

# Stat 8053: Factor Analysis, rev. November 18, 2013

(Notation is from Chapter 10 of Härdle and Simar.)

$F$  is a  $k \times 1$  vector of *unobservable*, or *latent* common factor variables. In the normal factor analysis model, we assume

$$F \sim N_k(0, I) \quad (1)$$

The dimension  $k$  is also unknown.

$X$  is a  $p \times 1$  vector of observable or *manifest* variables. The factor analysis model specifies the conditional distribution of  $Y|F$  as

$$X|F \sim N_p(\mu + QF, \Psi) \quad (2)$$

where  $Q$  is a  $p \times k$  matrix of *factor loadings* and  $\Psi$  is assumed to be a diagonal matrix with nonnegative entries. Thus the model assumes that the manifest variables  $X$  have a linear regression on the latent variables  $F$ .

Standard calculations based (1)–(2) give

$$X \sim N_p(\mu, QQ' + \Psi) \quad (3)$$

so  $\mu$  is the unconditional mean of  $X$ , and  $\Sigma = QQ' + \Psi$  is the covariance matrix. The goal is to learn about  $Q, k$ , and  $\Psi$  based on (3).

An alternative representation of the normal factor analysis model is the single equation

$$X = QF + U + \mu \quad (4)$$

This introduces a new quantity  $U \sim N_p(0, \Psi)$  often called the vector of *specific factors*, and  $F$  is distributed as in (1). This differs by our understanding of the data generating mechanism. For (1)–(3) we have a two-step process of generating first a subject at random with latent value  $F$ , and then given  $F$  we generate  $X$ , while in (4) we envision  $F$  and  $U$  generated simultaneously to produce the manifest variables  $x$ . In either, only  $X$  is observable.

**Estimation** The only estimates we consider are maximum likelihood, assuming  $X_1, \dots, X_n$  are iid copies from the distribution in (3). The likelihood was derived in class, and is given in the textbook. The data will consist of the  $n \times p$  matrix of manifest variables  $X$ , each of whose rows satisfies (3). The sufficient statistic for  $Q$  and  $\Psi$ , is the sample correlation matrix, which has  $p(p+1)/2$  unique elements. All parameters of interest are in  $\Sigma$ . The factor loading matrix  $Q$  has  $pk$  parameters for a  $k$ -factor solution, while  $\Psi$  has  $p$  parameters. Additional constraints on the parameters are introduced to get a unique solution, and these introduce an additional  $k(k-1)/2$  parameters (see the textbook for details). Estimation is possible as long as the number of unique elements in the correlation matrix exceeds the number of parameters and constraints.

## US Company Data

We continue with the US Companies data. As suggested in the last handout, all but two of the variables are converted to log-scale using the `transform` function in R. I will choose to keep all companies including 38 and 40 in the data. We create a new variable called `sector` which represents the type of company, as described in the textbook.

```
loc <- "http://www.stat.umn.edu/~sandy/courses/8053/Data/uscomp1.dat"
uscomp <- read.table(url(loc),header=TRUE)
uscomp <- transform(uscomp, Assets=log(Assets), Sales=log(Sales),
  MarketValue=log(MarketValue), Employees=log(Employees))
head(uscomp)
```

	Assets	Sales	MarketValue	Profits	CashFlow	Employees
1	9.893	9.114	9.272	1092.9	2576.8	4.374
2	8.532	7.847	7.545	239.9	578.3	3.086
3	9.519	8.486	8.428	485.0	898.9	3.153
4	7.018	6.945	6.170	59.7	91.7	1.335
5	7.398	6.553	6.521	74.3	135.9	1.030
6	8.640	7.134	7.602	310.7	407.9	1.825

```
snames <-c("Com", "Enr", "Fin", "HiTch", "Manu", "Med", "Oth", "Ret", "Tran")
sector <- rep(1:9, c(2 ,15, 17, 8, 10, 4, 7, 10, 6))
print(R <- cor(uscomp), digits=3)
```

	Assets	Sales	MarketValue	Profits	CashFlow	Employees
Assets	1.000	0.582	0.501	0.355	0.411	0.465
Sales	0.582	1.000	0.727	0.394	0.468	0.899
MarketValue	0.501	0.727	1.000	0.576	0.623	0.733
Profits	0.355	0.394	0.576	1.000	0.989	0.351
CashFlow	0.411	0.468	0.623	0.989	1.000	0.410
Employees	0.465	0.899	0.733	0.351	0.410	1.000

After fitting,  $\widehat{Q}\widehat{Q}' + \widehat{\Psi}$  should be “close” to  $R$ . In particular we want to reproduce the large correlations in this matrix, between Employees and Sales, and between Profits and Cash Flow. Each of these will require a separate factor (column of the  $Q$  matrix), so a solution of at least two factors is probably needed, and we will try a two-factor solution<sup>1</sup>.

