MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*†

Chris Fraley and Adrian E. Raftery

Technical Report No. 504

Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195-4322 USA

September 2006

MCLUST is a contributed R package for normal mixture modeling and model-based clustering. It provides functions for parameter estimation via the EM algorithm for normal mixture models with a variety of covariance structures, and functions for simulation from these models. Also included are functions that combine model-based hierarchical clustering, EM for mixture estimation and the Bayesian Information Criterion (BIC) in comprehensive strategies for clustering, density estimation and discriminant analysis. There is additional functionality for displaying and visualizing the models along with clustering and classification results. A number of features of the software have been changed in this version, and the functionality has been expanded to include regularization for normal mixture models via a Bayesian prior. A web page with related links including license information can be found at http://www.stat.washington.edu/mclust.

*Funded by National Institutes of Health grant 8 R01 EB002137–02 and by Office of Naval Research grant N00014-01-1-0745.

†Many users have offered comments and suggestions leading to improvements in the MCLUST documentation and software over the years. Special thanks go to Ron Wehrens for porting and maintaining earlier versions of MCLUST for R, and making a number of significant contributions.

Contents

List of Tables

List of Figures

1 Overview

The MCLUST software [10, 12] has evolved to include the following features:

- Model-based clustering (model and number of clusters selected via BIC).
- Normal mixture modeling via EM for ten covariance structures.
- Simulation from parameterized Gaussian mixtures.
- Discriminant analysis via MclustDA.
- Model-based hierarchical clustering for four covariance structures.
- Displays, including uncertainty plots and random projections.

This manuscript describes Version 3 of MCLUST for R, which allows regularization in normal mixture models via a Bayesian prior [13]. A number of features of the software have been changed as well, to reflect evoloution in its use. A comprehensive treatment of the methods used in MCLUST can be found in [11, 13].

This version of MCLUST is available as a contributed package (mclust) in the R language and can be obtained from http://cran.r-project.org/. Follow the instructions for installing R packages on your machine, and then do

```
> library(mclust)
```

inside R in order to use the software. Throughout this manual it will be assumed that these steps have been taken before running the examples.

2 Model-Based Cluster Analysis

MCLUST provides functionality for cluster analysis combining model-based hierarchical clustering (section 5), EM for Gaussian mixture models (section 3), and BIC (section 4).

2.1 Basic Cluster Analysis Example using Mclust

As an illustration, consider the two-dimensional faithful dataset (included in the R language distribution) shown in Figure 1. The following command performs a cluster analysis of the faithful dataset, and prints a summary of the result:

```
> faithfulMclust <- Mclust(faithful)
> faithfulMclust

 best model: EEE with 3 components
```

In this case, the best model is an equal-covariance model with 3 components or clusters. The clustering results can be displayed as follows:

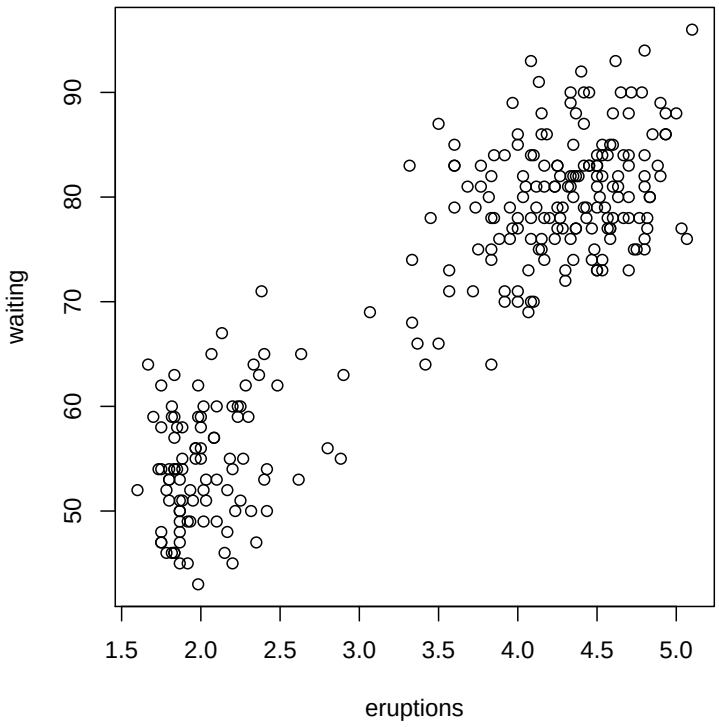


Figure 1: The two-dimensional `faithful` dataset.

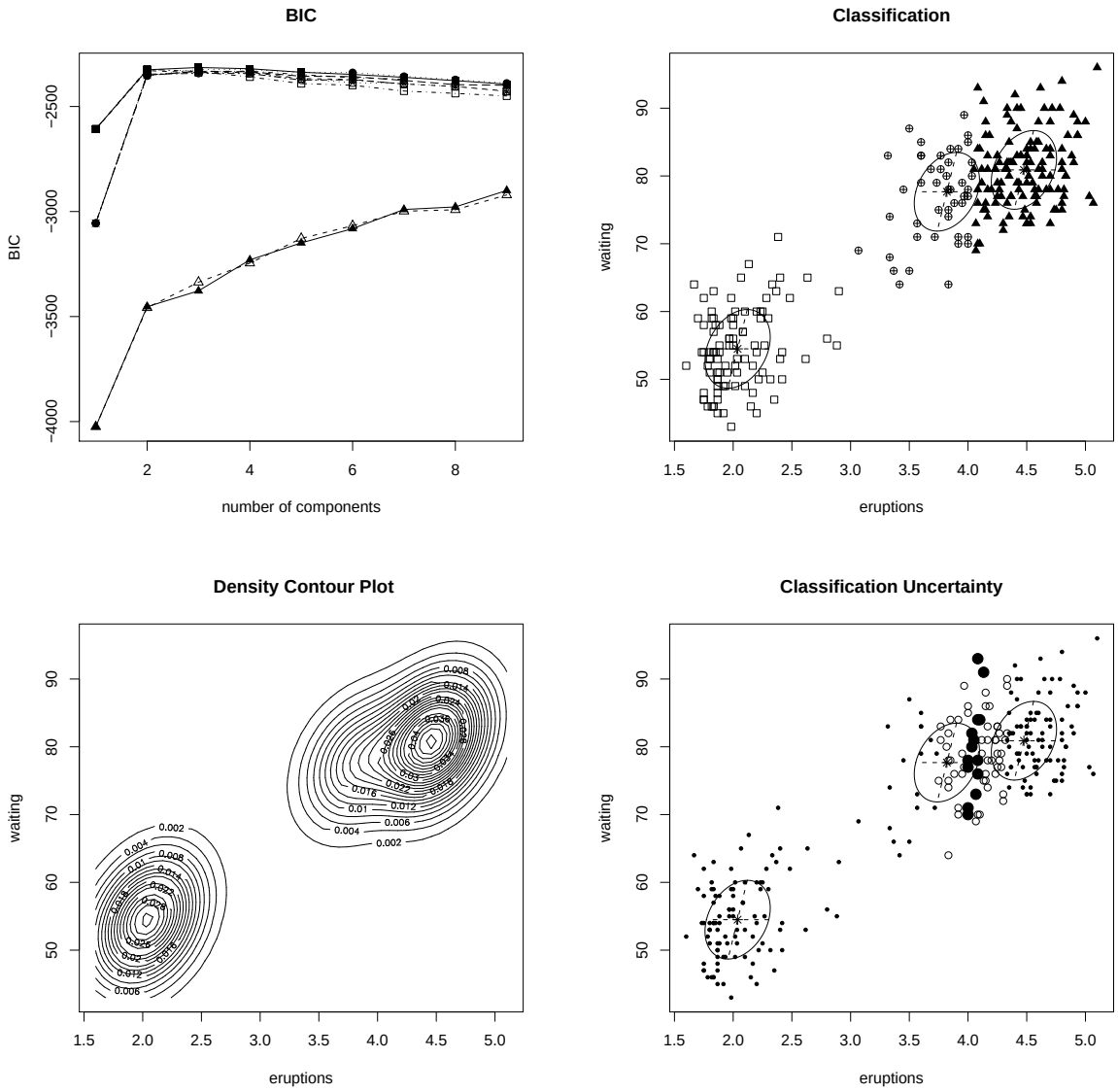


Figure 2: Plots associated with the function `Mclust` for the `faithful` dataset with the default arguments. Clockwise from upper left: BIC, classification, uncertainty, density. The ellipses superimposed on the classification and uncertainty plots correspond to the covariances of the components.

Table 1: Parameterizations of the covariance matrix Σ_k currently available in `MCLUST` for hierarchical clustering (HC) and/or EM for multidimensional data. ('•' indicates availability).

identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
`E`		•	•	(univariate)	equal		
`V`		•	•	(univariate)	variable		
`EII`	λI	•	•	Spherical	equal	equal	NA
`VII`	$\lambda_k I$	•	•	Spherical	variable	equal	NA
`EEI`	λA		•	Diagonal	equal	equal	coordinate axes
`VEI`	$\lambda_k A$		•	Diagonal	variable	equal	coordinate axes
`EVI`	λA_k		•	Diagonal	equal	variable	coordinate axes
`VVI`	$\lambda_k A_k$		•	Diagonal	variable	variable	coordinate axes
`EEE`	$\lambda D A D^T$	•	•	Ellipsoidal	equal	equal	equal
`EEV`	$\lambda D_k A D_k^T$		•	Ellipsoidal	equal	equal	variable
`VEV`	$\lambda_k D_k A D_k^T$		•	Ellipsoidal	variable	equal	variable
`VVV`	$\lambda_k D_k A_k D_k^T$	•	•	Ellipsoidal	variable	variable	variable

```
> plot(faithfulMclust, data = faithful)
```

The corresponding plots are shown in Figure 2. The covariance structures defining the models available in `MCLUST` are summarized in Table 1; these models in `MCLUST` are explained in detail in section A. The symbols used in the BIC plots to represent the various models are shown in Table 2.

