

INTRODUCTION TO LINEAR MODELS AND REGRESSION

Moutushi Chatterjee, notes by Ramdas Singh

Fourth Semester

Contents

1	RANDOM VECTORS	1
1.1	Definitions and Moments	1
1.1.1	Moment Generating Function	2
1.2	Linear Models	3
1.3	Multivariate Normal	3
1.4	Dirichlet Distribution	6
	Index	9

Chapter 1

RANDOM VECTORS

1.1 Definitions and Moments

January 8th.

Recall that for a random variable X , the r -th *raw moment* is defined as

$$\mu'_r = \mathbb{E}[X^r] \quad (1.1)$$

provided the expectation exists. Here, $\mu'_1 = \mathbb{E}[X] = \mu$ is the mean. The r -th *central moment* is defined as

$$\mu_r = \mathbb{E}[(X - \mu)^r]. \quad (1.2)$$

Similarly, for a set of sample points x_1, x_2, \dots, x_n , the r -th *sample raw moment* is defined as

$$m'_r = \frac{1}{n} \sum_{i=1}^n x_i^r \quad (1.3)$$

and the r -th *sample central moment* is defined as

$$m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r \quad (1.4)$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean. Note that $\mu_2 = \text{Var}(X)$ is the variance of X and $\frac{\mu'_3}{\mu_2}$ is the *skewness* of X . There is also *coefficient of variation* defined as $\frac{\sigma}{\mu}$ where $\sigma = \sqrt{\text{Var}(X)}$ is the standard deviation of X . For a sample, it is defined as $\frac{s}{\bar{x}}$ where $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ is the sample standard deviation. The *correlation coefficient* between x_i 's and y_i 's is defined as

$$\frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2 \cdot \frac{1}{n} \sum (y_i - \bar{y})^2}} \quad (1.5)$$

It is also known as the product-moment correlation coefficient.

These definitions can be extended to random vectors as follows. For a random vector $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$, the *mean vector* is defined as

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_p] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}. \quad (1.6)$$

The *dispersion matrix*, or covariance matrix, is defined as

$$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var}(X_p) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}. \quad (1.7)$$

Similar to the correlation coefficient for random variables, we define the *correlation matrix* for a random vector \mathbf{X} as

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} \quad (1.8)$$

Define D_σ as the diagonal matrix $\text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p)$. Then, we have the relation

$$\mathbf{R} = D_\sigma^{-1} \mathbf{\Sigma} D_\sigma^{-1}. \quad (1.9)$$

If $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$ is a random vector where \mathbf{X} is of dimension p and \mathbf{Y} is of dimension q , then we simply have

$$\mathbb{E}[\mathbf{Z}] = \begin{pmatrix} \mathbb{E}[\mathbf{X}] \\ \mathbb{E}[\mathbf{Y}] \end{pmatrix}, \quad \text{Cov}(\mathbf{Z}) = \begin{pmatrix} \mathbf{\Sigma}_\mathbf{X} & \mathbf{\Sigma}_{\mathbf{X}\mathbf{Y}} \\ \mathbf{\Sigma}_{\mathbf{Y}\mathbf{X}}^\top & \mathbf{\Sigma}_\mathbf{Y} \end{pmatrix}. \quad (1.10)$$

As for linear transformations, if $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ where \mathbf{A} is a $q \times p$ matrix and $\mathbf{b} \in \mathbb{R}^q$, then

$$\mathbb{E}[\mathbf{Y}] = \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{b}, \quad \text{Cov}(\mathbf{Y}) = \mathbf{A} \text{Cov}(\mathbf{X}) \mathbf{A}^\top. \quad (1.11)$$

In particular, if $\mathbf{a} = \mathbf{A}^\top$ is a p -dimensional vector, then Y is a univariate random variable written as $Y = \mathbf{a}^\top \mathbf{X} + b$.

