ASSIGNMENT: MULTIVARIATE ANALYSIS OF ADATASET IN R

I0P16a: Applied Multivariate Statistical Analysis

Jiangli Gui

R0867437

Introduction

The present study aimed to investigate the impact of different light regimes on the fruit quality and yield of tomatoes (*Solanum lycopersicum cv. Moneymaker*) using a two-group design. The study design is illustrated in the accompanying figure. To extract the metabolic fingerprints of the tomato fruits, proton nuclear magnetic resonance (^{1}H-NMR) was employed on samples taken on days 8, 15, 28, and 55, with three replicates per group. Multivariate data analysis was performed on the NMR data, following preprocessing, to ascertain the grouping structure and identify the key compounds affected by the varying light treatments.

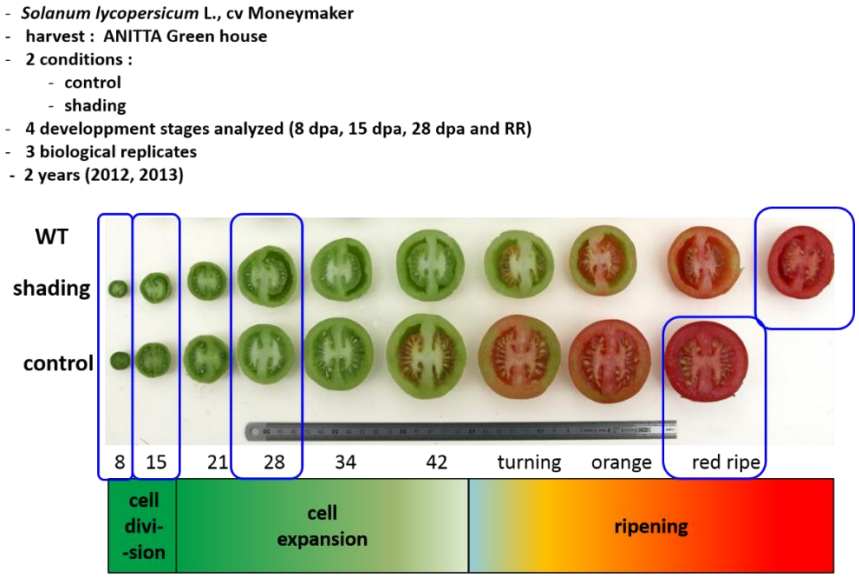

Figure 1. The study design (https://nmrprocflow.org/ex1)

Description of the problem

The present study aims to utilize proton nuclear magnetic resonance (NMR) spectra to distinguish between the "control" and "shadow" groups of tomatoes (*Solanum lycopersicum cv. Moneymaker*) and to identify the key metabolites affected by shadowing. Additionally, the study aims to investigate the metabolic changes that occur during the different growth stages of tomato plants. The high-dimensional nature of NMR spectra presents a typical discrimination problem, however, by utilizing literature, databases, and spiking, the corresponding chemicals of chemical shifts on the NMR spectra have been identified. By pinpointing the chemical shifts that are highly responsible for dissimilarity of known clusters, they can be linked to specific compounds.

Description of the dataset

The dataset used in this study includes 42 observations, each corresponding to a single tomato. The dataset includes two categorical variables, "Condition" and "Stage," and 38 numerical variables derived from NMR spectra after ppm calibration, baseline correction, spectra alignment, and bucketing. Some chemical shifts with no signatures have been removed, and the bucketing data has also undergone quantitative calibration, resulting in all variables being on the same scale in the unit of mg/DW. The data is considered to be clean for multivariate analysis without missing data.

