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
  - $E[\vec{X}]=\underset{\sim}{\mu}$
  - $Cov(\vec{X})=\Sigma$

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

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

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

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

- 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

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

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

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

$$D = \sup |F_N(x) - F_M(x)|$$

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 $X \sim N_{p,n}(\underset{\sim}{\mu}\underset{\sim}{1}_n', \Sigma, I_n)$. Then we have

$$Vec(X) \sim N_{pn}(Vec(\underset{\sim}{\mu}\underset{\sim}{1}_n'), I_n \otimes \Sigma)$$

- 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 $Beta(\frac{p}{2}, \frac{n-1-p}{2})$. We can use Kolmogorov-Smirnov's test as well as graphical approaches to test

ANOVA

- For comparing means between groups

Group 1	Group 2	⋯	Group k
y_{11}	y_{12}	⋯	y_{1k}
y_{12}	y_{22}	⋯	y_{2k}
⋮	⋮	⋱	
$y_{n_1,1}$	$y_{n_2,2}$	⋯	$y_{n_k,k}$

- k groups with n_j individuals in j^{th} group and $n = n_1+n_2+\dots+n_k$. y_{ij} is outcome for i^{th} individual in j^{th} group
- Can be represented as

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$
$$j = 1, 2, \dots, k \qquad i = 1, 2, \dots, n$$

- To test hypotheses

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$
$H_1: \mu_\ell \neq \mu_j$ for at least one $\ell \neq j$

We have

$$SST = \sum_{i,j}(y_{ij} - \bar{y})^2$$

$$SSE = \sum_{i,j}(y_{ij} - \bar{y}_j)^2$$

$$SSG = \sum_j n_j(\bar{y}_j - \bar{y})^2$$

$$SST = SSE + SSG$$

The hypothesis is tested using

$$F = \frac{SSG/k-1}{SSE/n-k} \sim F_{k-1, n-k}$$

ANOVA Matrix Formulation

- Can be written as

$$\underset{\sim}{Y}_{1\times n} = \underset{\sim}{B}_{1\times k} X_{k\times n} + \underset{\sim}{\varepsilon}_{1\times n}$$

$$\underset{\sim}{Y} = (y_{11}, y_{21}, \dots, y_{n_1,1}, \dots y_{1k}, y_{2k}, \dots, y_{n_k,k})$$

$$\underset{\sim}{B} = (\beta_1, \beta_2, \dots, \beta_k)$$

$$X = \begin{pmatrix} \underset{\sim}{1}_{n_1} & \underset{\sim}{0}_{n_1} & \cdots & \underset{\sim}{0}_{n_1} \\ \underset{\sim}{0}_{n_2} & \underset{\sim}{1}_{n_2} & \cdots & \underset{\sim}{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \underset{\sim}{0}_{n_k} & \underset{\sim}{0}_{n_k} & \cdots & \underset{\sim}{1}_{n_k} \end{pmatrix}$$

$$\underset{\sim}{\varepsilon} = (\varepsilon_{11}, \varepsilon_{12}, \dots, \varepsilon_{n_1,1}, \dots \varepsilon_{1k}, \varepsilon_{2k}, \dots, \varepsilon_{n_k,k})$$

- $\underset{\sim}{1}_{n_i}$ and $\underset{\sim}{0}_{n_i}$ are row vectors of 1s and 0s with length n_i;
- Matrix X referred to as design matrix
- We assume $E[\underset{\sim}{\varepsilon}] = 0$, $Cov(\underset{\sim}{\varepsilon}) = \sigma^2 I$
- Normality assumption: $\underset{\sim}{\varepsilon} \sim N_n(0, \sigma^2 I)$
  This also means $\underset{\sim}{Y} \sim N_n(\underset{\sim}{B}X, \sigma^2 I)$
- $\underset{\sim}{B}$ has the same interpretation as in regression models in general
- When dummy variables are involved, we need to be careful with interpretations of elements in $\underset{\sim}{B}$

ANOVA Estimation

- We want to minimize

$$Q = (\underset{\sim}{Y} - \underset{\sim}{B}X)(\underset{\sim}{Y} - \underset{\sim}{B}X)'$$

- Then we get if we minimize Q, we get

$$\hat{\underset{\sim}{B}} = \underset{\sim}{Y}X'(XX')^{-1} \text{ and } \hat{\underset{\sim}{Y}} = \underset{\sim}{Y}X'(XX')^{-1}X = \underset{\sim}{Y}H$$

- Residuals: $\underset{\sim}{Y} - \hat{\underset{\sim}{Y}} = \underset{\sim}{Y}(1-H)$

MANOVA model

Group 1	Group 2	⋯	Group k
$y_{11}, y_{12}, \dots, y_{1p}$	$y_{21}, y_{22}, \dots, y_{2p}$		$y_{k1}, y_{k2}, \dots, y_{kp}$
$y_{21}, y_{22}, \dots, y_{2p}$	$y_{21}, y_{22}, \dots, y_{2p}$		$y_{k1}, y_{k2}, \dots, y_{kp}$
⋮	⋮		⋮
$y_{n_1,1}, y_{n_1,2}, \dots, y_{n_1,p}$	$y_{n_2,1}, y_{n_2,2}, \dots, y_{n_2,p}$		$y_{n_k,1}, y_{n_k,2}, \dots, y_{n_k,p}$

