Topic 2: Matrix Algebra, Random Vectors, and Multivariate Normal

Reference:

1. Searle, S. R. (1982) Matrix Algebra useful for statistics. Wiley, ~~Inter-Science.~~
2. Horn, R. A. and Johnson, C. R. (1985) Matrix Analysis. Cambridge, ~~University Press.~~

Matrix: A $n \times p$ matrix X is a rectangular array of numbers, which can be expressed as

$$\mathrm{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & . & x_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ x_{n1} & x_{n2} & . & . & . & x_{np} \end{bmatrix} = (x_{ij}) = \begin{pmatrix} x_1^T \\ x_2^T \\ . \\ . \\ . \\ x_n^T \end{pmatrix} = (x_{(1)}, x_{(2)}, \cdots, x_{(p)}),$$

where $x_i = (x_{i1}, \cdots, x_{ip})^T$ and $x_{(j)} = (x_{1j}, \cdots, x_{nj})^T$, $i = 1, \cdots, n$, $j = 1, \cdots, k$.

Square Matrix: A $n \times p$ matrix X is called a square matrix if $n = p$.

Unit Vector: A $p \times 1$ vector with all numbers being one is called a unit vector and is denoted by 1_p.

Diagonal Matrix: A $p \times p$ diagonal matrix, denoted by D, is a square matrix with off-diagonal elements being zero.

$$D = diag(a_{11}, \cdots, a_{pp}) = \begin{pmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{pp} \end{pmatrix}$$

Identity Matrix: A $p \times p$ identity matrix, denoted by I_p, is a diagonal matrix with diagonal elements being one.

Symmetric Matrix: A square ($p \times p$) matrix **A** is symmetric if $A = A^T$, where "T" denotes the transpose.

Triangular Matrix: Upper triangular matrix $U =$ 

Lower triangular matrix $L =$ 

Trace: Let $A = (a_{ij})$ be a $p \times p$ matrix. The trace of A is defined as $tr(A) = \Sigma_{i=1}^{p} a_{ii}$.

Idempotent Matrix: A square matrix A is called an idempotent matrix if $A^2 = A$, where $A^2 = AA$.

Determinant and Cofactor: The determinant of a square matrix A, denoted by $\det(A)$ or $|A|$ is defined as

$\det(A) = \Sigma_{j=1}^{p} (-1)^{i+j} a_{ij} \det(A(i \mid j))$, where $\det(A(i \mid j))$ is the (i, j) cofactor of A.

Inverse: The inverse of a square matrix A, if exists, A^{-1} is unique and satisfies $A^{-1}A = AA^{-1} = I_p$.

Nonsingular: A is nonsingular iff the inverse of A exits iff $\det(A) \neq 0$.

Orthogonal: A $p \times p$ matrix A is orthogonal iff $AA^T = I_p$, which implies that $A^T = A^{-1}$.

Rank: The rank of a $p \times q$ matrix A, denoted by $rank(A)$, is the maximum number of linearly independent rows (columns) in A.

Eigenvalues and Eigenvectors: Let A be a $p \times p$ matrix and y be a nonzero $p \times 1$ vector such that $Ay = \lambda y$. Then, y is called the eigenvector of A and λ is the corresponding

eigenvalue of A .Here, $P(\lambda) = \det(A - \lambda \mathrm{I}_p)$ is the pth order polynomial in λ, which is called the characteristic polynomial of A .

Example 1. (See Appendix)

Partition Matrices:

Let $A_{m\times n} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ and $B_{n\times k} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$, where A_{ij} and $B_{j\ell}$ are separately $m_i \times n_j$ and $n_j \times k_\ell$ matrices. Then, $AB = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix}$.

Assume

~~Moreover, assume~~ that A is a $p \times p$ nonsingular matrix and $B = A^{-1}$. One has

$B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$ $(\triangleq A_{11\cdot 2}^{-1})$, $B_{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$ $(\triangleq A_{22\cdot 1}^{-1})$

$B_{12} = -A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1} = -A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1}$, and $B_{21} = -A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1} = -A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1}$.

Example 2. (See Appendix)

Spectral Decomposition Theorem (Jordan Decomposition Theorem):

Any symmetric $p \times p$ matrix A can be expressed as $A = \Gamma\Lambda\Gamma^T = \sum_{i=1}^{p} \lambda_i r_{(i)} r_{(i)}^T$, where Λ is a diagonal matrix of eigenvalues of **A** and Γ is an orthogonal matrix whose columns are the corresponding standardized eigenvectors .

Singular Value Decomposition Theorem:

Let A be a $p \times q$ matrix with $rank(A) = r$. Then, A can be expressed as $A = ULV^T$, where $U_{p\times r}$ and $V_{q\times r}$ are orthonormal matrices and L is a diagonal matrix with positive elements such that $A^TA = V\Lambda V^T$, $AA^T = U\Lambda U^T$, and $L = \Lambda^{1/2}$.

Quadratic Forms and Definiteness:

Let A be a $p \times p$ matrix and y be a $p \times 1$ vector. $q = y^TAy$ is called a quadratic form in y and A is the matrix of the quadratic form.

