Multivariate Statistics

Lecture 08

Fudan University

Outline

Outline

1. Distribution of T^2-Statistic

2. Uses of T^2-Statistic

Distribution of T^2-Statistic

Theorem 1

Let $T^2 = \mathbf{y}^\top \mathbf{S}^{-1}\mathbf{y}$, where $\mathbf{y}$ is distributed according to $\mathcal{N}_p(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ and $n\mathbf{S}$ is independently distributed as $\sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$ with $\mathbf{z}_1, \dots, \mathbf{z}_n$ independent, each with distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Then the random variable

$$\frac{T^2}{n} \cdot \frac{n-p+1}{p}$$

is distributed as a noncentral F-distribution with p and $n-p+1$ degrees of freedom and noncentrality parameter $\boldsymbol{\nu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\nu}$. If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is central F.

In the example of likelihood ratio criterion, we consider the special case of $\mathbf{y} = \sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$, $\boldsymbol{\nu} = \sqrt{N}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ and $n = N - 1$.

Distribution of T^2-Statistic

Corollary 1

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let

$$T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0).$$

The distribution of

$$\frac{T^2}{N-1} \cdot \frac{N-p}{p}.$$

is noncentral F with p and $N-p$ degrees of freedom and noncentrality parameter $N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$. If $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ then the F-distribution is central.

Distribution of T^2-Statistic

Theorem 2

Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are independent with $\mathbf{y}_\alpha$ distributed according to $\mathcal{N}(\mathbf{\Gamma w}_\alpha, \mathbf{\Phi})$, where $\mathbf{w}_\alpha$ is an r-component vector. Let $\mathbf{H} = \sum_{\alpha=1}^{m} \mathbf{w}_\alpha \mathbf{w}_\alpha^\top$ assumed non-singular, $\mathbf{G} = \sum_{\alpha=1}^{m} \mathbf{y}_\alpha \mathbf{w}_\alpha^\top \mathbf{H}^{-1}$ and

$$\mathbf{C} = \sum_{\alpha=1}^{m} (\mathbf{y}_\alpha - \mathbf{G}\mathbf{w}_\alpha)(\mathbf{y}_\alpha - \mathbf{G}\mathbf{w}_\alpha)^\top = \sum_{\alpha=1}^{m} \mathbf{y}_\alpha \mathbf{y}_\alpha^\top - \mathbf{G}\mathbf{H}\mathbf{G}^\top.$$

Then $\mathbf{C}$ is distributed as

$$\sum_{\alpha=1}^{m-r} \mathbf{u}_\alpha \mathbf{u}_\alpha^\top$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_{m-r}$ are independently distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{\Phi})$ independently of $\mathbf{G}$.

Distribution of T^2-Statistic

For large samples the distribution of T^2 given this corollary is approximately valid even if the parent distribution is not normal.

Theorem 3

Let $x_1, x_2, \ldots$ be a sequence of independently identically distributed random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let

$$\hat{\mathbf{x}}_N = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha, \qquad \hat{\mathbf{S}}_N = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top$$

and

$$T_N^2 = N(\bar{\mathbf{x}}_N - \boldsymbol{\mu}_0)^\top \mathbf{S}_N^{-1}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}_0).$$

Then the limiting distribution of T_N^2 as $N \to \infty$ is the χ^2-distribution with p degrees of freedom if $\boldsymbol{\mu} = \boldsymbol{\mu}_0$.

Distribution of T^2-Statistic

When the null hypothesis is true ($\boldsymbol{\mu}_0 = \boldsymbol{\mu}$), the likelihood ratio criterion holds that

$$\lambda^{\frac{2}{N}} = \frac{1}{1 + T^2/(N-1)} = \frac{1}{1 + T^2/n},$$

where $T^2 =$ and $n = N - 1$.

Then T^2 is distributed according to central F-distribution with degree of freedom p and $n - 1 - p$:

$$\begin{aligned} &\frac{T^2}{n} \cdot \frac{n-p+1}{p} \sim \frac{\chi^2(p)/p}{\chi^2(n-1-p)/(n-1-p)} \\ \Longrightarrow &\frac{T^2}{n} \sim \frac{\chi^2(p)}{\chi^2(n-1-p)} \\ \Longrightarrow &\lambda^{\frac{2}{N}} \sim \frac{\chi^2(n-1-p)}{\chi^2(n-1-p) + \chi^2(p)} \end{aligned}$$

Distribution of T^2-Statistic

Theorem 4

Let u be distributed according to the χ^2-distribution with a degrees of freedom and w be distributed according to the χ^2-distribution with b degrees of freedom. The density of $v = u/(u+w)$, when u and w are independent is

$$\frac{1}{B\left(\frac{a}{2}, \frac{b}{2}\right)} v^{\frac{a}{2}-1}(1-v)^{\frac{b}{2}-1}, \tag{1}$$

where $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$.

