STAT 755 Lab Report 4

Jiwan Bhandari

Introduction

Factor Analaysis like principal component analysis is an attempt to approximate the covariance matrix Σ. The factor model postulates that X is linearly dependent upon a few observable random variables $F_1, F_2, \ldots. F_m$ called common factors and p additional sources of variations $\epsilon_1 \epsilon_2, \ldots. \epsilon_p$ called errors or sometimes specific factors. In particular, the factor analysis model is

$$\underset{p*1}{X - \mu} = \underset{p*m}{L}\,\underset{m*1}{F} + \underset{p*1}{\epsilon}$$

, where L is the matrix of factor loadings or matrix of coefficients of linear transformation, F is matrix of common Factors. Constraining this linear relation such that $E(F) = 0$, $Cov(F) = I$, $E(\epsilon) = 0$ and $Cov(\epsilon) = \Psi$ where Ψ is a diagonal matrix, we get a factor model called orthogonal factor model. Imposing and applying these constraints on the preceeding linear transformation equation, gives a new linear equation that allows us to express the Σ, the variance-covariance matrix in terms of loadings and errors.

$$\Sigma = LL' + \Psi$$

Dataset

The dataset (n= 50, p=7) used in this exercise called salesperson dataset. The variables 3 variables the measure the performance of the sales staffs : growth of sales, profitability of sales, and new-acccount sales. These measure have been converted to scale, on which 100 indicates "avergae" performance. The other 3 variables are test scores which purported to measure creativity, mechanical reasoning, abstract reasoning, and mathematical ability respectively.

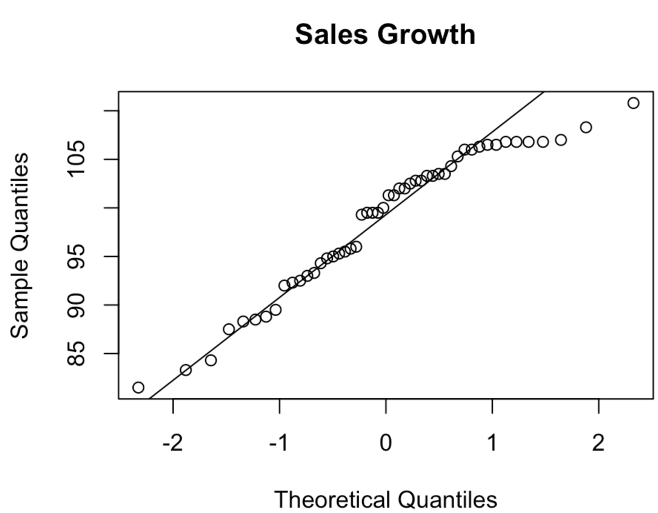
Correlation Matrix

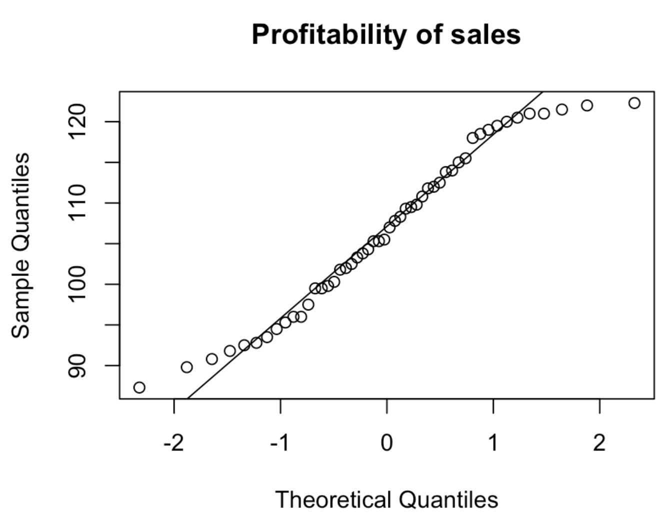
The units of the variables in the dataset aren't exactly commensurate and the variances differe significantly. In order to have few variables with large variance unduly influencing the determination of factor loading it's desirable to work with scales variances or correlations coefficients. The correlation matrix of the dataset is presented in the following table.

