MAT 4375

Multivariate Statistical Methods

Study Guide

Fall 2024

Vectors and Matrices

- A matrix is a rectangular array of scalar elements

$$A_{p\times q} = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix}$$

$p\times q$ dimension of matrix

p: number of rows

q: number of columns

- Row vector: $a_{ij} = (a_{11}\ a_{12}\ a_{13}\cdots a_{1q})$
- Column vector:

$$a_{i1} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{p1} \end{pmatrix}$$

Elementary Matrix Operations

- Sum and difference: A+B or A−B, done by doing element-wise addition and subtraction
- Multiplication (written AB)

$$AB = \left(\sum_k a_{ik}b_{kj}\right)$$

  - For multiplication to work, we need the matrices to be conformable
- Inner product: product of row vector and column vector, leads to scalar
- Outer product: product of column vector and row vector, leads to matrix
- Transpose: interchanging rows and columns of A, denoted A'
- A square matrix has same rows and columns. A square matrix A is symmetric if $A=A'$
- Properties of transpose:
  - $(A')'=A$
  - $(A+B)'=A'+B'$
  - $(AB)'=B'A'$
  - $(cA)'=cA'$
- A diagonal matrix is a square matrix with zero off-diagonal elements
- Upper triangular matrices have zero elements below the diagonal and lower triangular matrices have zero elements above the diagonal
- Kronecker Product:
  - $A=(a_{ij})$: $p\times q$
  - $B=(b_{ij})$: $m\times n$

$$A\otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ \vdots & \ddots & & \vdots \\ a_{p1}B & \cdots & & a_{pq}B \end{pmatrix},\ pm\times qn$$

- Properties of Kronecker Product
  - $(A+B)\otimes C = (A\otimes C)+(B\otimes C)$
  - $A\otimes(B+C) = (A\otimes B)+(A\otimes C)$
  - $A\otimes B \neq B\otimes A$
  - $(A\otimes B)(C\otimes D) = AC\otimes BD$
  - $(A\otimes B)' = A'\otimes B'$
- Vectorization, denote by vec(A), is done by stacking the columns of A and we form a column vector

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\text{vec}(A) = \begin{pmatrix} a \\ c \\ b \\ d \end{pmatrix}$$

Trace of Matrix

- Sum of diagonals of matrix

$$\text{trace}(A) = \sum_i a_{ii}$$

- Properties of trace
  - $\text{trace}(A') = \text{trace}(A)$
  - $\text{trace}(A+B) = \text{trace}(A)+\text{trace}(B)$
  - $\text{trace}(AB) = \text{trace}(BA)$
  - $\text{trace}(AA') = \text{trace}(A'A)$
  - $\text{trace}(ABC) = \text{trace}(CAB) = \text{trace}(BCA)$
  - $\text{trace}(A\otimes B) = \text{trace}(A)\,\text{trace}(B)$

Determinant of Matrix

- For 1×1 matrix, or scalar, the determinant is the value itself
- For 2×2 matrix $A=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, the determinant, denoted by $|A|$ is given by

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

- For 3×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

The determinant is calculated as

$$a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

- General formula

$$\det(A) = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}|A_{ij}| \text{ for all } j$$

or

$$\det(A) = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}|A_{ij}| \text{ for all } i$$

- Properties
  - $|A'| = |A|$
  - $|AB| = |A||B| = |BA|$
  - $|kA| = k^p|A|$, A $p\times p$
  - $|AA'| \geq 0$
- Positive semi-definite matrix: a symmetric matrix A is called positive semi-definite if $\vec{x}'A\vec{x} \geq 0$ for all $\vec{x}\in\mathbb{R}^p$ and $\vec{x}'A\vec{x} = 0$ for some nonzero $\vec{x}$. A positive semi-definite matrix can be written as XX' for some matrix X, hence $|A|\geq 0$
- Positive definite matrix: a symmetric matrix A is called positive definite if $\vec{x}'A\vec{x} > 0$ for any non-zero $\vec{x}\in\mathbb{R}^p$ vector. Again, A can be written as XX', but in this case X is nonsingular, $|A|>0$

Rank of Matrix

- Denoted P(A), is the number of linearly independent rows of the matrix
- Can use elementary row operations to get the row-echelon form (or reduced row-echelon form). The rank of the matrix is the number of non-zero rows of the reduced row-echelon form
- A square matrix $A_{p\times p}$ is said to be of full rank if $P(A)=p$. That is all rows are linearly independent
- Full rank $\Leftrightarrow |A|>0 \Leftrightarrow$ positive-definite

Inverse of Matrix

- A square matrix is invertible (or non-singular) if it is of full rank. Otherwise, it is singular
- For a non-singular matrix $A_{p\times p}$, there exists a unique matrix (inverse), often denoted by A^{-1} s.t. $AA^{-1} = I_p$
- Can obtain inverse of A using reduced row echelon form $[A|I_p] \rightarrow [I_p|A^{-1}]$
- Properties of inverse:
  - $(A')^{-1} = (A^{-1})'$, $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$, $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$
  - $(kA)^{-1} = \frac{1}{k}A^{-1}$ for $k\neq 0$, $|A^{-1}| = \frac{1}{|A|}$

Generalized Inverses

- Denoted by A^- s.t. $AA^-A = A$
- Not usually unique
- Moore-Penrose inverse is unique:
  - Moore-Penrose Conditions
    - AA^- and A^-A are symmetric
    - $AA^-A = A$ and $A^-AA^- = A^-$

