Estimating parameters of factor analysis model (maximum likelihood method) by using EM algorithm with application

Thanoon younis thanoon alshakerchy
Department financial and banking technical
Technical institute / Nainavah

Received **25 / 06 / 2008**
Accepted **15 / 10 / 2008**

الملخص

خوارزمية تعظيم التوقع Expectation Maximization Algorithm (EM) تستخدم لإيجاد مقدر يحمل نفس صفات مقدر الإمكان الأعظم مع الأخذ بنظر الاعتبار وجود نوعين من البيانات ، البيانات المشاهدة (observed data) والبيانات المخفية (المفقودة) (missing data)، تم في هذا البحث تقدير معلمات نموذج التحليل العاملي (طريقة الامكان الأعظم) باستخدام خوارزمية تعظيم التوقع وتم تطبيق التحليل العاملي (طريقة الأمكان الأعظم) على بيانات لمرضى مصابين بسرطان الثدي، وتبين من النتائج أهمية جميع المتغيرات ماعدا المتغيرين الأول (مستوى التعليم) و المتغير الخامس (استخدام الهرمونات العلاجية).

Summary

Expectation maximization algorithm (EM) is used to create estimator with the same qualities of maximum likelihood Estimator taking into consideration the existence of two types of data, Data viewing (observed data) and hidden data (missing data), in this research the estimating parameters of factor analysis model (maximum likelihood method) has been done by using expectation maximization algorithm and applied factor analysis (maximum likelihood method) on data for patients infected with breast cancer, and found from the results importance all of the variables in breast cancer variables except first variable (level of education) and fifth variable (hormone treatment used).

Introduction

Factor analysis is a mathematical model which attempts to explain the correlation between a large set of variables in terms of a small number of underlying factors. A major assumption of factor analysis is that it is not possible to observe these factor directly; the variables depend upon the factors but are also subjects to random errors. Such an assumption is particularly well-suited to subjects like psychology where it is not possible to measure exactly the concepts one is interested in (e.g. "intelligence") and in fact it is often ambiguous just how to define these concepts. (Mardia, Kent, & Biby, 1979).

Lawley (1940) developed the method of maximum likelihood factor analysis at a time when the field had considerable need for a sound theoretic statistical foundation. The centroid method was in use at the Thurstone laboratory for analyses of major factor analytic studies. The number of factors was a major problem. In that the maximum likelihood method provided a statistical test for significance of factors this was a most promising development. However, at that time, computing facilities were very limited so that the extensive computations required by the maximum likelihood method made this method unavailable. Application remained limited until developments of modern digital computers made them sufficiently powerful to accomplish the required computations. Lord (1956) used Whirlwind I computer in the analysis of a covariance matrix among 39 attributes. He used an iterative technique suggested by Lawley (1942). Convergence was very slow. This method had several other problems such as converging on secondary maximum. Jöreskog (1967) contributed advanced procedures which could use more powerful analysis methods such as the Newton-Raphson iterations. With the computer developments and Jöreskog's contributions, maximum likelihood factor analysis is quite feasible. This method has many desirable statistical properties such as consistency, normality, efficiency. (Tucker & MacCallum, 1997)

Aim of the research

1- estimation parameters of factor analysis model (Maximum likelihood method) by using expectation maximization algorithm.
2- selecting variables that have significant affecting on the study by using (maximum likelihood method) in factor analysis.

Factor Analysis Model

The factor analysis model is:

$$X = QF + U + u \qquad ...(1-1)$$

where Q is a ($p \times k$) matrix of the (nonrandom) loadings of the common factors $F(k \times 1)$ and U is a $(p \times 1)$ matrix of the (random) specific factors.

It is assumed that the common factors *F* are uncorrelated random variables and that the specific factors are uncorrelated and have zero covariance with the common factors. More precisely, it is assumed that: $EF = 0$, $\mathrm{Var}(F) = Ik$, $EU = 0$, $\mathrm{Cov}(U_i, U_j) = 0$, $i \neq j$, and $\mathrm{Cov}(F,U) = 0$.

