Multivariate Statistics

Lecture 14

Fudan University

Outline

Outline

Matrix Operations: Trace

The trace of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted $\operatorname{tr}(\mathbf{A})$, is the sum of diagonal elements in the matrix:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

The trace has the following properties

1. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, we have $\operatorname{tr}(\mathbf{A}) = \operatorname{tr}\left(\mathbf{A}^\top\right)$.
2. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, we have $\operatorname{tr}\left(\mathbf{A}^\top \mathbf{A}\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2$.
3. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$, $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$, we have $\operatorname{tr}(c_1\mathbf{A} + c_2\mathbf{B}) = c_1\operatorname{tr}(\mathbf{A}) + c_2\operatorname{tr}(\mathbf{B})$.
4. For $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{AB}$ is square, $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$.
5. For $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ such that $\mathbf{ABC}$ is square, we have $\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA}) = \operatorname{tr}(\mathbf{CAB})$.

Orthogonality

A nice property of orthogonal matrices is that operating on a vector with an orthogonal matrix will not change its Euclidean norm, that is

$$\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$$

for any $\mathbf{x} \in \mathbb{R}^n$ and orthogonal $\mathbf{U} \in \mathbb{R}^n$.

Orthogonal matrices can be used to represent a rotation.

A basis $\mathbf{x}_1, \ldots, \mathbf{x}_k$ of a subspace $\mathcal{W}$ of $\mathbb{R}^n$ is called orthonormal basis if all the elements have norm one and are orthogonal to one another.

In particular, if $\mathbf{A} \in \mathbb{R}^n$ is an orthogonal matrix then the columns of $\mathbf{A}$ form an orthogonal basis of $\mathbb{R}^n$.

Spectral Decomposition Theorem

Any symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be written as

$$\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^\top = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\top$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix elements of its main diagonal are $\lambda_1, \dots, \lambda_n$ and $\mathbf{X}$ is an orthogonal matrix whose columns are corresponding to standardized eigenvectors of $\mathbf{A}$.

Proof Sketch

1. The eigenvalues and eigenvectors of $\mathbf{A}$ are real.
2. Two eigenvectors corresponding to distinct eigenvalues of $\mathbf{A}$ are orthogonal.
3. If λ_i is an eigenvector of $\mathbf{A}$ with $m \geq 2$ algebra multiplicity, we can find m orthogonal eigenvectors in its eigenspace.

Schur Complement

Given matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{C} \in \mathbb{R}^{q\times p}$ and $\mathbf{D} \in \mathbb{R}^{q\times q}$ and suppose $\mathbf{D}$ is non-singular. Let

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}.$$

Then the Schur complement of the block $\mathbf{D}$ for $\mathbf{M}$ is

$$\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \in \mathbb{R}^{p\times p}.$$

Then we can decompose the matrix $\mathbf{M}$ as

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$$

and the inverse of $\mathbf{M}$ can be written as

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Cholesky Factorization

The symmetric positive-definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has the decomposition of the form

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top$$

where $\mathbf{L} \in \mathbb{R}^{n \times n}$ is a lower triangular matrix with real and positive diagonal entries such that

$$\mathbf{L} = \begin{bmatrix} + & 0 & \cdots & 0 \\ \cdot & + & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdots & + \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

The Gradient

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a smooth function that takes as input a matrix $\mathbf{X}$ of size $m \times n$ and returns a real value. Then the gradient of f with respect to $\mathbf{X}$ is

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Some Basic Results

1. For $\mathbf{X} \in \mathbb{R}^{m\times n}$, we have $\dfrac{\partial(f(\mathbf{X}) + g(\mathbf{X}))}{\partial \mathbf{X}} = \dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}} + \dfrac{\partial g(\mathbf{X})}{\partial \mathbf{X}}$.
2. For $\mathbf{X} \in \mathbb{R}^{m\times n}$ and $t \in \mathbb{R}$, we have $\dfrac{\partial t f(\mathbf{X})}{\partial \mathbf{X}} = t\dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$.
3. For $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{m\times n}$, we have $\dfrac{\partial \mathrm{tr}(\mathbf{A}^\top \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}$.
4. For $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^{n}$, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

   If $\mathbf{A}$ is symmetric, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

We can find more results in the matrix cookbook:
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

The Gradient of $\ln\det(\cdot)$

Consider the function $f(\mathbf{A}) = \ln(\det(\mathbf{A}))$ whose domain is $n \times n$ positive definite matrices. Then we have

$$\nabla f(\mathbf{A}) = (\mathbf{A}^{-1})^\top.$$

We usually write $\nabla f(\mathbf{A}) = \mathbf{A}^{-1}$ by further assuming the domain of f is symmetric.

This also can be viewed as the extension of $(\ln a)' = a^{-1}$ for $a > 0$.

Statistical Independence

The statistical independence of X and Y implies

$$
\begin{aligned}
&\Pr\{x_1 \le X \le x_2, y_1 \le Y \le y_2\} \\
={}& \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(u, v)\, \mathrm{d}u\, \mathrm{d}v \\
={}& \int_{y_1}^{y_2} f(u)\, \mathrm{d}u \int_{x_1}^{x_2} g(v)\, \mathrm{d}v \\
={}& \Pr\{x_1 \le X \le x_2\} \Pr\{y_1 \le Y \le y_2\}.
\end{aligned}
$$

Note that we say X and Y are uncorrelated if

$$
\begin{aligned}
& \mathrm{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = 0 \\
\iff\ & \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}
$$

Independent $\neq$ Uncorrelated

Note that

X are Y are independent implies X are Y uncorrelated.

However,

X are Y are uncorrelated do NOT implies X are Y are independent.

Transformation of Variables

Let the density of p dimensional random vector $\mathbf{x} = [x_1, \ldots, x_p]^\top$ be $f(\mathbf{x})$.

Consider the random vector p dimensional random vector $\mathbf{y} = [y_1, \ldots, y_p]^\top$ such that $y_i = u_i(\mathbf{x})$ for $i = 1, \ldots, p$. Let the density function of $\mathbf{y}$ be $g(\mathbf{y})$.

Assume the transformation $\mathbf{u}(\mathbf{x}) = [u_1(\mathbf{x}), \ldots, u_p(\mathbf{x})]^\top : \mathbb{R}^p \to \mathbb{R}^p$ from the space of $\mathbf{x}$ to the space of $\mathbf{y}$ is smooth and one-to-one.