Null Matrix

- Denoted by A^0 s.t. $AA^0 = 0$
- $I - A(A'A)^{-1}A'$ is a null matrix for A

Eigenvalues and Eigenvectors

- They play important roles in multivariate analysis
- Eigenvalues of a square matrix $A_{p\times p}$ are roots of $|A-\lambda I_p| = 0$
- Eigenvector $\vec{v}$ corresponding to eigenvalue λ of A can be found by solving $(A-\lambda I_p)\vec{v} = \vec{0}$
- Condition number: $\frac{\lambda_{max}}{\lambda_{min}}$
- Condition index: $\frac{\lambda_{max}}{\lambda}$
- A full rank matrix has all non-zero eigenvalues
- Large condition number means we have an ill-conditioned problem
- Determinant and Trace
  - $\text{trace}(A) = \sum_{i=1}^{p}\lambda_i$
  - $|A| = \prod_{i=1}^{p}\lambda_i$

  λ_i eigenvalues of $A_{p\times p}$

Matrix Decompositions

- Singular Value Decomposition: For any matrix A, there exist two orthogonal matrices U and V and a diagonal matrix D such that $A=UDV'$.
  - Diagonal values of D are singular values of A
  - Columns of U and V are called left and right singular vectors
- For a symmetric matrix A, there exists an orthogonal matrix T and a diagonal matrix Λ such that $A = T\Lambda T'$. The diagonal elements of Λ are the eigenvalues of A and the columns of T are the corresponding eigenvectors. This is known as eigen or spectral decomposition
- Cholesky Decomposition: For a symmetric positive definite matrix A, we have $A = TT'$, T lower triangular

Quadratic forms

If we have a vector $\vec{x}\in\mathbb{R}^n$, then we have $\vec{x}'A\vec{x}\in\mathbb{R}$ is a quadratic form for some $A\in\mathbb{R}^{n\times n}$

If we want to write $\sum_{i=1}^{n}(x_i-\bar{x})^2$ in matrix form, we use the following:

- 1_n is a vector of size n where all elements are equal to 1
- $J_{n\times n}$ is an $n\times n$ matrix where all elements are equal to 1 (write as J for simplicity)
- $J^2 = nJ$
- If we let $J^* = \frac{1}{n}J$, we get $J^{*2} = J^*$
- Define $C = I - \frac{1}{n}J = I - J^*$
- We get $\vec{x}'C\vec{x} = \sum_{i=1}^{n}(x_i-\bar{x})^2$

Some common Matrix Derivatives

- If we have a column vector

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

Then we have

$$\frac{\partial f(\vec{a})}{\partial \vec{a}} = \begin{pmatrix} \frac{\partial f(\vec{a})}{\partial a_1} \\ \frac{\partial f(\vec{a})}{\partial a_2} \\ \vdots \\ \frac{\partial f(\vec{a})}{\partial a_n} \end{pmatrix}$$

- For $f(\vec{a}) = \vec{c}'\vec{a}$, where $\vec{c}$ is a column vector of the same length as $\vec{a}$. We then get

$$\frac{\partial f(\vec{a})}{\partial \vec{a}} = \vec{c}$$

- For $f(\vec{a}) = \vec{a}'B\vec{a}$, $B=(b_{ij}) = (\vec{b}_1\ \vec{b}_2\ \vec{b}_3\cdots\vec{b}_n)$ a symmetric matrix. We have

$$\frac{\partial f(\vec{a})}{\partial \vec{a}} = 2B'\vec{a} = 2B\vec{a}$$

Probability Distributions

- If X is a random variable, its cumulative distribution function (CDF) is

$$F_X(x) = P(X\leq x) \quad \text{for all } x\in(-\infty,\infty)$$

Discrete Random Variables

- A random variable X is called a discrete random variable if it assumes only finite or denumerably infinite number of values
- $F_X(x)$ is a step function
- The probability mass function of a discrete random variable X that takes values x_i, where $i=1,2,3,\ldots$ is

$$P_X(x_i) = P(X=x_i) \quad \text{for all } i$$

- Common Discrete Distributions: Discrete Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative Binomial

Continuous Random Variables

- A random variable is continuous if it takes on any real numbers within its support
- A random variable is continuous if $F_X(x)$ is continuous
- The probability density function $f_X(x)$ for a continuous random variable is a function such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \quad \text{for } x\in(-\infty,\infty)$$

or $\frac{d}{dx}F_X(x) = f_X(x)$

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
- Common Continuous Distributions: Normal, Chi-Square, Exponential, Gamma, Beta, Student t, F

Expectation and Moments

- For a continuous random variable X, the expected value is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$$

- For a discrete random variable X, the expected value is

$$E[X] = \sum_i x_i P_X(x_i) \text{ for all } x_i$$

- Also known as the mean of distribution
- Is a measure of central tendency
- Other measures of central tendency include median and mode
- For a random variable X, the r^{th} non-central moment is defined as $E[X^r]$
- The expected value is therefore the first non-central moment

- The r^{th} central moment is defined as $E[(X-E[X])^r]$
- The variance of a random variable X is defined as its second central moment, or $E[(X-E[X])^2]$
- The square root of the variance is referred to as the standard deviation
- Other measures of variability include interquartile range (IQR) and range