---

<sup>1</sup>The four-factor solution cannot be fit as there are too many parameters relative to the number of variables. The three-factor model can be fit, but but there are as many parameters as there are unique elements in  $R$ .

```
(f2 <- factanal(uscomp, factor=2, rotation="varimax"))
```

Call:

```
factanal(x = uscomp, factors = 2, rotation = "varimax")
```

Uniquenesses:

Assets	Sales	MarketValue	Profits	CashFlow	Employees
0.638	0.040	0.340	0.011	0.005	0.160

Loadings:

	Factor1	Factor2
Assets	0.544	0.258
Sales	0.961	0.194
MarketValue	0.681	0.443
Profits	0.215	0.971
CashFlow	0.294	0.953
Employees	0.904	0.154

	Factor1	Factor2
SS loadings	2.631	2.175
Proportion Var	0.438	0.363
Cumulative Var	0.438	0.801

Test of the hypothesis that 2 factors are sufficient.

The chi square statistic is 13.6 on 4 degrees of freedom.

The p-value is 0.00871

In the above output:

1. The first argument to `factanal` is in this case the name of a data frame, and by default all columns are used to define  $X$ . You can also specify the columns using a one-sided formula, like `~ Assets + Sales + MarketValue + Profits + CashFlow + Employees`, and then using a `data=uscomp` argument. By default the program will convert the sample covariance matrix  $S$  to a correlation matrix before computing. If you want to override this behavior, you can choose the matrix yourself using the `covmat` argument. If you do provide a covariance matrix the program appears to convert it to a correlation matrix.
2. The *uniquenesses* are the estimates of the diagonal elements of  $\Psi$ . In the textbook, these are called *specific variances*. The larger the specific variance, the less a particular variable is determined by the latent factors. If the uniquenesses are close to

1, then that particular variable is not well “explained” by the common factors. In this example, **Assets** and **MarketValue** are least well represented by the two common factors, while **CashFlow**, **Sales** and **Profits** are very well represented.

3. The *loadings* are an estimate of  $Q$ , in this case computed as if  $k = 2$  factors were sufficient. Another bit of factor analysis jargon is the *communality*, which is one minus the specific variance, is equal to  $\sum_j q_{ij}^2$ , and so gives the same information as the specific variance. If any entries in  $\hat{Q}$  are shown as blank, they are really just *small*: the default is to display a blank if  $|q_{jk}| < .1$ . The **factanal** function does not compute standard errors for elements of  $\hat{Q}$ , although other programs do compute standard errors.

The displayed  $\hat{Q}$  depends on the argument **rotation**, since  $Q$  is unique only up to a rotation. The default in **factanal** that we have used here is the **varimax** rotation, which attempts to make the first column of  $\hat{Q}$  as close to a vector of 0s and 1s as possible, so it maximizes

$$V \propto \sum_{j=1}^k (\text{variance of squares of scaled factor loadings for factor } j)$$

The choice **rotation="none"** selects  $Q$  so that  $Q'\Psi^{-1}Q$  is a diagonal matrix. It’s hard for me to see why this would be a meaningful choice of rotation.

4. At the foot of the loadings, the **SS loadings** are the column sum of squares  $\sum_i q_{ij}^2$ , and this will depend on the rotation. If we define  $\text{tr}(R) = p$  to be the total variance, then **SS loadings**/6 is the proportion of the total variance “explained” by each factor, **Proportion Var**. The **Cumulative Var** will generally stay less than 1 because of the specific factors. The **Cumulative Var** for all the factors does not depend on the rotation.
5. Finally a likelihood ratio test is given, with null hypothesis that two factors are sufficient versus the alternative that more than two factors are required. The small  $p$ -value suggests that the two-factor model is not adequate. We could try the three-factor model.

