A Maximum Likelihood Mixture Approach for Multivariate Hypothesis Testing in case of Incomplete Data

Loc Nguyen1

1. Vietnam Institute of Mathematics

Received: June 30, 2013 / Accepted: June 30, 2013 / Not Published Yet.

**Abstract: Multivariate hypothesis testing becomes more and more necessary when data is in the process of changing from scalar and univariate format to multivariate format, especially financial and biological data is often constituted of *n-*dimension vectors. Likelihood ratio test is the best method that applies the test on mean of multivariate sample with known or unknown covariance matrix but it is impossible to use likelihood ratio test in case of incomplete data when the data incompletion gets popular because of many reasons in reality. Therefore, this research proposes a new approach that gives an ability to apply likelihood ratio test into incomplete data. Instead of replacing missing values in incomplete sample by estimated values, this approach classifies incomplete sample into groups and each group is represented by a potential or partial distribution. All partial distributions are unified into a mixture model which is optimized via expectation maximization (EM) algorithm. Finally, likelihood ratio test is performed on mixture model instead of incomplete sample. This research provides a thorough description of proposed approach and mathematical proof that is necessary to such approach. The comparison of mixture model approach and filling missing values approach is also discussed in this research.**

**Key words:** maximum likelihood, mixture model, multivariate hypothesis testing, incomplete data

1. Likelihood ratio test[[1]](#footnote-1)

Suppose data sample is *n-*dimension vector space *X*, which contains *m* observation vectors *X*1, *X*2,…, *Xm*where *Xi* = {*xi*1, *xi*2,…, *xin*}. Note that *Xi* is identically distributed random variable and it can be called observation, data point, or sample point. We conventionalize that upper-case letter denotes variable and lower-case letter denotes value or instance of variable, for example, *xi* is the instance of variable *Xi*. We test on mean of normal distribution when variance known or unknown, so *Xi* (s) conforms *N*(*μ*, Σ) and the null hypothesis *H*0*: μ = μ*0 and *H*1: no constraint on *μ*. Let *L*0(*μ*0, Σ *| X*) and *L*1(*μ*, Σ *| X*) be the likelihoods for null hypothesis *H*0 and alternative hypothesis *H*1, respectively.

(Where *n* is the number of dimensions of sample *X*)

We take the logarithm of likelihood functions so as to convert the repeated multiplication into repeated addition, so the log-likelihood function of *X* is [1, pp. 185-186]:

Where trace operator denoted tr(*A*) is the sum of all elements that are on the diagonal of matrix *A* and is sample mean and *S* is the sample covariance matrix.

*L*0(*μ*0, Σ | *X*) and *L*1(*μ*, Σ | *X*) get maximal if and only if *LogL*0(Θ | *X*) and *LogL*1(Θ | *X*) get maximal. It is proved that *LogL*0(*μ*0, Σ | *X*) and *LogL*1(*μ*, Σ | *X*) get maximal if and only if and Σ = *S* when and Σ = *S* are solutions of equation formed by setting first-order derivative with regard to *μ* and Σ to be zero.

Let *L*0*\** and *L*1*\** be the maximal likelihoods for null hypothesis *H*0 and alternative hypothesis *H*1. If we substitute and Σ = *S* to *L*0(*μ*0, Σ | *X*) and *L*1(*μ*, Σ | *X*), then *L*0(*μ*0, Σ | *X*) and *L*1(*μ*, Σ | *X*) is totally determined.

The likelihood ratio *R* [1, pp. 184-192] is defined as the ratio of the maximum likelihood of null hypothesis to the maximum likelihood of alternative hypothesis.

It is proved that *–*2*log*(*R*) is approximate to chi-square distribution χ2 with *n* degrees of freedom when Σ is known; so *H*0 is rejected in flavor of *H*1 if *–*2*log*(*R*) > χ2*α,n* with significant level *α*. This introduction is a brief principle of likelihood ration test. In general, likelihood ration test is a very effective testing method because it simplifies the complexity of *n*-dimension data by transforming multivariate testing criterion into univariate testing criterion based on scalar likelihood ratio.