The input to `Mclust` includes the number of mixture components and the covariance structures to consider. By default, `Mclust` compares BIC values for parameters optimized for up to nine components and all ten covariance structures currently available in `MCLUST`. The output includes the parameters of the maximum-BIC model (where the maximum is taken over all of the models and numbers of components considered), and the corresponding classification and uncertainty.

The object produced by `Mclust` is a list with a number of elements describing the selected model. The names of these elements can be displayed as follows:

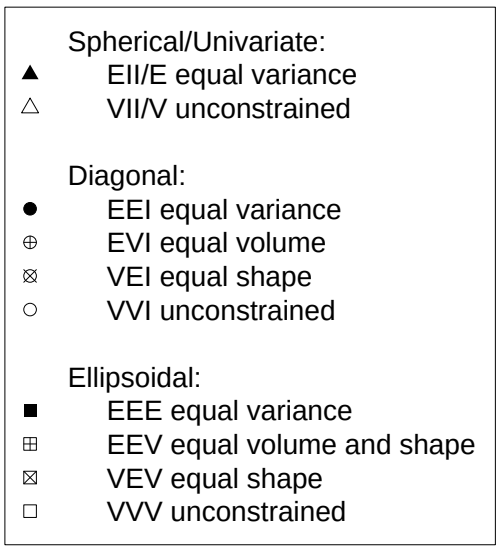
```
> names(faithfulMclust)
 [1] "modelName"      "n"              "d"              "G"
 [5] "BIC"            "bic"            "loglik"         "parameters"
 [9] "z"              "classification" "uncertainty"
```

A detailed description is provided in the `Mclust` help file.

2.2 `mclustBIC` and its `summary` function

To do further analysis, for example to see the results for the same dataset, but for a different set of models and/or different numbers of components, `Mclust` could be rerun. However this approach could involve unnecessary repetition of computations and could also take

Table 2: Symbols used to represent the different models in the BIC plots and their meaning.



considerable time when the dataset is large or the process is to be repeated many times. An alternative approach is to split the analysis into several parts using function `mclustBIC`.

For the `faithful` dataset, the following sequence of commands produces the same clustering result as the call to `Mclust`.

```
> faithfulBIC <- mclustBIC(faithful)
> faithfulSummary <- summary(faithfulBIC, data = faithful)
> faithfulSummary
classification table:
  1   2   3
130  97  45

best BIC values:
    EEE,3     EEE,4     VVV,2
-2314.386 -2320.207 -2322.192
```

Although the method used for printing is different, `faithfulSummary` has the same component names as `faithfulMclust`, except that it does not include `"BIC"`, the table of BIC values, which comprise the S-PLUS object `faithfulBIC` computed by `mclustBIC`:

```
> faithfulBIC

 BIC:
        EII       VII       EEI       VEI       EVI       VVI       EEE
1 -4024.721 -4024.721 -3055.835 -3055.835 -3055.835 -3055.835 -2607.623
```

```
2 -3452.998 -3458.300 -2354.601 -2350.607 -2352.618 -2346.065 -2325.220
3 -3377.712 -3336.542 -2323.008 -2332.698 -2332.204 -2342.371 -2314.386
4 -3230.246 -3245.732 -2323.676 -2331.829 -2334.756 -2343.068 -2320.207
5 -3149.389 -3128.214 -2337.730 -2348.284 -2355.885 -2374.251 -2336.967
6 -3081.401 -3067.580 -2338.116 -2363.073 -2357.745 -2372.728 -2347.296
7 -2990.334 -2998.496 -2356.458 -2370.071 -2375.850 -2393.086 -2361.216
8 -2978.088 -2991.847 -2371.814        NA -2395.992        NA -2376.920
9 -2899.778 -2920.951 -2388.617        NA -2399.085        NA -2393.733
         EEV       VEV       VVV
1 -2607.623 -2607.623 -2607.623
2 -2329.116 -2325.416 -2322.192
3 -2338.986 -2329.352 -2333.894
4 -2336.750 -2342.472 -2359.216
5 -2366.985 -2367.785 -2390.985
6 -2371.741 -2387.155 -2398.905
7 -2392.961 -2391.166 -2426.431
8 -2404.598 -2404.932 -2437.612
9 -2427.039 -2428.375 -2449.787

> plot(faithfulBIC)
```

The missing values are models and numbers of clusters for which parameter values could not be fit (using the default initialization). For multivariate data, the default initialization for all models uses the classification from hierarchical clustering based on an unconstrained model. For univariate data, the default is to divide the data into quantiles for initialization.

The `summary` method for `mclustBIC` allows specification of the models and numbers of clusters over which the best model is to be chosen, alllowing models other than the maximum BIC model to be extracted and analyzed.

2.3 Extended Cluster Analysis Example

As an example of an extended analysis, consider the `wreath` data shown in Figure 3. There are 1000 bivariate observations simultaed from a 14-component model in which the component covariance matrices are of equal size and shape, but differ in orientation. The BIC values can be obtained with a call to `mclustBIC` and then plotted:

```
> wreathBIC <- mclustBIC(wreath)
> plot(wreathBIC)
```

Refering to the BIC plot (shown on the left in Figure 4), the maximum BIC appears to be outside the range of the default values for the number of components in `mclustBIC` (and `Mclust`). More components (for example, up to 20) can be considered in the analysis without recomputingprevious results:

```
> wreathBIC <- mclustBIC(wreath, G = 1:20, x = wreathBIC)
> plot(wreathBIC, G = 10:20)
> summary(wreathBIC, wreath)
```

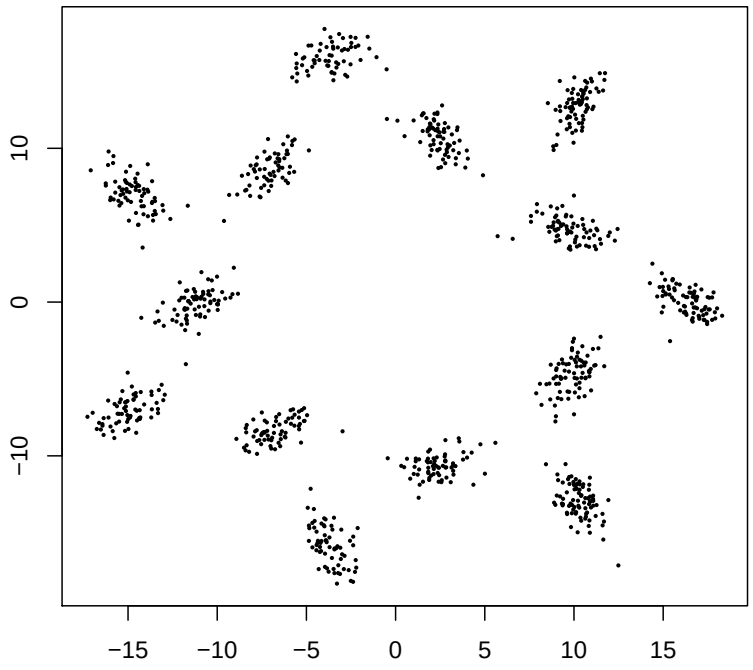


Figure 3: The two-dimensional `wreath` dataset, which consists of 1000 observations simulated from a 14-component normal mixture in which the component covariance matrices are of equal size and shape, but differ in orientation.

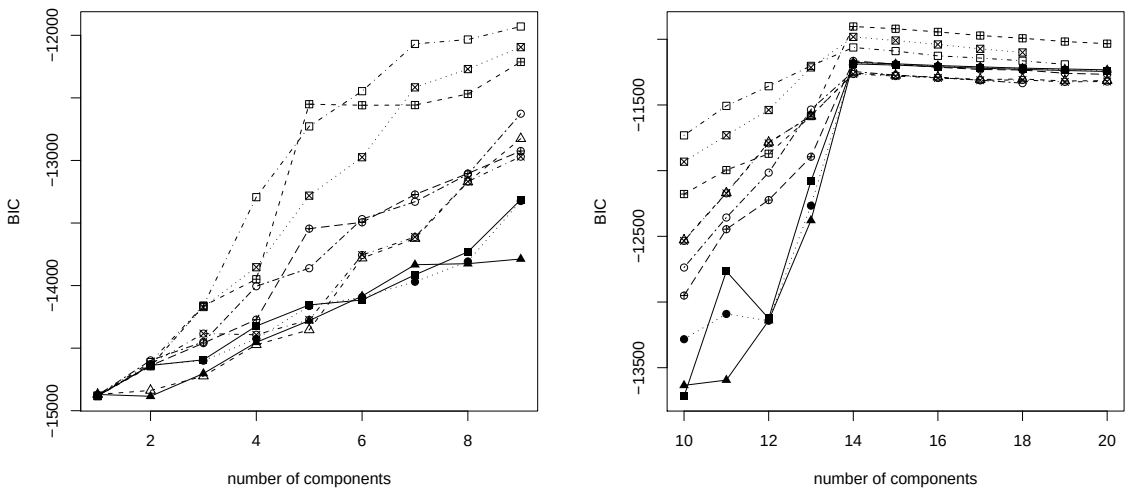


Figure 4: BIC for `wreath` dataset. LEFT: BIC for all models and up to 9 components (the default in `mclustBIC` and `Mclust`). RIGHT: BIC for 10:20 components, all models. There is a clear peak for all models at 14 components.

