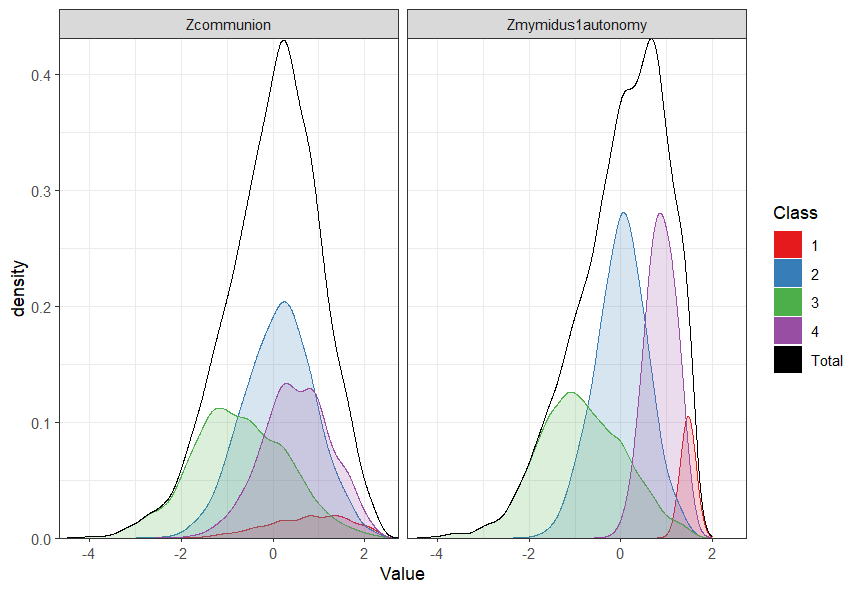
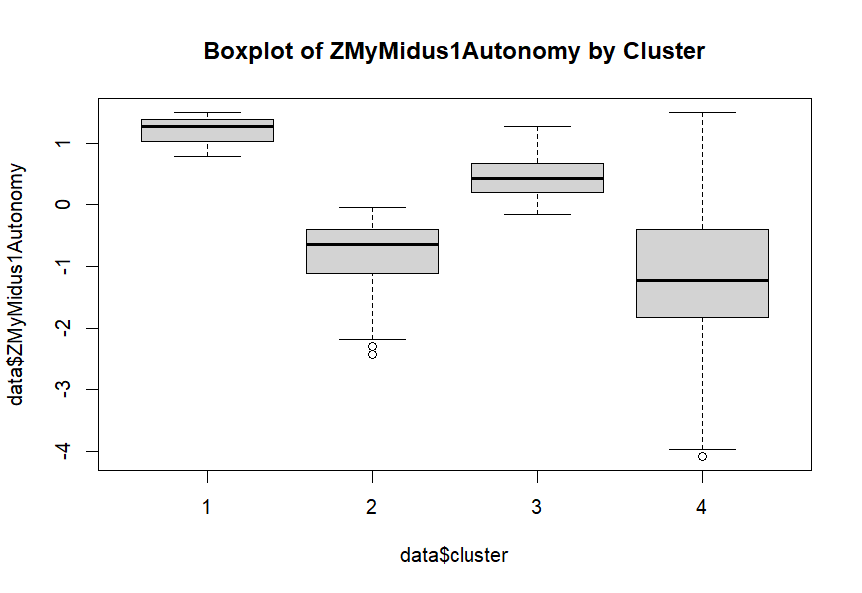
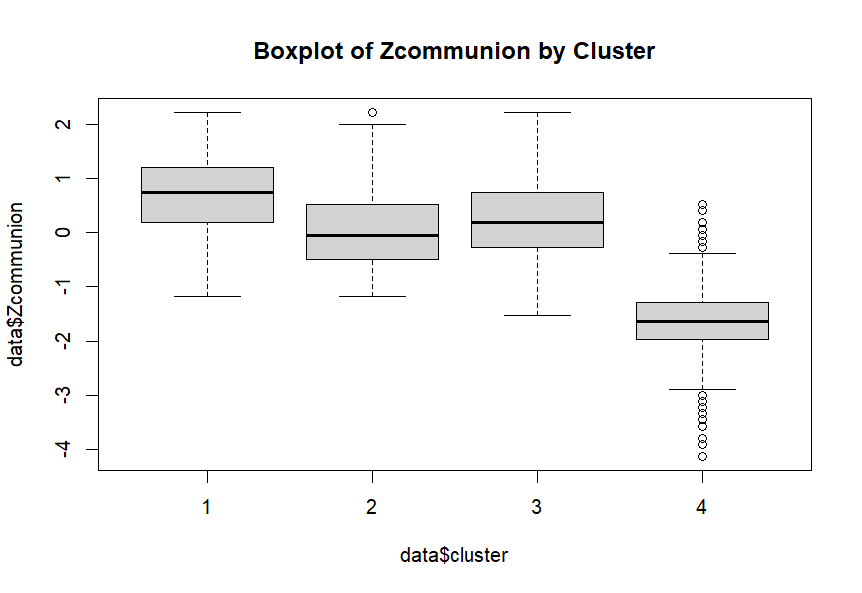
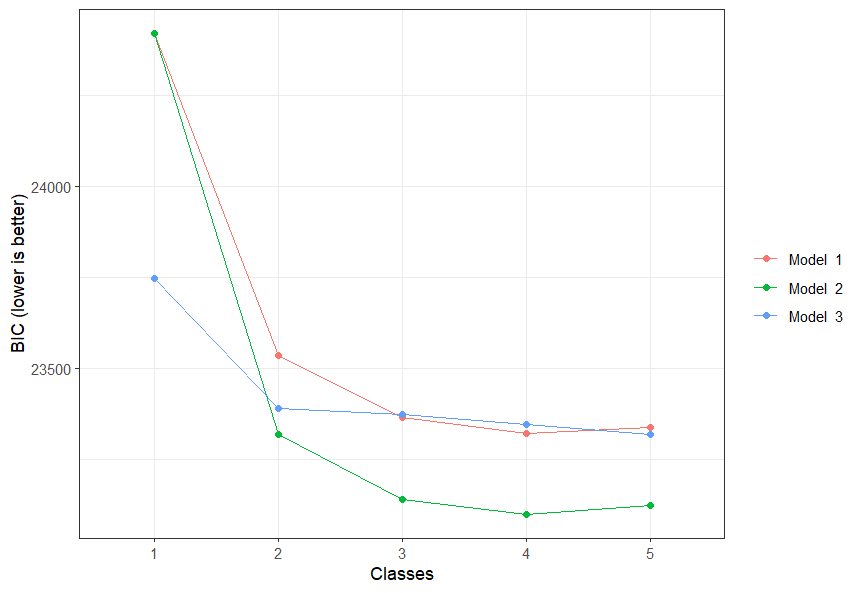
1. Density plot(4 types)