Normal Distribution

- Symmetric, unimodal
- Closed under linear combinations, conditioning, and marginalizing
- Most commonly used distributions are transformations of the normal distribution
- CLT
- Density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} I_{(-\infty,\infty)}(x)$$

$-\infty<\mu<\infty$, $\sigma > 0$

- $E[X]=\mu$ and $Var[X]=\sigma^2$

Distribution characterized by its mean and variance, written $X\sim N(\mu,\sigma^2)$

- Density usually denoted $\phi_{\mu,\sigma^2}(x)$ and CDF $\Phi_{\mu,\sigma^2}(x)$
- $\frac{X-\mu}{\sigma} \sim N(0,1)$ standard normal, pdf $\phi(x)$ and cdf $\Phi(x)$

Relationship of Normal with Other Distributions

- If $Z\sim N(0,1)$, then $Z^2\sim\chi^2(1)$
- If $Z_1, Z_2, \dots, Z_n \sim N(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$
- If $X_1, X_2, \dots, X_n \sim N(\mu,\sigma^2)$, then we have $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- If $Z\sim N(0,1)$ and $V\sim\chi^2(n)$, then $\frac{Z}{\sqrt{\frac{V}{n}}} \sim t(n)$
- If $X_1, X_2, \dots, X_n \sim N(\mu,\sigma^2)$, then $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$
- If $V\sim\chi^2(r)$ and $W\sim\chi^2(s)$, then $\frac{\frac{V}{r}}{\frac{W}{s}} \sim F(r,s)$

Random Samples and Statistics

- A random sample is a sequence of independent and identically distributed random variables from a given distribution (i.e. $X_1, X_2, \dots, X_n \overset{iid}{\sim} f(x)$)
- A statistic is a function of a random sample that is observable

$T=f(X_1, X_2, \dots, X_n)$

$T=\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$ is a statistic

$T=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)$ is not a statistic

- The sampling distribution of a statistic

i.e. $\bar{X}\sim N(\mu,\sigma_n^2)$

Estimation Methods

- Method of moments: equate sample and population moments of the same order, then solve for the estimators of parameters
- Maximum likelihood estimate:

The likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i;\vec{\theta})$$

Then we solve for $\vec{\theta}$ that maximizes $L(\vec{\theta})$

- Unbiased estimator: $E[\hat{\theta}-\theta]=0$
- Confidence interval: We say that $[L,U]$ is a $100(1-\alpha)\%$ confidence interval for a parameter θ if $P(L<\theta<U) = 1-\alpha$

Hypothesis Testing

- We usually formulate hypotheses

H_0: $\theta=\theta_0$

H_1: $\theta\neq\theta_0$ or $\theta>\theta_0$ or $\theta<\theta_0$

- Compute a test statistic T and the p-value is $P(T<t)$, where t is the critical value. Reject H_0 if the p-value is less than a significance level α
- Errors

	Accept	Reject
H_0 True	✓	Type I Error
H_0 False	Type II Error	✓

Random Vectors

- Random vectors are defined as a vector of random variables $\vec{X}=(X_1, X_2, \dots, X_p)$ and are associated with joint distributions, which is multivariate in nature
- The vector of random variables might consist of different outcomes taken from the same individuals, or a single variable/outcome measured at multiple time points
- If $\vec{X}=(X_1, X_2, \dots, X_p)$ is a random vector, its joint distribution is given by

$F_{\vec{X}}(x_1, x_2, \dots, x_p) = P(X_1\leq x_1, X_2\leq x_2, \dots, X_p\leq x_p)$

for $-\infty<x_i<+\infty$ and $i=1,2,\dots,p$

- The expected value of a random vector is given by

$E[\vec{X}] = (E[X_1], E[X_2], \dots, E[X_p])$

$= (\mu_1, \mu_2, \dots, \mu_p) = \vec{\mu}$

- For a random vector $\vec{X}=(X_1, X_2, \dots, X_p)$ with expected value $\vec{\mu}=(\mu_1, \mu_2, \dots, \mu_p)$, the 2^{nd} central moment, also known as the dispersion matrix, is given by

$E[(\vec{X}-\vec{\mu})(\vec{X}-\vec{\mu})']$

Note: if $\vec{X}$ is a column vector, then it is given by

$E[(\vec{X}-\vec{\mu})'(\vec{X}-\vec{\mu})]$

- Often denoted by

$$\Sigma_{p\times p} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & & \sigma_{pp} \end{pmatrix}$$

where $\sigma_{ii} = E[(X_i-\mu_i)^2] = \sigma_i^2$

and $\sigma_{ij} = E[(X_i-\mu_i)(X_j-\mu_j)]$

- Correlation matrix

$$\rho = (\rho_{ij}) = \frac{E[(X_i-\mu_i)(X_j-\mu_j)]}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

- Both Σ and ρ are symmetric and positive semi-definite matrices, in most practical cases they're positive definite
- The covariance matrix is often estimated by the sample moments corresponding to each of its elements
- The unbiased estimate of Σ is referred to as the sample covariance matrix

Moment Generating Functions

- For a univariate random variable

$M_X(t) = E[e^{tX}]$

- The mgf uniquely identifies distributions
- We also have

$$\frac{d^r}{dt^r} M_X(0) = E[X^r]$$

- The mgf of a random vector of length p (row vector) is defined as

$M_{\vec{X}}(\vec{t}) = M_{\vec{X}}(t_1, t_2, \dots, t_p) = E[e^{\vec{t}\vec{X}'}] = E[e^{\sum_{i=1}^{p} t_i X_i}]$