The BIC plot is shown on the right in Figure 4. Use `summary` to obtain the best model according to BIC, a 14-component `EEV` model is chosen, which is in agreement with how the data was simulated.

```
> wreathModel <- summary(wreathBIC, data = wreath)
> wreathModel

classification table:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
74 69 63 74 68 70 71 66 83 77 66 77 61 81

best BIC values:
   EEV,14    EEV,15    EEV,16
-10902.77 -10919.96 -10944.09
```

The model for the `wreath` dataset is shown in Figure 5. The `summary` function can also be used to restrict the set of models and/or numbers of clusters over which the best model is chosen according to BIC. For example, the following commands produce the best spherical model for the `wreath` data:

```
> wreathSphericalModel <- summary(wreathBIC, data = wreath,
                                  modelNames = c("EII", "VII"))
> wreathSphericalModel

classification table:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
```

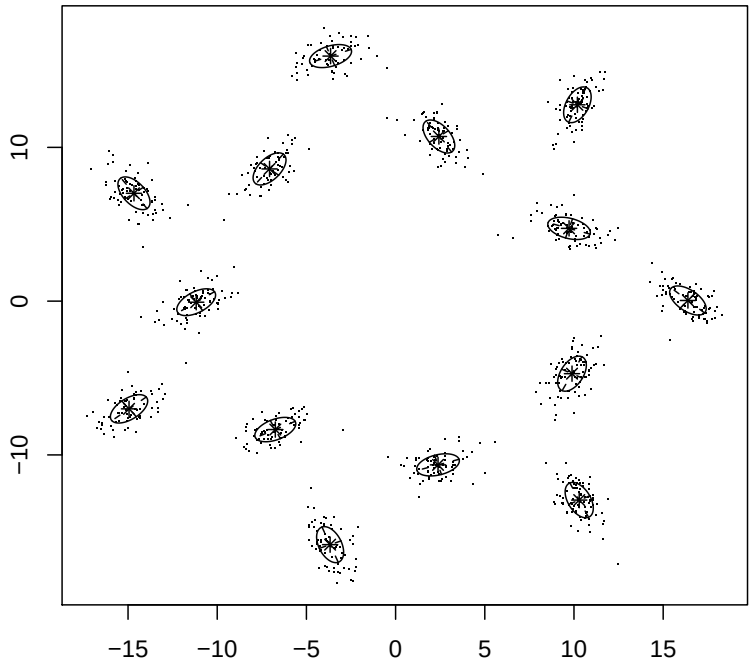


Figure 5: The 14-component `EEV` (equal size and shape) model obtained for the `wreath` dataset. The ellipses superimposed on the plot correspond to the covariances of the components.

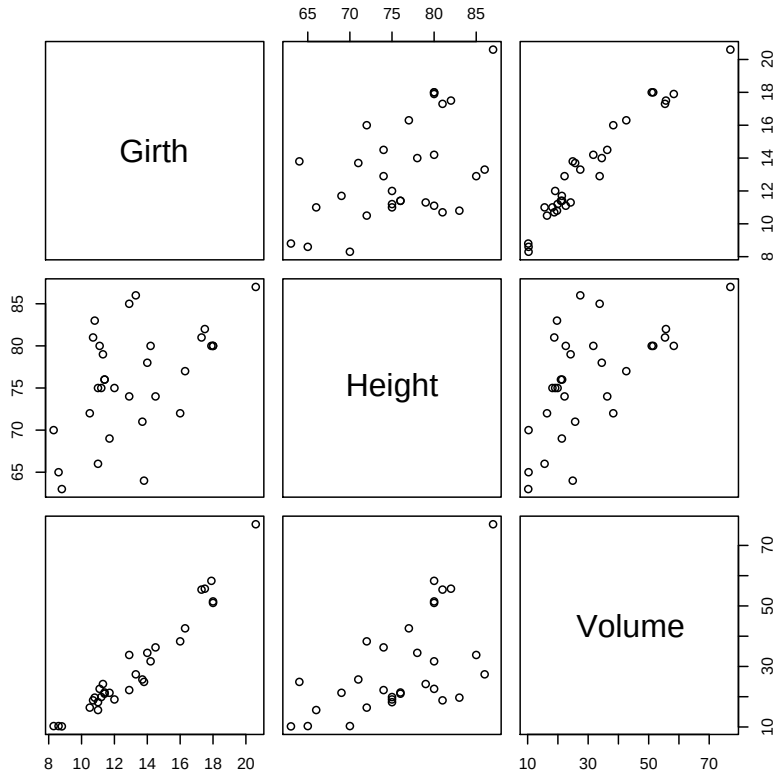


Figure 6: Pairs plot of `trees` dataset.

```
75 69 63 74 68 70 71 65 83 77 66 77 61 81

best BIC values:
   EII,14    EII,15    EII,16
-11175.90 -11186.51 -11200.04
```

2.4 Regularizing with a Prior

It is now possible in `MCLUST` to specify a prior distribution to regularize the fit to the data. We illustrate the use of a prior on the `trees` dataset (included in the `R` language distribution), for which a pairs plot is shown in Figure 6.

The following commands compute and plot the BIC curves for the `trees` dataset provided in `R` with and without a `prior`. Without the prior, the BIC plot shows a number of jagged peaks, and many BIC values are missing for some models due to failure in the EM computations caused by singularity and/or shrinking components. With the prior, the BICs are smoother and there are fewer EM failures. See Figure 7.

```
> treesBIC <- mclustBIC(trees) # default (no prior)
> plot(treesBIC)
> treesBICprior <- mclustBIC(trees, prior = priorControl())
```

```
> plot(treesBICprior)
```

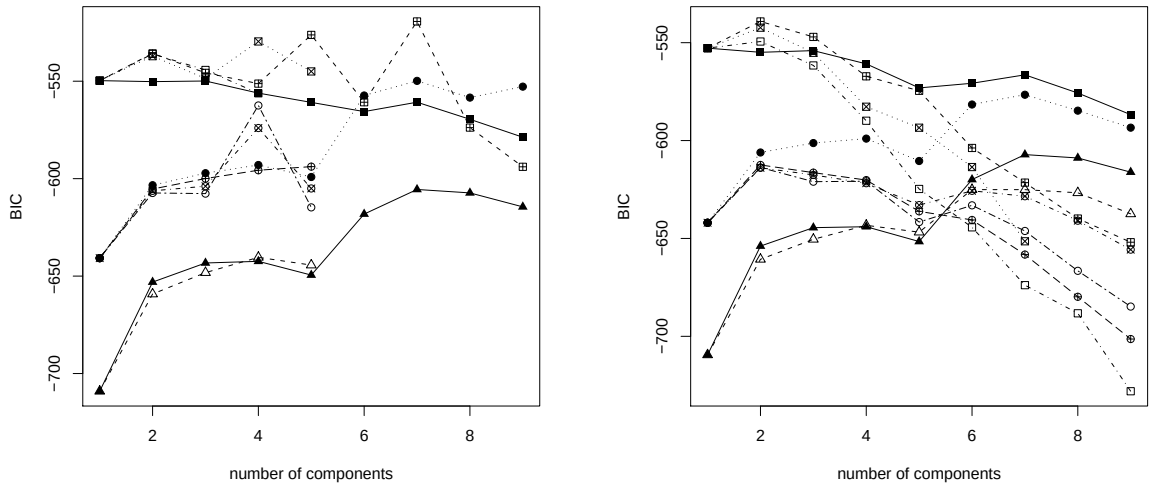


Figure 7: BIC without (left) and with the prior for the `trees` dataset.

A function `priorControl` is provided in `MCLUST` for specifying the prior and its parameters. When called with its defaults, it invokes another function called `defaultPrior` which can serve as a template for specifying alternative priors. An example of the result of a call to `defaultPrior` is shown below.

```
> defaultPrior(trees, G=2, modelName = "VVV")
$shrinkage
[1] 0.01

$mean
   Girth   Height   Volume
13.24839 76.00000 30.17097

$dof
[1] 5

$scale
           Girth   Height    Volume
Girth   6.203797  6.54109  31.42755
Height  6.541090 25.57640  39.47333
Volume 31.427545 39.47333 170.21710
```

For more detail on the prior and its specification, see Section A.3.

2.5 Clustering with Noise and Outliers

MCLUST allows model-based clustering with noise. In the following example, Poisson noise is added to the faithful dataset. A random initial estimate is used for the noise.

```
> b <- apply( faithful, 2, range)
> nNoise <- 500
> set.seed(0)
> poissonNoise <- apply(b, 2, function(x, n)
+                       runif(n, min = min(x)-.1, max = max(x)+.1), n = nNoise)
> faithfulNdata <- rbind(faithful, poissonNoise)
> set.seed(0)
> faithfulNoiseInit <- sample(c(TRUE,FALSE),size=nrow(faithful)+nNoise,
                                replace=TRUE,prob=c(3,1))
> faithfulNbic <- mclustBIC(faithfulNdata,
                            initialization = list(noise = faithfulNoiseInit))
> faithfulNsummary <- summary(faithfulNbic, faithfulNdata)
> faithfulNsummary

classification table:
  0   1   2
521 143 108

best BIC values:
    EVI,2     VVI,2     EEI,2
-7996.437 -7998.035 -8000.251
```

The data and classification are shown in Figure 8.

