R で因子分析：入門編

中澤　港 (minato-nakazawa@umin.net)

28 February 2014

本文書は 2011 年に公開した http://minato.sip21c.org/swtips/factor-in-R.pdf の和訳＋αである。
本文書はとてもラフな草稿なので，注意して読まれたい。ご提案やコメントは歓迎する。

1　参考文献

エディンバラ大学の心理学者，Timothy Bates 教授のウェブサイト[*1]は，大変助けになる。

群馬大学の青木繁伸教授により提供されているウェブサイト[*2]も非常に助けになる（英語圏の読者にとっては残念なことに，和文で書かれているが)。

2　因子分析の目的

真面目な説明　観察された変数の背後に隠れている因子を見いだすこと。この隠れた因子は直接測定できないが，観察された変数の「自然のグルーピング」になっている[*3]。

実用的な説明　互いに相関のある変数について，情報を集約して数を減らすこと。この意味では，主成分分析と似ている（向きは逆だが)。

3　因子分析の基本的な使い方

入力データ　ある程度のサンプルサイズと大きな変数をもつ数値行列で，通常，サンプルサイズは **300 より多い。 変数数に対する対象者の人数の比**は，通常，**2:1 から 10:1** の範囲をとる。原則として変数は正規分布に従うべきでだし，外れ値は含まない方がよい。他の変数と関連のない変数は分析に含めるべきではない。お互いに相関係数 1.0 の変数は含めることができない。どちらかを除外するか，適切であれば両者の和をとって合成変数として用いることは可能である。

出力　(1) 因子負荷量は，各変数がその元になる潜在因子と相関している程度を意味する（その際，さまざまな回転が用いられる[*4]），(2) 因子得点は，通常，各個人の応答と因子負荷量の積の和で（ただし複数の計算法があり，どの方法が最適かについて統一見解はない)，各個人の特性がどの程度その因子によって説明されるかを示す。

回転　回転の方法は 2 つに大別される。直交回転は，因子間の独立性を保ったまま因子ベクトルを回転させるが，斜交回転では因子間に相関が出てもいいことにしている。因子が理論的に相互依存を許してもいいときに，後者を考えるべきである。前者には最もよく使われていて単純なバリマックス回転が含まれる。バリマックス回転は，因子ごとの分散を最大化する。後者にはプロマックス回転やオブリミン回転が含まれる。

[*1] http://www.psy.ed.ac.uk/people/tbates/lectures/methodology/

[*2] http://aoki2.si.gunma-u.ac.jp/lecture/PFA/pfa6.html
に因子分析についての説明がある。
http://aoki2.si.gunma-u.ac.jp/R/kmo.html
http://aoki2.si.gunma-u.ac.jp/R/Bartlett.sphericity.test.html
は後述する KMO，MSA 及び Bartlett の球面性検定の関数定義。

[*3] データセット内のお互いに強く相関する変数のサブセットで，他の変数とは弱い相関をもつ。見つかった因子は，理論的に解釈可能な，隠れた「次元」に対応するはずである。

[*4] 最初の因子負荷量は，第一因子への負荷を最大にするように計算されるので，たいていの変数が 1 つ以上の因子に対して高い負荷量をもってしまい，因子の解釈が難しくなる。そこで，適切な**回転**をすると，この問題が解決することが多い。

因子分析のための道具　スクリープロット，バートレットの球面性検定，カイザー・マイヤー・オルキンのサンプリング適切性基準，平行分析 (Parallel Analysis) が便利。因子数がうまく決定できたら，各因子に含まれる変数が単一軸の加法的スコアになっているかどうかをチェックするために，クロンバックの α 係数を計算する（通常，それらの因子の和が信頼できるスコアであるためには，クロンバックの α が 0.7 より大きくなければいけない)。

推定された因子を解釈する際には，因子に適切な名前（意味）をつけることが必要である。因子がうまく推定できたと判定するには，因子負荷量が高い変数が少なくとも 3 つあるべきである。もし 1 つか 2 つしか因子負荷量が高い変数がないときは，因子数が多すぎるか，元の変数間に多重共線性が存在する可能性がある。