The model is given by: $\underset{p\times n}{Y} = \underset{p\times k}{B}\ \underset{k\times n}{X} + \underset{p\times n}{E}$

where

$$Y = \begin{pmatrix} y_{11} & y_{21} & \cdots & y_{n_1,1} & \cdots & y_{1k} & \cdots & y_{n_k,k} \\ y_{12} & y_{22} & \cdots & y_{n_1,2} & \cdots & y_{1k} & \cdots & y_{n_k,k} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ y_{1p} & y_{2p} & \cdots & y_{n_1,p} & \cdots & y_{1k,p} & \cdots & y_{n_k,p} \end{pmatrix}$$

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2k} \\ \vdots & \vdots & & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pk} \end{pmatrix} \qquad X = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & & \vdots \\ 0_{n_k} & 0_{n_k} & \cdots & 1_{n_k} \end{pmatrix}$$

$$E = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{21} & \cdots & \varepsilon_{n_1,1} & \cdots & \varepsilon_{1k,1} & \cdots & \varepsilon_{n_k,1} \\ \varepsilon_{12} & \varepsilon_{22} & \cdots & \varepsilon_{n_1,2} & \cdots & \varepsilon_{1k,2} & \cdots & \varepsilon_{n_k,2} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ \varepsilon_{1p} & \varepsilon_{2p} & & \varepsilon_{n_1,p} & & \varepsilon_{1k,p} & \cdots & \varepsilon_{n_k,p} \end{pmatrix}$$

- We assume the columns of E follow $N_p(0, \Sigma)$
- Can write as $E \sim N_{p,n}(0, \Sigma, I_n)$
- We have $Y \sim N_{p,n}(BX, \Sigma, I_n)$

MANOVA Estimation

- Want to minimize

$$Q = EE' = (Y-BX)(Y-BX)'$$

- If we do the minimization, we get

$$\hat{B} = YX'(XX')^{-1} \text{ and } \hat{Y} = YX'(XX')^{-1}X = YH$$

- The variance-covariance matrix is estimated by

$$\hat{\Sigma} = \frac{(Y-\hat{B}X)(Y-\hat{B}X)'}{n-k} = \frac{RR'}{n-k}$$

- We can write the likelihood of multivariate normal as

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{np} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\text{trace}\left((Y-M^*)(Y-M^*)'\Sigma^{-1}\right)\right)$$

$M^* = (\mu, \mu, \dots, \mu)$ $\quad E[Y] = M^*$

- For MANOVA, it becomes

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{np} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\text{trace}\left((Y-BX)(Y-BX)'\Sigma^{-1}\right)\right)$$

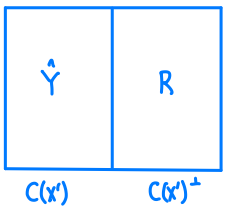
- Maximizing the likelihood also gives

$$\hat{B} = YX'(XX')^{-1} \qquad \hat{Y} = YX'(XX')^{-1}X = YH$$

- We also get the MLE for Σ is

$$\hat{\Sigma} = \frac{(Y-\hat{Y})(Y-\hat{Y})'}{n} = \frac{RR'}{n}$$

MANOVA Inference



$$\hat{B} = YX'(XX')^{-1} \qquad \hat{Y} = \hat{B}X = YX'(XX')^{-1}X = YH$$

$$R = Y - \hat{Y} = Y(1 - X'(XX')^{-1}X) = Y(1-H)$$

- $X'(XX')^{-1}X$ symmetric and idempotent, hence a projection matrix
- Projects Y onto the column space of X'
- The matrix $1 - X'(XX')^{-1}X$ projects onto the orthogonal complement of column space of X'
- $\hat{\Sigma} = S_p = \frac{1}{n-k}RR'$ is pooled sample variance-covariance matrix and is unbiased estimator of Σ
- We also have $YY' = \hat{Y}\hat{Y}' + RR'$
- $\hat{Y}$ and R are independent

MANOVA Tests

- Using vector formulation, the MANOVA model can be written as $\underset{\sim}{y}_{ij} = \underset{\sim}{\mu}_j + \underset{\sim}{\varepsilon}_{ij}$ where $\underset{\sim}{\varepsilon}_{ij} \sim N_p(0, \Sigma)$
- Matrix formulation: $Y = BX + E$, $\underset{\sim}{y}_{ij}$ corresponds to the columns of Y, ε_{ij} corresponds to columns of E
- Sample mean of group j: $\bar{\underset{\sim}{y}}_j$
- Overall mean. $\bar{\underset{\sim}{y}}$
- Then we have that

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}})(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}})'$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_j)(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_j)' + \sum_{j=1}^{k} n_j(\bar{\underset{\sim}{y}}_j - \bar{\underset{\sim}{y}})(\bar{\underset{\sim}{y}}_j - \bar{\underset{\sim}{y}})'$$

$$= SSE + SSG$$

- We have

$$YY' = Y\hat{Y}' + RR' \Rightarrow \underbrace{YCY'}_{SST} = \underbrace{\hat{Y}C\hat{Y}'}_{SSG} + \underbrace{RR'}_{SSE}$$