1. Different models about BIC

So we use model=2, classes=4

1. ANOVA towards 2 variables.
2. > anova\_results <- lapply(names(data)[names(data) != "cluster"], function(var) {
3. + aov(as.formula(paste(var, "~ cluster")), data = data)
4. + })
5. > lapply(anova\_results, summary)
6. [[1]]
7. Df Sum Sq Mean Sq F value Pr(>F)
8. cluster 1 1172 1171.6 1606 <2e-16 \*\*\*
9. Residuals 4342 3167 0.7
10. ---
11. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
12. [[2]]
13. Df Sum Sq Mean Sq F value Pr(>F)
14. cluster 1 636 635.9 799.4 <2e-16 \*\*\*
15. Residuals 4342 3454 0.8
16. ---
17. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

4. prediction on clusters:

> reg\_communion <- lm(Zcommunion ~ cluster, data = data)

> reg\_autonomy <- lm(ZMyMidus1Autonomy ~ cluster, data = data)

> summary(reg\_communion)

Call:

lm(formula = Zcommunion ~ cluster, data = data)

Residuals:

Min 1Q Median 3Q Max

-3.3146 -0.6650 -0.0088 0.5841 2.4894

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.40056 0.03720 37.65 <2e-16 \*\*\*

cluster -0.55520 0.01385 -40.08 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.854 on 4342 degrees of freedom

Multiple R-squared: 0.2701, Adjusted R-squared: 0.2699

F-statistic: 1606 on 1 and 4342 DF, p-value: < 2.2e-16

> summary(reg\_autonomy)

Call:

lm(formula = ZMyMidus1Autonomy ~ cluster, data = data)

Residuals:

Min 1Q Median 3Q Max

-3.5229 -0.6503 0.3540 0.7112 2.0726

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.07012 0.03886 27.54 <2e-16 \*\*\*

cluster -0.40905 0.01447 -28.27 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8919 on 4342 degrees of freedom