Mahalanobis Distance

All of the below equations

$$d^2 = (\mathbf{y}_1 - \mathbf{y}_2)^\top \mathbf{S}^{-1} (\mathbf{y}_1 - \mathbf{y}_2), \quad (1.12)$$

$$D^2 = (\bar{\mathbf{y}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}), \quad (1.13)$$

$$D^2 = (\bar{\mathbf{y}} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}), \quad (1.14)$$

are known as the *Mahalanobis distance*. Here, \mathbf{S} is the sample covariance matrix and $\mathbf{\Sigma}$ is the population covariance matrix. The Mahalanobis distance is a measure of distance that accounts for the correlations between variables and the different scales of the variables. Since $\mathbf{\Sigma}$ is positive semi-definite, we set $\mathbf{z} = \mathbf{\Sigma}^{-1/2} \mathbf{y} - \mathbf{\Sigma}^{-1/2} \boldsymbol{\mu}$ to get

$$(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1/2} \mathbf{\Sigma}^{-1/2} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{z}^\top \mathbf{z}. \quad (1.15)$$

This shows that the Mahalanobis distance is equivalent to the Euclidean distance in the transformed space where the data has been standardized by the covariance matrix. The covariance matrix of \mathbf{z} is

$$\text{Cov}(\mathbf{z}) = \mathbf{\Sigma}^{-1/2} \text{Cov}(\mathbf{y}) \mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1/2} \frac{1}{n} \mathbf{\Sigma} \mathbf{\Sigma}^{-1/2} = \frac{1}{n} \mathbf{I}_p. \quad (1.16)$$

Thus, the variables in \mathbf{z} are uncorrelated with equal variances; in other words, they have been standardized.

1.1.1 Moment Generating Function

Recall that in the univariate case, the moment generating function (MGF) of a random variable X is defined as

$$M_X(t) := \mathbb{E}[e^{tX}] = \mathbb{E} \left[\sum_{r=0}^{\infty} \frac{t^r X^r}{r!} \right]. \quad (1.17)$$

We extend this definition to the multivariate case as follows; for a random vector $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$, the *moment generating function* is defined as

$$M_{\mathbf{X}}(\mathbf{t}) := \mathbb{E}[e^{\mathbf{t}^\top \mathbf{X}}] = \mathbb{E}\left[e^{\sum_{i=1}^p t_i X_i}\right] \quad (1.18)$$

for all $\mathbf{t} = (t_1, t_2, \dots, t_p)^T \in \mathbb{R}^p$ such that the expectation exists. When the X_i 's are independent, we have

$$M_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^p M_{X_i}(t_i). \quad (1.19)$$

Similar to the univariate case, the MGF can be used to compute moments of the random vector. For example, the first moment (mean vector) can be obtained by taking the gradient of the MGF at $\mathbf{t} = \mathbf{0}$:

$$\mathbb{E}[\mathbf{X}] = \nabla_{\mathbf{t}} M_{\mathbf{X}}(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}}. \quad (1.20)$$

The second moment (covariance matrix) can be obtained by taking the Hessian of the MGF at $\mathbf{t} = \mathbf{0}$:

$$\text{Cov}(\mathbf{X}) = \nabla_{\mathbf{t}}^2 M_{\mathbf{X}}(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}} - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top. \quad (1.21)$$

Suppose $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ where \mathbf{A} is a $q \times p$ matrix and $\mathbf{b} \in \mathbb{R}^q$. Then, the MGF of \mathbf{Y} is given by

$$M_{\mathbf{Y}}(\mathbf{t}) = e^{\mathbf{t}^\top \mathbf{b}} M_{\mathbf{X}}(\mathbf{A}^\top \mathbf{t}). \quad (1.22)$$

1.2 Linear Models

Generally, we had

$$y = \beta_0 + \beta_1 x + \epsilon \quad (1.23)$$

where $\epsilon \sim N(0, \sigma^2)$. Here, we note that y is a random variable while x is a fixed variable. Moreover, $\sum_i (y_i - \bar{y}) = 0$ implies that $\sum_i \epsilon_i = 0$. In such a case, $\mathbb{E}[y] = \beta_0 + \beta_1 \mathbb{E}[x]$ and $\mathbb{E}[y | x] = \beta_0 + \beta_1 x$. A theorem states that if any one of the regression lines is constant, then $\text{Cov}(X, Y) = 0$.

A simple proof is given as follows: note that $\mathbb{E}[Y | X] = k \implies \mathbb{E}[Y] = k$. Also,

$$\mathbb{E}[XY] = \int_{\mathbb{R}} \int_{\mathbb{R}} xy f(x, y) dx dy = \int_{\mathbb{R}} x \int_{\mathbb{R}} y f_{Y|X}(y | x) dy f_X(x) dx = \int_{\mathbb{R}} x k f_X(x) dx = k \mathbb{E}[X]. \quad (1.24)$$

Thus,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = k \mathbb{E}[X] - \mathbb{E}[X]k = 0. \quad (1.25)$$

1.3 Multivariate Normal

January 29th.

