**Random Vector**

A random vector $U \in \mathbb{R}^k$ is a vector $(U_1, U_2, \cdots, U_k)^T$ of scalar random variables $U_i$ defined over a common probability space. The expectation and variance of $U$ are defined as:

$$\mathbb{E}U := \begin{pmatrix} \mathbb{E}U_1 \\ \mathbb{E}U_2 \\ \vdots \\ \mathbb{E}U_k \end{pmatrix}, \mathbb{V}\text{ar}U := \mathbb{E}(\{U-\mathbb{E}U\}\{U-\mathbb{E}U\}^T) = \begin{pmatrix} \mathbb{V}\text{ar}U_1 & \mathbb{C}\text{ov}(U_1, U_2) & \dots & \mathbb{C}\text{ov}(U_1, U_k) \\ \mathbb{C}\text{ov}(U_2, U_1) & \mathbb{V}\text{ar}U_2 & \dots & \mathbb{C}\text{ov}(U_2, U_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}(U_k, U_1) & \mathbb{C}\text{ov}(U_k, U_2) & \dots & \mathbb{V}\text{ar}U_k \end{pmatrix}.$$

It is easy to check that if $X = a + BU$ where $a\mathbb{R}^p$ and $B$ is a $p \times k$ matrix, then $X$ is a random vector in $\mathbb{R}^p$ with $\mathbb{E}X = a + B\ \mathbb{E}U$, $\mathbb{V}\text{ar}X = B\mathbb{V}\text{ar}(U)B^T$. You should also note that if $\Sigma = \mathbb{V}\text{ar}U$ for some random vector $U$, then $\Sigma$ is a $k \times k$ non-negative definite matrix, because for any $a \in \mathbb{R}^k$, $a^T\Sigma a = \mathbb{V}\text{ar}(a^TU) \geq 0$.

**Multivariate Normal**

**Definition.** A random vector $U \in \mathbb{R}^k$ is called a normal random vector if for $a \in \mathbb{R}^k$, $a^TU$ is a (one dimensional) normal random variable.

**Theorem 1.** *A random vector $U \in \mathbb{R}^k$ is a normal random vector if and only if one can write $U = m+AZ$ for some $m \in \mathbb{R}^k$ and $k\times k$ matrix $A$ where $Z = (Z_1, \cdots, Z_k)^T$ with $Z_i \overset{IID}{\sim} N(0,1)$.*

*Proof.* "If part". Suppose $U = m + AZ$ with $m$, $A$ and $Z$ as in the statement of the theorem. Then for any $a \in \mathbb{R}^k$,

$$a^TU = a^Tm + a^TAU = b_0 + \sum_{i=1}^{n} b_iZ_i$$

for some scalars $b_0, b_1, \cdots, b_k$. But a linear combination of independent (one dimensional) normal variables is another normal, so $a^TU$ is a normal variable.

"Only if part" Suppose $U$ is a normal random vector. It suffices to show that a $V = m + AZ$ with $Z$ as in the statement of the theorem, and suitably chosen $m$ and $A$, has the same distribution as $U$. For any $a \in \mathbb{R}^k$, the moment generating function $M_U(a)$ of $U$ at $a$ is $\mathbb{E}e^{a^TU} = \mathbb{E}e^X$ where $X = a^TU$ is normally distributed by definition. But, by one dimensional normal distribution theory, $\mathbb{E}e^X = e^{\mathbb{E}X+\frac{1}{2}\mathbb{V}\text{ar}X} = e^{a^T\mathbb{E}U+\frac{1}{2}a^T(\mathbb{V}\text{ar}U)a} = e^{a^T\mu+a^T\Sigma a}$ where we denote $\mathbb{E}U$ by $\mu$ and $\mathbb{V}\text{ar}U$ by $\Sigma$. Note that $\Sigma$ is non-negative definite and thus can be written as $\Sigma = AA^T$ for some $k \times k$ matrix $A$.

Write $V = \mu + AZ$ where $Z = (Z_1, \cdots, Z_k)^T$ with $Z_i \overset{\text{IID}}{\sim} N(0,1)$. Then, by the "if part", $V$ is a normal random vector, and because $Z_i$'s are IID with mean 0 and variance 1, $\mathbb{E}V = \mu$ and $\mathbb{V}\text{ar}V = \Sigma$. So by above discussion $M_V(a) = e^{a^T\mu+\frac{1}{2}a^T\Sigma a} = M_U(a)$. So $U$ and $V$ have the same moment generating function. Because this moment generating function is defined for all $a \in \mathbb{R}^k$, it uniquely determines the associated probability distribution. That is, $V$ and $U$ have the same distribution.