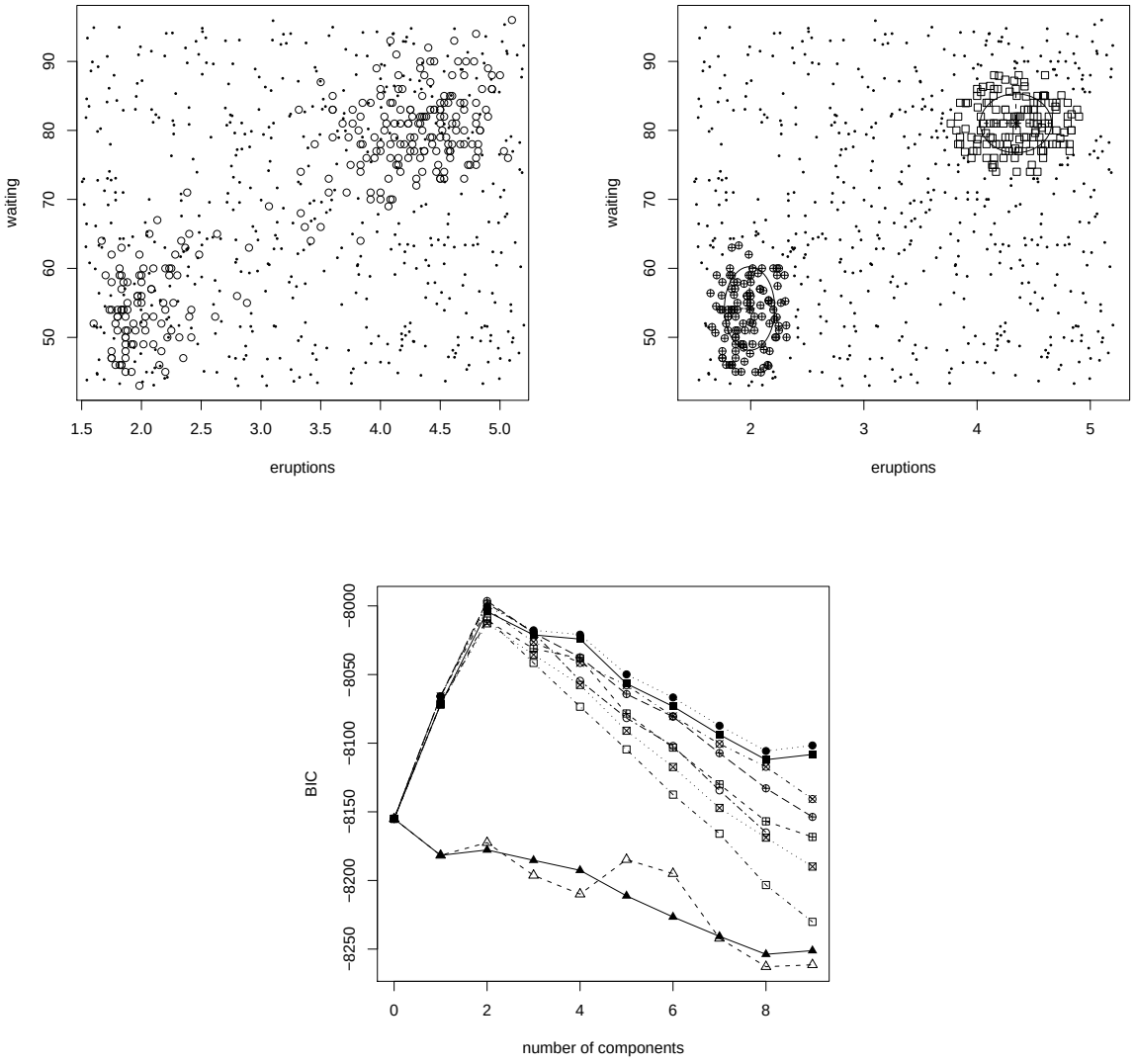


Figure 8: Cluster analysis of the `faithful` dataset with added Poisson noise. Upper Left: A projection the 272 obervations of the `faithful` dataset (circles) with 500 Poisson noise points (small dots). Upper Right: `MCLUST` classification starting with random noise estimate. Lower: BIC.

2.6 Further Considerations in Cluster Analysis

Clustering can be affected by parameters settings such as convergence tolerances within the clustering functions, although the defaults are most often adequate. It is also possible do model-based clustering starting with parameter estimates, conditional probabilities, or classifications other than those produced by model-based hierarchical clustering. The functions provided for mixture estimation (Section 3) and BIC (Section 4) can be used for this purpose.

Finally, it is important to take into account numerical issues in cluster analysis. The EM computations break down when the covariance corresponding to one or more components becomes ill-conditioned (singular or nearly singular). In general they cannot proceed if clusters contain only a few observations or if the observations they contain are very nearly colinear. Computations may also fail when one or more mixing proportions shrink to negligible values. The EM functions in `MCLUST` compute and monitor the conditioning of the covariances, and an error condition is issued (unless such warnings are turned off) when the associated covariance appears to be nearly singular, as determined by a threshold with the default value `emControl()$eps`.

3 EM for Mixture Models

`MCLUST` provides iterative EM (Expectation-Maximization) methods for maximum likelihood estimation in parameterized Gaussian mixture models. In the models considered here, an iteration of EM consists of an 'E'-step, which computes a matrix z such that z_{ik} is an estimate of the conditional probability that observation i belongs to group k given the current parameter estimates, and an 'M-step', which computes parameter estimates given z.

`MCLUST` functions `em` and `me` implement the EM algorithm for parameterized Gaussian mixtures. Function `em` starts with the E-step; besides the data and model specification, the model parameters (means, covariances, and mixing proportions) proportions must be provided. Function `me` starts with the M-step; besides the data and model specification, the conditional probabilities z must be provided. The output for both are the maximum likelihood estimates of the model parameters and z.

3.1 Individual `E` and `M` Steps

Functions `estep` and `mstep` implement the individual steps of the EM iteration. Conditional probabilities z and the log likelihood can be recovered from parameters via `estep`, while parameters can be recovered from conditional probabilities z using `mstep`. Below we apply `mstep` and `estep` to the `iris` dataset (included in the `R` language distribution).

```
> ms <- mstep( modelName = "VVV", data = iris[,-5], z = unmap(iris[,5]))
> names(ms)
[1] "modelName"  "prior"      "n"          "d"          "G"
[6] "z"          "parameters"

> es <- estep( modelName = "VVV", data = iris[,-5],
               parameters = ms$parameters)
```

```
> names(es)
[1] "modelName"  "n"          "d"          "G"          "z"
[6] "parameters" "loglik"
```

In this example, the initial estimate of z for the M-step is a matrix of indicator variables corresponding to a discrete classification (`iris[,5]`). The function `unmap` converts a discrete classification into the corresponding indicator variables. `MCLUST` allows specification of a prior, for which the EM algorithm will compute a posterior mode. See Sections 2.4 and A.3 for more details. In Section 7.1, we show how to use `mstep` and `estep` for discriminant analysis.

3.2 Uncertainty

The uncertainty in the classification associated with conditional probabilities z can be obtained by subtracting the probability of the most likely group for each observation from 1:

```
> meVVViris <- me(modelName = "VVV", data = iris[,-5], z = unmap(iris[,5]))
> uncer <- 1 - apply( meVVViris$z, 1, max)
```

The `R` function `quantile` applied to the uncertainty gives a measure of the quality of the classification.

```
> quantile(uncer)
          0%          25%          50%          75%         100%
0.000000e+00 0.000000e+00 1.907041e-08 1.392060e-03 3.361880e-01
```

In this case the indication is that the majority of observations are well classified. Note, however, that when groups intersect, uncertain classifications would be expected in the overlapping regions.

When a true classification is known, the relative uncertainty of misclassified observations can be displayed by function `uncerPlot`, as is done below for the `iris` example (see Figure 9):

```
> uncerPlot(z = meVVViris$z, truth = iris[,5])
```

It is also possible to plot an uncertainty curve for one-dimensional data (see Section 8) or an uncertainty surface for two-dimensional data (see Section 9.1).

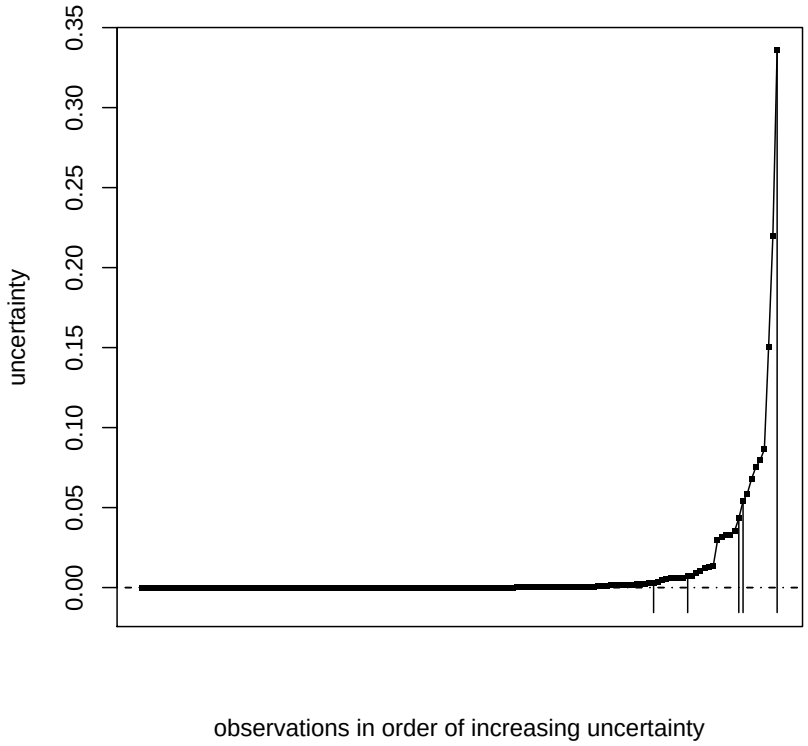


Figure 9: Uncertainty plot for the the 3-cluster mixture model fit of the `iris` dataset via EM based on unconstrained Gaussian mixtures. The vertical lines indicate misclassified observations. The plot was created with function `uncerPlot`, and shows the relative uncertainty of misclassified observations.

3.3 Control Parameters

Besides the initial values and the prior, other parameters can influence the outcome of `em` or `me`. These include:

`tol` Iteration convergence tolerance. The default is `emControl()$tol=c(1.e-5,`$\sqrt{\epsilon_M}$`)`, where ϵ_M is the relative machine precision, which has the value `2.220446e-16` on IEEE compliant machines. The first value is the tolerance for relative convergence of the loglikelihood in the EM algorithm, and the second value is the relative parameter convergence tolerance for the M-step for those models that have an iterative M-step (`"VEI"`, `"VEE"`, `"VVE"`, `"VEV"`).

`eps` A tolerance for terminating iterations due to ill-conditioning, such as near singularity in covariance matrices. The default is `emControl()$eps` which is set to the relative machine precision ϵ_M.

A function `emControl` is provided in `MCLUST` for setting these parameters and supplying default values. Although these are in some sense hidden by the defaults, they may have a significant effect on results in some instances and should be taken into consideration in analysis.

