Clustering Example 5: Model-based Clustering of Wine Data

Amy Wagaman, Nathan Carter

July 2020

Load necessary libraries.

```
library(mosaic)
library(mclust)
library(readr)
```

This requires you to have access to the `winedata.txt` file. It is available in the book's GitHub repository at the following URL.

https://github.com/ds4m/ds4m.github.io/tree/master/chapter-5-resources/winedata.txt

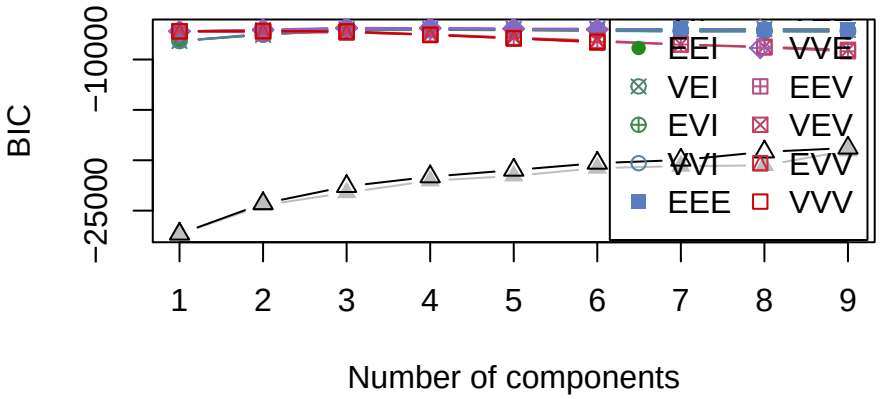
If you run this R code, place the data file in the same folder as the code file.

```
winedata <- read.csv("winedata.txt")
```

For summaries of the variables in the wine data, see the Clustering Example 3 file in the same folder as this one.

Plot BIC for various model-based methods.

```
mclustsol <- mclustBIC(winedata[, -1])
plot(mclustsol)
```



Compare two mixture models for a small range of k values.

This requires you to have access to the `mclustBICs.csv` file. It is available in the book's GitHub repository at the following URL.

https://github.com/ds4m/ds4m.github.io/tree/master/chapter-5-resources/mclustBICs.csv

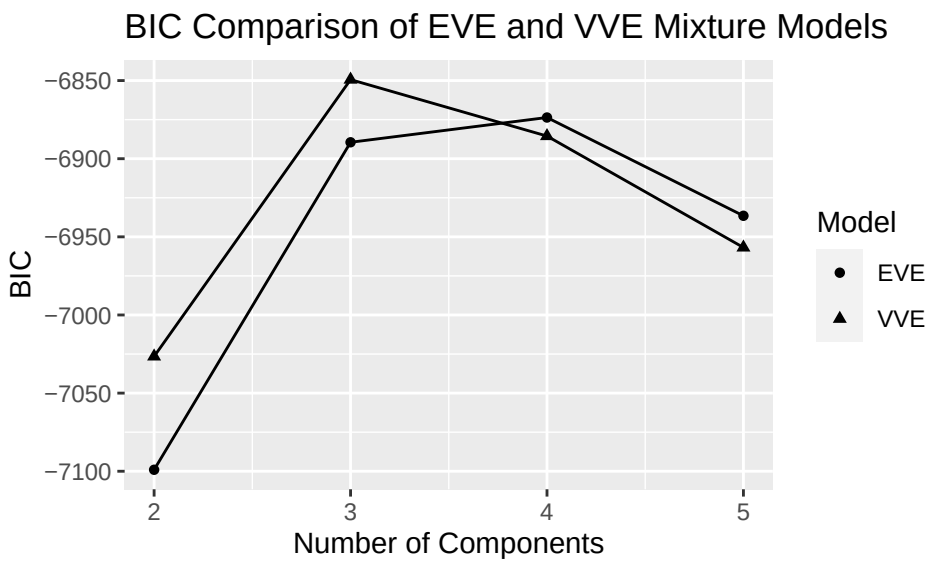
If you run this R code, place the data file in the same folder as the code file.

```
mclustBICs <- read_csv("mclustBICs.csv")
```

```
## Parsed with column specification:
## cols(
##   k = col_double(),
##   BIC = col_double(),
##   Model = col_character()
## )
```

```
mclustBICs2 <- mutate(mclustBICs, Model = factor(Model))
mclustBICs3 <- filter(mclustBICs2, Model == "EVE" | Model == "VVE", k < 6, k > 1.5)

gf_point(BIC ~ k, data = mclustBICs3, shape = ~ Model) %>%
  gf_labs(title = "BIC Comparison of EVE and VVE Mixture Models",
          x = "Number of Components") + geom_line()
```




```
summary(mclustsol)
```

```
## Best BIC values:
##              VVE,3       EVE,4       VVE,4
## BIC      -6849.387 -6873.64376 -6885.51547
## BIC diff     0.000   -24.25637   -36.12808
```