- We can obtain univariate mgf's by

$M_{X_i}(t_i) = M_{\vec{X}}(0, 0, \dots, 0, t_i, 0, \dots, 0)$

Random Matrices

- Matrix random variables are defined as

$$X_{p\times q} = \begin{pmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1q} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2q} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3q} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{p1} & X_{p2} & X_{p3} & \cdots & X_{pq} \end{pmatrix}$$

where the elements X_{ij}'s are themselves random variables

- The expectation for matrix random variables is

$$E[X] = \begin{pmatrix} E[X_{11}] & E[X_{12}] & E[X_{13}] & \cdots & E[X_{1q}] \\ E[X_{21}] & E[X_{22}] & E[X_{23}] & \cdots & E[X_{2q}] \\ E[X_{31}] & E[X_{32}] & E[X_{33}] & \cdots & E[X_{3q}] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E[X_{p1}] & E[X_{p2}] & E[X_{p3}] & \cdots & E[X_{pq}] \end{pmatrix}$$

- However, we run into problem when we try to consider higher order moments for matrix random variables
- The most common approach is to use the vectorized form of the matrix X, where the covariance matrix is represented as $\Sigma\otimes\Psi$, where Σ represents the covariance for column vectors and Ψ represents the covariance for row vectors

Multivariate Bias

- Suppose $\vec{\theta}=(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)$ is an estimator for a vector parameter $\vec{\theta}=(\theta_1, \theta_2, \dots, \theta_p)$. Multivariate bias is defined as

$E[\vec{\theta}-\vec{\theta}] = (E[\hat{\theta}_1]-\theta_1, E[\hat{\theta}_2]-\theta_2, \dots, E[\hat{\theta}_p]-\theta_p)$

- An estimator is referred to as an unbiased estimator if $E[\vec{\theta}-\vec{\theta}]=\vec{0}$
- Although the definition of bias is quite natural, we face a challenge when we attempt to compare two vector estimators with respect to their respective bias vectors
- For each vector estimator, we can calculate

$$D = \sqrt{(E[\vec{\theta}]-\vec{\theta})(E[\vec{\theta}]-\vec{\theta})'}$$

Compare the values for both estimators

- Can also use absolute bias across elements of the vector

Multivariate MSE

- For univariate estimators, it is defined as $E[(\hat{\theta}-\theta)^2]$ and can be rewritten as

$$MSE(\hat{\theta}) = E[(\hat{\theta}-E[\hat{\theta}])^2] + (E[\hat{\theta}-\theta])^2$$

$$= Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

- For vector estimators, the MSE is defined as

$MSE(\vec{\theta}) = E[(\hat{\theta}-\theta)'(\hat{\theta}-\theta)]$

This is an outer product, hence $MSE(\vec{\theta})$ is a matrix.

- We can rewrite the MSE as

$MSE(\vec{\theta}) = \Sigma_{\vec{\theta}} + Bias(\vec{\theta})' Bias(\vec{\theta})$

$\Sigma_{\vec{\theta}}$ variance covariance matrix of $\vec{\theta}$

- If we want to compare 2 vector estimators, we vectorize the matrix and calculate the Euclidean distance to compare MSE for two vector estimators
- $MSE(\vec{\theta})$ involves covariances in addition to variances, hence difficult to interpret
- To overcome difficulty in interpretation, we can use another definition of MSE

$MSE^{\#}(\vec{\theta}) = (Var(\hat{\theta}_1)+Bias(\hat{\theta}_1)^2, \dots, Var(\hat{\theta}_p)+Bias(\hat{\theta}_p)^2)$

This corresponds to the sum of the diagonal elements of $\Sigma_{\vec{\theta}}$ and the diagonal elements of $Bias(\vec{\theta})'Bias(\vec{\theta})$. Also equivalent to vector of element-wise MSEs

- Also easier to extend to matrix parameters

$$MSE = \begin{pmatrix} Var(\hat{\theta}_{11})+Bias(\hat{\theta}_{11})^2 & \cdots & Var(\hat{\theta}_{1p})+Bias(\hat{\theta}_{1p})^2 \\ \vdots & & \\ Var(\hat{\theta}_{p1})+Bias(\hat{\theta}_{p1})^2 & \cdots & Var(\hat{\theta}_{pp})+Bias(\hat{\theta}_{pp})^2 \end{pmatrix}$$

- Multivariate bias for matrix estimators is straightforward

$$Bias = \begin{pmatrix} E[\hat{\theta}_{11}-\theta_{11}] & \cdots & E[\hat{\theta}_{1p}-\theta_{1p}] \\ \vdots & & \vdots \\ E[\hat{\theta}_{p1}-\theta_{p1}] & \cdots & E[\hat{\theta}_{pp}-\theta_{pp}] \end{pmatrix}$$

- MSE for vector estimators are vectors. Use distance or average of element-wise MSEs to compare estimators
- Both bias and MSE for matrix estimators are matrices. We vectorize the matrices first and do the same as with vector estimators to compare estimators

Joint and Marginal Distributions

- If we have a random vector $\vec{X}=(X_1, X_2, \dots, X_p)$, the joint CDF is given by