Then we have $f(\mathbf{x}) = g(\mathbf{u}(\mathbf{x}))|\det(\mathbf{J}(\mathbf{x}))|$ where

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial u_1(\mathbf{x})}{\partial x_1} & \dfrac{\partial u_1(\mathbf{x})}{x_2} & \cdots & \dfrac{\partial u_1(\mathbf{x})}{\partial x_p} \\ \dfrac{\partial u_2(\mathbf{x})}{\partial x_1} & \dfrac{\partial u_2(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial u_2(\mathbf{x})}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial u_p(\mathbf{x})}{\partial x_1} & \dfrac{\partial u_p(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial u_p(\mathbf{x})}{\partial x_p} \end{bmatrix}.$$

Transformation of Variables

Similarly, we also have $g(\mathbf{y}) = f(\mathbf{u}^{-1}(\mathbf{y}))|\det(\mathbf{J}^{-1}(\mathbf{y}))|$ where

$$\mathbf{J}^{-1}(\mathbf{y}) = \begin{bmatrix} \frac{\partial u_1^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial u_1^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial u_1^{-1}(\mathbf{y})}{\partial y_p} \\ \frac{\partial u_2^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial u_2^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial u_2^{-1}(\mathbf{y})}{\partial y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_p^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial u_p^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial u_p^{-1}(\mathbf{y})}{\partial y_p} \end{bmatrix}.$$

Linear Transformation for Random Vector

Lemma

1. If $\mathbf{Z}$ is an $m \times n$ random matrix, $\mathbf{D}$ is an $l \times m$ real matrix, $\mathbf{E}$ is an $n \times q$ real matrix, and $\mathbf{F}$ is an $l \times q$ real matrix, then

$$\mathbb{E}[\mathbf{DZE} + \mathbf{F}] = \mathbf{D}\mathbb{E}[\mathbf{Z}]\mathbf{E} + \mathbf{F}.$$

2. If $\mathbf{y} = \mathbf{Dx} + \mathbf{f} \in \mathbb{R}^l$, where $\mathbf{D}$ is an $l \times m$ real matrix, $\mathbf{x} \in \mathbb{R}^m$ is a random vector, then

$$\mathbb{E}[\mathbf{y}] = \mathbf{D}\mathbb{E}[\mathbf{x}] + \mathbf{f}$$

and

$$\mathrm{Cov}[\mathbf{y}] = \mathbf{D}\mathrm{Cov}[\mathbf{x}]\mathbf{D}^\top.$$

Outline

If the density of a p-dimensional random vector $\mathbf{x}$ is

$$K \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top \mathbf{A}(\mathbf{x}-\mathbf{b})\right),$$

where $\mathbf{A} \in \mathbb{R}^{p\times p}$ is symmetric positive definite. Then the expectation of $\mathbf{x}$ is $\mathbf{b}$ and its covariance matrix is $\mathbf{A}^{-1}$.

Conversely, given a vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and a positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, there is a multivariate normal density

$$n(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

More General Linear Transformation

Theorem 5

Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$\mathbf{z} = \mathbf{D}\mathbf{x}$$

is distributed according to $\mathcal{N}_q(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top)$ for any $\mathbf{D} \in \mathbb{R}^{q \times p}$.

We do not require additional assumptions on $\mathbf{D}$ or $\boldsymbol{\Sigma}$.

Multivariate Normal Distribution (Marginal Distribution)

Because the numbering of the components of $\mathbf{x}$ is arbitrary, we can state the following theorem:

Theorem 3

If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ \mathbf{0}$, the marginal distribution of any set of components of $\mathbf{x}$ is multivariate normal with means, variances, and covariances obtained by taking the corresponding components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.

Multivariate Normal Distribution (Conditional Distribution)

Let $\mathbf{x}$ be distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ \mathbf{0}$. Let us partition

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \quad \text{with } \mathbf{x}^{(1)} \in \mathbb{R}^q \text{ and } \mathbf{x}^{(2)} \in \mathbb{R}^{p-q}.$$

The conditional density of $\mathbf{x}^{(1)}$ given that $\mathbf{x}^{(2)}$ is

$$\begin{aligned} & f(\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}) = \frac{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{f(\mathbf{x}^{(2)})} \\ =& \frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma}_{11.2})}} \exp\left(-\frac{1}{2}\left(\mathbf{x}^{(11.2)} - \boldsymbol{\mu}^{(11.2)}\right)^\top \boldsymbol{\Sigma}_{11.2}^{-1}\left(\mathbf{x}^{(11.2)} - \boldsymbol{\mu}^{(11.2)}\right)\right), \end{aligned}$$

where

$$\begin{aligned} \mathbf{x}^{(11.2)} =& \mathbf{x}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}^{(2)}, \\ \boldsymbol{\mu}^{(11.2)} =& \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}, \\ \boldsymbol{\Sigma}_{11.2} =& \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

Correlation Coefficient

Recall that for random vector $\mathbf{x} = [x_1, x_2, \ldots, x_p]^\top$, we define the covariance matrix as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

and the correlation coefficient between x_i and x_j as (suppose $\mathbf{\Sigma} \succ \mathbf{0}$)

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}.$$

Characteristic Function

The characteristic function of a p-dimensional random vector $\mathbf{x}$ is

$$\phi(\mathbf{t}) = \mathbb{E}\left[\exp(\mathrm{i}\,\mathbf{t}^\top \mathbf{x})\right]$$

defined for every real vector $\mathbf{t} \in \mathbb{R}^p$.

For the complex-valued function $g(z)$ be written as

$$g(z) = g_1(z) + \mathrm{i}\, g_2(z),$$

where $g_1(z)$ and $g_2(z)$ are real-valued, the expected value of $g(z)$ is

$$\mathbb{E}[g(z)] = \mathbb{E}[g_1(z)] + \mathrm{i}\,\mathbb{E}[g_2(z)].$$

Characteristic Function

Theorem 2

The characteristic function of $\mathbf{x}$ distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\phi(\mathbf{t}) = \exp\left(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}\mathbf{t}\right).$$

for every $\mathbf{t} \in \mathbb{R}^p$.

Sketch of the proof

1. The characteristic function of $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ is $\phi_0(\mathbf{t}) = \exp\left(-\frac{1}{2}\mathbf{t}^\top \mathbf{t}\right)$.
2. For $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have $\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu}$ such that $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$.
3. Using $\phi_0(\mathbf{t})$ to present the characteristic function of $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Characteristic Function and Moments

If the n-th moment of random variable x, denoted by $\mathbb{E}[x^n]$, exists and is finite, then its characteristic function is n times continuously differentiable and

$$\mathbb{E}[x^n] = \frac{1}{\mathrm{i}^n} \frac{\mathrm{d}^n \phi(t)}{\mathrm{d}t^n}\bigg|_{t=0},$$

which is because of

$$\begin{aligned}
\frac{\mathrm{d}^n \phi(t)}{\mathrm{d}t^n} &= \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbb{E}\left[\exp(\mathrm{i}\,tx)\right] \\
&= \mathbb{E}\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n}\exp(\mathrm{i}\,tx)\right] \\
&= \mathbb{E}\left[(\mathrm{i}\,x)^n \exp(\mathrm{i}\,tx)\right] \\
&= \mathrm{i}^n\,\mathbb{E}\left[x^n \exp(\mathrm{i}\,tx)\right].
\end{aligned}$$

Outline

The Maximum Likelihood Estimators

Given a sample of (vector) observations from a p-variate (non-singular) normal distribution, we ask for estimators of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the distribution.