Fit Gaussian finite mixture models to the data.

```
mod1 <- Mclust(winedata[, -1], x = mclustsol)
summary(mod1, parameters = TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVE (ellipsoidal, equal orientation) model with 3 components:
##
##  log-likelihood   n  df       BIC       ICL
##       -3015.333 178 158 -6849.387 -6850.694
##
```

```
## Clustering table:
##  1  2  3
## 59 69 50
##
## Mixing probabilities:
##         1         2         3
## 0.3300315 0.3903124 0.2796561
##
## Means:
##                            [,1]        [,2]        [,3]
## Alcohol                13.7474607  12.2884420  13.1132188
## MalicAcid               2.0093958   1.9349628   3.2824045
## Ash                     2.4524717   2.2414826   2.4395874
## AlcalinityofAsh        17.0056425  20.2329719  21.4025951
## Magnesium             106.2591851  93.9990767 100.0646503
## TotalPhenols            2.8408263   2.2829349   1.6680931
## Flavanoids              2.9836246   2.1105880   0.7895086
## NonflavanoidPhenols     0.2892543   0.3671547   0.4401329
## Proanthocyanins         1.8992977   1.6341503   1.1665817
## ColorIntensity          5.5361641   3.0977312   7.2299462
## Hue                     1.0619246   1.0596060   0.6915761
## ODDilutedWines          3.1576577   2.8052205   1.6972509
## Proline              1116.9517345 515.2032715 633.5416560
##
## Variances:
## [,,1]
##                            Alcohol     MalicAcid          Ash AlcalinityofAsh
## Alcohol              0.2058329968  -0.004024864 -0.009036939    -0.082837142
## MalicAcid           -0.0040248637   0.490091195  0.018910721     0.331081359
## Ash                 -0.0090369389   0.018910721  0.055228146     0.341310246
## AlcalinityofAsh     -0.0828371420   0.331081359  0.341310246     5.971450304
## Magnesium            0.1068697208  -0.991646554  0.571126451     2.017347189
## TotalPhenols         0.0052597707  -0.023068454  0.009333359     0.018738318
## Flavanoids           0.0065310325  -0.026137373  0.016676313     0.131007162
## NonflavanoidPhenols  0.0009463792   0.011285445  0.007911553     0.044245330
## Proanthocyanins      0.0149987005  -0.038452295 -0.006737493    -0.002275193
## ColorIntensity       0.1653294086  -0.148232480 -0.026834517    -0.255590507
## Hue                  0.0037878962  -0.022582987  0.007238772     0.004632895
## ODDilutedWines      -0.0121794833   0.014950157  0.014614274     0.168893091
## Proline             17.4628451708 -56.678967719 -3.337718902   -52.752038513
##                       Magnesium  TotalPhenols    Flavanoids
## Alcohol              0.10686972  0.0052597707  0.0065310325
## MalicAcid           -0.99164655 -0.0230684544 -0.0261373726
## Ash                  0.57112645  0.0093333589  0.0166763126
## AlcalinityofAsh      2.01734719  0.0187383181  0.1310071620
## Magnesium          139.89193989  0.9193181647  0.7113866941
## TotalPhenols         0.91931816  0.0971100170  0.0842443327
## Flavanoids           0.71138669  0.0842443327  0.1349705248
## NonflavanoidPhenols -0.11721737 -0.0061705519 -0.0069115449
## Proanthocyanins      1.30197911  0.0298378638  0.0625656802
## ColorIntensity       1.28678476  0.0493652310  0.1144048483
## Hue                  0.21238589  0.0007562286 -0.0006350309
## ODDilutedWines      -0.06193507  0.0110876746  0.0226408419
## Proline            732.49780801 13.8868978185 10.6322090581
```