$F_{\vec{X}}(x_1, x_2, \dots, x_p) = P(X_1\leq x_1, X_2\leq x_2, \dots, X_p\leq x_p)$

- The joint density function $f_{\vec{X}}(x_1, x_2, \dots, x_n)$ is defined in a way such that

$$F_{\vec{X}}(x_1, x_2, \dots, x_p) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p} f_{\vec{X}}(x_1, x_2, \dots, x_p)\,dx_p\cdots dx_2\,dx_1$$

- Similarly, the joint density function can be found from the CDF by differentiating w.r.t. all elements of the p-vector
- The marginal distribution of X_i in a p-variate random vector is obtained by taking p-1 integrals w.r.t. all the remaining random variables

Conditional Distributions and Independence

- Consider a p-variate vector $\vec{X}=(X_1, \dots, X_p)$, the conditional distribution of a sub-vector $\vec{X}^q=(X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_p)$ given X_i is

$$f_{\vec{X}^q|X_i}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p | X_i) = \frac{f_{\vec{X}}(x_1, x_2, \dots, x_p)}{f_{X_i}(x_i)}$$

- Conditional distribution of a sub-vector of random variables conditioned on another sub-vector of random variables can also be defined in a similar way
- A vector of random variables $\vec{X}=(X_1,\dots,X_p)$ are independent if and only if the joint distribution factorizes. That is

$$F_{\vec{X}}(x_1,x_2,\dots,x_p)=\prod_{i=1}^{p}F_{X_i}(x_i)$$

- The joint density also factorizes

$$f_{\vec{X}}(x_1,x_2,\dots,x_p)=\prod_{i=1}^{p}f_{X_i}(x_i)$$

- Alternatively, random variables are independent iff the joint MGF factorizes

$$M_{\vec{X}}(\vec{t})=\prod_{i=1}^{p}M_{X_i}(t_i)$$

- We also have if a vector of random variables $\vec{X}=(X_1,X_2,\dots,X_p)$ are independent, then

$$E[g(X_1)g(X_2)\cdots g(X_p)]=\prod_{i=1}^{p}E[g(X_i)]$$

- Independence also means that conditional distributions are the same as unconditional distributions

Properties of Expectations and Covariances

- Consider a random vector $\vec{X}=(X_1,X_2,\dots,X_p)$ and let $E[\vec{X}]=\vec{\mu}$, then

$$E[A\vec{X}']=AE[\vec{X}]=A\vec{\mu}'$$

for any non-random matrix $A:q\times p$, q any constant

- Similarly

$$E[\vec{X}B]=E[\vec{X}]B=\vec{\mu}B$$

for any non-random matrix $B:p\times r$, r any constant

- $E[\vec{X}']=E[\vec{X}]'$
- Let $Cov(\vec{X})=\Sigma$, then

$$Cov(A\vec{X}')=ACov(\vec{X}')A'=A\Sigma A'$$

for any non-random matrix $A:q\times p$

- Similarly, we can show that

$$Cov(\vec{X}B)=B'Cov(\vec{X})B=B'\Sigma B$$

for any non-random matrix $B:p\times r$

- $Cov(\vec{X})=Cov(\vec{X}')$
- Covariance of a non-random vector is 0
- Covariance matrix is always positive semi-definite

Multivariate Normal Distribution

Univariate normal distribution density

$$f_X(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)}$$

- A random vector $\vec{X}=(X_1,X_2,\dots,X_p)$ has a multivariate normal distribution iff every linear combination of $\vec{X}$ has a univariate normal distribution
- Multivariate normal distribution can also be defined with the aid of the univariate standard normal distribution, and by taking advantage of the fact that the normal distribution belongs to the location-scale family
- We first consider a random vector $\vec{U}=(U_1,U_2,\dots,U_p)$, $U_i\overset{iid}{\sim}N(0,1)$. We have

$$E[\vec{U}]=\vec{0}\quad\text{and}\quad Cov(\vec{U})=I_p$$

- The distribution of $\vec{U}$ is the same as the joint distribution of the U_i's and is given by

$$f_{\vec{U}}(u_1,u_2,\dots,u_p)=\prod_{i=1}^{p}f_{U_i}(u_i)=\left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}\sum_{i=1}^{p}u_i}$$

This can also be written as

$$\left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}\vec{u}\vec{u}'}=\left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}\mathrm{trace}(\vec{u}\vec{u}')}=\left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}\mathrm{trace}(\vec{u}'\vec{u})}$$

- This distribution is referred to as the standard p-variate normal distribution, often denoted by $\vec{U}\sim N_p(\vec{0},I_p)$
- A p-dimensional random vector $\vec{X}=(X_1,X_2,\dots,X_p)$ is said to have a multivariate normal distribution if it has the same distribution as $\vec{\mu}+\vec{U}B$, where $\vec{U}\sim N_p(\vec{0},I_p)$, B is a non-singular and symmetric matrix of dimension $p\times p$ and $\Sigma=B'B=BB$
- The multivariate normal distribution is often denoted as $\vec{X}\sim N_p(\vec{\mu},\Sigma)$ and its density function is given as

$$\left(\frac{1}{2\pi}\right)^{p/2}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})'}$$

- This can be derived using change-of-variables
- B is referred to as the square root of Σ, denote it by $\Sigma^{\frac{1}{2}}$
- The density function can also be written as

$$\left(\frac{1}{2\pi}\right)^{p/2}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathrm{trace}((\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})')}$$