Suppose our sample of N observations on the $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, which are distributed according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $N > p$. The likelihood function is

$$
\begin{aligned}
L &= \prod_{\alpha=1}^{N} n(\mathbf{x}_\alpha \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \frac{1}{(2\pi)^{\frac{pN}{2}} (\det(\boldsymbol{\Sigma}))^{\frac{N}{2}}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right].
\end{aligned}
$$

The Maximum Likelihood Estimators

Theorem 6

If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ constitute a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $p < N$, the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top$$

respectively.

Lemma 1

If $\mathbf{D} \in \mathbb{R}^{p\times p}$ is positive definite, the maximum of

$$f(\mathbf{G}) = -N \ln \det(\mathbf{G}) - \mathrm{tr}(\mathbf{G}^{-1}\mathbf{D})$$

with respect to positive definite matrices $\mathbf{G}$ exists, occurs at $\mathbf{G} = \frac{1}{N}\mathbf{D}$.

The Maximum Likelihood Estimators

Corollary 2

If on the basis of a given sample $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are maximum likelihood estimators of the parameters $\theta_1, \ldots, \theta_m$ of a distribution, then $\phi_1(\hat{\theta}_1, \ldots, \hat{\theta}_m), \ldots, \phi_m(\hat{\theta}_1, \ldots, \hat{\theta}_m)$ are maximum likelihood estimator of $\phi_1(\theta_1, \ldots, \theta_m), \ldots, \phi_m(\theta_1, \ldots, \theta_m)$ if the transformation from $\theta_1, \ldots, \theta_m$ to $\phi_1, \ldots, \phi_m$ is one-to-one. If the estimators of $\theta_1, \ldots, \theta_m$ are unique, then the estimators of $\theta_1, \ldots, \theta_m$ are unique.

Corollary 3

If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ constitutes a sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $\rho_{ij} = \sigma_{ij}/(\sigma_i \sigma_j)$. Then the maximum likelihood estimator of ρ_{ij} is

$$\hat{\rho}_{ij} = \frac{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)^2}\sqrt{\sum_{\alpha=1}^{N}(x_{j\alpha} - \bar{x}_j)^2}}$$

Distribution Theory

Theorem 2

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be independent, each distributed according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the mean of the sample

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N} \sum_{\alpha=1}^{N} \mathbf{x}_\alpha$$

is distributed according to $\mathcal{N}(\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma})$ and independent of

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top.$$

Additionally, we have $N\hat{\boldsymbol{\Sigma}} = \sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$, where $\mathbf{z}_\alpha \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for $\alpha = 1, \ldots, N-1$, and $\mathbf{z}_1, \ldots, \mathbf{z}_{N-1}$ are independent.

Distribution Theory

Theorem 2

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be independent, each distributed according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the mean of the sample

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha$$

is distributed according to $\mathcal{N}(\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma})$ and independent of

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top.$$

Additionally, we have $N\hat{\boldsymbol{\Sigma}} = \sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$, where $\mathbf{z}_\alpha \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for $\alpha = 1, \ldots, N-1$, and $\mathbf{z}_1, \ldots, \mathbf{z}_{N-1}$ are independent.

Consider the result of MLE for normal distribution:

1. We have

$$\mathbb{E}[\hat{\boldsymbol{\mu}}] = \mathbb{E}[\bar{\mathbf{x}}] = \mathbb{E}\left[\sum_{\alpha=1}^{N} \mathbf{x}_\alpha\right] = \boldsymbol{\mu}$$

and (not limited to normal distribution)

$$\mathbb{E}\left[\hat{\boldsymbol{\Sigma}}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{\alpha=1}^{N-1} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top\right] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

2. The sample covariance

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

is an unbiased estimator of $\boldsymbol{\Sigma}$.

Sufficiency

A statistic $\mathbf{t}$ is sufficient for a family of distributions of $\mathbf{y}$ or for a parameter $\boldsymbol{\theta}$ if the conditional distribution of $\mathbf{y}$ given $\mathbf{t}$ does not depend on $\boldsymbol{\theta}$.

The statistic $\mathbf{t}$ gives as much information about $\boldsymbol{\theta}$ as the entire sample $\mathbf{y}$.

Theorem 4

A statistic $\mathbf{t}(\mathbf{y})$ is sufficient for $\boldsymbol{\theta}$ if and only if the density $f(\mathbf{y} \mid \boldsymbol{\theta})$ can be factored as

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = g(\mathbf{t}(\mathbf{y}), \boldsymbol{\theta})h(\mathbf{y})$$

where $g(\mathbf{t}(\mathbf{y}), \boldsymbol{\theta})$ and $h(\mathbf{y})$ are nonnegative and $h(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$.

For the MLE of normal distribution, we apply this theorem with

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}, \quad \mathbf{y} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \quad \text{and} \quad \mathbf{t}(\mathbf{y}) = \{\bar{\mathbf{x}}, \mathbf{S}\}.$$

Sufficiency

Theorem 5

If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are observations from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

1. $\bar{\mathbf{x}}$ and $\mathbf{S}$ are sufficient for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$;
2. if $\boldsymbol{\mu}$ is given, $\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top$ is sufficient for $\boldsymbol{\Sigma}$;
3. if $\boldsymbol{\Sigma}$ is given, $\bar{\mathbf{x}}$ is sufficient for $\boldsymbol{\mu}$.

Asymptotic Normality

Multivariate central limit theorem.

Theorem 3

Let p-component vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots$ be i.i.d with means $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\nu}$ and covariance matrices $\mathbb{E}[(\mathbf{y}_\alpha - \boldsymbol{\nu})(\mathbf{y}_\alpha - \boldsymbol{\nu})^\top] = \mathbf{T}$. Then the limiting distribution of

$$\frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{y}_\alpha - \boldsymbol{\nu})$$

as $n \to +\infty$ is $\mathcal{N}(\mathbf{0}, \mathbf{T})$.

Bayes Procedure

If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independently distributed, each $\mathbf{x}_\alpha$ according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{\mu}$ has an a prior distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Phi})$, then the a posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is normal with mean

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}}+\frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu} \tag{1}$$

and covariance matrix

$$\boldsymbol{\Phi}-\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}.$$

If the loss function is

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) = (\boldsymbol{\theta}-\boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\boldsymbol{\theta}-\boldsymbol{\delta}(\mathbf{x}))$$

then the Bayes estimator of $\boldsymbol{\mu}$ is (1).

