Discovering Statistics Using R

Chapter17 Exploratory factor analysis

2014.10.08

김가경

factor analysis (요인분석)

정의: 수많은 변수들 중에서 몇 개의 latent variable (잠재된 변수)을 통해 변수의 그룹이나 클러스터를 찾아내는 것

목적: (1) to understand the structure of a set of variables (입력 변수 특성 파악) (ex: latent variable of ‘intelligence’); (2) to construct a questionnaire to measure an underlying variable (본래의 변수보다 더 적절한 새 변수의 생성) (ex: questionnaire to measure burnout); (3) to reduce a data set to a more manageable size while retaining as much of the original information as possible (불필요한 변수 제거) (ex: solution for multicollinearity(다중공선성)).

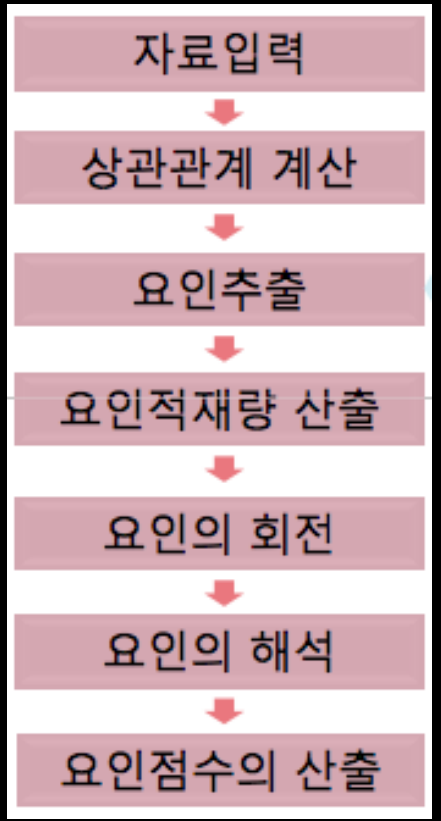
factor analysis (요인분석)

정의: 수많은 변수들 중에서 몇 개의 latent variable (잠재된 변수)을 통해 변수의 그룹이나 클러스터를 찾아내는 것

목적: (1) to understand the structure of a set of variables (입력 변수 특성 파악) (ex: latent variable of ‘intelligence’); (2) to construct a questionnaire to measure an underlying variable (본래의 변수보다 더 적절한 새 변수의 생성) (ex: questionnaire to measure burnout); (3) to reduce a data set to a more manageable size while retaining as much of the original information as possible (불필요한 변수 제거) (ex: solution for multicollinearity(다중공선성)).

factor analysis (요인분석) - 수행과정

1) 모든 변수들에 대한 상관행렬을 구한다.
2) 각각의 요인을 추출한다.
3) 보다 나은 해석을 위해 요인들을 회전한다.
4) 각 응답자에 대한 요인들의 점수를 산출한다.



factor analysis (요인분석) - 종류

1) 탐색적 요인분석(Exploratory Factor Analysis)

- 서로 관계가 알려져 있지 않은 측정변수와 잠재변수 간의 관계를 규명하기 위해 이용
- 측정항목들이 미리 의도한 해당 차원을 제대로 측정하고 있는지에 대해 사전지식을 갖고 있지 않기 때문에 탐색적(Exploratory)이라고 함
- 요인분석을 하기전까지 어떤항목들이 서로 묶이는지 알 수 없기 때문에 요인구조를 탐색하는 목적으로 사용됨
- 분석결과에 따라 일부항목을 제거하거나 추가하게 된다
- 새로운 구성개념의 척도 개발처럼 가설을 세우기에 충분한 증거들이 없을 때 주로 사용. 즉, 선행연구를 통한 이론적 배경이나 논리적 근거가 없기 때문에 이론 생성 과정(theory generating procedure)에 가까움.

factor analysis (요인분석) - 종류

2) 확인적 요인분석(Confirmatory Factor Analysis)

- 이론적 지식 혹은 경험에 근거하여 각 측정 변수와 잠재변수간의 관계를 사전에 가정하고 이 가정을 통계적으로 검증하기 위해 이용된다.
- 요인과 항목들간의 관계가 이미 정해진 상태에서 모델이 구성, 분석 됨.
- 학계에서는 기존 이론모델을 수정 혹은 변경한 모델에 대한 검증에 많이 이용된다.
- 확인적 요인분석은 LISREL이나 AMOS, PLS와 같은 구조방정식 모델을 이용해 분석하는 것이 일반적
- 관측변수(문항)과 잠재변수(요인)의 관계에 초점이 맞추어짐, 구조방정식을 통해서만 분석 가능

factor analysis (요인분석) - 종류 비교

	EFA	CFA
사용 프로그램	SPSS, SAS, R	Amos, LISREL, PLS
분석방법	탐색적, 경험적	확인적, 검증적
이론과정	이론생성	이론검증
선행연구 여부	없음	있음
지향성	데이터 지향적	이론 지향적
요인의 수와 항목	분석 전까지 알 수 없음	분석 전에 지정