The variables in the dataset are described as follows:

- Samplecode: corresponds to each single tomato
- Condition: binary variable, "Control" or "Shadow"
- Stage: categorical variable, 4 stages during the growth, "J08", "J15", "J28" or "J55" corresponds to day8, day15, day28 and day55 separately.
- B0_9509, B1_0206, B1_0442, B1_4839, B1_5269, B1_6215, B1_7200, B1_8600, B2_2965, B2_3430, B2_4450, B2_5720, B2_7995, B2_8925, B3_2066, B3_6060, B4_1155, B4_1645, B4_3025, B4_6508, B5_1040, B5_1870, B5_2385, B5_2710, B5_3085, B5_3925, B5_4175, B5_8275, B6_1525, B6_4235, B6_5231, B6_9035, B7_4290, B7_5585, B7_6185, B7_8650, B7_9539, B9_1275: Numerical variables, the bucketing value at each chemical shift location (ppm) with step 0.04. The character B indicates it is a bucketing value, the number before underscore is the number on the ones place on x-axis (ppm) on NMR spectra, while those after underscore are numbers after decimal point. So for example, the variable B0_9509 represents the integration value at 0.9509 ppm from NMR spectra with step 0.04.

Choice of methodology and motivation

The methodology chosen for this study is a combination of principal component analysis (PCA) and factor analysis, unsupervised k-means clustering, and supervised Partial Least Squares Discriminant Analysis (PLS-DA) and Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) with tree and random forest models. The motivation for using these methods is to reduce the dimensionality of the data using PCA and factor analysis, understand the grouping patterns and detect outliers. K-means clustering is used to extract the clustering structure from the data, and PLS-DA and OPLS-DA are used for powerful discriminate capacity and finding the feature metabolites that are significantly responsible for the dissimilarity among different clusters by Variable Importance in Projection (VIP). Additionally, tree and random forest models are used to find key compounds in the metabolic fingerprint of tomatoes during growth. Cross validation is applied to improve the performance of the models while not separating the data set partially for training and testing, as the focus of the study is not on prediction for new instances.

Results

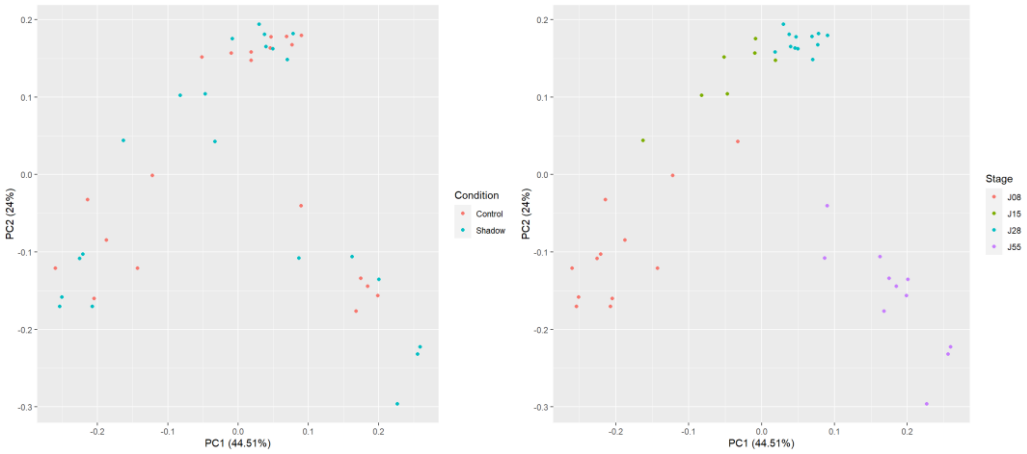

Figure 2. Two PCA plots based on scaling. Labeled by treatments, control versus shadow (left); Labeled by stages, J08, J15, J28 and J55 (right)