Outline

Chi-Squared Distribution

If $x_1, \ldots, x_n$ are independent, standard normal random variables, then the sum of their squares,

$$y = \sum_{i=1}^{n} x_i^2,$$

is distributed according to the (central) chi-squared distribution (χ^2-distribution) with n degrees of freedom.

We have $\mathbb{E}[y] = n$ and $\mathrm{Var}[y] = 2n$.

Chi-Squared Distribution

The probability density function of the (central) chi-squared distribution is

$$f(y;\, n) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} \exp\left(-\dfrac{y}{2}\right), & y > 0; \\ 0, & \text{otherwise}, \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t)\, \mathrm{d}t.$$

Noncentral Chi-Squared Distribution

If $x_1, \ldots, x_n$ are independent and each x_i are normally distributed random variables with means μ_i and unit variances, then the sum of their squares,

$$y = \sum_{i=1}^{n} x_i^2,$$

is distributed according to the noncentral Chi-squared distribution with n degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{n} \mu_i^2.$$

We have $\mathbb{E}[y] = n + \lambda$ and $\mathrm{Var}[y] = 2n + 4\lambda$.

Noncentral Chi-Squared Distribution

Theorem 1

If the n-component vector $\mathbf{y}$ is distributed according to $\mathcal{N}(\boldsymbol{\nu}, \mathbf{T})$ with $\mathbf{T} \succ \mathbf{0}$, then

$$\mathbf{y}^\top \mathbf{T}^{-1} \mathbf{y}$$

is distributed according to the noncentral χ^2-distribution with n degrees of freedom and noncentral parameter $\boldsymbol{\nu}^\top \mathbf{T}^{-1} \boldsymbol{\nu}$. If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is the central χ^2-distribution.

For the sample mean $\bar{\mathbf{x}} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma}\right)$, we have $\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$.

It follows from the theorem that

$$N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

has a (central) χ^2-distribution with p degrees of freedom.

Hypothesis Testing for the Mean (Covariance is Known)

Let $\chi_p^2(\alpha)$ be the number such that

$$\Pr\left\{N(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) > \chi_p^2(\alpha)\right\} = \alpha.$$

To test the hypothesis that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0$ is a specified vector, we use as our rejection region (critical region)

$$N(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) > \chi_p^2(\alpha).$$

F-Distribution

The F-distribution with d_1 and d_2 degrees of freedom is the distribution of

$$x = \frac{y_1/d_1}{y_2/d_2} = \frac{d_2 y_1}{d_1 y_2}$$

where y_1 and y_2 are independent random variables with Chi-square distributions with respective degrees of freedom d_1 and d_2.

Outline

The Generalized T^2-Statistic

The multivariate analog of t^2 is

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

where

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

Distribution of T^2-Statistics

Corollary 2

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 $N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$. If $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ then the F-distribution is central.

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

T^2-Statistic and Likelihood Ratio Criterion

We consider MLE for normal distribution. The likelihood function is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{pN}{2}} (\det(\boldsymbol{\Sigma}))^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\right).$$

The likelihood ratio criterion is

$$\lambda = \frac{\max\limits_{\boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\max\limits_{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Sigma}\in\mathbb{S}_p^{++}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}.$$

1. The denominator is the maximum over the entire parameter space.
2. The numerator is the maximum in the space restricted by the null hypothesis.
3. The likelihood ratio test is the procedure of rejecting the null hypothesis when λ is less than a predetermined constant.

T^2-Statistic and Likelihood Ratio Criterion

We have

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

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

The likelihood ratio test is defined by the critical region (region of rejection)

$$\lambda \leq \lambda_0, \tag{2}$$

where λ_0 is chosen so that the probability of (2) when the null hypothesis is true is equal to the significance level.

The inequality (2) also equivalent to

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

where $T_0^2 = (N-1)(\lambda_0^{-2/N} - 1)$.

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

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

Outline

The Distribution of the Sample Correlation Coefficient

Theorem 1

If the pairs $(z_{11}, z_{21}), \ldots, (z_{1n}, z_{2n})$ are independent and each pair are distributed according to

$$\begin{bmatrix} z_{1\alpha} \\ z_{2\alpha} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}\right), \quad \text{where } \alpha = 1, \ldots, n,$$

then given $z_{11}, z_{12}, \ldots, z_{1n}$, the conditional distributions of

$$b = \frac{\sum_{\alpha=1}^n z_{2\alpha} z_{1\alpha}}{\sum_{i=1}^n z_{1\alpha}^2} \quad \text{and} \quad \frac{u}{\sigma^2} = \sum_{\alpha=1}^n \frac{(z_{2\alpha} - bz_{1\alpha})^2}{\sigma^2}$$

are $\mathcal{N}\left(\beta, \sigma^2/c^2\right)$ and χ^2-distribution with $n-1$ degrees of freedom, respectively; and b and U are independent, where

$$\beta = \frac{\rho\sigma_2}{\sigma_1}, \quad \sigma^2 = \sigma_2^2(1-\rho^2) \quad \text{and} \quad c^2 = \sum_{i=1}^n z_{1\alpha}^2.$$

The Distribution of the Sample Correlation Coefficient

Theorem 2

if x and y are independently distributed, x having the distribution $\mathcal{N}(0,1)$ and y having the χ^2-distribution with m degrees of freedom, then

$$t = \frac{x}{\sqrt{y/m}}$$

has the density of t-distribution such that

$$f(t;m) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)}\left(1+\frac{t^2}{m}\right)^{-\frac{m+1}{2}}.$$

The Distribution of the Sample Correlation Coefficient

The conditional density of

$$t = \frac{cb/\sigma}{\sqrt{\frac{u/\sigma^2}{n-1}}} = \sqrt{n-1} \cdot \frac{r}{\sqrt{1-r^2}}$$

given $\mathbf{v}_1$ is

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

Then the conditional density of r given $\mathbf{v}_1$ is

$$k_N(r) = \frac{\Gamma\left(\frac{N-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{N-2}{2}\right)} (1-r^2)^{\frac{N-4}{2}}, \quad \text{where} \quad N = n+1.$$

We can verify that

$$\mathbb{E}\left[r^{2m}\right] = \frac{\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(m+\frac{1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{N-1}{2}+m\right)}.$$

The Asymptotic Distribution of Sample Correlation

The sample correlation coefficient can be written as $r = \frac{u_3}{\sqrt{u_1}\sqrt{u_2}}$.