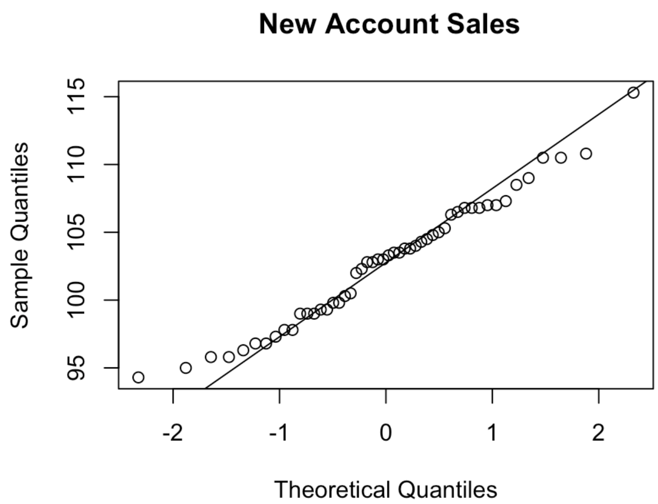
	Sales_Growth	Profitability	Account_Sales	Creativity_test	Mech_Reasoning_Test	Abs_reasoning_test	Math_test
Sales_Growth	1.000	0.926	0.884	0.572	0.708	0.674	0.927
Profitability	0.926	1.000	0.843	0.542	0.746	0.465	0.944
Account_Sales	0.884	0.843	1.000	0.700	0.637	0.641	0.853
Creativity_test	0.572	0.542	0.700	1.000	0.591	0.147	0.413
Mech_Reasoning_Test	0.708	0.746	0.637	0.591	1.000	0.386	0.575
Abs_reasoning_test	0.674	0.465	0.641	0.147	0.386	1.000	0.566
Math_test	0.927	0.944	0.853	0.413	0.575	0.566	1.000

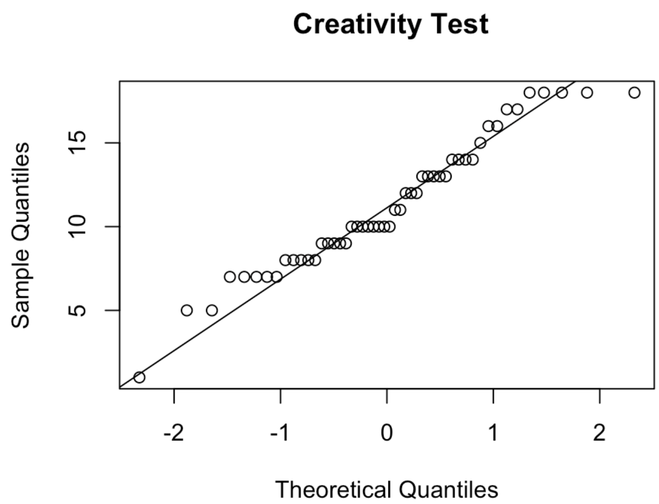
Normality Test

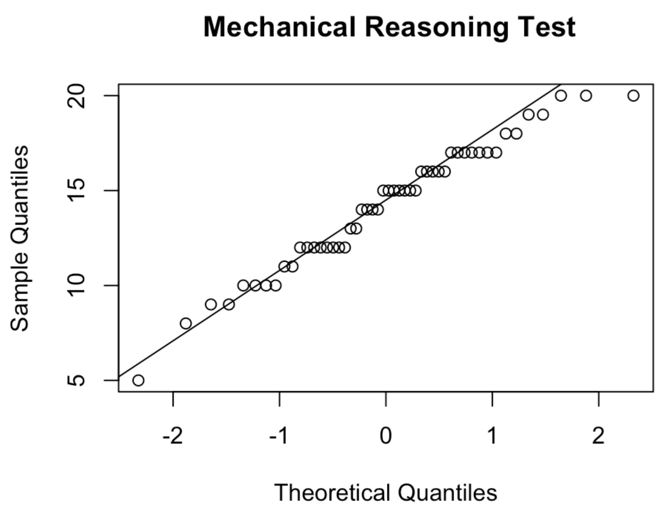
Normality of the data is very important property. In factor analysis normality of the underlying data allows us to estimate the loadings and factor using maximum likelihood estimation. So below we attempt to test normality of the data. First, we test the univariate normality using quantile-quantile plot.

