



통계 알고리즘 9

2019년 5월 정화민 교수



차원 축소 기법

- 차원 축소 기법:

대량의 빅데이터를 분석하는 데 있어, 분석대상이 되는 여러 변수들의 주요 정보는 최대한 유지하면서 데이터 세트의 변수의 개수를 줄이는 일련의 탐색적 분석 기법을 말한다.

- 차원 축소 기법의 개념 및 목적

차원 축소를 수행할 때, 축약되는 변수 세트는 원래의 전체 데이터의 변수들의 정보를 최대한 유지해야 한다는 점이 중요하다. 즉, 변수들 간에 내재한 특성이나 관계를 분석하여 이들을 잘 표현할 수 있는 새로운 선형 혹은 비선형 결합을 만들어내서 해당 결합변수만으로도 전체 변수를 적절히 설명할 수 있어야 한다는 것이다. 이러한 차원 축소 기법은 하나의 완결된 분석 기법으로 사용되기보다는 다른 분석과정을 위한 전 단계나 분석 수행 후 결과를 개선하기 위한 방법 혹은 효과적인 시각화 등의 목적으로 사용된다.

- 차원 축소 기법 주요 활용 분야

- 탐색적 데이터

- (1) 탐색적 데이터 분석
- (2) 변수 세트에서 주요 특징 추출 후 타 분석 기법의 설명변수로 투입 (주성분 회귀분석 등)
- (3) 텍스트 등 문서에서의 숨겨진 주제나 개념 추출
- (4) 이미지 및 사운드 등에서 주요한 데이터 특징 패턴 추출
- (5) 고객의 구매 및 거래 데이터의 아이템 축약 통한 추천 시스템 엔진 구현
- (6) 다차원 공간의 정보를 저차원 정보로 시각화
- (7) 공통요인을 추출하여 잠재된 데이터 구조를 발견

주요 차원축소 기법

기 법	기법 설명
주성분 분석 (PCA)	변수들의 공분산 행렬이나 상관행렬을 이용하여 원래의 데이터 세트의 변수들을 선형 변환하여 서로 직교하도록 선택된 새로운 변수들(주성분)을 만들어낸 뒤, 이를 통해 원래 변수를 설명하고자 하는 차원축소 기법
특이값 분해 (SVD)	주성분 분석이 행의 수와 열의 수가 같은 정방행렬에서만 사용될 수 있는 것에 비하여, 특이값 분해는 일반적인 $m \times n$ 차원의 행렬데이터에서 특이값을 추출하고 이를 통해 주어진 데이터 세트를 효과적으로 축약할 수 있는 기법
요인분석 (Factor Analysis)	데이터 안에 관찰할 수 없는 잠재적인 변수(Latent Variable)가 존재한다고 가정하고, 모형을 세운 뒤 관찰 가능한 데이터를 이용하여 해당 잠재 요인을 도출하여 데이터 안에 내재된 구조를 해석하는 기법. 주로 사회과학이나 설문조사 등에서 많이 활용된다.
독립성분분석 (ICA)	주성분 분석과는 달리, 변수들이 서로 독립적이라고 가정하며, 독립성분의 분포는 비정규분포를 따르게 되는 차원축소 기법
다차원 척도법 (MDS)	각 개체 간의 관계정보를 이용하여, 고차원의 데이터를 해석이 용이하도록 가시적인 저차원(주로 2차원)으로 사상한 뒤 각 좌표점으로부터 그룹 관계, 순서관계, 위상관계 등을 파악하는 다변량 시각화 기법

Source: NCS

주성분 분석

- 주성분 분석(PCA):

원래의 변수들을 선형 변환하여 주성분이라 불리는 서로 직교하도록 선택된 새로운 변수들을 만들어낸 뒤, 원래의 데이터를 적은 수의 대표적인 축약변수로 요약하는 차원축소 기법이다.

