4.3. Likelihood and maximum likelihood estimation.

Suppose $X_1, ..., X_n$ are iid random p-vectors $\sim MN(\mu, \Sigma)$. Their joint density is

$$\begin{aligned} f(x_1, ..., x_n) &= \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\{-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\} \\ &= \exp\Big\{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)'\Sigma^{-1}(x_i-\mu) - \frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma|\Big\} \end{aligned}$$

for $x_i \in R^p, i = 1, ..., n$. Thus, the likelihood based on observations $x_1, ..., x_n$ of $X_1, ..., X_n$ is

$$l_n(\mu, \Sigma) = \exp\Big\{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)'\Sigma^{-1}(x_i-\mu) - \frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma|\Big\},$$

which is formally same as the joint density function, but viewed as a function of the parameters (μ, Σ). The maximum likelihood estimation of μ and Σ, based on $X_1, ..., X_n$, is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \frac{n-1}{n}\mathbf{S} = \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})(X_i-\bar{X})',$$

respectively. The following is a technical proof, which shall not be required.

Proof. The proof uses some tricks involving matrix. Set $\tilde{\mathbf{S}} = [(n-1)/n]\mathbf{S}$. Write

$$\begin{aligned} &\log(l_n(\mu, \Sigma)) \\ \propto\ & -\frac{1}{2}\sum_{i=1}^n (X_i-\mu)'\Sigma^{-1}(X_i-\mu) - \frac{n}{2}\log|\Sigma| \\ =\ & -\frac{1}{2}\sum_{i=1}^n trace\{(X_i-\mu)'\Sigma^{-1}(X_i-\mu)\} - \frac{n}{2}\log|\Sigma| \\ =\ & -\frac{1}{2}\sum_{i=1}^n trace\{\Sigma^{-1}(X_i-\mu)(X_i-\mu)'\} - \frac{n}{2}\log|\Sigma| \qquad trace(\mathbf{AB}) = trace(\mathbf{BA}) \\ =\ & -\frac{n}{2}trace\{\Sigma^{-1}\frac{1}{n}\sum_{i=1}^n (X_i-\mu)(X_i-\mu)'\} - \frac{n}{2}\log|\Sigma| \\ =\ & -\frac{n}{2}trace\{\Sigma^{-1}[\tilde{\mathbf{S}} + (\bar{X}-\mu)(\bar{X}-\mu)']\} - \frac{n}{2}\log|\Sigma| \\ =\ & -\frac{n}{2}trace(\Sigma^{-1}\tilde{\mathbf{S}}) + \frac{n}{2}\log|\Sigma^{-1}| - \frac{n}{2}trace[\Sigma^{-1}(\bar{X}-\mu)(\bar{X}-\mu)'] \\ =\ & -\frac{n}{2}trace(\tilde{\mathbf{S}}^{1/2}\Sigma^{-1}\tilde{\mathbf{S}}^{1/2}) + \frac{n}{2}\log(|\tilde{\mathbf{S}}^{1/2}\Sigma^{-1}\tilde{\mathbf{S}}^{1/2}|) - \frac{n}{2}\log(|\tilde{\mathbf{S}}|) - \frac{n}{2}trace[(\bar{X}-\mu)\Sigma^{-1}(\bar{X}-\mu)'] \\ =\ & -\frac{n}{2}[trace(\tilde{\mathbf{S}}^{1/2}\Sigma^{-1}\tilde{\mathbf{S}}^{1/2}) - \log(|\tilde{\mathbf{S}}^{1/2}\Sigma^{-1}\tilde{\mathbf{S}}^{1/2}|)] - \frac{n}{2}trace[(\bar{X}-\mu)\Sigma^{-1}(\bar{X}-\mu)'] - \frac{n}{2}\log(|\tilde{\mathbf{S}}|) \\ =\ & -\frac{n}{2}[\sum_{k=1}^p \lambda_k - \sum_{k=1}^p \log(\lambda_k)] - \frac{n}{2}trace[(\bar{X}-\mu)\Sigma^{-1}(\bar{X}-\mu)'] - \frac{n}{2}\log(|\tilde{\mathbf{S}}|) \\ & \qquad\qquad\qquad \text{(where } \lambda_k \text{ are eigenvalues of } \tilde{\mathbf{S}}^{1/2}\Sigma^{-1}\tilde{\mathbf{S}}^{1/2}) \\ \leq\ & -\frac{n}{2}p - \frac{n}{2}\log(|\tilde{\mathbf{S}}|), \end{aligned}$$