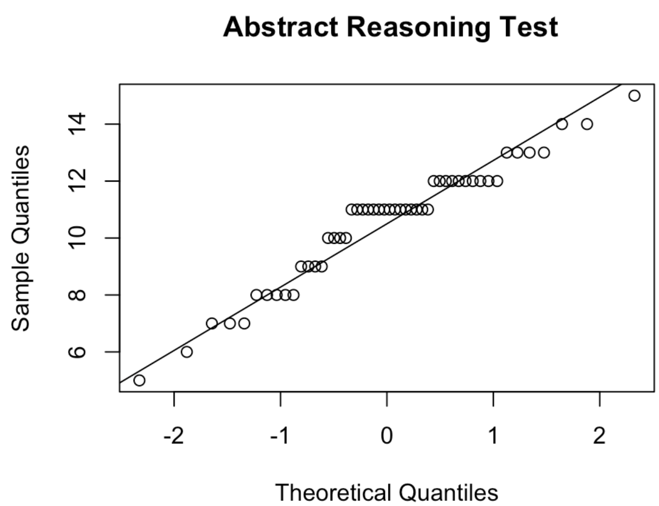


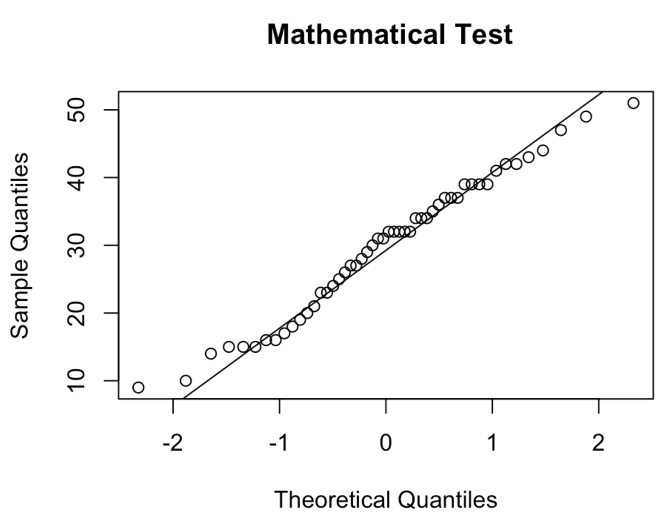












From the qqplots, it seems like profitability of sales is normally distributed. Moving on, we test multivariate normality of the dataset using Kolmogorow-Simrnow test. Specifically, we test whether the statistical distances have chi-square distribution.

```
D <- as.matrix(D)
df <- ncol(D)
cov_mat <- cov(D)
Sigma_inv <- solve(cov_mat)
DS <- sweep(D,2,colMeans(D))
d = rowSums(DS%*%Sigma_inv*DS)
testResult <- ks.test(d,"pchisq",df)
```

H_0 : Statistical distances have chi-square distribution.
H_1 : Statistical distances don't have chi-square distribution.
$Test\ Statistics$: Reject H_0 if $p - value < \alpha(0.05)$ otherwise don't reject H_0.
$Conclusion$: Since 0.11 > 0.05 and in-fact p-value in this case greater than any common level of α's we fail to reject the null hypothesis. Thus, the data does have chi-square distribution meaning the underlying data does have normal distribution.

Factor Analysis

In factor analysis, the number of common factor is usually determined by a priori considerations such as by theory or the work of other researchers. However, if it's not possible to determine the number of factors by domain knowledge the choice can be based o the estimated eigen values in much the same manner as with principal component analysis. The frequently encountered apporach is to choose the number of common factors to be equal to the numberof eighen values of correlation matrix greater than 1. This is usually a rule of thumb and shouldn't be applied indiscriminately. It's always a good idea to iteratively experiment several values and pick the one that is best. The estimated factor loadings, communalities, specific variances and proportion of total vsample varince explained by each factor for $m = 1, 2, 3$ factor solutions are shown in the following 3 tables. The tables show the factor solutions using 3 factor rotations - $'none'$, $'varimax'$, $'promax'$.