We try a 3-factor solution:

```
(f3 <-factanal(uscomp, factor=3, rotation="varimax", scores="regression"))
```

Call:

```
factanal(x = uscomp, factors = 3, scores = "regression", rotation = "varimax")
```

Uniquenesses:

Assets	Sales	MarketValue	Profits	CashFlow	Employees
--------	-------	-------------	---------	----------	-----------

0.513	0.091	0.321	0.008	0.005	0.005
-------	-------	-------	-------	-------	-------

Loadings:

	Factor1	Factor2	Factor3
Assets	0.337	0.217	0.571
Sales	0.809	0.187	0.468
MarketValue	0.628	0.433	0.312
Profits	0.179	0.969	0.146
CashFlow	0.227	0.944	0.229
Employees	0.968	0.156	0.184

	Factor1	Factor2	Factor3
SS loadings	2.183	2.123	0.750
Proportion Var	0.364	0.354	0.125
Cumulative Var	0.364	0.718	0.843

The degrees of freedom for the model is 0 and the fit was 0.0058

We get an exact fit because the three-factor model has as many free parameters as does a general  $\Sigma$ . The two-factor solution is not the first two columns of the three-factor solution. The uniqueness for **Assets** is smaller, but still relatively large. The cumulative variance increases from about 80% to about 84%, so it is not clear that a three-factor solution is much better than the two-factor solution.

**Factor scores** For each unit in the data there is a vector  $F$  of **factor scores**. Since  $F$  is a random variable, we would speak of predicting  $F$  rather than estimating it, as with mixed models. Now

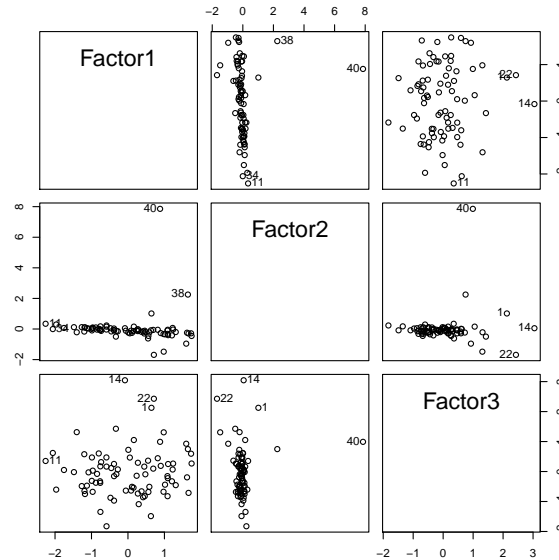
$$\text{Var} \begin{pmatrix} X - \mu \\ F \end{pmatrix} = \begin{pmatrix} \Sigma = QQ' + \Psi & Q \\ Q' & I_k \end{pmatrix}$$

and so regression prediction, which is justified by multivariate normality of  $X$  and  $F$ , of the factor score is

$$E(F|X) = Q'S_u^{-1}(X - \bar{x})$$

where  $S_u$  is the sample covariance (usually, correlation) matrix. We get the estimated factor scores by the argument **scores** = "regression" on the call to **factanal**. Here is a scatterplot:

```
library(car)
scatterplotMatrix(f3$scores, diagonal="none", reg.line=FALSE, smooth=FALSE, id.n=4)
```



Factor 1 appears to be successful at assorting the companies, but factor 2 seems to only serve to distinguish companies 38 and 40 from the others. These companies had enormous profits and cash flow relative to the other companies. Let's delete these two:

```
(f4 <-factanal(uscomp[-c(38, 40), ], factor=3, rotation="varimax", scores="regression"))
```

Call:

```
factanal(x = uscomp[-c(38, 40), ], factors = 3, scores = "regression", rotation = "varimax")
```

Uniquenesses:

Assets	Sales	MarketValue	Profits	CashFlow	Employees
0.592	0.100	0.396	0.070	0.062	0.005

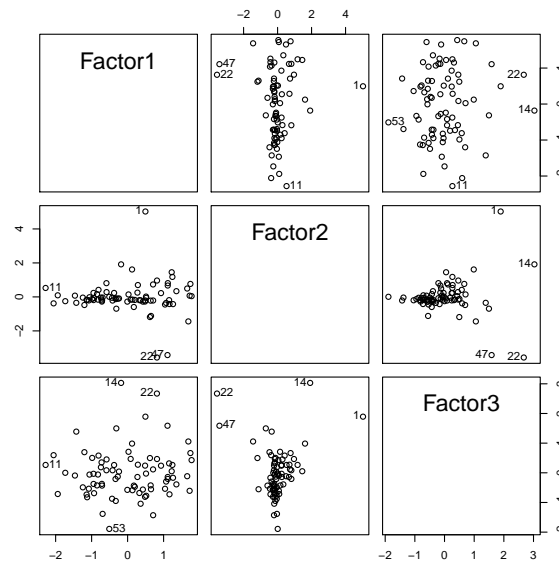
Loadings:

	Factor1	Factor2	Factor3
Assets	0.281	0.128	0.559
Sales	0.801		0.502
MarketValue	0.616	0.379	0.286
Profits		0.962	
CashFlow	0.219	0.900	0.284
Employees	0.975		0.190

	Factor1	Factor2	Factor3
SS loadings	2.103	1.910	0.763
Proportion Var	0.351	0.318	0.127
Cumulative Var	0.351	0.669	0.796

The degrees of freedom for the model is 0 and the fit was 0.0021

```
scatterplotMatrix(f4$scores, diagonal="none", reg.line=FALSE, smooth=FALSE, id.n=4)
```



The two analyses are remarkably similar: Factor 1 provides the discrimination among companies, while factor 2 separates out the few remaining companies that either have low profits and cash flow (47, and 22) and the one remaining company that is high on these, company 1.

## Intelligence

The following example was presented by Lawley and Maxwell, concerning a correlation of exam scores for  $n = 220$  male students.

```
loc<-"http://www.stat.umn.edu/~sandy/courses/8053/Data/LM.rda"
load(url(loc))
LM
```

	Gaelic	English	History	Arithmetic	Algebra	Geometry
Gaelic	1.000	0.439	0.410	0.288	0.329	0.248
English	0.439	1.000	0.351	0.354	0.320	0.329
History	0.410	0.351	1.000	0.164	0.190	0.181
Arithmetic	0.288	0.354	0.164	1.000	0.595	0.470
Algebra	0.329	0.320	0.190	0.595	1.000	0.464
Geometry	0.248	0.329	0.181	0.470	0.464	1.000

We provide without comment three solutions: PC (eigen decomposition), two-factor solution with no rotation, and two-factor solution with the varimax rotation.

```
print(f0 <- eigen(LM), digits=3)
```

\$values

```
[1] 2.733 1.130 0.615 0.601 0.525 0.396
```

\$vectors

```
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] -0.398 -0.422  0.237880  0.447  0.621 -0.1473
[2,] -0.416 -0.273  0.649785 -0.406 -0.370  0.1676
[3,] -0.313 -0.600 -0.671347 -0.099 -0.286 -0.0222
[4,] -0.447  0.389 -0.000831  0.232 -0.352 -0.6869
[5,] -0.450  0.353 -0.136085  0.402 -0.122  0.6910
[6,] -0.410  0.334 -0.227961 -0.640  0.508 -0.0205
```

```
(f1 <- factanal(factors=2, covmat=LM, n.obs=280, rotation="none"))
```

Call:

```
factanal(factors = 2, covmat = LM, n.obs = 280, rotation = "none")
```

Uniquenesses:

Gaelic	English	History	Arithmetic	Algebra	Geometry
0.510	0.594	0.644	0.377	0.431	0.628

Loadings:

Factor1	Factor2
---------	---------



Gaelic	0.553	0.429
English	0.568	0.288
History	0.392	0.450
Arithmetic	0.740	-0.273
Algebra	0.724	-0.211
Geometry	0.595	-0.132

	Factor1	Factor2
SS loadings	2.209	0.606
Proportion Var	0.368	0.101
Cumulative Var	0.368	0.469

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 2.99 on 4 degrees of freedom.  
The p-value is 0.56

```
(f2 <- update(f1, rotation="varimax"))
```

Call:

```
factanal(factors = 2, covmat = LM, n.obs = 280, rotation = "varimax")
```

Uniquenesses:

Gaelic	English	History	Arithmetic	Algebra	Geometry
0.510	0.594	0.644	0.377	0.431	0.628

Loadings:

	Factor1	Factor2
Gaelic	0.235	0.659
English	0.323	0.549
History		0.590
Arithmetic	0.771	0.170
Algebra	0.724	0.213
Geometry	0.572	0.210

	Factor1	Factor2
SS loadings	1.612	1.203

```
Proportion Var    0.269    0.201
Cumulative Var    0.269    0.469
```

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 2.99 on 4 degrees of freedom.  
The p-value is 0.56

## Officer ratings

This example consists of 14 ratings of 103 police officers by their superiors including an “overall” rating, presumably not just the average of the other 13 ratings. The data come from the Getting Started page of the SAS help files for **SAS proc factor**.

```
loc<-"http://www.stat.umn.edu/~sandy/courses/8053/Data/officerratings.csv"
data <- read.csv(url(loc),header=TRUE)
```

