4 MULTIVARIATE DISTRIBUTIONS

4.1 DISTRIBUTION AND DENSITY FUNCTION

Sections 4.1, 4.2, 7.1 in Härdle and Simar.

Let $X = (X_1, \ldots, X_p)^T$ be a random vector.

- For all $x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, the cumulative distribution function (cdf), or distribution function, of X is defined by

$$F(x) = P(X \leq x) = P(X_1 \leq x_1, \ldots, X_p \leq x_p)$$

- If X is continuous, the probability density function (pdf) or density, f, of X is a nonnegative function defined through the following equation:

$$F(x) = \int_{-\infty}^{x} f(u)\, du\,;$$

it always satisfies

$$\int_{-\infty}^{\infty} f(u)\, du = 1\,.$$

☛ The integrals are p-variate, $u \in \mathbb{R}^p$ but $f(u) \in \mathbb{R}$:

$$\int_{-\infty}^{x} f(u)\, du = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_p} f(u_1, \dots, u_p)\, du_1 \dots du_p\,.$$

- The marginal cdf a subset of X is obtained by the marginal of X computed at the subset, letting the other values equal to infinity.

☛ e.g. the marginal cdf of X_1 is

$$\begin{aligned}F_{X_1}(x_1) =& P(X_1 \leq x_1)\\ =& P(X_1 \leq x_1, X_2 \leq \infty, \ldots, , X_p \leq \infty)\\ =& F_X(x_1, \infty, \ldots, \infty)\end{aligned}$$

☛ e.g.the marginal cdf of (X_1, X_3) is

$$\begin{aligned}F_{X_1,X_3}(x_1, x_3) =& P(X_1 \leq x_1, X_3 \leq x_3)\\ =& P(X_1 \leq x_1, X_2 \leq \infty, X_3 \leq x_3, X_4 \leq \infty \ldots, , X_p \leq \infty)\\ =& F_X(x_1, \infty, x_3, \infty, \ldots, \infty)\,.\end{aligned}$$

- For a continuous random vector X, the marginal density of a subset of X is obtained from the joint density f of X by integrating out the other components.

    ☛ e.g. the marginal density X_1 is

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, u_2, \ldots, u_p)\, du_2 \ldots du_p$$

    ☛ e.g. the marginal density of (X_1, X_3) is

$$f_{X_1,X_3}(x_1, x_3) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, u_2, x_3, u_4, \ldots, u_p)\, du_2\, du_4 \ldots du_p\,.$$

- For two continuous random vectors X_1 and X_2, the conditional pdf of X_2 given X_1 is obtained by taking

$$f(x_2|x_1) = f(x_1, x_2)/f_{X_1}(x_1)\,.$$

(Defined only for values x_1 such that $f_{X_1}(x_1) > 0$)

- Two continuous random vectors X_1 and X_2 are independent if and only if

$$f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

☛ If X_1 and X_2 are independent then

$$f_{X_2|X_1}(x_2|x_1) = f(x_1, x_2)/f_{X_1}(x_1) = f_{X_1}(x_1) f_{X_2}(x_2)/f_{X_1}(x_1) = f_{X_2}(x_2)\,.$$

(Knowing the value of X_1 does not change the probability assessments on X_2 and vice versa)

- The mean $\mu \in \mathbb{R}^p$ of $X = (X_1, \ldots, X_p)^T$ is defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \int x f_{X_1}(x)\, dx \\ \vdots \\ \int x f_{X_p}(x)\, dx \end{pmatrix}.$$

☛ If X and Y are two p-vectors and α and β are constants then

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$

☛ If X is a $p \times 1$ vector which is independent of the $q \times 1$ vector Y then

$$E(XY^T) = E(X)E(Y^T).$$

☛ Hint: Remember to always check that matrix dimensions are compatible.

- As seen earlier, the covariance Σ of a vector X of mean μ is defined by

$$\Sigma = E\{(X - \mu)(X - \mu)^T\}\,.$$

We write

$$X \sim (\mu, \Sigma)$$

to denote a vector X with mean μ and covariance Σ.