$C = [1 - \frac{1}{n}J]$

- We wish to test

$H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \dots = \underset{\sim}{\mu}_k$ vs. $H_1: \underset{\sim}{\mu}_j \neq \underset{\sim}{\mu}_\ell$ for at least one $j \neq \ell$

- Under H_0, MLE for Σ is

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}})(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}})'$$

$$\Rightarrow n\hat{\Sigma}_0 = YCY' = SST$$

- Under $H_0 \cup H_1$, MLE for Σ is

$$\hat{\Sigma}_1 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_j)(\underset{\sim}{y}_{ij} - \bar{\underset{\sim}{y}}_j)'$$

$$\Rightarrow n\hat{\Sigma}_1 = RR' = SSE$$

- The Likelihood-Ratio test is given by

$$\frac{|n\hat{\Sigma}_1|}{|n\hat{\Sigma}_0|} = \frac{|SSE|}{|SST|} = \frac{|SSE|}{|SSE+SSG|}$$

$H = SSG \qquad E = SSE$

$$\Lambda = \frac{|E|}{|E+H|} = \frac{1}{|1+HE^{-1}|} \quad \text{(Wilk's Lambda)}$$

- Now, if we want to test $H_0: B = 0$ vs $H_1: B \neq 0$
- The LR test is then

$$\frac{|RR'|}{|YY'|} = \frac{|RR'|}{|RR'+\hat{Y}\hat{Y}'|} = \frac{|SSE|}{|SSE+SSG|} = \frac{|SSE|}{|SST|}$$

Again, this is Wilk's Lambda

$$\Lambda = \frac{1}{|1+HE^{-1}|}$$

- This can also be written as

$$\Lambda = \frac{1}{\pi(1+\lambda_i)} \qquad \lambda_i \text{ eigenvalues of } HE^{-1}$$

Other Tests for MANOVA

- Lawley-Hotelling trace test:

$$\text{Trace}(HE^{-1}) = \sum \lambda_i$$

reject for large values

- Pillai's Test: $\text{Trace}(H(E+H)^{-1}) = \sum\left(\frac{\lambda_i}{1+\lambda_i}\right)$
- Roy's maximum root test: uses maximum eigenvalue of HE^{-1} λ_{max}

General Linear Hypothesis

- Test hypotheses:

$H_0: GBF = 0$ vs. $H_1: GBF \neq 0$

G and F conformable matrices

- Wilk's Lambda:

$$\Lambda = \frac{|E|}{|E+H|} = \frac{1}{|I+HE^{-1}|}$$

$$H = G\hat{B}F(F'(XX')^{-1}F)^{-1}F'\hat{B}'G'$$

$$E = GY(I - X'(XX')^{-1}X)Y'G = GRR'G'$$

Longitudinal Data

Group 1	Group 2	...	Group k
$y_{11}, y_{12}, \ldots, y_{1p}$	$y_{21}, y_{22}, \ldots, y_{2p}$	...	$y_{n1}, y_{n2}, \ldots, y_{np}$
$y_{21}, y_{22}, \ldots, y_{2p}$	$y_{22}, y_{22}, \ldots, y_{22p}$	...	$y_{2n}, y_{2n2}, \ldots, y_{2np}$
$\vdots$			
$y_{n1}, y_{n2}, \ldots, y_{np}$	$y_{n21}, y_{n22}, \ldots, y_{n2p}$	...	$y_{nn1}, y_{nn2}, \ldots, y_{nnp}$

- Repeated measurements are from the same outcome at different time points
- There is often a functional relationship between the outcome and time
- MANOVA accounts for the correlation within individuals, but it assumes unstructured mean
- Time dependency often exists in most real life scenarios
- Repeated measurements can also be taken w.r.t. other continuous variables

GMANOVA Models

- GMANOVA arises as a linearly constrained MANOVA model
- Structure in the mean is induced in the model through a second design matrix
- GMANOVA accounts not only for the correlation between measurements across time, but it also allows us to model the response curves
- Often referred to as growth curve model (GCM), or multivariate bilinear regression model
- GMANOVA assumes the mean for each group to follow a polynomial function (of time) of degree $q-1$

$$E[y_j|t] = \beta_{0,j} + \beta_{1,j}t + \ldots + \beta_{q-1,j}t^{q-1}, \quad j = 1, 2, \ldots, k$$

Note: Equal degree of polynomials assumed

- Using matrix setup, the model can be formulated as

$$\underset{p\times n}{Y} = \underset{p\times q}{Z}\underset{q\times k}{B}\underset{k\times n}{X} + \underset{p\times n}{E}$$

- Y: observation/outcome matrix
- B: parameter matrix
- Z: time dependency within individual design matrix
- X: between individual design matrix

- $E \sim N_{p,n}(0, \Sigma, I_n)$ and $Y \sim N_{p,n}(ZBX, \Sigma, I_n)$
- It is assumed $p \geq q$ and $n \geq p(X) + p$
- $n = n_1 + n_2 + \ldots + n_k$