The column names in this data frame are very long, and to improve readability of the output, we will rename them with short names.

```
(names <- data.frame(vname=paste("Q", 1:14, sep=""),
                     description=names(data)))
```

	vname	description
1	Q1	Communication.Skills
2	Q2	Problem.Solving
3	Q3	Learning.Ability
4	Q4	Judgment.Under.Pressure
5	Q5	Observational.Skills
6	Q6	Willingness.to.Confront.Problems
7	Q7	Interest.in.People
8	Q8	Interpersonal.Sensitivity
9	Q9	Desire.for.Self.Improvement
10	Q10	Appearance
11	Q11	Dependability
12	Q12	Physical.Ability
13	Q13	Integrity
14	Q14	Overall.Rating

```
colnames(data) <- names$vname
```

The likely goal of this analysis is to convert the 13 questions into a small number of interpretable scales.

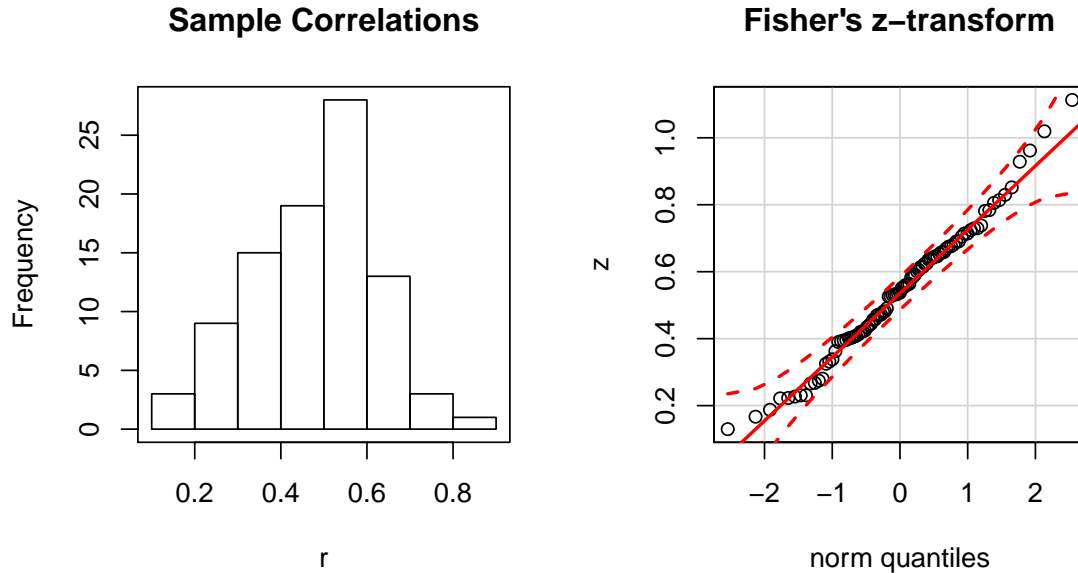
Let's look first at the correlation matrix:

```
print(R <- cor(data), digits=2)
```

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
Q1	1.00	0.63	0.55	0.55	0.54	0.53	0.44	0.50	0.56	0.49	0.55	0.22	0.51	0.68
Q2	0.63	1.00	0.57	0.62	0.43	0.50	0.40	0.44	0.41	0.39	0.45	0.32	0.38	0.58
Q3	0.55	0.57	1.00	0.49	0.62	0.52	0.27	0.19	0.57	0.40	0.51	0.23	0.31	0.59
Q4	0.55	0.62	0.49	1.00	0.37	0.40	0.62	0.61	0.48	0.23	0.55	0.35	0.59	0.66
Q5	0.54	0.43	0.62	0.37	1.00	0.73	0.26	0.17	0.60	0.42	0.56	0.43	0.39	0.58
Q6	0.53	0.50	0.52	0.40	0.73	1.00	0.22	0.13	0.53	0.48	0.49	0.49	0.33	0.59
Q7	0.44	0.40	0.27	0.62	0.26	0.22	1.00	0.81	0.49	0.27	0.61	0.38	0.75	0.61
Q8	0.50	0.44	0.19	0.61	0.17	0.13	0.81	1.00	0.37	0.26	0.54	0.22	0.69	0.58
Q9	0.56	0.41	0.57	0.48	0.60	0.53	0.49	0.37	1.00	0.45	0.60	0.38	0.57	0.67
Q10	0.49	0.39	0.40	0.23	0.42	0.48	0.27	0.26	0.45	1.00	0.51	0.38	0.41	0.57
Q11	0.55	0.45	0.51	0.55	0.56	0.49	0.61	0.54	0.60	0.51	1.00	0.45	0.65	0.77
Q12	0.22	0.32	0.23	0.35	0.43	0.49	0.38	0.22	0.38	0.38	0.45	1.00	0.38	0.44
Q13	0.51	0.38	0.31	0.59	0.39	0.33	0.75	0.69	0.57	0.41	0.65	0.38	1.00	0.67
Q14	0.68	0.58	0.59	0.66	0.58	0.59	0.61	0.58	0.67	0.57	0.77	0.44	0.67	1.00