4 Bayesian Information Criterion

`MCLUST` provides a function `bic` to compute the Bayesian Information Criterion (BIC) [21] given the maximized loglikelihood for model, the data dimensions, and the number of components in the model. The BIC is the value of the maximized loglikelihood with a penalty for the number of parameters in the model, and allows comparison of models with differing parameterizations and/or differing numbers of clusters. In general the larger the value of the BIC, the stronger the evidence for the model and number of clusters (see, e.g. [11]). The following shows the BIC calculation in `MCLUST` for the 3-cluster classification the `iris` dataset with the unconstrained variance model:

```
> meVVViris <- me(modelName = "VVV", data = iris[,-5], z = unmap(iris[,5]))

> bic( modelName = "VVV", loglik = meVVViris$loglik,
       n = nrow(iris), d = ncol(iris[,-5]), G = 3)
[1] -580.8397
```

5 Model-Based Hierarchical Clustering

`MCLUST` provides functions `hc` for model-based hierarchical agglomeration, and `hclass` for determining the resulting classifications. Function `hc` implements fast methods based on the multivariate normal classification likelihood [8]. We use the `iris` dataset distributed with `R` in our example. Figure 10 is a pairs plot of the `iris` dataset in which the three species are differentiated by symbol, obtained by the following command:

```
> clPairs(data = iris[,-5], classification = iris[,5])
```

Below we apply the hierarchical clustering algorithm for unconstrained covariances (`VVV`) to the `iris` dataset:

```
> hcVVViris <- hc(modelName = "VVV", data = iris[,-5])
```

The classification produced by `hc` for various numbers of clusters can be obtained with `hclass`. For example, for the classifications corresponding to 2 and 3 clusters:

```
> cl <- hclass(hcVVViris, 2:3)
```

The classifications can be displayed with the data using `clPairs`:

```
> clPairs(data = iris[,-5], classification = cl[,"2"])
> clPairs(data = iris[,-5], classification = cl[,"3"])
```

More options for displaying clustering and classification results are discussed in Section 9. The 3-group classification can be compared with the known 3-group classification into species, which is given in the 5th column of the `iris` data, using function `classError`:

```
> classError(cl[,"3"], truth = iris[,5])
$misclassifiedPoints
 [1] 102 107 114 115 120 122 124 127 128 134 139 143 147 150

$errorRate
[1] 0.09333333
```

Function `hc` starts by default with every observation of the data in a cluster by itself, and continues until all observations are merged into a single cluster. Arguments `partition` and `minclus` can be used to initialize the process at a chosen nontrivial partition, and to stop it before it reaches the final stage of merging.

Function `hc` for model-based hierarchical clustering based on the unconstrained (`VVV`) model is used to obtain initial values for the model-based clustering functions `Mclust` and `mclustBIC`.

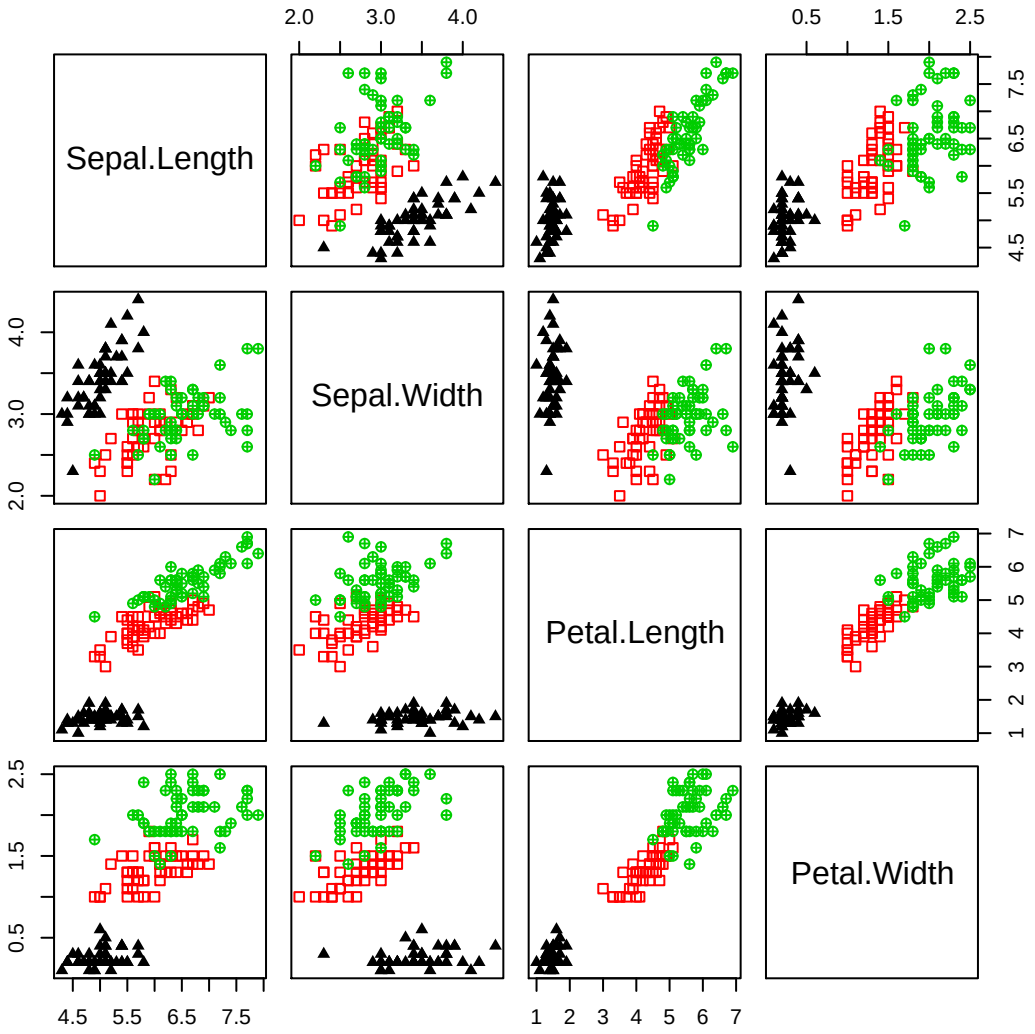


Figure 10: Pairs plot of the `iris` dataset showing classifcation into species.

6 Density Estimation

The clustering capabilities of MCLUST can also be viewed as a general strategy for multivariate density estimation. After applying the clustering functions to fit a model to the data, function dens can be used to get the density of a given point relative to that model. As an example, we use the two-dimensional faithful dataset (see Figure 1).

First, we use Mclust or (mclustBIC and summary) to get a model for the data, as was done in Section 2:

```
> faithfulMclust <- Mclust(faithful)
```

The faithful dataset is two dimensional, so for plotting the density can be computed over a grid. Function grid1 forms a one dimensional grid of a given size over a given range of values, while grid2 forms a two dimensional grid given two sequences of values.

```
> apply(faithful, 2, range)
     eruptions waiting
[1,]       1.6      43
[2,]       5.1      96
> x <- grid1( 100, range = range(faithful$eruptions))
> y <- grid1( 100, range = range(faithful$waiting))
> xy <- grid2(x,y)
> xyDens <- dens(modelName = faithfulMclust$modelName, data = xy,
                 parameters = faithfulMclust$parameters)
> xyDens <- matrix(xyDens, nrow = length(x), ncol = length(y))
```

The faithful dataset is two-dimensional, so the result can be plotted using S-PLUS functions contour, persp, or image.

```
> par(pty = "s")
> Z <- log(xyDens)
> persp(x = x, y = y, z = Z, box = FALSE)
> contour(x = x, y = y, z = Z, nlevels = 10)
> image(x = x, y = y, z = Z)
```

These plots are shown in Figure 11.

Probably the most common application for density estimation is discriminant analysis, for which a detailed discussion is given in Section 7.