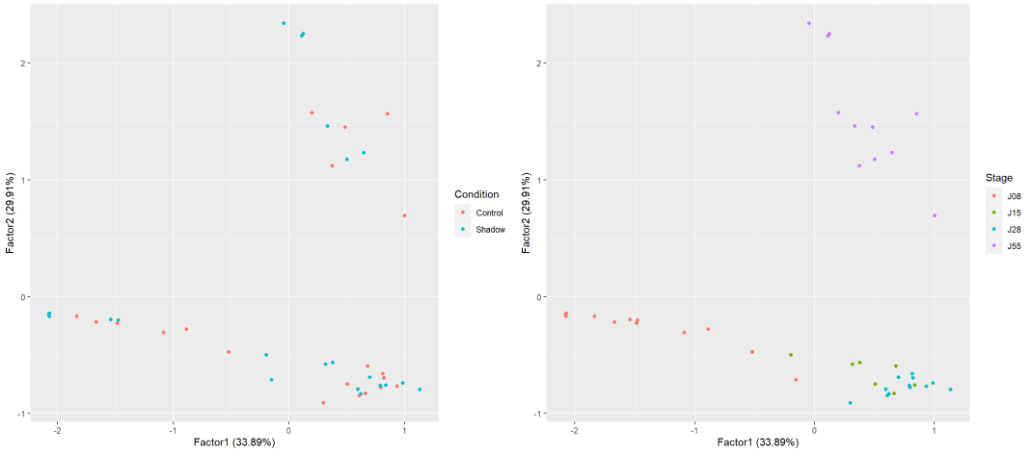

Figure 3. Two FA plots after varimax rotation. Labeled by treatments, control versus shadow (left); Labeled by stages, J08, J15, J28 and J55 (right)

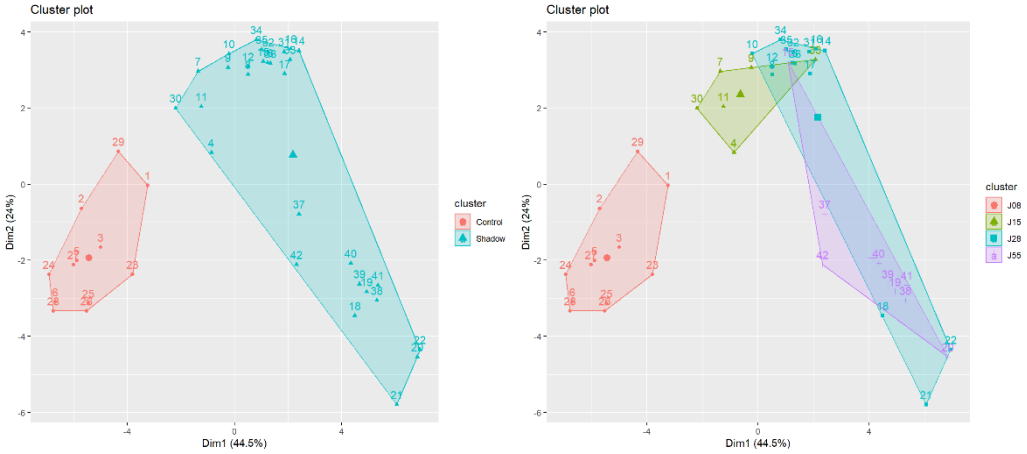

Figure 4. Two K-means plots. 2-means: Labled by treatments, control versus shadow (left); 4-means: Labled by stages, J08, J15, J28 and J55 (right)

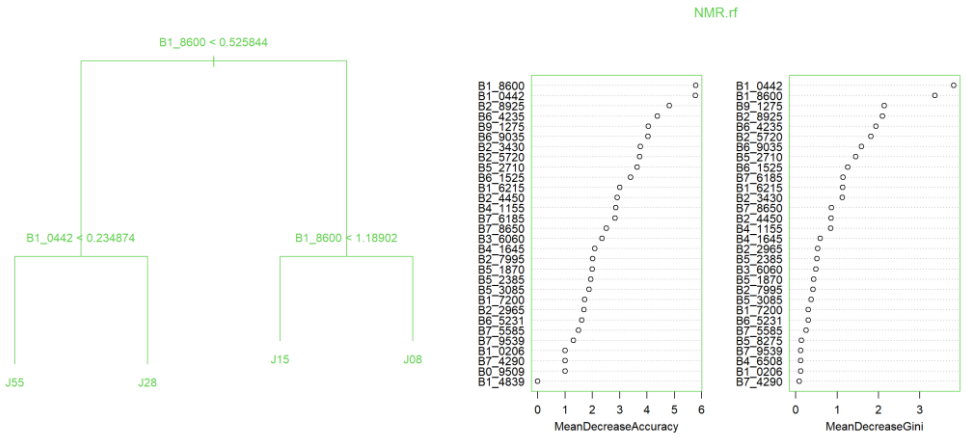

Figure 5. classification tree with 40 times cross validation (left) and variable importance identified by random forest with 100 trees grown(right)