Recall in the case of univariate normal distribution, we had $X \sim N(\mu, \sigma^2)$ with probability density function (PDF)

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), & x \in \mathbb{R}, \\ 0, & \text{otherwise.} \end{cases} \quad (1.26)$$

We extend this to the multivariate case as follows. A random vector $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$ is said to have a *multivariate normal distribution* with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma}$ (denoted as $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) if its probability density function (PDF) is given by

$$f(\mathbf{x}) = \begin{cases} \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), & \mathbf{x} \in \mathbb{R}^p, \\ 0, & \text{otherwise.} \end{cases} \quad (1.27)$$

where Σ is positive definite. The moment generating function (MGF) of \mathbf{X} is given by

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}^\top \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^\top \Sigma \mathbf{t}\right). \quad (1.28)$$

For further results, we make use of the following lemmas.

Lemma 1.1. *If Σ is a positive definite matrix, then there exists a non-singular matrix \mathbf{B} such that $\Sigma = \mathbf{B}^\top \mathbf{B}$.*

Lemma 1.2. *If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ and \mathbf{B} is as above ($\Sigma = \mathbf{B}^\top \mathbf{B}$) and satisfies $\mathbf{B}^\top \mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$, then $\mathbf{Y} \sim N_p(\mathbf{0}, \mathbf{I}_p)$.*

Proof. The Jacobian of the transformation is given by $|\mathbf{B}| = \sqrt{|\Sigma|}$. Also, $\mathbf{B}\Sigma^{-1}\mathbf{B}^\top = \mathbf{I}_p$. Thus,

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Y}^\top \mathbf{Y}. \quad (1.29)$$

Therefore, the PDF of \mathbf{Y} is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{B}^\top \mathbf{y} + \boldsymbol{\mu}) |\mathbf{B}| \quad (1.30)$$

$$= \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{y}^\top \mathbf{y}\right) |\mathbf{B}| \quad (1.31)$$

$$= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \mathbf{y}^\top \mathbf{y}\right). \quad (1.32)$$

This shows that $\mathbf{Y} \sim N_p(\mathbf{0}, \mathbf{I}_p)$. ■

Lemma 1.3. *If $X \sim (0, 1)$, then $M_X(t) = \exp\left(\frac{t^2}{2}\right)$.*

We can now prove the expression for the MGF of the multivariate normal distribution. We have

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}[e^{\mathbf{t}^\top \mathbf{X}}] = \mathbb{E}\left[e^{\mathbf{t}^\top (\mathbf{B}^\top \mathbf{Y} + \boldsymbol{\mu})}\right] = e^{\mathbf{t}^\top \boldsymbol{\mu}} \mathbb{E}\left[e^{(\mathbf{B}\mathbf{t})^\top \mathbf{Y}}\right] \\ &= e^{\mathbf{t}^\top \boldsymbol{\mu}} \prod_{i=1}^p \mathbb{E}\left[e^{(\mathbf{b}_i^\top \mathbf{t}) Y_i}\right] = e^{\mathbf{t}^\top \boldsymbol{\mu}} \prod_{i=1}^p \exp\left(\frac{(\mathbf{b}_i^\top \mathbf{t})^2}{2}\right) \\ &= e^{\mathbf{t}^\top \boldsymbol{\mu}} \exp\left(\frac{1}{2} \sum_{i=1}^p (\mathbf{b}_i^\top \mathbf{t})^2\right) = e^{\mathbf{t}^\top \boldsymbol{\mu}} \exp\left(\frac{1}{2} \mathbf{t}^\top \left(\sum_{i=1}^p \mathbf{b}_i \mathbf{b}_i^\top\right) \mathbf{t}\right) \\ &= e^{\mathbf{t}^\top \boldsymbol{\mu}} \exp\left(\frac{1}{2} \mathbf{t}^\top \Sigma \mathbf{t}\right). \end{aligned} \quad (1.33)$$