Theorem 5 [Serfling (1980), Section 3.3]

Let $\{\mathbf{u}(n)\}$ be a sequence of m-component random vectors and $\mathbf{b}$ a fixed vector such that

$$\lim_{n\to\infty} \sqrt{n}(\mathbf{u}(n) - \mathbf{b}) \sim \mathcal{N}(\mathbf{0}, \mathbf{T}).$$

Let $\mathbf{f}(\mathbf{u})$ be a vector-valued function of $\mathbf{u}$ such that each component $f_j(\mathbf{u})$ has a nonzero differential at $\mathbf{u} = \mathbf{b}$, and let

$$\frac{\partial f_j(\mathbf{u})}{\partial u_i}\bigg|_{\mathbf{u}=\mathbf{b}}$$

be the (i, j)-th component of $\mathbf{\Phi}_{\mathbf{b}}$. Then $\sqrt{n}(\mathbf{f}(\mathbf{u}(n)) - f(\mathbf{b}))$ has the limiting distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{b}}^\top \mathbf{T} \mathbf{\Phi}_{\mathbf{b}})$.

The Asymptotic Distribution of Sample Correlation

Applying Theorem 5 with $r = f(\mathbf{u}) = u_3 u_1^{-\frac{1}{2}} u_2^{-\frac{1}{2}}$, we have $f(\mathbf{b}) = \rho$ and

$$\boldsymbol{\Phi}_{\mathbf{b}} = \begin{bmatrix} \left.\frac{\partial r}{\partial u_1}\right|_{\mathbf{u}=\mathbf{b}} \\ \left.\frac{\partial r}{\partial u_2}\right|_{\mathbf{u}=\mathbf{b}} \\ \left.\frac{\partial r}{\partial u_3}\right|_{\mathbf{u}=\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \left.-\frac{1}{2} u_3 u_1^{-\frac{3}{2}} u_2^{-\frac{1}{2}}\right|_{\mathbf{u}=\mathbf{b}} \\ \left.-\frac{1}{2} u_3 u_1^{-\frac{1}{2}} u_2^{-\frac{3}{2}}\right|_{\mathbf{u}=\mathbf{b}} \\ \left.u_1^{-\frac{1}{2}} u_2^{-\frac{1}{2}}\right|_{\mathbf{u}=\mathbf{b}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1 \end{bmatrix}.$$

Thus, the covariance of the limiting distribution of $\sqrt{n}(r(n) - \rho)$ is

$$\begin{bmatrix} -\frac{1}{2}\rho & -\frac{1}{2}\rho & 1 \end{bmatrix} \begin{bmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{bmatrix} \begin{bmatrix} -\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1 \end{bmatrix} = (1-\rho^2)^2$$

and we have $\lim_{n\to\infty} \dfrac{\sqrt{n}(r(n) - \rho)}{1-\rho^2} \sim \mathcal{N}(0,1)$.

Partial Correlation Coefficients

Consider the normal distribution $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

then the conditional distribution of $\mathbf{x}^{(1)}$ given $\mathbf{x}^{(2)}$ is

$$\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(1)} + \mathbf{B}\big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big), \boldsymbol{\Sigma}_{11.2}\right),$$

where

$$\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \quad \text{and} \quad \boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Outline

The Wishart Distribution

Theorem 2

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be independently distributed, each according to $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, where $n \geq p$; let

$$\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top = \mathbf{T}^* \mathbf{T}^{*\top},$$

where $t_{ij}^* = 0$ for $i < j$, and $t_{ii}^* > 0$ for $i = 1, \ldots, p$. Then the density of $\mathbf{T}^*$ is

$$\frac{\prod_{i=1}^{p} t_{ii}^{*\,n-i} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}^*\mathbf{T}^{*\top}\right)\right)}{2^{\frac{p(n-2)}{2}} \pi^{\frac{p(p-1)}{4}} (\det(\mathbf{\Sigma}))^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}.$$

The Wishart Distribution

Theorem 3

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be independently distributed, each according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $n \geq p$. Then the density of $\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top$ is

$$\frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} (\det(\boldsymbol{\Sigma}))^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)} \tag{3}$$

for $\mathbf{A}$ positive definite, and 0 otherwise.

Corollary 2

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be independently distributed, each according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $N > p$; Then the density of $\mathbf{A} = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top$ is (3), where $n = N - 1$ and $\mathbf{x} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha$.

The multivariate gamma function is defined as

$$\Gamma_p(t) = \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\Big(t - \frac{1}{2}(i-1)\Big).$$

Then the Wishart density can be written as

$$\frac{(\det(\mathbf{A}))^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{A}\right)\right)}{2^{\frac{np}{2}} (\det(\mathbf{\Sigma}))^{\frac{n}{2}} \Gamma_p\left(\frac{n}{2}\right)}.$$

The Generalized Variance

The multivariate analog of the variance of the univariate distribution:

1. Covariance matrix $\boldsymbol{\Sigma}$.
2. The scalar $\det(\boldsymbol{\Sigma})$, which is called the generalized variance.

The generalized variance of the sample of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is

$$\det(\mathbf{S}) = \det\left(\frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \bar{\mathbf{x}}_\alpha)(\mathbf{x}_\alpha - \bar{\mathbf{x}}_\alpha)^\top\right)$$

Distribution of the Sample Generalized Variance

The distribution of $\det(\mathbf{S}) = \det(\mathbf{B})\det(\boldsymbol{\Sigma})/(N-1)^p$ is

$$\frac{\det(\boldsymbol{\Sigma})\prod_{i=1}^{p} t_{ii}^2}{(N-1)^p},$$

where $t_{11}^2, \dots, t_{pp}^2$ are independent and t_{ii}^2 are distributed according to χ^2-distribution with $N-i$ degrees of freedom.

The Inverted Wishart Distribution

If $\mathbf{A}$ has the distribution $\mathcal{W}(\mathbf{\Sigma}, m)$, then $\mathbf{B} = \mathbf{A}^{-1}$ has the density is

$$w^{-1}(\mathbf{B} \mid \mathbf{\Psi}, m) = \frac{(\det(\mathbf{\Psi}))^{\frac{m}{2}} (\det(\mathbf{B}))^{-\frac{m+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Psi}\mathbf{B}^{-1}\right)\right)}{2^{\frac{mp}{2}}\Gamma_p\left(\frac{m}{2}\right)}.$$

for $\mathbf{B}$ positive definite and 0 elsewhere, where $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}$.

1. We call $\mathbf{B}$ has the inverted Wishart distribution with m degrees of freedom and denote $\mathbf{B} \sim \mathcal{W}^{-1}(\mathbf{\Psi}, m)$.
2. We call $\mathbf{\Psi}$ the precision matrix or concentration matrix.
3. The derivation of $w^{-1}(\mathbf{\Psi}, m)$ are based on the determinant for Jacobian of transformation $\mathbf{A} = \mathbf{B}^{-1}$ is $(\det(\mathbf{B}))^{-(p+1)}$.