This isn't very helpful because there are too many numbers. One possibility is to imagine that the correlations are a sample from a common distribution. Let's look at a histogram and QQplot of the correlations.

```
r <- R[lower.tri(R)]
par(mfrow=c(1, 2))
hist(r, main="Sample Correlations", xlab="r")
box()
require(car)
z <- 0.5 * log((1 + r)/(1 - r))
qqPlot(z, main="Fisher's z-transform")
```



If all the population correlations are equal to some value  $\rho$ , and the estimates are independent, then the Fisher's  $z$ -transforms should be like a random sample from  $N(.5 \log[(1 + \rho)/(1 - \rho)], 1/(n - 3))$ . The observed sd of the Fisher  $z$ -transforms is 0.196, as compared to  $\sqrt{1/(n - 3)} = 0.107$ . From the qq-plot we might conclude that a few of the correlations are larger than would be expected if the correlations were an iid sample (a test is also possible here; how would you do it?). This is also consistent with the observed sd of the correlations larger than the theoretical sd.

If all the correlations were equal and  $\rho > 0$ , then

$$R = \rho 11' + (1 - \rho)I$$

This has the form of the factor analysis variance matrix with  $Q$  with a single column given by  $\sqrt{\rho}1$ , and  $\Psi = (1 - \rho)I$ , and with specific variances all equal to  $1 - \rho$ . The eigenvalues of  $R$  are  $k - (k - 1)(1 - \rho) \approx 14 - 13(1 - \bar{r}) = 7.24$  with multiplicity 1 and  $1 - \rho \approx 0.52$  with multiplicity  $k - 1$ . The eigenvector corresponding to the first eigenvalue is proportional to 1, and the other eigenvectors are arbitrary vectors orthogonal to 1.

```
ev <-eigen(R)
print(ev$values, digits=2)
```

```
[1] 7.33 1.77 1.01 0.75 0.68 0.45 0.39 0.31 0.29 0.26 0.25 0.20 0.18 0.14
```

The largest eigenvalue is as expected, but there appears to be a second eigenvalue that might be too large for this model to be acceptable.

```
print(ev$vector[1, ], digits=2)
```

```
[1] 0.286 -0.054 -0.330 -0.210 -0.181 -0.429 0.144 -0.166 0.456 0.447 -0.019 -0.193
[13] -0.207 0.089
```

Without a test and/or standard errors, it's hard to judge if this is proportional to a vector of 1s or not.

Let's try factor analysis.

```
(f3 <- factanal(~ ., data=data, factors=3, rotation="varimax"))
```

Call:

```
factanal(x = ~., factors = 3, data = data, rotation = "varimax")
```

Uniquenesses:

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
	0.371	0.285	0.407	0.364	0.304	0.332	0.196	0.157	0.407	0.623	0.305	0.689	0.267	0.204

Loadings:

	Factor1	Factor2	Factor3
Q1	0.449	0.357	0.548
Q2	0.295	0.261	0.748
Q3	0.583		0.497
Q4	0.267	0.554	0.508
Q5	0.791		0.255
Q6	0.744		0.337
Q7	0.189	0.864	0.148
Q8		0.875	0.279
Q9	0.645	0.367	0.204
Q10	0.544	0.213	0.190
Q11	0.599	0.549	0.185
Q12	0.491	0.258	
Q13	0.381	0.760	
Q14	0.616	0.535	0.361

	Factor1	Factor2	Factor3
SS loadings	3.751	3.439	1.898
Proportion Var	0.268	0.246	0.136
Cumulative Var	0.268	0.514	0.649

Test of the hypothesis that 3 factors are sufficient.  
 The chi square statistic is 73.98 on 52 degrees of freedom.  
 The p-value is 0.0242

```
(f4 <- factanal(~ ., data=data, factors=4, rotation="varimax"))
```