- We can also define a covariance matrix between a $p \times 1$ vector X of mean μ and a $q \times 1$ vector Y of mean ν by

$$\Sigma_{X,Y} = cov(X, Y) = E\{(X - \mu)(Y - \nu)^T\} = E(XY^T) - E(X)E(Y^T)\,.$$

The elements of this matrix are the pairwise covariances between the components of X and those of Y.

☛ We have

$$cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)$$

☛ We have

$$var(X+Y) = var(X) + cov(X,Y) + cov(Y,X) + var(Y)$$

☛ For matrices A and B and random vectors X and Y such that the below quantities are well defined we have

$$cov(AX,BY) = A\,cov(X,Y)B^T\,.$$

- The conditional expectation $E(X_2|X_1 = x_1)$ is defined by

$$E(X_2|X_1 = x_1) = \int x_2 f_{X_2|X_1}(x_2|x_1)\, dx_2$$

and the conditional (co)variance $var(X_2|X_1 = x_1)$ is defined by

$$var(X_2|X_1 = x_1) = E(X_2 X_2^T|X_1 = x_1) - E(X_2|X_1 = x_1)E(X_2^T|X_1 = x_1)\,,$$

if X_2 is a column vector.

- We also have the *law of total expectation* and *law of total variance*:

$$E(X_2) = E(E(X_2|X_1))$$
$$var(X_2) = E(var(X_2|X_1)) + var(E(X_2|X_1))$$

- The characteristic function of a random vector $X \in \mathbb{R}^p$ is a complex valued function $\varphi_X(\cdot)$ on $\mathbb{R}^p$ defined by

$$\varphi_X(t) = \mathbb{E}[e^{it^T X}],$$

where $t \in \mathbb{R}^p$ is the argument, and i is the imaginary number in the complex plane such that $i^2 = -1$.

- By definition, for any real number a, $e^{ia} \equiv cos(a) + isin(a)$. Hence,

$$\varphi_X(t) = \mathbb{E}[cos(t^T X)] + i\mathbb{E}[sin(t^T X)].$$

Note that these expectations (and hence the characteristic function) are always well-defined, as $|cos(t^T X)|, |sin(t^T X)| \leq 1$ for all values of t.

- *A characteristic function uniquely defines a probability distribution*: If $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^p$ are two random vectors such that $\varphi_X(t) = \varphi_Y(t)$ for all $t \in \mathbb{R}^p$, the distributions of X and Y must be the same.

- The non-trivial property above can be proved by Fourier method; see most graduate-level book on probability theory, such as Chung (2001).

Sections 4.4, 4.5, 5.1 in Härdle and Simar.

☛ In the univariate case, the density of a $N(\mu, \sigma^2)$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}.$$

☛ In the multivariate case with $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, its covariance matrix, as usual, can be denoted as

$$\Sigma = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1p} \\ & \vdots & \\ \sigma_{p1} & \ldots & \sigma_{pp} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \ldots & \sigma_{1p} \\ & \vdots & \\ \sigma_{p1} & \ldots & \sigma_p^2 \end{pmatrix}$$

where $\sigma_j^2 = \mathrm{var}(X_j)$. Its mean is also denoted as $\mu = (\mu_1, \ldots, \mu_p)^T$.

☛ If a p-vector X is normal with mean μ and covariance Σ, we write

$$X \sim N_p(\mu, \Sigma).$$

☛ Let $X \sim N_p(\mu, \Sigma)$, A a $q \times p$ matrix and b a $q \times 1$ vector. Then

$$Y = AX + b \sim N_q(A\mu + b, A\Sigma A^T)\,.$$

☛ Let $X = (X_1^T, X_2^T)^T \sim N_p(\mu, \Sigma)$ where

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

and

$$var(X_1) = \Sigma_{11}, \;\; var(X_2) = \Sigma_{22}.$$

Then one can prove that

$$\Sigma_{12} = 0 \text{ if and only if } X_1 \text{ and } X_2 \text{ are independent.}$$

☛ So between normal random variables/vectors, *zero covariances do imply independence*.