그 외 참고 자료 - http://blog.daum.net/dataminer9/66

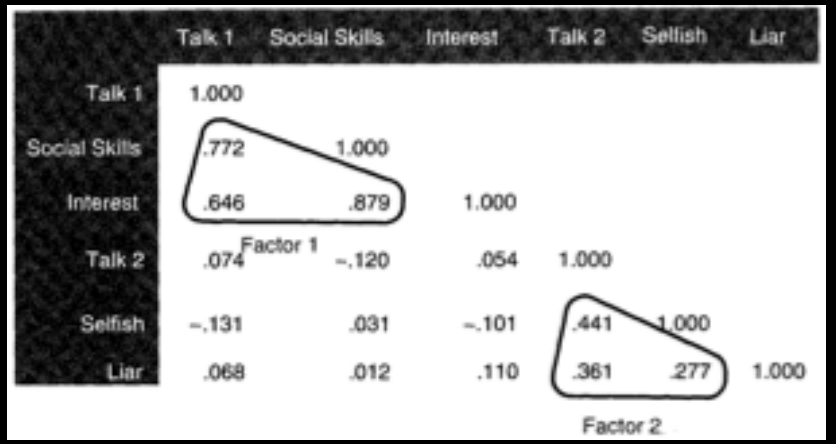
Factors (요인)

요인: 서로 상관계수가 높은 변수들끼리 모아서 작은 수의 변수집단으로 구분한것

R-matrix: a table of correlation coefficients between variables

Reduce this R-matrix down into its underlying dimensions by looking at **which variables seem to cluster together in a meaningful way**

FIGURE 17.2
An *R*-matrix



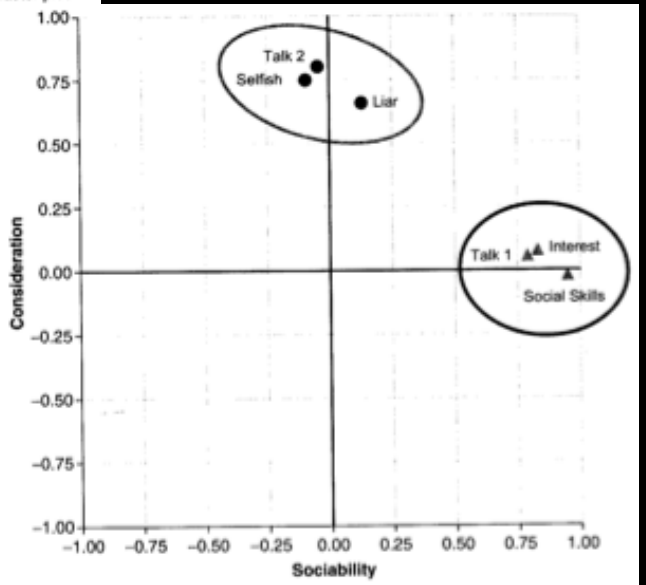
Factor1: sociability
Factor2: consideration

Talk1: proportion of time of talking about the other person during a conversation
Talk2: that about themselves

Graphical representation of factors

Factors: statistical entities can be visualized as classification axes along which measurement variables can be plotted

FIGURE 17.3
Example of a factor plot



Factor loading: the coordinate of a variable along a classification axis. Pearson correlation between a factor(요인) and a variable(변수)

Mathematical of factors

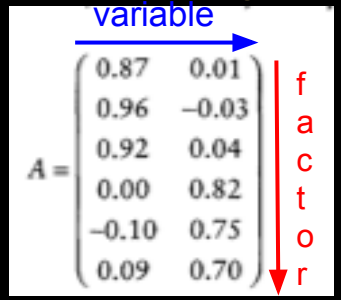
Linear model applies to the scenario of describing a factor

$$(17.3) \quad Y_i = b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + \varepsilon_i$$

$$\text{Sociability}_i = 0.87\text{Talk1}_i + 0.96\text{SocialSkills}_i + 0.92\text{Interest}_i + 0.00\text{Talk2}_i - 0.10\text{Selfish}_i + 0.09\text{Liar}_i + \varepsilon_i$$

$$\text{Consideration}_i = 0.01\text{Talk1}_i - 0.03\text{SocialSkills}_i + 0.04\text{Interest}_i + 0.82\text{Talk2}_i + 0.75\text{Selfish}_i + 0.70\text{Liar}_i + \varepsilon_i$$

Factor loadings can be placed in a factor matrix.



major assumption in factor analysis: algebraic factors represent real-world dimensions.

pattern vs. structure matrix

Both correlation coefficients and regression coefficients represent the relationship between a variable and linear model

1. Orthogonal rotation : situation in underlying factors are assumed to be independent, and the loading is the correlation between the and the variable , but is also the regression coefficient.
2. Oblique rotation : situations in underlying factors are assumed to be related or correlated. 2 loadings: 1)correlation coefficients between each variable and factor (in structure matrix) 2) regression coefficients for each variable on each (in pattern matrix)

Factor scores

factor scores: person's score on a factor, based on their scores for the constituent variables.

weighted average. In fact,this method is overly simplistic and rarely used, but it is probably easiest way to explain the principle

Talk1 (4), Social Skills (9), Interest (8), Talk2 (6), Selfish (8), and Liar (6).

higher on sociability than inconsideration

$$
\begin{aligned}
\text{Sociability} &= 0.87\text{Talk1} + 0.96\text{SocialSkills} + 0.92\text{Interest} \\
&\quad + 0.00\text{Talk2} - 0.10\text{Selfish} + 0.09\text{Liar} \\
&= (0.87 \times 4) + (0.96 \times 9) + (0.92 \times 8) + (0.00 \times 6) \\
&\quad - (0.10 \times 8) + (0.09 \times 6) \\
&= 19.22 \\
\text{Consideration} &= 0.01\text{Talk1} - 0.03\text{SocialSkills} + 0.04\text{Interest} \\
&\quad + 0.82\text{Talk2} + 0.75\text{Selfish} + 0.70\text{Liar} \\
&= (0.01 \times 4) - (0.03 \times 9) + (0.04 \times 8) + (0.82 \times 6) \\
&\quad + (0.75 \times 8) + (0.70 \times 6) \\
&= 15.21
\end{aligned}
\tag{17.4}
$$

단점: The scales of measurement used will influence the resulting scores, and if different variables use different measurement scales, then factor scores for different factors cannot be compared.