The random vectors F and U are unobservable. Define: $\mathrm{Var}(U) = \Psi = \mathrm{diag}(\psi_{11}, \ldots, \psi_{pp})$; then the variance matrix of X can be written as $\mathrm{Var}(X) = \Sigma = QQ^T + \Psi$, and we have for the ith component of the random vector X that $\sigma_{jj} = \mathrm{Var}(X_j) = \sum_{l=1}^{k} q_{jl}^2 + \Psi_{ij}$. The quantity $h_j^2 = \sum_{l=1}^{k} q_{jl}^2$ is called the communality and ψ_{jj} the specific variance. The objective of factor analysis is to find a small number, k, of common factors leading to large communalities and small specific variances. (härdle & Hlávka , 2007) (Härdle & simiak, 2007).

In general:

X: observed variables.
Q : matrix of factor loadings.
F : hidden variables with distribution N(0; I).
u : noise with distribution N(0; ψ).
U : mean of observed variables (assume zero).

The Expectation-Maximization Algorithm

The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. In ML estimation, we wish to estimate the model parameter(s) for which the observed data are the most likely.

Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation, or E-step, the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology.

In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing data.

Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration. (Borman, 2006).

Derivation of Expectation-Maximization Algorithm

As we have stated previously, the EM algorithm takes into account [hidden variables], according to the formula: (المشهداني, 2003)

$$f(z, y \mid \theta) = f(z \mid y; \theta) \times f(y \mid \theta) \quad \ldots (2-1)$$

Or can be written in the form:

$$f(y \mid \theta) = \frac{f(z, y \mid \theta)}{f(z \mid y, \theta)} \quad \ldots (3-1)$$

and taking the logarithm of the equation (3-1) we get:

$$\log f(y \mid \theta) = \log f(z, y \mid \theta) - \log f(z \mid y; \theta) \quad \ldots (4-1)$$

Taking expectation For x we get:

$$\log f(y \mid \theta) = \log \int (\log f(z, y \mid \theta) f(z \mid y; \theta^*))dx - \int \log f(z \mid y; \theta) f(z \mid y; \theta^*))dx$$

$$= Q(\theta, \theta^*) - H(\theta, \theta^*) \quad \ldots . (5-1)$$

As θ^* represents the estimator θ in a certain stage of expectation repetitions in the algorithm

Where as: --

$$Q(\theta, \theta^*) = \int (\log f(z, y \mid \theta) f(z \mid y; \theta^*))dx \quad \ldots (6-1)$$

$$H(\theta, \theta^*) = \int (\log f(z, y \mid \theta) f(z \mid y; \theta^*))dx \quad \ldots (7-1)$$

Expectation maximization algorithm is called expectation maximization and write in short (EM) because each repetition contains steps for $(i+1)^{st}$ repeating steps and can thus say that expectation maximization algorithm includes two steps:

1- (E-step)

Calculated through

$$Q(\theta, \theta^{(i)}) = E(\log L(\theta \mid z, y))$$

$$= \int \log L(\theta \mid z, y) * f(z \mid y; \theta^{(i)})dx \quad \ldots (8-1)$$

We note that in the calculation of expectation step, we find expectation take logarithm of the likelihood function $L(\theta \mid x)$ for all possible data, certified observed variables Y and parameter θ.

To calculate the expectation should be conditional function $f(z \mid y; \theta)$ known. In the family of exponential distributions expectation step account contains sufficient statistical expectation (essentially for the full data) with certified observed variables.

2 - (M-step)

contain maximization of the logarithm likelihood function For the complete data for θ. This maximization always be easier to use numerical methods.

