Module4L3

Neha Parulekar

February 22, 2016

## Additional packages needed

To run the code in M04\_Lesson\_03.Rmd you may need additional packages.

* If necessary install these packages.

install.packages("mclust");

require(mclust)

## Loading required package: mclust

## Package 'mclust' version 5.1

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

# Assingment

* Use the same dataset you use for M04 Lesson 02 for the partition (k-means,PAM) and hierarchical clustering from the he the [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/)

# set the working directory  
setwd("C:/Users/Neha/Desktop")  
  
# load the file  
ProtienExpressionData <- read.csv("Data\_Cortex\_Nuclear.csv")  
  
# Checking the data   
head(ProtienExpressionData)

## MouseID DYRK1A\_N ITSN1\_N BDNF\_N NR1\_N NR2A\_N pAKT\_N  
## 1 309\_1 0.5036439 0.7471932 0.4301753 2.816329 5.990152 0.2188300  
## 2 309\_2 0.5146171 0.6890635 0.4117703 2.789514 5.685038 0.2116362  
## 3 309\_3 0.5091831 0.7302468 0.4183088 2.687201 5.622059 0.2090109  
## 4 309\_4 0.4421067 0.6170762 0.3586263 2.466947 4.979503 0.2228858  
## 5 309\_5 0.4349402 0.6174298 0.3588022 2.365785 4.718679 0.2131059  
## 6 309\_6 0.4475064 0.6281758 0.3673881 2.385939 4.807635 0.2185778  
## pBRAF\_N pCAMKII\_N pCREB\_N pELK\_N pERK\_N pJNK\_N PKCA\_N  
## 1 0.1775655 2.373744 0.2322238 1.750936 0.6879062 0.3063817 0.4026984  
## 2 0.1728170 2.292150 0.2269721 1.596377 0.6950062 0.2990511 0.3859868  
## 3 0.1757222 2.283337 0.2302468 1.561316 0.6773484 0.2912761 0.3810025  
## 4 0.1764626 2.152301 0.2070042 1.595086 0.5832768 0.2967287 0.3770870  
## 5 0.1736270 2.134014 0.1921579 1.504230 0.5509601 0.2869612 0.3635021  
## 6 0.1762334 2.141282 0.1951875 1.442398 0.5663396 0.2898239 0.3638930  
## pMEK\_N pNR1\_N pNR2A\_N pNR2B\_N pPKCAB\_N pRSK\_N AKT\_N  
## 1 0.2969273 1.0220603 0.6056726 1.877684 2.308745 0.4415994 0.8593658  
## 2 0.2813189 0.9566759 0.5875587 1.725774 2.043037 0.4452219 0.8346593  
## 3 0.2817103 1.0036350 0.6024488 1.731873 2.017984 0.4676679 0.8143294  
## 4 0.3138320 0.8753903 0.5202932 1.566852 2.132754 0.4776707 0.7277046  
## 5 0.2779643 0.8649120 0.5079898 1.480059 2.013697 0.4834161 0.6877937  
## 6 0.2668369 0.8591209 0.5213066 1.538244 1.968275 0.4959000 0.6724022  
## BRAF\_N CAMKII\_N CREB\_N ELK\_N ERK\_N GSK3B\_N JNK\_N  
## 1 0.4162891 0.3696080 0.1789443 1.866358 3.685247 1.537227 0.2645263  
## 2 0.4003642 0.3561775 0.1736797 1.761047 3.485287 1.509249 0.2557270  
## 3 0.3998469 0.3680888 0.1739047 1.765544 3.571456 1.501244 0.2596135  
## 4 0.3856387 0.3629700 0.1794489 1.286277 2.970137 1.419710 0.2595358  
## 5 0.3675305 0.3553109 0.1748355 1.324695 2.896334 1.359876 0.2507050  
## 6 0.3694045 0.3571717 0.1797285 1.227450 2.956983 1.447910 0.2508402  
## MEK\_N TRKA\_N RSK\_N APP\_N Bcatenin\_N SOD1\_N MTOR\_N  
## 1 0.3196770 0.8138665 0.1658460 0.4539098 3.037621 0.3695096 0.4585385  
## 2 0.3044187 0.7805042 0.1571935 0.4309403 2.921882 0.3422793 0.4235599  
## 3 0.3117467 0.7851540 0.1608954 0.4231873 2.944136 0.3436962 0.4250048  
## 4 0.2792181 0.7344917 0.1622099 0.4106149 2.500204 0.3445093 0.4292113  
## 5 0.2736672 0.7026991 0.1548274 0.3985498 2.456560 0.3291258 0.4087552  
## 6 0.2840436 0.7043958 0.1568759 0.3910472 2.467133 0.3275978 0.4044899  
## P38\_N pMTOR\_N DSCR1\_N AMPKA\_N NR2B\_N pNUMB\_N