The function (1) is the density of beta distribution with parameters $a/2$ and $b/2$.

Outline

Testing the Hypothesis for the Mean

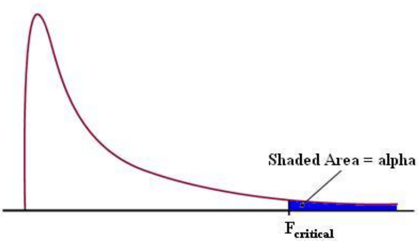
The likelihood ratio test of the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ on the basis of a sample of N from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined by the critical region

$$T^2 \geq T_0^2,$$

where $T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$.

If the significance level is α, then

$$T_0^2 = \frac{(N-1)p}{N-p} F_{p,N-p}(\alpha) \triangleq T_{p,N-1}^2(\alpha).$$



A Confidence Region for the Mean Vector

The probability of drawing a sample of N from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with sample mean $\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$ such that

$$N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq T^2_{p,N-1}(\alpha).$$

is $1 - \alpha$.

The set

$$\left\{\mathbf{m} : N(\bar{\mathbf{x}} - \mathbf{m})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \mathbf{m}) \leq T^2_{p,N-1}(\alpha)\right\}$$

corresponds to the interior and boundary of an ellipsoid. We state that $\boldsymbol{\mu}$ lies within this ellipsoid with confidence $1 - \alpha$.

Two-Sample Problems (Unknown Covariance)

Suppose $\mathbf{y}_1^{(i)}, \ldots, \mathbf{y}_{N_i}^{(i)}$ is a sample from $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$ for $i = 1, 2$. We wish to test the null hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$.

1. For $i = 1, 2$, we have

$$\bar{\mathbf{y}}^{(i)} = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} \mathbf{y}_\alpha^{(i)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(i)}, \frac{1}{N_i}\boldsymbol{\Sigma}\right).$$

2. Since

$$\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{y}}^{(1)} \\ \bar{\mathbf{y}}^{(2)} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \bar{\mathbf{y}}^{(1)} \\ \bar{\mathbf{y}}^{(2)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \begin{bmatrix} \frac{1}{N_1}\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \frac{1}{N_2}\boldsymbol{\Sigma} \end{bmatrix} \right),$$

we have

$$\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}, \left(\frac{1}{N_1} + \frac{1}{N_2}\right)\boldsymbol{\Sigma}\right).$$

Two-Sample Problems (Unknown Covariance)

Under the null hypothesis, we have

$$\sqrt{N_1 N_2/(N_1+N_2)}\left(\bar{\mathbf{y}}^{(1)}-\bar{\mathbf{y}}^{(2)}\right)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}).$$

Let

$$\begin{aligned}\mathbf{S}=\frac{1}{N_1+N_2-2}\Bigg(&\sum_{\alpha=1}^{N_1}\left(\mathbf{y}_\alpha^{(1)}-\bar{\mathbf{y}}^{(1)}\right)\left(\mathbf{y}_\alpha^{(1)}-\bar{\mathbf{y}}^{(1)}\right)^\top\\&+\sum_{\alpha=1}^{N_2}\left(\mathbf{y}_\alpha^{(2)}-\bar{\mathbf{y}}^{(2)}\right)\left(\mathbf{y}_\alpha^{(2)}-\bar{\mathbf{y}}^{(2)}\right)^\top\Bigg),\end{aligned}$$

then

$$(N_1+N_2-2)\mathbf{S}=\sum_{\alpha=1}^{N_1+N_2-2}\mathbf{z}_\alpha\mathbf{z}_\alpha^\top,$$

where $\mathbf{z}_\alpha$ are independent and $\mathbf{z}_\alpha\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$.

Two-Sample Problems (Unknown Covariance)

Let

$$T^2 = \frac{N_1 N_2}{N_1 + N_2}\left(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}\right)^\top \mathbf{S}^{-1}\left(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}\right),$$

then

$$\frac{T^2}{N_1 + N_2 - 2} \cdot \frac{N_1 + N_2 - p - 1}{p}$$

is distributed according to central F-distribution with p and $N_1 + N_2 - p - 1$ degrees of freedom.