- Using the trace property, we can also write

$$\left(\frac{1}{2\pi}\right)^{p/2}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathrm{trace}(\Sigma^{-1}(\vec{x}-\vec{\mu})'(\vec{x}-\vec{\mu}))}$$

- When $p=1$, the above reduces to univariate normal density

Properties of Multivariate Normal

- Consider $\underset{\sim}{X}\sim N_p(\underset{\sim}{\mu},\Sigma)$, we have
  - $E[\vec{X}]=\vec{\mu}$
  - $Cov(\vec{X})=\Sigma$

We can write Σ as

$$\Sigma_{p\times p}=\begin{pmatrix}\sigma_{11}&\sigma_{12}&\cdots&\sigma_{1p}\\ \sigma_{21}&\sigma_{22}&\cdots&\sigma_{2p}\\ \vdots&\vdots&\ddots&\vdots\\ \sigma_{p1}&\sigma_{p2}&\cdots&\sigma_{pp}\end{pmatrix}$$

- The multivariate normal distribution can also be defined using MGFs
- MGF of univariate standard normal: $M_U(t)=e^{\frac{1}{2}t^2}$
- If we have a random vector $\underset{\sim}{U}=(U_1,U_2,\dots,U_p)\sim N_p(\underset{\sim}{0},I_p)$, its mgf is $M_{\underset{\sim}{U}}(\underset{\sim}{t})=e^{\frac{1}{2}\underset{\sim}{t}\underset{\sim}{t}'}$, $\underset{\sim}{t}=(t_1,t_2,\dots,t_n)$
- If we have $\underset{\sim}{X}=\underset{\sim}{\mu}+\underset{\sim}{U}B$, we get $M_{\underset{\sim}{X}}(\underset{\sim}{t})=e^{\underset{\sim}{\mu}\underset{\sim}{t}'+\frac{1}{2}\underset{\sim}{t}\Sigma\underset{\sim}{t}'}$

This is the mgf for $\underset{\sim}{X}\sim N_p(\underset{\sim}{\mu},\Sigma)$

- Normal distribution is closed under linear combination
- Theorem: If $\underset{\sim}{X}=(X_1,X_2,\dots,X_p)\sim N_p(\underset{\sim}{\mu},\Sigma)$ and $C:p\times n$, then

$$\underset{\sim}{X}C\sim N_p(\underset{\sim}{\mu}C,C'\Sigma C)$$

Can be used to prove normality of individual elements

- If we have k independent p-variate normal random vectors $\underset{\sim}{X}_i\sim N_p(\underset{\sim}{\mu}_i,\Sigma)$ for $i=1,2,\dots,k$ and $A_i:p\times q$. Then

$$\sum_{i=1}^{k}\underset{\sim}{X}_iA_i\sim N_q\left(\sum_{i=1}^{k}\underset{\sim}{\mu}_iA_i,\sum_{i=1}^{k}A_i'\Sigma A_i\right)$$

Limit Theorems

$X_1,X_2,\dots,X_n$ any random sample

$E[X_i]=\mu$ $\quad Var[X_i]=\sigma^2$

- Univariate Weak Law of Large Numbers:

$$\bar{X}\xrightarrow{P}\mu$$

- Univariate Central Limit Theorem:

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\xrightarrow{d}N(0,1)$$

Now, consider iid p-variate random vectors $\underset{\sim}{X}_i$ with mean $\underset{\sim}{\mu}$ and covariance matrix Σ

- Multivariate Weak Law of Large Numbers

$$\underset{\sim}{\bar{X}}=(\bar{X}_1,\bar{X}_2,\dots,\bar{X}_p)\xrightarrow{P}\underset{\sim}{\mu}$$

- Multivariate CLT

$$\sqrt{n}(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu})\xrightarrow{d}N_p(\underset{\sim}{0},\Sigma)$$

Wishart Distribution

- For univariate random variables, we have

$$X\sim N(0,1)\Rightarrow X^2\sim\chi^2(1)$$

$$X_1,\dots,X_n\overset{iid}{\sim}N(0,1)\Rightarrow\sum_{i=1}^{n}X_i^2\sim\chi^2(n)$$

$$X_1,\dots,X_n\overset{iid}{\sim}N(\mu,\sigma^2)\Rightarrow\frac{(n-1)S^2}{S^2}\sim\chi^2(n-1)$$

- The Wishart distribution is a multivariate extension of chi-squared distribution. Let

$$\underset{\sim}{X}=\begin{pmatrix}X_1\\X_2\\X_3\\\vdots\\X_p\end{pmatrix}\sim N_p(\underset{\sim}{0},I_p)$$

Then we get that $W=\underset{\sim}{X}\underset{\sim}{X}'$ has the Wishart distribution, written as $W_p(1,I_p)$