	Factor Loadings (1 -factor solution)				
	Unrotated	Rot.(varimax)	Rot.(promax)	Communalities	Uniqueness
Variables	F_1	F_1	F_1		
Sales Growth	0.98	0.98	0.98	0.95	0.05
Profitability of sales	0.96	0.96	0.96	0.92	0.08
New Account Sales	0.9	0.9	0.9	0.81	0.19
Creativity Test	0.57	0.57	0.57	0.32	0.68
Mechanical Reasoning Test	0.71	0.71	0.71	0.51	0.49
Abstract Reasoning Test	0.61	0.61	0.61	0.38	0.62
Mathematical Test	0.95	0.95	0.95	0.91	0.09
Commulative Variance	**0.69**	**0.69**	**0.69**		

	Factor Loadings (2 -factor solution)							
	Unrotated		Rot.(varimax)		Rot.(promax)		Communalities	Uniqueness
Variables	F_1	F_2	F_1	F_2	F_1	F_2		
Sales Growth	0.7	0.67	0.85	0.45	0.9	0.11	0.93	0.07
Profitability of sales	0.67	0.69	0.87	0.42	0.93		0.93	0.07
New Account Sales	0.8	0.49	0.72	0.6	0.69	0.35	0.88	0.12
Creativity Test	0.98	-0.17	0.15	0.99	-0.12	1.06	1	0
Mechanical Reasoning Test	0.65	0.31	0.5	0.53	0.45	0.36	0.53	0.47
Abstract Reasoning Test	0.25	0.57	0.62		0.73	-0.23	0.39	0.61
Mathematical Test	0.56	0.81	0.95	0.28	1.06	-0.14	0.97	0.03
Commulative Variance	**0.48**	**0.8**	**0.51**	**0.8**	**0.57**	**0.78**		

	Factor Loadings (3 -factor solution)										
	Unrotated			Rot.(varimax)			Rot.(promax)			Communalities	Uniqueness
Variables	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3		
Sales Growth	0.9	0.38		0.79	0.37	0.44	0.77		0.2	0.96	0.04
Profitability of sales	0.78	0.6		0.91	0.32	0.18	1.12		-0.17	0.97	0.03
New Account Sales	0.93	0.2		0.65	0.54	0.44	0.46	0.38	0.26	0.91	0.09
Creativity Test	0.73	-0.12	0.67	0.26	0.96		-0.17	1.14	-0.11	1	0
Mechanical Reasoning Test	0.69	0.22	0.17	0.54	0.47	0.21	0.46	0.34		0.55	0.45
Abstract Reasoning Test	0.76	-0.13	-0.64	0.3		0.95		-0.11	1.08	1	0
Mathematical Test	0.76	0.61	-0.11	0.92	0.18	0.3	1.14	-0.23		0.96	0.04
Commulative Variance	**0.63**	**0.78**	**0.91**	**0.45**	**0.7**	**0.91**	**0.51**	**0.75**	**0.94**		

Residuals

We calcualted the residuals matrix for each value of common factor ($m = 1\ to\ 3$) using the following.

```
residual <- cor(D) - (var_fa$loading %*% t(var_fa$loadings) + diag(var_fa$uniquenesses))
```

Upon calcuating the residual matrices for each value across each rotation, we filtered the residual matrices as $(R - LL' - \Psi) > 0.01$. The result of this filtering showed, for each $(1, 2, 3)$ choice of m $no - rotation$ and $varimax$ rotation had same and lower values but $promax$ rotation had bigger values. We got the best value with $m = 3$ and $rotation = none\ or\ varimax$.

Cummulative variance

The cumulative propotion of the total (standarized) sample variance increases as we increase the number of factors. For $m = 1$, the commulative variance across each rotation type is 0.69. For $m = 2$ the cummulative variance for each rotation type is ~ 0.8. For $m = 3$, for $promax\ rotation$ we have cummulative variance of 0.94 and for other rotation type it is 0.91. So, clearly the proportion of the total variance explained by the 3-factor solution is appreicably larger than 2 and 1 factor solutions.