The regression method (회귀분석)

Factor loadings are adjusted to take account of the initial correlations between variables

=> differences in units of measurement and variable variances are stabilized

B: matrix of factor score coefficients

R-1: inverse of the original correlation or R-matrix

$$B = R^{-1}A$$

$$B = \begin{pmatrix} 4.76 & -7.46 & 3.91 & -2.35 & 2.42 & -0.49 \\ -7.46 & 18.49 & -12.42 & 5.45 & -5.54 & 1.22 \\ 3.91 & -12.42 & 10.07 & -3.65 & 3.79 & -0.96 \\ -2.35 & 5.45 & -3.65 & 2.97 & -2.16 & 0.02 \\ 2.42 & -5.54 & 3.79 & -2.16 & 2.98 & -0.56 \\ -0.49 & 1.22 & -0.96 & 0.02 & -0.56 & 1.27 \end{pmatrix} \begin{pmatrix} 0.87 & 0.01 \\ 0.96 & -0.03 \\ 0.92 & 0.04 \\ 0.00 & 0.82 \\ -0.10 & 0.75 \\ 0.09 & 0.70 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.343 & 0.006 \\ 0.376 & -0.020 \\ 0.362 & 0.020 \\ 0.000 & 0.473 \\ -0.037 & 0.437 \\ 0.039 & 0.405 \end{pmatrix}$$

$$\begin{aligned} \text{Sociability} &= 0.343\text{Talk1} + 0.376\text{SocialSkills} + 0.362\text{Interest} \\ &\quad + 0.000\text{Talk2} - 0.037\text{Selfish} + 0.039\text{Liar} \\ &= (0.343 \times 4) + (0.376 \times 9) + (0.362 \times 8) + (0.000 \times 6) \\ &\quad - (0.037 \times 8) + (0.039 \times 6) \\ &= 7.59 \\ \text{Consideration} &= 0.006\text{Talk1} - 0.020\text{SocialSkills} + 0.020\text{Interest} \\ &\quad + 0.473\text{Talk2} + 0.437\text{Selfish} + 0.405\text{Liar} \\ &= (0.006 \times 4) - (0.020 \times 9) + (0.020 \times 8) + (0.473 \times 6) \\ &\quad + (0.437 \times 8) + (0.405 \times 6) \\ &= 8.768 \end{aligned} \qquad (17.5)$$

mean = 0, variance = squared multiple correlation between the estimated factor scores and the true values. (**개개의 요인값 과 추정된 요인값 간의 차이를 제곱한 값 이 최소가 되도록 한다.**)

단점: scores can correlate with other factor scores from a different orthogonal factor.

Communality (공통분산)

common variance: share with other variables or measures

unique variance: specific to that measure => error or random variance

communality: common variance in a variable (한 변수의 분산이 추출된 요인들에 의해 설명되는 정도)

(1 = variable without specific variance (or random variance)

0 = share that none of its with any variable)

communality < .50: 그 변수를 무시하고 나머지 변수들을 중심으로 해석. 그 변수를 제거하고 요인분석을 다시 실시.

방법: 1) 모든 variance가 공통 variance라고 가정 (communality=1) => transpose original data into constituent ponents (= principal components analysis)

2) 각 variable의 communality 값을 측정함으로써 common variance 양을 측정 => 방법: squared multiple correlation (SMC) of each variable with all others

Factor analysis (FA) vs. principal component analysis (PCA) vs MANOVA

- Communality estimates
  - FA derives a mathematical model
  - PCA decomposes the original data into a set of linear variates
- Procedures
  - FA only can estimate the underlying factors, and it relies on various assumptions for these estimates to be accurate.
  - PCA is concerned only with establishing which linear components exist within the data and how a particular variable might contribute to that component.
- Theory
  - principal components analysis is a psychometrically sound procedure, is conceptually less complex than factor analysis, and bears numerous similarities to discriminant analysis

(MANOVA와의 비교 내용은 다음 시간에^^)

참고) 요인분석(FA)과 주성분(PCA)분석

*** 공통점**

[1] 모두 데이터를 축소한다. [2] 원래 데이터의 새로운 몇 개의 변수들로 만들어 낸다.

*** 차원을 많이 줄일 수록 좋은 이유**

- 불필요한 속성은 학습기를 오도하거나 무효하게 할 수 있다. 대부분 모델은 낮은 차원에서 유리
- 차원이 높아지면 조율하는데 더 많은 매개변수가 필요하며 과적합화 위험이 있다.
- 문제를 해결하기 위해 찾는 데이터는 인공적인 고차원을 가질 수 있는 반면 실제 차원은 작을 수 있다.
- 차원이 낮아지면 훈련이 빨라지고 시도할 수 있는 다양성이 많아지며 더 나은 결과가 나온다.
- 데이터를 시각화 할때, 2 또는 3차원으로 제한한다. 이를 시각화라고 한다.

Dimensionality reduction

Why do we need dimensionality reduction?