- We have the matrices are

$$Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1n} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{p1} & y_{p2} & y_{p3} & \cdots & y_{pn} \end{pmatrix}$$

$$B = \begin{pmatrix} b_{01} & b_{02} & b_{03} & \cdots & b_{0k} \\ b_{11} & b_{12} & b_{13} & \cdots & b_{1k} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ b_{q-1,1} & b_{q-1,2} & b_{q-1,3} & \cdots & b_{q-1,k} \end{pmatrix}$$

$$Z = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{q-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{q-1} \\ 1 & t_3 & t_3^2 & \cdots & t_3^{q-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^{q-1} \end{pmatrix}$$

$$X = \begin{pmatrix} \frac{1}{n_1}\underset{\sim}{1}_{n_1} & \underset{\sim}{0}_{n_2} & \underset{\sim}{0}_{n_3} & \cdots & \underset{\sim}{0}_{n_k} \\ \underset{\sim}{0}_{n_1} & \frac{1}{n_2}\underset{\sim}{1}_{n_2} & \underset{\sim}{0}_{n_3} & \cdots & \underset{\sim}{0}_{n_k} \\ \underset{\sim}{0}_{n_1} & \underset{\sim}{0}_{n_2} & \frac{1}{n_3}\underset{\sim}{1}_{n_3} & \cdots & \underset{\sim}{0}_{n_k} \\ \vdots & \vdots & \vdots & & \vdots \\ \underset{\sim}{0}_{n_1} & \underset{\sim}{0}_{n_2} & \underset{\sim}{0}_{n_3} & \cdots & \frac{1}{n_k}\underset{\sim}{1}_{n_k} \end{pmatrix}$$

Inference for GMANOVA

- Consider GCM $Y = ZBX + E$
- First approach involves a linear transformation by pre-multiplying the model using the matrix $(Z'V^{-1}Z)^{-1}Z'V^{-1}$ where V is an arbitrary non-singular matrix
- The transformation leads to the MANOVA model

$$Y^* = BX + E^*$$

$$Y^* = (Z'V^{-1}Z)^{-1}Z'V^{-1}Y$$

$$E^* = (Z'V^{-1}Z)^{-1}Z'V^{-1}E$$

- Y^* and E^* normally distributed
- MLE of B is

$$\hat{B} = (Z'V^{-1}Z)^{-1}Z'V^{-1}YX'(XX')^{-1}$$

- Fitted values and residuals

$$\hat{Y} = Z\hat{B}X = Z(Z'V^{-1}Z)^{-1}Z'V^{-1}YX'(XX')^{-1}X$$

$$R = Y - \hat{Y} = Y - Z(Z'V^{-1}Z)^{-1}Z'V^{-1}YX'(XX')^{-1}X$$

- MLE of Σ: $\hat{\Sigma} = \frac{1}{n}RR'$
- When V is non-random, R is a linear combination of a multivariate normal random variable, and hence is normally distributed

Bilinear Projections

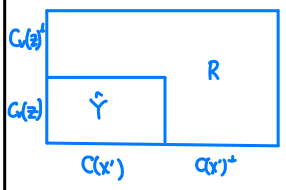
- If we look at the fitted values and residuals we obtained from Potthoff and Roy's transformations

$$\hat{Y} = Z(Z'V^{-1}Z)^{-1}ZV^{-1}YX'(XX')^{-1}X$$

$$= P_Z Y P_X$$

$$R = Y - P_Z Y P_X$$

$$P_Z = Z(Z'V^{-1}Z)^{-1}ZV^{-1} \quad P_X = X'(XX')^{-1}X$$



- Decompositions of orthogonal complement to the design space has been done to facilitate usage and interpretation of residuals
- MLEs of parameters given by

$$\hat{B} = (Z'S^{-1}Z)^{-1}Z'S^{-1}YX'(XX')^{-1}$$

$$S = Y(I - X'(XX')^{-1}X)Y$$

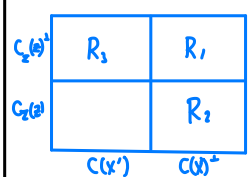
- Estimated mean trajectory

$$\hat{Y} = Z(Z'S^{-1}Z)^{-1}Z'S^{-1}YX'(XX')^{-1}X$$

$$= P_{Z,S}YP_X$$

$$P_{Z,S} = Z(Z'S^{-1}Z)^{-1}Z'S^{-1}$$

- The vector operator can be applied and the residual space can be decomposed into 3 orthogonal spaces



$$R_1 = (I - P_{Z,S})Y(I - P_X)$$

$$R_2 = P_{Z,S}Y(I - P_X)$$

$$R_3 = (I - P_{Z,S})YP_X$$

Dimension Reduction

- One challenge with multivariate data and its analysis is difficulty in data visualization
- Understanding relationships between variables within the same data matrix or between data matrices is often challenging
- In some cases, even pairwise explorations are not feasible
- High-dimensionality can also cause a problem in subsequent confirmatory analyses
- Too many variables can cause curse of dimensionality
- Principal component analysis (PCA) is one of the statistical methods we use to overcome the challenges
- Exploratory in nature and is often used as dimension reduction and data visualization technique

Principal Component Analysis

- Involves projections/transformations where new sets of uncorrelated variables are created
- Main objective: reduce dimension of data while keeping much of the variation in the original data
- New variables are linear combinations of the original variables such that the variances of them are sequentially maximized
- Suppose $\underset{\sim}{X} = (X_1, X_2, \ldots, X_p)'$ is a random vector with $E[\underset{\sim}{X}] = \underset{\sim}{\mu}$ and $Cov(\underset{\sim}{X}) = \Sigma$, the new set of variables are

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p = \underset{\sim}{a}_1'\underset{\sim}{X}$$