☛ The density of a multinormal distribution with mean μ and covariance $\Sigma > 0$ is given by

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}. \tag{1}$$

☛ If Σ is invertible and $\sigma_{ij} = 0$ for all pairs $i \neq j$, we have

$$\Sigma^{-1} = diag(\sigma_1^{-2}, \ldots, \sigma_p^{-2}),$$

and the density satisfies

$$\begin{aligned} f(x) =& \frac{1}{\sqrt{(2\pi)^p}\prod_{j=1}^p \sigma_j} \exp\left\{ -\frac{1}{2}\sum_{j=1}^p (x_j-\mu_j)^2/\sigma_j^2\right\} \\ =& \frac{1}{\sqrt{(2\pi)^p}\prod_{j=1}^p \sigma_j} \prod_{j=1}^p \exp\left\{ -\frac{1}{2}(x_j-\mu_j)^2/\sigma_j^2\right\} \\ =& \prod_{j=1}^p \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -(x_j-\mu_j)^2/(2\sigma_j^2)\right\}, \end{aligned}$$

which is the product of the densities of p univariate $N(\mu_j, \sigma_j^2)$.

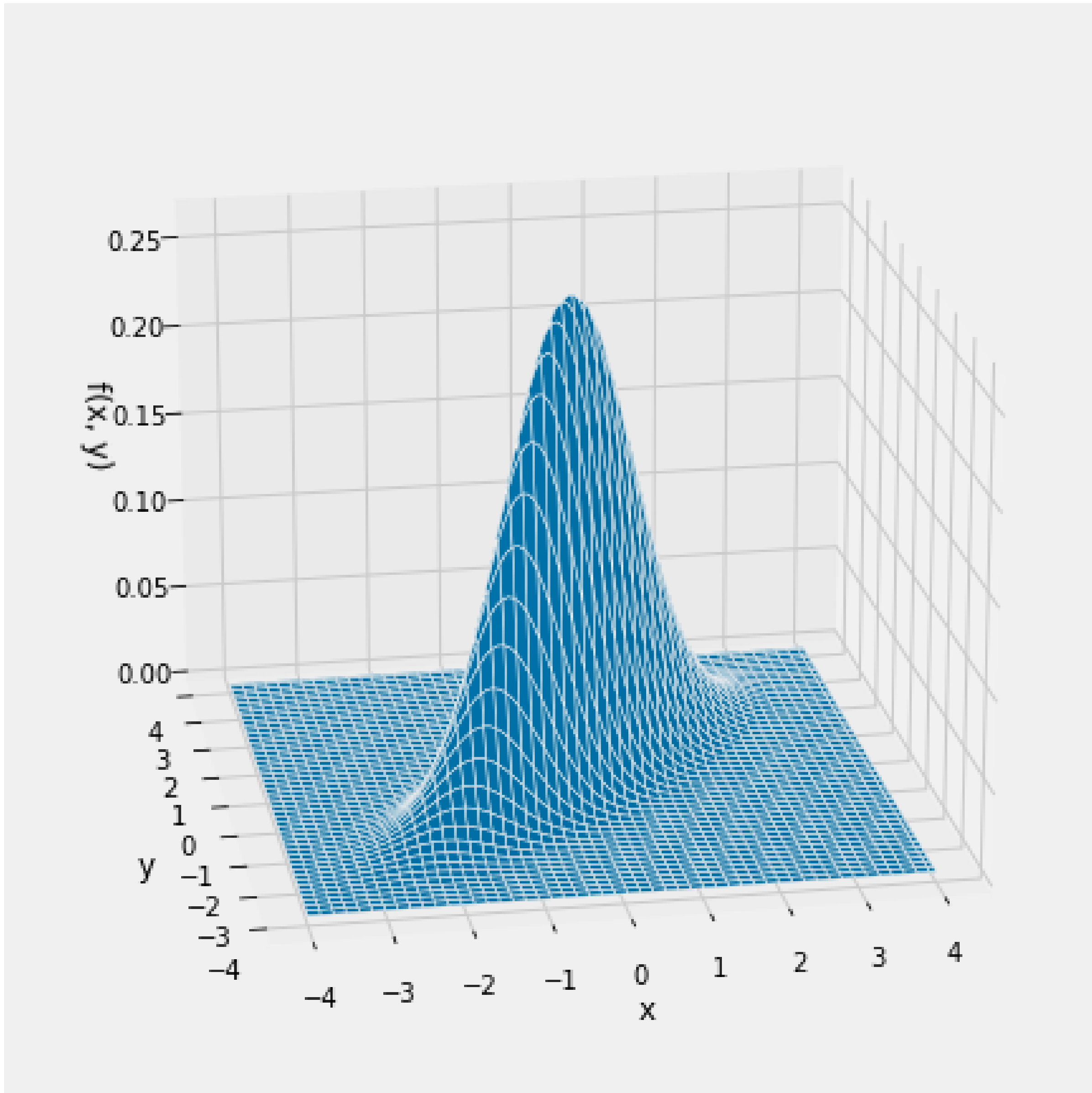
☛ When we define the multivariate normal density, we need to assume that Σ is *strictly* positive definite, i.e. $\Sigma > 0$. In this case, Σ is invertible, and the density exists.