Call:

```
factanal(x = ~., factors = 4, data = data, rotation = "varimax")
```

Uniquenesses:

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
0.244	0.369	0.408	0.224	0.311	0.321	0.174	0.148	0.418	0.512	0.305	0.581	0.276	0.200

Loadings:

	Factor1	Factor2	Factor3	Factor4
Q1	0.324	0.213	0.606	0.488
Q2	0.278	0.227	0.695	0.140
Q3		0.472	0.562	0.219
Q4	0.574	0.266	0.603	-0.112
Q5		0.693	0.347	0.290
Q6		0.676	0.403	0.243
Q7	0.873	0.204	0.143	
Q8	0.874		0.246	0.158
Q9	0.344	0.541	0.286	0.300
Q10	0.181	0.397	0.166	0.520
Q11	0.530	0.519	0.225	0.307
Q12	0.269	0.586		
Q13	0.742	0.328	0.134	0.220
Q14	0.515	0.498	0.410	0.346

Factor1	Factor2	Factor3	Factor4
---------	---------	---------	---------

SS loadings	3.367	2.790	2.242	1.111
Proportion Var	0.240	0.199	0.160	0.079
Cumulative Var	0.240	0.440	0.600	0.679

Test of the hypothesis that 4 factors are sufficient.  
The chi square statistic is 53.61 on 41 degrees of freedom.  
The p-value is 0.0897

The three-factor solution is inadequate, while the four-factor solution provides a reasonable approximation to the correlation matrix, explaining about 70% of the variability.

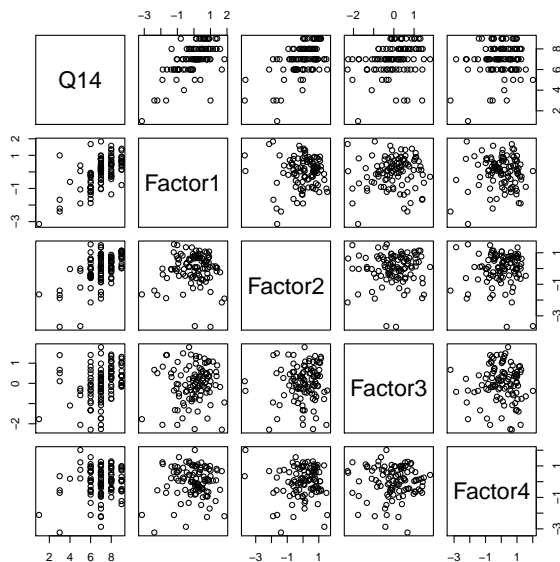
One interpretation of the loadings comes from computing the correlation between  $X$  and  $F$ :

$$\text{Cov}(X, F) = \text{Cov}(QF + U, F) = Q$$

With  $X$  in correlation scale,  $\hat{Q}$  estimates correlations. At least with this rotation the overall rating Q14, has correlation of about .5 or less with each of the factors. correlates with any of the individual ratings.

As an exercise, let's refit omitting Q14, the overall score, and look at a scatterplot matrix of Q14 and the factor scores for the 4-factor solution:

```
f5 <- factanal(~ . - Q14, data, factors=4, scores="regression")
pairs(cbind(Q14=data$Q14, f5$scores))
```



```
summary(lm(data$Q14 ~ f5$scores))
```