Estimating parameters of factor analysis model (maximum likelihood method)(Q and ψ) using EM :-

$$p(Q,\Psi)=p(x_i,\, f_i \mid Q,\Psi)$$

but the likelihood function is:

$$L(Q,\Psi)=\prod_i^N p(x_i,\, f_i \mid Q,\Psi)$$

taking the logarithm for both sides

$$\log(L(Q,\Psi))=\log\prod_i^N p(x_i,\, f_i \mid Q,\Psi)$$

$$=\sum_i^N \log p(x_i,\, f_i \mid Q,\Psi)$$

$$=\sum_i^N \log\, p(x_i \mid f_i, Q,\Psi)\, P(f_i \mid Q,\Psi)$$

$$=\sum_i^N \log\, p(x_i \mid f_i, Q,\Psi)+\sum_i^N \log P(f_i \mid Q,\Psi)$$

but the distribution of (f) is independent of Q and ψ

$$=\sum_i^N \log\, p(x_i \mid f_i, Q,\Psi)+\sum_i^N \log P(f_i) \qquad \ldots(9-1)$$

the second term in equation (9-1) is independent of Q and ψ ,it suffices (for the purpose of estimating Q and ψ)to only deal with the term .

$$L=\sum_i^N \log p(x_i \mid f_i, Q,\Psi)$$

now the expected and covariance values for distribution P(x|f) is given by :-

$$E(x \mid f)=E\{(Qf+u) \mid f\}=Qf \qquad \ldots(10-1)$$

$$Cov(x \mid f)=E\{(x-Qf)(x-Qf)^T \mid f\}=E\{uu^T \mid f\} \qquad \ldots(11-1)$$

Hence we can expand L as follow :-

$$L=\sum_i^N \log \frac{1}{(2\pi)^{\frac{d}{2}} \mid \Psi \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-Qf_i)^T \Psi^{-1}(x-Qf_i)\}$$

$$=-\frac{ND}{2}\log(2\pi)-\frac{N}{2}\log\mid \Psi \mid-\frac{1}{2}\sum_i^N (x_i^T \Psi^{-1} x_i-2x_i^T \Psi^{-1} Q\, f_i+f_i^T Q^T \Psi^{-1} Q\, f_i)$$

$$=-\frac{ND}{2}\log(2\pi)-\frac{N}{2}\log\mid \Psi \mid-\frac{1}{2}\sum_i^N (x_i^T \Psi^{-1} x_i-2x_i^T \Psi^{-1} Q\, f_i+Tr[Q^T \Psi^{-1} Q\, f_i\, f_i^T])$$

In the last step we have exploited the relation $X^T AX=Tr[\,AXX^T\,]$. Taking the expectation of L according to $p(x_i \mid f_i, Q,\Psi)$, we get:

$$E\{L\}=k-\frac{N}{2}\log|\Psi|-\frac{1}{2}\sum_{i}^{N}(x_i^T\Psi^{-1}x_i-2x_i^T\Psi^{-1}QE\{f_i|x_i\}+Tr[Q^T\Psi^{-1}QE\{f_i\,f_i^T|x\}]) \quad ...(12-1)$$

Maximizing eq. (12 -1) w.r.t Q

$$\frac{\partial E\{L\}}{\partial Q}=-\frac{1}{2}\sum_{i}^{N}(-2\Psi^{-1}x_iE\{f_i|x_i\}^T+2\Psi^{-1}QE\{f_i\,f_i^T|x\}])=0$$

Where we have used the relation

$$\frac{\partial A^TXB}{\partial X}=AB^T, and \frac{\partial Tr[X^TAXB]}{\partial X}=AXB+A^TXB^T$$ Hence

$$Q\sum_{i}^{N}E\{f\,f^T|x\}=\sum_{i}^{N}x_iE\{f_i|x_i\}^T$$ and we arrive at the update equation

$$Q=(\sum_{i}^{N}x_iE\{f_i|x_i\}^T)(\sum_{i}^{N}E\{f_i\,f_i^T|x\})^{-1} \quad ...(13-1)$$

Maximizing equation (12-1) for Ψ^{-1}

$$\frac{\partial E\{L\}}{\partial\Psi^{-1}}=\frac{N}{2}\Psi-\frac{1}{2}\sum_{i}^{N}(x_ix_i^T-2x_iE\{f_i|x_i\}^TQ^T+QE\{f_i\,f_i^T|x\}Q^T)$$

$$=\frac{N}{2}\Psi-\frac{1}{2}\sum_{i}^{N}x_ix_i^T+(\sum_{i}^{N}x_iE\{f_i|x_i\}^TQ^T-\frac{1}{2}Q(\sum_{i}^{N}E\{f_i\,f_i^T|x\})Q^T$$

$$=0$$

Where we have used the relation $\frac{\partial\log|X|}{\partial X}=(X^{-1})^T, and \frac{\partial A^TXB}{\partial X}=AB^T$