- If $\underset{\sim}{X}\sim N_p(\underset{\sim}{0},\Sigma)$, then $W=\underset{\sim}{X}\underset{\sim}{X}'\sim W_p(1,\Sigma)$
- If $\underset{\sim}{X}\sim N_p(\underset{\sim}{\mu},\Sigma)$, then $W=\underset{\sim}{X}\underset{\sim}{X}'$ has the non-central Wishart distribution
- If we let $\underset{\sim}{X}_1,\underset{\sim}{X}_2,\dots,\underset{\sim}{X}_n\overset{iid}{\sim}N_p(\underset{\sim}{\mu},\Sigma)$ be the columns of X, we can write $X\sim N_{p,n}(\mu^*,\Sigma,I_n)$ where $\mu^*=(\underset{\sim}{\mu},\underset{\sim}{\mu},\dots,\underset{\sim}{\mu})$
- A random matrix $W:p\times n$ is said to have the Wishart distribution if and only if it can be written as $W=XX'$, where $X\sim N_{p,n}(\mu^*,\Sigma,I_n)$, Σ positive definite nonsingular matrix. This is written as $W_p(n,\Sigma,\Delta)$, $\Delta=\mu^*\mu^{*\prime}$ non-centrality parameter. When $\mu^*=0$, we have the central wishart distribution $W_p(n,\Sigma)$
- We can write $W=XX'=\sum_{i=1}^{n}\underset{\sim}{X}_i\underset{\sim}{X}_i'$ $\quad\underset{\sim}{X}_i\overset{iid}{\sim}N(\underset{\sim}{\mu},\Sigma)$
- Elements of W are sums of independent random variables. Diagonal elements have $\chi^2(n)$ distribution
- Wishart density:

$$f_W(w)=c|\Sigma|^{-\frac{n}{2}}|W|^{\frac{(n-p-1)}{2}}e^{-\frac{1}{2}\mathrm{trace}(\Sigma^{-1}W)}$$

$$c=\left((\sqrt{2})^{np}\Gamma_p\left(\frac{n}{2}\right)\right)^{-1}\qquad\Gamma_p(\cdot)\text{ multivariate Gamma}$$

- Properties:
  - If $W_1\sim W_p(n,\Sigma,\Delta_1)$ and $W_2\sim W_p(m,\Sigma,\Delta_2)$ and independent, then we have $W_1+W_2\sim W_p(n+m,\Sigma,\Delta_1+\Delta_2)$
  - If $W\sim W_p(n,\Sigma,\Delta)$ and $A:r\times p$, then $AWA'\sim W_r(n,A\Sigma A',A\Delta A')$

Sampling from Multivariate Normal

- If we let $\underset{\sim}{X}_1,\underset{\sim}{X}_2,\dots,\underset{\sim}{X}_n\overset{iid}{\sim}N_p(\underset{\sim}{\mu},\Sigma)$, we get $\underset{\sim}{\bar{X}}=\frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{X}_i\sim N_p(\underset{\sim}{\mu},\frac{1}{n}\Sigma)$
- We can also get if $S=\frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})'$, we get $(n-1)S\sim W_p(n-1,\Sigma)$
- We also have that $\underset{\sim}{\bar{X}}$ and S are jointly sufficient for $\underset{\sim}{\mu}$ and Σ

Estimation of $\underset{\sim}{\mu}$ and Σ

- We can obtain that the method of moments estimators for $\underset{\sim}{\mu}$ and Σ are $\hat{\underset{\sim}{\mu}}=\underset{\sim}{\bar{X}}$ and $\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})(\underset{\sim}{X}_i-\underset{\sim}{\bar{X}})'$ by solving the system $E[\underset{\sim}{X}_i]=\underset{\sim}{\bar{X}}$ and $E[\underset{\sim}{X}_i\underset{\sim}{X}_i']=\frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{X}_i\underset{\sim}{X}_i'$
- If we want to do MLE, the likelihood is

$$L(\underset{\sim}{\mu},\Sigma|X)=\left(\frac{1}{\sqrt{2\pi}}\right)^{np}|\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\underset{\sim}{x}_i-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}_i-\underset{\sim}{\mu})\right)$$

- The log-likelihood is given by

$$\ell(\underset{\sim}{\mu},\Sigma|X)=-np\log(\sqrt{2\pi})-\frac{n}{2}\log|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}(\underset{\sim}{x}_i-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}_i-\underset{\sim}{\mu})$$

- If we take derivatives w.r.t. $\underset{\sim}{\mu}$, we get

$$\frac{\partial}{\partial\underset{\sim}{\mu}}\ell(\underset{\sim}{\mu},\Sigma|X)=-\frac{1}{2}\sum_{i=1}^{n}-2\Sigma^{-1}(\underset{\sim}{x}_i-\underset{\sim}{\mu})$$

If we set this to 0, we get $\hat{\underset{\sim}{\mu}}=\underset{\sim}{\bar{X}}$

- To get the MLE for Σ, we need the following matrix derivatives

$$\frac{\partial|A|}{\partial A}=(A')^{-1}|A|\quad\text{and}\quad\frac{\partial\,\mathrm{trace}(AB)}{\partial B}=A'$$

Rewrite the log-likelihood as

$$\ell(\underset{\sim}{\mu},\Sigma|X)=-np\log(\sqrt{2\pi})+\frac{n}{2}\log|\Sigma^{-1}|-\frac{1}{2}\mathrm{trace}\left((\underset{\sim}{x}_i-\underset{\sim}{\mu})(\underset{\sim}{x}_i-\underset{\sim}{\mu})'\Sigma^{-1}\right)$$

We can get

$$\frac{\partial\ell}{\partial\Sigma^{-1}}=\frac{n}{2}\frac{1}{|\Sigma^{-1}|}\cdot\Sigma\cdot|\Sigma^{-1}|-\frac{1}{2}\sum_{i=1}^{n}(\underset{\sim}{x}_i-\underset{\sim}{\mu})(\underset{\sim}{x}_i-\underset{\sim}{\mu})'$$

If we set this to 0, we get

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{x}_i-\underset{\sim}{\mu})(\underset{\sim}{x}_i-\underset{\sim}{\mu})'$$

- Note: $\hat{\Sigma}$ is not an unbiased estimator for Σ.