2. Mixture model test on incomplete data

Suppose data sample is *n-*dimension vector space, which contains *m* observation vectors *X*1, *X*2,…, *Xm*where *Xi* = {*xi*1, *xi*2,…, *xin*}. Thus, *X =* {*X*1, *X*2,…, *Xm*} compose a matrix whose each row is observation *Xi*. In case of incomplete data, *X* is sparse matrix and observation *Xi* is not always complete vector when it can lack some components *xij* (s). Likelihood ration test cannot apply into incomplete sample as *X* because it is impossible to calculate statistics such as mean and variance in terms of incomplete sample. This research tries to overcome this drawback by discovering potential probability distributions under observations *Xi* (s) regardless of data incompletion. Suppose we test on mean of normal distribution when variance known or unknown, so *Xi* (s) conforms *N*(*μ*, Σ) and the null hypothesis *H*0: *μ = μ*0 is in flavor of alternative hypothesis *H*1: no constraint on *μ*. Based on such potential distributions, the maximum likelihood of these hypotheses is determined.

Firstly, we estimate the number of potential probability distributions under observations *Xi* (s). Observations are classified into *k* classes following two conditions below.

1. Class *c*1 represents observations whose components *xij* (s) are completed. Note that class *c*1 may not exist.
2. Classes *c*2, *c*3,…, *ck* represent observations lacking the same components *xij* (s). For example, we have four observations *X*1 = {*x*11=1, *x*12=2, *x*13 (*empty*), *x*14 (*empty*)}, *X*2 *=* {*x*21=2, *x*22=3, *x*23(*empty*), *x*24 (*empty*)}, *X*3 = {*x*31 (*empty*), *x*32(*empty*), *x*33=1, *x*34 =2}and *X*4 = {*x*41 (*empty*), *x*42(*empty*), *x*43=2, *x*44=3}. Thus, *X*1 and *X*2 belong to the same class (class 1); *X*3 and *X*4 belong to the same class (class 2) according to this condition. Note that empty value is considered as missing value. Two or many classes can overlap together. For example, if *X*5 = {*x*11=3, *x*12=(*empty*), *x*13 =(*empty*), *x*14 (*empty*)} then *X*5 belongs to different class – class 3, thus, class 1 and class 3 are overlapped.

The number of potential probability distributions is initialized to be *k*. Note that *k* should be much smaller than the size *m* of sample. Note that potential probability distribution is also called partial probability distribution. Let *pi* be the potential probability distributions corresponding to class *k*. Suppose *pi* conforms normal distribution *N*(*μi*, Σ*i*) with mean *μi* and variance Σ*i*. The probabilistic mixture model [2, p. 3] is defined as below.

|  |  |
| --- | --- |
|  | **(1)** |

Where *pj* is the potential distribution and *xi* is instance of variable *Xi*. Let Θ = {*α*1, *α*2,…, *αk*, *θ*1, *θ*2,…, *θk*} and *θj* = {*μj,* Σ*j*} be probabilistic parameters of sample and potential distribution *pj*, respectively. Equation (1) indicates that the distribution of data point *xi* is constituted of potential (or partial) distributions *pj* (s). Each potential distribution *pj* is weighted by the weight *αj* such that . The weight *αj* is the probability of *pj* if we consider *pj* as random variable. The weights *αj* (s) are learned and updated from sample, which will be discussed later. The likelihood function of *X* is:

We take the logarithm of likelihood function so as to convert the repeated multiplication into repeated addition, so the log-likelihood function of *X* is [2, p. 3]:

Let *Y*1, *Y*2,…, *Ym* be variables indicating that data point *Xi* comes from which potential distribution. The value of each *Yi* ranges in {1, 2, 3,…, *m*}. Concretely, if *Yi = yi* then data point *Xi* conforms distribution . When *X =* {*X*1, *X*2,…, *Xm*} is incomplete data, *Y =* {*Y*1, *Y*2,…, *Ym*} is observational data that let us exploit *X*. The probability of *Yi = yi* is the prior probability of which indicates that how weighted data point *xi* comes from distribution , so we have *P*(*yi*) = . Similarly, the conditional probability of data point *xi* given *yi* is the partial probability *pj* given *xi*, so we have *P*(*xi | yi*) = . Let (*X, Y*) be the complete data, the log-likelihood function is re-written as below [2, p. 3].

|  |  |
| --- | --- |
|  | **(2)** |

Our goal is to find out the optimal parameter Θ\* = {*α*1\*, *α*2\*,…, *αk\**, *θ*1\*, *θ*2\*,…, *θk\**} that maximizes the log-likelihood function in (2). Expectation maximization (EM) algorithm being the iterative process is applied in order to find out Θ\*. Let be the optimal parameter at *tth* iteration. Note that *θjt =* {*μjt,* Σ*jt*} is the optimal parameter of partial probability *pj* at *tth* iteration. In general, starting with initial estimate Θ0, each iteration in EM algorithm has two steps [3, p. 8]:

* E-step: computing the conditional expectation based on the previous estimate Θ*t*
* M-step: finding out current estimate Θ*t+*1 *=* Θ\* that maximizes such conditional expectation. Note that Θ*t+*1 is reserved for next iteration.