각 주성분의 분산의 크기는 제1 주성분에서 p번째 주성분까지 순차적으로 감소하게 된다. 따라서 차원축소를 위해서는 적절한 주성분 개수를 선택해야 한다. 만일 3개의 주성분 (q=3)을 선택했다면, 3개의 주성분들의 누적분산이 원래의 p개 변수들의 누적분산에 대한 비율만큼 데이터의 변동을 잘 설명하는 것이 되며, 이 비율이 커질수록 적은 수의 주성분만으로 원래의 변수들을 잘 설명할 수 있다는 것이 된다. Source: NCS

$$Z_1 = \gamma_{11}X_1 + \gamma_{12}X_2 + \dots + \gamma_{1p}X_p = \gamma_1^T X$$

$$Z_2 = \gamma_{21}X_1 + \gamma_{22}X_2 + \dots + \gamma_{2p}X_p = \gamma_2^T X$$

$$\dots \dots \dots$$

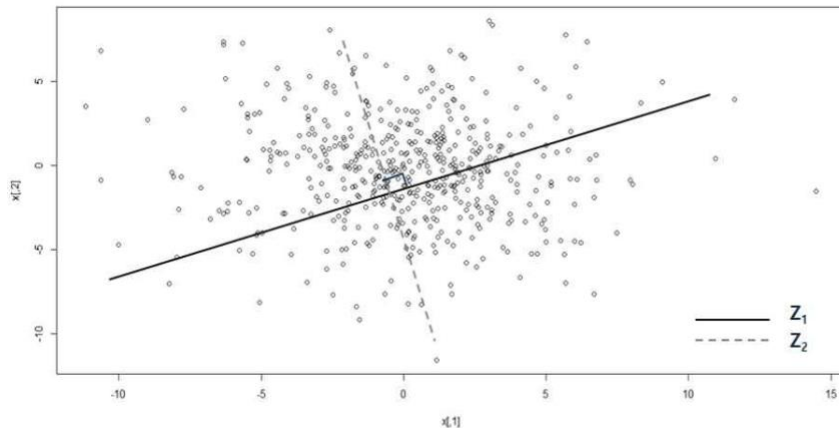
$$Z_q = \gamma_{q1}X_1 + \gamma_{q2}X_2 + \dots + \gamma_{qp}X_p = \gamma_q^T X$$

$$\text{maximize}_{\gamma_{11}, \dots, \gamma_{1p}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \gamma_{1j} x_{ij} \right)^2 \right\}$$

$$\text{sbj } \sum_{j=1}^p \gamma_{1j}^2 = 1$$

주성분 분석

X_1, X_2, \dots, X_p 는 평균이 0이 되게 중심화하였다고 가정하며, Z_1, Z_2, \dots, Z_q 는 X_1, X_2, \dots, X_p 의 선형결합으로 만들어진 새로운 주성분 변수이고, k 번째 주성분 변수의 분산은 $\text{Var}(Z_k) = \gamma_k^T \Sigma_p \gamma_k$, k 번째 주성분과 또 다른 l 번째 주성분 간의 공분산은 $\text{Cov}(Z_k, Z_l) = \gamma_k^T \Sigma_p \gamma_l$,이며, $\gamma_k^T \gamma_k = \sum_{p,j=1} \gamma_{kj}^2 = 1$ 를 만족하는 단위벡터로 적재계수를 나 타낸다. 제 1주성분 $Z_1 = \gamma_1^T X$ 은 $\text{Var}(\gamma_1^T X)$ 를 최대화하는 선형조합이며, 제 2주성분 $Z_2 = \gamma_2^T X$ 은 $\text{Cov}(\gamma_1^T X, \gamma_2^T X) = \gamma_1^T \Sigma_p \gamma_2 = 0$ 이면서, 동시에 $\text{Var}(\gamma_2^T X)$ 를 최대화하는 두 번째 선형조합이다.



주성분 분석의 시각화

굵은 실선으로 표시된 직선이 제 1주성분에 해당하는 Z_1 그래프이며, 점선으로 표시된 직선이 제 2주성분에 해당하는 Z_2 그래프이다. 제 1주성분 Z_1 은 X_1, X_2 분포를 가장 많이 설명하는 방향으로 축이 설정되어 있으며, 제 2주성분 Z_2 는 제 1주성분과 직교를 이루면서 제 1주성분 다음으로 변수의 분포를 설명하는 방향으로 축이 형성되어 있다