RAPTOR\_N  
## 1 0.3353358 0.8251920 0.5769155 0.4480993 0.5862714 0.3947213 0.3395706  
## 2 0.3248347 0.7617176 0.5450973 0.4208761 0.5450973 0.3682546 0.3219592  
## 3 0.3248517 0.7570308 0.5436197 0.4046298 0.5529941 0.3638799 0.3130859  
## 4 0.3301208 0.7469798 0.5467626 0.3868603 0.5478485 0.3667707 0.3284919  
## 5 0.3134148 0.6919565 0.5368605 0.3608164 0.5128240 0.3515510 0.3122063  
## 6 0.2962764 0.6744186 0.5397231 0.3542143 0.5143164 0.3472241 0.3031321  
## TIAM1\_N pP70S6\_N NUMB\_N P70S6\_N pGSK3B\_N pPKCG\_N CDK5\_N  
## 1 0.4828639 0.2941698 0.1821505 0.8427252 0.1926084 1.443091 0.2947000  
## 2 0.4545193 0.2764306 0.1820863 0.8476146 0.1948153 1.439460 0.2940598  
## 3 0.4471972 0.2566482 0.1843877 0.8561658 0.2007373 1.524364 0.3018807  
## 4 0.4426497 0.3985340 0.1617677 0.7602335 0.1841694 1.612382 0.2963818  
## 5 0.4190949 0.3934470 0.1602002 0.7681129 0.1857183 1.645807 0.2968294  
## 6 0.4128243 0.3825783 0.1623303 0.7796946 0.1867930 1.634615 0.2880373  
## S6\_N ADARB1\_N AcetylH3K9\_N RRP1\_N BAX\_N ARC\_N ERBB4\_N  
## 1 0.3546045 1.339070 0.1701188 0.1591024 0.1888517 0.1063052 0.1449893  
## 2 0.3545483 1.306323 0.1714271 0.1581289 0.1845700 0.1065922 0.1504709  
## 3 0.3860868 1.279600 0.1854563 0.1486963 0.1905322 0.1083031 0.1453302  
## 4 0.2906795 1.198765 0.1597991 0.1661123 0.1853235 0.1031838 0.1406558  
## 5 0.3093450 1.206995 0.1646503 0.1606870 0.1882214 0.1047838 0.1419830  
## 6 0.3323671 1.123445 0.1756929 0.1505939 0.1838235 0.1064762 0.1395645  
## nNOS\_N Tau\_N GFAP\_N GluR3\_N GluR4\_N IL1B\_N P3525\_N  
## 1 0.1766677 0.1251904 0.1152909 0.2280435 0.1427556 0.4309575 0.2475378  
## 2 0.1783090 0.1342751 0.1182345 0.2380731 0.1420366 0.4571562 0.2576322  
## 3 0.1762129 0.1325604 0.1177602 0.2448173 0.1424450 0.5104723 0.2553430  
## 4 0.1638042 0.1232096 0.1174394 0.2349467 0.1450682 0.4309959 0.2511031  
## 5 0.1677096 0.1368377 0.1160478 0.2555277 0.1408705 0.4812265 0.2517730  
## 6 0.1748445 0.1305147 0.1152432 0.2368495 0.1364536 0.4785775 0.2444853  
## pCASP9\_N PSD95\_N SNCA\_N Ubiquitin\_N pGSK3B\_Tyr216\_N SHH\_N  
## 1 1.603310 2.014875 0.1082343 1.0449792 0.8315565 0.1888517  
## 2 1.671738 2.004605 0.1097485 1.0098831 0.8492704 0.2004036  
## 3 1.663550 2.016831 0.1081962 0.9968476 0.8467087 0.1936845  
## 4 1.484624 1.957233 0.1198832 0.9902247 0.8332768 0.1921119  
## 5 1.534835 2.009109 0.1195244 0.9977750 0.8786678 0.2056042  
## 6 1.507777 2.003535 0.1206872 0.9201782 0.8436793 0.1904695  
## BAD\_N BCL2\_N pS6\_N pCFOS\_N SYP\_N H3AcK18\_N EGR1\_N  
## 1 0.1226520 NA 0.1063052 0.1083359 0.4270992 0.1147832 0.1317900  
## 2 0.1166822 NA 0.1065922 0.1043154 0.4415813 0.1119735 0.1351030  
## 3 0.1185082 NA 0.1083031 0.1062193 0.4357769 0.1118829 0.1333618  
## 4 0.1327812 NA 0.1031838 0.1112620 0.3916910 0.1304053 0.1474442  
## 5 0.1299541 NA 0.1047838 0.1106939 0.4341538 0.1184814 0.1403143  
## 6 0.1315752 NA 0.1064762 0.1094457 0.4398331 0.1166572 0.1407664  
## H3MeK4\_N CaNA\_N Genotype Treatment Behavior class  
## 1 0.1281856 1.675652 Control Memantine C/S c-CS-m  
## 2 0.1311187 1.743610 Control Memantine C/S c-CS-m  
## 3 0.1274311 1.926427 Control Memantine C/S c-CS-m  
## 4 0.1469011 1.700563 Control Memantine C/S c-CS-m  
## 5 0.1483799 1.839730 Control Memantine C/S c-CS-m  
## 6 0.1421804 1.816389 Control Memantine C/S c-CS-m