Communalities

The communalities for each indicate the percentage of sample variances of each variables accounted for by the factors. In general, for each value $(1, 2, 3)$ of common factor, we have high communalities. However, the communalities do increase with m. For instance, for $m = 1$, the communality of variable $Creativity\ test$ is 0.32 but for $m = 2$ the communality goes to 1 meaning that with 2 factors we captured almost all the variance in $Creativity\ test$. Overall, with $ m=3$ we have best/highest communalities values.

Uniqueness

Uniqueness are the specific variances or the error variances. For good fit of factor model it's desirable to have low uniqueness values. From the tables above we observe that the uniqueness values decrese as we increase the number of common factors. $m = 3$ has the lowest uniqueness values meaning that factors have captured most of variablility.

Analysing loadings

For **m=1** i.e 1-factor solution, the factor loadings are unaffected by rotation. Variables $Sales\ Growth$, $Profitability\ of\ sales$, $New\ Account\ Sales$, $Mathematical\ Test$ have high factor loadings on F_1 while $Creativity\ Test$, $Mechanical\ Reasoning\ Test$ and $Abstract\ Reasoning\ Test$ have relatively low loading on F_1. The communalities values and commulative variance indicate that we do need additional factors to capture the variability of certain variables.

For **m=2** i.e 2-factor solution, for $unrotated factor\ loadings$ almost all variables load highly on the first factor except for $Abstract\ Reasoning\ Test$. In general, the factor loadings don't indicate an apparent groupings with unrotated factor loadings. $Varimax\ \&\ promax$ rotations appear to crank up the factor loadings overall but more importantly they have injected more contrast. From the rotated loadings it's now more apparent that $sales\ growth, profitability\ of\ sales$ and $Mathematical\ Test$ load highly on F_1. It's also obvious that $Creativity\ test$ loads almost entirely on F_2. So, we clearly have $two\ groups$ and the remaining variables align highly towards F_1 than F_2. F_1 could be interpreted as $Quantitative\ performance factor$ because it captures the relationship between mathematical test score of a candidate and sales performance of the candidate. F_2 could be interpreted as $creativity factor$ as it mostly dominated by $creativity\ test$.

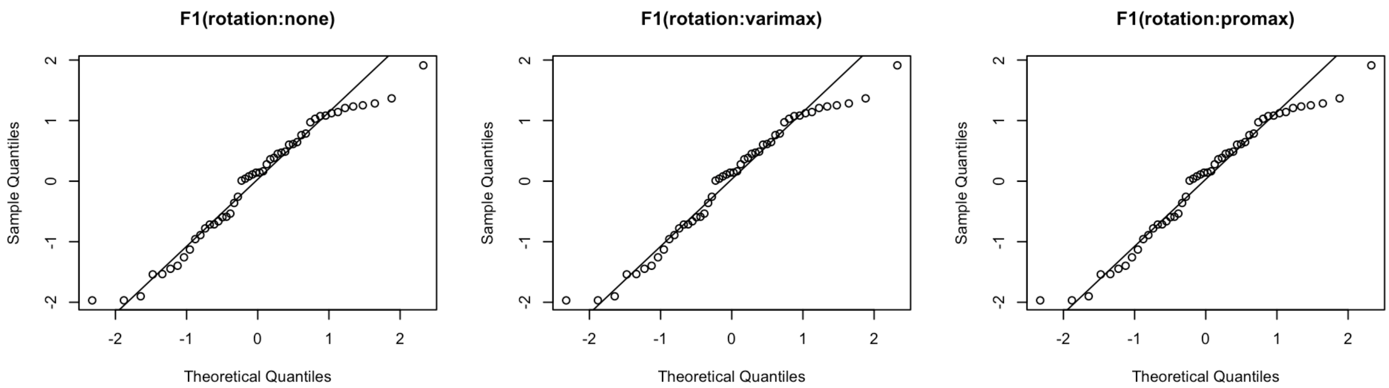
For **m=3**,i.e. 3-factor solution $proxmax$ rotation clearly caputres more total standarized sample variance than $unroated$ and $varimax$ rotated loadings. With 3 factors we do have variables that have no bearing on some of the factors. For $unrotated$ factors loadings we notice very little contrast in F_1 to make out the groups. F_2 and F_3 do indicate groups with F_2 mostly influenced by $Profitability\ of\ sales$ and $Mathematical\ test$ while F_3 is mostly about $Creativity\ Test$ and $Abstract\ Resoning\ Test$. F_3 can also be interpretted as $factor\ of\ tests$ as it only captures test variables. $Varimax$ rotated factor loadings have more contrast than $unrotated$ loadings. F_1 is dominated more by $Sales\ Growth$, $Profitability\ of\ Sales$, $Mathematical\ Test$, F_2 is mostly about the variance in $creativitytest$ and F_3 is mostly about variance in $Abstract\ Reasoning\ Test$. $Promax$ rotated loading give a slightly better picture of groups. F_1 is influenced by $Mathematical\ Test$, $Profitability\ of\ Sales$ and $Sales\ Growth$. F_2 and F_3 are mostly about $Creativity\ test$ and $Abstract\ Reasoning\ Test$. So, for $3 - factor$ solution rotated factor loadings suggest $3 groups$.

