p            : 0.663891
   skew           : 205.6956
   p.value.skew   : 2.23961e-43

   g2p            : 18.50902
   kurtosis       : 56.6385
   p.value.kurt   : 0

   small.skew     : 206.2493
   p.value.small  : 1.702512e-43

   Result         : Data is not multivariate normal.
---------------------------------------
```

Now we apply the Hotelling T^2 test:

```
> library(ICSNP)
> HotellingsT2(eu_data, mu = c(0,0), test = "chi")

        Hotelling's one sample T2-test
```

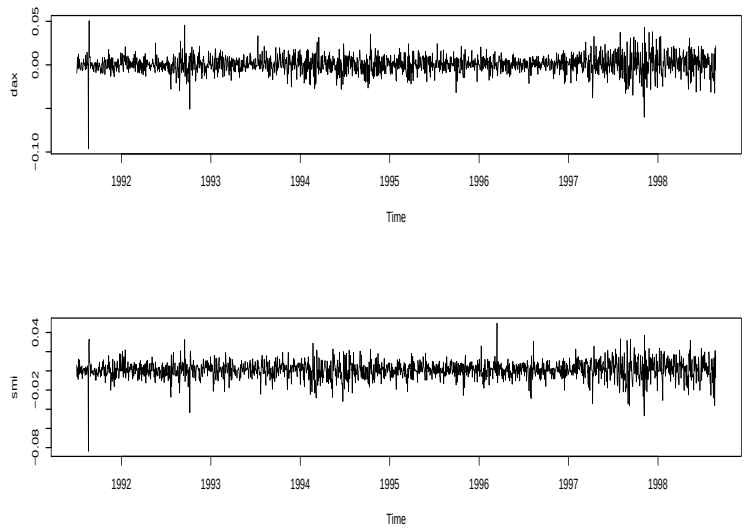

Figure 7.1: Time series plots of log returns of DAX and SMI

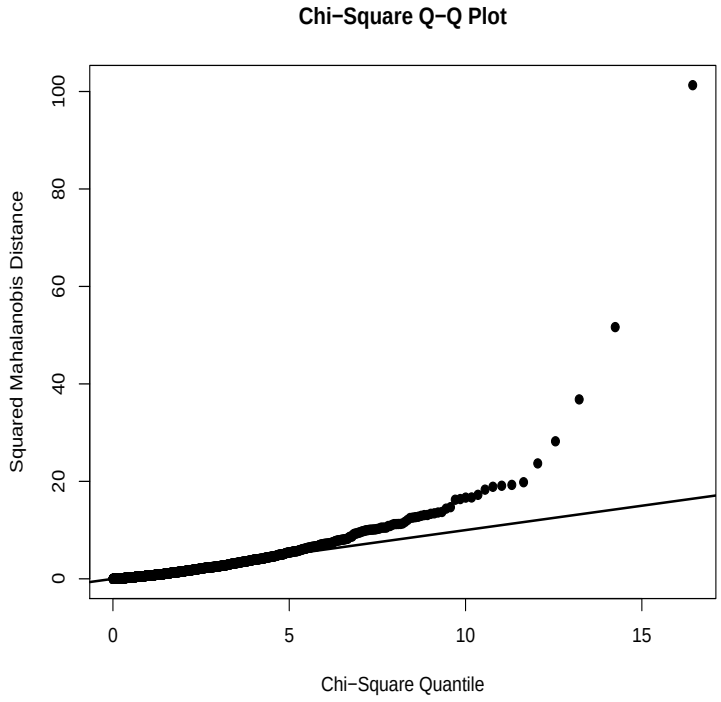

Figure 7.2: Quantile-Quantile (QQ) plot of log returns of DAX and SMI

```
data:  eu_data
T.2 = 14.5389, df = 2, p-value = 0.0006965
alternative hypothesis: true location is not equal to c(0,0)
```

and univariate tests:

```
> t.test(eu_data[,1])

        One Sample t-test

data:  eu_data[, 1]
t = 2.7292, df = 1858, p-value = 0.006408
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.0001834832 0.0011206003
sample estimates:
   mean of x
0.0006520417

> t.test(eu_data[,2])

        One Sample t-test

data:  eu_data[, 2]
t = 3.8124, df = 1858, p-value = 0.0001421
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.0003971393 0.0012386600
sample estimates:
   mean of x
0.0008178997
```

Why not just do p separate (univariate) t-tests for each component of $\mathbf{X}$?

- Does not take into account interrelationships (correlations) among the components of $\mathbf{X}$.
- The multivariate test controls the type I error rate, the probability of rejecting H_0 when H_0 is true.
  - When doing multiple univariate tests, the probability that at least one univariate null hypothesis will be rejected by chance alone increases with the number of tests being performed.
  - There are many procedures to account for this problem (e.g., Bonferroni) but generally such procedures are not as attractive as performing a single multivariate test. However, there are pros and cons for both approaches.