A is positive definite (p.d.) if $y^TAy > 0$ for all $y \neq 0$.

A is positive semidefinite (p.s.d) if $y^TAy \geq 0$ and $y^TAy = 0$ for some $y \neq 0$.

Moments of Random Vectors and Matrices:

$$\text{Let } x=\begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_p \end{pmatrix}, y=\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_q \end{pmatrix}, \text{ and } U=\begin{pmatrix} u_{11} & u_{12} & . & . & . & u_{1q} \\ u_{21} & u_{22} & . & . & . & u_{2q} \\ . & & & & & . \\ . & & & & & . \\ . & & & & & . \\ u_{p1} & u_{p2} & . & . & . & u_{pq} \end{pmatrix}.$$

$$E[x]=\begin{pmatrix} E[x_1] \\ E[x_2] \\ . \\ . \\ . \\ E[x_p] \end{pmatrix}, \; V(x)=\begin{pmatrix} \text{var}[x_1] & \text{cov}(x_1,x_2) & . & . & . & \text{cov}(x_1,x_p) \\ \text{cov}(x_2,x_1) & \text{var}[x_2] & . & . & . & \text{cov}(x_2,x_p) \\ . & & & & & . \\ . & & & & & . \\ . & & & & & . \\ \text{cov}(x_p,x_1) & \text{cov}(x_p,x_2) & . & . & . & \text{var}[x_p] \end{pmatrix},$$

$$E[U]=\begin{pmatrix} E[u_{11}] & E[u_{12}] & . & . & . & E[u_{1q}] \\ E[u_{21}] & E[u_{22}] & . & . & . & E[u_{2q}] \\ . & & & & & . \\ . & & & & & . \\ . & & & & & . \\ E[u_{p1}] & E[u_{p2}] & . & . & . & E[u_{pq}] \end{pmatrix}, Cov(x,y)=E[(x-E[x])(y-E[y])^T].$$

When $x=y$, $Cov(x,y)\triangleq V(x)$.

Property 2.1.

(a) $V(x)=E[xx^T]-E[x](E[x])^T$.

(b) $E[Ax+b]=AE[x]+b$ and $V(Ax+b)=AV(x)A^T$, where A and b are separately a $p\times p$ constant matrix and a $p\times 1$ constant vector.

(c) $Cov(Ax,By)=ACov(x,y)B^T$.

(d) $E[y^TAy]=tr(AV(y))+(E[y])^TA(E[y])$.

Example 3. (See Appendix)

Example 4. (See Appendix)

Example 5. (See Appendix)

Characteristic Function: The characteristic function of a $p\times 1$ random vector x is

defined ~~to be~~ as $\varphi_x(t) = E[e^{it^T x}]$, where t is a $p \times 1$ constant vector in some neighborhood of **0** vector.

Vector and Matrix Differentiation:

Let $u = g(x)$.

$$\frac{\partial u}{\partial x} \triangleq \begin{bmatrix} \frac{\partial u}{\partial x_1} \\ \frac{\partial u}{\partial x_2} \\ \vdots \\ \frac{\partial u}{\partial x_p} \end{bmatrix} \quad \text{and} \quad \frac{\partial^2 u}{\partial x \partial x^T} \triangleq \begin{pmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial u}{\partial x_1}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial u}{\partial x_1}\right) & \cdots & \frac{\partial}{\partial x_p}\left(\frac{\partial u}{\partial x_1}\right) \\ \frac{\partial}{\partial x_1}\left(\frac{\partial u}{\partial x_2}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial u}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_p}\left(\frac{\partial u}{\partial x_2}\right) \\ \vdots & \vdots & & \vdots \\ \frac{\partial}{\partial x_1}\left(\frac{\partial u}{\partial x_p}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial u}{\partial x_p}\right) & \cdots & \frac{\partial}{\partial x_p}\left(\frac{\partial u}{\partial x_p}\right) \end{pmatrix}.$$

Properties 2.2.

(a) Let $u = a^T y$, $\frac{\partial u}{\partial y} = a$.

(b) Let $u = y^T y$, $\frac{\partial u}{\partial y} = 2y$.

(c) Let $u = y^T A y$, $\frac{\partial u}{\partial y} = Ay + A^T y$. (When A is symmetric, $\frac{\partial u}{\partial y} = 2Ay$.)

Transformations of Random Vectors: Let x be a continuous $p \times 1$ random vector with p.d.f. $f_X(x)$, and let $y = u(x)$ be a 1-1 transformation. Then, the p.d.f. of y is

$$f_Y(y) = f_X(u^{-1}(y))\left|\det(J)\right| \quad \text{, where } J = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \end{pmatrix}.$$

Measures of Multivariate Scatter (Dispersion):

(a) The generalized variance: $\det(V(x))$.

(b) The total variation: $tr(V(x))$.