The Inverted Wishart Distribution

If the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$ is in the same probability distribution family as the prior probability distribution $p(\boldsymbol{\theta})$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior.

Theorem 6

If $\mathbf{A}$ has the distribution $\mathcal{W}(\boldsymbol{\Sigma}, n)$ and $\boldsymbol{\Sigma}$ has the a prior distribution $\mathcal{W}^{-1}(\boldsymbol{\Psi}, m)$, then the conditional distribution of $\boldsymbol{\Sigma}$ given $\mathbf{A}$ is the inverted Wishart distribution $\mathcal{W}^{-1}(\mathbf{A} + \boldsymbol{\Psi}, n + m)$.

Corollary 4

If $n\mathbf{S}$ has the distribution $\mathcal{W}(\boldsymbol{\Sigma}, n)$ and $\boldsymbol{\Sigma}$ has the a prior distribution $\mathcal{W}^{-1}(\boldsymbol{\Psi}, m)$, then the conditional distribution of $\boldsymbol{\Sigma}$ given $\mathbf{S}$ is the inverted Wishart distribution $\mathcal{W}^{-1}(n\mathbf{S} + \boldsymbol{\Psi}, n + m)$.

Outline

The Estimation in Multivariate Linear Regression

Theorem 1

Suppose $\mathbf{x}_\alpha$ is an observation from $\mathcal{N}_q(\mathbf{B}\mathbf{z}_\alpha, \mathbf{\Sigma})$ for $\alpha = 1, \dots, N$, where $[\mathbf{z}_1, \dots, \mathbf{z}_N] \in \mathbb{R}^{N \times q}$ of rank q is given and $N \geq p + q$, the maximum likelihood estimator of $\mathbf{B}$ is given by

$$\hat{\mathbf{B}} = \mathbf{C}\mathbf{A}^{-1},$$

where

$$\mathbf{C} = \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{z}_\alpha^\top \qquad \text{and} \qquad \mathbf{A} = \sum_{\alpha=1}^{N} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top;$$

the maximum likelihood estimator of $\mathbf{\Sigma}$ is give by

$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)(\mathbf{x}_\alpha - \hat{\mathbf{B}}\mathbf{z}_\alpha)^\top.$$

Properties of the Estimators

The density then can be written as

$$\frac{1}{(2\pi)^{\frac{Np}{2}}(\det(\boldsymbol{\Sigma}))^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left(N\hat{\boldsymbol{\Sigma}}+(\hat{\mathbf{B}}-\mathbf{B})\mathbf{A}(\hat{\mathbf{B}}-\mathbf{B})^{\top}\right)\right)\right).$$

Then $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\Sigma}}$ form a sufficient set statistics for $\mathbf{B}$ and $\boldsymbol{\Sigma}$.

The Best Linear Unbiased Estimator

A linear unbiased estimator F is best if it has minimum variance over all linear unbiased estimators; that is, if $\mathbb{E}[(F-\beta_{ig})^2] \leq \mathbb{E}[(G-\beta_{ig})^2]$ for $G = \sum_{\alpha=1}^{N} \mathbf{g}_\alpha^\top \mathbf{x}_\alpha$ and $\mathbb{E}[G] = \beta_{ig}$.

The least squares estimator $\hat{\mathbf{B}}$ is the best linear unbiased estimator of $\mathbf{B}$.

1. Let $\tilde{\beta}_{ig} = \sum_{\alpha=1}^{N} \sum_{j=1}^{p} f_{j\alpha} x_{j\alpha}$ be arbitrary unbiased estimator of β_{ig}.
2. Then we have

$$
\begin{aligned}
&\mathbb{E}\left[(\tilde{\beta}_{ig}-\beta_{ig})^2\right] \\
=&\mathbb{E}\left[(\hat{\beta}_{ig}-\beta_{ig})^2\right] + 2\mathbb{E}\left[(\hat{\beta}_{ig}-\beta_{ig})(\tilde{\beta}_{ig}-\hat{\beta}_{ig})\right] + \mathbb{E}\left[(\tilde{\beta}_{ig}-\hat{\beta}_{ig})^2\right] \\
=&\mathbb{E}\left[(\hat{\beta}_{ig}-\beta_{ig})^2\right] + \mathbb{E}\left[(\tilde{\beta}_{ig}-\hat{\beta}_{ig})^2\right] \\
\geq&\mathbb{E}\left[(\hat{\beta}_{ig}-\beta_{ig})^2\right].
\end{aligned}
$$

Testing Equality of Means with Common Covariance

Let $\mathbf{x}_\alpha^{(g)}$ be an observation from the g-th population $\mathcal{N}(\boldsymbol{\mu}^{(g)}, \boldsymbol{\Sigma})$ for $\alpha = 1, \ldots, N_g$, $g = 1, \ldots, q$.

We wish to test the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g.$$

The likelihood function is

$$L = \prod_{g=1}^{q} \frac{1}{(2\pi)^{\frac{pN_g}{2}} (\det(\boldsymbol{\Sigma}))^{\frac{N_g}{2}}} \exp\left(-\frac{1}{2} \sum_{\alpha=1}^{N_g} \left(\mathbf{x}_\alpha^{(g)} - \boldsymbol{\mu}^{(g)}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_\alpha^{(g)} - \boldsymbol{\mu}^{(g)}\right) \right).$$

1. The space Ω is the parameter space in which $\boldsymbol{\Sigma}$ is positive definite and each $\boldsymbol{\mu}^{(g)}$ is any vector.
2. The space ω is the parameter space in which $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g$ (positive definite) and $\boldsymbol{\Sigma}$ is any positive definite matrix.

Testing Equality of Means with Common Covariance

Let

$$N = \sum_{g=1}^{q} N_g, \quad \mathbf{A}_g = \sum_{\alpha=1}^{N_g} \left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\right)\left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}^{(g)}\right)^\top, \quad \mathbf{A} = \sum_{g=1}^{q} \mathbf{A}_g,$$

and

$$\mathbf{B} = \sum_{g=1}^{q} \sum_{\alpha=1}^{N_g} \left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}\right)\left(\mathbf{x}_\alpha^{(g)} - \bar{\mathbf{x}}\right)^\top.$$

The maximum likelihood estimators of $\boldsymbol{\mu}^{(g)}$ and $\boldsymbol{\Sigma}$ in Ω are given by

$$\hat{\boldsymbol{\mu}}_\Omega^{(g)} = \bar{\mathbf{x}}^{(g)} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_\Omega = \frac{1}{N}\mathbf{A}.$$

The maximum likelihood estimators of $\boldsymbol{\mu}^{(g)}$ and $\boldsymbol{\Sigma}$ in ω are given by

$$\hat{\boldsymbol{\mu}}_\omega^{(g)} = \bar{\mathbf{x}} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_\omega = \frac{1}{N}\mathbf{B}.$$

Testing Equality of Means with Common Covariance

The likelihood ratio criterion for testing H_0 is

$$\lambda_0 = \frac{\left(\det\left(\hat{\boldsymbol{\Sigma}}_\Omega\right)\right)^{\frac{N}{2}}}{\left(\det\left(\hat{\boldsymbol{\Sigma}}_\omega\right)\right)^{\frac{N}{2}}} = \frac{(\det(\mathbf{A}))^{\frac{N}{2}}}{(\det(\mathbf{B}))^{\frac{N}{2}}}.$$

The critical region is

$$\lambda_0 \leq \lambda_0(\epsilon),$$

where $\lambda_0(\epsilon)$ is defined so that above inequality holds with probability ϵ when H_0 is true.