**Notation.** If a random $k$-vector $U$ is a normal random vector, then by above proof, its distribution is completely determined by its mean $\mu = \mathbb{E}U$ and variance $\Sigma = \mathbb{V}\text{ar}U$. We shall

denote this distribution by $N_k(\mu, \Sigma)$. Note that $U \sim N_k(\mu, \Sigma)$ then means that $U = \mu + AZ$ for $Z$ as in the above theorem, where $A$ satisfies $\Sigma = AA^T$.

**Theorem 2 (Linear transformations).** *If $U \sim N_k(\mu, \Sigma)$ and $a \in \mathbb{R}^p$ and $B$ is $p \times p$ matrix, then $V = a + BU \sim N_p(a + B\mu, B\Sigma B^T)$.*

*Proof.* Write $U = \mu + AZ$ where $Z$ is as in Theorem 1 and $A$ satisfies $\Sigma = AA^T$. Then, $V = a + BU = (a + B\mu) + (BA)Z$. This proves the result.

**Theorem 3 (Density).** *A $N_k(\mu, \Sigma)$ distribution with a positive definite $\Sigma$, admits the following pdf:*

$$f(x) = (2\pi)^{-k/2}(\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \;\; x \in \mathbb{R}^k.$$

*Proof.* Suppose $X \sim N_k(\mu, \Sigma)$ then $X = \mu + AZ$ where $AA^T = \Sigma$ and $A$ is non-singular (because $\Sigma$ is p.d.). The joint pdf of $Z_1, \cdots, Z_k$ is given by

$$g(z_1, \cdots, z_k) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z_i^2}{2}\right\} = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}z^T z\right\}, z = (z_1, \cdots, z_k)^T \in \mathbb{R}^k.$$

Therefore, by multivariate change of variable formula for the transformation $X = \mu + AZ$ with inverse transformation $Z = A^{-1}(X - \mu)$ and hence Jacobian $J = (\det A)^{-1}$, the pdf of $X$ is,

$$\begin{aligned} f(x) &= |(\det A)^{-1}| g(A^{-1}(x-\mu)), \;\; x \in \mathbb{R}^k \\ &= (2\pi)^{-k/2} |\det A|^{-1} \exp\left\{-\frac{1}{2}(x-\mu)^T(AA^T)^{-1}(x-\mu)\right\}, \;\; x \in \mathbb{R}^k \\ &= (2\pi)^{-k/2}(\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \;\; x \in \mathbb{R}^k. \end{aligned}$$

as $\det \Sigma = (\det A)^2$.

**Theorem 4 (independence).** *Let $U \sim N_k(\mu, \Sigma)$. Suppose for some $1 \leq p < k$, $\Sigma$ can be partitioned as*

$$\Sigma = \begin{pmatrix} \Sigma_{(11)} & 0_{p\times(k-p)} \\ 0_{(k-p)\times p} & \Sigma_{(22)} \end{pmatrix}$$

*where $\Sigma_{(11)}$ is $p \times p$, $\Sigma_{(22)}$ is $(k-p) \times (k-p)$ and $0_{m\times n}$ denotes the matrix of zeros of the specified dimensions. Similarly partition*

$$U = \begin{pmatrix} U_{(1)} \\ U_{(2)} \end{pmatrix}, \text{ and } \mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}$$

*into vectors of length $p$ and $(k-p)$. Then*

$$U_{(1)} \sim N_p(\mu_{(1)}, \Sigma_{(11)}), \;\; U_{(2)} \sim N_{k-p}(\mu_{(2)}, \Sigma_{(22)})$$

*and $U_{(1)}$ and $U_{(2)}$ are independent.*

*Proof.* Because $\Sigma$ is non-negative definite, so must be any of its diagonal blocks. Therefore there are $p \times p$ and $(k-p) \times (k-p)$ matrices $A_{(1)}$ and $A_{(2)}$ such that $\Sigma_{(11)} = A_{(1)}A_{(1)}^T$ and $\Sigma_{(22)} = A_{(2)}A_{(2)}^T$. Clearly, $\Sigma = AA^T$ with

$$A = \begin{pmatrix} A_{(1)} & 0_{p\times(k-p)} \\ 0_{(k-p)\times p} & A_{(2)} \end{pmatrix}.$$

Because $U$ is normal, we must have $U = \mu + AZ$ where $Z = (Z_1, \cdots, Z_k)^T$ with $Z_i \overset{\text{IID}}{\sim} N(0,1)$. Write $Z_{(1)} = (Z_1, \cdots, Z_p)^T$ and $Z_{(2)} = (Z_{p+1}, \cdots, Z_k)^T$. Then $Z_{(1)}$ and $Z_{(2)}$ are independent and

$$\begin{pmatrix} U_{(1)} \\ U_{(2)} \end{pmatrix} = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix} + \begin{pmatrix} A_{(1)} & 0_{p\times(k-p)} \\ 0_{(k-p)\times p} & A_{(2)} \end{pmatrix} \begin{pmatrix} Z_{(1)} \\ Z_{(2)} \end{pmatrix} = \begin{pmatrix} \mu_{(1)} + A_{(1)}Z_{(1)} \\ \mu_{(2)} + A_{(2)}Z_{(2)} \end{pmatrix}.$$

From this the result follows.

