Gaussaian Mixture Model

Josephine Choi

null

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tableone)  
path<-file.choose()  
data<-read.csv(path)  
#data - Use diabetes\_may1\_2019\_ijaz\_raw\_clean.csv

data<-data%>%select(Age\_at\_Exam,Sex,sBP,BMI,LDL,HDL,A1c,TG,FBS,Total\_Cholesterol,Diabetes,Depression,HTN,OA,COPD)  
head(data)

## Age\_at\_Exam Sex sBP BMI LDL HDL A1c TG FBS Total\_Cholesterol Diabetes  
## 1 70 1 148 26.2 3.3 1.5 5.8 1.2 5.0 5.4 0  
## 2 38 0 107 31.7 2.4 0.8 5.8 1.1 5.8 3.7 0  
## 3 74 0 126 49.0 1.5 1.0 5.2 1.9 6.0 3.4 0  
## 4 61 1 120 62.1 2.5 1.1 5.5 1.0 5.1 4.0 0  
## 5 63 0 147 32.0 2.6 1.3 5.9 1.2 6.5 4.5 0  
## 6 60 1 129 35.8 3.0 1.3 6.2 1.1 5.0 4.8 0  
## Depression HTN OA COPD  
## 1 0 1 0 0  
## 2 0 0 0 0  
## 3 0 1 1 1  
## 4 0 1 0 0  
## 5 0 0 0 0  
## 6 0 1 0 0

summary(data)

## Age\_at\_Exam Sex sBP BMI   
## Min. :18.00 Min. :0.0000 Min. : 64.0 Min. : 10.60   
## 1st Qu.:54.00 1st Qu.:0.0000 1st Qu.:118.0 1st Qu.: 25.70   
## Median :64.00 Median :0.0000 Median :129.0 Median : 29.00   
## Mean :63.03 Mean :0.4486 Mean :129.4 Mean : 30.44   
## 3rd Qu.:73.00 3rd Qu.:1.0000 3rd Qu.:140.0 3rd Qu.: 33.50   
## Max. :90.00 Max. :1.0000 Max. :266.0 Max. :739.00   
## LDL HDL A1c TG   
## Min. :0.100 Min. :0.400 Min. :4.300 Min. : 0.200   
## 1st Qu.:1.900 1st Qu.:1.100 1st Qu.:5.600 1st Qu.: 0.900   
## Median :2.500 Median :1.300 Median :5.900 Median : 1.300   
## Mean :2.626 Mean :1.353 Mean :5.875 Mean : 1.482   
## 3rd Qu.:3.300 3rd Qu.:1.600 3rd Qu.:6.200 3rd Qu.: 1.800   
## Max. :7.700 Max. :4.200 Max. :6.500 Max. :13.600   
## FBS Total\_Cholesterol Diabetes Depression   
## Min. :2.500 Min. : 1.700 Min. :0.0000 Min. :0.0000   
## 1st Qu.:5.100 1st Qu.: 3.800 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :5.600 Median : 4.600 Median :1.0000 Median :0.0000   
## Mean :5.625 Mean : 4.643 Mean :0.5067 Mean :0.2324   
## 3rd Qu.:6.200 3rd Qu.: 5.400 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :7.000 Max. :11.400 Max. :1.0000 Max. :1.0000   
## HTN OA COPD   
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :1.0000 Median :0.0000 Median :0.0000   
## Mean :0.6299 Mean :0.3104 Mean :0.1036   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

names(data)

## [1] "Age\_at\_Exam" "Sex" "sBP"   
## [4] "BMI" "LDL" "HDL"   
## [7] "A1c" "TG" "FBS"   
## [10] "Total\_Cholesterol" "Diabetes" "Depression"   
## [13] "HTN" "OA" "COPD"

#Cleaning up the data   
data<- data %>%mutate(Sex = replace(Sex, Sex == "MALE", "Male"))  
data<- data %>%mutate(Sex = replace(Sex, Sex == "FEMALE", "Female"))

#Table 1   
myvars=c('Age\_at\_Exam','Sex','sBP','BMI','LDL','HDL','A1c','TG','Total\_Cholesterol','Diabetes','HTN','OA','CoPD','FBS')  
tab <- CreateTableOne(vars = myvars, data = data)