Hence

$$\Psi=\frac{1}{N}[\sum_{i}^{N}x_ix_i^T-2(\sum_{i}^{N}x_iE\{f_i|x_i\}^T)Q^T+Q(\sum_{i}^{N}E\{f_i\,f_i^T|x\})Q^T] \quad ...(14-1)$$

substituting the value of Q from the update equation we get

$$\Psi=\frac{1}{N}[\sum_{i}^{N}x_ix_i^T-2(\sum_{i}^{N}x_iE\{f_i|x_i\}^T)Q^T+(\sum_{i}^{N}x_iE\{f_i|x_i\}^T)(\sum_{i}^{N}E\{f_i\,f_i^T|x\})^{-1}(\sum_{i}^{N}E\{f_i\,f_i^T|x\})Q^T]$$

giving use the update equation

$$\Psi=\frac{1}{N}diag[\sum_{i}^{N}x_ix_i^T-(\sum_{i}^{N}x_iE\{f_i|x_i\}^T)Q^T] \quad ...(15-1)$$

(The application side):

The data is collected by the consultation clinic of breast diseases and Hazem Al-Hafiz hospital of cancer tumors and nuclear treatment as considered adopted health institution in diagnosis and treatment of cancer diseases in general and breast cancer in particular, through the information in drum of every patient in these healthy institution the variables appointed which are believed that have an impact in breast cancer.

What have been collected is (134) diseased case (infected with the disease) but actually the study sample (93) diseased case and neglected

(41) cases for different reasons according to accuracy of mentioned information in it or it is contain of unknown paragraphs .
The shape of sample size taken from the consultation clinic of breast diseases 56% of the total number of reviews were the ratio of housewives in sample 79% , 21% were married. The size of the sample taken from the Hazem Al-Hafiz hospital of cancer tumors and nuclear treatment 44% of the total number which are all diagnosed with breast cancer disease and mostly of housewives with the rate of 83% and married in that sample constituted (88%) , the study included eight variables which is believed that have impact on the disease , these variables represent predict variables and explanatory which have been identified after review of some specialist physicians in breast cancer disease , this data consists of (8) variables and as follow: (2005, العبيدي)

X_1 : Level of education. which contain 1: high studies, 2: graduate, 3: secondary or diploma, 4: intermediate, 5:primary.
X_2 : Marital state. which contain 1: bachelor, 2: married, 3: widower, 4: separated.
X_3 : Age at marriage. which contain, 0:not married, 1: married before 35 year, 2: married after 35 year.
X_4 : Number of pregnancy.
X_5 : Hormone treatment used.
X_6 : Natural lactation.
X_7 : Time of natural lactation.
X_8 : Disease in breast. which contain 0: not found, 1: node, 2: strange secretion, 3: Tumour, 4: pain, 5:1+2, 6:1+3, 7:1+4, 8:2+3, 9:2+4, 10:3+4.

Statistical analysis :

Statistical program Spss (V11.5) is used to find the factor analysis (maximum likelihood method) and application on data of breast cancer, and were the factor analysis result as follows:-

Table (1)

Total Variance Explained for simple correlation matrix

% Cumulative	% of Variance	Eigen values	factors
38.442	38.442	3.075	1
54.124	15.682	1.255	2
66.814	12.691	1.015	3

1. factors were identified that could be considered a significant to explain the nature and direction of the relationship between the studied variables with three factors based on eigen values of simple correlation matrix, which was chosen on the basis that this value greater than one (λ j>1) (1987, الراوي) notes that in table (1) : --

Table (2)

Total Variance Explained for reduce correlation matrix

% Cumulative	% of Variance	Eigen values	factors
25.814	25.814	2.065	1
39.705	13.891	1.111	2
50.900	11.195	0. 896	3

