**Multivariate Testing Report**

Loc Nguyen

Vietnam Institute of Mathematics

*Draft version*

**Abstract**

Hypothesis testing is the most important subject in domain of statistics. Given a population with distribution *fϴ*(*X*) where *ϴ =* {*θ1, θ2,…, θk*} is a set of parameters. Given a data sample includes a lot of observations or sample points, *D* = (*X1, X2, …, Xm*). A function of sample points, e.g: or , is called a (sample) statistic. A statistic is considered an estimate of parameter *θi*. Hypothesis is defined as a statement about parameters of one or more populations. Hypothesis testing is the procedure that decides whether a hypothesis is accepted or rejected based on one or more sample statistic (s). If sample point *Xi* is single value, in other words, *fϴ* is univariate function, then this procedure is univariate hypothesis testing. Otherwise, sample point *Xi* is a vector and *fϴ* is multivariate function, this procedure is multivariate hypothesis testing. When sample point *Xi* = {*xi1, xi2,…, xin*} is a vector, data sample *D* = (*X1, X2, …, Xm*) is called *n*-dimension sample which is identical to vector space. Note that terms *n-dimension* and *multivariate* have the same meaning in this report.

Note that in this report terms *sample* and *data sample* have the same meaning. Each *sample point* is also called an *observation* or *data point*. Sample point *Xi* is considered as random variable in hypothesis testing context. Random variable *Xi* is referred as random vector in *n-*dimension vector space by default, *Xi* = {*xi1, xi2,…, xin*} if there is no additional explanation. In other words, sample can be considered as *m* rows and *n* columns matrix ***X*** = (*X1, X2, …, Xm*) composed of *m* row vector *Xi*. Bold and uppercase letter, for example ***X***, often denotes a multivariate sample or matrix. Lowercase letter, for example *xij*, often denotes the *jth* component of random vector *Xi* and such component is considered as partial singular random variable and so, terms *component variable* and *partial variable* are the same. However, some parameters are denoted by lowercase letters although they are vectors, for example, mean *μ*.

This report gives an overview of multivariate hypothesis testing in which some details of univariate (or scalar) testing are ignored. This report includes four sections.

* *Section 1* summarizes knowledge of matrix algebra. Almost statistical concepts in case of multivariate data relate to matrix algebra.
* *Section 2* describes multivariate probability distribution which is an extension of univariate distribution. Some popular distributions such as normal distribution and binominal distribution are extended in multi-dimension data.
* *Section 3* discusses statistical estimation in multi-dimension data. Please pay attention to this section because parameter estimation is the base of hypothesis testing.
* *Section 4* exploits aspects of multivariate hypothesis testing. This section is the most important part in this report.
* *Section 5* is the conclusion.

**1. Matrix algebra**

Statistical study has a strong attachment to matrix operations and so this section is a survey of matrix algebra and matrix calculus. Section has four parts: basic concepts of matrix, matrix analysis, matrix calculus and geometrical aspects of multivariate data. The last part gives us an imagination of multivariate distribution such as distance, space shape and structure, contour and etc.

**1.1. Basic concepts of matrix**

Vector *X* is a range of *n* values denoted or so-called *n-*dimension vector where *T* denotes transpose operation which change row to column and otherwise. Each *xi* is partial value (or partial component, partial element) of vector *X*. If *X* is random variable, *xi* is also random variable. The dot product or scalar product of two vectors *X* and *Y* denoted *XY* or *XYT* is a scalar value which is the sum of multiplications of their components.

The length (or module, magnitude, norm) of vector *X* is:

The cosine of angle between two vectors X and Y is the ratio of its product to multiplication of their lengths.

Cosine ranges in interval [*–1, 1*]. The cross product of two vectors *X* and *Y* denoted is the vector that is perpendicular to both *X* and *Y* according to right-hand rule. The length of cross product vector is:

Where *sine*(*X, Y*) is calculated via *cosine*(*X, Y*). Note that the semantic of cross product is the same to the semantic of normal multiplication like *2 \* 3 = 6*. Given an integer *k > 1*, the *kth* power of vector *X* is defined as below:

Where *Π* denotes cross product or scalar product (like = *Z* or *2 \* 3 = 6*) and so we infer that *Xk* is scalar if *k* is even and *Xk* is vector if *k* is odd. We have summarization table for *k = 2, 3, 4, 5* as following.

|  |  |  |
| --- | --- | --- |
| Power | Value | Type |
| *X2* |  | Scalar |
| *X3* |  | Vector |
| *X4* |  | Scalar |
| *X5* |  | Vector |

Matrix *A* is a table including *m* rows and *n* columns, whose cell or element is scalar value.

Matrix *A* denoted (*aij*) or *A*(*mxn*) can be considered as a set of *m* row vectors or a set of *n* column vectors. In multivariate analysis, matrix *A*(*mxn*) is considered as a set of *m* row vectors in *n*-dimension vector spaceif there is no additional explanation.

Matrix *A*(*nxn*) having the same number of rows and columns is called square matrix. Square matrix is very popular in statistical study. Vector can be considered as *1-*row or *1-*column matrix, so *X*(*1xn*) denote row vector and *X*(*nx1*) denotes column vector. By default, vector *X* is column vector if there is no additional note.