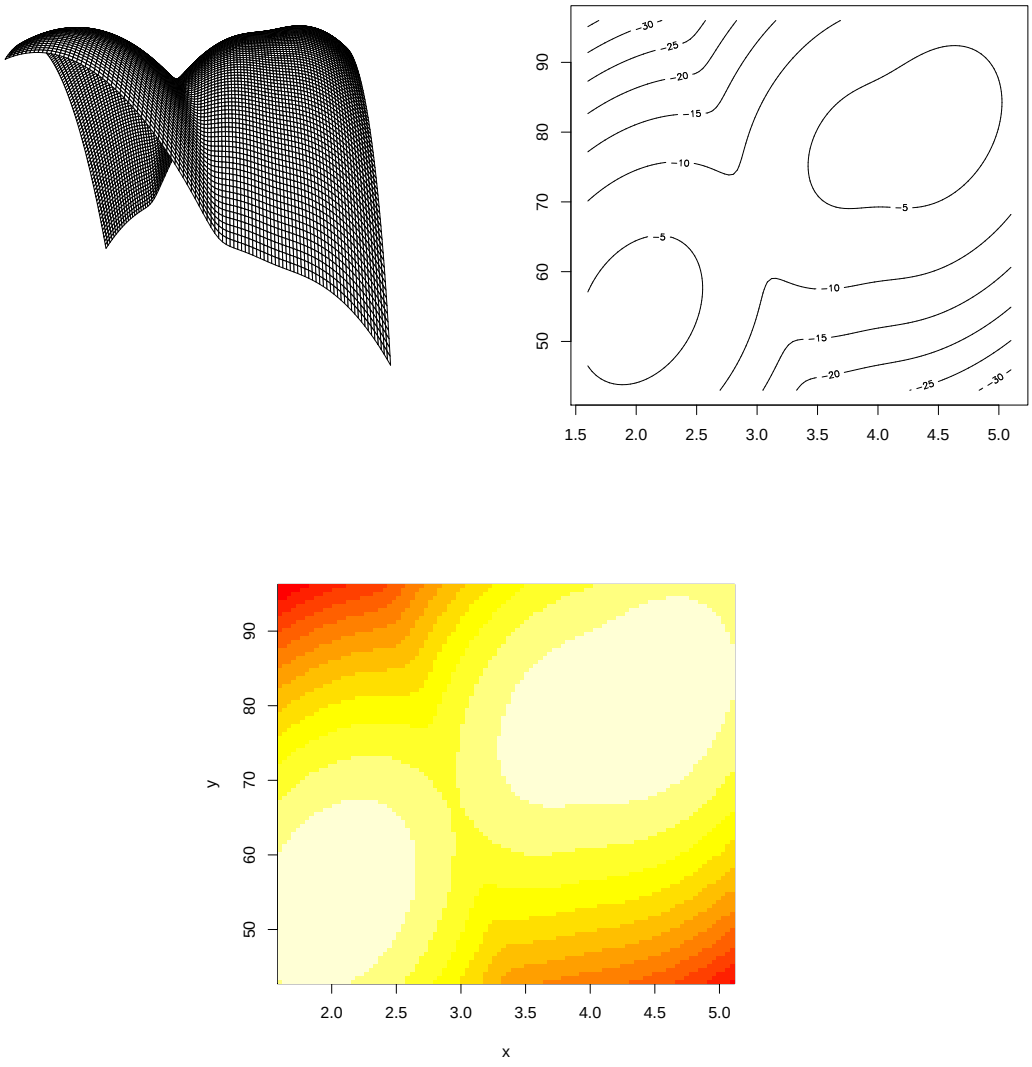


Figure 11: Perspective, contour and image plots of an `MCLUST` density estimate for the `faithful` dataset.

7 Discriminant Analysis

In discriminant analysis, observations of known classification are used to classify others. MCLUST provides several approaches to discriminant analysis. We demonstrate some possible methods applied to the faithful dataset using the three-group model-based classification shown in Figure 2 as the ground truth:

```
> faithfulMclust <- Mclust(faithful)
> faithfulClass <- faithfulMclust$classification
```

7.1 Discriminant Analysis using mstep and estep

MCLUST functions mstep and estep implementing the individual steps of the EM algorithm for Gaussian mixtures can be used for discriminant analysis. The idea is to produce a density estimate for the training data which is a mixture model, in which each known class is modeled by a single Gaussian term.

First, the parameterization giving the best model fit to the training data must be chosen. Most commonly, this would be done by leave-one-out cross validation. Leaving out one training observation at a time, function cv1EMtrain fits each model using mstep, then classifies the observation that was left out using estep. The output of cv1EMtrain is the error rate for each model; that is, the fraction of left-out observations correctly classified by the model fit to the remaining observations.

Using the odd numbered observations in the faithful dataset as a training set, the result is:

```
> odd <- seq(from=1, to=nrow(faithful), by=2)
> round(cv1EMtrain(data = faithful[odd,], labels = faithfulClass[odd]),3)
  EII   VII   EEI   VEI   EVI   VVI   EEE   EEV   VEV   VVV
0.162 0.162 0.037 0.037 0.044 0.044 0.015 0.015 0.015 0.022
```

The crossvalidation error achieves a minimum for the elliptical, equal shape models EEE, EEV, and VEV. Of these we choose the most parsimonious model EEE. When there are two training classes, the EEE model corresponds to linear discriminant analysis, while the VVV model corresponds to quadratic discriminant analysis (e.g. [17]).

To classify the even data points, we first compute the parameters corresponding to the EEE model for the odd data points using mstep, then use estep to get conditional probabilities z and a classification:

```
> modelEEE <- mstep(modelName = "EEE", data=faithful[odd,],
                    z=unmap(faithfulClass[odd]))
> classEEE <- map(estep(modelName = "EEE", data=faithful,
                        parameters = modelEEE$parameters)$z)
> classError(classEEE[odd], faithfulClass[odd])$errorRate
[1] 0.007352941

> even <- seq(from=2, to=nrow(faithful), by=2)
```

```
> classError(classEEE[even], faithfulClass[even])$errorRate
[1] 0.007352941
> classError(classEEE[even], faithfulClass[even])$misclassified
[1] 17
```

The error rates for the training [odd-numbered] data and the test [even-numbered] data are identical (.735%); two data points are misclassified:

```
> classError(classEEE, faithfulClass)$misclassified
[1] 34 71
```

The classification and the misclassified observations are shown in Figure 12. Not surprisingly, the misclassified observations fall in the reason where the cluster overlap.

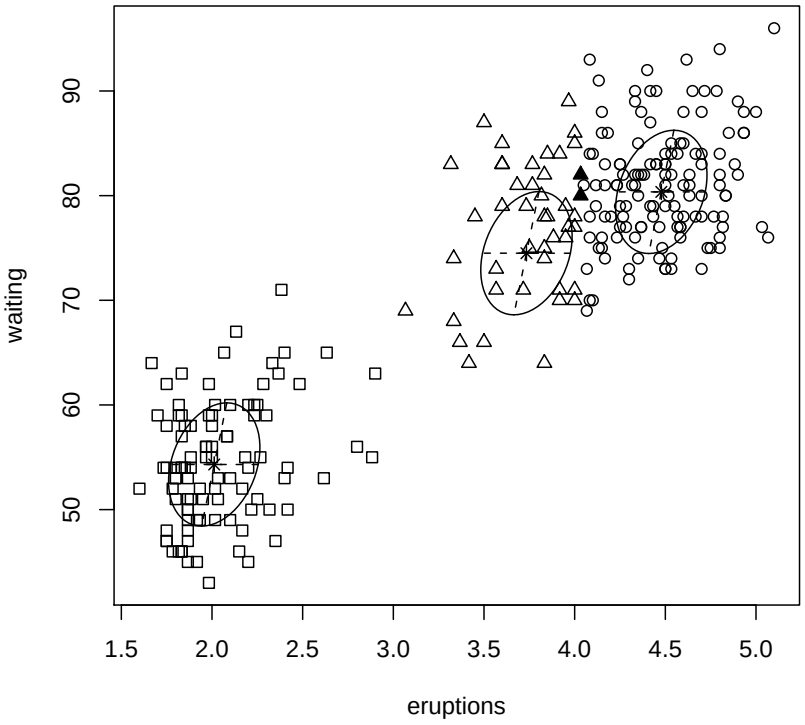


Figure 12: Classification errors from discriminant analysis for the `faithful` dataset using `mstep` and `estep`. Filled symbols are the misclassified data points.

Another option for model selection that is faster to compute than crossvalidation is to select the best fitting model via BIC after using `mstep` to fit each model to the training data. A function `bicEMtrain` is provided within `MCLUST` for this purpose. For the `faithful` dataset, BIC for the models fitted to the odd-numbered observations is:

```
> round(bicEMtrain(faithful[odd,], labels = faithfulClass[odd]),0)
```

```
  EII   VII   EEI   VEI   EVI   VVI   EEE   EEV   VEV   VVV
-1761 -1771 -1187 -1196 -1194 -1204 -1183 -1193 -1202 -1210
```

BIC chooses `EEE` as the best model, so in this case the training and test errors are the same as for crossvalidation, and the classification results are as shown in Figure 12.

Although in this case crossvalidation and BIC happen to choose the same model, for other datasets the models selected, and hence the discriminant results, could be different. Cross-validation is much more computationally intensive procedure for model-selection than BIC, although timing comparisons between `cv1EMtrain` and `bicEMtrain` should not be considered a valid algorithmic comparison because there are more efficient ways to compute crossvalidation using updating schemes and compiled code.

7.2 Mixture Discriminant Analysis via MclustDA

In Section 7.1, discriminant analysis was accomplished modeling the training data by a mixture density with a single Gaussian component for each class. That section also showed how to choose the appropriate cross-cluster constraints to give the lowest training error rate using either leave-one-out crossvalidation or BIC. A more flexible alternative is to use model-based clustering to fit a Gaussian mixture model as a density estimate for each class in the training set. We illustrate the methods in this section with the 2-group model from model-based clustering for the `faithful` dataset as ground truth:

```
> faithfulBIC2 <- mclustBIC(faithful, G=2)
> faithfulClass2 <- summary(faithfulBIC2, faithful)$classification
```

7.2.1 `mclustDA`

If both training and test sets are given in advance, function `mclustDA` can be used for discriminant analysis. Its input is the training data and associated class labels, and the test data (optionally with labels). The output of `mclustDA` includes the mixture models for the training data, the classification of both the test data and training data under the model, posterior probabilities for the test data, and the training error rate.

```
> faithfulMclustDA <- mclustDA(train = list(data = faithful[odd,],
                                    labels = faithfulClass2[odd]),
                               test = list(data = faithful[even,],
                                    labels = faithfulClass2[even]))
XXX EEI
  1   2
> faithfulMclustDA

Modeling Summary:
  trainClass mclustModel numGroups
1          1         XXX         1
2          2         EEI         2

Test Classification Summary:
```

```
  1   2
101  35

Training Classification Summary:

 1  2
75 61

Training Error: 0

Test Error: 0.007352941
```

The error rates for `mclustDA` classification are 0% and .735% for the training [odd-numbered] and test [even-numbered] data, respectively.

These discriminant analysis results can be plotted as follows:

```
> plot(discrim, trainData = faithful[odd,], testData = faithful[even,])
```

Figure 13 shows some plots of the results: The training models are shown in Figure 14.

7.2.2 `mclustDAtrain` and `mclustDAtest`

Often more flexibility is required in discriminant analysis. For example, a suitable training set may need to be chosen and/or it may be desirable to test additional data after a training density has already been established. Since training typically takes much more time that testing, it can be advantageous to separate training and testing computations. Function `mclustDAtrain` allows users to choose training model parameterizations, and selects from among all available models as a default. The output of `mclustDAtrain` is a list, each element being the model for each class.

In the simplest case, a single Gaussian could be fit to each training class. This is similar to the discriminant analysis procedure of Section 7.1, except that in `MclustDA` a model for each class of the training data is chosen separately, instead of choosing a parameterized mixture model (which may have cross-cluster constraints) for all of the training data. `MclustDA` uses BIC (see section 4), which adds a penalty term to the maximized loglikelihood that increases with the number of parameters, to select the model.