Linear Combinations: Let A be a $q \times p$ constant matrix, b be a $q \times 1$ constant vector, and x be a $p \times 1$ random vector with $x \sim (\mu, \Sigma)$, where Σ is a p.d. matrix. Then, $y = Ax + b \sim (A\mu + b, A\Sigma A^T)$.

Scaling transformation: $y = (diag(\Sigma))^{-1/2}(x - \mu) \sim (0, (diag(\Sigma))^{-1/2}\Sigma(diag(\Sigma))^{-1/2})$, where $P = (diag(\Sigma))^{-1/2}\Sigma(diag(\Sigma))^{-1/2}$ is a correlation matrix.

Mahalanobis transformation: $z = \Sigma^{-1/2}(x - \mu) \sim (0, \mathrm{I}_p)$.

Distances or Dissimilarities:

Euclidean Distance: $\|x_1 - x_2\|_2 = ((x_1 - x_2)^T (x_1 - x_2))^{1/2}$.

Mahalanobis Distance: $\|z_1 - z_2\|_2 = ((x_1 - x_2)^T \Sigma^{-1} (x_1 - x_2))^{1/2}$.

Karl Pearson Distance: $\|y_1 - y_2\|_2 = ((x_1 - x_2)^T (\mathrm{diag}(\Sigma))^{-1} (x_1 - x_2))^{1/2}$.

Multivariate Normal Distribution:

Definition. The p-variate random vector x follows a multivariate normal distribution if for any $p \times 1$ constant vector b, $U = b^T x$ is either a constant or a univariate normal.

Theorem 2.1. If x is p-variate normal and $y = Ax + b$ for a $q \times p$ constant matrix A and a $q \times 1$ constant vector b, then y is q-variate normal.

Corollary 2.1.1. If z is p-variate standard normal with p.d.f. $f_z(z) = (2\pi)^{\frac{-p}{2}} e^{\frac{-1}{2} z^T z}$, and $y = Az + b$, then y is q-variate normal with mean b and variance matrix AA^T.

Corollary2.1.2. Let x be p-variate normal with $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$, where x_1 and x_2 are separately $q \times 1$ and $(p - q) \times 1$ random vectors, respectively.

Let $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then, $x_1 \sim N_q(\mu_1, \Sigma_{11})$ and $x_2 \sim N_{p-q}(\mu_2, \Sigma_{22})$.

Theorem 2.2. For any $p \times 1$ constant vector μ and $p \times p$ constant matrix Σ, which is a p.s.d. matrix, there exists a unique multivariate normal distribution with mean μ and variance Σ.

Proof: (By the uniqueness of characteristic function $\varphi_x(t) = e^{it^T\mu - t^T\Sigma t/2}$ for $N_p(\mu, \Sigma)$.)

Lemma 2.1. If a $p \times p$ constant matrix Σ is p.s.d., there exists a $p \times p$ matrix M such that $\Sigma = M.^2$ (M is often denoted by $\Sigma^{1/2}$.)

Theorem 2.3. For any $p \times 1$ constant vector μ and $p \times p$ constant matrix Σ, which is a p.s.d. matrix, a random vector $y \sim N_p(\mu, \Sigma)$ can be generated as $y = \Sigma^{1/2} z + \mu$.

Proof: (Corollary 2.1.1 and Theorem 2.2)

Theorem 2.4. Let $x \sim N_p(\mu, \Sigma)$, where Σ is p.d. The density function of x is

$$f_X(x) = (2\pi)^{\frac{-p}{2}} (\det(\Sigma))^{\frac{-1}{2}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) I_{\mathbb{R}}(x).$$

Theorem 2.5. Let x be p-variate normal with $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$, where x_1 and x_2 are separately $q \times 1$ and $(p-q) \times 1$ random vectors. Let $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. The following two conditions are equivalent:

(a) $\Sigma_{12} = 0$, i.e., x_1 and x_2 are uncorrelated.

(b) x_1 and x_2 are independent.

Example 6. (Appendix)

Theorem 2.6. Let $y \sim N_n(X\beta, \sigma^2 \mathrm{I}_n)$. Then, the least squares estimators $\hat{\beta}$ and the residual vector e have the following properties:

(a) $\begin{pmatrix}\hat{\beta}\\ e\end{pmatrix}$ is multivariate normal.

(b) $\hat{\beta}$ and e are independent.

(c) $\begin{pmatrix}\hat{\beta}\\ e\end{pmatrix} \sim N_{p+n}\left(\begin{pmatrix}\beta\\ 0\end{pmatrix}, \begin{pmatrix}\sigma^2\left(X^TX\right)^{-1} & 0\\ 0 & \sigma^2\left(\mathrm{I}_n - H\right)\end{pmatrix}\right)$.

Theorem 2.7. Let $X = \begin{pmatrix}X_1\\ X_2\end{pmatrix} \sim N_p(\mu, \Sigma)$ with Σ being a positive definite matrix. The conditional distribution of X_2 given $X_1 = x_1$ is $(p-q)$-variate normal with mean ~~and variance being~~ $\mu_{2\cdot 1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)$ and $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, respectively.