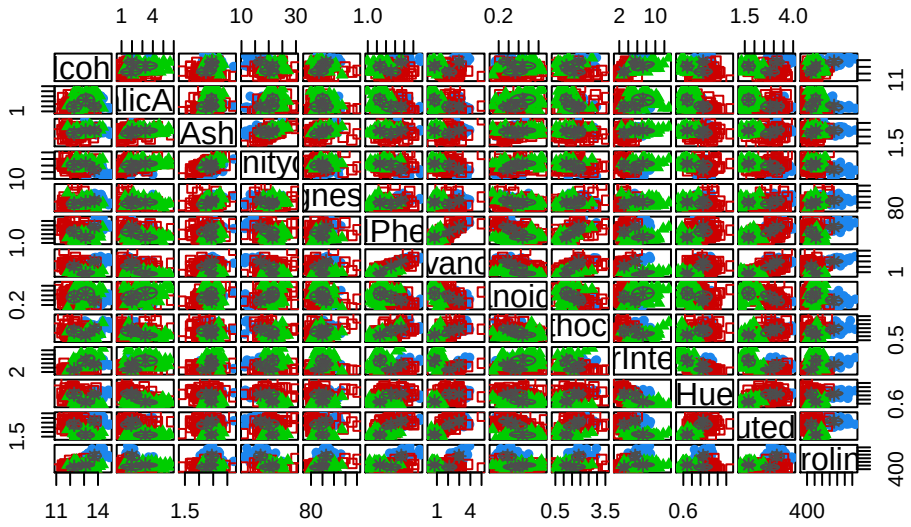
```
##                      NonflavanoidPhenols Proanthocyanins ColorIntensity
## Alcohol                    0.0009463792     0.014998700   0.1653294086
## MalicAcid                  0.0112854453    -0.038452295  -0.1482324804
## Ash                        0.0079115530    -0.006737493  -0.0268345172
## AlcalinityofAsh            0.0442453302    -0.002275193  -0.2555905072
## Magnesium                 -0.1172173734     1.301979114   1.2867847621
## TotalPhenols              -0.0061705519     0.029837864   0.0493652310
## Flavanoids                -0.0069115449     0.062565680   0.1144048483
## NonflavanoidPhenols        0.0072891701    -0.007172903  -0.0004530542
## Proanthocyanins           -0.0071729033     0.164934604   0.1228256872
## ColorIntensity            -0.0004530542     0.122825687   1.2827176196
## Hue                        0.0008467983     0.002076420  -0.0181024064
## ODDilutedWines            -0.0092570244     0.011252864  -0.0639638945
## Proline                   -3.0396915543    23.979311959  89.6000335505
##                                Hue ODDilutedWines      Proline
## Alcohol              0.0037878962   -0.012179483    17.462845
## MalicAcid           -0.0225829875    0.014950157   -56.678968
## Ash                  0.0072387724    0.014614274    -3.337719
## AlcalinityofAsh      0.0046328950    0.168893091   -52.752039
## Magnesium            0.2123858908   -0.061935066   732.497808
## TotalPhenols         0.0007562286    0.011087675    13.886898
## Flavanoids          -0.0006350309    0.022640842    10.632209
## NonflavanoidPhenols  0.0008467983   -0.009257024    -3.039692
## Proanthocyanins      0.0020764204    0.011252864    23.979312
## ColorIntensity      -0.0181024064   -0.063963894    89.600034
## Hue                  0.0133148261   -0.001197288     8.488849
## ODDilutedWines      -0.0011972880    0.137646988    -8.638897
## Proline              8.4888492803   -8.638897429 48054.698849
## [,,2]
##                          Alcohol     MalicAcid          Ash AlcalinityofAsh
## Alcohol              0.290957199   0.041494976 -0.015172597     -0.13922279
## MalicAcid            0.041494976   0.959616485  0.031392560      0.55750842
## Ash                 -0.015172597   0.031392560  0.100832278      0.66495472
## AlcalinityofAsh     -0.139222786   0.557508416  0.664954718     11.66688422
## Magnesium           -0.151719216  -0.652094100  1.080065801      4.59938073
## TotalPhenols        -0.040256458  -0.002113848  0.029119508      0.04787564
## Flavanoids          -0.060147746  -0.004874673  0.039501254      0.25897013
## NonflavanoidPhenols  0.007548535   0.015915807  0.013514996      0.08314979
## Proanthocyanins     -0.040749045  -0.009251307 -0.006148156      0.01893329
## ColorIntensity       0.125146552  -0.080664399 -0.033632185     -0.40238687
## Hue                  0.004211772  -0.034179000  0.007228388      0.02452953
## ODDilutedWines      -0.042241674   0.026423741  0.028446620      0.31725455
## Proline              8.626596290 -27.985876625 -1.659845775    -26.09488891
##                        Magnesium TotalPhenols   Flavanoids NonflavanoidPhenols
## Alcohol              -0.15171922 -0.040256458 -0.060147746         0.007548535
## MalicAcid            -0.65209410 -0.002113848 -0.004874673         0.015915807
## Ash                   1.08006580  0.029119508  0.039501254         0.013514996
## AlcalinityofAsh       4.59938073  0.047875643  0.258970125         0.083149785
## Magnesium           234.48630651  1.361037876  1.054059766        -0.148757464
## TotalPhenols          1.36103788  0.264992918  0.257799261        -0.022078478
## Flavanoids            1.05405977  0.257799261  0.402859513        -0.030553714
## NonflavanoidPhenols  -0.14875746 -0.022078478 -0.030553714         0.013721881
## Proanthocyanins       1.84347909  0.127681628  0.219205900        -0.025794233
## ColorIntensity        0.53161273  0.007950886  0.047959574         0.003634641
```

```
## Hue                   0.21105543 -0.002551795 -0.004842160         0.001549258
## ODDilutedWines        0.05715343  0.110329268  0.160712163        -0.021481045
## Proline             359.20264305  6.843821755  5.239795551        -1.499604179
##                     Proanthocyanins ColorIntensity          Hue ODDilutedWines
## Alcohol                -0.040749045    0.125146552  0.004211772   -0.042241674
## MalicAcid              -0.009251307   -0.080664399 -0.034179000    0.026423741
## Ash                    -0.006148156   -0.033632185  0.007228388    0.028446620
## AlcalinityofAsh         0.018933288   -0.402386869  0.024529529    0.317254548
## Magnesium               1.843479090    0.531612726  0.211055426    0.057153426
## TotalPhenols            0.127681628    0.007950886 -0.002551795    0.110329268
## Flavanoids              0.219205900    0.047959574 -0.004842160    