REPORT

수강과목	:	다변량통계학(I)
담당교수	:	최용석
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2020.06.09.

HW3 for Multivariate Statistics I

June 2, 2020

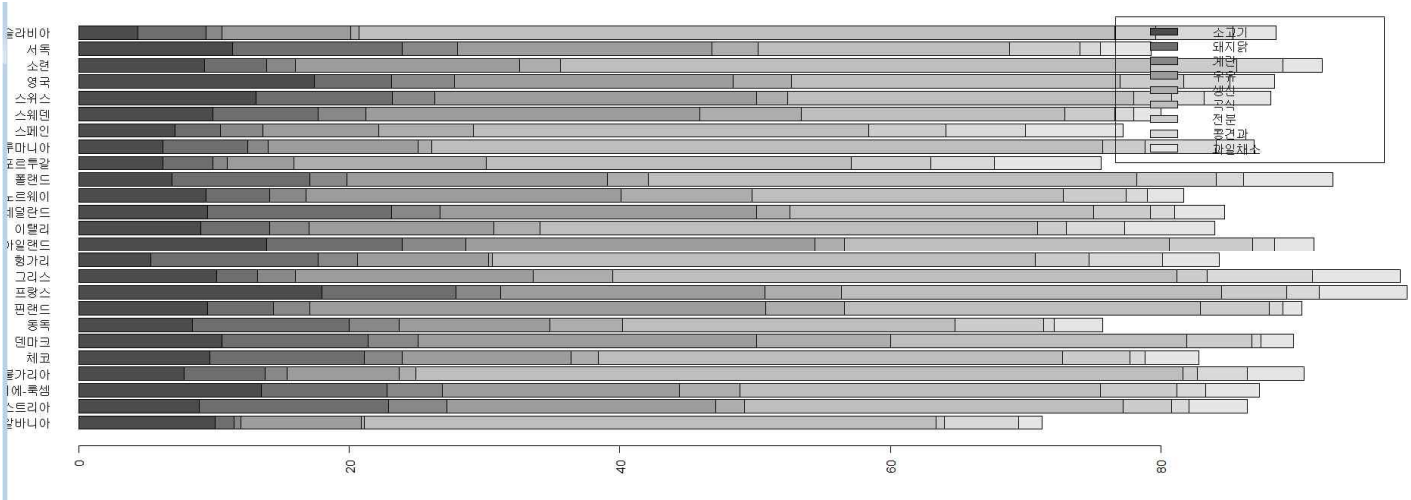
Chapter 3. Factor Analysis

Consider the Exercise 3.9. in page 191.

Solve the problems (1) ~ (5), in Exercise 3.9.

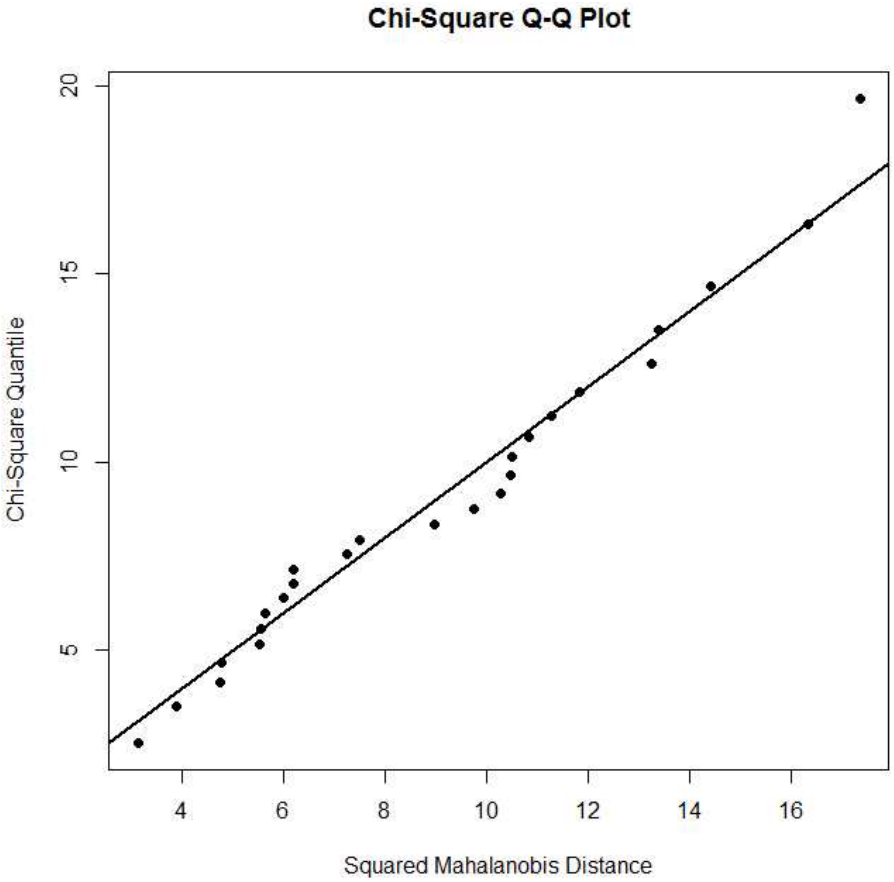
3.9 [자료 1.3.3] (protein.txt)은 유럽 25개국의 9가지 단백질 섭취원에 대한 평균섭취량 자료가 있다.

Barplot



Starplot

위의 그래프를 보면 콩견과, 곡식은 주로 동유럽국가(루마니아, 유고슬라비아, 불가리아, 알바니다 등)에서 섭취량이 높고 소고기, 돼지닭, 계란, 우유, 전분은 주로 서북유럽국가(아일랜드, 핀랜드, 스웨덴, 영국, 덴마크 등)에서 섭취량이 높다. 생선과 과일채소는 주로 남유럽(포르투갈, 스페인 등)에서 섭취량이 높다.




```
> mvn(X, mvnTest = "mardia", multivariatePlot =  "qq")
$multivariateNormality
            Test          Statistic           p value Result
1 Mardia Skewness   168.086605971262 0.418584258072361    YES
2 Mardia Kurtosis -0.523571666842167 0.600576492382073    YES
3             MVN               <NA>              <NA>    YES

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk  소고기      0.9302    0.0879    YES
2 Shapiro-Wilk  돼지닭      0.9410    0.1565    YES
3 Shapiro-Wilk   계란       0.9588    0.3915    YES
4 Shapiro-Wilk   우유       0.9609    0.4323    YES
5 Shapiro-Wilk   생선       0.9012    0.0195    NO
6 Shapiro-Wilk   곡식       0.8927    0.0128    NO
7 Shapiro-Wilk   전분       0.9450    0.1930    YES
8 Shapiro-Wilk  콩견과      0.9026    0.0209    NO
9 Shapiro-Wilk 과일채소     0.9269    0.0738    YES

$Descriptives
          n   Mean   Std.Dev Median  Min  Max 25th 75th        Skew   Kurtosis
소고기   25  9.828  3.347078    9.5  4.4 18.0  7.8 10.6  0.77847879  0.2225442
돼지닭   25  7.896  3.694081    7.8  1.4 14.0  4.9 10.8  0.03504453 -1.4190709
계란     25  2.936  1.117617    2.9  0.5  4.7  2.7  3.7 -0.39426844 -0.6646212
우유     25 17.112  7.105416   17.6  4.9 33.7 11.1 23.3  0.21508664 -0.7787642
생선     25  4.284  3.402533    3.4  0.2 14.2  2.1  5.8  1.09632534  0.7974802
곡식     25 32.248 10.974786   28.0 18.6 56.7 24.3 40.1  0.81881605 -0.4981383
전분     25  4.276  1.634085    4.7  0.6  6.5  3.1  5.7 -0.58880387 -0.6590063
콩견과   25  3.072  1.985682    2.4  0.7  7.8  1.5  4.7  0.63338857 -0.8391066
과일채소 25  4.136  1.803903    3.8  1.4  7.9  2.9  4.9  0.54556050 -0.8604672
```

multivariateNormality 값을 살펴보면 Mardia Skewness, Mardia Kurtosis의 result가 YES이다. (alpha=0.05)
따라서, 다변량정규성을 만족한다.

(1) PCFA를 실시하여 스크리그림을 통하여 인자개수를 정하고 총 기여율을 구하라.

protein.txt.의 PCFA

[1단계] 다변량 자료행렬 X를 준비한다.

```
> head(X)
```

	소고기	돼지닭	계란	우유	생선	곡식	전분	콩견과	과일채소
알바니아	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
오스트리아	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
벨기에-룩셈	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
불가리아	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
체코	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
덴마크	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4

```
> dim(X)
[1] 25  9
```

[자료 1.3.3] (protein.txt)은 크기가 25×6인 다변량 자료행렬이다.