EM algorithm stops when it meets the terminating condition, for example, the difference of previous estimate Θ*t* and current estimate Θ*t+*1 is smaller than some pre-defined threshold *ε*, namely, |Θ*t+*1 *–* Θ*t*| < *ε*. In E-step, the conditional expectation *Q*(Θ, Θ*t*) is determined as below [2, p. 4].

|  |  |
| --- | --- |
|  | **(3)** |

Note that *Q*(Θ, Θ*t*) is the function of variable Θ and value Θ*t* is known estimate in previous iteration. *Q*(Θ, Θ*t*) sums over all possible instances of *Y=*{*Y*1, *Y*2,…, *Ym*} and so Ψ denotes all combinational values over *Y*. Now we need to specify the probability of *y* given *X* and Θ*t*, *P*(*y | X*, Θ*t*). Note that *y =* {*y*1, *y*2,…, *ym*} is instance of observation variable data *Y =* {*Y*1, *Y*2,…, *Ym*}. Applying Bayesian rule, we have [2, p. 3]:

|  |  |
| --- | --- |
|  | **(4)** |

Suppose *y*1, *y*2,… and *ym* are mutually independent, the probability of *y* given *X* is following product [2, p. 4].

Consequently, the conditional expectation *Q*(Θ, Θ*t*) is expended as below [2, p. 4].

(because Ψ denotes all combinational values over *Y*)

(where = 1 if and only if *yi* = *c* and where *c* ∈ {1,2,…, *k*} represents classes of potential probabilities *pc*)

In general, we have:

|  |  |
| --- | --- |
|  | **(5)** |

Note that *P*(*yi = c | xi*) is computed following equation (4). Because equation (5) has two parameters *αc* and *θc =* {*μc, Σc*}, the conditional expectation *Q*(Θ, Θ*t*) is split into two parts as below.

|  |  |
| --- | --- |
|  | **(6)** |

Let

It is easy to recognize that both *G*(*αc*) and *H*(*θc*) are less than or equal to 0 because of . Hence, the expectation *Q*(Θ*,* Θ*t*) is maximal if and only if both *G*(*αc*) and *H*(*θc*) are minimal. In other words, we find out extreme points *αc\** and *θc\** = {*μc*\*, *∑c\**} that minimize both *G*(*αc*) and *H*(*μc*,*∑c*) when *Q*(Θ*,* Θ*t*) is considered as function of variables *αc* and *θc* = {*μc*,*∑c*}.

Applying Lagrange function into *G*(*αc*) with constraint , extreme point *αc\** is the solution of the equation formed by setting first-order derivative of sum of *G*(*αc*) and Lagrange constraint to be zero.

We have equation (7) to determine *λ*.

|  |  |
| --- | --- |
|  | **(7)** |

Summing (7) over *k* classes {1, 2,…, *k*}, we have [2, p. 5]:

Substituting *λ = –m* into equation (7), extreme point *αc\** is totally determined.

Where *P*(*yi = c | xi*) is calculated following equation (4). Now we find out another extreme point *θc\** = {*μc*\*,Σ*c\**}. Suppose potential distribution *pc*(*xi|θc*) is normal, it is expended as below.

It implies that

(where *n* is the dimension of sample space *X*)

The first-order partial derivative of *H*(*μc*, *∑c*) with respect to *μc* is [4, p. 35]:

The optimal *μc*\* maximizing *H*(*μc*, Σ*c*) is the solution of equation created by setting partial derivatives of *H*(*μc*, Σ*c*) with regard to mean *μc* to be zero. Note that *Ø* denotes zero vector (0, 0,…, 0).

The first-order partial derivative of *H*(*μc*, *∑c*) with respect to *∑c* is:

Due to:

And

Because author [2, p. 5] mentioned:

Where tr(*A*) is trace operator which takes sum of diagonal elements of matrix .