## Warning in ModuleReturnVarsExist(vars, data): The data frame does not have: CoPD  
## Dropped

tab

##   
## Overall   
## n 9664   
## Age\_at\_Exam (mean (SD)) 63.03 (13.44)  
## Sex = 1 (%) 4335 (44.9)   
## sBP (mean (SD)) 129.40 (17.25)  
## BMI (mean (SD)) 30.44 (12.11)  
## LDL (mean (SD)) 2.63 (0.96)   
## HDL (mean (SD)) 1.35 (0.40)   
## A1c (mean (SD)) 5.87 (0.39)   
## TG (mean (SD)) 1.48 (0.83)   
## Total\_Cholesterol (mean (SD)) 4.64 (1.12)   
## Diabetes (mean (SD)) 0.51 (0.50)   
## HTN (mean (SD)) 0.63 (0.48)   
## OA (mean (SD)) 0.31 (0.46)   
## FBS (mean (SD)) 5.63 (0.71)

tab1 <- CreateTableOne(vars = myvars, data = data, strata='Diabetes')

## Warning in ModuleReturnVarsExist(vars, data): The data frame does not have: CoPD  
## Dropped

tab1

## Stratified by Diabetes  
## 0 1 p test  
## n 4767 4897   
## Age\_at\_Exam (mean (SD)) 60.36 (13.93) 65.63 (12.40) <0.001   
## Sex = 1 (%) 2034 (42.7) 2301 (47.0) <0.001   
## sBP (mean (SD)) 128.80 (17.60) 129.99 (16.89) 0.001   
## BMI (mean (SD)) 29.19 (6.44) 31.65 (15.68) <0.001   
## LDL (mean (SD)) 2.90 (0.93) 2.36 (0.91) <0.001   
## HDL (mean (SD)) 1.43 (0.42) 1.28 (0.37) <0.001   
## A1c (mean (SD)) 5.69 (0.35) 6.06 (0.34) <0.001   
## TG (mean (SD)) 1.41 (0.84) 1.55 (0.81) <0.001   
## Total\_Cholesterol (mean (SD)) 4.96 (1.08) 4.34 (1.06) <0.001   
## Diabetes (mean (SD)) 0.00 (0.00) 1.00 (0.00) <0.001   
## HTN (mean (SD)) 0.53 (0.50) 0.73 (0.44) <0.001   
## OA (mean (SD)) 0.27 (0.44) 0.35 (0.48) <0.001   
## FBS (mean (SD)) 5.30 (0.59) 5.94 (0.66) <0.001

## Including Plots

You can also embed plots, for example:

data$comorbidities<-data$HTN+data$OA+data$COPD+data$Depression  
  
  
#Logistic Regression (dependent - Diabetes)  
mod.1<-data%>%glm(formula=factor(Diabetes)~FBS+sBP+Total\_Cholesterol+TG+BMI+LDL+HDL+Age\_at\_Exam,family=binomial)  
  
#Ijaz's model  
  
  
#addint the quadratic elements   
#assumgin FBS = Fasting Plasma Glucose (FPG) - dependent variable   
data$BMI2<-data$BMI^2  
data$A1c2<-data$A1c^2  
  
mod.2<-data%>%lm(formula=FBS~Age\_at\_Exam+sBP+BMI+LDL+HDL+A1c+TG+factor(Diabetes)+OA+factor(Sex)+BMI2+BMI\*OA+A1c2)  
  
#mod.1<-data%>%glm(formula=factor(Diabetes)~FBS+sBP+Total\_Cholesterol+BMI+LDL+HDL+Age\_at\_Exam+factor(Sex),family=binomial)  
summary(mod.1)

##   
## Call:  
## glm(formula = factor(Diabetes) ~ FBS + sBP + Total\_Cholesterol +   
## TG + BMI + LDL + HDL + Age\_at\_Exam, family = binomial, data = .)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6624 -0.8590 0.3018 0.8508 3.1060   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.640007 0.322258 -23.708 < 2e-16 \*\*\*  
## FBS 1.439976 0.040508 35.548 < 2e-16 \*\*\*  
## sBP -0.004815 0.001452 -3.317 0.000909 \*\*\*  
## Total\_Cholesterol 0.025852 0.141590 0.183 0.855122   
## TG 0.032572 0.067017 0.486 0.626952   
## BMI 0.033559 0.003810 8.808 < 2e-16 \*\*\*  
## LDL -0.587217 0.143039 -4.105 4.04e-05 \*\*\*  
## HDL -0.567499 0.152463 -3.722 0.000197 \*\*\*  
## Age\_at\_Exam 0.020881 0.001979 10.550 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 13395 on 9663 degrees of freedom  
## Residual deviance: 10282 on 9655 degrees of freedom  
## AIC: 10300  
##   
## Number of Fisher Scoring iterations: 4

summary(mod.2)