4　因子分析の基本モデル

300 人で変数 10 個（$X_1, X_2, ..., X_{10}$）の場合を考えよう。これら 10 個の変数の背後に，もし 2 個の潜在因子（F_1 と F_2）があるとしたら，各変数は，これらの因子によって次のように説明される。

$$X_1 = \beta_{1\cdot1}F_1 + \beta_{2\cdot1}F_2 + \epsilon_1$$
$$X_2 = \beta_{1\cdot2}F_1 + \beta_{2\cdot2}F_2 + \epsilon_2$$
$$\vdots$$
$$X_{10} = \beta_{1\cdot10}F_1 + \beta_{2\cdot10}F_2 + \epsilon_{10}$$

ここで，β は，各変数と潜在因子との相関を意味し，これを**因子負荷量 (Factor loadings)** と呼ぶ。ϵ は誤差分散を意味する。言い換えると，推定された因子では説明できなかった**独自性 (uniqueness)** でもある[*5]。しかし，潜在因子 F_1 と F_2 は測定された値ではない。だから，我々は，主因子法，最小残差法，最尤法などの様々な方法で，反復計算させながら推定しなくてはいけない[*6]。

回転する前は，因子 F_1 と F_2 は独立と仮定されている。いま，n 番目（n は区間 [1, 300] の整数）の人の i 番目の変数の値を $X_i(n)$ と書くと，その人の因子得点（ここでは $FS_1(n)$ と $FS_2(n)$）は，次のように得られる（ただし，これは最も単純な方法である。因子得点として提案されている指標値は，この他にもいくつかある)。計算に使う変数は，β の絶対値が十分大きい（通常，0.3 とか 0.4，あるいは 0.5 以上とする）ものに限るのが普通。

$$FS_1(n) = \sum_{i=1}^{10} \beta_{1\cdot i} X_i(n)$$
$$FS_2(n) = \sum_{i=1}^{10} \beta_{2\cdot i} X_i(n)$$

5　いくつの因子を推定すべきか？

この問題には以下のようにいくつかの基準が提案されているが，100% これが良いという検定法などは存在しない。

スクリープロットを描く　最初に可能な限り多くの因子を仮定して因子分析を行い，各因子によって説明される分散を代表するものとしての固有値（あるいは同じ意味で因子負荷量の二乗和）を，大きい順に線でつないだ折れ線グラフがスクリープロットである。折れ線が急に激しく落ち込む変数があれば，その直前が適切な因子数と考えられる。

[*5] 独自性を 1 から引いたものを共通性 (communality) という。後述する `rela` パッケージの関数では，共通性が出力される。

[*6] 主成分分析では，各主成分は，測定された変数の線形結合として定式化されるので，反復推定は必要ない。

パラレル分析をする　実際のスクリープロットを，ランダムにリサンプルしたデータから計算したスクリープロットと比較する。2 つのプロットが交差する点が適切な因子数であると考える。

固有値が 1 を超えている間　固有値が 1 を超えている間は，変数 1 つよりも情報量が多いと考えられるので。

6　因子分析の適切性をチェックする

因子分析の適切性をチェックするための方法がいくつかある。

サンプルサイズの適切性の基準　サンプルサイズは 50 では非常に乏しい (very poor)。100 でも乏しい (poor)。200 ならまあまあ (fair)，300 なら十分 (good)，500 なら非常に良い (very good)。1,000 を超えたら極めて優れている (excellent) といえる (Comfrey and Lee, 1992, p.217)。

KMO と MSA　KMO とは，Kaiser-Meyer-Olkin が提唱した因子分析全体についてのサンプリング適切性基準であり，MSA とは Measures of Sampling Adequacy の頭語で，それぞれの変数についての個別のサンプリング適切性基準である。データセットの中に，十分な数の因子が存在するかどうかを示す指標値である。技術的には，変数間の相関係数の偏相関係数に対する比を計算する。もし偏相関係数が生の相関係数と同じような値なら，それらの変数は互いに分散をあまり共有していないことを意味する。KMO の範囲は 0.0 から 1.0 で，0.5 以上が望ましい[*7]。また，MSA が 0.5 未満の変数は，その変数がどの因子グループにも属していないことを示すので，因子分析から除くべきである。