Given matrices *A*(*mxn*) and *B*(*mxn*) and scalar constant *c*, following are matrix operations.

*AT* = (*aji*) when *T* is transpose operation

*A + B =* (*aij) +* (*bij*)

*A – B =* (*aij*) + (*bij*)

*cA =* (*caij*)

*AB = A*(*mxk*)*B*(*kxn*) *= C*(*mxn*) *=* (*cij*) *=* ()

Given square matrix A(nxn), the *nth* power of *A* is

Following are properties of matrix operation.

*A + B = B + A*

*A*(*B + C*) *= AB + AC*

*A*(*BC*) *=* (*AB*)*C*

(*AT*)*T = A*

(*AB*)*T = BTAT*

Note that matrix multiplication is not commutative. Following is the list of some special vectors and matrices.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Definition** | **Notation** | **Example** |
| Scalar |  | *a* | *1* |
| Column vector |  | *a* |  |
| Row vector |  | *aT* | *(1, 2)* |
| Vector of ones |  | *1n* |  |
| Vector of zeros  or zero vector |  | *0n* |  |
| Square matrix |  | *A(nxn)* |  |
| Diagonal matrix | Square matrix and  *aij = 0, ,* | *diag*(*aii*) |  |
| Identity matrix | Diagonal matrix and  *aij = 1,* | *In* |  |
| Unit matrix | Square matrix and  *aij = 1* | *1n1nT* |  |
| Null matrix  or zero matrix | Square matrix and  *aij = 0* | *0n0nT* |  |
| Symmetric matrix | Square matrix and  *aij = aji* |  |  |
| Upper triangular matrix | *aij = 0,* |  |  |
| Idempotent matrix | Square matrix and  *AA = A* |  |  |
| Orthogonal matrix | Square matrix and  *AAT = In = ATA* |  |  |

Now we survey characteristics of matrix. The rank of matrix *A*(*mxn*) denotes *rank*(*A*) is the maximum number of linear independent rows or columns, *rank*(*A*) *min*(*m, n*). Non-singular matrix is square matrix and all its rows (columns) are independent; so the rank of non-singular is its number of rows (columns). The trace of matrix *A* denoted *tr*(*A*) is the sum of all is diagonal, *tr*(*A*) = . Following are properties of *trace* and *rank*.

*tr*(*A + B*) *= tr*(*A*) *+ tr*(*B*)

*tr*(*cA*) *= ctr*(*A*)

*tr*(*AB*) *= tr*(*BA*) *given A*(*mxn*) *and B*(*nxm*)

*tr*(*ABC*) *= tr*(*BCA*) *= tr*(*CBA*)given *A*(*mxn*)*, B*(*nxp*)and *C*(*pxm*)

*rank*(*A*) *min*(*m, n*) *given A*(*mxn*)

*rank*(*A*) *0*

*rank*(*A*) *= rank*(*AT*)

*rank*(*A*(*nxn*)) = *n* if *A* is non-singular

*rank*(*ATA*) *= rank*(*AAT*) *= rank*(*A*)

*rank*(*A+B*) *= rank*(*A*) *+ rank*(*B*)

*rank*(*AB*) *min*(*rank*(*A*)*, rank*(*B*))

*rank*(*ABC*) *= rank*(*B*)given *A*, *C* are non-singular

Given a square matrix *A*(*nxn*), determinant of matrix A denoted *det*(*A*) or |*A*| is defined as the sum over all permutations *σ* (s) of indexes {*1, 2,…, n*}. The *ith* value of permutation *σ* is denoted permutations *σi*. The set of all permutations *σ* (s) given *n* indexes {*1, 2,…, n*} is denoted *Sn*. *Sn* has *n!* elements.

Above formula is Leibniz formula where *sgn*(σ) denotes the sign of permutation σ.

Permutation σ is even (or odd) when the new permutation can be obtained by the even (or odd, respectively) number of switches of numbers. For example, given indexes {*1, 2, 3*}, the permutation (*123*) is even because there is *0* number of switches and the permutation (*132*) is even because there is *1* number of switches. According to Leibniz formula, the determinants of 2x2 and 3x3 matrices are:

*|A*(*2x2*)*| = a11a22 – a12a21*

*|A*(*3x3*)*| = a11a22a33 – a11a23a32 – a12a21a33 + a12a23a31 + a13a21a32 – a13a22a31*

Leibniz formula gets a huge of operators due to *n!* permutations. So matrix determinant is computed more effectively by applying Laplace expansion. Given element *aij* of square matrix *A*(*nxn*), the algebra complement denoted *Mij* is the sub-matrix including *n – 1* rows and *n – 1* columns, which is create by removing *ith* row and *jth* column from *A*.

According to Laplace extension, determinant of square matrix *A*(*nxn*) is computed followed determinant of algebra complement given arbitrary *ith* row.

Above formula is recursive formula with regard to column expansion. According to row expansion, determinant of square matrix *A*(*nxn*) given arbitrary *jth* column is:

Note that the expression (*–1*)*i+j|Mij|* is called cofactor *cij*. Cofactor matrix of *A* denoted *cofactor*(*A)* is defined as matrix whose elements are cofactor *cij* (s). The adjoint matrix of *A* denoted *adj*(*A*) is the transposition of cofactor matrix of *A*.