0.160712163
## NonflavanoidPhenols    -0.025794233    0.003634641  0.001549258   -0.021481045
## Proanthocyanins         0.397162067    0.055879900 -0.003299242    0.121953613
## ColorIntensity          0.055879900    1.025931508 -0.020505202   -0.072269929
## Hue                    -0.003299242   -0.020505202  0.041007971   -0.002452333
## ODDilutedWines          0.121953613   -0.072269929 -0.002452333    0.215689168
## Proline                11.823144132   44.247495205  4.190420761   -4.267192790
##                          Proline
## Alcohol                 8.626596
## MalicAcid             -27.985877
## Ash                    -1.659846
## AlcalinityofAsh       -26.094889
## Magnesium             359.202643
## TotalPhenols            6.843822
## Flavanoids              5.239796
## NonflavanoidPhenols    -1.499604
## Proanthocyanins        11.823144
## ColorIntensity         44.247495
## Hue                     4.190421
## ODDilutedWines         -4.267193
## Proline             23730.772720
## [,,3]
##                         Alcohol     MalicAcid          Ash AlcalinityofAsh
## Alcohol              0.364795175   0.064640215 -0.014255292    -0.035993225
## MalicAcid            0.064640215   1.202742989  0.009619805     0.182864164
## Ash                 -0.014255292   0.009619805  0.034969282     0.269439543
## AlcalinityofAsh     -0.035993225   0.182864164  0.269439543     4.687651663
## Magnesium           -0.090609764  -0.365698545  0.624920131     2.699827797
## TotalPhenols         0.031616437  -0.055151917  0.023026263     0.034940535
## Flavanoids           0.083433862  -0.103556511 -0.009829708     0.135032979
## NonflavanoidPhenols  0.004430543   0.021256335  0.001934607     0.032945039
## Proanthocyanins      0.077373081  -0.072496459 -0.000510812     0.034275071
## ColorIntensity       0.769007331  -0.245044458 -0.061093951     0.022939765
## Hue                 -0.017715605  -0.035787308  0.004035577     0.005182554
## ODDilutedWines      -0.023346744  -0.013926442  0.013095158     0.123688178
## Proline              4.869717266 -15.801209810 -0.937506525   -14.738807160
##                       Magnesium TotalPhenols   Flavanoids NonflavanoidPhenols
## Alcohol              -0.09060976  0.031616437  0.083433862        4.430543e-03
## MalicAcid            -0.36569855 -0.055151917 -0.103556511        2.125633e-02
## Ash                   0.62492013  0.023026263 -0.009829708        1.934607e-03
## AlcalinityofAsh       2.69982780  0.034940535  0.135032979        3.294504e-02
## Magnesium           134.94975622  0.783495550  0.608435055       -8.524479e-02
## TotalPhenols          0.78349555  0.112711973  0.005760379        3.875283e-03
## Flavanoids            0.60843506  0.005760379  0.179195550       -5.819478e-03
```

```
## NonflavanoidPhenols  -0.08524479  0.003875283 -0.005819478        1.605783e-02
## Proanthocyanins       1.06047025  0.047307104  0.066022386       -4.840697e-05
## ColorIntensity        0.29043216  0.184482532  0.584663452        2.722327e-02
## Hue                   0.12098019  0.000579047 -0.016193650       -1.250589e-03
## ODDilutedWines        0.03632290 -0.003086967 -0.016412517       -6.006693e-03
## Proline             202.78506111  3.863769830  2.957297389       -8.467936e-01
##                    Proanthocyanins ColorIntensity          Hue ODDilutedWines
## Alcohol               7.737308e-02     0.76900733 -0.017715605   -0.023346744
## MalicAcid            -7.249646e-02    -0.24504446 -0.035787308   -0.013926442
## Ash                  -5.108120e-04    -0.06109395  0.004035577    0.013095158
## AlcalinityofAsh       3.427507e-02     0.02293976  0.005182554    0.123688178
## Magnesium             1.060470e+00     0.29043216  0.120980192    0.036322896
## TotalPhenols          4.730710e-02     0.18448253  0.000579047   -0.003086967
## Flavanoids            6.602239e-02     0.58466345 -0.016193650   -0.016412517
## NonflavanoidPhenols  -4.840697e-05     0.02722327 -0.001250589   -0.006006693
## Proanthocyanins       1.223608e-01     0.49064410 -0.010005284   -0.015100071
## ColorIntensity        4.906441e-01     5.78108574 -0.166574834   -0.231453078
## Hue                  -1.000528e-02    -0.16657483  0.015609471    0.007313253
## ODDilutedWines       -1.510007e-02    -0.23145308  0.007313253    0.103382839
## Proline               6.674755e+00    24.97418230  2.366371149   -2.409193145
##                         Proline
## Alcohol                4.8697173
## MalicAcid            -15.8012098
## Ash                   -0.9375065
## AlcalinityofAsh      -14.7388072
## Magnesium            202.7850611
## TotalPhenols           3.8637698
## Flavanoids             2.9572974
## NonflavanoidPhenols   -0.8467936
## Proanthocyanins        6.6747551
## ColorIntensity        24.9741823
## Hue                    2.3663711
## ODDilutedWines        -2.4091931
## Proline            13399.5878181
```