群馬大学の青木繁伸教授は，前述したウェブサイトで，KMO と MSA を計算するための次の関数定義を公表している。

```
kmo <- function(x)
{
        x <- subset(x, complete.cases(x))       # Remove the cases with any missing value
        r <- cor(x)                             # Correlation matrix
        r2 <- r^2                               # Squared correlation coefficients
        i <- solve(r)                           # Inverse matrix of correlation matrix
        d <- diag(i)                            # Diagonal elements of inverse matrix
        p2 <- (-i/sqrt(outer(d, d)))^2          # Squared partial correlation coefficients
        diag(r2) <- diag(p2) <- 0               # Delete diagonal elements
        KMO <- sum(r2)/(sum(r2)+sum(p2))
        MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
        return(list(KMO=KMO, MSA=MSA))
}
```

バートレットの球面性検定　変数間の相関が偶然期待されるより大きいという仮説を検定する。技術的には行列が単位行列であるかどうかを検定する。p 値が有意である場合，対角以外のすべての相関がゼロであるという帰無仮説が棄却される。

[*7] Kaiser (1974) の提案によれば，0.5 未満では不適切，0.5 以上 0.6 未満は悲惨なレベル (miserable)，0.6 以上 0.7 未満は良くも悪くもなく (mediocre)，0.7 以上 0.8 未満は並 (middling)，0.8 以上 0.9 未満は賞賛に値し (meritorious)，0.9 以上なら極めて優れている (marvelous)。

バートレットの球面性検定についても，群馬大学の青木繁伸教授が前述したウェブサイトで次の関数定義を公表している。

```
Bartlett.sphericity.test <- function(x)
{
       method <- "Bartlett's test of sphericity"
       data.name <- deparse(substitute(x))
       x <- subset(x, complete.cases(x)) # Remove the cases with any missing value
       n <- nrow(x)
       p <- ncol(x)
       chisq <- (1-n+(2*p+5)/6)*log(det(cor(x)))
       df <- p*(p-1)/2
       p.value <- pchisq(chisq, df, lower.tail=FALSE)
       names(chisq) <- "X-squared"
       names(df) <- "df"
       return(structure(list(statistic=chisq, parameter=df, p.value=p.value,
              method=method, data.name=data.name), class="htest"))
}
```

7　R で因子分析を実行するための関数

factanal　この関数は標準でインストールされる。因子負荷量を計算するのに最尤法を用いる。推定すべき因子数は明示的に指定せねばならない。バリマックス回転とプロマックス回転が可能である。入力データは行列またはデータフレーム。

paf　この関数は **rela** パッケージに含まれているので，**rela** パッケージをインストールし，使用前にメモリにロードする必要がある。因子負荷量を計算するのに主因子法を用いる。適切な因子数は，固有値の基準によって自動的に決定され（固有値をいくつ以上にするかは，`eigencrit=`オプションで指定できる。デフォルトは 1 である），KMO と MSA が自動的に計算されるので，初心者用と言われている。回転は提供されていない。入力データは行列。

fa　この関数は **psych** パッケージに含まれている。`fm=`オプションで因子負荷量の計算方法を指定できる（`"minres"`で最小残差法，`"ml"`で最尤法，`"pa"`で主因子法）。推定する因子数は `nfactors=`オプションで指定せねばならない。`rotate=`オプションでさまざまな回転方法を指定できる（`"none"`，`"varimax"`，`"quartimax"`，`"bentlerT"`，`"geominT"`，`"oblimin"`，`"simplimax"`，`"bentlerQ"`，`"geominQ"`，`"cluster"`が可能）。

alpha　この関数は **psych** パッケージに含まれている。クロンバックの α 係数を計算する。

cortest.bartlett　この関数も **psych** パッケージに含まれている。バートレットの球面性検定を実行する。

fa.parallel　この関数も **psych** パッケージに含まれている。パラレル分析を実行し，返り値として，`$nfact` に推定すべき適切な因子数を返す。

sem　確証的因子分析 (confirmatory factor analysis; CFA) には，**sem** パッケージを用いることができる。もちろん sem は構造方程式モデリングのパッケージであり，CFA 以上のことができる。