By default, `mclustDAtrain` will fit up to nine components for each possible model. Results for the odd-numbered observations in the `fauthful` dataset are as follows:

```
> faithfulTrain <- mclustDAtrain(data = faithful[odd,],
                                 labels = faithfulClass2[odd])
XXX EEI
  1   2
```

The training models are shown in Figure 14.

The density of observations under the training models can be obtained using `mclustDAtest`, while the classification and posterior probabilities of the test or other data can be recovered

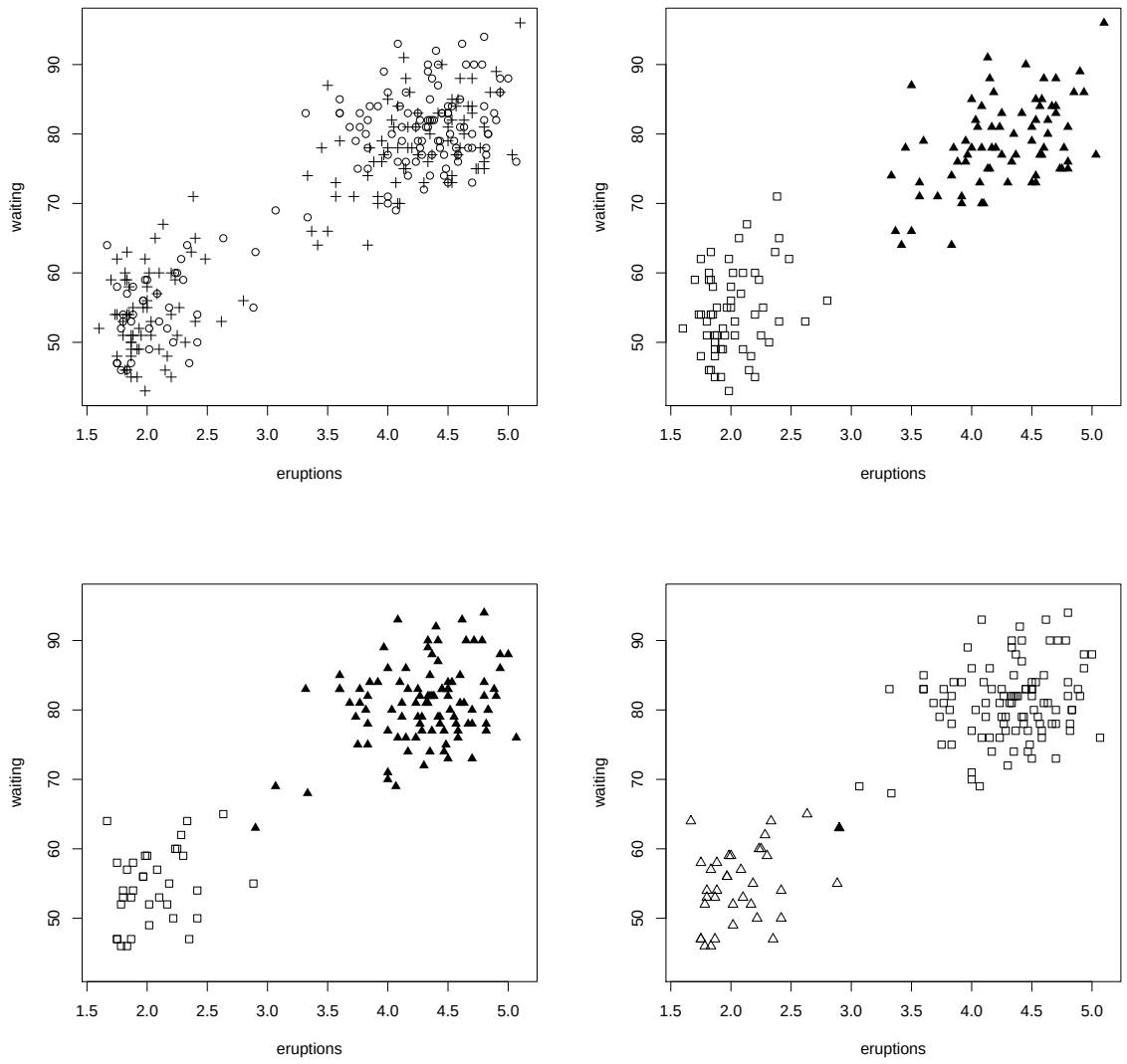


Figure 13: Plots associated with `mclustDA` on the `faithful` dataset. Upper Left: the training [odd-numbered/circles] and test [even-numbered/crosses] `faithful` data. Upper Right: the training data with known classification. Lower Left: the `mclustDA` classification of the test data. Lower Right: the errors (filled symbols) in using the `mclustDA` model to classify the test data.

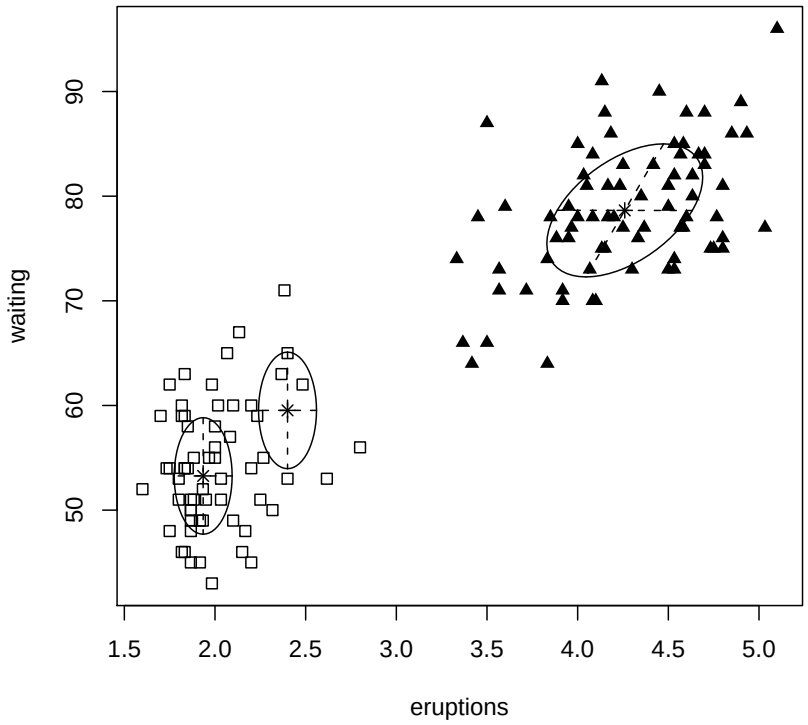


Figure 14: `mclustDA` training models for the odd numbered observations of the `faithful` dataset, using the two-group classification from model-based clustering as ground truth. One of the classes is modeled by a two-group equal variance diagonal model, and the other by a single unconstrained normal.

from the `summary` function for `mclustDAtest`. The test (even-numbered) classfication and error can be obtained as follows:

```
> faithfulEvenTest <- mclustDAtest(models=faithfulTrain, data=faithful[even,])

> names(summary(faithfulEvenTest))
[1] "classification" "z"

> classError(summary(faithfulEvenTest)$classification,
                     faithfulClass2[even])$errorRate

[1] 0.007352941
```

The training (odd-numbered) classfication and error can be obtained as follows:

```
> faithfulOddTest <- mclustDAtest(models=faithfulTrain, data=faithful[odd,])

> classError(summary(faithfulOddTest)$classification,
                     faithfulClass2[odd])$errorRate
[1] 0
```

8 One-Dimensional Data

The MCLUST functions for clustering, density estimation and discriminant analysis can be applied to one-dimensional as well as multidimensional data. Analysis is somewhat simplified since there are only two possible models — equal variance (denoted E) or varying variance (denoted V).

8.1 Clustering

Cluster analysis for one-dimensional data can be carried out as for two and higher dimensions. As an example, we use the precip dataset (included in the R language distribution):

```
> precipMclust <- Mclust(precip)

> plot(precipMclust, precip)
```

Figure 15 shows the BIC, classification, uncertainty, and density for this example.

The analysis can also be divided into two parts: BIC computation via mclustBIC and model computation via summary, as shown below for the rivers dataset (included in the R language distribution):

```
> riversBIC <- mclustBIC(rivers)
> plot(riversBIC)
> riversModel <- summary(riversBIC, rivers)
> riversModel

classification table:
 1  2  3
76 52 13

best BIC values:
      V,3       V,4       V,5
-2015.579 -2022.513 -2035.102
```

There is a special plotting function mclust1Dplot for one-dimensional model-based clustering. As an example, we compare graphical results the 2-component maxmimum BIC model with the 3-component model:

```
> riversModel2 <- summary(riversBIC, rivers, G = 2)

> mclust1Dplot(data = rivers, what = "classification",
               parameters=riversModel$parameters, z=riversModel$z)
> mclust1Dplot(data = rivers, what = "density",
               parameters=riversModel$parameters, z=riversModel$z)
> abline(v = riversModel$parameters$mean, lty = 3)
```