*adj*(*A*) = (*cofactor*(*A*))*T* = (*cji*)

Given square matrix *A*(nxn) and |*A*| 0, the inverse of *A* denoted *A-1* is the one that multiplication of A and itself is equal to identical matrix.

*AA-1 = A-1A = In*

The inverse *A-1* is determined according to adjoint matrix.

If there is existence of the inverse of *A*, in other words |*A*| 0, then A is invertible or non-singular. The generalized inverse (G-inverse) of matrix *A* denoted *A–* is the one satisfying following condition.

*AA–A = A*

Note that the inverse *A-1* is a concrete case of G-inverse *A–* and *A–* is the general concept of the inverse *A-1*. Following are properties of matrix determinant with note that determinant exists if and only if matrix is square matrix.

*|In| = 1*

*|AT| = |A|*

*|A–1| = |A|–1*

*|AB| = |A||B|* if both *A* and *B* are *nxn* matrices

*|cA| = cn|A|* where *c* is scalar constant

*|A| =* if *A* is triangular matrix

**1.2. Matrix analysis**

Matrix analysis consists of techniques decomposing matrix so as to transform a complex matrix into a set of simple matrices because it is easy to process partially on simple matrices instead of whole complex matrix. Matrix analysis starts with concepts of eigenvalues and eigenvectors.

Given square matrix *A*(*nxn*), if there is a scalar value *λ* *> 0* and a vector *γ* such that

*A = λ γ*

Then λ and *γ* are eigenvalue and eigenvector of matrix A, respectively. If |A| 0 then there are *n* eigenvalues and *n* respective eigenvectors and at that time, the vector space determined by matrix *A* is decomposed into *n* disjoint sub-spaces and each sub-space is specified by a pair of eigenvalue and eigenvector. There is a question how to find out *n* eigenvalues and *n* respective eigenvectors and so, eigenvalues are solutions of following equation:

|*A – λIn*| = 0

Determinant |*A – λIn*| is expanded as a *nth* order polynomial which has *n* solutions {*λ1, λ2,…, λn*}. Suppose a solution of |*A – λIn*| is *λi*, the respective eigenvector *γi* is vector solution of following equation:

*A – λiIn = 0*

Note that eigenvectors are mutually orthogonal. Let *Λ* = *diag*(*λ1, λ2,…, λn*) be the eigenvalue matrix and let *Γ* be the orthogonal matrix created by *n* eigenvectors. *Γ* is called eigenvector matrix.

The Jordan decomposition theorem states that

This theorem is the most important theorem in matrix analysis and it is a base of many decomposition techniques. Note that Jordan decomposition exists if and only if *A* is square matrix and non-singular, |A| 0 and such matrix A is call *diagonalized* matrix.

Matrices *Γ* and *Λ* are much simpler than matrix *A.* Moreover, Γ and *Λ* are orthogonal and diagonal matrices and so they have many valuable properties.

When matrix *A*(*mxn*) is not square matrix, there is a technique so-called singular value decomposition (SVD), a generalization of Jordan decomposition, which is used to decompose matrix *A*(*mxn*). Let *Λ* = *diag*(*λ11/2, λ21/2,…, λr1/2*) be the eigenvalue matrix where *λi* is eigenvalue of *AAT* and *ATA*. Note that *λi* (s) are in descending ordering, *λ1  λ2 … λr1/2*. In other words, *r* eigenvalues *λi* (s) are solutions of one of two following equations:

|*AAT – λIn*| = 0

|*ATA – λIn*| = 0

The number of eigenvalues is *r = rank*(*AAT*) *= rank*(*ATA*) *min*(*m, n*). There are *r* eigenvectors *γi* (s) corresponding to *r* eigenvalues, which are solutions of equation:

*AAT – λiIn = 0*

There are *r* eigenvectors *δi* (s) corresponding to *r* eigenvalues, which are solutions of equation:

*ATA – λiIn = 0*

Let *Γ* (*mxr*) and *Δ*(*nxr*) be eigenvector matrices of *r* eigenvectors *γi* (s) and *r* eigenvectors *δi* (s). Note that both *Γ* and *Δ* are column orthogonal. It is easy to infer that columns of *Γ* and *Δ* are mutually orthogonal eigenvectors and each eigenvector *γi* (*δi*) has *m* (*n*) components. The singular value decomposition (SVD) theorem states that:

*A* = *ΓΛΔT*

Let *A–* = *ΔΛ-1Γ, we have:*

*AA–A =* (*ΓΛΔT*)(*ΔΛ-1ΓT*)(*ΓΛΔT*) = *ΓΛΔT = A*

We infer that *AA–A* is G-inverse of *A*.

According to SVD theorem, any matrix can be decomposed into three simpler matrices. SVD is often used to reduce vector space. For instance, when you choose *k* *r* largest eigenvalues, the number of columns of *Γ*, *Λ* and *Δ* are reduced to be *k* *r*. Let *Λ’*(*kxk*), *Γ’*(*axk*) and *Δ’*(*bxk*) are eigenvalue and eigenvector matrices of *k* eigenvalues, we have:

*A’* = *Γ’Λ’Δ’T*

Where *A’*(*axb*) is reduced matrix of *A*.