주성분 분석 참조논문

〈Table 3〉 Factor analysis & Reliability of variables					
Variables	Items of variables	Factor loadings	Eigen values	% of Variance	Cronbach Alpha
개방성	개방성3	.801	2.327	16.619	.740
	개방성4	.760			
	개방성5	.719			
	개방성2	.713			
모험심	모험심2	.793	2.097	14.979	.739
	모험심1	.755			
	모험심4	.743			
독창성	독창성5	.841	2.320	16.569	.836
	독창성1	.793			
	독창성2	.686			
융통성	융통성2	.800	2.646	18.903	.812
	융통성4	.731			
	융통성1	.725			
	융통성5	.691			
조직몰입	조직몰입11	.817	4.910	27.277	.881
	조직몰입9	.794			
	조직몰입8	.793			
	조직몰입12	.759			
	조직몰입10	.710			
	조직몰입13	.674			
	조직몰입7	.659			
	조직몰입6	.620			
	조직몰입2	.534			
	조직몰입5	.527			
인지된 성과	인지성과1	.863	2.579	14.329	.819
	인지성과3	.818			
	인지성과2	.716			
	인지성과4	.694			
직무만족	직무만족2	.860	2.563	14.241	.718
	직무만족1	.788			
	직무만족4	.753			
	직무만족5	.556			

Note: 요인추출 방법: 주성분 분석, 회전 방법: 베리맥스(Varimax)

The e-Business Studies

Volume 17, Number 4, August, 30, 2016 : 285-301

Received: 20160807 / Accepted: 20160828
 Revised: 20160826 / Published: 20160828

The Study on R&D Employee's Creativity and Performance in ICT Companies Using Cluster Analysis

/ Hwa Min Jeong**

[ABSTRACT]
 The objective of this study is to analyze the relationship between employee's creativity and perceived performance of project in R&D department of ICT companies. Creativity was measured by creative personality (openness, adventurousness), creative ability (originality, flexibility) and organizational commitment. We conducted cluster analysis of R&D workers' creativity. As the result of this research, five groups emerged: the high balanced creativity-high performance group (group 1), the low balanced creativity-low performance group (group 5), and the imbalance creativity-intermediate performance group (group 2, 3, 4). It means that well-balanced development of creativity has positive effects on R&D performance. After investigating the differences in age, education level and work experience emerged between the groups, this research is expected to be highly suggestive for human resources management planning of the R&D department.

[CONTENTS]
 ABSTRACT
 I. Introduction
 II. Literature review
 III. Research model and methodology
 IV. Results of analysis
 V. Conclusion and implications
 Reference
 국문초록

[Key Words]
 Cluster Analysis, ICT, R&D, Creativity, Performance

I. Introduction

오늘날 기업은 극한 경쟁의 상황에서 자신이 가지고 있는 자원을 활용하여 어떻게 글로벌 시장에서의 경쟁우위를 창출하는가를 매우 중요한 이슈로 간주하고 있다. 글로벌화와 기술의 발달 및 변화가 이러한 극한 경쟁을 유발하는 주요 원인으로 파악되고 있기 때문에(Hitt, Ireland, and Hoskisson, 2007) 기업의 생존 및 발전을 위해서는 국제화와 환경의 변화에 적합한 기술의 개발이 필수적이라고 할 수 있다. 이 중 경쟁우위의 기반인 기술을 확보하기 위해서는 기업내부에 강력한 R&D 조직이 있어야 하며 이를 뒷받침 해줄 자본과 조직시스템이 갖춰져야 한다. 그러나 글로벌화 된 기업들의 경우에도 각 기업의 역량에 따라 연구개발의 성과가 다르게 나타나며 동일 기업이라 할지라도 프로젝트의 성과와 예측치 사이에 차이가 발생하기도 한다. 그동안 많은 선행 연구들이 프로젝트의 성과를 좌우하는 영향요인으로서 개인의 특성을 연구해 왔다. 특히 끊임없이 회사의 제품이나 서비스를 발전시킬 새로운 지식, 기술을 개발하고 프로세스를 향상 하는 ICT 기업의 R&D 연구원들에게 요구되는 개인의 특성은 창의성이라고 할 수 있다. 기업이 혁신할 수 있는 기본 원료를 얼마나 가지고 있는가는 곧 기업 구성원이 가지고 있는 창의적 자원에 따라 결정되기 때문이다(Amabile, 1983; Davis, 1986; Grossman, 1982; Saw, 1990). 따라서 본 연구는 개인의 R&D 성과에 영향을 미치는 요인들을 분석하기 위해 개인 창의적 특성에 대해 보다 구체적으로 접근할 필요가 있으며, 국내 R&D 연구원들의 창의적 특성과 성과


ISSN 1229-9936 (Print), ISSN 2466-1716 (Online)

Hyeon Um / Hwa Min Jeong 285

주성분분석(Principle component analysis)과 요인적재치의 단순화를 위하여 직교회전방식인 베리맥스(varimax)를 채택하였으며, 고유값(eigen value)은 1.0 이상을 기준으로 요인을 분석하였다.