##   
## Call:  
## lm(formula = FBS ~ Age\_at\_Exam + sBP + BMI + LDL + HDL + A1c +   
## TG + factor(Diabetes) + OA + factor(Sex) + BMI2 + BMI \* OA +   
## A1c2, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7620 -0.3743 0.0052 0.4010 1.9093   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.071e+00 1.070e+00 3.804 0.000143 \*\*\*  
## Age\_at\_Exam 4.103e-03 5.137e-04 7.988 1.53e-15 \*\*\*  
## sBP 2.829e-03 3.581e-04 7.900 3.09e-15 \*\*\*  
## BMI 9.443e-03 1.122e-03 8.415 < 2e-16 \*\*\*  
## LDL 2.187e-02 6.735e-03 3.247 0.001170 \*\*   
## HDL -1.510e-02 1.769e-02 -0.854 0.393226   
## A1c -4.078e-01 3.714e-01 -1.098 0.272247   
## TG 6.194e-02 7.862e-03 7.879 3.67e-15 \*\*\*  
## factor(Diabetes)1 4.171e-01 1.425e-02 29.271 < 2e-16 \*\*\*  
## OA 5.090e-02 3.376e-02 1.508 0.131674   
## factor(Sex)1 1.601e-01 1.301e-02 12.306 < 2e-16 \*\*\*  
## BMI2 -1.180e-05 1.734e-06 -6.802 1.09e-11 \*\*\*  
## A1c2 7.631e-02 3.204e-02 2.382 0.017243 \*   
## BMI:OA -2.309e-03 1.006e-03 -2.296 0.021717 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5855 on 9650 degrees of freedom  
## Multiple R-squared: 0.3134, Adjusted R-squared: 0.3125   
## F-statistic: 338.8 on 13 and 9650 DF, p-value: < 2.2e-16

data\_num<-data%>%select(Age\_at\_Exam,sBP,BMI,LDL,HDL,A1c,TG,FBS,Total\_Cholesterol,Diabetes,comorbidities,OA)  
#data\_num<-data%>%select(Age\_at\_Exam,sBP,BMI,LDL,HDL,A1c,TG,FBS,Total\_Cholesterol)  
#cor(data\_num)

## McFadden’s R Squared

#McFadden’s R squared in R  
# <- glm(data$Diabetes~1, family="binomial")  
#logLik(mod.1)/logLik(nullmod)

* McFadden’s R Squared - 0.2 to 0.4 indicates a good fit (Haider, p.362)
* Therefore, 0.24 indicates that this logit model is adequate ## VIF

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

vif(mod.1)

## FBS sBP Total\_Cholesterol TG   
## 1.044599 1.078539 38.780859 5.162892   
## BMI LDL HDL Age\_at\_Exam   
## 1.113874 28.765208 6.129004 1.147769

* Using VIF here to check for multicollinearity
* Total\_Cholesterol and LDL has scores over 5 (8.71 and 8.17 respectively). This is confirmed with the correlation matrix (see below). Therefore, Total\_Cholesterol should be dropped