Given a invertible matrix *A*(*nxn*) and a random vector *X,* the quadratic form of *X* denoted *Q*(*X*) is:

Quadratic form is positive definite or positive semi-definite if *Q*(*X*) > 0 or *Q*(*X*) 0, respectively. Matrix *A* is called positive definite denoted *A > 0* or positive semi-definite denoted *A 0* if quadratic form is positive definite or semi-definite, respectively. Otherwise if *Q*(*X*) < 0, quadratic form and matrix A are indefinite. Because *A* is invertible, it is decomposed into *A = ΓΛΓT* where *Λ* and *Γ* are eigenvalue matrix and eigenvector matrix. Let *Y = ΓTX*, we have:

Where *λi* (s) are eigenvalues.

We have some properties of quadratic form:

*A* > *0* if and only if all eigenvalues *λi* (s) > 0

If *A* > *0* then *A-1* exists and |*A*| > 0

Given invertible and symmetric matrices *A* and *B* and given a constraint *XTBX = 1*, the maximum (minimum) of *XTAX* is the largest (smallest) eigenvalue of *B-1A*. The vector maximizing (minimizing) *XTAX* is the eigenvector which corresponds to largest (smallest) eigenvalue of *B-1A*.

Now we research partitioned matrix which is useful in matrix analysis. Given matrix *A*(mxn) are composed of other matrices *Aij*(*mixnj*).

Where *Aij* (s) are matrices having *mi* rows and *ni* columns. *Aij* (s) are called groups or sub-matrices and *A* is called partitioned matrix. Partitioning matrix technique is to perform matrix operations and determine properties and characteristics of matrix such as determinant |A|, inverse *A-1*, etc according to sub-matrices. Given partitioned matrix *B*(*mxn*) = as similar to *A*, we have:

If *A* is invertible, the inverse of *A* is:

If *A11* is invertible, the determinant of *A* is:

If *A22* is invertible, the determinant of *A* is:

Suppose partitioned matrix *B* is composed of a invertible matrix *A*(*nxn*) and two (*nx1*) vectors *a* and *b* as following.

The determinant of *B* is:

The inverse of *B* is:

**1.3. Matrix calculus**

Given a function *f* from *n-*dimension domain *Rn* to one-dimension image *R*, in other words, domain of *f* is vector space and image of *f* is a set of scalar number.

*f: Rn → R*

Thus, *f* is a scalar function of random vector *X*. Note that there is another kind of multivariate function which is mapping from vector (or matrix) to vector (or matrix) but this report focuses on scalar function from vector (or matrix) to scalar value. The column derivative of *f* with respective to *X* is defined as below:

So the column derivative of *f* with respective to *X* is a column vector whose components are partial derivatives where *xi* (s) are components of *X*. Similarly, the row derivative of *f* with respective to *X* is:

We have:

Given constant square matrix *A*(*nxn*), random vector *X*, scalar function *u,* scalar function *v,* scalar function *k* with scalar variable*,* vector function *U*,vector function *V*,vector constant *a* and scalar constant *c*, following are some properties of vector derivatives. Note that vector function is function transforming a vector into another vector. Only derivative of scalar function is discussed in this report, which relates to scalar function but some concepts relate to vector function.

Where and are matrices determined as below:

Note that *uv* and *UV* denotes the multiplication (not product mapping). The second order derivative of quadratic form is also called Hessian matrix.

If random variable is matrix and *f: Rmxn → R* is the scalar function of matrix, the derivative of matrix variable ***X***(*mxn*) is defined as a matrix of partial derivatives as below:

Let *u* and *v* be scalar functions of matrix.Let *k* bescalar function with scalar variable. Let ***U*** and ***V*** are function of matrix ***X*** which transform ***X*** into another matrix. Let *c* be constant scalar. Let *A, B, C, D* are matrix (or vector) constants because vector can be considered as *1-*row or *1-*column matrix. Following are some properties of matrix derivatives.

Let ***G*** = ***g***(***X***) is the matrix representing polynomial with scalar coefficients. Examples of ***G*** are polynomials such as *sin*(***X***), *cos*(***X***), *e****X***, *ln****X*** using Taylor series.

The derivative of ***G*** with respect to matrix ***X*** is a matrix representing polynomial with scalar coefficients, defined as below:

**1.4. Geometrical aspects of multivariate data**

Geometrical aspects of multivariate data starts with concept of distance. Distance *d* is defined as the function from *Rn x Rn* to *R+*:

Distance *d* is always greater than or equal to *0*, which satisfies three following axioms.

The Euclidean distance between two points is defined as following equation

Where *A* is positive definite matrix (A > 0). *A* is called a metric and the space on which A is defined is called metric space. This report focuses on Euclidean distance. If *A* is identical matrix *A = In*, Euclidean distance defined by .

Given a point *X0* and a scalar constant *d*, a *n-*dimension sphere with radius *d* is defined as following equation.

(*X – X0*)*T*(*X – X0*) *= d2*

**Figure 1.1.** *d*-radius sphere

This sphere is a set of points which are far from *X0* a distance *d*. Given a point *X0*, a matrix constant *A*(*nxn*) and a scalar constant *d*, a *n-*dimension ellipsoid is defined as following equation.