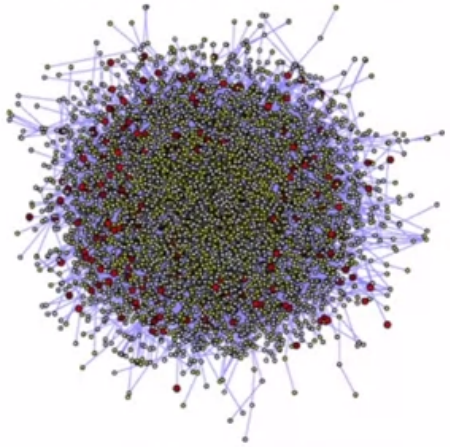
Interesting data often includes many interacting variables

Dimensionality reduction lets you combine information from redundant variables

This makes data easier to visualize

Dimensionality Reduction also helps you avoid....

The human protein interaction network



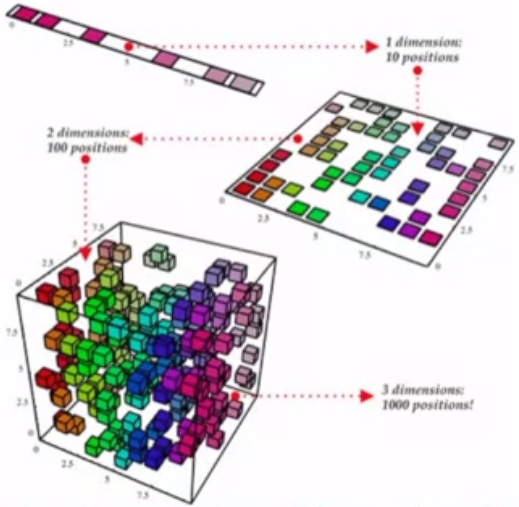
Kim, Philip M., Jan O. Korbel, and Mark B. Gerstein. "Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context." Proceedings of the National Academy of Sciences 104.51 (2007): 20274-20279.

The Curse of Dimensionality

Adding additional dimensions tends to make data sparse

To sample the same portion of the space, you need exponentially more data

Sparse data makes it harder to develop models and validate statistical tests



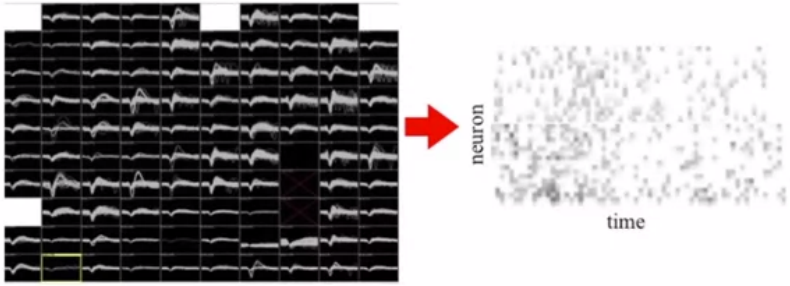
http://www.iro.umontreal.ca/~bengioy/yoshua_en/research.html

2,3차원으로 축소하는 것은 시각화와 해석에 도움이 됨.

Curse of dimensionality를 피하게 함

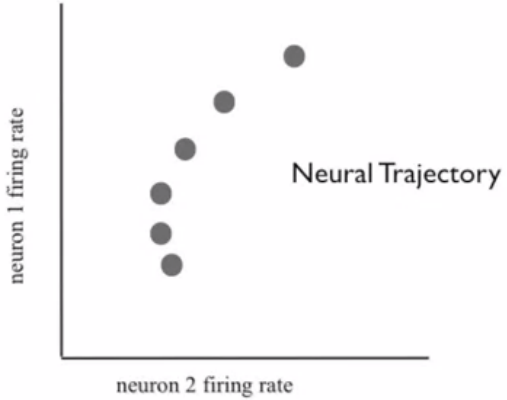
Dimensionality

Neural data can be very high dimensional...



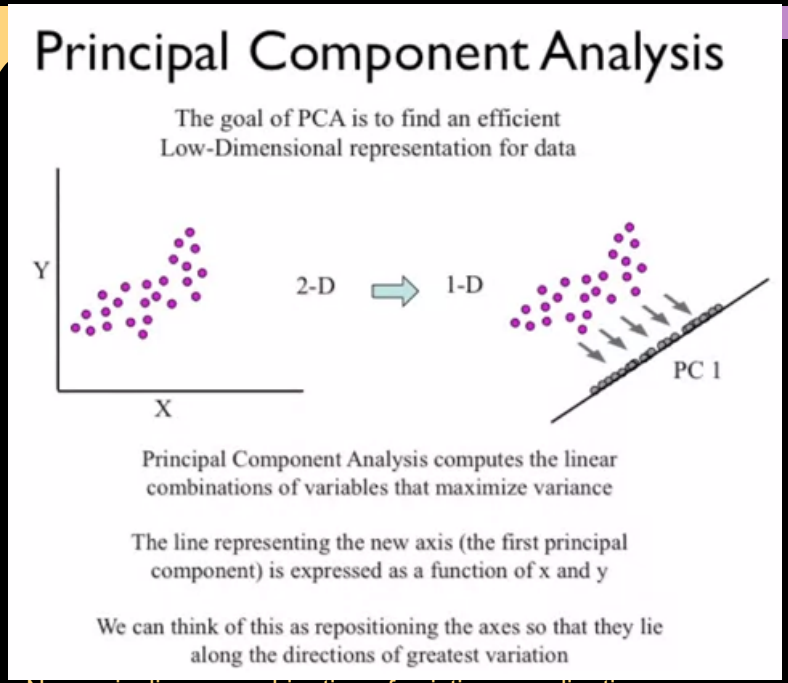
For example, let's take the firing patterns of an ensemble of neurons during a given period of time