# Removing the NA columns and columns without protien expression (last four cols)from the data   
  
RemoveColms <- c("MouseID","Genotype","Treatment","Behavior","class")  
ModifiedProtienExpressionData <- ProtienExpressionData[, !(names(ProtienExpressionData) %in% RemoveColms)]  
  
ModifiedProtienExpressionData <- ModifiedProtienExpressionData[complete.cases(ModifiedProtienExpressionData),]  
  
head(ModifiedProtienExpressionData)

## DYRK1A\_N ITSN1\_N BDNF\_N NR1\_N NR2A\_N pAKT\_N pBRAF\_N  
## 76 0.6497813 0.8286964 0.4058618 2.921435 5.167979 0.2071741 0.1766404  
## 77 0.6164807 0.8419742 0.3885837 2.862575 5.194163 0.2234335 0.1677253  
## 78 0.6374243 0.8528818 0.4005615 2.968155 5.350820 0.2087902 0.1732608  
## 79 0.5768145 0.7553900 0.3483463 2.624901 4.727509 0.2058923 0.1611920  
## 80 0.5425448 0.7579173 0.3500507 2.634509 4.735602 0.2105263 0.1656711  
## 81 0.5699176 0.7610777 0.3439100 2.598085 4.764640 0.2039635 0.1644400  
## pCAMKII\_N pCREB\_N pELK\_N pERK\_N pJNK\_N PKCA\_N pMEK\_N  
## 76 3.728084 0.2392826 1.666579 0.9694663 0.3212598 0.4065617 0.2552931  
## 77 3.648240 0.2210300 1.565150 0.9918455 0.3429185 0.3969099 0.2708155  
## 78 3.814545 0.2223002 1.741732 0.9899991 0.3289762 0.3979296 0.2637951  
## 79 3.778530 0.1941528 1.505475 0.8191669 0.3115476 0.3858223 0.2415623  
## 80 3.871971 0.1942973 1.531613 0.8150569 0.3150006 0.3915249 0.2541418  
## 81 3.843465 0.1930528 1.575707 0.8364507 0.3065019 0.3737475 0.2734358  
## pNR1\_N pNR2A\_N pNR2B\_N pPKCAB\_N pRSK\_N AKT\_N BRAF\_N  
## 76 1.0563430 1.0993001 1.925284 2.231671 0.4268591 0.8732283 0.5127734  
## 77 1.0327897 1.0787124 1.893047 2.330987 0.4091845 0.7890129 0.5115021  
## 78 1.0922888 1.0940433 1.915958 1.952540 0.4540749 0.7937538 0.5332047  
## 79 0.9484140 0.9248222 1.755842 2.297212 0.4848177 0.7421831 0.4591940  
## 80 0.9553702 0.9384650 1.744844 2.477854 0.4765017 0.7158796 0.4863068  
## 81 0.9453351 0.9258517 1.740481 2.374415 0.4718326 0.7214429 0.4957693  
## CAMKII\_N CREB\_N ELK\_N ERK\_N GSK3B\_N JNK\_N MEK\_N  
## 76 0.3432196 0.1676290 1.852756 3.366492 1.503675 0.2460192 0.3054243  
## 77 0.3448069 0.1666953 1.852532 3.431159 1.471674 0.2420601 0.3001717  
## 78 0.3520484 0.1725590 1.965786 3.434249 1.541802 0.2519519 0.3459075  
## 79 0.3375099 0.1619822 1.364714 2.814990 1.440795 0.2349024 0.2749746  
## 80 0.3386679 0.1605996 1.398174 2.891018 1.469402 0.2348698 0.2700327  
## 81 0.3269873 0.1589846 1.400356 2.794923 1.425852 0.2349143 0.2923625  
## TRKA\_N RSK\_N APP\_N Bcatenin\_N SOD1\_N MTOR\_N P38\_N  
## 76 0.8510061 0.1651794 0.4415573 2.578215 0.3408574 0.4876640 0.4010499  
## 77 0.8336481 0.1626609 0.4372532 2.538026 0.3376824 0.4726180 0.4023176  
## 78 0.8813931 0.1652777 0.4434600 2.609615 0.3423985 0.4939030 0.4109132  
## 79 0.7664522 0.1505813 0.4076081 2.311999 0.3121120 0.4663055 0.3947398  
## 80 0.7554378 0.1495548 0.4113603 2.320861 0.3094782 0.4723318 0.3936662  
## 81 0.7505010 0.1439546 0.4093743 2.272322 0.3047206 0.4647072 0.4164997  
## pMTOR\_N DSCR1\_N AMPKA\_N NR2B\_N pNUMB\_N RAPTOR\_N TIAM1\_N  
## 76 0.7813648 0.5340332 