8 例 1

Tomothy Bates 教授が提供している SPSS データ[*8]の変数 p1-p40 を分析してみる。Bates 教授は学部学生のための pdf 文書[*9]も提供してくれている。

最も簡単な方法は次のようにする。因子数は自動的に決定される。因子負荷量は res$Factor.Loadings に保存されている。

```
library(foreign)
y <- read.spss("http://www.subjectpool.com/ed_teach/y3method/factorexdata05.sav")
x <- as.data.frame(y)
for (i in 1:length(x)) { x[,i] <- ifelse(x[,i]==999,NA,x[,i]) }
# // Comments // ============================================================
# The data \verb!x! consists of 538 cases with 102 variables.
# it can be saved as "factorexdata05.txt" by the following line
# write.table(x,"factorexdata05.txt",quote=FALSE,sep="\t",row.names=FALSE)
# if so, the data can be read by:
# x <- read.delim("factorexdata05.txt")
# ============================================================================
Ps <- x[,4:43] # Extract variables p1-p40
Ps <- subset(Ps, complete.cases(Ps)) # Omit missings (511 cases remain)
library(rela)
res <- paf(as.matrix(Ps))
summary(res) # Automatically calculate KMO with MSA, determine the number of factors,
             # calculate chi-square of Bartlett's sphericity test, communalities and
             # factor loadings.  Communalities are 1 minus uniquenesses.
barplot(res$Eigenvalues[,1]) # First column of eigenvalues.
resv <- varimax(res$Factor.Loadings) # Varimax rotation is possible later.
print(resv)
barplot(sort(colSums(loadings(resv)^2),decreasing=TRUE)) # screeplot using rotated SS loadings.
scores <- as.matrix(Ps) %*% as.matrix(resv$loadings) # Get factor scores in a simple manner.
library(psych)
cortest.bartlett(Ps) # Bartlett's sphericity test.
res2 <- fa.parallel(Ps)
res3 <- fa(Ps, fm="minres", nfactors=8, rotate="oblimin")
print(res3) # Factor loadings as $loadings
```

9 例 2

石田基広 (2014)『とある弁当屋の統計技師②因子分析大作戦』（共立出版）に掲載されている因子分析は factanal() 関数で因子数 2 の決め打ちなので，他の方法でやってみる。

9.1 因子分析大作戦のパッケージのインストール

サポートサイト[*10]に書かれている通り，

[*8] http://www.subjectpool.com/ed_teach/y3method/factorexdata05.sav

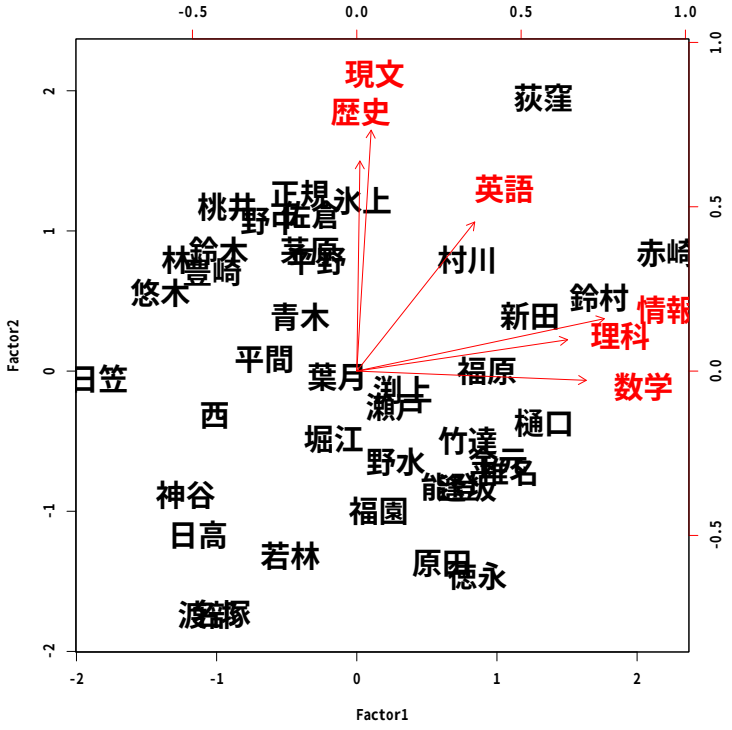
[*9] http://www.subjectpool.com/ed_teach/y3method/factorex05.pdf と http://www.subjectpool.com/ed_teach/y3method/fa.pdf