It implies [4, p. 45]

Where *∑c* is symmetric and invertible matrix.

The optimal Σ*c*\* maximizing *H*(*μc*,Σ*c*) is the solution of equation created by setting partial derivatives of *H*(*μc*,Σ*c*) with regard to mean Σ*c* to zero. Note that [0] denotes zero matrix , we have:

In general, the optimal parameters *αc\** and *θc\** = {*μc*\*,Σ*c\**} for each potential (partial) probability *pc* form a following triple shown as equation (8). Note that *P*(*yi = c | xi*) is computed following equation (4).

|  |  |
| --- | --- |
|  | **(8)** |

Of course, we have *k* groups (*α*1\*, *θ*1\*), (*α*2\*, *θ*2\*),… and (*α*2\*, *θ*2\*) like (8) and so the global optimal parameter Θ\* = {*α*1\*, *α*2\*,…, *αk\**, *θ*1\*, *θ*2\*,…, *θk\**} in M-step of EM algorithm is totally specified. The basic idea is to divide the global optimal parameter *Θ\** into *k* groups of partial parameters (*αj\**, *θj\**) and find out such partial parameters in particular.

Back the test on mean of normal distribution when the null hypothesis *H*0: *μ = μ*0 is in flavor of alternative hypothesis *H*1: no constraint on *μ*. The likelihood ratio *R*(*X*) is defined as the ratio of the maximum likelihood of null hypothesis to the maximum likelihood of alternative hypothesis.

Where and are maximum likelihoods of null hypothesis and alternative hypothesis, respectively.

This research proposes a 4-step testing process based on mixture model in order to test on mean of normal distribution in case of incomplete data.

1. Specifying *k* classes and *k* respective partial probabilities *pc*(*xc | θc*) with and *θc* = {*μc*, Σ*c*}. The weights (s) of partial probabilities *pc* (s) are initialized by the ratio of the number *Xi*(s) belonging to class *c* to *m*, the number of total *Xi* (s),
   1. Class *c*1 represents observations whose components *xij* (s) are completed. Note that class *c*1 may not exist.
   2. Classes *c*2, *c*3,…, *ck* represent observations lacking the same components *xij* (s).
2. Specifying the likelihood functions of null hypothesis and alternative hypothesis such as *L*0(*X*) and *L*1(*X*).

Where parameter *θc* = {*μc*, Σ*c*} is constant and is assigned by sample mean *μ*0 and known covariance matrix Σ (or sample covariance matrix *S*).

* 1. .

1. Applying EM algorithm into find out optimal parameters *αc\** and *θc\** = {*μc*\*, Σ*c\**} for each potential (partial) probability *pc*.
   1. Optimal parameters *αc\** with respect to null hypothesis: . Note that *θc* = {*μc*, Σ*c*} is constant.
   2. Optimal parameters *αc\** and *θc\** = {*μc*\*, Σ*c\**} with respect to alternative hypothesis:
2. Substituting *αc\** and *θc\** = {*μc*\*, Σ*c\**} into likelihood functions *L*0(*X*) and *L*1(*X*) so that the likelihood ratio is totally determined. It is proved that *–*2*log*(*R*) is approximate to chi-square distribution χ2 with *n* degrees of freedom when population covariance Σ is known; so *H*0 is rejected in flavor of *H*1 if –2*log*(*R*) > χ2*α,n* with significant level *α*.

Please pay attention to step 1 and step 3 because it is slightly complicated. Suppose data sample is *n-*dimension vector space *X*, which contains *m* observation vectors *X*1, *X*2,…, *Xm*where *Xi* = {*xi*1, *xi*2,…, *xin*}. Thus, *X =* {*X*1, *X*2,…, *Xm*} compose a matrix whose each row is observation *Xi*. Suppose *Xi* (s) conforms *N*(*μ*, Σ). If class *c* is composed of *u* non-empty components where *cu* = , then the mean *μc* contains only *c*1*th*, *c*2*th*,…, *cuth* components which correspond with *c*1*th*, *c*2*th*,…, *cuth* columns in matrix *X*. Similarly, the covariance matrix Σ*c* contains only variances among *c*1*th*, *c*2*th*,…, *cuth* components which correspond with *c*1*th*, *c*2*th*,…, *cuth* columns in matrix *X*. All arithmetical operators in step 3 are performed based on such *u* non-empty components . This important thing is the tip of this research. The mean *μc* in the likelihood function of null hypothesis is constant and is assigned by *c*1*th*, *c*2*th*,…, *cuth* components of sample mean *μ*0*.* The covariance matrix Σ*c* in the likelihood function of null hypothesis is constant and is assigned by *c*1*th*, *c*2*th*,…, *cuth* components of population covariance Σ or sample covariance *S* in case of unknown Σ. It means that it is not necessary to computed the optimal parameters *μc\** and Σ*c\** with respect to likelihood function of null hypothesis *L*0(*X*).