(*X – X0*)*TA*(*X – X0*) *= d2*

**Figure 1.2.** Ellipsoid with center *X0*, matrix *A* and constant *d*.

Note that *A* is invertible matrix. Suppose matrix *A* has *n* eigenvalues *λ1 λ2 … λn* and *n* respective orthogonal eigenvectors *γ1, γ2,…, γn*. This ellipsoid has following properties:

* The principle axes of ellipsoid have the same direction to eigenvectors *γ1, γ2,…, γn*. Of course, the number of axes is equal to the number of eigenvectors.
* The half of length of each axes is equal to where *λi* (s) are eigenvalues.
* Let *x0i* be the element *i* of center *X0 =* {*x01, x02,…, x0n*}. The *n*-dimension rectangle surrounding ellipsoid is determined by equation: where *aii* is the element (*i, i*) of *A–1*.
* The coordinate of tangency point (*xi*) between ellipsoid and its surrounding rectangle in the positive of *jth* axis is determined (*xi*) = where *aij* and *ajj* is the element (*i, j*) and (*j, j*) of *A–1*.

As aforementioned, the length (or norm) of vector *X* is:

and the cosine of angle between two vector *X* and *Y* is:

These concepts are extended in metric space given the metric matrix *A*(*nxn*). The length (or norm) of *X* becomes:

and the cosine of angle between two vectors *X* and *Y* becomes:

Given invertible matrix *A*(*nxn*) representing a vector space trsformation, the imagine of vector space via *A* denoted *Im*(*A*) is:

The kernel of vector space via *A* denoted *Ker*(*A*) is:

Note that in practice, both *Im*(*A*) and *Ker*(*A*) are column matrices whose columns are vectors. It is easy to infer that *Ker*(*A*) is the orthogonal complement of *Im*(*A*) and otherwise. We have:

*rank*(*Ker*(*A*)) *+ rank*(*Im*(*A*)) *= n*

Given a vector *a*, if *a* is orthogonal with all vector *X* (*aTAX = 0n, X* ) if and only if *a* *Ker*(*A*). Note that *Im*(*A*) and *Ker*(*A*) are also called image space and null space, respectively and *rank*(*Im*(*A*)) and *rank*(*Ker*(*A*)) are dimensions of column space and null space, respectively.

**2. Multivariate distributions**

Let *X* = (*x1, x2,…, xn*)T be *n-*dimension random vector. The multivariate distribution function *F* of *X* is defined as the accumulative probability.

*F*(*X0*) *= P*(*X < X0*) = *F*(*x1 < x01, x2 < x02,…, xn < x0n*)

The density function *f* of *X* is defined as the derivative of accumulative function *F*. In practice, density function is used in lieu of accumulative function.

Probability density function (pdf) satisfies three following conditions.

Note that *R* is the region of random vector *X*, thus, *R* is the *n-*dimension space.

The marginal probability density function of partial variable *xi* where *xi* is a component of *X* is the integral of *f* over points so that *xi* receives concrete value. Let be the region of *X* so that *xi = a* and so the marginal density function of *xi* is defined as following.

Partial random variables are independent if and only if the probability density function is the product of marginal density function of partial variables.

The marginal density function can be extended with *k* partial variables *xi+1, xi+2,…, xi+k*. Note that it is not necessary for *k* partial variables to be in successive order such as *i+1, i+2,.., i+k*.

The condition probability of *xj* given a set of *k* partial variables *xi+1, xi+2,…, xi+k* is the ratio of marginal density function of *xj*, *xi+1, xi+2,…, xi+k* to the marginal density function of *xi+1, xi+2,…, xi+k*.

The partial variable *xj* is independent from a set of *k* partial variables *xi+1, xi+2,…, xi+k* if and only marginal density function of *xj*, *xi+1, xi+2,…, xi+k* equals marginal density function of *xj*.

Please pay attention to the concept of independence because it relates to many probabilistic theorems. The condition probability can be extended to *c* partial variables *xj+1, xj+2,…, xj+c*. Note that it is not necessary for *c* partial variables to be in successive order such as *j+1, j+2,.., j+c*.

**2.1. Some important parameters of multivariate distribution**

Now we survey some important parameters of probability distribution such as population mean and population variance. We should distinguish between population parameters and their estimates extracted from sample. For example, sample mean and sample variance is estimates of population mean and population variance, respectively. Population parameters are theoretical parameters which calculated from probability density function. This section discusses only theoretical parameters. The partial mean of variable *xi* is the expected value of *xi*.

Note that mean is also called expectation or expected value. The conditional mean of variable *xj* given a set of *k* partial variables *xi+1, xi+2,…, xi+k* is defined as below:

The partial variance of variable *xi* is:

The standard deviation of *xi* is the squared-root of its variance.

The partial population covariance between variables *xi* and *xj* is:

We recognize that = . The correlation coefficient of variables *xi* and *xj* is calculated by normalizing their covariance over their standard deviations. Correlation coefficient ranges in interval [*-1 … 1*]. If it equals *1* then two variables is totally proportional. If it equals –*1* then two variables is inversely proportional. If it equals *0* then two variables are uncorrelated.