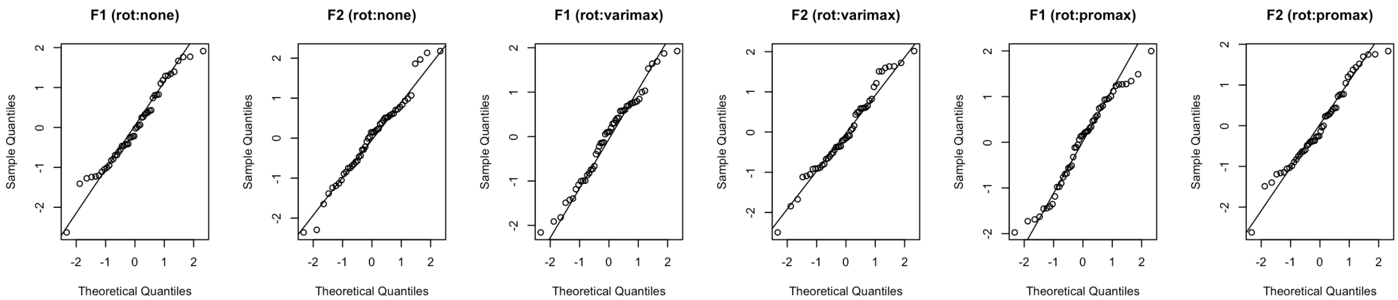
Optimal Number of factors

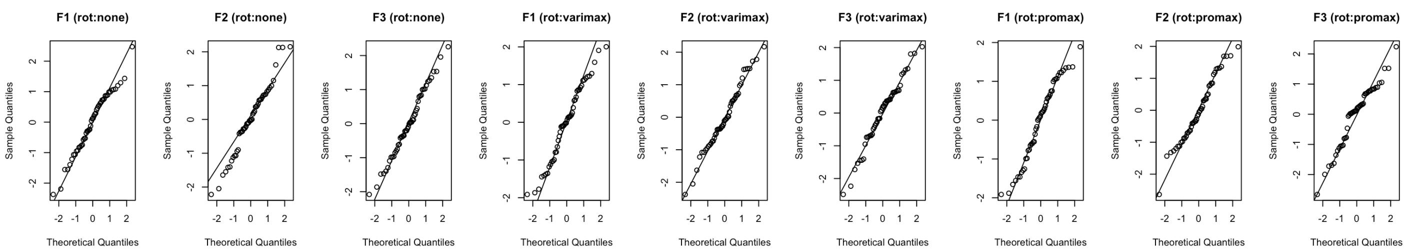
Based on the above discussion on residuals, cummulative variances, communalities, uniqueness, and loadings analysis, I think it's warranted that 3 factors are optimal rather than $2\ or\ 1$. 3 factors, overall, give the maximum cummulative variance, lowest residuals, highest communalities. The amount of variance embodied by the final factor is $\sim 20\%$ which implies that F_3 does infact account for significant amount variance. With 3 factors we see more contrast in the loadings implying underlying variable groupings which simplifies the interpretation of loadings.

Normality of Scores

Below we test the normality of factor scores for 1-factor, 2-factor and 3-factor solutions using qqplots.