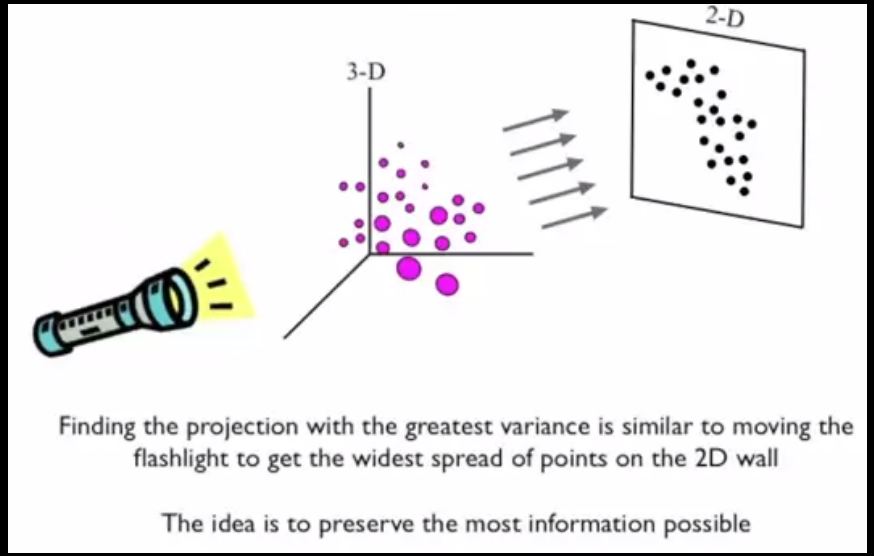
Representing network states



But what happens if we have hundreds of neurons?

Principal Componet Analysis





New axis: linear combination of existing coordination

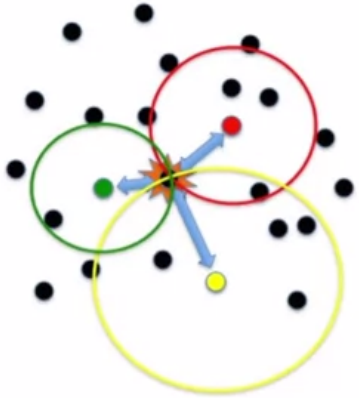
shift axis around to find the direction of maximum variance (change the angle of the flashlight to obtain the maximum spread of points on the wall)

=> 이때 noise 주의, 인접한 지점 간의 관계를 유지하는 것에도 유의

Using dimension reduction and taking the whole ensembl of neurons together => predictable patterns!

Pairwise distance

Defining a space using pair-wise distances



Instead of using cartesian <X,Y>, we can represent each point using <distance to yellow,distance to red, distance to green>, creating a new coordinate axis relative to these three points. In this way, the colored points form the 'basis' for a new space.

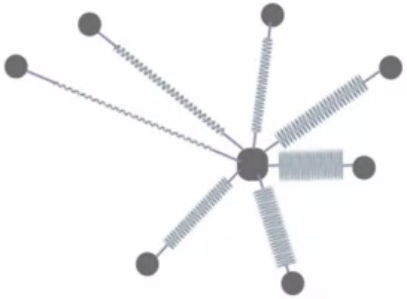
A mass-spring model of nearest neighbor relationships

Each point is 'attached' to other points with springs

Spring stiffness (k) is proportional to the SIMILARITY between the points

Similar points pull each other into tight clusters, while exerting small forces on points that are very different

Eventually the system will settle into an equilibrium state with similar points close together and different points far apart



NOTE: for simplicity, only the springs for ONE point are shown in this diagram!

But how do we measure the 'similarity' between two neural activity patterns?

Low dimensional representation preserving local neighborhoods in the high dimensional space

Pairwise distance => reduce into lower dimensional distance

참고) 요인분석(FA)과 주성분(PCA)분석

* 차이점

[1] 생성되는 변수의 수

FA : 몇 개라고 지정할 수 없다. 데이터의 의미에 따라 다르다. 데이터에 서로 성관성을 갖는 변수들의 군집의 개수로 나뉘어질 것이다. PCA : 보통 2개. 제1주성분, 제2주성분 이라고 불린다.

[2] 생성되는 변수의 의미 (이름)

FA : 분석가가 적절한 이름을 붙인다. PCA : 보통 2개의 변수.

[3] 생성된 변수들의 관계

FA : 새 (잠재)변수들은 기본적으로 대등한 관계. 단, 분류/예측에 그 다음 단계로 사용된 다면 그 때 중요성의 의미가 부여될 것. PCA : 변수들 간의 중요성의 순위가 존재 (제1주성분이 가장 중요).

[4] 분석방법의 의미

FA : 목표 필드를 고려하고 데이터가 주어지면 변수들을 비슷한 성격들로 묶어서 새로운 [잠재] 변수들을 만들어 낸다. PCA : 목표 변수를 고려. 목표 변수를 잘 예측/분류하기 위하여 원래 변수들의 선형 결합으로 이루어진 몇 개의 주성분(변수)들을 찾아냄.

Factor eigen values and the scree plot

eigen values (고유값)

-각각의 요인으로 설명할 수 있는 변수들의 분산 총합으로 각 요인별로 모든 변수의 요인적재값을 제곱하여 더한 값임 얼마나 표현할 수 있는가를 나타내는 비율

-변수 속에 담겨진 정보(분산)가 어떤 요인에 의하여 어느 정도 표현될 수 있는가를 말해주는 비율로, 먼저 추출된 요인이 고유값은 항상 다음에 추출되는 요인의 고유값의 값보다 큼.