Call:

```
lm(formula = data$Q14 ~ f5$scores)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.0447	-0.5038	0.0656	0.5447	2.0988

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.0000	0.0751	93.25	< 2e-16
f5\$scoresFactor1	0.9356	0.0820	11.41	< 2e-16
f5\$scoresFactor2	0.8460	0.0793	10.67	< 2e-16
f5\$scoresFactor3	0.3658	0.0853	4.29	4.3e-05
f5\$scoresFactor4	0.2068	0.0779	2.65	0.0093

Residual standard error: 0.762 on 98 degrees of freedom

Multiple R-squared: 0.751, Adjusted R-squared: 0.74

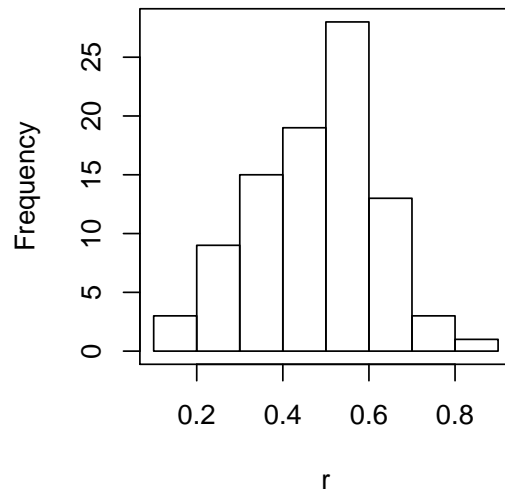
F-statistic: 73.7 on 4 and 98 DF, p-value: <2e-16

Here is a simulation for comparison:

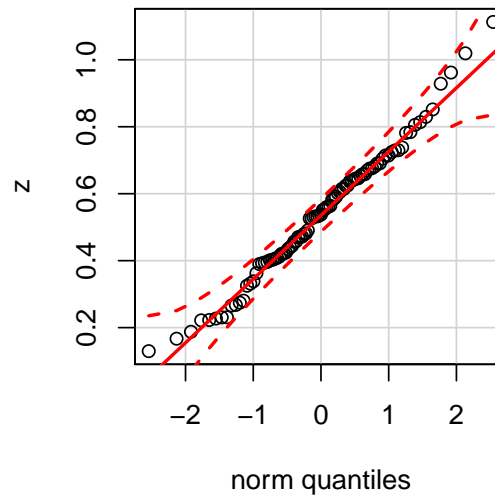
```
r <- 0.48
p <- 14
S <- r * outer(rep(1, p), rep(1, p)) + diag(rep(1-r, p))
library(MASS)
set.seed(44)
X <- mvrnorm(163, rep(0, p), S)
r <- R[lower.tri(R)]
par(mfrow=c(1, 2))
hist(r, main="Sample Correlations, Simulation", xlab="r")
box()
z <- 0.5 * log((1 + r)/(1 - r))
qqPlot(z, main="Fisher's z-transform, Simulation")
```



**Sample Correlations, Simulation**



**Fisher's z-transform, Simulation**



```
esim <- eigen(R)
esim$values
```

```
[1] 7.3322 1.7730 1.0053 0.7510 0.6783 0.4525 0.3876 0.3078 0.2852 0.2634 0.2458 0.2009
[13] 0.1762 0.1407
```

```
esim$vectors[, 1]
```

```
[1] 0.2865 0.2601 0.2512 0.2776 0.2593 0.2519 0.2621 0.2401 0.2828 0.2258 0.3041 0.1998
[13] 0.2818 0.3321
```

```
(f3 <- factanal(X, data=X, factors=1, rotation="varimax"))
```

Call:

```
factanal(x = X, factors = 1, data = X, rotation = "varimax")
```

Uniquenesses:

```
[1] 0.558 0.386 0.512 0.432 0.474 0.485 0.387 0.473 0.431 0.452 0.544 0.465 0.475 0.474
```

Loadings:

```
      Factor1
[1,] 0.665
[2,] 0.784
[3,] 0.699
[4,] 0.753
[5,] 0.725
[6,] 0.718
[7,] 0.783
[8,] 0.726
[9,] 0.754
[10,] 0.740
[11,] 0.676
[12,] 0.731
[13,] 0.725
[14,] 0.725
```

```
      Factor1
SS loadings      7.452
Proportion Var   0.532
```

Test of the hypothesis that 1 factor is sufficient.

The chi square statistic is 66.73 on 77 degrees of freedom.

The p-value is 0.792

```
(f4 <- factanal(X, data=X, factors=2, rotation="varimax"))
```

Call:

```
factanal(x = X, factors = 2, data = X, rotation = "varimax")
```

Uniquenesses:

```
[1] 0.460 0.385 0.502 0.414 0.453 0.486 0.378 0.378 0.423 0.453 0.532 0.457 0.448 0.452
```

Loadings:

```
      Factor1 Factor2
```

[1,]	0.329	0.657
[2,]	0.639	0.455
[3,]	0.474	0.523
[4,]	0.666	0.378
[5,]	0.649	0.356
[6,]	0.542	0.469
[7,]	0.661	0.431
[8,]	0.382	0.690
[9,]	0.639	0.410
[10,]	0.585	0.453
[11,]	0.453	0.513
[12,]	0.625	0.391
[13,]	0.660	0.340
[14,]	0.466	0.575

	Factor1	Factor2
SS loadings	4.482	3.297
Proportion Var	0.320	0.236
Cumulative Var	0.320	0.556

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 46.63 on 64 degrees of freedom.  
The p-value is 0.95