☛ More generally, one can define a normal distribution even when the covariance Σ is non-invertible, i.e.

$$\Sigma \geq 0 \text{ and } \det(\Sigma) = 0.$$

☛ In the latter case, we would end up with a *degenerate* normal distribution where a density function cannot be defined on $\mathbb{R}^p$; all probability mass lies in a lower dimensional subspace of $\mathbb{R}^p$

Figure 1: A "near-degenerate" bivariate centered (i.e., $\mu = (0, 0)^T$) normal density, $\sigma_{11} = \sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = 0.8$. If σ_{12} becomes 1, all the mass collapses onto the line $x = y$.



☛ In fact, even a constant number $c \in \mathbb{R}$ is trivially a degnerate normal random variable with variance 0!

☛ A general definition of the normal random vector $X \sim N_p(\mu, \Sigma)$ can be achieved via its characteristic function, which has the form

$$\varphi_X(t) = \exp\left(it^T\mu - \frac{1}{2}t^T\Sigma t\right) \text{ for } t \in \mathbb{R}^p.$$

☛ Σ is not required to be strictly positive definite, i.e. A probability density function on $\mathbb{R}^p$ may not exist.

☛ We can see that the value of the density of a $N_p(\mu, \Sigma)$ is constant when

$$(x-\mu)^T\Sigma^{-1}(x-\mu)$$

is constant.

☛ Now for positive constant c,

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = c$$

corresponds to an ellipsoid.

☛ The quantity

$$\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

is called the Mahalanobis distance between x and μ.