0.4318460 0.6180227 0.3614173 0.3219598 0.4272966  
## 77 0.7865236 0.5332189 0.4298712 0.6031760 0.3602575 0.2960515 0.4258369  
## 78 0.8127906 0.5646987 0.4411791 0.6402316 0.3743311 0.2996754 0.4452145  
## 79 0.7406028 0.5229710 0.3919178 0.5850547 0.3451857 0.2867141 0.3999323  
## 80 0.7392088 0.5422067 0.3899470 0.5874000 0.3465570 0.2903189 0.4058379  
## 81 0.7389223 0.5352928 0.3895569 0.6014251 0.3515921 0.2887998 0.4003563  
## pP70S6\_N NUMB\_N P70S6\_N pGSK3B\_N pPKCG\_N CDK5\_N S6\_N  
## 76 0.2520560 0.2420860 1.0611810 0.1637581 1.862368 0.3211749 0.5410410  
## 77 0.2520172 0.2421199 1.0639853 0.1699811 1.692204 0.3204733 0.5392264  
## 78 0.2523028 0.2410417 1.0594529 0.1752296 1.654840 0.3178561 0.5618754  
## 79 0.2774580 0.2087179 0.9690072 0.1537790 2.027937 0.2998254 0.4667623  
## 80 0.3546715 0.2127144 0.9729058 0.1630624 1.969799 0.3030077 0.4917972  
## 81 0.2940325 0.2123696 0.9852211 0.1682191 1.960879 0.3097367 0.4834203  
## ADARB1\_N AcetylH3K9\_N RRP1\_N BAX\_N ARC\_N ERBB4\_N nNOS\_N  
## 76 2.159547 0.1346388 0.1581778 0.1961242 0.1320008 0.1611201 0.2109882  
## 77 2.167098 0.1362732 0.1615790 0.1933479 0.1390077 0.1615293 0.2107984  
## 78 2.182447 0.1359140 0.1675078 0.1922378 0.1348037 0.1810336 0.2025336  
## 79 1.795148 0.1260913 0.1443627 0.1769768 0.1189823 0.1636318 0.1953106  
## 80 1.763485 0.1243475 0.1537410 0.1881059 0.1286353 0.1679717 0.1969923  
## 81 1.744163 0.1255589 0.1606433 0.1913810 0.1279185 0.1779682 0.1884625  
## Tau\_N GFAP\_N GluR3\_N GluR4\_N IL1B\_N P3525\_N pCASP9\_N  
## 76 0.2103795 0.1281453 0.2924614 0.1525974 0.5272423 0.3480621 1.687551  
## 77 0.1933976 0.1345829 0.2914885 0.1518843 0.5197872 0.3651188 1.662872  
## 78 0.2043000 0.1295549 0.2700616 0.1536287 0.5063087 0.3497527 1.716362  
## 79 0.1941257 0.1193564 0.2769394 0.1346346 0.5265029 0.3166002 1.586368  
## 80 0.1971166 0.1286353 0.2833706 0.1372110 0.5425677 0.3403555 1.619190  
## 81 0.2043592 0.1250621 0.2711128 0.1366120 0.5537755 0.3327745 1.674242  
## PSD95\_N SNCA\_N Ubiquitin\_N pGSK3B\_Tyr216\_N SHH\_N BAD\_N  
## 76 2.436587 0.1681717 1.139154 0.9877739 0.2397524 0.1390524  
## 77 2.354728 0.1747042 1.227851 0.9816546 0.2490305 0.1337874  
## 78 2.437317 0.1560008 1.226103 0.9941456 0.2479560 0.1423236  
## 79 2.382140 0.1450486 1.056498 0.9700673 0.2332252 0.1336368  
## 80 2.386217 0.1509446 1.132737 0.9807979 0.2444693 0.1333582  
## 81 2.380651 0.1505216 1.176354 1.0024839 0.2475161 0.1358669  
## BCL2\_N pS6\_N pCFOS\_N SYP\_N H3AcK18\_N EGR1\_N H3MeK4\_N  
## 76 0.1129261 0.1320008 0.1293628 0.4869115 0.1251522 0.1468649 0.1435166  
## 77 0.1216068 0.1390077 0.1430844 0.4678333 0.1128567 0.1611316 0.1457194  
## 78 0.1302614 0.1348037 0.1476734 0.4625013 0.1164328 0.1605935 0.1428788  
## 79 0.1073210 0.1189823 0.1212896 0.4791095 0.1028311 0.1442380 0.1416812  
## 80 0.1128511 0.1286353 0.1426175 0.4383545 0.1106140 0.1556674 0.1464082  
## 81 0.1100969 0.1279185 0.1338177 0.4613140 0.1142573 0.1559240 0.1421386  
## CaNA\_N  
## 76 1.627181  
## 77 1.562096  
## 78 1.571868  
## 79 1.646608  
## 80 1.607631  
## 81 1.506520