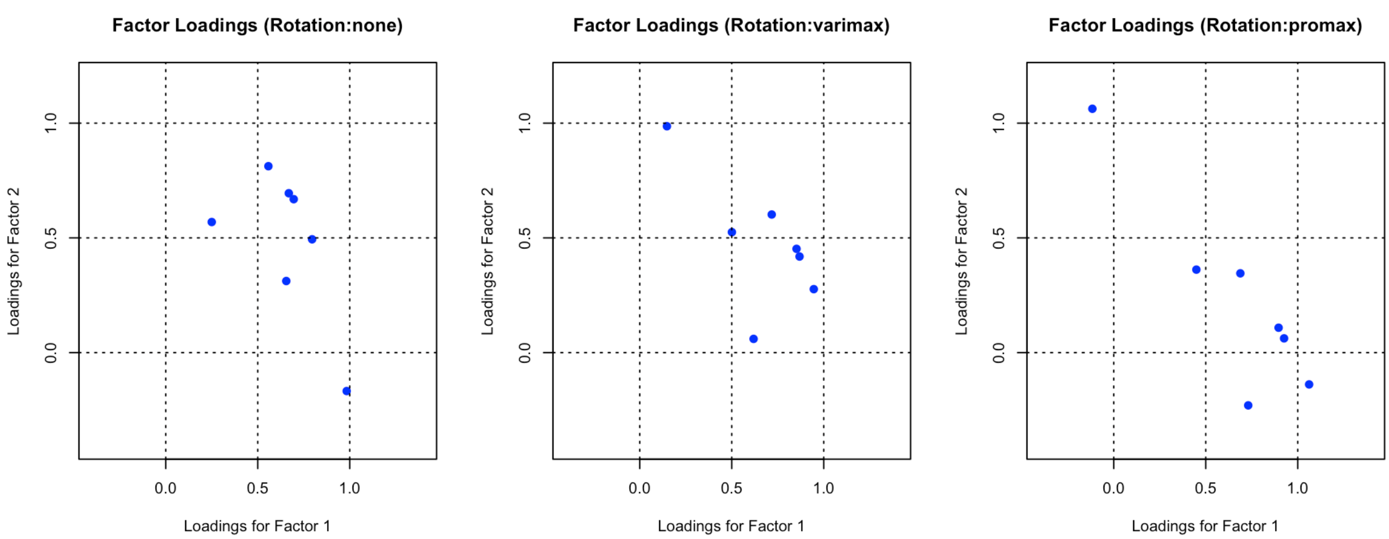




From the qqplots, since the scores are linear and fall mainly in the $qqline$, we can infer that the scores are normally distributed.

FA with 2 common factors

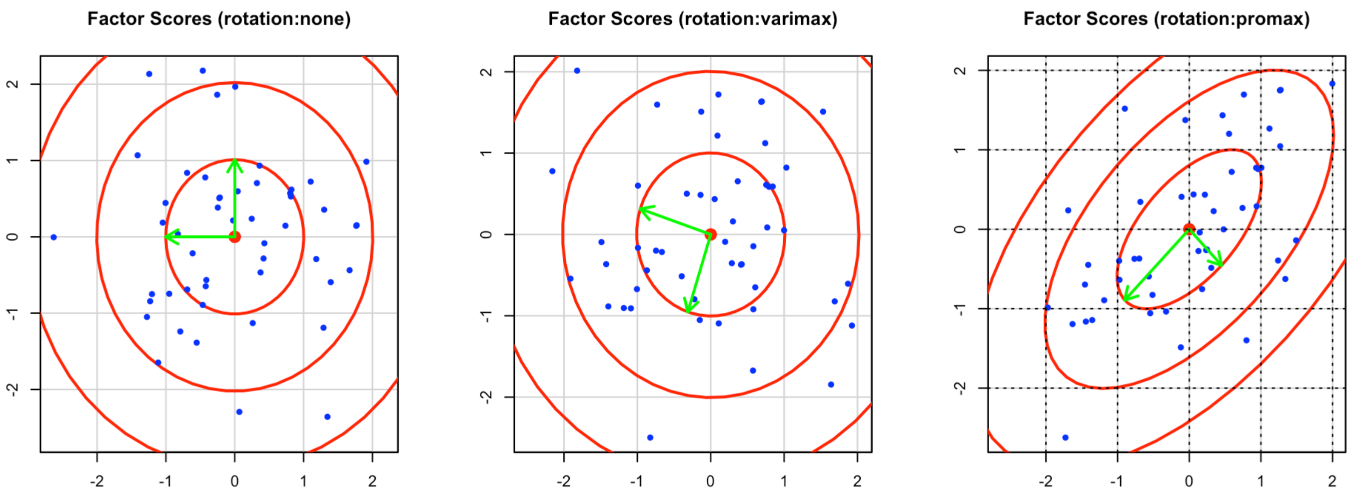
The second table above list the factor loadings, communalities, specific variances and cummulative variances for the 2 factor solution using 'promax' rotation. Below we plot the factor loadings for each rotation type.