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## Age\_at\_Exam sBP BMI LDL HDL A1c TG FBS  
## Age\_at\_Exam 1.00 0.18 -0.08 -0.20 0.07 0.29 -0.06 0.20  
## sBP 0.18 1.00 0.06 0.06 0.01 0.07 0.09 0.13  
## BMI -0.08 0.06 1.00 -0.02 -0.14 0.09 0.10 0.11  
## LDL -0.20 0.06 -0.02 1.00 0.13 -0.17 0.08 -0.12  
## HDL 0.07 0.01 -0.14 0.13 1.00 -0.13 -0.35 -0.18  
## A1c 0.29 0.07 0.09 -0.17 -0.13 1.00 0.09 0.45  
## TG -0.06 0.09 0.10 0.08 -0.35 0.09 1.00 0.14  
## FBS 0.20 0.13 0.11 -0.12 -0.18 0.45 0.14 1.00  
## Total\_Cholesterol -0.16 0.08 -0.03 0.91 0.35 -0.16 0.26 -0.12  
## Diabetes 0.20 0.03 0.10 -0.28 -0.19 0.48 0.09 0.46  
## comorbidities 0.33 0.16 0.09 -0.13 -0.02 0.16 0.09 0.12  
## OA 0.28 0.05 0.05 -0.06 0.05 0.09 0.03 0.06  
## Total\_Cholesterol Diabetes comorbidities OA  
## Age\_at\_Exam -0.16 0.20 0.33 0.28  
## sBP 0.08 0.03 0.16 0.05  
## BMI -0.03 0.10 0.09 0.05  
## LDL 0.91 -0.28 -0.13 -0.06  
## HDL 0.35 -0.19 -0.02 0.05  
## A1c -0.16 0.48 0.16 0.09  
## TG 0.26 0.09 0.09 0.03  
## FBS -0.12 0.46 0.12 0.06  
## Total\_Cholesterol 1.00 -0.28 -0.09 -0.02  
## Diabetes -0.28 1.00 0.18 0.09  
## comorbidities -0.09 0.18 1.00 0.63  
## OA -0.02 0.09 0.63 1.00  
##   
## n= 9664   
##   
##   
## P  
## Age\_at\_Exam sBP BMI LDL HDL A1c TG FBS   
## Age\_at\_Exam 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  
## sBP 0.0000 0.0000 0.0000 0.4473 0.0000 0.0000 0.0000  
## BMI 0.0000 0.0000 0.0738 0.0000 0.0000 0.0000 0.0000  
## LDL 0.0000 0.0000 0.0738 0.0000 0.0000 0.0000 0.0000  
## HDL 0.0000 0.4473 0.0000 0.0000 0.0000 0.0000 0.0000  
## A1c 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  
## TG 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  
## FBS 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000   
## Total\_Cholesterol 0.0000 0.0000 0.0020 0.0000 0.0000 0.0000 0.0000 0.0000  
## Diabetes 0.0000 0.0007 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  
## comorbidities 0.0000 0.0000 0.0000 0.0000 0.1306 0.0000 0.0000 0.0000  
## OA 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0098 0.0000  
## Total\_Cholesterol Diabetes comorbidities OA   
## Age\_at\_Exam 0.0000 0.0000 0.0000 0.0000  
## sBP 0.0000 0.0007 0.0000 0.0000  
## BMI 0.0020 0.0000 0.0000 0.0000  
## LDL 0.0000 0.0000 0.0000 0.0000  
## HDL 0.0000 0.0000 0.1306 0.0000  
## A1c 0.0000 0.0000 0.0000 0.0000  
## TG 0.0000 0.0000 0.0000 0.0098  
## FBS 0.0000 0.0000 0.0000 0.0000  
## Total\_Cholesterol 0.0000 0.0000 0.0356  
## Diabetes 0.0000 0.0000 0.0000  
## comorbidities 0.0000 0.0000 0.0000  
## OA 0.0356 0.0000 0.0000

# Gaussian Mixture Model - loading mclust

### Thoughts

* Using BIC to decide the number of components
* will start with only a couple of indpendent variables
* ideally will see only a couple clusters - one for diabetic pt and one for non-diabetic patient

### Benefit

* of Gaussian Mixture Model - more flexible than K-mean , with two parameters to describe the cluster shape (mean and standard deviation) - K-mean makes naive use of the mean value as the cluster center
* fastest algorithm among the learning mixture models

### Disadvantage

* need sufficient pts per mixture, else estimating the covariance becomes difficult
* need to select the # of gaussians (components)

### Steps

* select # of clusters (accomplished by BIC)
* compute the probability
* based on probability, a new set of Guassian distributed will be computed

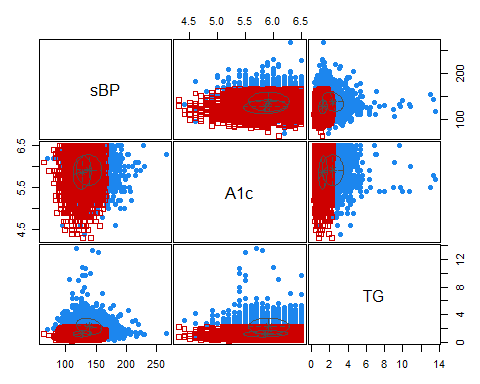
# Load the package  
library(mclust)

## Package 'mclust' version 5.4.10  
## Type 'citation("mclust")' for citing this R package in publications.

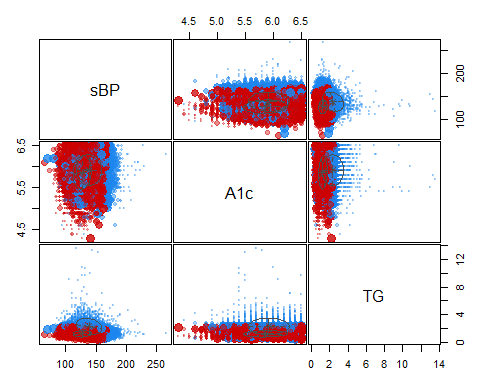
#testing on (3-sBP, 4-BMI,7-A1c,8-TG,9-FBS)  
  
X<-data[,c(3,4,7,8,9)]  
  