The critical region is

$$T^2 \geq \frac{(N_1 + N_2 - 2)p}{N_1 + N_2 - p - 1} F_{p, N_1+N_2-p-1}(\alpha)$$

with significance level α.

Two-Sample Problems (Unknown Covariance)

The probability of

$$
\begin{aligned}
T^2 =& \frac{N_1 N_2}{N_1 + N_2} \left(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)} \right)^\top \mathbf{S}^{-1} \left(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)} \right) \\
\leq & \frac{(N_1 + N_2 - 2)p}{N_1 + N_2 - p - 1} F_{p, N_1 + N_2 - p - 1}(\alpha)
\end{aligned}
$$

is $1 - \alpha$.

A confidence region for $\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$ with confidence level $1 - \alpha$ is the set of vectors $\mathbf{m}$ satisfying

$$
\begin{aligned}
& \frac{N_1 N_2}{N_1 + N_2} \left(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)} - \mathbf{m} \right)^\top \mathbf{S}^{-1} \left(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)} - \mathbf{m} \right) \\
\leq & \frac{(N_1 + N_2 - 2)p}{N_1 + N_2 - p - 1} F_{p, N_1 + N_2 - p - 1}(\alpha).
\end{aligned}
$$

A Problem of Several Samples

There is a theoretical reason for believing the gene structures of three species of Iris virginica to be such that the mean vectors of the three populations are related as

$$3\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(3)} + 2\boldsymbol{\mu}^{(2)},$$

where $\boldsymbol{\mu}^{(i)}$ is the mean vector of the i-th population.

A Problem of Several Samples

Let $\left\{\mathbf{x}_\alpha^{(i)}\right\}$ for $\alpha = 1, \ldots, N_i$, $i = 1, \ldots, q$ be independent samples from $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i = 1, \ldots, q$, respectively. Let us test the hypothesis

$$H : \sum_{i=1}^{q} \beta_i \boldsymbol{\mu}^{(i)} = \boldsymbol{\mu}.$$

where $\beta_1, \ldots, \beta_q$ are given scalars and $\boldsymbol{\mu}$ is a given vector.

A Problem of Several Samples

The criterion is

$$T^2 = c\left(\sum_{i=1}^{q} \beta_i \bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}\right) \mathbf{S}^{-1} \left(\sum_{i=1}^{q} \beta_i \bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}\right)^{\top}$$

where

$$\bar{\mathbf{x}}^{(i)} = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} \mathbf{x}_{\alpha}^{(i)}, \qquad c = \left(\sum_{i=1}^{q} \frac{\beta_i^2}{N_i}\right)^{-1}$$

and

$$\mathbf{S} = \frac{1}{\sum_{i=1}^{q} N_i - q} \sum_{i=1}^{q} \sum_{\alpha=1}^{N_i} \left(\mathbf{x}_{\alpha}^{(i)} - \bar{\mathbf{x}}^{(i)}\right)\left(\mathbf{x}_{\alpha}^{(i)} - \bar{\mathbf{x}}^{(i)}\right)^{\top}.$$

This T^2 has the T^2-distribution with $\sum_{i=1}^{q} N_i - q$ degrees of freedom.

A Problem of Symmetry

Consider testing the hypothesis

$$H : \mu_1 = \mu_2 = \cdots = \mu_p$$

on the basis of sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

A Problem of Symmetry

Let $\mathbf{C}$ be any $(p-1) \times p$ matrix of rank $p-1$ such that

$$\mathbf{C}\mathbf{1}_p = \mathbf{0}_{p-1}.$$

Then we have

$$\mathbf{y}_\alpha = \mathbf{C}\mathbf{x}_\alpha \sim \mathcal{N}\left(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top\right)$$

and the hypothesis H is equivalent to $\mathbf{C}\boldsymbol{\mu} = \mathbf{0}_{p-1}$ (why?).

A Problem of Symmetry

We can construct the T^2 statistic

$$T^2 = N\bar{\mathbf{y}}^\top \mathbf{S}^{-1}\bar{\mathbf{y}}$$

where

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{y}_\alpha = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{C}\mathbf{x}_\alpha = \mathbf{C}\bar{\mathbf{x}}$$

$$\mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})^\top = \frac{1}{N-1}\sum_{\alpha=1}^{N}\mathbf{C}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top\mathbf{C}^\top.$$

Two-Sample Problems (Unequal Covariance)