Proof：

Let $Y \triangleq \begin{pmatrix}Y_1\\ Y_2\end{pmatrix} = \begin{pmatrix}\mathrm{I}_q & 0\\ -\Sigma_{21}\Sigma_{11}^{-1} & \mathrm{I}_{p-q}\end{pmatrix}\begin{pmatrix}X_1\\ X_2\end{pmatrix}$. By Theorem 2.5 and $Cov(Y_1, Y_2) = 0$, it implies that Y_1 and Y_2 are independent. Thus, the distribution of Y_2 given $X_1 = x_1$ is same with the unconditional distribution of Y_2, where $Y_2 \sim N_{p-q}(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$.

Conditioning on $X_1 = x_1$, one has $X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1 \sim N_{p-q}(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$,

i.e., $X_2 \mid X_1 = x_1 \sim N_{p-q}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22.1})$.

Remark. From Theorem 2.6, one has $\det(\Sigma) = \det(\Sigma_{11})\det(\Sigma_{22\cdot 1})$, which is consistent with the property $\det\begin{pmatrix}A_{11} & 0\\ 0 & A_{22\cdot 1}\end{pmatrix} = \det(C_1 A C_2) = \det(A)$ with $C_1 = \begin{pmatrix}\mathrm{I}_q & 0\\ -A_{21}A_{11}^{-1} & \mathrm{I}_{p-q}\end{pmatrix}$ and $C_2 = \begin{pmatrix}\mathrm{I}_q & -A_{11}^{-1}A_{12}\\ 0 & \mathrm{I}_{p-q}\end{pmatrix}$.

Theorem 2.8. Assume that $X \sim N_p(\mu, \Sigma)$ with a p.d. Σ. ~~where Σ is p.d.~~ Let $\chi^2 = (X-\mu)^T\Sigma^{-1}(X-\mu)$.

Then, $\chi^2 \sim \chi_p^2$.

Proof:

Let $Z = \Sigma^{-1/2}(X - \mu)$. $Z \sim N_p(0, I_p)$, and, hence $\chi^2 = Z^T Z \sim \chi_p^2$.

Example 7. (See Appendix)

Theorem 2.9. (Multivariate Central Limit Theorem) Suppose that $X_1, \cdots, X_n$ are independent and identically distributed p-variate random vectors with mean μ and variance Σ. Let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then , $\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N_p(0, \Sigma)$ as $n \to \infty$.

Spherical Distribution: The p-variate random vector z with density function $f_Z(z)$ is spherical or spherically symmetric if $f_Z(z)$ depends on z only through $z^T z$. (an equivalent statement: $\Gamma z \sim z$ for any orthogonal matrix Γ.

Remark: Let $u = \Gamma z$, where $\Gamma = \begin{pmatrix} ct^T \\ \Gamma_1 \end{pmatrix}$ with $\Gamma^T \Gamma = I_p$ and $c = (tt^T)^{-1/2}$. The c.h.f. of z is

$$\phi_Z(t) = \iiint e^{u_1/c} f_Z(\sum_{i=1}^{p} u_i^2) du_1 \ldots du_p \triangleq g(t^T t).$$

Elliptical Distribution: Assume that z follows a p-variate spherical distribution, A and m are separately $p \times p$ constant matrix and $p \times 1$ fixed vector. Then, $x = Az + m$ is said to be elliptical or elliptically symmetric.

Note: Let $z \sim (0, cI_p)$ and $x \sim (\mu, \Sigma)$, one can derive that the characteristic function of x is $\phi_x(t) = E[e^{it^T X}] = e^{it\mu} \cdot g(t^T \Sigma t)$.

Theorem 2.10. Let x follow a p-variate ellipitical distribution with variance matrix Σ. If diagonality of Σ implies independence of all components of x, then x is multivariate normal.

Proof: (See Muirhead (1982)).

Without loss of generality, we assume that $E[x] = 0$. By the diagonality of Σ and the independence of x_i's, one can derive $g(\sum_{i=1}^{p} t_i^2 \sigma_i^2) = \prod_{i=1}^{p} g(t_i^2 \sigma_i^2)$. Thus, $g(z) = e^{kz}$ for $k \le 0$, which implies that $\phi_x(t) = e^{-t^T \Sigma t/2}$ because it is the characteristic function. By the uniqueness of the c.h.f, x is shown to be multivariate normal.

Wishart Distribution ($W_p(\Sigma,n)$): Let $X_{n\times p}$ be a data matrix with x_i's being a random sample from a $N_p(0,\Sigma)$. Then, $X^TX=\sum_{i=1}^n x_i x_i^T \sim W_p(\Sigma,n)$.

Theorem 2.11. Let $x_1,\cdots,x_n$ be a random sample from $N_p(0,\Sigma)$, $X=(x_1,...,x_n)^T$, and C be a $n\times n$ idempotent matrix with $rank(C)=r$. Then, $X^TCX\sim W_p(\Sigma,r)$.