2. variables affecting was discriminatory in each factor through rotated factor loading matrix (to be chosen this value when it is larger than 0.5 in order to be significant influence) and the results were as follows: --

Table (2)

Rotated factor loadings

rotated Factor loadings			**variables**
3	2	1	
0.008	0.188	0.031	1
-0.207	0.256	0.708	2
0.171	0.029	0.818	3
-0.142	0.924	0.354	4
0.228	-0.059	-0.023	5
-0.382	0.260	0.590	6
-0.249	0.262	0.639	7
0.737	0.128	-0.108	8

As for communalities the table (4) shows the estimated communalities by using multiple correlation coefficients and derived communalities from rotated factor loadings matrix.

Table (4)
Explain initial and extracted communality

Extracted communality	Initial communality	variables
0.037	0.081	X1
0.609	0.539	X2
0.699	0.419	X3
0.999	0.431	X4
0.056	0.069	X5
0.562	0.493	X6
0.539	0.475	X7
0.571	0.225	X8

Interpretation the results :-

In general we note the importance of all variables in breast cancer disease except the first variable (level of education) and fifth variable (hormone treatment used), which do not show a significant influence in each of three factors.

As for the affected variables that have significant impact its importance is different from the sequence factor and the loading of variables with factors indicated that the second variable (marital state), third (age at marriage) and sixth (Natural lactation) and seventh (Time of natural lactation) have primary importance in the disease of breast cancer, because this variables showed in the first factor, which I interpreted as per 25,814% of the total variance, and on this basis the fourth variable (number of pregnancy) came in second, third place and the final find the eighth variable (Disease in breast).

Conclusions :

1- Estimating parameters of factor analysis model (maximum likelihood method) and as shown in equations (13-1) and (15-1).
2- The diagnosis of affected factors has been done in breast cancer disease through three factors interpreted per 50.9% from the total variance.
3- Results showed the affect of all the variables on breast cancer disease, except first variable (level of education) and fifth variable (hormone treatment used), where the communality of these two variables less possible.
4- We can determine the degree of importance of each variable of the remaining variables affecting the disease through a matrix of rotated factor loading and as shown in table 3, where results showed the

importance of variables second (marital state), third (age at marriage), sixth (Natural lactation) and seventh (Time of natural lactation) first-class followed by the fourth variable (number of pregnancy) came in second while eighth variable (Disease in breast) came third and last.

References

Arabic References :

1) الـراوي، خاشـع محمـود (1987): "**المـدخل إلـى تحليـل الانحـدار**". مديريـة دار الكتـب للطباعة والنشر –جامعة الموصل – لطبعة الأولى.

2) العبيدي، ندوى خزعل (2005): "**دراسة مقارنة لبعض طرائق اختبار مجموعة جزئية في نماذج الانحدار المتعددة مع تطبيق في سرطان الثدي**"، رسالة ماجستير، علوم الحاسبات والرياضيات، جامعة الموصل.

3) المشـهداني، احمـد إدريـس (2003): "**اسـتخدام خوارزميـة تعظـيم التوقـع فـي الترجمـة العربيـة–الإنكليزيـة باتجـاهين**"، رسـالة ماجسـتير، علـوم الحاسـبات والرياضـيات، جامعـة الموصل.

English References :

4) Borman, Sean (October 14, 2006): "**The Expectation Maximization Algorithm A short tutorial**", http://www.seanborman.com/publications/EM_algorithm.pdf

5) härdle, Wolfgang & Hlávka, Zdeněk (2007): "**Multivariate Statistics: Exercices and solutions**", Berlin and Prague.

6) Mardia, K. V, Kent, J. T. & Biby, J. M (1979): "**Multivariate Analysis**". Academic Press, London.

7) W. Härdle & L. simiak (2007): "**Applied Multivariate Statistical Analysis**", Berlin and Louvain -la- Neuve.

8) Ledyard R Tucker and Robert C. MacCallum: (1997) "**chapter 9 FACTOR FITTING BY STATISTICAL FUNCTIONS**", http://www.unc.edu/~rcm/book/ch9.pdf .