The mean of random vector *X* is composed of *n* partial population variances.

The covariance of given random vectors *X* and *Y* is the *nxn* matrix whose elements are partial covariance (s).

We recognize that the covariance *ΣXY* is symmetric matrix. If variables *X* and *Y* are independent, then we infer that *Covar*(*X, Y*) *= 0* but the backward inference is not asserted. The correlation coefficient matrix of variables *X* and *Y* is composed of partial correlation coefficients.

The variance of random vector *X* is defined as the covariance of *X* and itself.

Let *μ = μX* and *Σ = ΣXX*, we denote *X* (*μ, Σ*) if *X* has mean *μ* and variance (covariance) *Σ*. Let *A* and *B* be constants matrices. Let *X, Y* and *Z* be random vectors. Let *a* and *b* be constant vectors. Let *α* and *β* be scalar numbers. We have some following properties.

**2.2. Multinormal distribution**

Normal distribution is the heart of hypothesis because almost parametric tests are based on the assumption that sample conforms normal distribution. Multi-dimension normal distribution so-called multinormal distribution is used as an extension of one-dimension normal distribution when sample observations are vectors. Let *μ* and *Σ* be mean and covariance matrix, multinormal density function is:

Where |.| denotes determinant of matrix.

The multinormal density function *N*(*μ, Σ*) is constant on ellipsoids following equation (*X – μ*)*TΣ–1*(*X – μ*) = *d2*. The half-lengths of axes of contour ellipsoid are where *λi* (s) are eigenvalues of covariance matrix *Σ-1*.

If random vector *X* conforms multinormal distribution, we denotes *X* *N*(*μ, Σ*). If is zero vector and is identity matrix, multinormal distribution becomes standard multinormal distribution or Gaussian distribution.

Suppose *X* *N*(*μ, Σ*) and let *Z* = , we have *Z* *N*(*Ø, I*). Otherwise if *X* *N*(*Ø, I*) and let *Z = Σ1/2X + μ*, we have *Z* *N*(*μ, Σ*). Suppose *X* *N*(*μ, Σ*) and let *A* be non-singular *nxn* matrix and let *Y = AX + c* be the random vector where c is the constants vector, we have:

Suppose *X* *N*(*μ, Σ*) and let *U =* (*X – μ*)*TΣ–1*(*X – μ*) be random vector, thus, *U* conforms chi-square distribution with *n* degrees of freedom, U χ2n.

**3. Multivariate estimation of parameters**

Parameters such as mean, variance, median, etc are essential aspects of probability distribution. Section *2* mentions two important parameters: mean and variance. However parameters are theoretical aspects of distribution and so, they are often estimated from sample. Given a sample, the function of observations is called statistic, for example, sample mean and sample covariance *S* are statistics. Statistical inference includes two methods: parameter estimation and hypothesis testing. Both of them are based on these statistics. Hypothesis testing is main subject in next section and this section focuses on parameter estimation. As aforementioned, parameters are theoretical aspects of distribution and so statistics are used to estimate these parameters, thus, statistics are called estimates of parameters. For example, sample mean , sample covariance *S* are estimates of population mean *μ* and population covariance *Σ*, respectively. Note that statistics are considered random variables. When sample is multivariate data, statistics and parameters are vectors or matrix. Hereafter, we browse some concepts of parameter estimation.

**3.1. Basic concepts of parameter estimation**

Let *θ* and *Θ* represent multivariate parameter (such as population mean *μ* and population variance *Σ*) and it estimate (such as sample mean and sample covariance *S*). Estimate *Θ* is unbiased estimate of parameter θ if and only if expectation of Θ equals θ.

Given a *n-dimension* sample ***X*** = {*X1, X2,…, Xm*}, the sample mean is the average vector over *m* observation vectors *Xi* (s).

We prove that sample mean is unbiased estimate of theoretical mean *μ*. Suppose *m* observation vectors *Xi* (s) are independently distributed with the same theoretical mean *μ.*

Similarly, the sample covariance *S* defined as below is unbiased estimate of theoretical covariance Σ.

Where is the partial covariance between *kth* component and *lth* component of observation vectors. Note that each sample point *Xi =* (*xi1, xi2,…, xin*) is n-dimension vector whose components are singular random variable.

Where and the sample mean of *kth* component and *lth* component, respectively.

Therefore, sample covariance *S* is re-written as below:

If sample is considered as matrix ***X***(*mxn*) consisting of *m* sample row vector *Xi* (s) we have:

Where *xij* (s) are partial or component variables. Note that ***X*** is called sample matrix.

Sample covariance *S* is defined equivalently as below:

Let ***H*** = , we have:

Where ***H*** is symmetric and idempotent due to:

The expected value of *S* is:

Due to:

and:

We determine that sample covariance *S* is unbiased estimate of population covariance *Σ* because the expected value of *S* equals *Σ*.

The sample correlation matrix *R* is defined as below

Where is the partial covariance between *kth* component and *lth* component of observation vectors and *sj* is the partial standard deviation of *jth* component of observation vectors.