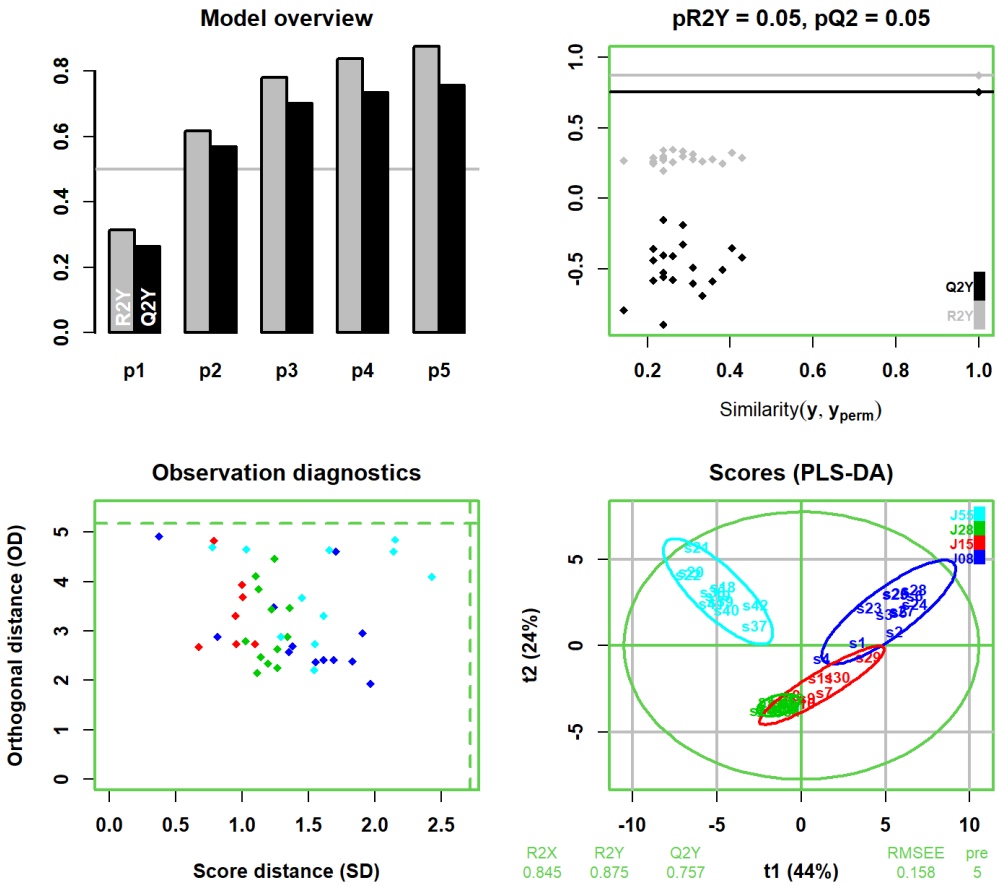

Figure 6. PLSDA for tomatoes in 4 different stages. Top left: inertia barplot: it indicates that 5 orthogonal components are enough to explain the overall variance; Top right: significance diagnostic: whether the generated R2Y and Q2Y by this model are significant enough compared to random premutation; Bottom left: outlier diagnostics; Bottom right: x-score plot: 4 clusters are projected by 2 orthogonal components with model indications R2X, R2Y, Q2Y, and RMSEE