Proof:

Let $C=\Gamma_n\begin{bmatrix} I_r & 0 \\ 0 & 0\end{bmatrix}\Gamma_n^T$ and $Y=\Gamma_n^TX$ with $\Gamma_n=(\gamma_{(1)},\ldots,\gamma_{(n)})$, $\Gamma_n^T\Gamma_n=I_n$. $y_j=\sum_{i=1}^n \gamma_{i(j)}x_i \sim N_p(0,\Sigma)$, y_j's are independent

One has $X^TCX=Y^T\begin{bmatrix} I_r & 0 \\ 0 & 0\end{bmatrix}Y=\sum_{j=1}^r y_j y_j^T$. Since y_j's are i.i.d random vectors from a $N_p(0,\Sigma)$, it implies that $X^TCX\sim W_p(\Sigma,r)$.

Example 8. (Appendix)

Hotelling T^2 Distribution ($T^2(p,n)$): Let $d\sim N_p(0,I_p)$ and $M\sim W_p(I,n)$ be mutually independent. Then, $nd^TM^{-1}d\sim T^2(p,n)$.

Example 9. (Appendix)

Theorem2.12. Let $M=\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22}\end{pmatrix}\sim W_p(\Sigma,n)$ with $\dim(M_{11})=a$. Then,

$M_{22.1}=M_{22}-M_{21}M_{11}^{-1}M_{12}\sim W_{p-a}(\Sigma_{22.1},n-a)$.

Proof:

Let $M=X^TX$ with $X=(X_1,X_2)$ being a data matrix from a $N_p(0,\Sigma)$ and $M_{ij}=X_i^TX_j$.

Then, $M_{22.1}=X_2^TPX_2$ with $P=I_n-X_1(X_1^TX_1)^{-1}X_1$. $X_{2.1}=X_2-\Sigma_{21}\Sigma_{11}^{-1}X_1$

Since $PX_1=0$, one has $M_{22.1}=X_2^TPX_2=X_{2.1}^TPX_{2.1}$, where $X_{2.1}=X_2-X_1\Sigma_{11}^{-1}\Sigma_{12}$.

From the proof of Theorem 2.7, we know that $X_{2.1}$ is a data matrix from $N_{p-a}(0,\Sigma_{22.1})$.

It is implied ~~and, hence,~~ by Theorem 2.11~~, it can be derived~~ that $M_{22.1}\sim W_{p-a}(\Sigma_{22.1},n-a)$.

Theorem 2.13. $T^2(p,n) \sim \frac{np}{n-p+1} F_{p,n-p+1}$.

Proof:

Let $d \sim N_p(0, I_p)$ be independent of $M \sim W_p(I_p, n)$, and $D = (D_1, d)$ be a nonsingular matrix with $d^T D_1 = 0$. One has $\alpha = n d^T M^{-1} d = \frac{n d^T M^{-1} d}{d^T d} d^T d \sim T^2(p,n)$.

From $N = D^{-1} M (D^{-1})^T \sim W_p(D^{-1}(D^{-1})^T, n)$, one has $(N_{22.1})^{-1} = d^T M^{-1} d$ with $N_{22.1} \sim W_1((d^T d)^{-1}, n-p+1)$.

Thus, conditioning on d, $\beta = (d^T d / d^T M^{-1} d) \sim \chi^2_{n-p+1}$, which does not depend on d.

By using $d^T d \sim \chi^2_p$, we have $\alpha = n \frac{(d^T d / p)}{(\beta/(n-p+1))} \frac{p}{(n-p+1)} \sim \frac{np}{(n-p+1)} F_{p,n-p+1}$.

Remark: It implies from Theorem 2.13 that $\frac{n(n-p)}{(n-1)p} (\bar{x}_n - \mu)^T (S_n^2)^{-1} (\bar{x}_n - \mu) \sim F_{p,n-p}$.

Theorem 2.14. Let $d \sim N_p(0, I_p)$ be independent of $M \sim W_p(I_p, n)$. Then, $\frac{|M|}{|M + dd^T|} \sim Beta(\frac{n-p+1}{2}, \frac{p}{2})$.

Proof:

$$\frac{|M|}{|M+dd^T|} = |I_p + M^{-1} dd^T|^{-1} = (1 + d^T M^{-1} d)^{-1} = \frac{\beta}{d^T d + \beta} \sim Beta(\frac{n-p+1}{2}, \frac{p}{2}).$$

Wilks' Lambda Distribution ($\Lambda(p,n,m)$): Let $A \sim W_p(I, n)$ and $B \sim W_p(I, m)$ be mutually independent with $n \geq p$. Then, $\Lambda = |A| / |A+B| = |I + A^{-1} B|^{-1} \sim \Lambda(p,n,m)$.

Theorem 2.15. $\Lambda(p,n,m) \sim \prod_{i=1}^{m} u_i$, where u_i's are independent Beta random variables with parameters $(\frac{n+i-p}{2}, \frac{p}{2}), i = 1, \cdots, m$.