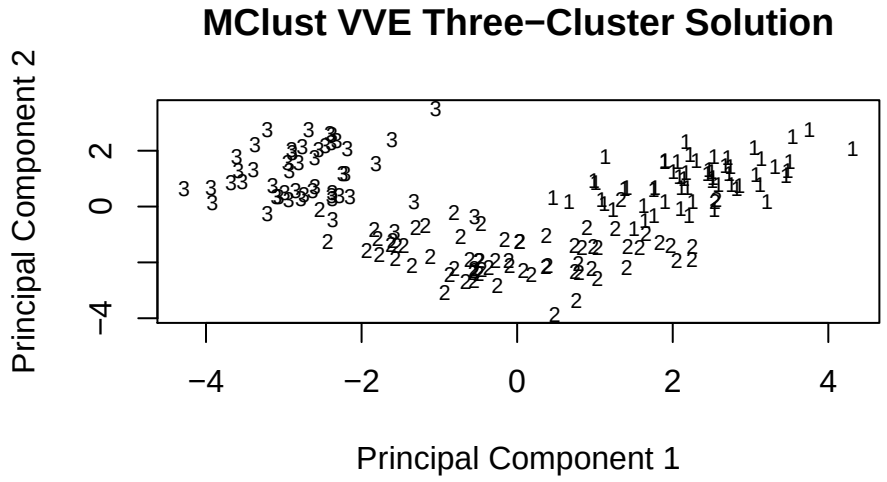
Plotting with so many variables does not lead to readable results.

```
plot(mod1, what = "classification")
```



Plot the clustering from the VVE model with $k = 3$, using principal component space.

```
set.seed(1)
winePCAs <- princomp(winedata[, -c(1)], cor = TRUE)
plot(winePCAs$scores[, 1:2], type = "n",
     xlab = "Principal Component 1", ylab = "Principal Component 2",
     main="MClust VVE Three-Cluster Solution")
text(winePCAs$scores[, 1:2], labels = mod1$classification, cex = 0.7)
```



Compare that solution to the original wine cultivars.

```
tally(winedata$Cultivar ~ mod1$classification)
```

```
##                   mod1$classification
## winedata$Cultivar  1  2  3
##                 1 59  0  0
##                 2  0 69  2
##                 3  0  0 48
```