최소 고유값 (minimum eigenvalue)

- 가장 많이 사용하는 방법의 하나로 적용하기 매우 간단함
- 주성분 분석법에서는 요인들은 고유값이 '1'보다 적을 경우에 의미 없는 것으로 간주하고 무시
- 공통요인분석법을 선택할 경우 고유치는 약간 하향 조정되어야. 이것만 기준으로 하기엔 위험

스크리검정 (Scree test) (Figure 17.4)

- 요인 수가 증가하면 고유값이 점점 작아지다가 일정 수준에 이르 면 완만
- 고유값이 하락하다가 급격한 하락에서 완만한 하락으로 추세가 바뀌는 지점에서 요인의 수를 결정하는 방식

Factor rotation (1)

- 요인분석의 중요한 개념은 요인의 회전임
- 요인추출 주성분요인 또는 공통요인에 의해 얻어진 최초 요인행렬은 측정변수들의 분산을 어느 정도 설명할 수 있으나, 대부분 각 변 수들과 요인들간의 관계가 명확하게 나타나지 않음
- 회전되지 않는 요인은 단순히 자료를 감축시키는 과정으로 요인들의 중요성에 따라 요인들을 추출하기 때문에 변수의 형태 에 따른 의미있는 정보를 얻기 어려움
- 요인회전의 궁극적인 목적은 요인을 해석하기 쉽고 의미있는 요인패턴을 갖도록 분산을 재분배시키는 과정임
- 요인을 회전시키는 방법에는 직각회전방식(orthogonal)과 비직각 회전방식(oblique)이 있음

Factor rotation (2)

직각 회전

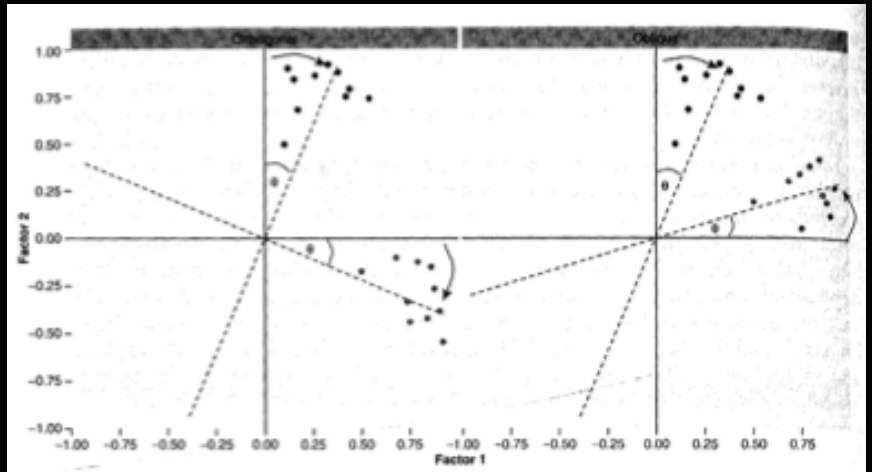
- **요인들간의 상관관계가 없다고 가정**하고 요인을 회전시키는 방법으로 각 요인간의 각도를 **90°로 유지하면서 회전**시킴
- 이 방식은 변수들간의 독립성을 유지시키면서 회전시킴
- 여기에 Varimax 방식과 Quartimax방식이 있음

쿼티멕스회전 (QUARTIMAX)	-요인행렬의 행을 단순화시키는 방식 -한 변수가 어떤 요인에 대해 높은 요인적재량을 가지면 다른 요인에 대해서는 낮은 요인적재량을 갖게 함 -단순한 요인구조를 얻는 데는 문제가 있는 반면, **많은 변수에 대해 문항간 높은 적재량을 갖는 변수들의 일반적 요인**을 만들어 낼 수 있음
베리멕스회전 (VARIMAX)	-**요인행렬의 열을 단순화시키는 방식**으로 대부분 이 방법을 사용함 -요인행렬의 각 열에 1 또는 0에 가까운 요인적재량을 보임 -변수와 요인간의 관계가 명확해지고 해석하기에 용이하기 때문에 **단순한 요인구조를 산출**할 때 사용함
이쿼멕스회전 (EQUIMAX)	-쿼티멕스 회전과 베리멕스 회전을 절충한 방법 -행과 열을 동시에 간략히 하려는 방법임 -널리 인정받지 못하고 있고, 따라서 자주 사용되지 않음

비직각 회전

- 이 방식은 요인들간의 **상관관계가 있을 경우**에 사용함
- 요인간의 각도를 **90°이외의 사각(사선)을 유지하면서 변수를 회전**시키는 방법임
- 요인간의 상관관계를 인정하기 때문에 다소 설득력이 떨어지지만 경험적인 근거를 가지고 요인구조를 만들어 낼 수 있기 때문에 사회현상 분석에 많이 사용할 수 있음

오블리민회전 (OBLIMAX)	-분석자가 단순히 이론적으로 더 의미있는 구조, 차원을 얻는 데 관심이 있다면 사각회전을 많이 사용함



Research example

FIGURE 17.6 The **R** anxiety questionnaire (RAQ)

SD = Strongly Disagree, D = Disagree, N = Neither, A = Agree, SA = Strongly Agree