Proof:

Let $B = X^T X$ with $x_1, \cdots x_m$ being a sample from a $N_p(0, I_p)$, and $M_i = M_{i-1} + x_i x_i^T, i =$

$= 1, \cdots, m$, with $M_0 = A$. One has $\Lambda(p,n,m) = \prod_{i=1}^{m} u_i$ with $u_i = \frac{|M_{i-1}|}{|M_i|}, i = 1, \cdots, m.$

By Theorem 2.14, one has $u_i \sim Beta(\frac{n+i-p}{2}, \frac{p}{2}), i = 1, \cdots, m.$

Using the independence of M_i and u_i, and the independence of u_i and $x_{i+1}, \cdots, x_m$, it follows that u_i's are independent.

Appendix

Example 1.

Let X be a $n \times p$ matrix with $rank(X) = p < n$, $H = X(X^T X)^{-1} X^T$.

(a) H is an idempotent matrix. ($H^2 = \left(X(X^T X)^{-1} X^T\right)\left(X(X^T X)^{-1} X^T\right) = H$)

(b) The eigenvalues of H are either 0 or 1.

Let λ and y be any corresponding eigenvalue and eigenvector of H. One has $Hy = \lambda y$, and, hence, $H(Hy) = H(\lambda y)$.

Since H is an idempotent matrix, ~~$\therefore H(Hy) = H(\lambda y)$.~~

~~Since~~ the left-hand side $= H^2 y = Hy = \lambda y$ and the right-hand side $= \lambda Hy = \lambda^2 y$, we ~~get~~ have $\lambda(\lambda-1)y = 0$. It follows that $\lambda = 0$ or 1.

~~$\lambda(\lambda - 1)y = 0$, which implies that $\lambda = 0$ or 1.~~

(c) $rank(H) = trace(H) = p$.

$trace(H) = trace(X(X^T X)^{-1} X^T) = trace((X^T X)^{-1} X^T X) = trace(I_p) = p$.

Let $y = (X^T X)^{-1/2} X^T$. One has

$$rank(H) = rank(X(X^T X)^{-1} X^T) = rank(YY^T) = rank(Y)$$
$$= rank((X^T X)^{-1/2} X^T) = rank(X) = p.$$

Example 2. (Least Squares Estimation)

Let $y = (y_1, \cdots, y_n)^T \sim (X\beta, \sigma^2 \mathrm{I}_n)$, where X is a $n \times p$ known matrix with $p < n$ and

$rank(X) = p$, β is a $p \times 1$ unknown parameter vector, and σ^2 is an unknown positive parameter.

The least squares estimation criterion for β is to find a minimizer of the following sum of squares: $S(\beta) = (y - X\beta)^T (y - X\beta)$. The minimizer, say $\hat{\beta}$, of $S(\beta)$ can be derived to be $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, and is called the least squares estimator of β.

The estimated or predicted vector is $\hat{y} = X\hat{\beta} \triangleq Hy$, where $H = X\left(X^T X\right)^{-1} X^T$ is called a hat matrix. The residual vector is defined to be $e = y - \hat{y} = \left(\mathrm{I}_n - H\right) y$.

Let $X=(X_1,X_2)$, $\beta=(\beta^T{}_1,\beta^T{}_2)^T$ $H=X(X^TX)^{-1}X^T$ and $H_{22\cdot1}=X_2^T(I-H_1)X_2$ with $H_1=X_1(X_1^TX_1)^{-1}X_1^{-1}$.

One has $\hat{\beta}_2=(-H_{22\cdot1}^{-1}(X_2^TX_1)(X_1^TX_1)^{-1},H_{22\cdot1}^{-1})\begin{pmatrix}X_1^T\\X_2^T\end{pmatrix}Y$, which implies that

$$\hat{\beta}_2=(-H_{22\cdot1}^{-1}X_2^TH_1+H_{22\cdot1}^{-1}X_2^T)Y=(H_{22\cdot1}^{-1}X_2^T(I-H_1))Y=(e_2^Te_2)^{-1}e_2^TY$$

$$(\text{or}\quad =(e_2^Te_2)^{-1}e_2^T((I-H_1)Y)),$$

where $e_2=(I-H_1)X_2$.

Reference:

1. Rao, C. R. and Toutenburg, H. (1999). Linear Models: Least Squares and Alternatives. Second Edition. Springer-Verlag.
2. Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). Applied Regression Analysis: A Research Tool. Second Edition. Springer-Verlag.
3. Sen , A. K. and Srivastava , M. S. (199 1) Regression Analysis: Theory, Methods, and Applications. Springer-Verlag.

Example 3.

Let Σ be a variance matrix of a $p\times1$ random vector X.

(a) Σ is at least a p.s.d. matrix.

Let a be any $p\times1$ constant vector, $a^T\Sigma a=a^TV(X)a=V(a^TX)\geq0$.

(b) By the spectral decomposition theorem , Σ can be expressed as $\Sigma=\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T$.

One has $\lambda_i{}'s\geq0$ ($\because\ \lambda_i=\gamma_{(i)}^T\Sigma\gamma_{(i)}\geq0$).

Without loss of generality, let $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$. If $\lambda_p>0$, then Σ is non-singular. ($\because$ $rank(\Sigma)=rank(\Gamma\Lambda\Gamma^T)=rank(\Lambda)=p$.).