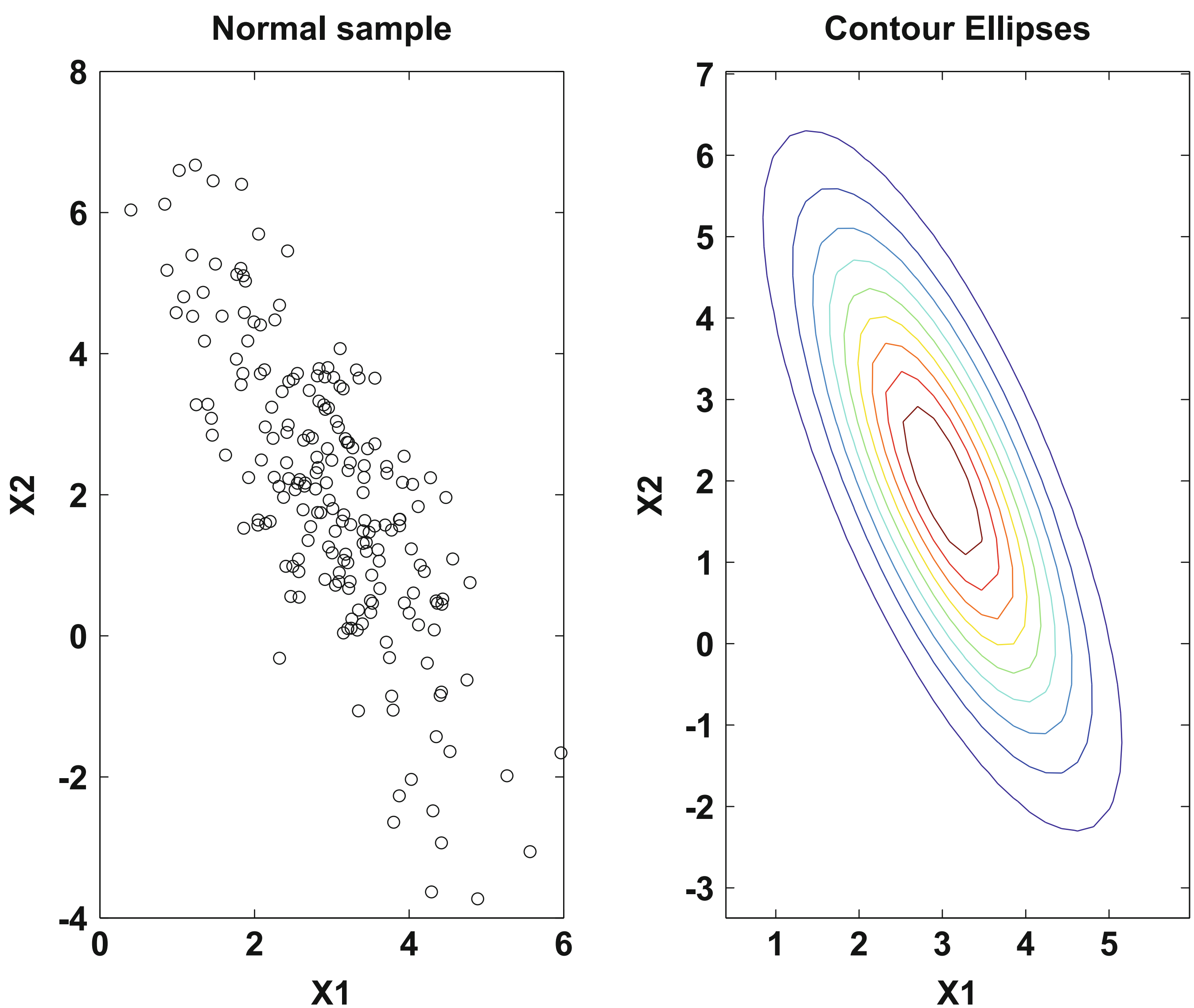


Fig. 4.3 Scatterplot of a normal sample and contour ellipses for $\mu = \binom{3}{2}$ and $\Sigma = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix}$

☛ If $X \sim N_p(\mu, \Sigma)$ and A and B are matrices with p columns, then

$$AX \text{ and } BX \text{ are independent} \iff A\Sigma B^T = 0\,. \tag{2}$$

☛ If $X_1, \ldots, X_n$ are i.i.d.$\sim N_p(\mu, \Sigma)$, then

$$\bar{X} \sim N_p(\mu, \Sigma/n) \tag{3}$$

☛ If $Z_1, \ldots, Z_n$ are independent $N(0, 1)$ then

$$X = \sum_{k=1}^{n} Z_k^2 \sim \chi_n^2$$

is said to be a chi-square random variable with n degree of freedom.