with the equality holding only when $\bar{X} = \mu$ and $\lambda_k = 1, k = 1, ..., p$, since the function $x - \log(x), x > 0$ is minimized as 1 only when $x = 1$. . It implies $\tilde{\mathbf{S}}^{1/2}\Sigma^{-1}\tilde{\mathbf{S}}^{1/2} = I_p$, the $p \times p$ identity matrix. Therefore $\log l_n(\mu, \Sigma)$ is maximized only when $\mu = \bar{X}$ and $\Sigma = \tilde{S}$, proving that the maximum likelihood estimator of (μ, Σ) is $(\bar{X}, \tilde{S})$. □

It is clear that

$$\bar{X} \sim MN(\mu, (1/n)\Sigma).$$

The $p \times p$ matrix $(n-1)\mathbf{S} = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$ is called a *Wishart random matrix* of degree of freedom $n-1$, which has the same distribution as $\sum_{i=1}^{n-1} Z_i Z_i'$ where Z_i are iid $\sim MN(0, \Sigma)$. Its distribution is denoted as $W_{n-1}(\Sigma)$. Moreover, $\bar{X}$ and $\mathbf{S}$ are independent.

4.4. Checking normality.

(i). Checking univariate normality.

Let $x_1, ..., x_n$ be the observations. And let $x_{(1)} \leq ... \leq x_{(n)}$ represent the order statistics.

Graphical approach: (qqnorm) plot the quantiles of the observations $x_1, ..., x_n$ against the *corresponding* quantiles of the standard normal distribution. To be precise, the plots are $(q_{(i)}, x_{(i)}), i = 1, ..., n$, where $q_{(i)} = \Phi^{-1}((i - 1/2)/n)$, the quantile/percentile of the standard normal distribution at level $(i - .5)/n$. If the plot is associate with a straight line, it's the evidence of normality and the slope and intercept represent the standard deviation and mean.

Quantitative approach: Define

$$r_Q = \frac{\sum_{i=1}^{n}(x_{(i)} - \bar{x})(q_{(i)} - \bar{q})}{\sqrt{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2 \sum_{i=1}^{n}(q_{(i)} - \bar{q})^2}},$$

where is the sample correlation between $(q_{(i)}, x_{(i)}), i = 1, ..., n$. The justification criterion is: *r_Q being close to 1 is the evidence of normality of the underlying distribution.* The smaller the r_Q, the stronger the evidence against normality.

Remark. The explanation of the above justification criterion is as follows. Correlation of two random variables is 1 if and only if one is a linear function of the other with positive slope. Sample correlation of two variables being close to 1 indicates they are nearly positively linearly related. Hence r_Q being close to 1 implies the $x_{(i)} \approx aq_{(i)} + b$, $i = 1, ..., n$ for some constant $a > 0$ and b. Since $x_{(i)} \approx F_X^{-1}(i/n)$ and $q_{(i)} = \Phi^{-1}((i-1/2)/n)$, $i = 1, ..., n$, it follows that $F_X^{-1}(t) \approx a\Phi^{-1}(t) + b$, assuming n is reasonably large. According to the claim presented and proved on page 2, F_X is close to the distribution $aN(0,1) + b$, which is $N(b, a^2)$.

Example 1.1. (continued, see Appendix A.) Ten American Companies. In this example, r_Q is computed for each *single* variable:

	Sales	Profit	Asset
r_Q:	0.936	0.948	0.936
p-value:	> 0.10	> 0.10	> 0.10

where the p-value is found by checking Table 4.2 on page 182 of the textbook, which is partly listed in the following:

	Significance levels		
Sample size n	.01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
⋮	⋮	⋮	⋮
300	.9935	. 9953	.9960

These are critical points for the qq-plot correlation of coefficient test of normality at significance levels $\alpha = 0.01$, 0.05 and 0.10:

Note that here the p-values are those of hypotheses testing. To be specific, we consider, for example, for the variable profit the hypothesis

$$\begin{cases} H_0: & \text{the variable profit follows normal distribution} \\ H_a: & \text{the variable profit does not follow normal distribution.} \end{cases}$$

The p-value being greater than 0.10 implies the present data does not contain significant evidence (at level 10%) against the assumption that the profit variable follows a normal distribution.