For example, sample *X* has four observations *X1* = {*x*11=1, *x*12=2, *x*13 (*empty*), *x*14 (*empty*)}, *X*2 = {*x*21=2, *x*22=3, *x*23 (*empty*), *x*24 (*empty*)}, *X*3 = {*x*31 (*empty*), *x*32 (*empty*), *x*33=1, *x*34 =2}and *X*4 = {*x*41 (*empty*), *x*42 (*empty*), *x*43=2, *x*44=3}. These observations formed a 4*x*4 matrix shown in table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *X*1 | *x*11 = 1 | *x*12 = 2 | *x*13 = ? | *x*14 = ? |
| *X*2 | *x*21 = 2 | *x*22 = 3 | *x*23 = ? | *x*24 = ? |
| *X*3 | *x*31 = ? | *x*32 = ? | *x*33 = 1 | *x*34 = 2 |
| *X*4 | *x*41 = ? | *x*42 = ? | *x*43 = 2 | *x*44 = 3 |

**Table 1.** An example of incomplete sample *X* where question mark (?) denotes missing value (or empty value).

Suppose we test on mean of normal distribution and so *Xi* (s) conforms *N*(*μ*, Σ) and the null hypothesis *H*0: *μ = μ*0 = (0, 0, 0, 0)*T* and *H*1: no constraint on *μ*. Suppose variance is known, Σ = . Sample *X* is classified into two classes such as class 1 containing *X*1, *X*2 and class 2 containing *X*3, *X*4. Let *p*1 and *p*2 are partial probabilities attaching to class 1 and class 2, respectively. Both *p*1 and *p*2 conform normal distribution, *p*1 *N*(*μ*1, Σ1) and *p*2 *N(μ*2, Σ2). Let *Y =* {*Y*1, *Y*2, *Y*3, *Y4*} be the set of observations where *Yi* indicates *Xi* belongs to which class (1 or 2). Because *X*1, *X*2 belong to class 1 and *X*3, *X*4 belong to class 2, we have *Y =* {*Y*1 = 1, *Y*2 = 1, *Y*3 = 2, *Y*4 = 2}*, p*1(*x*1) *=*0.5, *p*1(*x*2) = 0.5, *p*1(*x*3) = 0, *p*1(*x*4) = 0, *p*2(*x*1) =0, *p*2(*x*2) = 0, *p*2(*x*3) = 0.5 and *p*2(*x*4) = 0.5.

When class 1 lacks columns (components) 3 and 4 and relates only *X*1 and *X*2, the mean *μ*1 receives two first components of *μ* and the covariance *Σ*1 containing variances between *X*1 and *X*2 is the sub-matrix of Σ. Similarly, the mean *μ*2 receives two last components of *μ* and the covariance Σ2 is the sub-matrix of Σ which contains variances between *X3* and *X4*. We have:

Let *α*1 and *α*1 be the weights of partial probabilities *p*1 and *p*2, respectively. Hence, *αi* is initialized as the number of *Xi* (s) belonging to class *i*. We have:

At the first iteration of EM algorithm, the probability of *y* given *X* denoted *P*( *yi | xi*) is calculated according to equation (4), for example:

In the similar way, we have:

* *P*(*y*1 = 1 *| x*1) *=* 1, *P*(*y*1 = 1 *| x*2) *=* 1, *P*(*y*1 = 1 *| x*3) *=* 0, *P*(*y*1 = 1 | *x*4) = 0, *P*(*y*1 = 2 | *x*1) = 0, *P*(*y*1 = 2 | *x2*) = 0, *P*(*y*1 = 2 | *x*3) = 1, *P*(*y*1 = 2 | *x*4) = 1.
* *P*(*y*2 = 1 | *x*1) = 1, *P*(*y*2 = 1 | *x*2) = 1, *P*(*y*2 = 1 | *x*3)= 0, *P*(*y*2 = 1 | *x*4) = 0, *P*(*y*2 = 2 | *x*1) = 0, *P*(*y*2 = 2 | *x2*) = 0, *P*(*y*2 = 2 | *x*3) = 1, *P*(*y*2 = 2 | *x*4) = 1.
* *P*(*y*3 = 1 | *x*1) = 1, *P*(*y*3 = 1 | *x*2) = 1, *P*(*y*3 = 1 | *x*3) = 0, *P*(*y*3 = 1 | *x*4) = 0, *P*(*y*3 = 2 | *x*1) = 0, *P*(*y*3 = 2 | *x*2) = 0, *P*(*y*3 = 2 | *x*3) = 1, *P*(*y*3 = 2 | *x*4) = 1.
* *P*(*y*4 = 1 | *x*1) = 1, *P*(*y*4 = 1 | *x*2) = 1, *P*(*y*4 = 1 | *x*3) = 0, *P*(*y*4 = 1 | *x*4) = 0, *P*(*y*4 = 2 | *x*1) = 0, *P*(*y*4 = 2 | *x*2) = 0, *P*(*y*4 = 2 | *x*3) = 1, *P*(*y*4 = 2 | *x*4) = 1.

Parameters *μ*1, Σ1, *μ*2, Σ2, *α*1 and *α*2 are re-calculated according to equation (8).

Where *μ*1\*, Σ1\*, *μ*2\*, Σ2\*, *α*1\* and *α*2\* are optimal parameters that maximizing log-likelihood function. Because the deviation between *α*1\* and *α*10is zero and the deviation between *α*2\* and *α*20is zero, algorithm is stopped at the first iteration. Let *R*(*X*) be the likelihood ratio.

(where *L*0\*(*X*) and *L*1\*(*X*) are maximum likelihoods of null hypothesis and alternative hypothesis, respective)

(where *LogL*0\*(*X*) and *LogL*1\*(*X*) are maximum likelihoods of null hypothesis and alternative hypothesis, respective)

Thus, hypothesis *H*0: *μ*0 = (0, 0, 0, 0)*T* cannot be rejected in flavor of *H*1: no constraint at significant level 0.05% because *–*2*logR =* 0 < 9.49 = χ20.05,4.

3. Mixture model vs. filling missing values

Filling missing values method is to try to estimate missing values (or empty values) so as to transform incomplete data into complete data and after that, hypothesis testing is done on complete data as usual. We discuss some methods of filling missing values and compare them with mixture model. Suppose data sample is *n-*dimension vector space, which contains *m* observation vectors *X*1, *X*2,…, *Xm*where *Xi* = {*xi*1, *xi*2,…, *xin*}. Thus, *X =* {*X*1, *X*2,…, *Xm*} composes a matrix whose each row is observation *Xi*. In case of incomplete data, *X* is sparse matrix and observation *Xi* is not always complete vector when it can lack some components *xij* (s). Table 2 is an example of sample *X*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | v = 1 | v = 2 | v = 3 | v = 4 |
| u = 1 | *x*11 = 1 | *x*12 = 2 | *x*13 = ? | *x*14 = ? |
| u = 2 | *x*21 = 2 | *x*22 = 3 | *x*23 = ? | *x*24 = ? |
| u = 3 | *x*31 = ? | *x*32 = ? | *x*33 = 1 | *x*34 = 2 |
| u = 4 | *x*41 = ? | *x*42 = ? | *x*43 = 2 | *x*44 = 3 |

**Table 2.** Incomplete sample where question mark (?) denotes missing value (or empty value), *u* denotes row index and *v* denotes column index.

In the easiest way, if each column is considered as random variable, the missing value will be estimated as the mean of column vector. This way is called average method.

According to average method, we have *x*31 = *x*41 = (1+2) / 2 = 1.5, *x*32 = *x*42 = (2+3) / 2 = 2.5, *x*13 = *x*23 = (1+2) / 2 = 1.5, *x*14 = *x*24 = (2+3) / 2 = 2.5. The strong point of average method is easy to calculated mean value and its cost is low but it is based on the assumption that column is random variable but this thing is not totally exact because the null hypothesis requires that only *Xi* (s) being row vectors are random variables. On the contrary, mixture model approach follow hard on the requirement of null hypothesis.