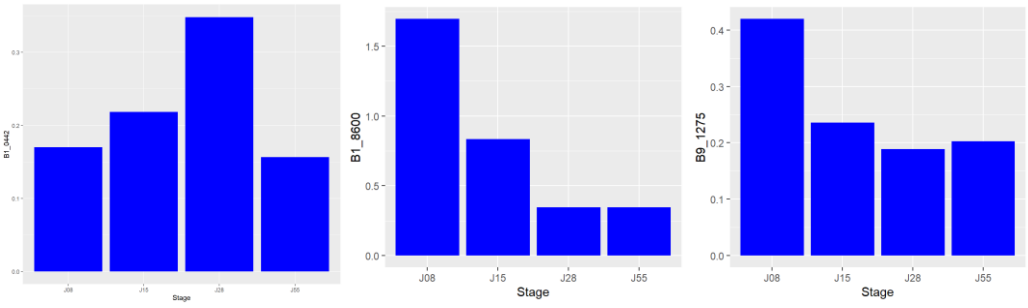

Figure 7. Fluctuation of metabolites during growth. B1_0442 (Valine), B1_8600(Quinate) and B9_1275(Trigonelline)

Interpretation

The first two principal components (PCs) were able to explain a significant amount of variance in the data, with 68.51% of the total variance being captured. However, when analyzing the PCA plots, it was found that the "control" and "shadow" groups did not show a clear dissimilarity, while four distinct group structures were observed when the PCA plots were labeled by stages. A similar pattern was also observed when performing factor analysis (FA), with a cumulative variance of 63.80%.

The scree plot further indicated that four factors were able to effectively explain the entire data set, with lower uniquenesses for 38 variables. Due to the ability to provide insights into the combination of factors after rotation, FA was determined to be more suitable for this data set. It was found that B1_6215, B1_8600, B4_1155 and B5_2385 were assigned to the first factor, and B2_3430, B2_7995 and B5_2710 were assigned to the second factor. The k-means method, however, performed poorly in this data set, with a 47.62% misclassification rate when splitting the data into two clusters and a 21.43% misclassification rate for four clusters. The poor performance was clearly evident from the plots, even with the presence of a significant overlap in the four-means plot.

Orthogonal partial least squares discriminant analysis (OPLSDA) was then applied to study the dissimilarity between the "control" and "shadow" groups, but R returned an error stating that "No model was built because the predictive component was not significant". This was consistent with the results obtained previously. Therefore, at this stage, the focus was not on studying the influence of metabolite profiles of tomatoes under different lighting conditions, but rather on the dynamic metabolic fluctuations during ripening. A partial least squares discriminant analysis (PLSDA) model was built for tomatoes in four stages, with an R2X of 0.845, an R2Y of 0.875 and a Q2Y of 0.757. This indicated that around 85% of variance could be explained by this model and that the prediction property for new instances was 75.7%.

Since feature variables could not be identified without OPLSDA, tree-based models were used to find these variables instead. A very simple tree model was obtained through the tree algorithm, with B1_8600 and B1_0442 being used to identify tomatoes in four stages. Finally, a random forest was applied to find the key compounds again, with B1_0442, B1_8600 and B9_1275 being identified as the top three compounds that experienced significant changes during growth.

Overall, the study found that there were no significant differences in the metabolic fingerprint of tomatoes grown in different lightness strengths, but there were clear grouping patterns based on the stage of tomato growth. The key compounds that experience significant changes during tomato growth were identified as B1_0442 (Valine), B1_8600(Quinate) and B9_1275(Trigonelline).