BIC<-mclustBIC(X)  
BIC

## Bayesian Information Criterion (BIC):   
## EII VII EEI VEI EVI VVI EEE  
## 1 -354126.4 -354126.4 -211948.9 -211948.9 -211948.9 -211948.9 -209265.1  
## 2 -341123.6 -332238.0 -209175.2 -207517.8 -199802.0 -198497.7 -208976.3  
## 3 -321870.0 -309780.4 -207283.1 -196554.8 -196011.1 -194701.9 -207048.9  
## 4 -309025.7 -298628.2 -207027.2 -194019.0 -194359.2 -192741.0 -206903.3  
## 5 -302882.4 -290464.3 -206988.7 -193477.9 -193078.3 -191854.7 -206540.9  
## 6 -299283.6 -285638.0 -206685.0 -193335.5 -192420.8 -191058.3 -206592.1  
## 7 -293896.1 -278271.3 -206739.7 -193116.7 -191708.9 -190486.9 -206648.0  
## 8 -291642.7 -275315.8 -206727.1 -193074.5 -191427.8 -190691.0 -206617.2  
## 9 -288581.2 -271610.4 -195947.4 -192693.0 -191182.5 -190129.1 -197001.5  
## VEE EVE VVE EEV VEV EVV VVV  
## 1 -209265.1 -209265.1 -209265.1 -209265.1 -209265.1 -209265.1 -209265.1  
## 2 -196795.7 -197258.5 -195953.6 -198617.9 -195766.7 -196666.3 -195290.8  
## 3 -194533.8 -194067.6 -192677.7 -196775.0 -193043.4 -193389.0 -192218.5  
## 4 -193689.4 -192663.0 -191664.0 -196813.0 -192567.9 -193074.8 -191718.8  
## 5 -193320.6 -192027.7 -191455.4 -195480.2 -191798.9 -192046.2 -191066.2  
## 6 -193062.9 -191903.5 -190965.0 -195481.5 -192240.2 -191740.4 -190756.1  
## 7 -192823.9 -191421.9 -190569.0 -195306.3 -191613.6 -191347.1 -190642.4  
## 8 -192771.2 -191191.8 -190405.0 -192903.5 -191324.9 -191240.9 -190385.6  
## 9 -191835.9 -190983.2 -190281.6 -192557.0 -191100.2 -191266.5 -190380.0  
##   
## Top 3 models based on the BIC criterion:   
## VVI,9 VVE,9 VVV,9   
## -190129.1 -190281.6 -190380.0

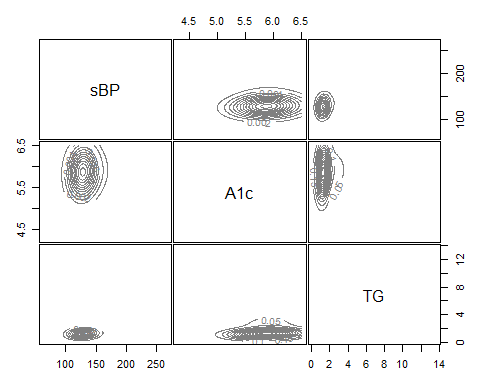
# Select 4 continuous variables and look for two distinct groups.  
mcl.model <- Mclust(data[,c(3,7,8)],2)  
# Plot our results.  
plot(mcl.model, what = "classification", main = "Mclust Classification")



plot(mcl.model,what="uncertainty",main="Mclust Uncertainty")



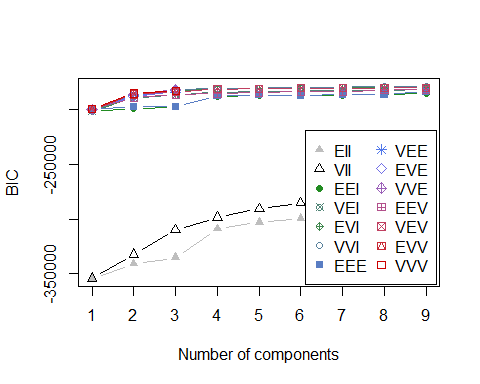
plot(mcl.model,what="density",main="Mclust density")



table(data$Diabetes,mcl.model$classification)

##   
## 1 2  
## 0 720 4047  
## 1 889 4008

BIC <- mclustBIC(data[,c(3,4,6,7,8)])  
plot(BIC)



summary(BIC)

## Best BIC values:  
## VVI,9 VVI,8 VVI,7  
## BIC -178479.6 -178526.04221 -179087.7810  
## BIC diff 0.0 -46.39474 -608.1335