**Sample mean and variance of normal data**

**Theorem 5.** *Let $X_i \overset{IID}{\sim} N(\mu, \sigma^2)$. Then*

1. $W = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$,
2. $V = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$
3. *and $V$ and $W$ are independent.*

*Proof.* Suffices to prove for $\mu = 0$ and $\sigma^2 = 1$ (otherwise take $\tilde{X}_i = (X_i - \mu)/\sigma$). Let $X$ denote the random vector $(X_1, \cdots, X_n)^T \in \mathbb{R}^n$. Then, $X_n \sim N_n(0_{n\times 1}, I_n)$ where $I_n$ denotes the $n \times n$ identity matrix with 1 on the diagonals and zero elsewhere. Let $B$ be a $n \times m$ orthogonal matrix whose first row equals $(1/\sqrt{n})1_{1\times n}$ where $1_{m\times n}$ denote the matrix of the given dimensions (such an orthogonal matrix can be constructed by the Gram-Schmidt method). Take $Y = BX$. Then

$$Y \sim N_n(0_{n\times 1}, BB^T) = N_n\left(0_{n\times 1}, I_n\right),$$

as, $B$ is orthogonal. Therefore the coordinate variables $Y_1, Y_2, \cdots, Y_n$ of $Y$ are IID $N(0,1)$. Now argue,

1. $Y_1 = (1/\sqrt{n})1_{1\times m}X = (1/\sqrt{n})\sum_{i=1}^n X_i = \sqrt{n}\bar{X}$. So $W = Y_1 \sim N(0,1)$.
2. $Y^TY = X^TX$ (since $B$ is orthogonal), and so $Y_1^2 + Y_2^2 + \cdots + Y_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2$. But $Y_1^2 = n\bar{X}^2$. Hence

$$Y_2^2 + \cdots + Y_n^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = V.$$

   So, $V = Y_2^2 + \cdots + Y_n^2 \sim \chi^2_{n-1}$ as this is a sum of squares of $n-1$ IID standard normal variables.

3. $W = Y_1$ and $V$ is a function of $Y_2, \cdots, Y_n$, hence these two are independent as $Y_i$'s are independent.

**Quadratic forms**

A $k \times k$ symmetric matrix $H$ is called idempotent if $H^2 = H$. The eigenvalues of an idempotent matrix are either 0 or 1. To see this, note that if $\lambda$ is an eigenvalue of an idempotent matrix $H$ then $Hv = \lambda v$ for some $v \neq 0$. Pre-multiply both sides by $H$ to get $H^2 v = \lambda H v = \lambda^2 v$. But $H^2 = H$ and so $H^2 v = Hv = \lambda v$. Thus $\lambda^2 v = \lambda v$, and because $v \neq 0$ this implies $\lambda^2 = \lambda$. So $\lambda \in \{0, 1\}$.

Because the rank of a symmetric matrix is equal to the number of its non-zero eigenvalues, the only full-rank idempotent matrix is the identity matrix. If $H$ is idempotent with rank $m$ then $I_k - H$ is idempotent with rank $k - m$ (because $(I_k - H)(I_k - H) = I_k - H - H + H^2 = I_k - H$).

For any set of linearly independent vectors $v_1, \cdots, v_m \in \mathbb{R}^k$, the linear space $\{a_1 v_1 + \cdots + a_m v_m : a_1, \cdots, a_m \in \mathbb{R}\}$ is the same as the space of vectors $\mathcal{C}(V) = \{Vb : b \in \mathbb{R}^m\}$ where $V = [v_1 : v_2 : \cdots : v_m]$ is the $k \times m$ matrix with columns $v_i$. This space is called the column space of $V$. The matrix $H = V(V^T V)^{-1} V^T$ is idempotent and gives the orthogonal projection matrix onto $\mathcal{C}(V)$ (that means for any $x \in \mathbb{R}^k$, the Euclidean distance $\|x - y\|$, $y \in \mathcal{C}(V)$ is minimized at $y = Hx$ and $Hx$ and $x - Hx$ are orthogonal.).