[2단계] 표준화자료행렬 Z와 상관행렬 R을 계산한다.

```
> Z
                  소고기      돼지닭        계란        우유        생선        곡식        전분      콩견과    과일채소
알바니아      0.08126490 -1.75848885 -2.17963852 -1.15573814 -1.200282130  0.9159176 -2.24957717  1.2227536 -1.35040507
오스트리아   -0.27725673  1.65237315  1.22045441  0.39237676 -0.641874675 -0.3870690 -0.41368721 -0.8923886  0.09091397
벨기에-룩셈   1.09707621  0.38006748  1.04150215  0.05460623  0.063482111 -0.5146342  0.87143577 -0.4895043 -0.07539207
불가리아     -0.60590157 -0.51325352 -1.19540109 -1.24018077 -0.906383469  2.2280161 -1.94359551  0.3162641  0.03547862
체코         -0.03824231  0.94854448 -0.12168754 -0.64908235 -0.671264541  0.1869740  0.44306145 -0.9931096 -0.07539207
덴마크        0.23064892  0.78612248  0.68359763  1.11013912  1.650534878 -0.9428885  0.32066878 -1.1945517 -0.96235764
동독         -0.42664075  1.00268515  0.68359763 -0.84611516  0.327990905 -0.6968701  1.36100643 -1.1441912 -0.29713346
핀랜드       -0.09799591 -0.81102719 -0.21116367  2.33455726  0.445550369 -0.5419696  0.50425778 -1.0434702 -1.51671112
프랑스        2.44153235  0.54248948  0.32569311  0.33608167  0.416160503 -0.3779572  0.32066878 -0.3384228  1.31049162
그리스        0.11114171 -1.32536352 -0.12168754  0.06868001  0.474940235  0.8612468 -1.27043586  2.3810458  1.31049162
헝가리       -1.35282164  1.21924781 -0.03221141 -1.04314796 -1.170892264  0.7154581 -0.16890188  1.1723931  0.03547862
아일랜드      1.21658342  0.56955981  1.57835893  1.22272929 -0.612484809 -0.7515408  1.17741743 -0.7413070 -0.68518090
이탈리       -0.24737993 -0.75688652 -0.03221141 -0.48019709 -0.259806416  0.4147689 -1.33163219  0.6184273  1.42136232
네덜란드     -0.09799591  1.54409181  0.59412150  0.88495877 -0.524315210 -0.8973295 -0.04650921 -0.6405859 -0.24169812
노르웨이     -0.12787272 -0.86516785 -0.21116367  0.87088500  1.591755146 -0.8426588  0.19827612 -0.7413070 -0.79605159
폴랜드       -0.87479279  0.62370048 -0.21116367  0.30793413 -0.377365880  0.3509863  0.99382843 -0.5398649  1.36592697
포르투갈     -1.08393041 -1.13587119 -1.64278174 -1.71868901  2.914299118 -0.4781870  0.99382843  0.8198694  2.08658649
루마니아     -1.08393041 -0.43204252 -1.28487722 -0.84611516 -0.965163201  1.5810786 -0.71966887  1.1220326 -0.74061625
스페인       -0.81503919 -1.21708219  0.14674085 -1.19795945  0.798228762 -0.2777275  0.87143577  1.4241958  1.69853906
스웨덴        0.02151130 -0.02598752  0.50464537  1.06791780  0.945178092 -1.1615716 -0.35249087 -0.8420280 -1.18409903
스위스        0.97756900  0.59663015  0.14674085  0.94125386 -0.583094943 -0.6057521 -0.90325786 -0.3384228  0.42352606
영국          2.26227153 -0.59446452  1.57835893  0.49089316  0.004702379 -0.7242054  0.25947245  0.1651825 -0.46343951
소련         -0.15774952 -0.89223819 -0.74802044 -0.07205771 -0.377365880  1.0343709  1.29981010  0.1651825 -0.68518090
서독          0.46966334  1.24631815  1.04150215  0.23756527 -0.259806416 -1.2435778  0.56545411 -0.7916675 -0.18626277
유고슬라비아 -1.62171287 -0.78395685 -1.55330561 -1.07129551 -1.082722666  2.1551217 -0.78086520  1.3234747 -0.51887486
attr(,"scaled:center")
  소고기   돼지닭    계란    우유    생선    곡식    전분  콩견과 과일채소
   9.828   7.896   2.936  17.112   4.284  32.248   4.276   3.072   4.136
attr(,"scaled:scale")
   소고기    돼지닭     계란     우유     생선     곡식     전분   콩견과 과일채소
 3.347078 3.694081 1.117617 7.105416 3.402533 10.974786 1.634085 1.985682 1.803903
```

```
> R
                 소고기      돼지닭        계란        우유        생선        곡식        전분      콩견과    과일채소
소고기      1.00000000  0.1530027  0.58560895  0.5029311  0.06095745 -0.49987746  0.13542594 -0.3494486 -0.07422123
돼지닭      0.15300271  1.0000000  0.62040916  0.2814839 -0.23400923 -0.41379691  0.31377205 -0.6349618 -0.06131670
계란        0.58560895  0.6204092  1.00000000  0.5755331  0.06557136 -0.71243682  0.45223071 -0.5597810 -0.04551755
우유        0.50293110  0.2814839  0.57553312  1.0000000  0.13788370 -0.59273662  0.22241118 -0.6210875 -0.40836414
생선        0.06095745 -0.2340092  0.06557136  0.1378837  1.00000000 -0.52423080  0.40385286 -0.1471529  0.26613865
곡식       -0.49987746 -0.4137969 -0.71243682 -0.5927366 -0.52423080  1.00000000 -0.53326231  0.6509973  0.04654808
전분        0.13542594  0.3137721  0.45223071  0.2224112  0.40385286 -0.53326231  1.00000000 -0.4743116  0.08440956
콩견과     -0.34944855 -0.6349618 -0.55978097 -0.6210875 -0.14715294  0.65099727 -0.47431155  1.0000000  0.37496971
과일채소   -0.07422123 -0.0613167 -0.04551755 -0.4083641  0.26613865  0.04654808  0.08440956  0.3749697  1.00000000
```

소고기와 돼지닭, 계란, 우유, 전분은 서로 양의 관계를 보이며 이는 곡식, 콩견과, 과일채소와는 음의 관계를 보이고 있다.(전분과 과일채소는 약한 양의 상관이다.) 특히, 곡식과 콩견과, 생선과 과일채소 사이에는 양의 상관을 보이고 있다. 전분은 곡식과 매우 강한 음의 상관을 보여 이는 서로 대체할 수 있는 단백질 섭취원을 보여준다.

[3단계] R의 스펙트럼분해를 통해 고윳값과 고유벡터를 구한다.

고윳값

```
> round(eigen.R$values, 2)
 [1] 4.01 1.63 1.13 0.95 0.46 0.33 0.27 0.12 0.10
```

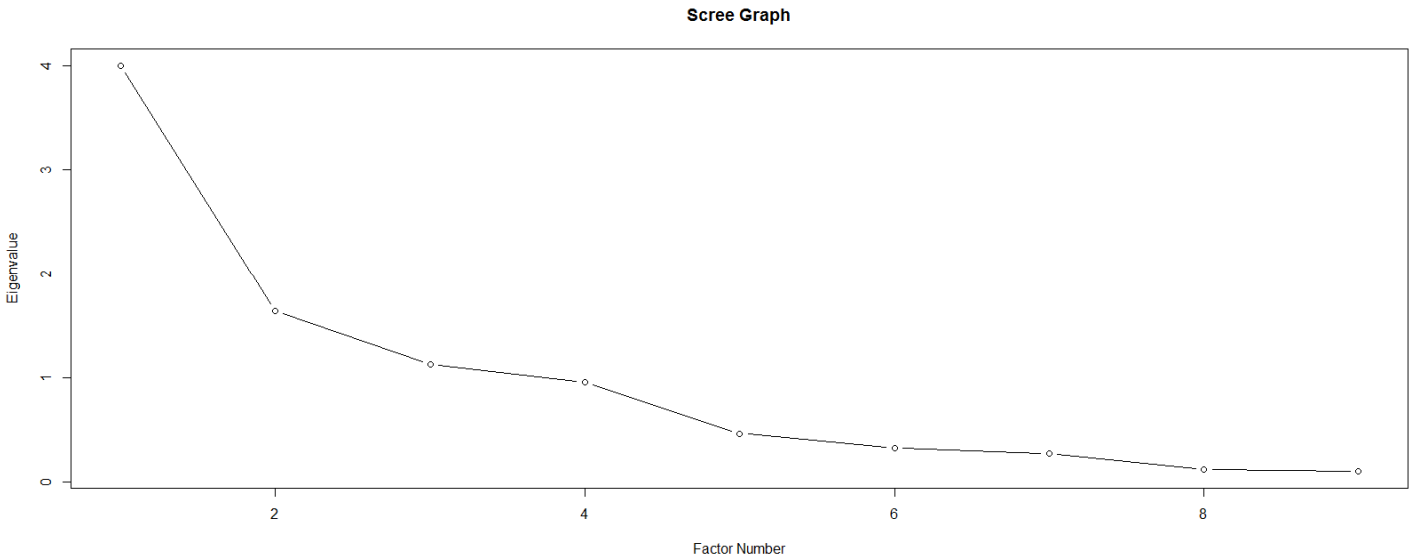
고유벡터

```
> round(V,3)
        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]
 [1,] -0.303 -0.056  0.298  0.646 -0.322  0.460  0.150 -0.020 -0.246
 [2,] -0.311 -0.237 -0.624 -0.037  0.300  0.121 -0.020 -0.028 -0.592
 [3,] -0.427 -0.035 -0.182  0.313 -0.079 -0.361 -0.443 -0.491  0.333
 [4,] -0.378 -0.185  0.386 -0.003  0.200 -0.618  0.462  0.081 -0.178
 [5,] -0.136  0.647  0.321 -0.216  0.290  0.137 -0.106 -0.449 -0.313
 [6,]  0.438 -0.233 -0.096 -0.006 -0.238 -0.081  0.405 -0.703 -0.152
 [7,] -0.297  0.353 -0.243 -0.337 -0.736 -0.148  0.153  0.115 -0.122
 [8,]  0.420  0.143  0.054  0.330 -0.151 -0.447 -0.407  0.184 -0.518
 [9,]  0.110  0.536 -0.408  0.462  0.234 -0.119  0.450  0.092  0.203
```

[4단계] 인자의 수 m에 따른 공통인자 기여율을 구한다.

```
> round(gof, 3)
 [1] 44.516 18.167 12.532 10.607  5.154  3.613  3.018  1.292  1.101
```

m=3개 고윳값(l1=4.01, l2=1.63, l3=1.13)의 총 기여율 : 44.516+18.167+12.532=75.215%



스크리그림을 보면 팔꿈치가 4에서 이루어진다. 인자수는 3으로 하는 것이 시각적으로 타당해 보인다.
3요인이 선택되었을 때의 적합도는 75.215%이다.

[5단계] 인자적재행렬의 추정

인자적재행렬

```
> round(L, 3)
          요인1  요인2  요인3
소고기   -0.606 -0.072  0.316
돼지닭   -0.622 -0.303 -0.663
계란     -0.854 -0.045 -0.193
우유     -0.756 -0.236  0.410
생선     -0.272  0.827  0.341
곡식      0.876 -0.299 -0.102
전분     -0.595  0.451 -0.258
콩견과    0.841  0.183  0.058
과일채소  0.221  0.686 -0.433
```

요인1 : 곡식, 콩견과에서 양 소고기, 돼지닭, 계란, 우유, 전분에서 음의 값을 가지므로 서북유럽과 동유럽간의 관계를 보여준다. (생선과 과일채소의 값은 절댓값이 낮다.)

요인2 : 생선과 과일채소에서의 절댓값이 높으므로 남유럽에 대한 요인이다.

요인3 : 돼지닭에서의 절댓값이 높으므로 돼지닭에서 단백질을 섭취하는 곳을 보여준다.