As previously mentioned, Example 1.1 is not a good example in that it is not a random sample from certain given population. For the purpose of demonstration, we may *pretend* it is a random sample from all listed companies in NYSE. Of course it's only a pretension.

(ii). Checking bivariate or multivariate normality.

Suppose our data consist of observations $X_1, ..., X_n$ of p variables. In other words, each X_i is of p dimension. To check whether the population distribution is multivariate (p dimension) normal, we

1. check the normality for each component or, in other words, each single variable, by using graphical approach qqplot or numerical approach r_Q test.

2. check the pairwise relations for any pairs of components by using scatter plot: No systematic dependence other than association with a straight line subject to error corresponds to evidence of normality. This is because, according to Proposition 4.1, a component of a bivariate normal random variable is a linear function of another component plus another independent normal random variable. See the Remark following Proposition 4.1.

3. use χ^2 plot: plot the quantiles of the squares of statistical distances:

$$\{(X_1 - \bar{X})'\mathbf{S}^{-1}(X_1 - \bar{X}),\ ...,\ (X_n - \bar{X})'\mathbf{S}^{-1}(X_n - \bar{X})\}$$

against the corresponding quantiles of χ^2_p. If the plot is associated with the straight line at 45° through the origin, it's evidence of normality. This is because $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_p$ if $X \sim MN(\mu, \Sigma)$.

Remark. In theory, lower dimension normality does not ensure higher dimension normality. In practice, normality at one or two dimension is usually deemed enough evidence for high dimension normality.

4.5. Transformation.

In practice, the target variable very often clearly does not follow normal distribution. Very often we take reasonable transformation of the variable so that the transformed variable follows normal distribution and the normality based statistical methodology can be applied to the transformed variable. Among a variety of transformations, the Box-Cox transformation, also called power transformation, is the most popular. Suppose the variable x is positive, The Box-Cox transformation is the following class of transformation, indexed by λ, of x:

$$x^{(\lambda)} = \begin{cases} (x^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log x & \lambda = 0. \end{cases}$$

Essentially, it is simply power function or log function. It is presented in the above seemingly more complex form in order to show that log-transformation is a limit of power transformation.

A natural question then arises: what should be the most ideal λ for the Box-Cox transformation? A crude but intuitive rule of thumb is

choose $\lambda < 1$ if the distribution of the observations is skewed to the right.

choose $\lambda > 1$ if the distribution of the observations is skewed to the left.

A numerical criterion is: choose λ so that

$$l(\lambda) \equiv -\frac{n}{2}\log[\frac{1}{n}\sum_{j=1}^{n}(x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2] + (\lambda - 1)\sum_{j=1}^{n}\log x_j$$

is maximized, where $\overline{x^{(\lambda)}}$ is the sample average of the transformed data.

For multivariate observations $X_1, ..., X_n$ of p dimension, p power transformations, one on each variable/component, may be considered all at once. Let $X_i = (x_{i1}, ..., x_{ip})'$. We are to choose λ_k to perform power transformation on the k-th variable. Specifically, the observations of the k-th variable: $(x_{1k}, ..., x_{nk})$ shall be transformed to $(x_{1k}^{(\lambda_k)}, ..., x_{nk}^{(\lambda_k)})$, where

$$x_{ik}^{(k)} = \begin{cases} (x_{ik}^{\lambda_k} - 1)/\lambda_k & \lambda_k \neq 0 \\ \log x_{ik} & \lambda_k = 0. \end{cases} \qquad i = 1, ..., n.$$

An extension of the above criterion is: choose $(\lambda_1, ..., \lambda_p)$ so that

$$l(\lambda_1, ..., \lambda_p) \equiv -\frac{n}{2}\log(|\mathbf{S}(\lambda)|) + \sum_{k=1}^{p}(\lambda_k - 1)\sum_{i=1}^{n}\log x_{ik}$$

is maximized, where $\mathbf{S}(\lambda)$ is the sample covariance matrix based on the transformed data.