Therefore, sample standard deviation *sj* is the root-square of covariance *sjj*. For convenience, let *x* and *y* represent the *kth* and *lth* components of observation vectors we have:

Note that the sample correlation coefficient ranges in interval [*-1, 1*]. It is proved that *rkl* is unbiased estimate of theoretical correlation coefficient *ρkl* = where *σkl, σk* and *σl* are theoretical covariance between *kth* and *lth*, standard deviation of *kth* component and standard deviation of *lth* component, respectively. The definition of theoretical correlation coefficient is described in section *2* about multivariate distribution. Suppose *S* denotes covariance matrix and sample matrix, correlation coefficient matrix R is equivalently defined:

Where D is diagonal matrix whose diagonal elements (*j, j*) are partial standard deviation *sj* of *jth* component.

If *Θ* is biased estimate, the deviation *E*(*Θ*) – *θ* is called a bias. Let *Var*(*Θ*) be the variance of estimate *Θ*, the standard error of *Θ* denoted *se*(*Θ*) is defined as its standard deviation.

In general, we often prefer unbiased estimates but biased estimates are used in some cases that it is impossible to draw unbiased estimates. If so, the best biased estimate is the one whose mean square error is smallest. The mean square error denoted *MSE*of estimate *Θ* is defined the expectation of square of bias.

Let *Θ1* and *Θ2* be biased estimates, the relative efficiency defined as below is used to choose which one is better.

If relative efficiency is less than *1*, *Θ1* is more efficient than *Θ2* and otherwise. As aforementioned sample mean and sample covariance *S* are unbiased estimates but there is a question that how to find out such estimates: unbiased estimates and efficient biased estimates. Methods to determine statistical estimates are described later.

**3.2. Enhancement of sample statistic**

**3.3. Parameter estimation methods**

There are three methods to determine statistical estimates: moment method, maximum likelihood method and Bayesian method. Firstly, moment method is described because of its simplicity. The population or theoretical *kth* moment is defined as the expectation of *kth* power of random variable.

Because *X* is random vector, please see section *1* mentioning the power of vector. The sample *kth* moment is defined as below given a set of sample observations *X1, X2,…, Xm*.

For example, the first and second population moments are population mean *E*(*X*) *= μ* and population *E*(*X2*) = *E*(*XXT*) = *Σ + μμT*, respectively. The first and second sample moments are and . The basic idea of moment method is to equate population moment to sample moment. The estimates are solutions of such equations. For example, we set first and second population moments to be equal to first and second sample moments, respectively in order to find out the estimates of *Σ* and *μ*.

is unbiased estimate of population mean *μ* and now we evaluate whether or not is unbiased estimate of population covariance *Σ*.

So, is biased estimate of covariance *Σ*. It means that moment method is not effective estimate method. Now we discuss a more popular method so-called maximum likelihood. The likelihood function is the joint probability which is the product of conditional probabilities of observations *Xi* (s) given the parameters. So likelihood function is considered as function of parameters. Let *Θ* = {*θ1, θ2,…, θk*} represent a set of parameters such population mean *μ*, population covariance *Σ*, etc, the likelihood function *L*(*Θ*) is defined as below:

Let be sample estimate of *Θ*, we recognize that is extreme value at which *L*(*Θ*) gets maximal.

Because it is too difficult to work with the likelihood function in the form of product of condition probabilities, we should take the logarithm of *L*(*Θ*) so that the log function converts the repeated multiplication to repeated addition. The logarithm of *L*(*Θ*) so-called log-likelihood is denoted *LogL*(*Θ*).

The essence of maximizing the likelihood function is to find the peak of the curve of *LogL*(*Θ*). This can be done by setting the first-order partial derivative of *LogL*(*Θ*) with respect to each parameter *θi* *Θ* to *0* and solving this equation to find out parameter *θi*. The reason is that the slope of the curve *LogL*(*Θ*) is equal to 0 at its peak. The number of equations corresponds with the number of parameters. For example, suppose observations *Xi* (s) conform normal distribution, we need to find out sample estimates of population mean *μ* and population covariance *Σ*. The log log-likelihood function is:

The first-order partial derivative of *LogL*(*Θ*) with respect to *μ* is:

The first-order partial derivative of *LogL*(*Θ*) with respect to *Σ* is:

Partial derivative is re-written:

Let *A =* , partial derivative is re-written:

The estimates and are solutions of a set of equations which created by setting first-order partial derivatives of *LogL*(*Θ*) with respect to *μ* and *Σ* to be zero.

Where *Ø* denotes zero vector and [0] = denotes zero matrix.

Therefore the estimates of population mean and covariance are and , respectively. This result is similar to one from moment method when the estimate of covariance is biased estimate.

**4. Multivariate hypothesis testing**

As aforementioned, hypothesis is defined as a statement about parameters of one or more populations. Hypothesis testing is the procedure that decides whether a hypothesis is accepted or rejected based on one or more sample statistic (s). Almost parametric hypothesis tests rely on two conditions:

* Sample observations conform normal distribution.
* Underlying distributions such as normal, chi-squared, student, F, etc are used to determine the critical region for tests.