[6단계] 특성분산의 추정

```
> round(Psi, 3)
  소고기   돼지닭     계란     우유     생선     곡식     전분   콩견과 과일채소
   0.528    0.083    0.231    0.205    0.126    0.133    0.376    0.255    0.294
```

세 요인들에 대한 변수들의 설명력을 보면 다른 변수들에 비하여 소고기의 설명력은 조금 떨어진다.

(0에 가까울수록 설명력이 높음)

[7단계] 잔차행렬

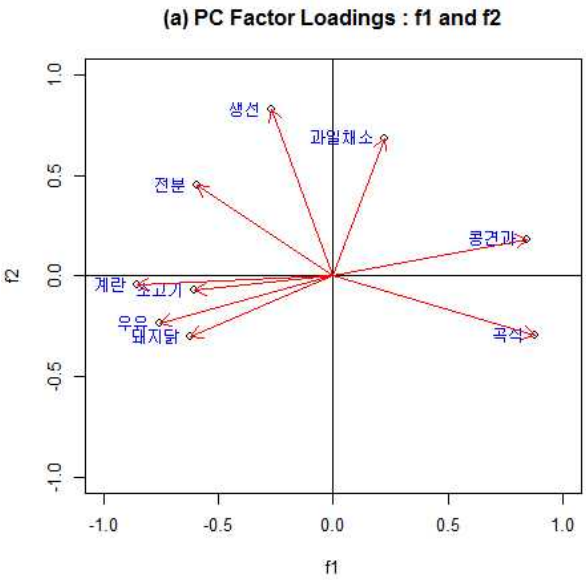
```
> round(Rm,3)
          소고기 돼지닭   계란   우유   생선   곡식   전분 콩견과 과일채소
소고기     0.000 -0.036  0.126 -0.101 -0.152  0.042 -0.111  0.155    0.246
돼지닭    -0.036  0.000 -0.052  0.011  0.074 -0.027 -0.090 -0.018   -0.003
계란       0.126 -0.052  0.000 -0.002 -0.063  0.003 -0.085  0.178    0.091
우유      -0.101  0.011 -0.002  0.000 -0.012  0.041 -0.015  0.035    0.098
생선      -0.152  0.074 -0.063 -0.012  0.000 -0.005 -0.043 -0.090   -0.093
곡식       0.042 -0.027  0.003  0.041 -0.005  0.000  0.096 -0.026    0.013
전분      -0.111 -0.090 -0.085 -0.015 -0.043  0.096  0.000 -0.041   -0.205
콩견과     0.155 -0.018  0.178  0.035 -0.090 -0.026 -0.041  0.000    0.088
과일채소   0.246 -0.003  0.091  0.098 -0.093  0.013 -0.205  0.088    0.000
```

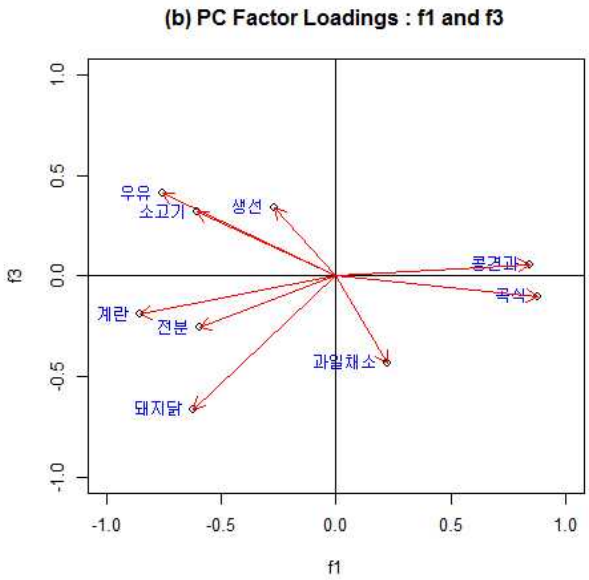
대각원소는 0이 되고 비대각원소가 전반적으로 작아 m=3개의 공통인자를 가지는 인자모형은 적절하다고 할 수 있다.

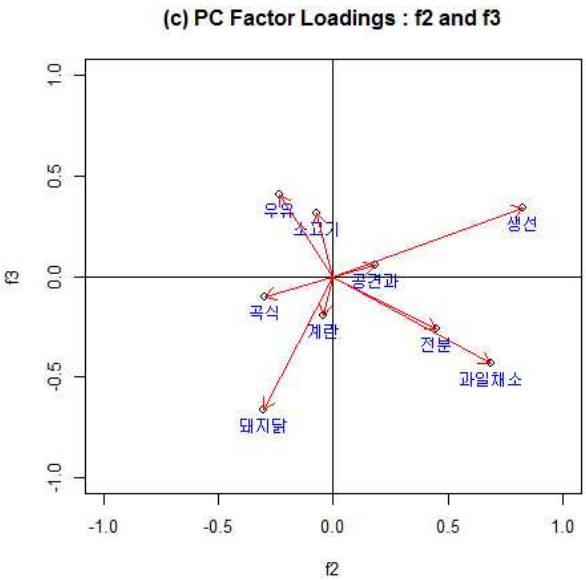
(2) 인자적재값과 인자적재그림을 통하여 인자를 해석하라.

인자적재행렬

```
> round(L, 3)
         요인1  요인2  요인3
소고기   -0.606 -0.072  0.316
돼지닭   -0.622 -0.303 -0.663
계란     -0.854 -0.045 -0.193
우유     -0.756 -0.236  0.410
생선     -0.272  0.827  0.341
곡식      0.876 -0.299 -0.102
전분     -0.595  0.451 -0.258
콩견과    0.841  0.183  0.058
과일채소  0.221  0.686 -0.433
```







1요인은 쇠고기, 계란, 우유, 돼지닭에서 단백질을 얻는 서북유럽과 콩견과, 곡식에서 단백질을 얻는 동유럽 사이의 요소이다. 2요인에서는 생선과 과일채소에서의 절댓값이 높으므로 남유럽에 대한 요인이다. 3요인은 돼지닭이 가장 크기 때문에 돼지닭에 대한 요소이다.

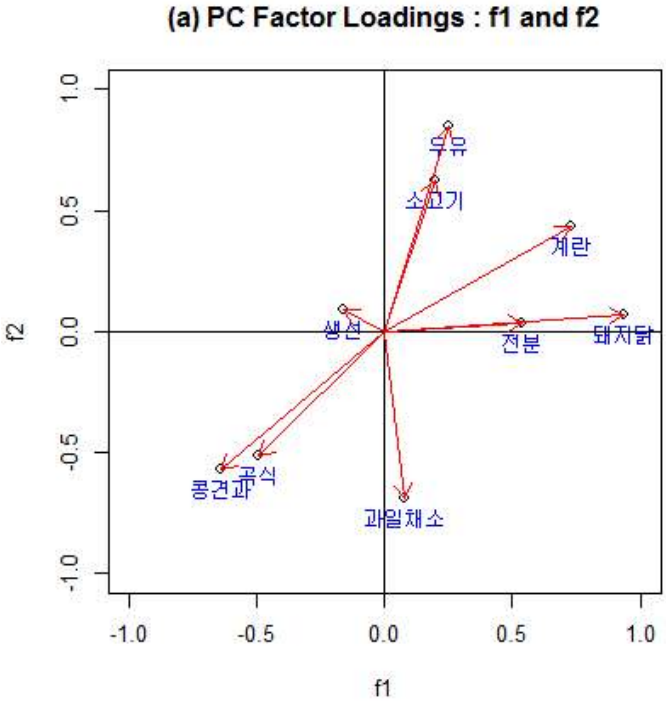
쇠고기, 계란, 우유, 돼지닭은 서로 연관되어있고 콩견과와는 음의 상관을 가진다. 또한 곡식과 전분도 음의 상관을 띈다.

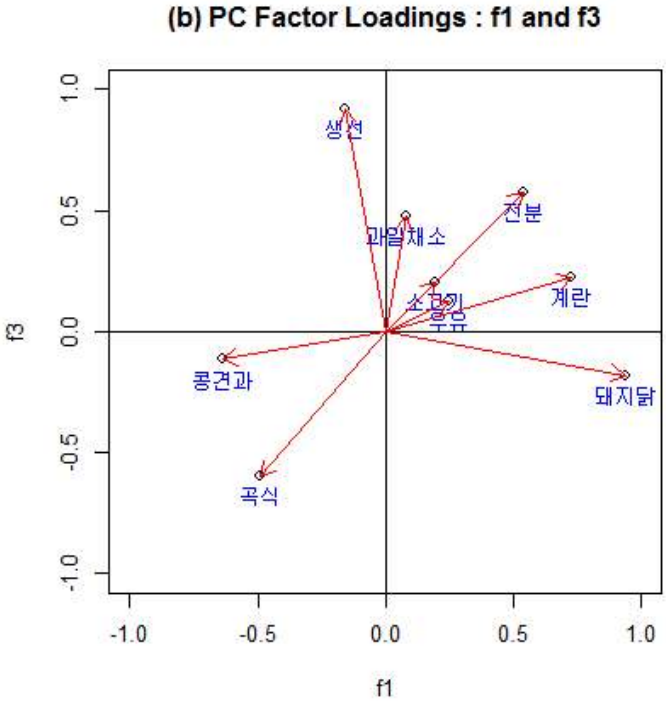
변수들간의 더 확실한 해석을 위해 직교회전을 하였다.