(c) Find a vector η satisfying $\eta^T\eta=1$ such that the variance of $U=\eta^TX$ is maximized .

Since η can be expressed as $\eta=\sum_{i=1}^{p}\eta_i\gamma_{(i)}$, it can be derived that

$Var(U)=\eta^TV(X)\eta=\eta^T\sum_{i=1}^{p}\lambda_i\gamma_{(i)}\gamma_{(i)}^T\eta=\sum_{i=1}^{p}\lambda_i\eta_i^2\leq\lambda_1\sum_{i=1}^{p}\eta_i^2=\lambda_1$. Equality holds if and only if $\eta=\gamma_{(1)}$.

Example 4. (Example 2 continued)

Let $e = Y - X\hat{\beta} = (I_n - H)Y$.

(a) $E[\hat{\beta}] = E[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^TE[Y] = (X^TX)^{-1}X^TX\beta = \beta$ and

$$V(\hat{\beta}) = (X^TX)^{-1}X^TV(Y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

(b) $E[e] = (I-H)E[Y] = (I-H)X\beta = (X-X)\beta = 0$ ($\because HX = X$) and

$V[e] = (I-H)(\sigma^2 I_n)(I-H) = \sigma^2(I-H)$.

(c) $Cov(\hat{\beta}, e) = Cov((X^TX)^{-1}X^TY, (I_n - H)Y) = (X^TX)^{-1}X^T(\sigma^2 I_n)(I_n - H) = 0$.

(d) $E[e^Te] = E[Y^T(I-H)Y] = tr((I_n - H)(\sigma^2 I_n)) + \beta^TX^T(I_n - H)X\beta = (n-p)\sigma^2$

($\because I_n - H$ is idempotent) .

Example 5.

Let $\begin{pmatrix} x \\ y \end{pmatrix} \sim \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$ with Σ_{11} and Σ_{22} being non-singular. Moreover, let $\eta = a^Tx$ and $\phi = b^Ty$.

Consider $\rho(\eta, \phi) = \dfrac{a^T\Sigma_{12}b}{(a^T\Sigma_{11}a)^{1/2}(b^T\Sigma_{22}b)^{1/2}}$ subject to $a^T\Sigma_{11}a = b^T\Sigma_{22}b = 1$, the objective here is to find (a,b) such that $\rho(\eta,\phi)$ is maximized.

Let $a^{*T} = \Sigma_{11}^{\frac{1}{2}}a$ and $b^{*T} = \Sigma_{11}^{\frac{1}{2}}b$. $\rho(\eta,\phi)$ can be re-expressed as

$\rho(\eta,\phi) = a^{*T}(\Sigma_{11}^{\frac{-1}{2}}\Sigma_{12}\Sigma_{22}^{\frac{-1}{2}})b^{*T} \triangleq a^{*T}Kb^{*T}$, where $K = \Sigma_{11}^{\frac{-1}{2}}\Sigma_{12}\Sigma_{22}^{\frac{-1}{2}}$.

By the singular value decomposition theorem, $K = ULV^T$ with $U = (\alpha_1, \cdots, \alpha_p)$, $V = (\beta_1, \cdots, \beta_q)$, and $L = \begin{pmatrix} diag(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_k}) & 0_{k\times(q-k)} \\ 0_{(p-k)\times k} & 0_{(p-k)\times(q-k)} \end{pmatrix}$ satisfying $KK^T = UL^2U^T$ and $K^TK = VL^2V^T$, where $\lambda_1 \geq \cdots \geq \lambda_k > 0$.

Since a^* and b^* can be expressed as $a^* = \sum_{i=1}^{p} a_i^* \alpha_i$ and $b^* = \sum_{i=1}^{q} b_i^* \beta_i$, one has

$$\rho(\eta,\phi) = (\sum_{i=1}^{k} a_i^* b_i^* \sqrt{\lambda_i}) \le \left[(\sum_{i=1}^{k} a_i^{*2} \sqrt{\lambda_i})(\sum_{i=1}^{k} b_i^{*2} \sqrt{\lambda_i}) \right]^{\frac{1}{2}}.$$

(Equality holds if $a_i^* = c b_i^*, i = 1, \cdots, k, \quad c > 0$.)

$$\le \sqrt{\lambda_1}.$$

(Equality holds if $a_1^* = b_1^* = 1, a_2^* = \cdots = a_p^* = 0$, and $b_2^* = \cdots = b_q^* = 0$.)

Thus, $\rho(\eta,\phi)$ is maximized when $(a,b) = (\Sigma_{11}^{\frac{-1}{2}} \alpha_1, \Sigma_{22}^{\frac{-1}{2}} \beta_1)$.

Example 6. Let $X_1, \cdots, X_n$ be a random sample from $N_p(\mu, \sigma^2)$. Then, the sample mean

$\bar{X}_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$ and the sample variance $S_n^{\,2} = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}{(n-1)}$ are independent.