Reference

Data source: NMRProcFlow platform(https://nmrprocflow.org/ex)

Bénard C., Bernillon S., Biais B., Osorio S., Maucourt M., Ballias P., Deborde C., Colombié S., Cabasson C., Jacob D., Vercambre G., Gautier H., Rolin D., Génard M., Fernie A., Gibon Y., Moing A. 2015 Metabolomic profiling in tomato reveals diel compositional changes in fruit affected by source-sink relationships. Journal of Experimental Botany Vol. 66, No. 11 pp. 3391–3404 doi:10.1093/jxb/erv151

Ropls package:
https://master.bioconductor.org/packages/release/bioc/vignettes/ropls/inst/doc/ropls-vignette.html#1_The_ropls_package

Appendix

R code

```
library(ggfortify)
library(nFactors)
library(factoextra)
library(ropls)
library(tree)
library(randomForest)
library(ggplot2)

setwd("D:/KU Leuven/Third semester/Applied Multivariate Statistical Analysis/Assignment")

NMR<-read.table("Jiangli_Gui.csv",header=T,sep=',',na.strings="NA")

names(NMR)
names(NMR)[names(NMR) == "锘緎 amplecode"] <- "Samplecode"
names(NMR)

attach(NMR)
dim(NMR)

NMR.mat <- NMR[,-c(1:3)]
#pairs(NMR.mat)

### PCA
NMR.pca <- prcomp(NMR.mat, scale = FALSE)
plot(NMR.pca)
NMR.pca <- prcomp(NMR.mat, scale = TRUE)
plot(NMR.pca)
plot(NMR.pca, type = "l")
abline(h=1,lty='dashed')

autoplot(NMR.pca, data = NMR, colour = 'Condition')
autoplot(NMR.pca, data = NMR, colour = 'Stage')

### FA
eigenvalues <- eigen(cor(NMR.mat)) # get eigenvalues
aparallel <- parallel(subject=nrow(NMR.mat),var=ncol(NMR.mat),
                      rep=100,cent=.05)
scree <- nScree(x=eigenvalues$values, aparallel=aparallel$eigen$qevpea)
plotnScree(scree)

NMR.fa <- factanal(NMR.mat, 4, rotation = "varimax", lower = 0.01, scores = "regression")
```

```
NMR.fa
autoplot(NMR.fa, data = NMR, colour = "Condition")
autoplot(NMR.fa, data = NMR, colour = "Stage")
NMR.fa$uniquenesses
apply(NMR.fa$loadings^2,1,sum)

### try unsupervised method to cluster
### K-means
fviz_nbclust(NMR.mat, kmeans, method = "wss")

set.seed(100)
NMR.km.two <- kmeans(NMR.mat, 2, nstart = 5)
NMR.km.two
fviz_cluster(NMR.km.two, NMR.mat)
#autoplot(NMR.pca, data = NMR, colour = 'Condition') # Comparison
NMR.km.two$cluster[NMR.km.two$cluster == "1"] <- "Control"
NMR.km.two$cluster[NMR.km.two$cluster == "2"] <- "Shadow"
pred_clusters <- NMR.km.two$cluster
mis_rate_two <- mean(NMR[,2] != pred_clusters)
mis_rate_two #0.4761905

set.seed(100)
NMR.km.four <- kmeans(NMR.mat, 4, nstart = 5)
NMR.km.four
fviz_cluster(NMR.km.four, NMR.mat)
#autoplot(NMR.pca, data = NMR, colour = 'Stage') # Comparison
NMR.km.four$cluster[NMR.km.four$cluster == "1"] <- "J55"
NMR.km.four$cluster[NMR.km.four$cluster == "2"] <- "J08"
NMR.km.four$cluster[NMR.km.four$cluster == "3"] <- "J28"
NMR.km.four$cluster[NMR.km.four$cluster == "4"] <- "J15"
pred_clusters <- NMR.km.four$cluster
mis_rate <- mean(NMR[,3] != pred_clusters)
mis_rate #0.2142857

set.seed(100)
NMR.km.eight <- kmeans(NMR.mat, 8, nstart = 5)
NMR.km.eight
fviz_cluster(NMR.km.eight, NMR.mat)

### try supervised method
### PLSDA/OPLSDA
# treatment <- NMR[ ,2]
stage <- NMR[ ,3]
```

```
# NMR.oplsda <- opls(NMR.mat, treatment, predI = 1, orthoI = NA)
# Error: No model was built because the predictive component
#               was not significant

# NMR.plsda <- opls(NMR.mat, stage)
NMR.plsda.cv <- opls(NMR.mat, stage, crossvalI = 40)
# NMR.pca.new <- opls(NMR.mat)

### tree based method to find feature variables
### TREE
NMR.stage <- NMR[ ,-(1:2)]
NMR.stage$Stage <- as.factor(NMR.stage$Stage)

NMR.tree <- tree(Stage~., data = NMR.stage)
summary(NMR.tree)
plot(NMR.tree)
text(NMR.tree)
# "B1_8600" "B1_0442"

set.seed(100)
NMR.tree.cv <- cv.tree(NMR.tree,FUN=prune.misclass )
NMR.tree.cv #4
NMR.tree.prune <- prune.misclass (NMR.tree,best=4)
plot(NMR.tree.prune)
text(NMR.tree.prune)

### Random Forest
NMR.rf =randomForest(Stage~.,data= NMR.stage, mtry=38,importance =TRUE, ntree = 100)
features <- importance(NMR.rf)

subset(features[ ,1],features[ ,1]>3)
subset(features[ ,2],features[ ,2]>3)
subset(features[ ,3],features[ ,3]>3)
subset(features[ ,4],features[ ,4]>3)

varImpPlot(NMR.rf)

# "B1_8600" "B1_0442" "B9_1275"

B1_0442 <- aggregate(B1_0442 ~ Stage, data = NMR, mean)
B1_8600 <- aggregate(B1_8600 ~ Stage, data = NMR, mean)
B9_1275 <- aggregate(B9_1275 ~ Stage, data = NMR, mean)

ggplot(data = B1_0442, aes(x = Stage, y = B1_0442)) +
```

```
    geom_bar(stat = 'identity', fill='blue') +
    geom_line()

ggplot(data = B1_8600, aes(x = Stage, y = B1_8600)) +
    geom_bar(stat = 'identity', fill='blue') +
    geom_line()

ggplot(data = B9_1275, aes(x = Stage, y = B9_1275)) +
    geom_bar(stat = 'identity', fill='blue') +
    geom_line()
```