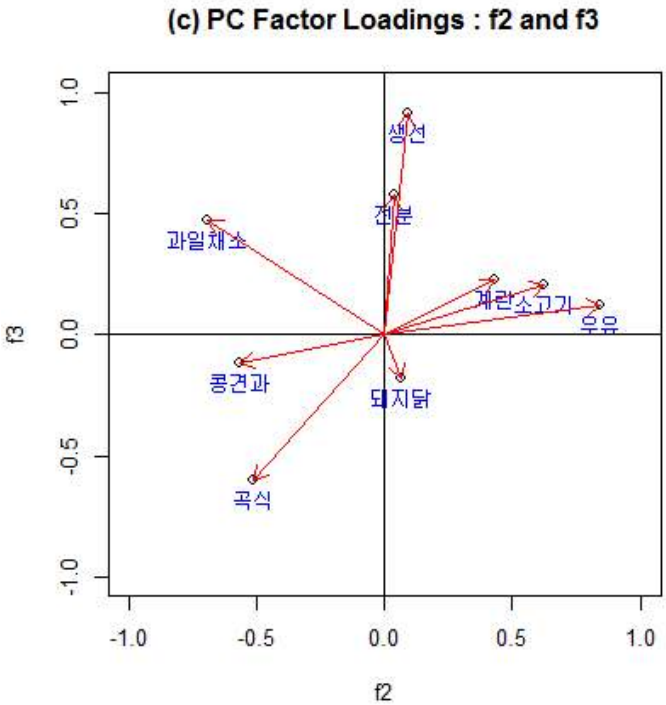
```
> round(L, 3)
           RC3    RC1    RC2
소고기     0.195  0.626  0.205
돼지닭     0.938  0.068 -0.180
계란       0.728  0.434  0.224
우유       0.250  0.847  0.124
생선      -0.158  0.089  0.917
곡식      -0.491 -0.515 -0.601
전분       0.539  0.037  0.576
콩견과    -0.639 -0.568 -0.117
과일채소   0.079 -0.690  0.473

> round(pcfa$values, 3)
[1] 4.006 1.635 1.128 0.955 0.464 0.325 0.272 0.116 0.099
> round(gof, 3)
[1] 44.516 18.167 12.532 10.607  5.154  3.613  3.018  1.292  1.101
> round(Psi,3)
  소고기   돼지닭     계란     우유     생선     곡식     전분   콩견과 과일채소
   0.528    0.083    0.231    0.205    0.126    0.133    0.376    0.255    0.294
> round(Rm, 3)
        소고기 돼지닭   계란   우유   생선   곡식   전분 콩견과 과일채소
소고기   0.000 -0.036  0.126 -0.101 -0.152  0.042 -0.111  0.155    0.246
돼지닭  -0.036  0.000 -0.052  0.011  0.074 -0.027 -0.090 -0.018   -0.003
계란     0.126 -0.052  0.000 -0.002 -0.063  0.003 -0.085  0.178    0.091
우유    -0.101  0.011 -0.002  0.000 -0.012  0.041 -0.015  0.035    0.098
생선    -0.152  0.074 -0.063 -0.012  0.000 -0.005 -0.043 -0.090   -0.093
곡식     0.042 -0.027  0.003  0.041 -0.005  0.000  0.096 -0.026    0.013
전분    -0.111 -0.090 -0.085 -0.015 -0.043  0.096  0.000 -0.041   -0.205
콩견과   0.155 -0.018  0.178  0.035 -0.090 -0.026 -0.041  0.000    0.088
과일채소 0.246 -0.003  0.091  0.098 -0.093  0.013 -0.205  0.088    0.000
```

특성분산과 잔차행렬, 교윳값 기여율은 동일하다. 하지만 인자적재행렬에서의 인자적재값이 달라졌다.
첫 번째 요인은 직교회전 전의 f3, 두 번째 요인은 직교회전 전의 f1, 세 번째 요인은 직교회전 전의 f2와 데이터 구조가 비슷하다. 직교회전을 하고 생성된 세 요인들을 순서대로 f1, f2, f3로 다시 정의하고 인자적재그림을 살펴본다.



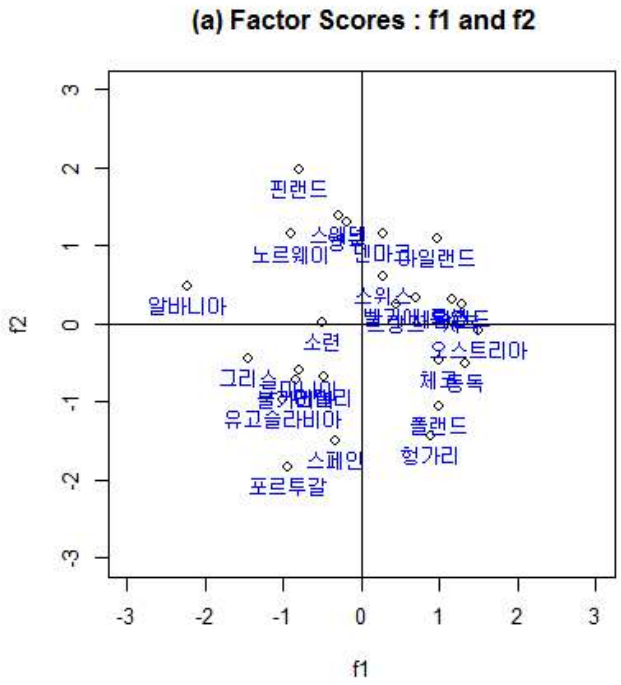


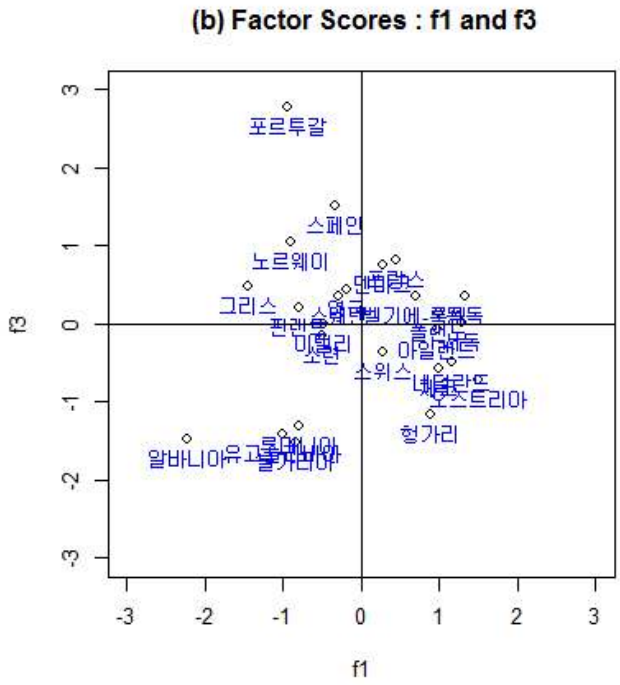


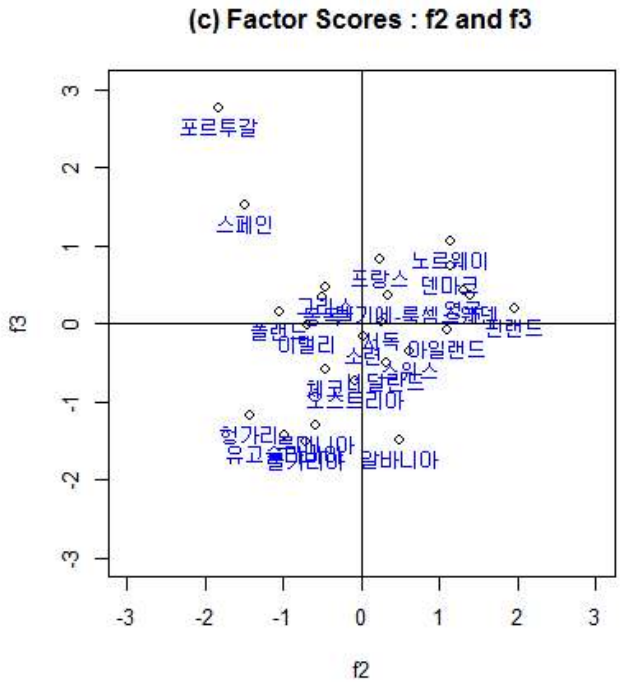
1요인은 돼지닭이 가장 크기 때문에 돼지닭에 대한 요소이다. 2요인에서는 쇠고기, 계란, 우유에서 단백질을 얻는 서북유럽과 콩견과, 곡식, 과일채소에서 단백질을 얻는 동유럽 및 남유럽 사이의 요소이다. 3요인은 생선에서의 값이 높으므로 생선에 대한 요소이다.

곡식과 계란와의 강한 음의 관계를 보인다. 우유와 소고기, 콩견과와 곡식은 높은 상관을 보인다.

(3) 인자점수그림을 통해 유럽 25개국 군집의 형성과 특성을 살펴보라.







1요인은 돼지닭에 대한 요소이며 알바니아가 돼지닭에서 단백질 섭취가 낮고 오스트리아는 돼지닭에서 단백질 섭취가 높다.
2요인은 쇠고기, 계란, 우유에서 단백질을 얻는 서북유럽과 콩견과, 곡식, 과일채소에서 단백질을 얻는 동유럽과 남유럽 사이의 요소로 아일랜드, 핀랜드, 스웨덴, 영국, 덴마크 등 주로 서북유럽 국가들이 상위에 위치해있고 루마니아, 유고슬라비아, 불가리아, 알바니다, 포르투갈, 스페인 등 주로 동유럽과 남유럽이 하위에 위치해있다.
3요인에서는 생선에 대한 요소이며 포르투갈 및 스페인이 생선에서 단백질 섭취가 높고, 알바니아는 생선에서 단백질 섭취가 낮다.

(4) (1)의 인자개수에 대해 MLFA를 실시하고 (2)~(3)을 시행한 후에 결과를 서로 비교하라.
데이터는 다변량 정규성을 만족하므로 MLFA를 사용한다.
또한 해석의 편의를 위해 직교회전을 하여 MLFA를 구한다.