As evident from the above plots, rotations (varimax and promax) do make it slightly easier to notice groups than unrotated loadings. Grouping the variables is to some extent is also a subjective matter and based on the promax rotation plot I would say 3 groups is more appropriate than 2.

We have already analyzed loadings for 2-factor solution above.

In the following, we plot the factor scores for each rotation type.



The normalized scores scatter plots above don't show any extreme outliers ($> 3 * \sigma \ or < -3 * sigma$). The scores also show normality in that most of the data is concerntrated inside the smaller circles.

Below we show the residual matrices for each rotation type.

$$Residual_{norotation} = \begin{pmatrix} 0.00 & -0.00 & 0.00 & 0.00 & 0.04 & 0.12 & -0.00 \\ -0.00 & 0.00 & -0.03 & 0.00 & 0.09 & -0.10 & 0.01 \\ 0.00 & -0.03 & 0.00 & 0.00 & -0.04 & 0.16 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.00 & -0.00 & 0.00 \\ 0.04 & 0.09 & -0.04 & -0.00 & 0.00 & 0.04 & -0.04 \\ 0.12 & -0.10 & 0.16 & -0.00 & 0.04 & 0.00 & -0.04 \\ -0.00 & 0.01 & 0.01 & 0.00 & -0.04 & -0.04 & 0.00 \end{pmatrix}$$

$$Residual_{varimax} = \begin{pmatrix} 0.00 & -0.00 & 0.00 & 0.00 & 0.04 & 0.12 & -0.00 \\ -0.00 & 0.00 & -0.03 & 0.00 & 0.09 & -0.10 & 0.01 \\ 0.00 & -0.03 & 0.00 & 0.00 & -0.04 & 0.16 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.00 & -0.00 & 0.00 \\ 0.04 & 0.09 & -0.04 & -0.00 & 0.00 & 0.04 & -0.04 \\ 0.12 & -0.10 & 0.16 & -0.00 & 0.04 & 0.00 & -0.04 \\ -0.00 & 0.01 & 0.01 & 0.00 & -0.04 & -0.04 & 0.00 \end{pmatrix}$$

$$Residual_{promax} = \begin{pmatrix} 0.12 & 0.09 & 0.23 & 0.56 & 0.27 & 0.04 & -0.01 \\ 0.09 & 0.07 & 0.18 & 0.58 & 0.31 & -0.20 & -0.03 \\ 0.23 & 0.18 & 0.28 & 0.41 & 0.20 & 0.22 & 0.17 \\ 0.56 & 0.58 & 0.41 & -0.15 & 0.26 & 0.48 & 0.68 \\ 0.27 & 0.31 & 0.20 & 0.26 & 0.19 & 0.14 & 0.15 \\ 0.04 & -0.20 & 0.22 & 0.48 & 0.14 & -0.20 & -0.24 \\ -0.01 & -0.03 & 0.17 & 0.68 & 0.15 & -0.24 & -0.18 \end{pmatrix}$$

The residual matrix for no rotation and varimax rotation are the same. 'promax' rotation seems to yeild the highest values in the residual matrix. It's also worth noting that the diagonal entires for no-rotation and varimax rotation are 0 while for 'promax' rotation they are not. We want very low values in residual matrices, ideally zeros.