Outline

Principal Components

Let random vector $\mathbf{x}$ of p component has mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

Let $\boldsymbol{\beta}$ be a p-component column vector such that $\|\boldsymbol{\beta}\|_2 = 1$.

1. The variance of $\boldsymbol{\beta}^\top \mathbf{x}$ is

$$\mathbb{E}\left[(\boldsymbol{\beta}^\top \mathbf{x})^2\right] = \boldsymbol{\beta}^\top \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]\boldsymbol{\beta} = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}.$$

2. Maximizing $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}$ must satisfy

$$(\boldsymbol{\Sigma} - \lambda_1 \mathbf{I})\boldsymbol{\beta} = \mathbf{0},$$

where λ_1 is the largest root of

$$\det(\boldsymbol{\Sigma} - \lambda \mathbf{I}) = 0.$$

3. Let $\boldsymbol{\beta}^{(1)} = \underset{\|\boldsymbol{\beta}\|_2=1}{\arg\max}\, \boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}$.

Principal Components

At the $(r+1)$-th step, we want to find a vector such that $\boldsymbol{\beta}^\top \mathbf{x}$ has maximum variance and lacks correlation with $u_1 \ldots, u_r$, that is

$$0 = \mathbb{E}\big[\boldsymbol{\beta}^\top \mathbf{x} u_i\big] = \mathbb{E}\big[\boldsymbol{\beta}^\top \mathbf{x}\mathbf{x}^\top \boldsymbol{\beta}^{(i)}\big] = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}^{(i)} = \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}^{(i)}$$

for $i = 1, \ldots, r$, where $u_i = {\boldsymbol{\beta}^{(i)}}^\top \mathbf{x}$

Finally, we obtain $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(p)}$ and $\lambda_1 \geq \cdots \geq \lambda_p$ such that

$$\boldsymbol{\Sigma}\mathbf{B} = \mathbf{B}\boldsymbol{\Lambda}$$

where $\mathbf{B} = [\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(p)}]$ satisfying $\mathbf{B}^\top \mathbf{B} = \mathbf{I}$ and

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$

Principal Components

The transformation

$$\mathbf{u} = \mathbf{B}^\top \mathbf{x}$$

leads to the r-th component of $\mathbf{u}$ has maximum variance of all normalized linear combinations uncorrelated with $u_1, \ldots, u_{r-1}$.

The vector $\mathbf{u}$ is defined as the vector of principal components of $\mathbf{x}$.

Maximum Likelihood Estimators of Principal Components

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be N observations from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ has p different characteristic roots and $N > p$. Then a set of maximum likelihood estimators of $\lambda_1, \ldots, \lambda_p$ and $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(p)}$ consists of the roots $\lambda_1 > \cdots > \lambda_p$ of

$$\det(\hat{\mathbf{\Sigma}} - \lambda \mathbf{I}) = 0$$

and a set of vectors $\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(p)}$ satisfying $\|\hat{\boldsymbol{\beta}}^{(i)}\|_2 = 1$ and

$$(\hat{\mathbf{\Sigma}} - \lambda_i \mathbf{I})\hat{\boldsymbol{\beta}}^{(i)} = \mathbf{0}$$

for $i = 1, \ldots, p$, where $\hat{\mathbf{\Sigma}}$ is the the maximum likelihood estimate of $\mathbf{\Sigma}$.

Canonical Correlations

We still consider random vector $\mathbf{x}$ of p components has zero means and the covariance matrix $\mathbf{\Sigma} \succ \mathbf{0}$.

We partition $\mathbf{x}$ into two subvectors of p_1 and p_2 components $(p_1 \leq p_2)$

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}.$$

The covariance matrix is partitioned into p_1 and p_2 rows and columns

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

Here we shall develop a transformation of $\mathbf{x}^{(1)}$ and another transformation of $\mathbf{x}^{(2)}$ to a new system that exhibit clearly the intercorrelations between $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$.

Outline

Canonical Correlations

Consider linear combinations

$$u = \boldsymbol{\alpha}^\top \mathbf{x}^{(1)} \quad \text{and} \quad v = \boldsymbol{\gamma}^\top \mathbf{x}^{(2)}.$$

We ask for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ that maximize the correlation between u and v.

1. We require $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ such that

$$1 = \mathbb{E}\left[u^2\right] = \mathbb{E}\left[\boldsymbol{\alpha}^\top \mathbf{x}^{(1)} \mathbf{x}^{(1)\top} \boldsymbol{\alpha}\right] = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha},$$
$$1 = \mathbb{E}\left[v^2\right] = \mathbb{E}\left[\boldsymbol{\gamma}^\top \mathbf{x}^{(2)} \mathbf{x}^{(2)\top} \boldsymbol{\gamma}\right] = \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}.$$

2. The correlation between u and v is

$$\mathbb{E}[uv] = \mathbb{E}\left[\boldsymbol{\alpha}^\top \mathbf{x}^{(1)} \mathbf{x}^{(2)\top} \boldsymbol{\gamma}\right] = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}.$$

3. Then the problem is

$$\max_{\substack{\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} = 1 \\ \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} = 1}} \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}.$$

Canonical Correlations

The solution of

$$\max_{\substack{\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}=1 \\ \boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}=1}} \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{12}\boldsymbol{\gamma}.$$

must satisfy

$$\begin{bmatrix} -\lambda\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda\boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{bmatrix} = \mathbf{0},$$

where λ is the root of

$$\det\left(\begin{bmatrix} -\lambda\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda\boldsymbol{\Sigma}_{22} \end{bmatrix}\right) = 0.$$

Denote the largest root and the corresponds vectors be λ_1, $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\gamma}^{(1)}$

Then we consider $u = \boldsymbol{\alpha}^\top \mathbf{x}^{(1)}$ and $v = \boldsymbol{\gamma}^\top \mathbf{x}^{(2)}$ for $\mathbf{x}^{(2)}$ with maximum correlation, such that u is uncorrelated with $u_1 = {\boldsymbol{\alpha}^{(1)}}^\top \mathbf{x}^{(1)}$ and v is uncorrelated with $v_1 = {\boldsymbol{\gamma}^{(1)}}^\top \mathbf{x}^{(2)}$.