특정분산

```
> round(Psi, 3)
  소고기 돼지닭   계란   우유   생선   곡식   전분 콩견과 과일채소
   0.490  0.005  0.265  0.400  0.012  0.135  0.633  0.356    0.783
```

세 요인들에 대한 변수들의 설명력을 보면 다른 변수들에 비하여 과일채소의 설명력은 조금 떨어진다.
(0에 가까울수록 설명력이 높음)

잔차행렬

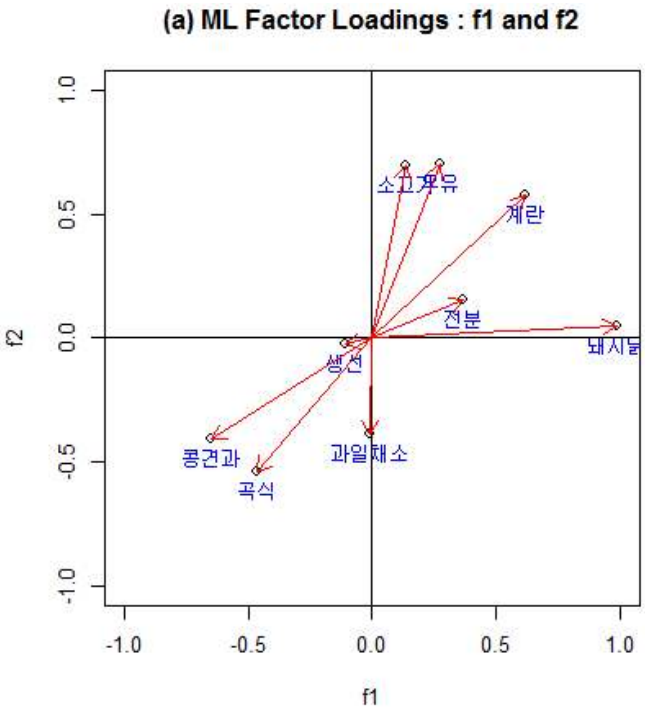
```
> round(Rm, 3)
        소고기 돼지닭   계란   우유   생선   곡식   전분 콩견과 과일채소
소고기   0.000  0.000  0.090 -0.038 -0.001 -0.007 -0.065  0.044    0.172
돼지닭   0.000  0.000  0.000  0.000  0.000  0.000 -0.001 -0.001    0.001
계란     0.090  0.000  0.000 -0.024 -0.001 -0.023  0.066  0.113    0.145
우유    -0.038  0.000 -0.024  0.000  0.002  0.023 -0.072 -0.113   -0.182
생선    -0.001  0.000 -0.001  0.002  0.000  0.000 -0.001 -0.002    0.001
곡식    -0.007  0.000 -0.023  0.023  0.000  0.000 -0.006 -0.012   -0.014
전분    -0.065 -0.001  0.066 -0.072 -0.001 -0.006  0.000 -0.066    0.031
콩견과   0.044 -0.001  0.113 -0.113 -0.002 -0.012 -0.066  0.000    0.268
과일채소 0.172  0.001  0.145 -0.182  0.001 -0.014  0.031  0.268    0.000
```

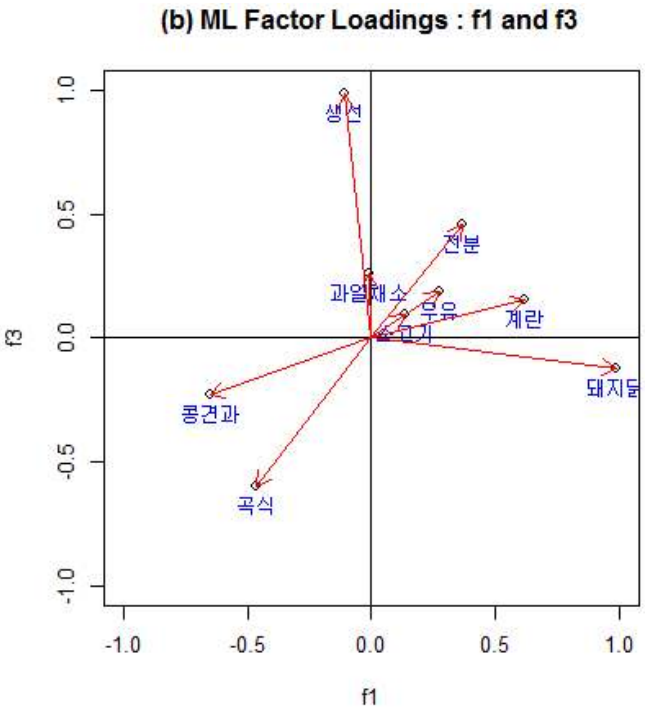
대각원소는 0이 되고 비대각원소가 전반적으로 작아 m=3개의 공통인자를 가지는 인자모형은 적절하다고 할 수 있다.

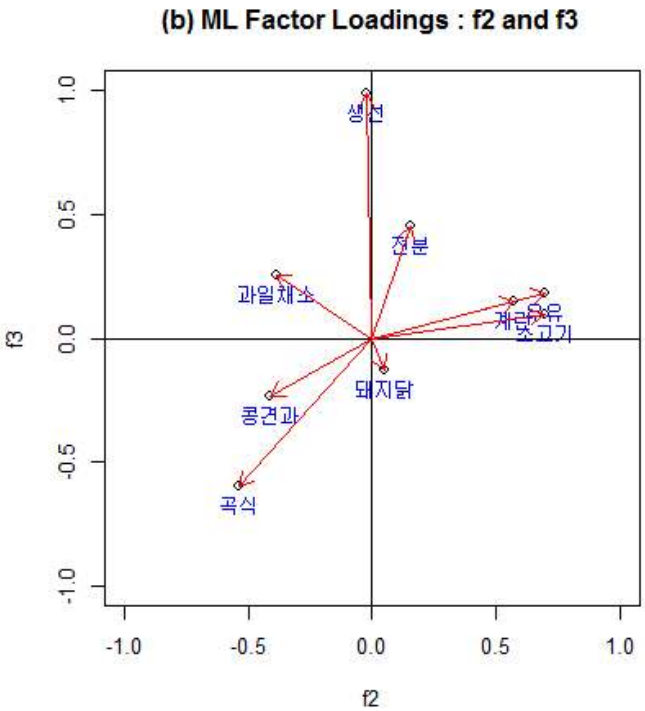
인자적재행렬

```
> round(Lm, 3)
        Factor1 Factor2 Factor3
소고기    0.133   0.695   0.096
돼지닭    0.988   0.049  -0.125
계란      0.618   0.574   0.152
우유      0.274   0.700   0.187
생선     -0.111  -0.026   0.988
곡식     -0.468  -0.538  -0.598
전분      0.369   0.155   0.456
콩견과   -0.650  -0.410  -0.231
과일채소 -0.011  -0.388   0.257
```

인자적재그림



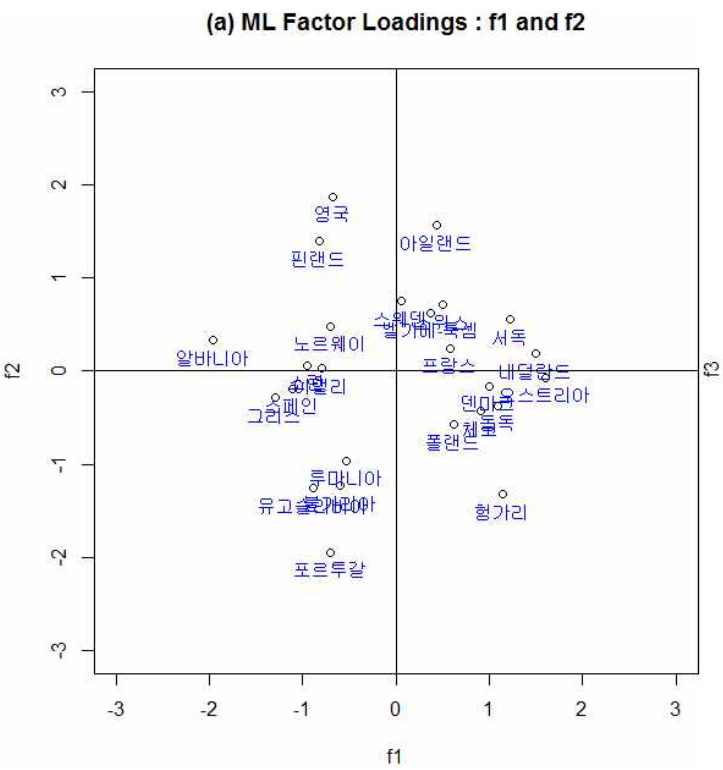


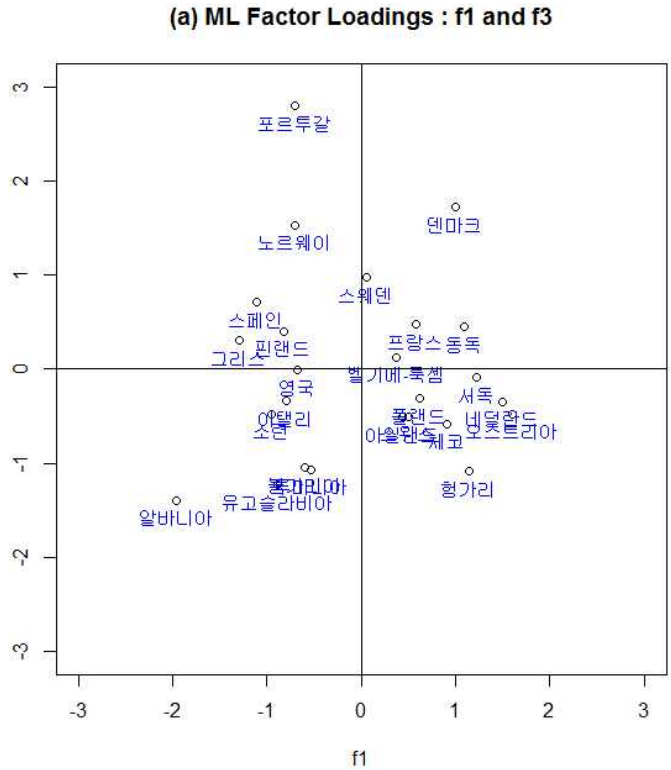


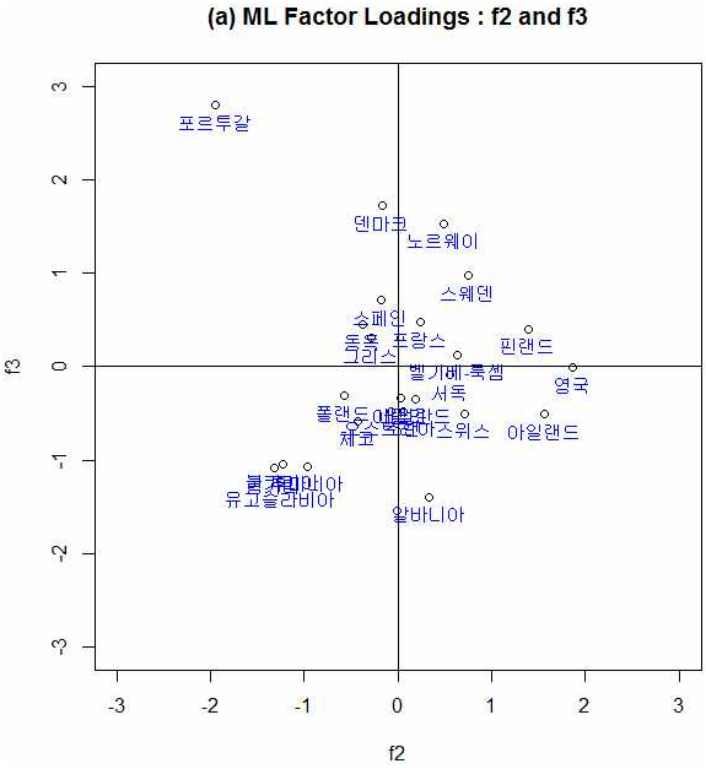
인자적재행렬과 인자적재그림을 살펴보면

요인1 : 돼지닭에서의 절댓값이 높으므로 돼지닭에서 단백질을 섭취하는 곳을 보여준다.
요인2 : 쇠고기, 계란, 우유에서 단백질을 얻는 서북유럽과 콩견과, 곡식, 과일채소에서 단백질을 얻는 동유럽과 남유럽을 보여준다.
요인3 : 생선에서의 절댓값이 높으므로 생선에서 단백질을 섭취하는 곳을 보여준다.