[*10] http://ishida-m.github.io/misaki/

```
install.packages("Misaki", repos="http://rmecab.jp/R")
library(Misaki)
demo(part2) # 因子分析の部分
```

とする。因子分析を実施しているコードと出力される結果は以下の通り。factanal() はデフォルトでバリマックス回転するので，回転させたくない場合は，rotation="none"というオプションを付ける。

```
> # もし demo(part2) をしないで直接実行する場合は，data(tests) が必要。
> result <- factanal(六教科, factors=2, scores="regression")
> result$loadings

Loadings:
     Factor1 Factor2
英語  0.449   0.567
歴史          0.799
現文          0.916
情報  0.942   0.199
理科  0.802   0.119
数学  0.873

                Factor1 Factor2
SS loadings       2.496   1.855
Proportion Var    0.416   0.309
Cumulative Var    0.416   0.725

> biplot(result$scores, result$loading, cex = 2)
```



biplot() で表示されるグラフは，個人についての因子得点と変数についての因子負荷量が，ともに横軸を第１因子，縦軸を第２因子としてプロットされ（左と下の目盛が因子得点，右と上の目盛が因子負荷量を意味する）わかりやすい。

rela パッケージの paf 関数を使うと，以下のようになる。推定される因子数は 2 つだが，因子負荷量はかなり異なり，因子の解釈もおそらく異なる。サンプルサイズ 40 は「非常に乏しい」が，KMO は 0.7 を超えているので「並」のサンプリング適切性基準はあり，MSA はどの変数についても 0.5 を超えている。

```
> library(rela)
> summary(paf(as.matrix(六教科)))
$KMO
[1] 0.7157

$MSA
         MSA
英語 0.87987
歴史 0.62913
現文 0.58698
情報 0.67863
理科 0.84021
数学 0.70709

$Bartlett
[1] 134.01

$Communalities
     Initial Communalities Final Extraction
英語               0.49330          0.54513
歴史               0.56582          0.67665
現文               0.61918          0.79076
情報               0.79924          0.91818
理科               0.62605          0.67215
数学               0.71366          0.74745

$Factor.Loadings
        [,1]     [,2]
英語 0.68824 -0.26730
歴史 0.43847 -0.69599
現文 0.49364 -0.73964
情報 0.90487  0.31527
理科 0.76242  0.30146
数学 0.72851  0.46553

$RMS
[1] 0.01485
```