☛ If $X \sim N_p(\mu, \Sigma)$ and Σ is invertible, then

$$Y = (X-\mu)^T \Sigma^{-1} (X-\mu) \sim \chi^2_p \,. \tag{4}$$

Proof:

1. First write $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$ with the spectral decomposition, i.e. if for $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$,
$$\Sigma = \Gamma\Lambda\Gamma^T$$
is the spectral decomposition of Σ, then
$$\Sigma^{1/2} = \Gamma\Lambda^{1/2}\Gamma^T,$$
where $\Sigma^{1/2} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_p^{1/2})$.

2. One can then define $\Sigma^{-1/2} = \Gamma\Lambda^{-1/2}\Gamma^T$, where
$$\Lambda^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \ldots, \lambda_p^{-1/2}).$$

3. But $\Sigma^{-1/2}(X-\mu) \equiv Z \sim N_p(0, I_p)$, since one can easily check that
$$cov(Z) = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \Gamma\Lambda^{-1/2}\Gamma^T\Gamma\Lambda\Gamma^T\Gamma\Lambda^{-1/2}\Gamma^T = I_p$$

4. So $Y = Z^T Z \sim \chi^2_p$, by the definition of a chi-square distribution.

☛ The Wishart distribution, denoted $W_p(\Sigma, n)$, is a generalisation to multiple dimensions of the chi-square distribution.

☛ Definition: If M is an $p \times n$ matrix whose columns are independent and all have a $N_p(0, \Sigma)$ distribution, $\mathcal{M} = MM^T$ is a Wishart-distributed matrix with parameters p, Σ and n. We write

$$\mathcal{M} \sim W_p(\Sigma, n).$$

☛ Depends on three parameters: p, the scale matrix Σ ($p \times p$) and the number of degrees of freedom n.

☛ If we write $M = [m_1| \dots |m_n]$, where each m_i is a p-dimensional column vector, MM^T can be equivalently expressed as

$$MM^T = \sum_{i=1}^{n} m_i m_i^T.$$

☛ *Remark*: Σ doesn't have to be strictly positive definite, nor is there any restriction on the relative sizes of n and p.

☛ Generalizing the chi-square distribution: If σ is a scalar, a $W_1(\sigma^2, n)$ 1-by-1 random matrix is distributed precisely the same as the scalar σ^2 times a χ^2_n random variable.

☛ A Wishart-distributed $\mathcal{M}$ must be non-negative definite: By definition, $\mathcal{M}$ can be represented as MM^T for some M with independent normal columns. Hence it must be that

$$x^T \mathcal{M} x = x^T M M^T x \geq 0$$

for any p-vectors x.

☛ A Wishart distribution can also be defined by its characteristic function, which has the form

$$\varphi_{\mathcal{M}}(T) = |I_p - 2iT\Sigma|^{-n/2}.$$

Here, T is a matrix with the same dimensions as Σ and acts just like the usual "t" in the definition of a characteristic function.

☛ If $\mathcal{M}$ is non-singular (i.e. positive definite) with probability 1, it is said to have a non-singular Wishart distribution.

☛ Proposition 8.2 of the IMS lecture notes by Morris Eaton states exactly when $\mathcal{M}$ is non-singular:

Suppose $\mathcal{M}$ is Wishart-distributed with parameters Σ, p, n. Then $\mathcal{M}$ has a non-singular Wishart distribution if and only if $n \geq p$ and $\Sigma > 0$, in which case $\mathcal{M}$ has the density

$$f_{\Sigma,n}(\mathcal{M}) = \frac{|\mathcal{M}|^{\frac{n-p-1}{2}} \exp(-\frac{1}{2}\mathbf{tr}\left(\mathcal{M}\Sigma^{-1}\right))}{2^{pn/2}\pi^{p(p-1)/4}|\Sigma|^{n/2}\prod_{i=1}^{p}\Gamma((n+1-i)/2)}$$