Multiple R-squared: 0.1555, Adjusted R-squared: 0.1553

F-statistic: 799.4 on 1 and 4342 DF, p-value: < 2.2e-16

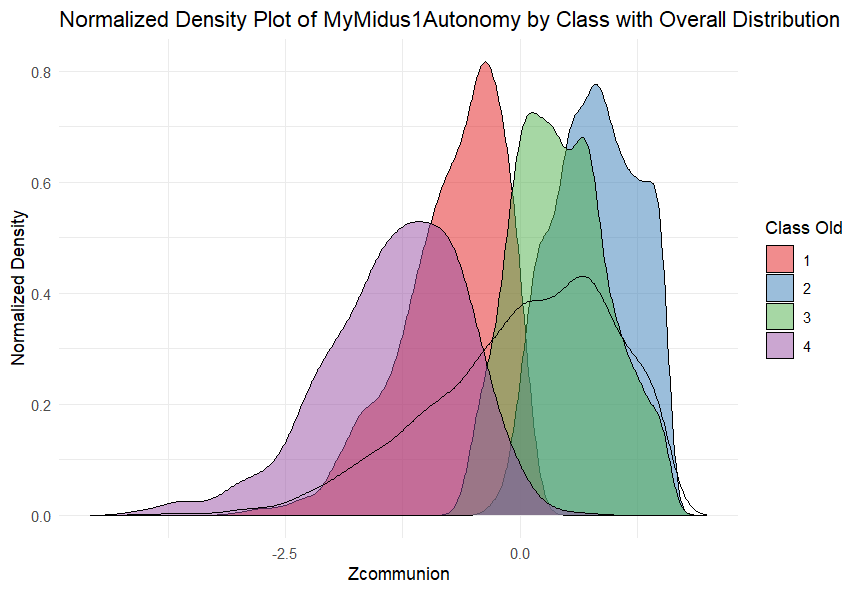
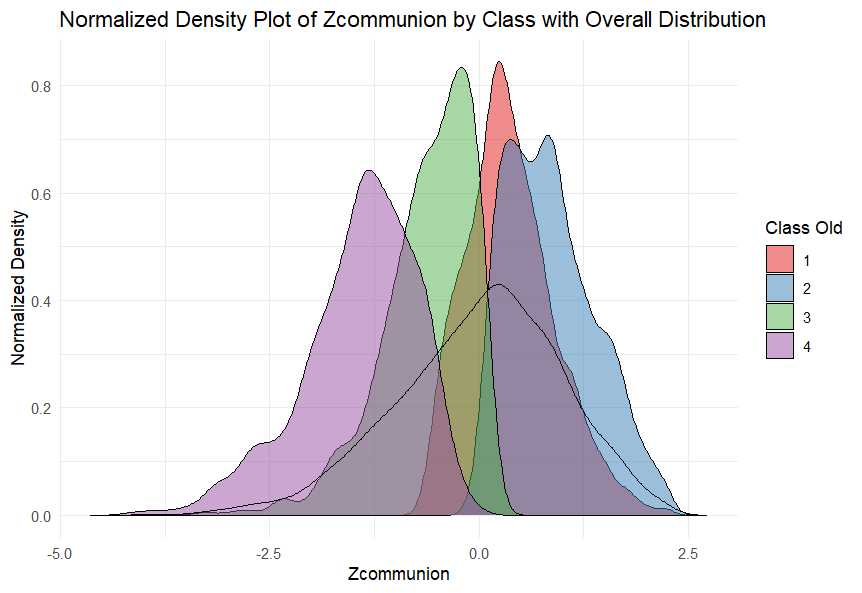
5. calculate the similarity between 2 clustering methods:

> ari <- adjustedRandIndex(clustering$Old\_Class, clustering$Class)

> print(paste("Adjusted Rand Index:", ari))

[1] "Adjusted Rand Index: 0.259178429194821"

It shows that they differ a lot.



6.BIC COMPARE

tidyLPA analysis using mclust:

Model Classes AIC BIC Entropy prob\_min prob\_max n\_min n\_max BLRT\_p

2 4 22867.39 22988.54 0.59 0.70 0.99 0.06 0.40 0.01

For Kmeans :

> k <- 4

> kmeans\_result <- kmeans(data, centers = k, nstart = 25)

> WCSS <- sum(kmeans\_result$withinss)

> n <- nrow(data)

> d <- ncol(data)

> num\_parameters <- k \* d # every k center has d parameters

> logLikelihood <- -0.5 \* n \* (d \* log(2 \* pi) + d \* log(WCSS / n) + n)

> BIC\_value <- log(n) \* num\_parameters - 2 \* logLikelihood

> print(BIC\_value)

[1] 2659.269

> AIC\_value <- 2 \* num\_parameters + WCSS

> print(AIC\_value)

[1] 2608.257

We can conclude that Kmeans is better than LPA in terms of informational criterion.

> ari <- adjustedRandIndex(clustering$Old\_Class, clustering$New\_class)

> print(paste("Adjusted Rand Index:", ari))

[1] "Adjusted Rand Index: 0.973987650852238"

I did 25 repeated K-means and get the average it shows it is very similar to the results you get, which means the results through SPSS is trustworthy.