인자점수그림

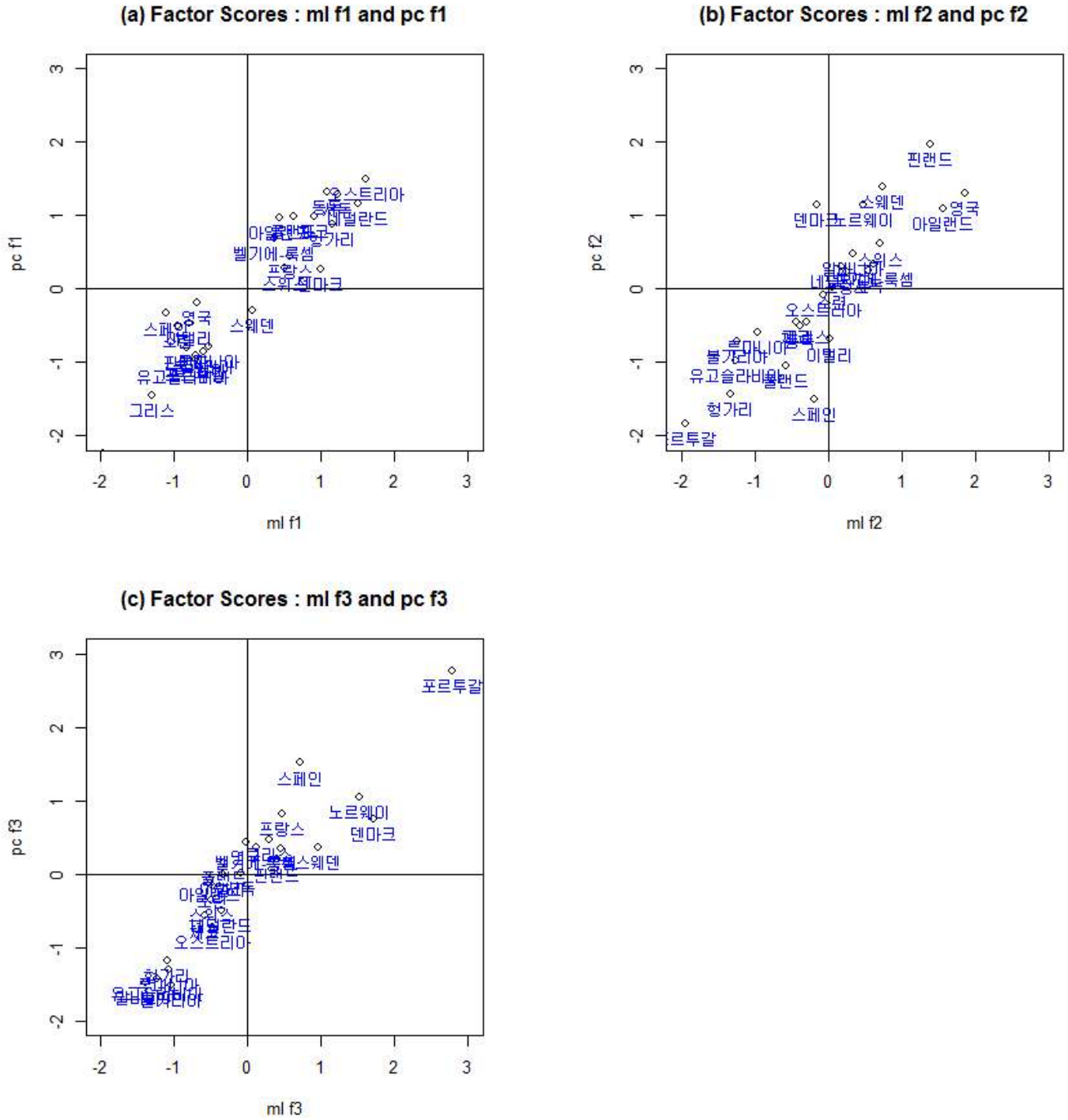






인자점수그림을 비교해보니 f1은 돼지닭에 대한 요소이며 알바니아가 돼지닭에서 단백질 섭취가 낮고 오스트리아는 돼지닭에서 단백질 섭취가 높다. 2요인은 서북유럽과 동유럽, 남유럽사이의 요소이며 계란, 소고기, 우유의 섭취 비율이 높은 아일랜드, 핀랜드, 스웨덴, 영국, 덴마크 등 주로 서북유럽 국가들이 상위에 위치해있고 곡식 및 콩견과의 섭취 비율이 높은 루마니아, 유고슬라비아, 불가리아, 알바니다 등 주로 동유럽 국가와 과일채소의 섭취비율이 높은 포르투갈이 하위에 위치해있다. f3에서는 생선에 대한 요소이며 포르투갈이 생선에서 단백질 섭취가 높고, 알바니아는 생선에서 단백질 섭취가 낮다.

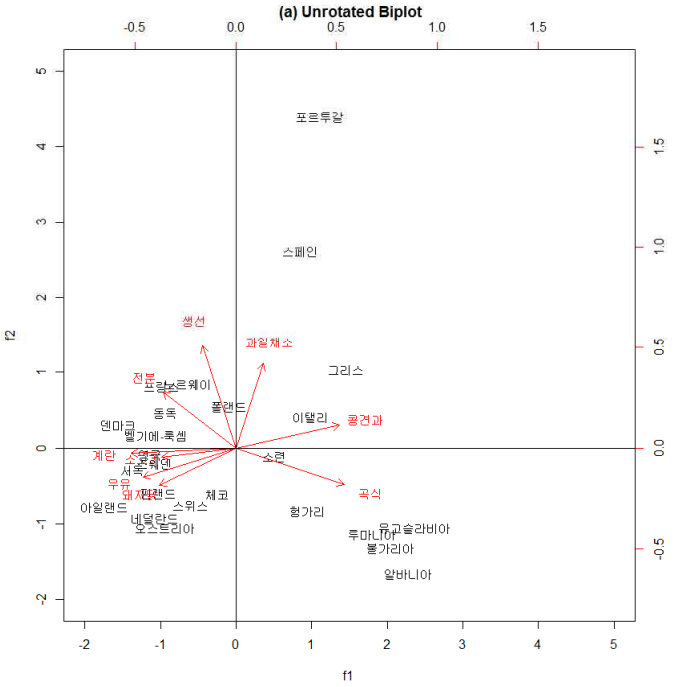
PCFA와 MLFA 두가지의 경우 모두 타원의 모양으로 다변량 정규성을 만족한다.

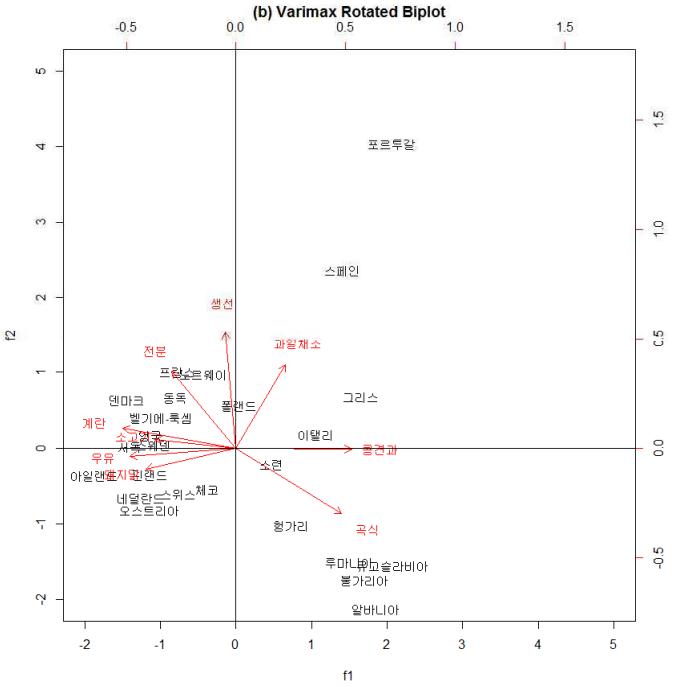


서로의 세 인자에 대하여 비교를 해본결과 두 인자가 직선에 가까우므로 서로의 요인이 비슷하다.

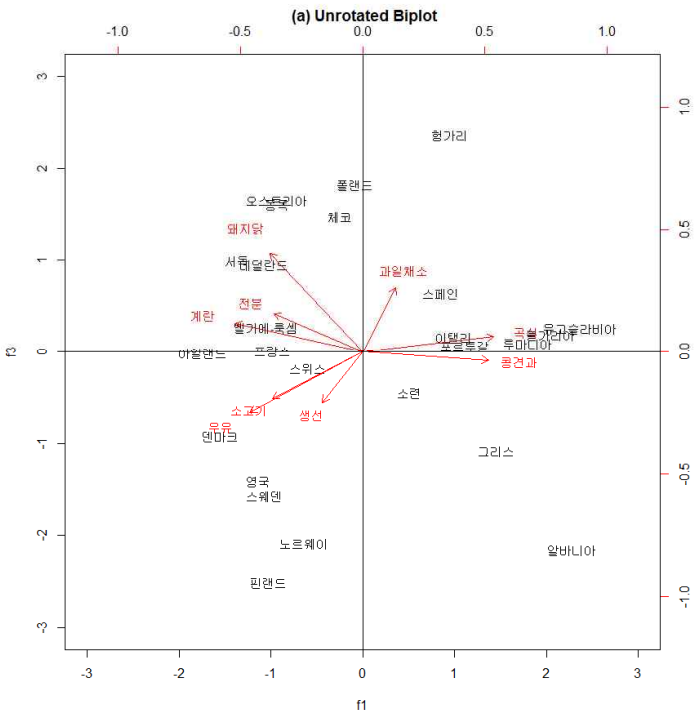
(5) 인자행렬도를 통해 단백질 섭취원 인자와 25개국 개체 간의 연관성을 살펴보라.