Now suppose $z = \mathbf{a}^\top \mathbf{X}$ where $\mathbf{a} \in \mathbb{R}^p$. Then, the MGF of z is given by

$$M_z(t) = \mathbb{E}[e^{tz}] = \mathbb{E}[e^{t\mathbf{a}^\top \mathbf{X}}] = M_{\mathbf{X}}(t\mathbf{a}) = \exp\left(t\mathbf{a}^\top \boldsymbol{\mu} + \frac{t^2}{2} \mathbf{a}^\top \Sigma \mathbf{a}\right). \quad (1.34)$$

Thus, $z \sim N(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \Sigma \mathbf{a})$. Similarly, for $\mathbf{Z} = \mathbf{A}\mathbf{X}$ where \mathbf{A} is a $q \times p$ matrix, we have

$$M_{\mathbf{Z}}(\mathbf{t}) = \mathbb{E}[e^{\mathbf{t}^\top \mathbf{Z}}] = \mathbb{E}[e^{\mathbf{t}^\top \mathbf{A}\mathbf{X}}] = M_{\mathbf{X}}(\mathbf{A}^\top \mathbf{t}) = \exp\left(\mathbf{t}^\top \mathbf{A} \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^\top \mathbf{A} \Sigma \mathbf{A}^\top \mathbf{t}\right). \quad (1.35)$$

So, $\mathbf{Z} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\Sigma\mathbf{A}^\top)$.

Theorem 1.4. *$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ if and only if every linear combination of the components of \mathbf{X} follows univariate normal distribution with suitable parameters.*

Proof. The forward direction is already shown above. For the converse, suppose $Y = \mathbf{L}^\top \mathbf{X} \sim N(\mathbf{L}^\top \boldsymbol{\mu}, \mathbf{L}^\top \boldsymbol{\Sigma} \mathbf{L})$ for any $\mathbf{L} \in \mathbb{R}^p$. Note that the MGF of Y is

$$M_Y(t) = \exp \left(t \mathbf{L}^\top \boldsymbol{\mu} + \frac{t^2}{2} \mathbf{L}^\top \boldsymbol{\Sigma} \mathbf{L} \right). \quad (1.36)$$

Plugging in $t = 1$, we get

$$M_Y(1) = \exp \left(\mathbf{L}^\top \boldsymbol{\mu} + \frac{1}{2} \mathbf{L}^\top \boldsymbol{\Sigma} \mathbf{L} \right). \quad (1.37)$$

Alternatively, we have $M_Y(1) = \mathbb{E}[e^{\mathbf{L}^\top \mathbf{X}}] = M_{\mathbf{X}}(\mathbf{L})$. Thus, by the uniqueness property of MGFs, we must have $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. ■

Theorem 1.5. Suppose $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where we partition \mathbf{X} as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \quad (1.38)$$

where \mathbf{X}_1 is of dimension r and \mathbf{X}_2 is of dimension $s = p - r$. Correspondingly, we partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \quad (1.39)$$

Then \mathbf{X}_1 and \mathbf{X}_2 are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ (or equivalently, $\boldsymbol{\Sigma}_{21} = \mathbf{0}$).

Proof. The converse direction is straightforward: if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}| |\boldsymbol{\Sigma}_{22}|$ and the PDF of \mathbf{X} can be factored into the product of the PDFs of \mathbf{X}_1 and \mathbf{X}_2 , implying independence. The forward direction is shown as follows. Suppose \mathbf{X}_1 and \mathbf{X}_2 are independent. Then, the MGF of \mathbf{X} can be written as

$$M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}_1}(\mathbf{t}_1) M_{\mathbf{X}_2}(\mathbf{t}_2) \quad (1.40)$$

where $\mathbf{t} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix}$. Plugging in the expressions for the MGFs, we have

$$\begin{aligned} & \exp \left(\mathbf{t}^\top \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t} \right) \\ &= \exp \left(\mathbf{t}_1^\top \boldsymbol{\mu}_1 + \frac{1}{2} \mathbf{t}_1^\top \boldsymbol{\Sigma}_{11} \mathbf{t}_1 \right) \exp \left(\mathbf{t}_2^\top \boldsymbol{\mu}_2 + \frac{1}{2} \mathbf{t}_2^\top \boldsymbol{\Sigma}_{22} \mathbf{t}_2 \right). \end{aligned} \quad (1.41)$$