The converse is also true, every idempotent matrix $H$ of rank $m$ is the orthogonal projection matrix onto $\mathcal{C}(V)$ for some $k \times m$ matrix $V$ with independent columns. This can be seen by writing $H = \sum_{i=1}^m v_i v_i^T$ where $v_i$'s are eigenvectors corresponding to the $m$ non-zero eigenvalues (which must equal 1) of $H$. Take $V = [v_1 : v_2 : \cdots : v_m]$. Then $H = VV^T = V(V^T V)^{-1} V^T$ as $V^T V = I_m$.

**Theorem 6 (quadratic form).** *Suppose $X \sim N_k(0, I_k)$ and $H$ is idempotent with rank $m$. Then $Y = X^T H X \sim \chi^2_m$.*

*Proof.* Write $H = \sum_{i=1}^m v_i v_i^T$ where $v_i$'s are eigenvectors corresponding to the $m$ non-zero eigenvalues of $H$. Then $H = VV^T$ where $V = [v_1 : \cdots : v_m]$. Then $Y = (V^T X)^T (V^T X) = Z^T Z$ where $Z = V^T X \sim N_m(0, I_m)$ as $V^T V = I_m$. Therefore $Y = Z^T Z = Z_1^2 + \cdots + Z_m^2 \sim \chi^2_m$.

**Theorem 7 (a variant of Thm 6).** *Suppose $X \sim N_k(0, H)$ where $H$ is idempotent with rank $m$, then $X^T X \sim \chi^2_m$.*

*Proof.* Let $Y = HZ$ where $Z \sim N_k(0, I_k)$. Then $Y \sim N_k(0, HH^T)$. But because $H$ is idempotent, $HH^T = H$ and hence $Y$ has the same distribution as $X$. So $X^T X$ has the same distribution as $Y^T Y = (HZ)^T (HZ) = Z^T H^T H Z = Z^T H Z$ (because $H^T H = H^2 = H$). But by Theorem 6, $Z^T H Z \sim \chi^2_m$.

**Least-squares estimate and residual sum of squares**

Consider the Gaussian linear model

$$Y_i = z_i^T \beta + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

where $z_i \in \mathbb{R}^p$ are fixed, and $\epsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2)$. The least squares estimate of $\beta$ is $\hat{\beta}_{\text{LS}} = (Z^T Z)^{-1} Z^T Y$ with $Y = (Y_1, \cdots, Y_n)^T$ and $Z$ denoting $n \times p$ matrix with $i$-th row given by

$z_i^T$. Note that $\mathcal{C}(Z)$ is now the column space of $z_{\cdot 1}, z_{\cdot 2}, \cdots, z_{\cdot p}$ where $z_{\cdot j} = (z_{1j}, z_{2j}, \cdots, z_{nj})^T$ is the vector of observations from all $n$ cases on the $j$-th attribute of the input variable. Denote $H = Z(Z^TZ)^{-1}Z^T$. Then $\hat{Y} = Z\hat{\beta}_{\text{LS}} = HY$ is the projection of $Y$ onto $\mathcal{C}(Z)$. This is expected, because in least squares we minimize $\|Y - Z\beta\|^2$ over $\beta$ which is same as minimizing $\|Y - v\|^2$ over $v \in \mathcal{C}(Z)$. Notice that $HZ = Z$ and so $(I_n - H)Z = 0$.

Define the residuals $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - z_i^T\hat{\beta}_{\text{LS}}$ and take $\hat{\epsilon} = (\hat{\epsilon}_1, \cdots, \hat{\epsilon}_n)^T$. Then $\hat{\epsilon} = Y - Z\hat{\beta}_{\text{LS}} = (I_n - H)Y$. An estimate of $\sigma^2$ is given by $s^2 = \frac{1}{n-p}\sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n-p}\hat{\epsilon}^T\hat{\epsilon}$. Here is the counterpart of Theorem 5 for the least-squares.

**Theorem 8.** *Let $Y$, $Z$, $\hat{\beta}_{LS}$ and $H$ be as above. Then*

1. $\hat{\beta}_{LS} \sim N_p(\beta, \sigma^2(Z^TZ)^{-1})$.
2. $\hat{\epsilon} \sim N_n(0, \sigma^2(I_n - H))$
3. $\hat{\beta}_{LS}$ *and* $\hat{\epsilon}$ *are independent.*
4. $\frac{1}{\sigma^2}\hat{\epsilon}^T\hat{\epsilon} \sim \chi^2_{n-p}$.