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p = \underset{\sim}{a}_2'\underset{\sim}{X}$$

$$\vdots$$

$$Z_p = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p = \underset{\sim}{a}_p'\underset{\sim}{X}$$

- Z_i's called principal components
- a_{ij}'s called loadings
- $\underset{\sim}{a}_i$'s called loading vectors
- Dimension reduction is achieved because most of the variation in the original data is often contained in the first few PCs, hence we often use the leading PCs in subsequent confirmatory analysis
- We want to transform the data such that the linear combinations maximize the variance sequentially.
- The first PC is obtained by maximizing its variance, while putting constraint $\underset{\sim}{a}'\underset{\sim}{a} = 1$
- Variance of first PC:

$$Var(Z_1) = \underset{\sim}{a}_1'\Sigma\underset{\sim}{a}_1$$

- For the i^{th} PC, we maximize the variance $(\underset{\sim}{a}_i'\Sigma\underset{\sim}{a}_i)$ under the constraints $\underset{\sim}{a}_i'\underset{\sim}{a}_i = 1$ and $\underset{\sim}{a}_i'\underset{\sim}{a}_j = 0$ for any $j < i$

Obtaining Loading Vectors

- Σ can be decomposed as $\Gamma\Lambda\Gamma'$, where Γ is orthogonal and columns are eigenvectors of Σ, Λ is diagonal matrix with eigenvalues of Σ on the diagonals
- We have

$$Var(Z_1) = \frac{\underset{\sim}{a}_1'\Sigma\underset{\sim}{a}_1}{\underset{\sim}{a}_1'\underset{\sim}{a}_1}$$

$$= \frac{\underset{\sim}{a}_1'\Gamma\Lambda\Gamma'\underset{\sim}{a}_1}{\underset{\sim}{a}_1'\Gamma\Gamma'\underset{\sim}{a}_1'}$$

$$= \frac{\underset{\sim}{b}_1'\Lambda\underset{\sim}{b}_1}{\underset{\sim}{b}_1'\underset{\sim}{b}_1} \qquad \underset{\sim}{b}_1 = \Gamma'\underset{\sim}{a}_1$$

$$= \frac{\sum\lambda_i b_{1i}^2}{\sum b_{1i}^2} \leq \frac{\lambda_1\sum b_{1i}^2}{\sum b_{1i}^2} = \lambda_1$$

λ_1 is maximum eigenvalue of Σ

- Equality is achieved when $\underset{\sim}{a}_1 = \underset{\sim}{\gamma}_1$, $\underset{\sim}{\gamma}_1$ eigenvector of Σ corresponding to λ_1

- Variance of first PC is the largest eigenvalue and maximum variance is achieved by choosing the loading vector $\underset{\sim}{a}_1'$ to be the normalizing eigenvector corresponding to λ_1
- Similarly, if we use $\underset{\sim}{a}_2 = \underset{\sim}{\gamma}_2$, we get the variance to be λ_2, second largest eigenvalue of Σ
- Then if $\underset{\sim}{a}_i = \underset{\sim}{\gamma}_i$, the variance is λ_i
- The loading vectors are orthogonal to each other, ensuring the PCs are uncorrelated
- We also have $Var(Z_1) \geq Var(Z_2) \geq \ldots \geq Var(Z_p)$
- The PCs preserve the total variation contained in the original data

$$Var(Z_i) = \lambda_i; \quad \sum_{i=1}^{p} Var(Z_i) = trace(\Sigma) = \sum_{i=1}^{p} Var(X_i)$$

- In practice, we don't use all the PCs, hence there will be a loss of some information
- Proportion of total variance contained in k^{th} PC is

$$PV_k = \frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$$

- All of the above also valid for using correlation matrix. In practice, the correlation matrix is used more since variables often have different units of measurements and levels of variability
- Σ often unknown, so we use sample covariance matrix S or sample correlation matrix R
- PCs obtained using S or Σ are in most cases, different from R or P
- Total variation explained by PCs obtained from S or Σ are also different from those obtained using R or P

PCA of Random Sample

- Consider a random sample $X = (\underset{\sim}{x}_1\ \underset{\sim}{x}_2 \ldots \underset{\sim}{x}_p)$ where $E[\underset{\sim}{x}_i] = \underset{\sim}{\mu}$ and $Cov(\underset{\sim}{x}_i) = \Sigma$
- The new, transformed data is calculated from the PCs and is given by

$$Z = (\underset{\sim}{z}_1\ \underset{\sim}{z}_2 \cdots \underset{\sim}{z}_p) = X\hat{\Gamma}$$

$\hat{\Gamma}$ matrix of normalized eigenvectors of $\hat{\Sigma}$

Cluster Analysis

- An exploratory statistical method used to uncover categories in the data, where individuals are homogenous within a group and are different between groups
- Often referred to as class discovery
- Unsupervised: group identifier often unknown and is not used in the analysis
- Assign individuals to groups (clusters) with the objective of maximizing homogeneity (similarity) within groups/clusters and maximizing difference between them
- Patterns in the data and potential clusters can sometimes be discovered through other exploratory analysis
- There may also be prior knowledge/evidence that motivates the study leading to class discovery
- Cluster analysis can also be simply used as a means of hypothesis generating