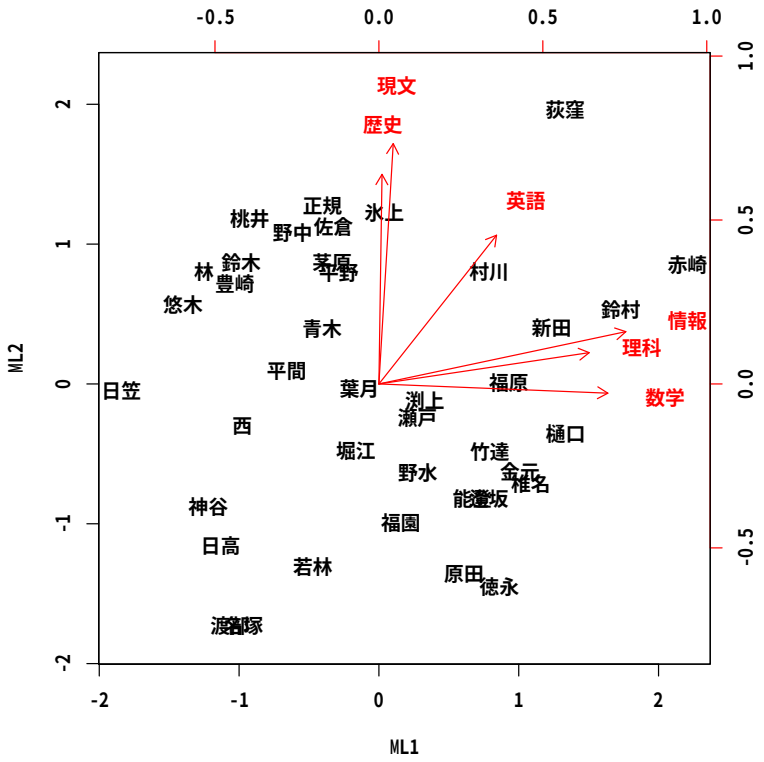
次に，**psych** パッケージを使ってみる。

```
> cortest.bartlett(cor(六教科), n=40)$p.value # バートレットの球面性検定
[1] 3.474469e-21
> # 実は cortest.bartlett(六教科) でも六教科が平方行列ではないので自動的に
> # 相関係数行列を求め，サンプルサイズも実際の値を使って計算してくれる
> res2 <- fa.parallel(六教科)
Loading required package: parallel
Loading required package: MASS
Parallel analysis suggests that the number of factors =  2  and the number of components =  2
```

バートレットの球面性検定の結果の p 値はきわめて小さく，変数間に関連があるといえるので，変数の背後に共通する潜在因子を考えてよい。続けてパラレル分析をした結果，適切な因子数は 2 であることが示唆された。そこで，因子数を 2 と指定して fa() 関数を使って因子分析を実施する。

```
> print(res3 <- fa(六教科, nfactors=2, rotate="varimax", fm = "ml"))
Factor Analysis using method =  ml
Call: fa(r = 六教科, nfactors = 2, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
      ML1   ML2   h2    u2 com
英語 0.45  0.57 0.52 0.477 1.9
歴史 0.01  0.80 0.64 0.362 1.0
現文 0.05  0.92 0.84 0.158 1.0
情報 0.94  0.20 0.93 0.074 1.1
理科 0.80  0.12 0.66 0.343 1.0
数学 0.87 -0.04 0.76 0.237 1.0

                       ML1  ML2
SS loadings           2.50 1.85
Proportion Var        0.42 0.31
Cumulative Var        0.42 0.73
Proportion Explained  0.57 0.43
Cumulative Proportion 0.57 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  15  and the objective function was  3.71 with
 Chi Square of  134.01
The degrees of freedom for the model are 4  and the objective function was  0.06

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.05

The harmonic number of observations is 40 with the empirical chi square 0.44 with prob < 0.98
The total number of observations was 40 with MLE Chi Square = 2.16 with prob < 0.71

Tucker Lewis Index of factoring reliability =  1.061
RMSEA index =  0  and the 90 % confidence intervals are  NA 0.178
BIC =  -12.6
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                               ML1  ML2
Correlation of scores with factors             0.97 0.94
Multiple R square of scores with factors       0.94 0.88
Minimum correlation of possible factor scores  0.88 0.77
> biplot(res3$scores, res3$Structure)
```

バリマックス回転で最尤推定にしたので，結果は factanal() と同様であるが，この関数では 40 というサンプルサイズが十分とは言えないことが示されている。