Let $\{\mathbf{x}_\alpha^{(i)}\}$ for $\alpha = 1, \ldots, N_i$, $i = 1, \ldots, q$ be independent samples from $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_i)$ for $i = 1, 2$, respectively. We wish to test the hypothesis

$$H : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}.$$

We cannot use the technique in the case of equal covariance, because

$$\sum_{\alpha=1}^{N_1} \left(\mathbf{x}_\alpha^{(1)} - \bar{\mathbf{x}}^{(1)}\right)\left(\mathbf{x}_\alpha^{(1)} - \bar{\mathbf{x}}^{(1)}\right)^\top + \sum_{\alpha=1}^{N_2} \left(\mathbf{x}_\alpha^{(2)} - \bar{\mathbf{x}}^{(2)}\right)\left(\mathbf{x}_\alpha^{(2)} - \bar{\mathbf{x}}^{(2)}\right)^\top$$

does not correspond to normal distributed variables $\mathbf{z}_\alpha$ with covariance

$$\frac{1}{N_1}\boldsymbol{\Sigma}_1 + \frac{1}{N_2}\boldsymbol{\Sigma}_2.$$

Two-Sample Problems ($N_1 = N_2$)

If $N_1 = N_2 = N$, we can use the T^2-test in an obvious way.

1. Let $\mathbf{y}_\alpha = \mathbf{x}_\alpha^{(1)} - \mathbf{x}_\alpha^{(2)}$, then $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent and
$$\mathbf{y}_\alpha \sim \mathcal{N}\big(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\big).$$
2. Define
$$\begin{aligned} \bar{\mathbf{y}} &= \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{y}_\alpha = \bar{\mathbf{x}}_\alpha^{(1)} - \bar{\mathbf{x}}_\alpha^{(2)}, \\ (N-1)\mathbf{S} &= \sum_{\alpha=1}^{N} (\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})^\top \\ &= \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha^{(1)} - \mathbf{x}_\alpha^{(2)} - \bar{\mathbf{x}}_\alpha^{(1)} + \bar{\mathbf{x}}_\alpha^{(2)})(\mathbf{x}_\alpha^{(1)} - \mathbf{x}_\alpha^{(2)} - \bar{\mathbf{x}}_\alpha^{(1)} + \bar{\mathbf{x}}_\alpha^{(2)})^\top. \end{aligned}$$
3. Then $T^2 = N\bar{\mathbf{y}}^\top \mathbf{S}^{-1}\bar{\mathbf{y}}$ is suitable for testing the hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ and has the T^2-distribution with $N-1$ degrees of freedom.

Two-Sample Problems ($N_1 \neq N_2$)

For the case of $N_1 \neq N_2$, we let $N_1 < N_2$ and define

$$\mathbf{y}_\alpha = \mathbf{x}_\alpha^{(1)} - \sqrt{\frac{N_1}{N_2}}\mathbf{x}_\alpha^{(2)} + \frac{1}{\sqrt{N_1 N_2}}\sum_{\beta=1}^{N_1}\mathbf{x}_\beta^{(2)} - \frac{1}{N_2}\sum_{\gamma=1}^{N_2}\mathbf{x}_\gamma^{(2)}$$

for $\alpha = 1, \ldots, N_1$. We have

$$\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$$

and

$$\mathrm{Cov}(\mathbf{y}_\alpha, \mathbf{y}_{\alpha'}) = \begin{cases} \boldsymbol{\Sigma}_1 + \frac{N_1}{N_2}\boldsymbol{\Sigma}_2, & \alpha = \alpha', \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

Two-Sample Problems ($N_1 \neq N_2$)

We test $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ by using

$$T^2 = N_1 \bar{\mathbf{y}}^\top \mathbf{S}^{-1} \bar{\mathbf{y}},$$

which has T^2-distribution with $N_1 - 1$ degrees of freedom, where

$$\begin{aligned}
\bar{\mathbf{y}} =& \frac{1}{N_1} \sum_{\alpha=1}^{N_1} \mathbf{y}_\alpha = \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}, \\
\mathbf{S} =& \frac{1}{N_1 - 1} \sum_{\alpha=1}^{N_1} (\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})^\top.
\end{aligned}$$

Lemma 3

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be independent samples from $\mathcal{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$ for $i = 1, \ldots, m$. Define

$$\mathbf{z}_1 = \sum_{\alpha=1}^{N} a_\alpha \mathbf{x}_\alpha \quad \text{and} \quad \mathbf{z}_2 = \sum_{\alpha=1}^{N} b_\alpha \mathbf{x}_\alpha,$$

then

$$\mathrm{Cov}(\mathbf{z}_1, \mathbf{z}_2) = \sum_{\alpha=1}^{N} a_\alpha b_\alpha \boldsymbol{\Sigma}_\alpha.$$