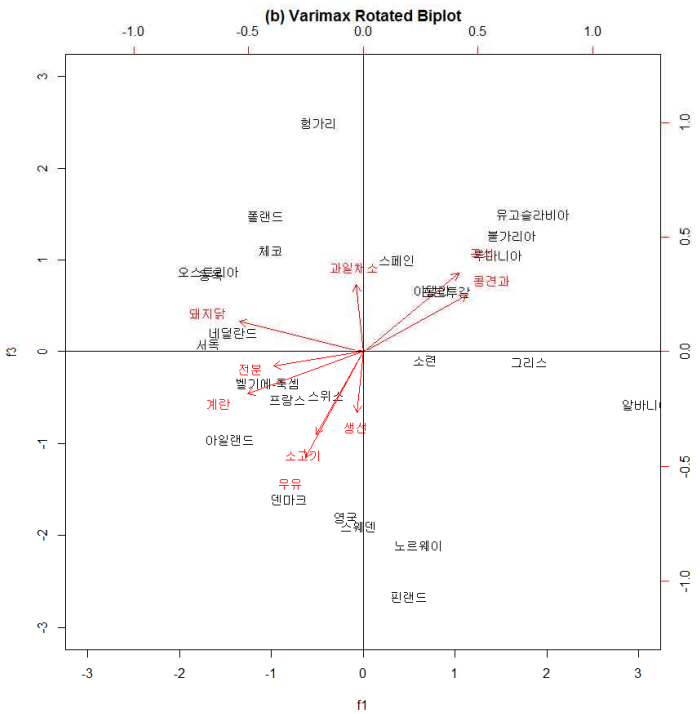
f1과 f2의 인자행렬도 (62.68%의 설명력)



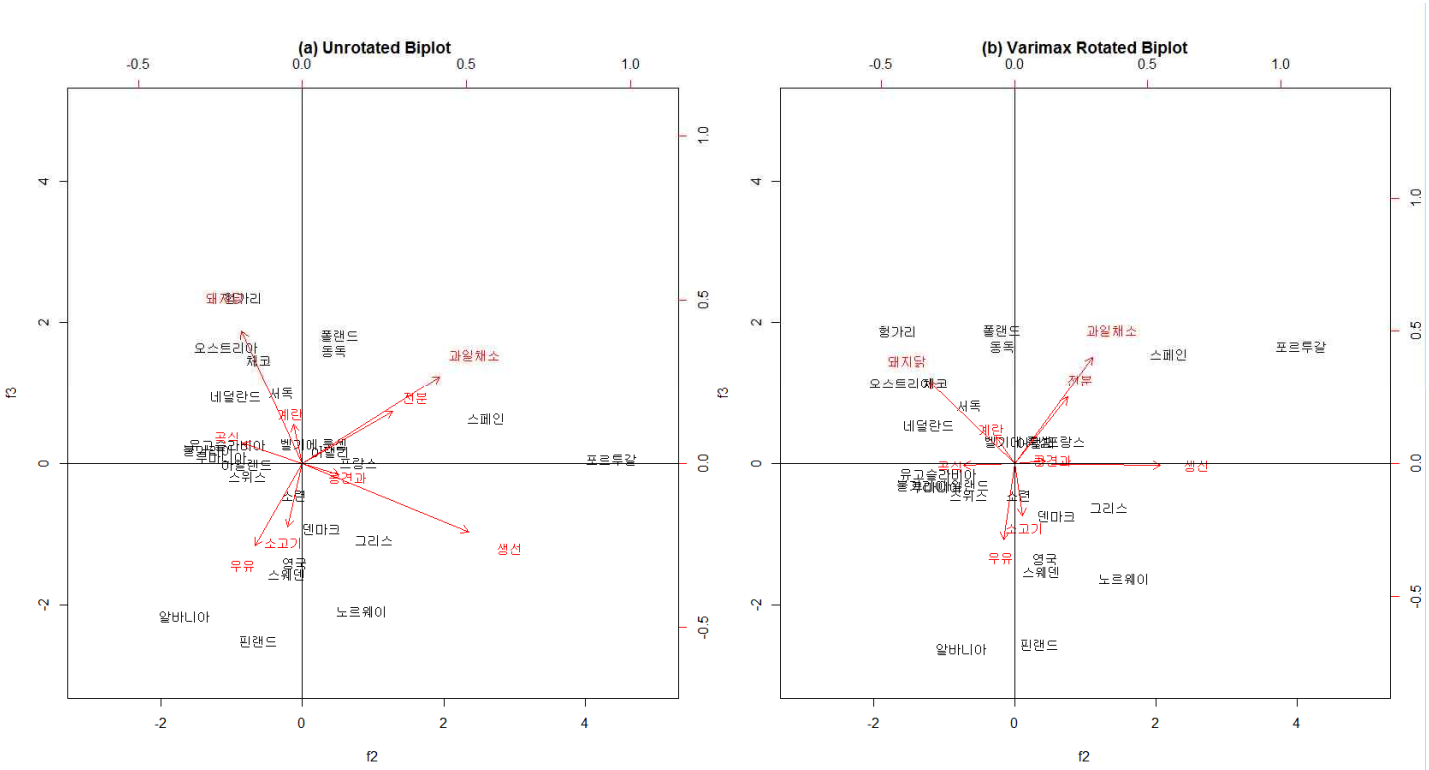


f1과 f3의 인자행렬도 (57.05%의 설명력)





f2과 f3의 인자행렬도 (30.70%의 설명력)



PCFA로 인자행렬도 만들어 단백질 섭취원 인자와 25개국 개체 간의 연관성을 살펴보고자한다. 회전의 유무에 따라 큰 차이는 존재하지는 않지만 회전을 했을 때가 하지 않았을 때보다 변수들의 관계가 더 분병하다.

가장 설명력이 높은 f1과 f2의 인자행렬도를 살펴보면 계란, 소고기, 우유, 돼지닭은 높은 연관을 보이고 있고 콩견과 및 곡식과는 음의 상관을 보이고 있다. 또한 생선과는 연관성이 낮아 보인다.