		SD	D	N	A	SA
1	Statistics make me cry	O	O	O	O	O
2	My friends will think I'm stupid for not being able to cope with R	O	O	O	O	O
3	Standard deviations excite me	O	O	O	O	O
4	I dream that Pearson is attacking me with correlation coefficients	O	O	O	O	O
5	I don't understand statistics	O	O	O	O	O
6	I have little experience of computers	O	O	O	O	O
7	All computers hate me	O	O	O	O	O
8	I have never been good at mathematics	O	O	O	O	O
9	My friends are better at statistics than me	O	O	O	O	O
10	Computers are useful only for playing games	O	O	O	O	O
11	I did badly at mathematics at school	O	O	O	O	O
12	People try to tell you that R makes statistics easier to understand but it doesn't	O	O	O	O	O
13	I worry that I will cause irreparable damage because of my incompetence with computers	O	O	O	O	O
14	Computers have minds of their own and deliberately go wrong whenever I use them	O	O	O	O	O
15	Computers are out to get me	O	O	O	O	O
16	I weep openly at the mention of central tendency	O	O	O	O	O
17	I slip into a coma whenever I see an equation	O	O	O	O	O
18	R always crashes when I try to use it	O	O	O	O	O
19	Everybody looks at me when I use R	O	O	O	O	O
20	I can't sleep for thoughts of eigenvectors	O	O	O	O	O
21	I wake up under my duvet thinking that I am trapped under a normal distribution	O	O	O	O	O
22	My friends are better at R than I am	O	O	O	O	O
23	If I am good at statistics people will think I am a nerd	O	O	O	O	O

Sample size

- Reliability of factor analysis => dependent on sample size.
- Correlation coefficients fluctuate from sample to sample, much more so in small samples than in large - at least 10-15 participants per variable.
- Simulated data (Monte Carlo studies)
  - Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1970): ratio of the squared correlation between variables to the squared partial correlation between variables

    0: sum of partial correlations is large relative to the sum of correlations, 1: patterns of correlations are relatively compact and so tor analysis should yield distinct and reliable factors.

    >.5 as barely acceptable, .5< values < .7 : mediocre

    .7 < values < .8 : good, .8 < values < .9 : great, values > .9 : superb

    sample >= 300 provide a stable factor solution

    - Other measures by Guadagnoli and Velicer (1988), MacCallum, Widaman, Zhang, and Hong (1999)

Correlation between variables

correlations of the variables의 문제 => 문제되는 변수를 제거해야 함! (1) correlations that are not high enough (2) correlations that are too high. cor() function (see Chapter 6) to create a correlation matrix of all variables. 을 통해 확인 correlations < .3: 인지 확인 필요

Correlation between variables

Bartlett's test to examines whether the population correlation matrix resembles an identity matrix. (**요인들의 제곱한 값의 합이 최소가 되도 록 한다.)**

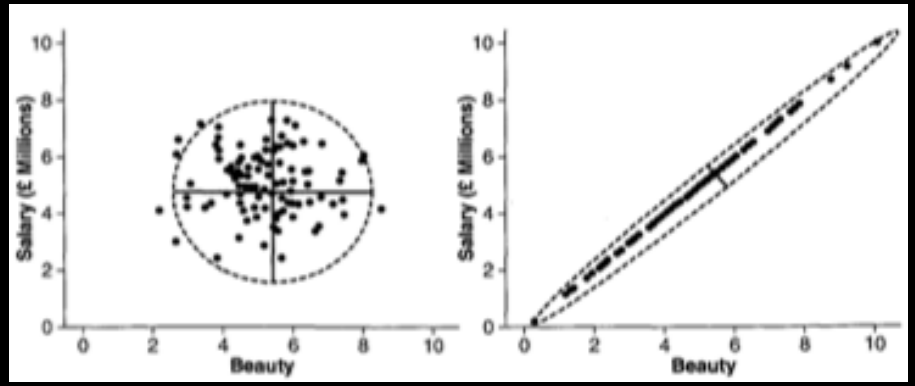
=> Every variable correlates very badly with all other variables (all correlation coefficients =~ 0)

Multicollinearity causes problems in factor analysis because it becomes impossible to determine the unique contribution to a factor of the variables that are highly correlated.

=> Eliminate highly correlating variables (Deteminant of R-matrix > 0.00001 여야 함!)

Determinant & distribution of data

determinant : the ‘area’ of the data => **low correlation so the determinant (area) is 0; the biggest value it can be is 1.**



assumption of normality is most important if you wish to generalize the results of your analysis beyond the sample collected

dichotomous variables you should construct the correlation matrix from polychoric correlation coefficients => polychor() 함수로 계산

실습

FIGURE 17.6 The R anxiety questionnaire (RAQ)

SD = Strongly Disagree, D = Disagree, N = Neither, A = Agree, SA = Strongly Agree

		SD	D	N	A	SA
1	Statistics make me cry	O	O	O	O	O
2	My friends will think I'm stupid for not being able to cope with R	O	O	O	O	O
3	Standard deviations excite me	O	O	O	O	O
4	I dream that Pearson is attacking me with correlation coefficients	O	O	O	O	O
5	I don't understand statistics	O	O	O	O	O
6	I have little experience of computers	O	O	O	O	O
7	All computers hate me	O	O	O	O	O
8	I have never been good at mathematics	O	O	O	O	O
9	My friends are better at statistics than me	O	O	O	O	O
10	Computers are useful only for playing games	O	O	O	O	O
11	I did badly at mathematics at school	O	O	O	O	O
12	People try to tell you that R makes statistics easier to understand but it doesn't	O	O	O	O	O
13	I worry that I will cause irreparable damage because of my incompetence with computers	O	O	O	O	O
14	Computers have minds of their own and deliberately go wrong whenever I use them	O	O	O	O	O
15	Computers are out to get me	O	O	O	O	O
16	I weep openly at the mention of central tendency	O	O	O	O	O
17	I slip into a coma whenever I see an equation	O	O	O	O	O
18	R always crashes when I try to use it	O	O	O	O	O
19	Everybody looks at me when I use R	O	O	O	O	O
20	I can't sleep for thoughts of eigenvectors	O	O	O	O	O
21	I wake up under my duvet thinking that I am trapped under a normal distribution	O	O	O	O	O
22	My friends are better at R than I am	O	O	O	O	O
23	If I am good at statistics people will think I am a nerd	O	O	O	O	O