Equating the exponents, we get

$$\mathbf{t}_1^\top \boldsymbol{\mu}_1 + \mathbf{t}_2^\top \boldsymbol{\mu}_2 + \frac{1}{2} \begin{pmatrix} \mathbf{t}_1^\top & \mathbf{t}_2^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix} = \mathbf{t}_1^\top \boldsymbol{\mu}_1 + \frac{1}{2} \mathbf{t}_1^\top \boldsymbol{\Sigma}_{11} \mathbf{t}_1 + \mathbf{t}_2^\top \boldsymbol{\mu}_2 + \frac{1}{2} \mathbf{t}_2^\top \boldsymbol{\Sigma}_{22} \mathbf{t}_2. \quad (1.42)$$

Simplifying, we obtain

$$\mathbf{t}_1^\top \boldsymbol{\Sigma}_{12} \mathbf{t}_2 + \mathbf{t}_2^\top \boldsymbol{\Sigma}_{21} \mathbf{t}_1 = 0. \quad (1.43)$$

Since this holds for all \mathbf{t}_1 and \mathbf{t}_2 , we must have $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{21} = \mathbf{0}$. ■

February 5th.

For a multivariate normal vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, partition it as above. Let us make the transformation $\mathbf{X} \rightarrow \mathbf{Y}$ prescribed as $\mathbf{Y}_{r \times 1}^{(1)} = \mathbf{X}_{r \times 1}^{(1)}$ and $\mathbf{Y}_{(p-r) \times 1}^{(2)} = \mathbf{X}_{(p-r) \times 1}^{(2)} + \mathbf{B} \mathbf{X}_{r \times 1}^{(1)}$, where \mathbf{B} is an appropriate matrix of constant values. This is simply the transformation given by

$$\mathbf{Y} = \mathbf{M} \mathbf{X}, \quad \mathbf{M} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{B} & \mathbf{I}_{p-r} \end{pmatrix}. \quad (1.44)$$

Then, $\mathbf{Y} \sim N_p(\mathbf{M}\boldsymbol{\mu}, \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^\top)$. We can choose \mathbf{B} such that $\boldsymbol{\Sigma}_{21} + \mathbf{B}\boldsymbol{\Sigma}_{11} = \mathbf{0}$, which gives $\mathbf{B} = -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$. With this choice of \mathbf{B} , we have

$$\text{Cov}(\mathbf{Y}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{pmatrix}. \quad (1.45)$$

Thus, $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are independent. Note that $\mathbf{Y}^{(2)} = \mathbf{X}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}$. If we define $\boldsymbol{\Sigma}_{22\cdot 1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, then we have

$$\mathbf{Y}^{(1)} \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \quad \mathbf{Y}^{(2)} \sim N_{p-r}(\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{22\cdot 1}). \quad (1.46)$$

To find the distribution of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)}$, we can use the fact that $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are independent. We have

$$\begin{aligned} f_{\mathbf{X}^{(1)}|\mathbf{X}^{(2)}}(\mathbf{x}^{(1)} | \mathbf{x}^{(2)}) &= f_{\mathbf{Y}^{(1)}|\mathbf{Y}^{(2)}}(\mathbf{y}^{(1)} | \mathbf{y}^{(2)}) = f_{\mathbf{Y}^{(1)}}(\mathbf{y}^{(1)}) \\ &= \frac{1}{(2\pi)^{r/2}|\boldsymbol{\Sigma}_{11}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}^{(1)} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}^{(1)} - \boldsymbol{\mu}_1)\right) \end{aligned} \quad (1.47)$$

which can be rewritten as

$$\frac{1}{(2\pi)^{r/2}|\boldsymbol{\Sigma}_{11}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(1)} - (\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}_2)))^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}^{(1)} - (\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}_2)))\right). \quad (1.48)$$