☛ If a $p \times p$ random matrix $\mathcal{Y} \sim W_p(\Sigma, n)$ and B is a $q \times p$ matrix then

$$B\mathcal{Y}B^T \sim W_q(B\Sigma B^T, n)\,.$$

☛ If a $p \times p$ random matrix $\mathcal{Y} \sim W_p(\Sigma, n)$ and a is a $p \times 1$ vector such that $a^T\Sigma a \neq 0$, then

$$a^T\mathcal{Y}a/a^T\Sigma a \sim \chi^2_n\,.$$

☛ If $\mathcal{Y} \sim W_p(\Sigma, n)$, then $\mathbb{E}[\mathcal{Y}] = n\Sigma$

☛ If $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ are independently and correspondingly distributed as

$$W_p(\Sigma, n_1), \ldots, W_p(\Sigma, n_k),$$

then $\sum_{i=1}^k \mathcal{Y}_i \sim W_p(\Sigma, \sum_{i=1}^k n_i)$.

☛ Recall the unbiased sample covariance matrix

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T .$$

It can be proved that

$$(n-1)S \sim W_p(\Sigma, n-1).$$

☛ Essentially, it says that if $X_1, \ldots, X_n$ be iid $N(\mu, \Sigma)$ random vectors with sample mean $\bar{X}$, then

$$\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

is distributed as $\sum_{i=1}^{n-1} Z_i Z_i^T$, where the Z_i's are iid $N(0, \Sigma)$ random vectors.

☛ See Theorem 3.3.2 in *An Introduction to Multivariate Statistical Analysis* by Anderson for the proof. (to be covered later if time allows)

☛ The Hotelling's T^2 distribution, denoted $T^2_{p,n}$, is a generalisation to multiple dimensions of the student t_n distribution with n degrees of freedom.

☛ Definition: If $X \sim N_p(0, I_p)$ is independent of $\mathcal{M} \sim W_p(I_p, n)$, then

$$nX^T \mathcal{M}^{-1} X \sim T^2_{p,n}.$$

☛ Depends on two parameters: p and the number of degrees of freedom n.

☛ Recall that a t-distributed variable $X \sim t_n$ of degree n is defined by

$$X = \frac{Y}{\sqrt{Z/n}},$$

where Y and Z are independent random variables, $Y \sim N(0,1)$ and $Z \sim \chi^2_n$; hence, X^2 is distributed as a $T^2_{1,n}$ distribution when $p = 1$.

☛ Hotelling's T^2 and the F-distribution is related by

$$T^2_{p,n} = \frac{np}{n-p+1} F_{p,n-p+1}.$$

(Recall that the square of a univariate t-distribution with n degree of freedom is same as an $F_{1,n}$ distribution)

☛ If $X \sim N_p(\mu, \Sigma)$ is independent of $\mathcal{M} \sim W_p(\Sigma, n)$ with $\mathcal{M}$ being non-singular, then

$$n(X-\mu)^T\mathcal{M}^{-1}(X-\mu) \sim T^2_{p,n}.$$

(Theorem 5.8 in Härdle and Simar)

Proof:

1. Let $\Sigma^{1/2}$ be the square root of Σ obtained by the spectral decomposition. One can write

   $$X-\mu = \Sigma^{1/2}Y$$

   and

   $$\mathcal{M} = \Sigma^{1/2}\mathcal{L}\Sigma^{1/2},$$

   where Y and $\mathcal{L}$ are independent with $Y \sim N_p(0, I_p)$ and $\mathcal{L} \sim W_p(I_p, n)$.