Another way to predict empty values is to use Bayesian rule as estimation tool. It is better than average method because it takes advantage of probability distribution under *2-*dimension data. Let *z* = *xuv* be the random variable that are dependent on row index *u* and column index *v* which, in turn, are random variables. According to Bayesian rule, the probability of *z* given *u* and *v* is:

Assuming that *u* and *v* are mutually independent given *z*, *P*(*z | u*, *v*) is re-written.

Let *θ* = *P*(*z*) = {*θ*1, *θ*2,…, *θk*} be the parameter of *P*(*z | u*, *v*) where *θi* corresponds with the *ith* value of *z*, we have:

The goal of Bayesian estimate method is to determine *P*(*z | u*, *v*) via finding out optimal parameter *θ*. This method is iterative process whose iteration includes two steps: calculating *P*(*z | u*, *v*) based on *θ* and re-calculating *θ*. Let *t* and *θt* denotes *tth* iteration and the estimate of parameter *θ* at *tth* iteration. Note that *θ*0 can be initialized by 0.5. These two steps are described as below.

1. The current probability *P*(*z | x, y*) is calculated based on previous estimate of parameter *θt-*1.
2. Assigning current *P*(*z | x, y*) to current estimate of parameter *θt*.

If the deviation | *θt* – *θt-*1| is less than a threshold *ε*, then algorithm is stopped; otherwise go back step 1.

With example in table 2, suppose ranges of *u, v* and *z* are {1, 2, 3, 4}, {1, 2, 3, 4} and {1, 2, 3}, the conditional probabilities *P*(*u | z*) and *P*(*v | z*) are calculated and shown in table 3.

|  |  |  |  |
| --- | --- | --- | --- |
|  | z = 1 | z = 2 | z = 3 |
| u = 1 | 0.5 | 0.25 | 0 |
| u = 2 | 0 | 0.25 | 0.5 |
| u = 3 | 0.5 | 0.25 | 0 |
| u = 4 | 0 | 0.25 | 0.5 |
| v = 1 | 0.5 | 0.25 | 0 |
| v = 2 | 0 | 0.25 | 0.5 |
| v = 3 | 0.5 | 0.25 | 0 |
| v = 4 | 0 | 0.25 | 0.5 |

**Table 3.** Conditional probabilities *P*(*u | z*) and *P*(*v | z*)

Suppose the parameter is initialized by 0.5, we have *θ*0 = *P*0(*z | u*, *v*) = 0.5. It means that the prior probabilities of *z* is 0.5. It is easy to calculate the posterior probabilities *P*(*z | u, v*), for example:

By the similar way, the probabilities *P*(*z | u, v*), at the first iteration, over all values of *u*, *v* and *z* are calculated and shown in table 4.

|  |  |  |  |
| --- | --- | --- | --- |
|  | z = 1 | z = 2 | z = 3 |
| u=1, v=1 | 0.8 | 0.2 | 0 |
| u=1, v=2 | 0 | 1 | 0 |
| u=1, v=3 | 0.8 | 0.2 | 0 |
| u=1, v=4 | 0 | 1 | 0 |
| u=2, v=1 | 0 | 1 | 0 |
| u=2, v=2 | 0 | 0.2 | 0.8 |
| u=2, v=3 | 0 | 1 | 0 |
| u=2, v=4 | 0 | 0.2 | 0.8 |
| u=3, v=1 | 0.8 | 0.2 | 0 |
| u=3, v=2 | 0 | 1 | 0 |
| u=3, v=3 | 0.8 | 0.2 | 0 |
| u=3, v=4 | 0 | 1 | 0 |
| u=4, v=1 | 0 | 1 | 0 |
| u=4, v=2 | 0 | 0.2 | 0.8 |
| u=4, v=3 | 0 | 1 | 0 |
| u=4, v=4 | 0 | 1 | 0 |