\***Cluster some of your data using EM based clustering.**

# Fitting the data   
fit <- Mclust(ModifiedProtienExpressionData)  
fit

## 'Mclust' model object:  
## best model: ellipsoidal multivariate normal (XXX) with 1 components

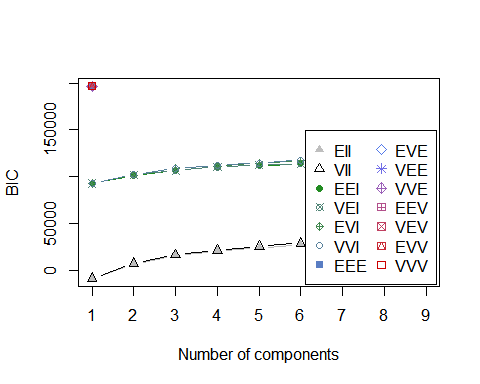
# Getting the summary  
summary(fit)

## ----------------------------------------------------  
## Gaussian finite mixture model fitted by EM algorithm   
## ----------------------------------------------------  
##   
## Mclust XXX (ellipsoidal multivariate normal) model with 1 component:  
##   
## log.likelihood n df BIC ICL  
## 108040.3 552 3080 196634.9 196634.9  
##   
## Clustering table:  
## 1   
## 552

# Applying, summarizing and plotting BIC  
BIC = mclustBIC(ModifiedProtienExpressionData)  
summary(BIC)

## Best BIC values:  
## EEE,1 EEV,1 EVE,1  
## BIC 196634.9 196634.9 196634.9  
## BIC diff 0.0 0.0 0.0

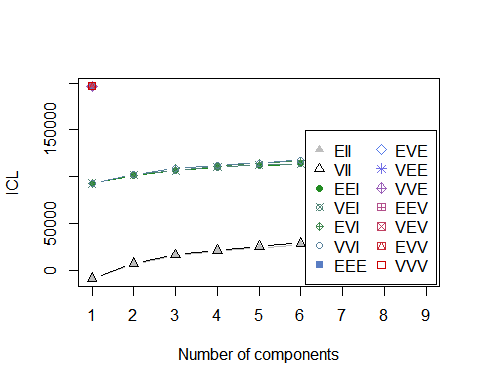
plot(BIC)



# Applying, summarizing and plotting ICL  
ICL = mclustICL(ModifiedProtienExpressionData)  
summary(ICL)

## Best ICL values:  
## EEE,1 EEV,1 EVE,1  
## ICL 196634.9 196634.9 196634.9  
## ICL diff 0.0 0.0 0.0

plot(ICL)

 ***How did you choose a model for EM?***  
**Evaluate the model performance.**  
I choose **mclust** model based clustering as it gives BIC, ICL, degrees of freedom as well as log.likelihood. In EM it is essential we give how the data is generated. The model here has 70 cloumns and there are chances that each might have different distributions. In EM we only include columns with either gaussian, binomail or poisson and not a mixture of those. As the data was large it is not an essential means of clustering.

\***Cluster some of your data using EM based clustering that you also used for k-means,PAM. and hierarchical clustering.** As the BIC value is too high, we can say the model performance is not good, as a good model should have low BIC value. When compaired to other models, k-means, PAM and Heirarchial, this model does not make much sence. There is only one component in the model.