Hypothesis Testing for $\underset{\sim}{\mu}$

- For univariate case, if we have a sample from $N(\mu,\sigma^2)$, where σ^2 is fixed but unknown. If we want to test

$$H_0:\mu=\mu_0\quad\text{vs.}\quad H_1:\mu\neq\mu_0$$

We reject H_0 when

$$T=\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right|>t_{1-\frac{\alpha}{2}}(n-1)$$

We also have $T^2\sim F_{1,n-1}$, so we can also reject H_0 when $T^2>F_{1-\alpha}(1,n-1)$

- For multivariate normal, if we want to test

$$H_0:\underset{\sim}{\mu}=\underset{\sim}{\mu}_0\quad\text{vs.}\quad H_1:\underset{\sim}{\mu}\neq\underset{\sim}{\mu}_0$$

we use the Hotelling's T^2 statistic

$$T^2=n(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu}_0)'S^{-1}(\underset{\sim}{\bar{X}}-\underset{\sim}{\mu}_0)$$

- The Hotelling's T^2 statistic is also the likelihood ratio test statistic
- The distribution for T^2 is not a standard distribution, so we need to scale it
- Roy's Union-Intersection Test statistic

$$RF = \left(\frac{n-p}{p(n-1)}\right) T^2 \sim F(p, n-p)$$

- We reject H_0 when $RF > F_{1-\alpha}(p, n-p)$

Two Sample Test

- Let $X = (\underset{\sim}{x}_1\ \underset{\sim}{x}_2 \dots \underset{\sim}{x}_{n_1})$, $Y = (\underset{\sim}{y}_1\ \underset{\sim}{y}_2 \dots \underset{\sim}{y}_{n_2})$ random samples from two independent p-variate normal distributions with common covariance matrix Σ. Say $N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $N_p(\underset{\sim}{\mu}_2, \Sigma)$, respectively
- We want to test the following:

$$H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 \quad \text{vs.} \quad H_1: \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$$

- For univariate case, we reject H_0 when

$$T = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \left| \frac{\bar{x}-\bar{y}}{s_p} \right| > t_{1-\frac{\alpha}{2}}(n_1+n_2-2)$$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

s_1^2, s_2^2 sample variance for groups

- We can also reject H_0 when

$$T^2 > F_{1-\alpha}(1, n_1+n_2-2)$$

- For multivariate case, we have the two sample Hotelling's T^2 statistic is

$$T^2 = \left(\frac{n_1 n_2}{n_1+n_2}\right)(\bar{x}-\bar{y})' S_p^{-1} (\bar{x}-\bar{y})$$

$$S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

S_1, S_2 sample covariance matrices for groups

- Two-sample Roy's UIT

$$RF = \left(\frac{f-p+1}{fp}\right) T^2 \sim F(p, f-p+1)$$

$f = n_1+n_2-2$

Reject H_0 when $RF > F_{1-\alpha}(p, f-p+1)$

Testing for Normality

- Shapiro-Wilk's test: Consider a univariate random sample $X_1, X_2, \dots, X_n$ and suppose $Y_1, Y_2, \dots, Y_n$ are the corresponding order statistics. The Shapiro-Wilk's test for normality is given by

$$W = \frac{\sum_{i=1}^{n} a_i y_{(i)}}{s}$$

s sample standard deviation
a_i function of order statistic of standard normal distribution

- Kolmogorov-Smirnov's test: uses the difference between the empirical distributions of two random samples, and reject when the difference is large

$$D = \sup_x |F_Y(x) - F_W(x)|$$

reject when D is large

- To test for normality, compare the empirical distribution corresponding to the data to standard normal
- In practice, we usually use graphical methods, by themselves or combined with formal tests

Testing Multivariate Normality

- One way is to vectorize the random sample $X = (\underset{\sim}{x}_1\ \underset{\sim}{x}_2 \cdots \underset{\sim}{x}_n)$.
- We have if $\underset{\sim}{x}_i \overset{iid}{\sim} N_p(\underset{\sim}{\mu}, \Sigma)$, then $X \sim N_{p,n}(\underset{\sim}{\mu}^*, \Sigma, I_n)$. Then we have

$$Vec(X) \sim N_{pn}(Vec(\underset{\sim}{\mu}^*), \Sigma \otimes I_n)$$

- Then we can use univariate tests and traditional graphical methods like qq-plot to test for normality
- Limitation: vectorized data does not necessarily consist of independent samples
- Another way is to use Small's graphical method. Let c_i be defined as

$$c_i = \frac{n}{(n-1)^2}(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})' S^{-1} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})$$

- When $\underset{\sim}{x}_i \overset{iid}{\sim} N_p(\underset{\sim}{\mu}, \Sigma)$, it can be shown that c_i's are approximately distributed as $Beta(\frac{1}{2}, \frac{n-1-p}{2})$.

We can use Kolmogorov-Smirnov's test as well as graphical approaches to test