Distances

- The key in cluster analysis is similarity (homogeneity) and separation/heterogeneity (distance). We sometimes use the word dissimilarity to refer to distance matrices
- Most commonly used distance matrices are based on the Euclidean distance, and those based on the correlation or covariance matrix
- There are situations where one needs to use other similarity matrices
- Choice of an appropriate similarity/distance matrix that optimally quantifies similarities between individuals is, therefore, fundamental to performing cluster analysis

Some common Distances

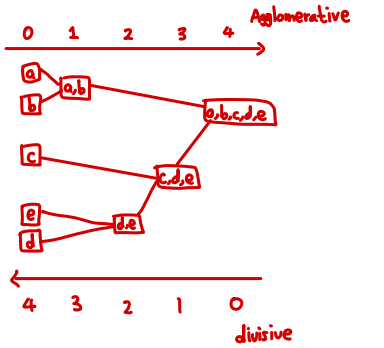
- For any distance function d, we need
  - $d \geq 0$
  - $d(x_i, x_i) = 0$
- Let $\underset{\sim}{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ represent p measurements for the i^{th} individual. The distance between individuals i and j can be calculated using
  - Euclidean Distance: $\sqrt{\sum_{k=1}^{p}(x_{ki} - x_{kj})^2}$
  - Manhattan Distance: $\sum_{k=1}^{p}|x_{ki} - x_{kj}|$
  - Correlation Based Distance: $1 - \rho(\underset{\sim}{x}_i, \underset{\sim}{x}_j)$
    $\rho(\underset{\sim}{x}_i, \underset{\sim}{x}_j)$ can be Pearson's, Spearman's, or Kendall's correlations
- For categorical variables, suppose we have binary data as follows

individual i \ individual j	Present(+)	Absent(-)
Present(+)	a	b
Absent(-)	c	d

- Jacard:
  - Similarity: $\frac{a}{a+b+c}$
  - distance: $1 - \frac{a}{a+b+c}$
- Simple Matching
  - Similarity: $\frac{a+d}{a+b+c+d}$
  - distance: $\frac{b+c}{a+b+c+d}$

Hierarchical Clustering

- 2 strategies: agglomerative or divisive
- Distances between individuals or clusters of individuals are calculated at each iteration



- Distance matrices are calculated iteratively until all observations are grouped together in one cluster (agglomerative clustering) or until all individuals become separate clusters on their own (divisive clustering)
- To calculate distances between clusters, the mechanism by what we calculate the distance between clusters at each iteration of the clustering algorithm is referred to as the linkage method
- Linkage methods:
  - Single Linkage: distance determined by considering the distance between the closest pair of individuals in the clusters
  - Complete Linkage: distance between the farthest pair of individuals in the clusters
  - Average Linkage: distances between all pairs of individuals are first calculated, then the distances between clusters are then determined by taking all the distances. Clusters with the smallest average distance are joined together in the subsequent iteration

Gower Similarity

- Useful when there's a mix of categorical and continuous variables
- Gower's similarity first calculates similarity with respect to each of the variables, and takes the average similarity across the variables
- Let S_{ijk} be the similarity between individuals i and j, with respect to the k^{th} variable. Gower's similarity between individuals i and j is then calculated as

$$S_{ij} = \frac{\sum_k w_{ijk} S_{ijk}}{\sum_k w_{ijk}}$$

$w_{ijk} = 0$ when a given variable is missing for either of the individuals, $w_{ijk} = 1$ otherwise

- To calculate S_{ijk}, then we have that
  - If the k^{th} variable is categorical, the simple matching similarity is used.
  - If k^{th} variable is continuous, then $S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$, where x_{ik} and x_{jk} are the values of the variable corresponding to the i^{th} and j^{th} individuals, respectively. R_k represents the range of the k^{th} variable
- Gower's distance: $d_{ij} = 1 - S_{ij}$

K-means Clustering

- Data first partitioned into k groups $C_1, C_2, \ldots, C_k$
- Individuals then assigned to the k clusters, where class assignment is determined by using some optimization criteria
- Most commonly used criteria is the group (within cluster) sum of squares (WSS), where data is partitioned in such a way that the total WSS is minimized
- General idea: find all possible partitions, and choose the one that leads to minimum total WSS
- Not computationally feasible
- The algorithm:
  1. Select initial partition of data, using initially guessed number of clusters k
  2. Move each of the individuals around and calculate the change in the total WSS for all the moves made
  3. Accept the move that lead to the maximum improvement in the WSS- these are our newly established clusters
  4. Repeat 2 and 3 until no improvement is possible

Canonical Correlation Analysis

- Involves projections in direction of maximum correlation between two sets of variables
- Involves 2 data matrices X and Y
- Main objective is to discover relationships between the variables in the 2 datasets
- CCA seeks to find linear combinations of variables that sequentially maximize correlation between the two sets of variables

Covariance Matrices

- Consider 2 random vectors of length p and q

$$\underset{\sim}{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad \underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_q \end{pmatrix}$$

- $\begin{pmatrix} \underset{\sim}{x} \\ \underset{\sim}{y} \end{pmatrix}$ combined vector
- The variance covariance matrix of the combined vector can be written as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Σ_{11} covariance matrix of $\underset{\sim}{x}$