**Table 4.** Posterior probabilities of *z*, *P*(*z | u, v*)

At next iteration posterior probabilities *z*, *P*(*z | u, v*) are computed by similar way except that the parameter *θ* is assigned by values in table 3. In other words, the parameter for next iteration is assigned as current posterior probabilities *z*, *P*(*z | u, v*). Now we estimate missing values such as *x*31, *x*41, *x*32, *x*42, *x*13, *x*23, *x*14 and *x*24. The posterior probabilities for *x*31 = 1, *x*31 = 2, *x*31 = 4 are *P*(*z* = 1 | *u* = 3, *v* = 1) = 0.8, *P*(*z* = 2 | *u* = 3, *v* =1) = 0.2 and *P*(*z* = 3 | *u* = 3, *v* = 1) = 0, respectively. Because *P*(*z* = 1 | *u* = 3, *v* = 1) = 0.8 is maximal, *x*31 receives value 1, *x*31 = 1. By the similar way, we have *x*41 = 2, *x*32 = 2, *x*42 = 3, *x*13 = 1, *x*23 = 2, *x*14 = 2 and *x*24 = 3. These estimates are much better than ones resulted from average method even if the algorithm have just come over one iteration. Table 5 shows the complete sample whose missing values are estimated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | v = 1 | v = 2 | v = 3 | v = 4 |
| u = 1 | *x*11 = 1 | *x*12 = 2 | *x*13 = 1 | *x*14 = 2 |
| u = 2 | *x*21 = 2 | *x*22 = 3 | *x*23 = 2 | *x*24 = 3 |
| u = 3 | *x*31 = 1 | *x*32 = 2 | *x*33 = 1 | *x*34 = 2 |
| u = 4 | *x*41 = 2 | *x*42 = 3 | *x*43 = 2 | *x*44 = 3 |

**Table 5.** Complete sample whose missing values are estimated by Bayesian method.

However, the drawback of Bayesian estimate method in hypothesis testing is that it gives three assumptions as below.

* Sample scalar values *xij* (s) are considered as random values while vector *Xi* (s) are random variables in null hypothesis. Scalar distribution is different from vector distribution with regard to the modality of data. When the modality of data is transformed, the meaning of data is changed unpredictably.
* Row index *u* and column index *v* are also considered as random variables but this assumption doesn’t exist in null hypothesis.
* Row variable *u* and column variable *v* are mutually independent given sample value *xij*.

In general, mixture model approach is better than filling missing values approaches in flavor of hypothesis testing because it does not give any additional assumption about sample space except assumptions that null hypothesis issues such as the normality of sample data. In other words, it conserves attributes of sample data.

4. Conclusions

The basic idea of this research is to analyze the global distribution as a set of partial (or potential) distributions. Because it is impossible to apply the global distribution into incomplete sample, the research applies each partial distribution into a sub-sample of incomplete sample where sub-sample is defined as a piece of incomplete sample so that such sub-sample contains complete observation vectors. Partial distributions are unified into a mixture model and expectation maximization (EM) algorithm is used to find out optimal parameters of mixture model such that these optimal parameters maximize the likelihood function. The ratio of the maximum likelihood of null hypothesis to the maximum likelihood of alternative hypothesis is totally determined based on these optimal parameters. Consequently, this ratio is used to test null hypothesis in flavor of alternative hypothesis by the normal way that is to take advantage of chi-square distribution. The excellent of this research is that there is no requirement of filling missing values existing in incomplete sample. When missing values are replaced by estimated values, inherent attributes of sample data changed or disturbed even if such estimated values are considered as the best-predicted values. In other words, this method maintains inherent attributes of sample data. Moreover, the higher the density of missing values is, the more effective this research is. Especially, if the data incompletion occurs frequently or conforms a particular period or rule in process of collecting samples, then the corporation of partial distributions and likelihood ratio test delivers the best result. However if missing values occur randomly such as pepper-and-salt noise, this approach is not good choice for test because pepper-and-salt noise data has no inherent features and so partial distributions cannot be specified and estimated precisely.

References

1. W. Hardle and L. Simar, Applied Multivariate Statistical Analysis, Berlin: Research Data Center, School of Business and Economics, Humboldt University, 2013, p. 486.
2. J. A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models,” University of Washington, Berkeley, 1998.
3. B. Sean, “The Expectation Maximization Algorithm - A short tutorial,” Sean Borman's Homepage, 2009.
4. L. Nguyen, Matrix Analysis and Calculus, 1st ed., C. Evans, Ed., Hanoi: Lambert Academic Publishing, 2015, p. 72.

1. **1Corresponding author:** Loc Nguyen, postdoctoral, research field: computer science, statistics, and mathematics. E-mail: ng\_phloc@yahoo.com. [↑](#footnote-ref-1)