この結果から，英語は第 1 因子，第 2 因子両方の影響を受け，歴史と現文は第 2 因子のみ，他の 3 教科は第 1 因子のみの影響を受けるという因子構造を想定し，**sem** パッケージを使って[*11]確証的因子分析をするには，次のコードを打つ[*12]。

```
library(Misaki); data(tests) # 既に実行済みなら繰り返す必要はない
library(sem) # sem パッケージのロード
cor1 <- cor(六教科) # 相関係数行列の計算
model1 <- specifyModel() # モデルの指定
英語 <- 文系, a1
歴史 <- 文系, a2
現文 <- 文系, a3
英語 <- 理系, b1
情報 <- 理系, b2
理科 <- 理系, b3
数学 <- 理系, b4
英語 <-> 英語, e1, NA
歴史 <-> 歴史, e2, NA
現文 <-> 現文, e3, NA
情報 <-> 情報, e4, NA
理科 <-> 理科, e5, NA
数学 <-> 数学, e6, NA
文系 <-> 文系, NA, 1
理系 <-> 理系, NA, 1

sem1 <- sem(model1, cor1, N=40) # sem 実行に最低限必要なのはこの３つ。
summary(sem1, fit.indices=c("GFI", "AGFI", "RMSEA", "CFI", "AIC", "BIC"))
```

[*11] `install.packages(sem, dep=TRUE)` により，予め sem パッケージをインストールしておく必要がある。

[*12] `http://minato.sip21c.org/cfa-test.R` にコードを掲載してある。

なお，CFA の場合のモデルの指定は，もっと簡単な方法もある[*13]が，いずれにせよモデル指定の最後に空行が必要である。出力は以下のようになる。AGFI が 0.9 に達しないので十分とは言いにくいが，CFI や RMSEA の値などからすると，そこそこうまく因子構造を説明できるモデルといえる。

```
 Model Chisquare =  8.509691   Df =  8 Pr(>Chisq) = 0.3853279
 Goodness-of-fit index =  0.9381964
 Adjusted goodness-of-fit index =  0.8377655
 RMSEA index =  0.04041812   90% CI: (NA, 0.1949356)
 Bentler CFI =  0.9960645
 AIC =  34.50969
 BIC =  -21.00134

 Normalized Residuals
      Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
-0.1188000  0.0000001  0.1023000  0.3528000  0.6272000  1.5080000

 R-square for Endogenous Variables
  英語   歴史   現文   情報   理科   数学
0.5033 0.7145 0.7516 0.9157 0.6638 0.7240

 Parameter Estimates
   Estimate   Std Error  z value  Pr(>|z|)
a1 0.51486041 0.13399171 3.842480 1.217975e-04 英語 <--- 文系
a2 0.84530493 0.15181150 5.568122 2.574997e-08 歴史 <--- 文系
a3 0.86695423 0.15136758 5.727476 1.019357e-08 現文 <--- 文系
b1 0.44876380 0.12781506 3.511040 4.463569e-04 英語 <--- 理系
b2 0.95691786 0.12322908 7.765358 8.141525e-15 情報 <--- 理系
b3 0.81476524 0.13469909 6.048780 1.459463e-09 理科 <--- 理系
b4 0.85091108 0.13199780 6.446403 1.145356e-10 数学 <--- 理系
e1 0.46041583 0.11888891 3.872656 1.076558e-04 英語 <--> 英語
e2 0.28545947 0.15148233 1.884441 5.950539e-02 歴史 <--> 歴史
e3 0.24839031 0.15469287 1.605700 1.083399e-01 現文 <--> 現文
e4 0.08430837 0.07118676 1.184326 2.362839e-01 情報 <--> 情報
e5 0.33615766 0.09212222 3.649040 2.632223e-04 理科 <--> 理科
e6 0.27595046 0.08360466 3.300659 9.645810e-04 数学 <--> 数学

 Iterations =  22
```

[*13]

```
model1 <- cfa(covs=NULL, reference.indicators=FALSE)
文系: 英語, 歴史, 現文
理系: 英語, 情報, 理科, 数学
```