Proof:

Let $X = (X_1, \cdots, X_n)^T$ and $I_n - H = I_n - \dfrac{1}{n} 1_n 1_n^{\,T}$.

One has $\bar{X}_n = \dfrac{1}{n} 1_n^{\,T} X$ and $S^2 = \dfrac{X^T (I_n - H) X}{(n-1)} = \dfrac{((I_n - H)X)^T ((I_n - H)X)}{(n-1)}$.

From $\begin{pmatrix} 1_n^T \\ I_n - H \end{pmatrix} X \sim N_{n+1}\left(\begin{pmatrix} n\mu \\ 0 \end{pmatrix}, \begin{pmatrix} n\sigma^2 & 0 \\ 0 & (I_n - H)\sigma^2 \end{pmatrix} \right)$, it implies that $\bar{X}_n$ and $(I_n - H)X$ are independent. Since $S_n^{\,2}$ is a function of $(I_n - H)X$, $\bar{X}_n$ and $S_n^{\,2}$ are shown to be independent.

Example 7.

Let $X_1, \cdots, X_n$ be a random sample from a $N(\mu, \sigma^2)$. Show that $\dfrac{(n-1)S_n^{\,2}}{\sigma^2} \sim \chi_{n-1}^2$.

Proof：

Since $I_n - \dfrac{1}{n} 1_n 1_n^{\,T}$ is an idempotent matrix with rank $(n-1)$, by the spectral decomposition theorem, one has

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}{\sigma^2}=\frac{x^T\left(\mathrm{I}_n-\frac{1}{n}1_n1_n^T\right)x}{\sigma^2}=\tilde{z}^T\left(\mathrm{I}_n-\frac{1}{n}1_n1_n^T\right)\tilde{z}=\tilde{z}^T\left(\Gamma_{n-1}\Gamma_{n-1}^T\right)\tilde{z},$$

where $\Gamma_{n-1}=(r_{(1)},\cdots,r_{(n-1)})$ with $\Gamma_{n-1}1_n=0$ and $\tilde{z}=\frac{x-\mu 1_n}{\sigma}\sim N_n(0,I_n)$.

Since $z=\Gamma_{n-1}^T\tilde{z}\sim N_{n-1}(0,\mathrm{I}_{n-1})$, it implies that z_i^2's are i.i.d. χ_1^2 random variables, and,

hence, $\tilde{z}^T\left(\Gamma_{n-1}\Gamma_{n-1}^T\right)\tilde{z}=\sum_{i=1}^{n-1}z_i^2\sim\chi_{n-1}^2$.

Example 8.

Let $x_1,\cdots,x_n$ be a random sample from a $N_p(\mu,\Sigma)$ and $X=(x_1,...,x_n)^T$ be a data matrix . Then, $(n-1)S_n^{\,2}=X^T(I_n-\frac{1}{n}1_n1_n^T)X\sim W_p(\Sigma,(n-1))$.

Proof:

Let $\tilde{\mathrm{X}}=(\tilde{x}_1,\ldots,\tilde{x}_n)^T$ with $\tilde{x}_i=x_i-\mu$, ($i=1,\cdots,n$) where $\tilde{\mathrm{X}}$ is a data matrix from a $N_p(0,\Sigma)$.

One has $(n-1)S_n^{\,2}=X^T(I_n-\frac{1}{n}1_n1_n^T)X=\tilde{X}^{\mathrm{T}}(I_n-\frac{1}{n}1_n1_n^T)\tilde{\mathrm{X}}$. ← $\tilde{X}=X-1_n\mu^T$

Since $(I_n-\frac{1}{n}1_n1_n^T)$ is an idempotent matrix with $rank(I_n-\frac{1}{n}1_n1_n^T)=(n-1)$, it ~~implies~~ follows

from Theorem 2.11 that $(n-1)S_n^{\,2}\sim W_p(\Sigma,(n-1))$.

Example 9. Let $\bar{\mathrm{x}}_{\mathrm{n}}=\frac{1}{\mathrm{n}}\mathrm{X}^{\mathrm{T}}1_{\mathrm{n}}$. Then, $n(\bar{x}_n-\mu)^T(S_n^{\,2})^{-1}(\bar{x}_n-\mu)\sim T^2(p,(n-1))$.

$\Sigma^{-\frac{1}{2}}(\sqrt{n}(\bar{X}_n-\mu))\sim N_p(0,I_p)$

& $(n-1)S_n^2\sim W_p(\Sigma,(n-1))$,

which implies that $(n-1)\Sigma^{-\frac{1}{2}}S_n^2\Sigma^{-\frac{1}{2}}\sim W_p(I_p,n-1)$.

$\Rightarrow (n-1)\left(\Sigma^{-\frac{1}{2}}\sqrt{n}(\bar{X}_n-\mu)\right)^T\left((n-1)\Sigma^{-\frac{1}{2}}S_n^2\Sigma^{-\frac{1}{2}}\right)^{-1}\left(\Sigma^{-\frac{1}{2}}\sqrt{n}(\bar{X}_n-\mu)\right)$

$\sim T^2(p,n-1)$