Σ_{22} covariance matrix of $\underset{\sim}{y}$

$\Sigma_{12} = \Sigma_{21}'$ covariances between $\underset{\sim}{x}$ variables and $\underset{\sim}{y}$ variables

- Σ_{11} and Σ_{22} within covariances
- Σ_{12} and Σ_{21} cross covariances

Canonical Variates

- CCA provides two sets of linear combinations $z_i = \underset{\sim}{v}_i' \underset{\sim}{x}$ and $w_i = \underset{\sim}{u}_i' \underset{\sim}{y}$, where the column vectors $\underset{\sim}{v}_i$ and $\underset{\sim}{u}_i$ are obtained by sequentially maximizing the correlation between the z_i's and w_i's. The two pairs of linear combinations are given by:

$$z_1 = \underset{\sim}{v}_1' \underset{\sim}{x} \quad \cdots \quad z_r = \underset{\sim}{v}_r' \underset{\sim}{x}$$

$$w_1 = \underset{\sim}{u}_1' \underset{\sim}{y} \quad \cdots \quad w_r = \underset{\sim}{u}_r' \underset{\sim}{y}$$

$r = \min(p, q)$

- For instance, $\underset{\sim}{v}_1$ and $\underset{\sim}{u}_1$ are selected such that the correlation between z_1 and w_1 is maximized

$$Cor(z_1, w_1) = \frac{\underset{\sim}{v}_1' \Sigma_{12} \underset{\sim}{u}_1}{\sqrt{\underset{\sim}{v}_1' \Sigma_{11} \underset{\sim}{v}_1} \sqrt{\underset{\sim}{u}_1' \Sigma_{22} \underset{\sim}{u}_1}}$$

- In order to ensure that the correlation is bounded by 1, the above mathematical problem is maximized under the constraints $\underset{\sim}{v}_1' \Sigma_{11} \underset{\sim}{v}_1 = 1$ and $\underset{\sim}{u}_1' \Sigma_{22} \underset{\sim}{u}_1 = 1$
- Similarly, the k^{th} pair of linear combinations are derived by maximizing

$$Cor(z_k, w_k) = \frac{\underset{\sim}{v}_k' \Sigma_{12} \underset{\sim}{u}_k'}{\sqrt{\underset{\sim}{v}_k' \Sigma_{11} \underset{\sim}{v}_k} \sqrt{\underset{\sim}{u}_k' \Sigma_{22} \underset{\sim}{u}_k}}$$

under $\underset{\sim}{v}_k' \Sigma_{11} \underset{\sim}{v}_k = \underset{\sim}{u}_k' \Sigma_{22} \underset{\sim}{u}_k = 1$

- We also need z_i and z_k are uncorrelated for $i \neq k$, likewise with w_i and w_k

Some key Terms for CCA

- The linear combinations z_i and w_i are called canonical variates
- Correlations between pairs of canonical variates are called canonical correlations
- The vectors $\underset{\sim}{v}_i$ and $\underset{\sim}{u}_i$ are called canonical weight vectors or canonical coefficient vectors, and the elements of the loading vectors are referred to as canonical weights or canonical coefficients
- The correlations between the original variables and their corresponding canonical variates are referred to as canonical loadings, structural coefficients, or structural correlations
- The correlations between the $\underset{\sim}{x}$ variables and the $\underset{\sim}{w}$ variables and the correlations between the $\underset{\sim}{y}$ variables and $\underset{\sim}{z}$ variables are referred as to the canonical cross-loadings or canonical cross-correlations

CCA and Eigenvalues

- We can convert finding the linear combinations into a problem involving eigenvalues and eigenvectors
- Rewrite the correlation as

$$Cor(z_1, w_1) = \frac{\underset{\sim}{v}_1' \Sigma_{11}^{-\frac{1}{2}} \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} \Sigma_{22}^{-\frac{1}{2}} \underset{\sim}{u}_1}{\sqrt{\underset{\sim}{v}_1' \Sigma_{11}^{-\frac{1}{2}} \Sigma_{11}^{-\frac{1}{2}} \underset{\sim}{v}_1} \sqrt{\underset{\sim}{u}_1' \Sigma_{22}^{-\frac{1}{2}} \Sigma_{22}^{-\frac{1}{2}} \underset{\sim}{u}_1}}$$

Let $\underset{\sim}{a}' = \underset{\sim}{v}_1' \Sigma_{11}^{-\frac{1}{2}}$ and $\underset{\sim}{b} = \Sigma_{22}^{-\frac{1}{2}} \underset{\sim}{u}_1$. Then

$$Cor(z_1, w_1) = \frac{\underset{\sim}{a}' \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} \underset{\sim}{b}}{\sqrt{\underset{\sim}{a}'\underset{\sim}{a}} \sqrt{\underset{\sim}{b}'\underset{\sim}{b}}}$$

under $\underset{\sim}{a}'\underset{\sim}{a} = 1$ and $\underset{\sim}{b}'\underset{\sim}{b} = 1$

Let $\Omega = \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$. Then

$$Cor(z_1, w_1) = \frac{\underset{\sim}{a}' \Omega \underset{\sim}{b}}{\sqrt{\underset{\sim}{a}'\underset{\sim}{a}} \sqrt{\underset{\sim}{b}'\underset{\sim}{b}}}$$