The first condition is kept in multivariate tests but the second condition is the cause of testing complication because it is impossible to survey all of distributions such as normal, chi-squared, student, F, etc in multi-dimension data when sample points are vectors and some sample statistics are matrices. Statistical theory gives us an excellent method to eliminate multivariate distributions from testing process. In other words, likelihood function together with one-dimension distribution based on such likelihood function are used to calculate the critical region for tests instead of using underlying multivariate distributions. This method is called likelihood ratio test.

**4.1. Likelihood ratio test**

Suppose data sample is *n-*dimension vector space ***X***, which contains *m* observation vectors *X1, X2,…, Xm*where *Xi* = {*xi1*, *xi2*,…, *xin*}. Note that *Xi* is identically distributed random vector and it can be called observation, data point, or sample point. We test on mean of normal distribution when covariance known or unknown, so *Xi* (s) conforms *N*(*μ, Σ*) and the null hypothesis *H0: μ = μ0* and *H1*: no constraint on *μ*. Let *L0*(*μ0, Σ* *| X*) and *L1*(*μ, Σ | X*) be the likelihoods for null hypothesis *H0* and alternative hypothesis *H1*, respectively. Note that *L0* is function of covariance *Σ* and *L1* is function of both mean *μ* and covariance *Σ* because *μ0* is constant.

Let *L0\** and *L1\** be the maxima of *L0* and *L1*, respectively. The likelihood ratio *R* is defined the ratio of null maximum *L0\** to alternative maximum *L1\**.

We take logarithm of *R* as below:

Note that *LogL0\** and *LogL1\** are maxima of *LogL0* and *LogL1* which, in turn, are logarithms of *L0* and *L1*, respectively.

When population covariance Σ is known, likelihoods *LogL0* and *LogL1* get maximal if *μ* = . Therefore, maxima *LogL0\** and *LogL1\** are totally determined.

It is proved that *–2Log*(*R*) is approximate to chi-square distribution *χ2* with *n* degrees of freedom when *Σ* is known; so *H0* is rejected in flavor of *H1* if *–2Log*(*R*) > *χ2α,n* with significant level *α*.

**4.2. Test on sample correlation coefficient**

Suppose data sample is *n-*dimension vector space ***X***, which contains *m* observation vectors *X1, X2,…, Xm*where *Xi* = {*xi1*, *xi2*,…, *xin*}. For convenience, let *x* and *y* represent the *kth* and *lth* components of observation vectors, the sample correlation coefficient between two components *x* and *y* is:

Where *xi* and *xj* denote *kth* and *lth* components of observation *Xi*.

Let *ρxy* be the theoretical correlation coefficient. Note that the sample correlation coefficient *ρxy* ranges in interval [*-1, 1*]. If *ρxy* = 0, two components *x* and *y* are independent. When *rxy* is a statistic, we test on *ρxy* so as to determine whether or not two components *x* and *y*. It means that we test on *ρxy = 0* and *H1*: no constraint on *ρxy*, based on statistic *rxy*. Although, it is very complicated to discover distribution of statistic *rxy*, there are some transformations on *rxy* can be applied. According to Fisher transformation, the statistic is approximate to normal distribution with mean and variance

Given *H0*: *ρxy = 0* and significant level *α*, *H0* is rejected if the *z*-value, is greater than critical value defined by significant level *α*.

When sample size is small (*m* 25), Hotelling transformation is used:

Hotelling distribution *w\** is approximate to normal distribution with variance . Given *H0*: *ρxy = 0* and significant level *α*, *H0* is rejected if the *z*-value, is greater than critical value defined by significant level *α*.

Suppose *x* and *y* are normally distributed, t-distribution can be used to test on *H0*: *ρxy = 0*. The statistic conforms *t-*distribution with *m – 2* degrees of freedom.

We reject *H0*: *ρxy = 0* if *|T| t1-α/2,m-2* with significant level *α*.

**4.3. Multivariate Bayesian test**

**5. Conclusion**

By the way, we have already survey over almost basic concepts of multivariate statistics, especially, multivariate hypothesis testing, which aim us to draw some comments:

* Statistical inference, namely parameter estimation and hypothesis testing, in multi-dimension is less efficient and its cost is high because matrix computation is very complicated. Moreover, matrix calculus such as derivative and primitive of vector and matrix is a hazard. Please pay attention to choosing whether or not using multivariate inference instead of univariate inference. You should transform data from multi-dimension format to singular or one-dimension format if possible. When it is not able to perform transformation task because the inherence of data is multi-dimensional, multivariate statistical inference is the best choice if you take advantage of computer program because it is easy to get mistakes when calculating matrix operations manually.
* Hypothesis testing is based on two conditions (1) assumption of normality (2) parameter estimation. The concept of biased and unbiased estimate is significantly attentive because it is the heart of parameter estimation. Likelihood function plays important role in multivariate hypothesis testing due to its excellent ability to interpret the semantic of statistical conclusion from multi-dimension point of view to one-dimension point of view.
* Normal distribution is always the heart of parametric hypothesis testing when it is the first assumption on almost tests even though hypothesis testing is extended from one-dimension to multi-dimension

In conclusion, multivariate statistical inference is not significantly different (as the terminology of hypothesis testing) from univariate statistical inference except that matrix algebra and matrix calculus are intricate problems and you should expend your imagination in multi-dimension space.