r code 참조

Table 17.1 Summary of exploratory factor analysis results for the **R** anxiety questionnaire ($N = 2571$)

	Varimax rotated factor loadings			
Item	*Fear of computers*	*Fear of statistics*	*Peer evaluation*	*Fear of maths*
I have little experience of computers	**.80**	−.01	−.07	.10
R always crashes when I try to use it	**.68**	.33	−.08	.13
I worry that I will cause irreparable damage because of my incompetence with computers	**.65**	.23	−.10	.23
All computers hate me	**.64**	.33	−.08	.16
Computers have minds of their own and deliberately go wrong whenever I use them	**.58**	.36	−.07	.14
Computers are useful only for playing games	**.55**	.00	−.12	.13
Computers are out to get me	**.46**	.22	−.19	.29
I can't sleep for thoughts of eigen vectors	−.04	**.68**	−.14	.08
I wake up under my duvet thinking that I am trapped under a normal distribution	.29	**.66**	−.07	.16
Standard deviations excite me	−.20	**−.57**	.37	−.18
People try to tell you that R makes statistics easier to understand but it doesn't	**.47**	**.52**	−.08	.10
I dream that Pearson is attacking me with correlation coefficients	.32	**.52**	.04	.31
I weep openly at the mention of central tendency	.33	**.51**	−.12	.31
Statistics makes me cry	.24	**.50**	.06	.36
I don't understand statistics	.32	**.43**	.02	.24
I have never been good at mathematics	.13	.17	.01	**.83**
I slip into a coma whenever I see an equation	.27	.22	−.04	**.75**
I did badly at mathematics at school	.26	.21	−.14	**.75**
My friends are better at statistics than me	−.09	−.20	**.65**	.12
My friends are better at R than I am	−.19	.03	**.65**	−.10
If I'm good at statistics my friends will think I'm a nerd	−.02	.17	**.59**	−.20
My friends will think I'm stupid for not being able to cope with R	−.01	−.34	**.54**	.07
Everybody looks at me when I use R	−.15	−.37	**.43**	−.03
Eigenvalues	3.73	3.34	1.95	2.55
% of variance	16.22	14.52	8.48	11.10
α	.82	.82	.57	.82

Note: Factor loadings over .40 appear in bold.

A principal components analysis (PCA) was conducted on the 23 items with orthog- onal rotation (varimax). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis KMO = .93 (‘superb’ according to Kaiser, 1974), and all KMO values for individual items were > .77, which is well above the acceptable limit of .5. Bartlett’s test of sphericity, X2 (253) = 19,334, *P* < .001, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Four components had eigenvalues over Kaiser’s criterion of 1 and in combination explained 50.32% of the variance. The scree plot was slightly ambiguous and showed inflexions that would justify retaining both. two and four components. Given the large sample size, and the convergence of the scree plot and Kaiser’s criterion on four components, four components were retained in the final analysis. Table 17.1 shows the factor loadings after rotation. The items that cluster on the same components suggest that component 1 represents a fear of computers, component 2 a fear of statistics, component 3 a fear of maths and component 4 peer evaluation concerns.

Reliability analysis

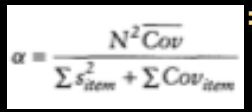
Measure (or questionnaire) should consistently reflect the its construct.

Test-retest reliability: 다른 시간에 조사한 설문지에 같은 점수를 가진 것

People who are the same in terms of the construct being measured should get the same score. => 예)두 사람이 통계 공포증이 있다면, RAQ(R anxiety questionnaire) 점수가 유사할 것임

Individual items (=sets of items) should produce results consistent with the overall questionnaire. => 예)통계 공포를 가진 사람이라면 RAQ 점수가 높을 것임

Split-half reliability: 무작위로 데이터를 두 개로 나눈 후 각각의 반에 대해서 참여자의 점수를 측정. 이 수치가 유사하거나 같다면 신뢰도가 있는 것.

Cronbach's alpha  => 항목간의 variance(대각 행렬요소), 항목과 다른 항목과 covariance(오프 대각 행렬요소)

=> 분자: average of the off- diagonal elements, 분모: sum of the item variances and

Reliability analysis - Tips

- Reliability is really the consistency of a measure.
- Reliability analysis can be used to measure the consistency a questionnaire.
- Remember to deal with ***reverse-scored items***. Use the ***keys option*** when you run the analysis.
- Run separate reliability analyses for all subscales your questionnaire.
- **Cronbach's alpha** indicates the overall reliability of a questionnaire and values around .8 are good (or .7 for ability tests and such like).
- The raw alpha when an item is dropped tells you whether removing an item will improve the overall reliability: values greater than the overall reliability indicate that removing that item will improve the overall reliability of the scale. Look for items that dramatically increase the value alpha
- If you do remove items, rerun your factor analysis to check that the factor structure still holds!