This procedure is continued. At r-th step, we have

$$u_1 = {\boldsymbol{\alpha}^{(1)}}^\top \mathbf{x}^{(1)}, \dots, u_r = {\boldsymbol{\alpha}^{(r)}}^\top \mathbf{x}^{(1)}$$
$$v_1 = {\boldsymbol{\gamma}^{(1)}}^\top \mathbf{x}^{(2)}, \dots, v_r = {\boldsymbol{\gamma}^{(r)}}^\top \mathbf{x}^{(2)}$$

and each of them are uncorrelated. Let the correlation between u_i and v_i be λ_i.

We obtain $\boldsymbol{\alpha}^{r+1}$ and $\boldsymbol{\gamma}^{(r+1)}$ by maximizing the correlation between $u = \boldsymbol{\alpha}^\top \mathbf{x}^{(1)}$ and $v = \boldsymbol{\gamma}^\top \mathbf{x}^{(2)}$ such that u is uncorrelated with $u_1, \dots, u_r$ and v is uncorrelated with $v_1, \dots, v_r$.

Let $\mathbf{A} = [\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(p_1)}]$, $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2] = [\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(p_2)}]$ and

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{p_1} \end{bmatrix}.$$

All of conditions can be summarized as

$$\begin{aligned} \mathbf{A}^\top \boldsymbol{\Sigma}_{11} \mathbf{A} &= \mathbf{I}, \\ \mathbf{A}^\top \boldsymbol{\Sigma}_{12} \boldsymbol{\Gamma}_1 &= \boldsymbol{\Lambda}, \\ \boldsymbol{\Gamma}_1^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\Gamma}_1 &= \mathbf{I}, \\ \boldsymbol{\Gamma}_2^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\Gamma}_1 &= \mathbf{0}, \\ \boldsymbol{\Gamma}_2^\top \boldsymbol{\Sigma}_{22} \boldsymbol{\Gamma}_2 &= \mathbf{I}. \end{aligned}$$

Canonical Correlations

Each $\boldsymbol{\alpha}^{(i)}$, $\boldsymbol{\gamma}^{(i)}$ can be obtained by solving

$$\begin{bmatrix} -\lambda_i \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda_i \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{bmatrix} = \mathbf{0},$$

where λ_i is the i-th largest root of

$$\det\left(\begin{bmatrix} -\lambda \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda \boldsymbol{\Sigma}_{22} \end{bmatrix} \right) = 0.$$

This can be written as generalized eigenvalue problems

$$\left(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \lambda^2 \boldsymbol{\Sigma}_{11}\right)\boldsymbol{\gamma} = \mathbf{0}$$

and

$$\left(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \lambda^2 \boldsymbol{\Sigma}_{22}\right)\boldsymbol{\alpha} = \mathbf{0}.$$

Let

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}$$

be a random vector where $\mathbf{x}^{(1)}$ has p_1 components and $\mathbf{x}^{(2)}$ has p_2 components.

In the r-th pair of canonical variates is the pair of linear combinations

$$u_r = \boldsymbol{\alpha}^{(r)}{}^\top \mathbf{x}^{(1)} \quad \text{and} \quad v_r = \boldsymbol{\gamma}^{(r)}{}^\top \mathbf{x}^{(2)},$$

each of unit variance and uncorrelated with the first $r-1$ pairs of canonical variates and having maximum correlation.

The correlation between u_r and v_r is the r-th canonical correlation.

Outline

Factor Analysis

Let the observable vector $\mathbf{t}$ be written as

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\mathbf{t}$, $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$ are column vectors of d components, $\mathbf{x}$ is column vector of q components ($q \leq d$), and $\mathbf{W}$ is a $d \times q$ matrix.

We assume $\boldsymbol{\epsilon}$ is distributed independently of $\mathbf{x}$ and with mean $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and covariance matrix $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{\Psi}$ is diagonal.

1. The model is similar to regression, but $\mathbf{x}$ is unobserved.
2. There are two kinds of models:
    - $\mathbf{x}$ is a nonrandom vector
    - $\mathbf{x}$ is a random vector: $\mathbf{t}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \boldsymbol{\epsilon}_\alpha$

Probabilistic Principle Component Analysis

Let $\mathbf{t}_1, \dots, \mathbf{t}_N$ be N independent observation and we have

$$\mathbf{t}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \boldsymbol{\epsilon}_\alpha,$$

where $\mathbf{x}_\alpha \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\epsilon}_\alpha \sim \mathcal{N}_d(\mathbf{0}, \sigma^2\mathbf{I})$ are independent.

Then, we have $\mathbf{t}_\alpha \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

The log-likelihood function is

$$-\frac{Nd\ln(2\pi)}{2} - N\ln\det(\mathbf{C}) - \mathrm{tr}\left(\mathbf{C}^{-1}\sum_{\alpha=1}^{N}(\mathbf{t}_\alpha - \boldsymbol{\mu})(\mathbf{t}_\alpha - \boldsymbol{\mu})^\top\right).$$

The Maximum Likelihood Estimators

The maximum likelihood estimators of $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2 are

$$\boldsymbol{\mu} = \bar{\mathbf{t}} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{t}_\alpha, \quad \hat{\mathbf{W}} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \hat{\sigma}^2\mathbf{I})\mathbf{R} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{d-q}\sum_{j=q+1}^{d} \lambda_j,$$

where $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ with columns are the principal eigenvectors of

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{t}_\alpha - \bar{\mathbf{t}})(\mathbf{t}_\alpha - \bar{\mathbf{t}})^\top,$$

$\boldsymbol{\Lambda}_q \in \mathbb{R}^{q\times q}$ is diagonal matrix with corresponding eigenvalues $\lambda_1, \ldots, \lambda_q$ and $\mathbf{R}$ is any $q \times q$ orthogonal matrix.

The EM Algorithm

The update of the EM algorithm

1. In E-step, we take the expectation

$$l_C = \mathbb{E}\left[\ln\left(\prod_{\alpha=1}^{N} p(\mathbf{x}_\alpha \,|\, \mathbf{t}_\alpha)\right)\right].$$

2. In the M-step, we maximized l_C with respect to $\mathbf{W}$ and σ^2:

$$\begin{aligned}\tilde{\mathbf{W}} =& \hat{\mathbf{\Sigma}}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\mathbf{\Sigma}}\mathbf{W})^{-1},\\ \tilde{\sigma}^2 =& \frac{1}{d}\mathrm{tr}\left(\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\right).\end{aligned}$$

Note that the computational complexity of EM is $\mathcal{O}(Ndq)$, while MLE requires $\mathcal{O}(Nd^2 + d^3)$.