단백질 섭취원 인자와 25개국 개체 간의 관계를 살펴보면 포르투갈과 스페인(서유럽국가)은 과일채소 및 생선섭취비율이 높았다. 계란, 소고기, 우유, 돼지닭의 섭취 비율은 아일랜드, 핀랜드, 스웨덴, 영국, 덴마크 등 주로 서북유럽 국가들이다. 곡식 및 콩견과의 섭취 비율이 높은 국가는 루마니아, 유고슬라비아, 불가리아, 알바니다 등으로 동유럽 국가들이다. f1과 f3의 인자행렬도에도 비슷한 결과를 가진다.

```
setwd("D:/2020 1학기 정호재/다변량통계학(1)/200402 다변량 실습1/Rdata")
Data1.3.3<-read.table("protein1.txt", header=T)
X<-Data1.3.3

# Barplot of 25 Countries
X<-t(X)
par(las=2)                # label style 1,2,3
par(mar=c(4,4,1,2))       # 여백 mar=(c(아래,왼쪽,위,오른쪽))
barplot(X, legend=rownames(X), horiz=TRUE)

# Star Plot
X<-scale(Data1.3.3)
stars(X, key.loc=c(0, 2), full=FALSE)

# MVN tests based on the Skewness and Kurtosis Ststistics
install.packages("MVN")
library("MVN") # for mardia test
mvn(X, mvnTest = "mardia", multivariatePlot =  "qq")

#################################################################
#[Step 1] Data Matrix X
Data1.3.3<-read.table("protein.txt", header=T)
X=Data1.3.3[,-c(1,2)]
rownames(X)<-Data1.3.3[,"국가"]
p=ncol(X)
n=nrow(X)
head(X)
dim(X)
Z<-scale(X)
Z

#[Step 2] Covariance Matrix S(or Correlation Matix R)
R=cor(X)
R

#[Step 3] Spectral Decomposition (# of factor)
eigen.R=eigen(R)
round(eigen.R$values, 2)
V=eigen.R$vectors
round(V,3)

#[Step 4] Number of factors : m (# of factor)
gof=eigen.R$values/p*100
round(gof, 3)
```

```
plot(eigen.R$values, type="b", main="Scree Graph", xlab="Factor Number", ylab="Eigenvalue")
#[Step 5]Factor LoadinPCAgs and Communality
V2=V[,1:3]
L=V2%*%diag(sqrt(eigen.R$values[1:3]))
rownames(L) = colnames(X)
colnames(L) = c("요인1","요인2","요인3")
round(L, 3)
round(diag(L%*%t(L)), 3)

#[Step 6]Specific Variance : 특정분산(Psi) ( 1- Communality )
Psi=diag(R-L%*%t(L))
round(Psi, 3)

#{Step 7] Residual Matrix ( 전체 = 공통성 + 특정분산 + 잔차 )
Rm = R-(L%*%t(L) + diag(Psi))
round(Rm,3)

############################################################
par(mfrow=c(2,2))
lim<-range(pretty(L))
#요인 1과 2
plot(L[,1], L[,2],main="(a) PC Factor Loadings : f1 and f2",  xlab="f1", ylab="f2",
     xlim=lim, ylim=lim)
text(L[,1], L[, 2], labels=rownames(L), cex=0.8, col="blue", pos=2)
abline(v=0, h=0)
arrows(0,0, L[,1], L[, 2], col=2, code=2, length=0.1)

#요인 1과 3
plot(L[,1], L[,3],main="(b) PC Factor Loadings : f1 and f3",  xlab="f1", ylab="f3",
     xlim=lim, ylim=lim)
text(L[,1], L[, 3], labels=rownames(L), cex=0.8, col="blue", pos=2)
abline(v=0, h=0)
arrows(0,0, L[,1], L[, 3], col=2, code=2, length=0.1)

#요인 2과 3
plot(L[,2], L[,3],main="(c) PC Factor Loadings : f2 and f3",  xlab="f2", ylab="f3",
     xlim=lim, ylim=lim)
text(L[,2], L[, 3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,2], L[, 3], col=2, code=2, length=0.1)
```

```
#######################################################
library(psych)
pcfa<-principal(Z, nfactors=3, rotate="varimax")
round(pcfa$values, 3)
gof=pcfa$values/p*100 # Goodness-of fit(적합도)
round(gof, 3)

# Residual Matrix
L=pcfa$loading[, 1:3]
round(L, 3)
Psi=pcfa$uniquenesses
round(Psi,3)
Rm = R-(L%*%t(L) + diag(Psi))
round(Rm, 3)

#######################################################
# Plot of PC Factor Loadings
par(mfrow=c(2,2))
lim<-range(pretty(L))
#요인 1과 2
plot(L[,1], L[,2],main="(a) PC Factor Loadings : f1 and f2",  xlab="f1", ylab="f2",
     xlim=lim, ylim=lim)
text(L[,1], L[, 2], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,1], L[, 2], col=2, code=2, length=0.1)

#요인 1과 3
plot(L[,1], L[,3],main="(b) PC Factor Loadings : f1 and f3",  xlab="f1", ylab="f3",
     xlim=lim, ylim=lim)
text(L[,1], L[, 3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,1], L[, 3], col=2, code=2, length=0.1)

#요인 2과 3
plot(L[,2], L[,3],main="(c) PC Factor Loadings : f2 and f3",  xlab="f2", ylab="f3",
     xlim=lim, ylim=lim)
text(L[,2], L[, 3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,2], L[, 3], col=2, code=2, length=0.1)
```

```
#######################################################
# Factor Scores : Regression Method #관측치와 관련한 그래프를 그려보자
fpc=pcfa$scores

# Plot of Factor Scores : PFA (173p)
par(mfrow=c(2,2))
par(pty="s")
lim<-range(pretty(fpc))
plot(fpc[,1], fpc[,2],main=" (a) Factor Scores : f1 and f2",  xlab="f1", ylab="f2",
     xlim=lim, ylim=lim)
text(fpc[,1], fpc[,2], labels=rownames(fpc), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
#7,8이 CO NO값 다른 값보다 높음

plot(fpc[,1], fpc[,3],main=" (b) Factor Scores : f1 and f3",  xlab="f1", ylab="f3",
     xlim=lim, ylim=lim)
text(fpc[,1], fpc[,3], labels=rownames(fpc), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(fpc[,2], fpc[,3],main="(c) Factor Scores : f2 and f3",  xlab="f2", ylab="f3",
     xlim=lim, ylim=lim)
text(fpc[,2], fpc[,3], labels=rownames(fpc), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

#######################################################
##### MLFA using the factanal( ) ####
library(psych)
mlfa<-factanal(Z, factors = 3, rotation="varimax", score="regression")

# Residual Matrix
Lm=mlfa$loading[, 1:3]
round(Lm, 3)
Psi=mlfa$uniquenesses
Rm = R-(Lm%*%t(Lm) + diag(Psi))
round(Rm, 3)
```

```
#######################################################
# ML Factor Loadings Plot
par(mfrow=c(2,2))
lim<-range(pretty(L))
plot(Lm[,1], Lm[,2],main="(a) ML Factor Loadings : f1 and f2",  xlab="f1", ylab="f2",
     xlim=lim, ylim=lim)
text(Lm[,1], Lm[, 2], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, Lm[,1], Lm[, 2], col=2, code=2, length=0.1)

plot(Lm[,1], Lm[,3],main="(b) ML Factor Loadings : f1 and f3",  xlab="f1", ylab="f3",
     xlim=lim, ylim=lim)
text(Lm[,1], Lm[, 3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, Lm[,1], Lm[, 3], col=2, code=2, length=0.1)

plot(Lm[,2], Lm[,3],main="(b) ML Factor Loadings : f2 and f3",  xlab="f2", ylab="f3",
     xlim=lim, ylim=lim)
text(Lm[,2], Lm[, 3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, Lm[,2], Lm[, 3], col=2, code=2, length=0.1)

#######################################################
# Factor Scores : Regression Method
fml=mlfa$scores
round(fml, 3)
# Plot of Factor Scores : MLFA
par(mfrow=c(2,2))
par(pty="s")
lim<-range(pretty(fml))
plot(fml[,1], fml[,2],main=" (a) Factor Scores : f1 and f2",  xlab="f1", ylab="f2",
     xlim=lim, ylim=lim)
text(fml[,1], fml[,2], labels=rownames(fml), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(fml[,1], fml[,3],main=" (b) Factor Scores : f1 and f3",  xlab="f1", ylab="f3",
     xlim=lim, ylim=lim)
text(fml[,1], fml[,3], labels=rownames(fml), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(fml[,2], fml[,3],main="(c) Factor Scores : f2 and f3",  xlab="f2", ylab="f3",
     xlim=lim, ylim=lim)
text(fml[,2], fml[,3], labels=rownames(fml), cex=0.8, col="blue", pos=1)
```

```
abline(v=0, h=0)
#######################################################
# Plot of Factor Scores : Pairs(MLFA, PCFA) #MLFA와 PCFA의 차이
par(pty="s")
par(mfrow=c(2,2))
plot(fml[,1], fpc[,1],main="(a) Factor Scores : ml f1 and pc f1",  xlab="ml f1", ylab="pc f1",
     xlim=lim, ylim=lim)
text(fml[,1], fpc[,1], labels=rownames(fml), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
#두요인 비슷

plot(fml[,2], fpc[,2],main="(b) Factor Scores : ml f2 and pc f2",  xlab="ml f2", ylab="pc f2",
     xlim=lim, ylim=lim)
text(fml[,2], fpc[,2], labels=rownames(fml), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
#두요인 비슷

plot(fml[,3], fpc[,3],main="(c) Factor Scores : ml f3 and pc f3",  xlab="ml f3", ylab="pc f3",
     xlim=lim, ylim=lim)
text(fml[,3], fpc[,3], labels=rownames(fml), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
#두요인 비슷

#######################################################(1-2)
# Biplot based on the Singular Value Decomposition
svd.Z <- svd(Z)
U <- svd.Z$u
V <- svd.Z$v
D <- diag(svd.Z$d)
F <- (sqrt(n-1)*U)[,1:2]  # Factor Scores Matrix : F
L <- (sqrt(1/(n-1))*V%*%D)[,1:2] # Factor Loadings Matrix : Lambda
C<- rbind(F, L)
rownames(F)<-rownames(X)
rownames(L)<-colnames(X)

# Godness-of-fit
eig <- (svd.Z$d)^2
per <- eig/sum(eig)*100
gof <- sum(per[1:2])
per
gof
```

```
# Biplot: Joint Plot of Factor Loadings and Scores #회전 안했을 때
par(mfrow=c(1,2))
par(pty="s")
lim1 <- range(pretty(L))
lim2 <- range(pretty(F))
biplot(F,L, xlab="f1",ylab="f2", main=" (a) Unrotated Biplot",
       xlim=lim2,ylim=lim2,cex=0.8,pch=16)
abline(v=0,h=0)

# Varimax Rotated Biplot: Joint Plot of Rotated Factor Loadings and Scores
#회전 했을 때

varimax<-varimax(L)
Lt = varimax$loadings
T=varimax$rotmat
T
Ft= F%*%T
biplot(Ft,Lt, xlab="f1",ylab="f2", main="(b) Varimax Rotated Biplot",
       xlim=lim2,ylim=lim2,cex=0.8,pch=16)
abline(v=0,h=0)




##########################################################(1-3)
svd.Z <- svd(Z)
U <- svd.Z$u
V <- svd.Z$v
D <- diag(svd.Z$d)
F <- (sqrt(n-1)*U)[,c(1,3)]  # Factor Scores Matrix : F
L <- (sqrt(1/(n-1))*V%*%D)[,c(1,3)] # Factor Loadings Matrix : Lambda
C<- rbind(F, L)
rownames(F)<-rownames(X)
rownames(L)<-colnames(X)

# Godness-of-fit
eig <- (svd.Z$d)^2
per <- eig/sum(eig)*100
gof <- sum(per[c(1,3)])
per
gof
```

```
# Biplot: Joint Plot of Factor Loadings and Scores #회전 안했을 때
par(mfrow=c(1,2))
par(pty="s")
lim1 <- range(pretty(L))
lim2 <- range(pretty(F))
biplot(F,L, xlab="f1",ylab="f3", main=" (a) Unrotated Biplot",
       xlim=lim2,ylim=lim2,cex=0.8,pch=16)
abline(v=0,h=0)

# Varimax Rotated Biplot: Joint Plot of Rotated Factor Loadings and Scores
#회전 했을 때

varimax<-varimax(L)
Lt = varimax$loadings
T=varimax$rotmat
T
Ft= F%*%T
biplot(Ft,Lt, xlab="f1",ylab="f3", main="(b) Varimax Rotated Biplot",
       xlim=lim2,ylim=lim2,cex=0.8,pch=16)
abline(v=0,h=0)


#########################################################(2-3)
svd.Z <- svd(Z)
U <- svd.Z$u
V <- svd.Z$v
D <- diag(svd.Z$d)
F <- (sqrt(n-1)*U)[,2:3]  # Factor Scores Matrix : F
L <- (sqrt(1/(n-1))*V%*%D)[,2:3] # Factor Loadings Matrix : Lambda
C<- rbind(F, L)
rownames(F)<-rownames(X)
rownames(L)<-colnames(X)

# Godness-of-fit
eig <- (svd.Z$d)^2
per <- eig/sum(eig)*100
gof <- sum(per[2:3])
per
gof
```

```
# Biplot: Joint Plot of Factor Loadings and Scores #회전 안했을 때
par(mfrow=c(1,2))
par(pty="s")
lim1 <- range(pretty(L))
lim2 <- range(pretty(F))
biplot(F,L, xlab="f2",ylab="f3", main=" (a) Unrotated Biplot",
        xlim=lim2,ylim=lim2,cex=0.8,pch=16)
abline(v=0,h=0)

# Varimax Rotated Biplot: Joint Plot of Rotated Factor Loadings and Scores
#회전 했을 때

varimax<-varimax(L)
Lt = varimax$loadings
T=varimax$rotmat
T
Ft= F%*%T
biplot(Ft,Lt, xlab="f2",ylab="f3", main="(b) Varimax Rotated Biplot",
        xlim=lim2,ylim=lim2,cex=0.8,pch=16)
abline(v=0,h=0)
```