Thus, $\mathbf{X}^{(1)} | \mathbf{X}^{(2)} = \mathbf{x}^{(2)} \sim N_r(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11})$. For the distribution of $\mathbf{X}^{(2)}$ given $\mathbf{X}^{(1)}$, we have

$$\begin{aligned} f_{\mathbf{X}^{(2)}|\mathbf{X}^{(1)}}(\mathbf{x}^{(2)} | \mathbf{x}^{(1)}) &= f_{\mathbf{Y}^{(2)}|\mathbf{Y}^{(1)}}(\mathbf{y}^{(2)} | \mathbf{y}^{(1)}) = f_{\mathbf{Y}^{(2)}}(\mathbf{y}^{(2)}) \\ &= \frac{1}{(2\pi)^{(p-r)/2}|\boldsymbol{\Sigma}_{22\cdot 1}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}^{(2)} - (\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1))^\top (\boldsymbol{\Sigma}_{22\cdot 1})^{-1}(\mathbf{y}^{(2)} - (\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1))\right) \end{aligned} \quad (1.49)$$

which can be rewritten as

$$\frac{1}{(2\pi)^{(p-r)/2}|\boldsymbol{\Sigma}_{22\cdot 1}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(2)} - (\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}^{(1)} - \boldsymbol{\mu}_1)))^\top (\boldsymbol{\Sigma}_{22\cdot 1})^{-1}(\mathbf{x}^{(2)} - (\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}^{(1)} - \boldsymbol{\mu}_1)))\right). \quad (1.50)$$

Thus, $\mathbf{X}^{(2)} | \mathbf{X}^{(1)} = \mathbf{x}^{(1)} \sim N_{p-r}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}^{(1)} - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22\cdot 1})$.

1.4 Dirichlet Distribution

Just like the multivariate normal distribution is a generalization of the univariate normal distribution, the Dirichlet distribution is a generalization of the univariate beta distribution. Recall that the PDF of the gamma distribution with parameters α and λ is given by

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0, \\ 0, & \text{otherwise} \end{cases} \quad (1.51)$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ is the gamma function, and the PDF of the beta distribution with parameters α and β is given by

$$f(x) = \begin{cases} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, & x \in (0, 1), \\ 0, & \text{otherwise.} \end{cases} \quad (1.52)$$

Also recall that if X_1 and X_2 are independent random variables such that $X_1 \sim \Gamma(\alpha, 1)$ and $X_2 \sim \Gamma(\beta, 1)$, then $X_1 + X_2 \sim \Gamma(\alpha + \beta, 1)$ and $\frac{X_1}{X_1 + X_2} \sim \text{Beta}(\alpha, \beta)$. We can extend this to the multivariate case as follows. Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)$ be such that $Y_i > 0$ for all i and $\sum_{i=1}^k Y_i = 1$. Also let $\alpha_1, \alpha_2, \dots, \alpha_k$

be positive parameters. Then, \mathbf{Y} is said to have a *Dirichlet distribution* with parameters $\alpha_1, \alpha_2, \dots, \alpha_k$ (denoted as $\mathbf{Y} \sim \text{Dir}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)$) if its PDF is given by

$$f(y) = \begin{cases} \frac{\Gamma(\alpha_1 + \alpha_2 + \dots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_k)} y_1^{\alpha_1-1} y_2^{\alpha_2-1} \dots y_k^{\alpha_k-1}, & y_i > 0 \text{ for all } i, \sum_{i=1}^k y_i = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (1.53)$$

If we let X_1, X_2, \dots, X_k be independent random variables such that $X_i \sim \Gamma(\alpha_i, 1)$ for all i , then the joint distribution of X_1, X_2, \dots, X_k is given by

$$f(x_1, x_2, \dots, x_k) = \prod_{i=1}^k \frac{1}{\Gamma(\alpha_i)} x_i^{\alpha_i-1} e^{-x_i}, \quad x_i > 0 \text{ for all } i. \quad (1.54)$$

If we then let $Y_i = \frac{X_i}{\sum_{j=1}^k X_j}$ for all i , then $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k) \sim \text{Dir}(\boldsymbol{\alpha})$. The moments of the Dirichlet distribution are given by

$$\mathbb{E}[Y_i] = \frac{\alpha_i}{\alpha_0}, \quad (1.55)$$

$$\text{Var}(Y_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \quad (1.56)$$

$$\text{Cov}(Y_i, Y_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \quad (1.57)$$

where $\alpha_0 = \sum_{i=1}^k \alpha_i$.

Index

central moment, 1
coefficient of variation, 1
correlation coefficient, 1
correlation matrix, 2

Dirichlet distribution, 7
dispersion matrix, 2

Mahalanobis distance, 2

mean vector, 1
moment generating function, 3
multivariate normal distribution, 3

raw moment, 1

sample central moment, 1
sample raw moment, 1
skewness, 1