R에서 주성분 분석

- R의 datasets 패키지에서 기본으로 제공되는 USArrests 데이터
주성분 분석 알파벳순으로 미국의 50개 주를 포함하며, 각 주의 100,000명의 인구당 체포된 세 가지 강력 범죄수(murder, assault, rape)와 각 주마다 도시에 거주하는 인구의 비율(UrbanPop)%로 구성.
- 주성분 분석을 실행하는 함수로는 대표적으로 princomp와 prcomp가 있다.



```
> FA<-princomp(USArrests, cor=T)
> summary(FA)
Importance of components:

              Comp.1      Comp.2      Comp.3      Comp.4
Standard deviation  1.5748783  0.9948694  0.5971291  0.41644938
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion 0.6200604 0.8675017 0.9566425 1.00000000
>
```

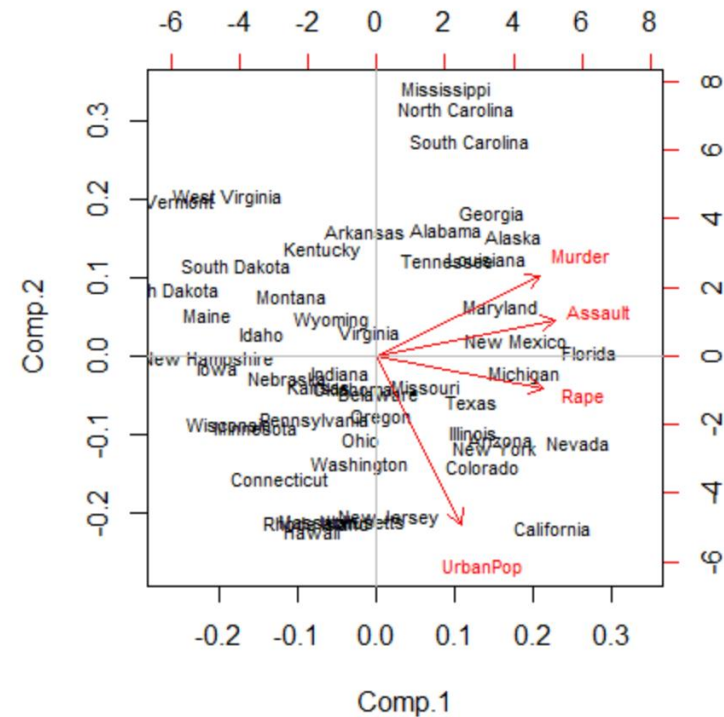
- 위의 R 프로그래밍 결과값을 보면 주성분의 표준편차를 볼 수 있고 이 값들을 제공하면 각 주성분의 분산이 된다.
- 각 표준편차 크기 순서대로 Z1, Z2, Z3, Z4 주성분 변수가 만들어진 것을 볼 수 있으며, 이 중 2개의 주성분 Z1, Z2의 **누적설명비율이 약 86.8%**으로 대부분 설명하고 있어 **2개의 요인으로 축소할 수 있다.**

R 에서 주성분 분석

```
> FA$center
Murder  Assault  UrbanPop   Rape
7.788  170.760   65.540   21.232
> FA$scale
Murder  Assault  UrbanPop   Rape
4.311735 82.500075 14.329285  9.272248
> FA$loadings
```

```
Loadings:
      Comp.1  Comp.2  Comp.3  Comp.4
Murder    0.536   0.418   0.341   0.649
Assault    0.583   0.188   0.268  -0.743
UrbanPop    0.278  -0.873   0.378   0.134
Rape        0.543  -0.167  -0.818
```

```
> plot(FA$scores[,1], FA$scores[,2], xlab="z1", ylab="z2")
> biplot(pc1, cex=0.7)
> abline(v=0, h=0, col="gray")
```



제1주성분(가로축 방향)에는 Assault, Rape, Murder가 많은 영향을 미치고, 제2주성분(세로축 방향)에는 UrbanPop과 Rape가 많은 영향을 미치고 있다. 또한, 각 데이터 좌표들과 해당 변수 벡터에 의하여 뉴멕시코, 메릴랜드 등은 살인과 폭행의 비율이 높고, 미시간, 텍사스 등은 강간의 비율이 높은 지역임을 확인할 수 있음.