2. Then

$$\begin{aligned} n(X-\mu)^T\mathcal{M}^{-1}(X-\mu) &= nY^T\Sigma^{1/2}\Sigma^{-1/2}\mathcal{L}^{-1}\Sigma^{-1/2}\Sigma^{1/2}Y \\ &= nY^T\mathcal{L}^{-1}Y \sim T^2_{p,n} \end{aligned}$$

- If $X_1, \ldots, X_n$ are i.i.d.$\sim N_p(\mu, \Sigma)$, then the sample mean vector $\bar{X}$ and the unbiased sample covariance matrix S are such that

$$n(\bar{X} - \mu)^T S^{-1}(\bar{X} - \mu) \sim T^2_{p,n-1}.$$

Proof sketch:

1. By Cochran's theorem (Theorem 5.7 in the Härdle and Simar), S is independent of $\bar{X}$.
2. From a previous result we know that

$$\mathcal{M} \equiv \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \sim W_p(\Sigma, n-1).$$

3. Moreover, we can write

$$(\bar{X} - \mu) = \frac{1}{\sqrt{n}} Y,$$

where $Y \sim N(0, \Sigma)$ is independent of $\mathcal{M}$.

4. Putting these together we have

$$\begin{aligned} n(\bar{X} - \mu)^T S^{-1}(\bar{X} - \mu) &= n \left(\frac{1}{\sqrt{n}} \right)^2 (n-1) Y^T \mathcal{M}^{-1} Y \\ &= (n-1) Y^T \mathcal{M}^{-1} Y \sim T^2_{p,n-1} \end{aligned}$$

☛ Hotelling's T^2 statistic is typically used for the following hypothesis testing problem:

☛ Suppose $X_1, \ldots, X_n$ is an iid random sample from the $N_p(\mu, \Sigma)$ population with $\Sigma > 0$ *unknown*. Test

$$H_0 : \mu = \mu_0 \quad \text{VS} \quad H_1 : \text{no constraints}.$$

☛ This is the multivariate version of the univariate testing problem tackled by t statistics.

☛ When H_0 is true,

$$n(\bar{X} - \mu_0)^T S^{-1} (\bar{X} - \mu_0) \sim T^2_{p,n-1}.$$

☛ Naturally, we can use $n(\bar{X} - \mu_0)^T S^{-1} (\bar{X} - \mu_0)$ as the test statistic, and calibrate the cutoff threshold using the null $T^2_{p,n-1}$ distribution (or equivalently, the $F_{p,n-p}$ distribution).

☛ By defining the constrained and unconstrained parameter spaces

$$\Omega_0 = \{(\mu, \Sigma) : \mu = \mu_0, \Sigma > 0\} \quad \text{and} \quad \Omega_1 = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma > 0\},$$

the testing problem can be written as

$$H_0 : (\mu, \Sigma) \in \Omega_0 \quad \text{VS} \quad H_1 : (\mu, \Sigma) \in \Omega_1.$$

☛ Alternatively, consider the *likelihood* for this problem, which is the joint density of the data

$$L(\mathcal{X}; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

thought of as a function in the parameters $\theta \equiv (\mu, \Sigma)$.

☛The likelihood ratio (LR) statistic is

$$\lambda = \frac{\max_{\theta \in \Omega_0} L(\mathcal{X}; \theta)}{\max_{\theta \in \Omega_1} L(\mathcal{X}; \theta)};$$

one tends to favor H_1 if λ is low.

- Hence, if the distribution of λ under H_0 is known (the null distribution), one can implement a likelihood ratio test by rejecting H_0 if λ is lower than a threshold calibrated based on this null distribution.

- However, the exact null distribution of λ may be hard to derive.

- Turns out, it can be shown (Härdle and Simar, section 7.1) that

$$-2\log(\lambda) = n\log\left(1 + \frac{n(\bar{X}-\mu_0)^T S^{-1}(\bar{X}-\mu_0)}{n-1}\right)$$

- Hence, $-2\log(\lambda)$ is really one-to-one and increasing function in the Hotelling's T^2 statistic $n(\bar{X}-\mu_0)^T S^{-1}(\bar{X}-\mu_0)$.

- Rejecting for small values of λ is same as rejecting for large values of $n(\bar{X}-\mu_0)^T S^{-1}(\bar{X}-\mu_0)$;

Hotelling's T^2 test is equivalent to the LR test!