*Proof.* From what we have discussed above,

$$\begin{aligned}
\begin{pmatrix}\hat{\beta}_{\text{LS}} \\ \hat{\epsilon}\end{pmatrix} &= \begin{pmatrix}(Z^TZ)^{-1}Z^T \\ I_n - H\end{pmatrix} Y \\
&\sim N_{n+p}\left(\begin{pmatrix}(Z^TZ)^{-1}Z^T \\ I_n - H\end{pmatrix} Z\beta, \begin{pmatrix}(Z^TZ)^{-1}Z^T \\ I_n - H\end{pmatrix} (\sigma^2 I_n) \begin{pmatrix}Z(Z^TZ)^{-1} & I_n - H\end{pmatrix}\right) \\
&= N_{n+p}\left(\begin{pmatrix}(Z^TZ)^{-1}Z^TZ\beta \\ (I_n - H)Z\beta\end{pmatrix}, \begin{pmatrix}\sigma^2(Z^TZ)^{-1}Z^TZ(Z^TZ)^{-1} & \sigma^2(Z^TZ)^{-1}Z^T(I_n - h) \\ \sigma^2(I_n - H)Z(Z^TZ)^{-1} & \sigma^2(I_n - H)(I_n - H)\end{pmatrix}\right) \\
&= N_{n+p}\left(\begin{pmatrix}\beta \\ 0_{n\times 1}\end{pmatrix}, \begin{pmatrix}\sigma^2(Z^TZ)^{-1} & 0_{p\times n} \\ 0_{n\times p} & \sigma^2(I_n - H)\end{pmatrix}\right)
\end{aligned}$$

because $HZ = Z$, $(I_n - H)Z = 0_{n\times p}$ and $H^2 = H$ (and so $(I_n - H)^2 = I_n - H$). This proves assertions 1, 2 and 3 (see theorem 4 above). To prove the last assertion, note that $\frac{1}{\sigma}\hat{\epsilon} \sim N_{n-p}(0, I_n - H)$ and $I_n - H$ is idempotent. So, by Theorem 7, $\frac{1}{\sigma^2}\hat{\epsilon}^T\hat{\epsilon} \sim \chi^2_{n-p}$.

**Theorem 9 (confidence intervals).** *Let $Y$, $Z$, $\hat{\beta}_{LS}$ and $H$ be as above. Let $\alpha$ be any number in $(0, 1)$. Then*

1. *For any vector $a \in \mathbb{R}^p$, $a \neq 0_{p\times 1}$, $a^T\hat{\beta}_{LS} \mp z_{n-p}(\alpha)\frac{s}{\sqrt{n_a}}$ is a $(1-\alpha)$ confidence interval for $a^T\beta$, where $n_a = \frac{1}{a^T(Z^TZ)^{-1}a} > 0$.*
2. *In particular, for every $j = 1, 2, \cdots, p$, $\hat{\beta}_{LS,j} \mp z_{n-p}(\alpha)\frac{s}{\sqrt{n_{jj}}}$ is a $(1-\alpha)$ confidence interval for $\beta_j$, where $n_{jj}$ is the inverse of the $j$-th diagonal entry of $(Z^TZ)^{-1}$.*

*Proof.* The second assertion follows from the first with $a = (0, \cdots, 0, 1, 0, \cdots, 0)^T$, where the 1 occurs at the $j$-th position. To prove the first assertion, note that by Theorem 8,

1. $a^T\hat{\beta}_{\text{LS}} \sim N(a^T\beta, \sigma^2 a^T(Z^TZ)^{-1}a)$ and hence $W = \frac{a^T\hat{\beta}_{\text{LS}} - a^T\beta}{\sigma\sqrt{a^T(Z^TZ)^{-1}a}} \sim N(0, 1)$.

2. $V = \frac{(n-p)s^2}{\sigma^2} = \frac{1}{\sigma^2}\hat{\epsilon}^T\hat{\epsilon} \sim \chi^2_{n-p}$.

3. $W$ and $V$ are independent.

Hence $T = \frac{W}{\sqrt{V/(n-p)}} = \frac{a^T\hat{\beta}_{\text{LS}} - a^T\beta}{s\sqrt{a^T(Z^TZ)^{-1}a}} = \frac{a^T\hat{\beta}_{\text{LS}} - a^T\beta}{s/\sqrt{n_a}} \sim t_{n-p}$. Therefore,

$$P\left(a^T\beta \in a^T\hat{\beta}_{\text{LS}} \mp z_{n-p}(\alpha)\frac{s}{\sqrt{n_a}}\right) = P(-z_{n-p}(\alpha) \leq T \leq z_{n-p}(\alpha)) = 1 - \alpha.$$