We can obtain

$$Cor(z_1, w_1)^2 = (\underset{\sim}{a}' \Omega \underset{\sim}{b})^2 \leq \underset{\sim}{a}' \Omega \Omega' \underset{\sim}{a}$$

$$\Omega \Omega' = \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}$$

Let $\Psi = \Omega \Omega'$ and $\Psi = \Gamma \Lambda \Gamma'$ be its eigen decomposition. Then

$$Cor(z_1, w_1)^2 \leq \underset{\sim}{a}' \Omega \Omega' \underset{\sim}{a} = \underset{\sim}{a}' \Gamma \Lambda \Gamma' \underset{\sim}{a}'$$

$$= \sum_{i=1}^{n} \lambda_i a_i^2 \leq \lambda_1 \sum_{i=1}^{n} a_i^2 = \lambda_1$$

λ_i eigenvalues of $\Omega \Omega'$

λ_1 maximum eigenvalue

Then maximum correlation is $\sqrt{\lambda_1}$ and the standardized eigenvector corresponding to λ_1 satisfies $\underset{\sim}{a}'\underset{\sim}{a} = 1$

$$\underset{\sim}{a}' = \underset{\sim}{v}_1' \Sigma_{11}^{\frac{1}{2}} \Rightarrow \underset{\sim}{v}_1' = \underset{\sim}{a}_1' \Sigma_{11}^{-\frac{1}{2}}$$

Similarly, we can get

$$Cor(w_1, z_1) = \left(\frac{\underset{\sim}{a}' \Omega \underset{\sim}{b}}{\sqrt{\underset{\sim}{a}'\underset{\sim}{a}} \sqrt{\underset{\sim}{b}'\underset{\sim}{b}}} \right)^2 \leq \underset{\sim}{b}' \Omega' \Omega \underset{\sim}{b}$$

Leads to same eigenvalues but different eigenvectors

Choose $\underset{\sim}{b}$ standardized eigenvector of $\Omega'\Omega$ corresponding to λ_1

$$\underset{\sim}{b} = \Sigma_{22}^{\frac{1}{2}} \underset{\sim}{u}_1 \Rightarrow \underset{\sim}{u}_1 = \Sigma_{22}^{-\frac{1}{2}} \underset{\sim}{b}$$

To get k^{th} pair of canonical variates, we use the standardized eigenvectors corresponding to the k^{th} largest eigenvalues of $\Omega\Omega'$ and $\Omega'\Omega$

Correlations sequentially maximized

$$\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots \geq \sqrt{\lambda_r}$$

We can obtain that

$\Omega\Omega'$ and $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ are similar

$\Omega'\Omega$ and $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ are similar

Then we obtain

$\underset{\sim}{a} = \Sigma^{\frac{1}{2}} \underset{\sim}{a}'$ $\underset{\sim}{a}'$ eigenvector of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

$\underset{\sim}{b} = \Sigma^{\frac{1}{2}} \underset{\sim}{b}'$ $\underset{\sim}{b}'$ eigenvector of $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

We can then get

$\underset{\sim}{v}_1 = \underset{\sim}{a}'$ and $\underset{\sim}{u}_1 = \underset{\sim}{b}'$

CCA and SVD

- We can use the Singular value decomposition of $\Omega = \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} = PSQ'$
  - P and Q orthogonal matrices
  - S diagonal matrix with singular values as its diagonal elements
- Then we have the solutions are

$$V = \Sigma_{11}^{-\frac{1}{2}} P \qquad U = \Sigma_{22}^{-\frac{1}{2}} Q$$

where $V = (\underset{\sim}{v}_1\ \underset{\sim}{v}_2 \cdots \underset{\sim}{v}_r)$ and $U = (\underset{\sim}{u}_1\ \underset{\sim}{u}_2 \cdots \underset{\sim}{u}_r)$

Uncorrelatedness of Canonical Variates

- We can easily show the canonical variates $\underset{\sim}{z}$ and $\underset{\sim}{w}$ are uncorrelated and have unit variance

$$\begin{aligned} Cov(\underset{\sim}{z}) &= Cov(V'\underset{\sim}{x}) \\ &= V' Cov(\underset{\sim}{x}) V \\ &= P'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{11}\Sigma_{11}^{-\frac{1}{2}}P \\ &= P'P = I \end{aligned}$$

$$\begin{aligned} Cov(\underset{\sim}{w}) &= Cov(U'\underset{\sim}{y}) \\ &= U' Cov(\underset{\sim}{y}) U \\ &= Q'\Sigma_{22}^{-\frac{1}{2}}\Sigma_{22}\Sigma_{22}^{-\frac{1}{2}}Q \\ &= Q'Q = I \end{aligned}$$

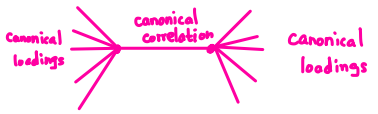
Some more Remarks on CCA

- We often use correlation matrices to conduct CCA because variables may be on different scales
- We cannot obtain the population covariance matrices in practice, so we can use the sample covariance matrices

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}$$

Presenting CCA Results

- The main CCA results are the canonical variates and the corresponding correlation between the pairs of canonical variates
- We also explore and interpret the contributions of the original set of variables to the corresponding canonical variates
- One way to present is using finger plots



- Canonical coefficients can also be plotted with bar graphs or a helio-plot