Rotated component matrix(Varimax)

	Factor1	Factor2	Factor3	Factor4
B0_9509	0.335	0.510	0.571	0.304
B1_0206	0.738		0.598	0.272
B1_0442	0.462	-0.542	0.611	0.303
B1_4839	0.283	0.313	0.165	0.886
B1_5269	0.446	-0.116	0.188	0.835
B1_6215	**-0.943**			
B1_7200		0.791	0.430	0.184
B1_8600	**-0.926**	-0.222	-0.184	
B2_2965		0.244	0.616	
B2_3430	0.243	**0.946**		
B2_4450	0.303	0.879	0.326	
B2_5720	-0.443	0.826		
B2_7995	0.220	**0.964**	0.112	
B2_8925	0.341	0.872	0.271	
B3_2066	0.114	0.675	0.225	
B3_6060	-0.594	0.473	0.240	0.299
B4_1155	**0.884**	0.395		
B4_1645	0.306	0.845	0.371	
B4_3025	0.124		0.165	-0.370
B4_6508	**0.928**	0.144		
B5_1040		-0.119		0.162
B5_1870	0.631	0.547	0.195	
B5_2385	**0.925**	0.197		
B5_2710	0.150	**0.929**		
B5_3085	-0.688	0.530	-0.110	
B5_3925	0.184	0.298	0.616	
B5_4175	-0.521		-0.151	
B5_8275	0.828	0.124	0.252	
B6_1525		**0.959**		0.170
B6_4235	**-0.940**	-0.208	-0.162	
B6_5231	-0.855	-0.318		
B6_9035	0.651		0.643	0.135
B7_4290	0.684	0.307	0.498	0.290
B7_5585	**-0.943**		-0.115	
B7_6185	0.183	0.779		-0.448
B7_8650	0.153	**0.924**		
B7_9539	-0.572	-0.317	-0.366	0.271
B9_1275	**-0.910**	-0.122		