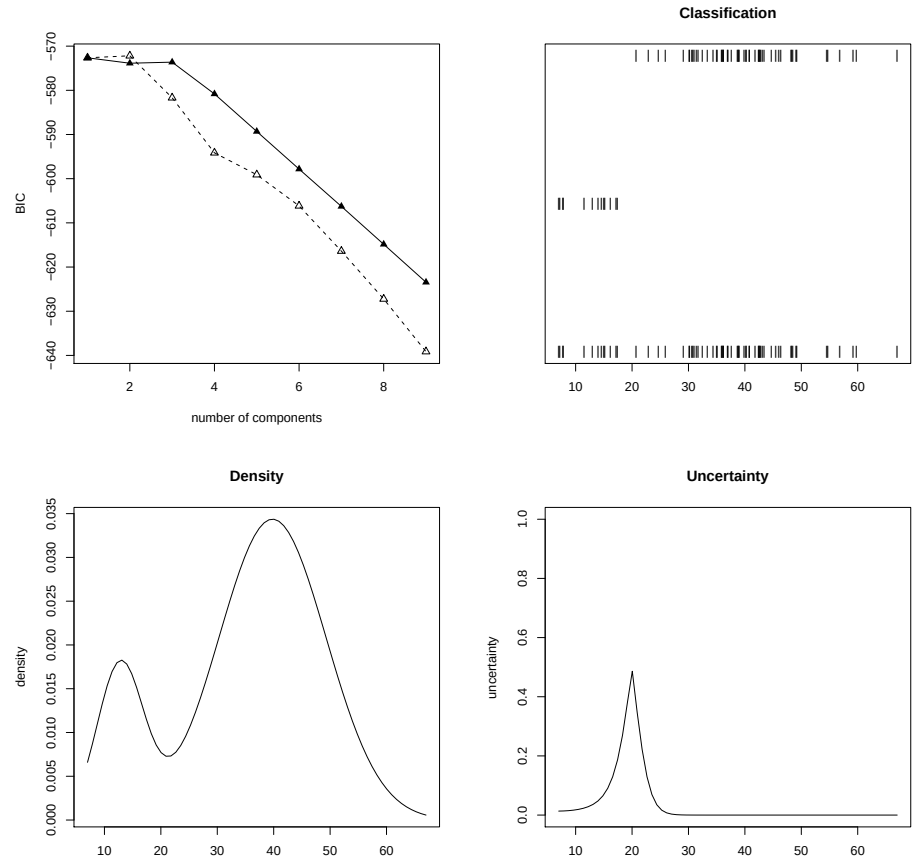


Figure 15: Model-based clustering of the one-dimensional R dataset precip. Clockwise from upper left: BIC, classification, uncertainty, and density from Mclust applied to the simulated one-dimensional example. In the classification plot, all of the data is displayed at the bottom, with the separated classes shown different levels above.

```
> mclust1Dplot(data = rivers, what = "classification",
               parameters=riversModel2$parameters, z=riversModel2$z)
> mclust1Dplot(data = rivers, what = "density",
               parameters=riversModel2$parameters, z=riversModel2$z)
> abline(v = riversModel2$parameters$mean, lty = 3)
```

Vertical lines are added at the means of each component. Figure 16 shows the classification and density corresponding to the two and three components cases for this example. A density estimates can also be computed and plotted directly:

```
> points <- seq(from = min(rivers), to = max(rivers), length = 1000)
> riversDens3 <- dens(modelName = riversModel$modelName, data = points,
                      parameters = riversModel$parameters)
> abline(v = riversModel$parameters$mean, lty = 3)
```

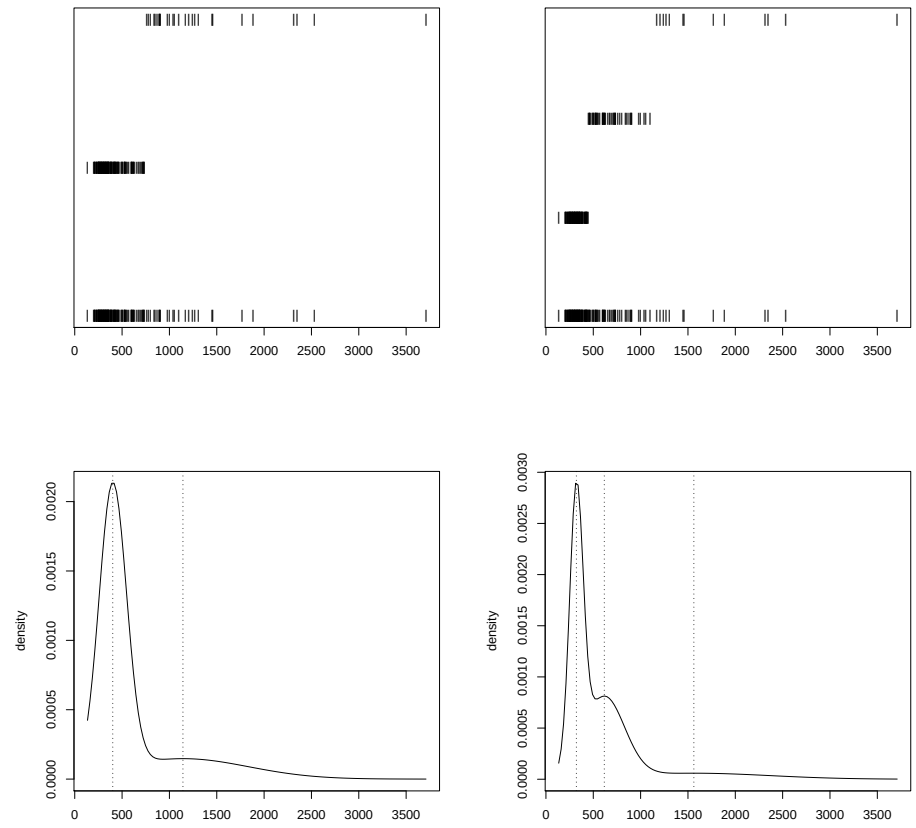


Figure 16: 2 and 3 component classifications and densities for the one-dimensional R dataset `rivers`. Vertical lines have been added to the density plots to show the location of the component means.

```
> plot(points, riversDens3, type = "l")
> riversDens2 <- dens(modelName = riversModel2$modelName, data = points,
                      parameters = riversModel2$parameters)
> plot(points, riversDens2, type = "l")
> abline(v = riversModel2$parameters$mean, lty = 3)
```

8.2 Discriminant Analysis

To illustrate discriminant analysis on one-dimensional data, we use simulated data from a normal mixture consisting of two components with variance 1 centered at -9 and 9, respectively, and one component with variance 4 centered at 0:

```
> set.seed(0)
> x <- c(rnorm(300, -9), rnorm(400, 0, sd = 2), rnorm(300, 9))
```

We use the following simulated data as a test set:

```
> set.seed(1)
> y <- c(rnorm(100, -9), rnorm(100, 0, sd = 2), rnorm(100, 9))
```

The density of the distribution from which the training data is drawn is shown in Figure 17.

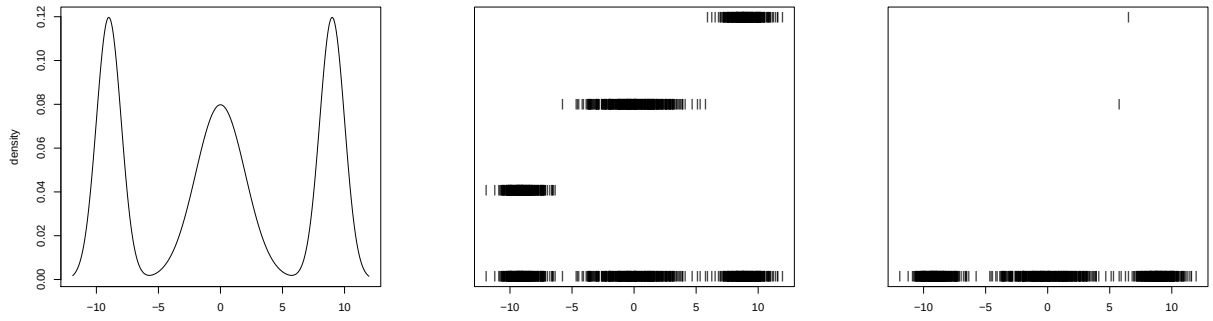


Figure 17: LEFT: Density for the one-dimensional simulation data. There are two components with variance 1 centered at -9 and 9, respectively, and one component with variance 4 centered at 0. CENTER: Training step classification. RIGHT: Misclassified training observations.

Discriminant Analysis via EM. In discriminant analysis via EM (Section 7.1), if we assume that each component consitutes as separate group,

```
> xClass <- c(rep(1,300),rep(2,400),rep(3,300))
> yClass <- c(rep(1,100),rep(2,100),rep(3,100))
```

then both leave-one-out crossvalidation and BIC choose the equal variance model `E` in the training stage:

```
> round(cv1EMtrain(x,labels=xClass),3)
    E     V
0.006 0.002
> round(bicEMtrain(x,labels=xClass),3)
        E         V
-5786.672 -5607.173
```

The varying variance model `V` is chosen by both cross-validation and BIC. The training and test errors for the data with this model are as follows:

```
> modelV <- mstep(modelName = "V", data = x, z = unmap(xClass))

> classV <- map(estep(modelName = "V", data = c(x,y),
                      parameters = modelV$parameters)$z)

> classError(classV[1:length(x)],xClass)$errorRate ## training error
```

```
[1] 0.002

> classError(classV[1:length(x)],xClass)$misclassified
[1] 391 990

> classError(classV[-(1:length(x))],yClass)$errorRate ## test error
[1] 0
```

The classification and classification errors for the training data are shown in Figure 17.

Discriminant Analysis via `mclustDA`. To illustrate the discriminant analysis via `MclustDA` (Section 7.2): we used the same simultated one-dimensional data but assume that observations are grouped by component variance:

```
> xClass <- c(rep(1,300),rep(2,400),rep(1,300))
> yClass <- c(rep(1,100),rep(2,100),rep(1,100))
```

The training stage fits a two component equal-variance model to one group, and a one-component model to the other:

```
> xTrain <- mclustDAtrain(x, labels = xClass)
 E X
 2 1
```

The classification error rates are the same as we obtained for discriminat analysis via EM with the 3-class grouping:

```
> xTest <- summary(mclustDAtest(x,xtrain))

> classError(xTest$classification,xClass)$errorRate ## training error
[1] 0.002
> classError(xTest$classification,xClass)$misclassified
[1] 391 990

> yTest <- summary(mclustDAtest(y,xTrain))

> classError(yTest$classification,yClass)$errorRate  ## testing error
[1] 0
```

The classification and classification errors for the training data are as shown in Figure 17.

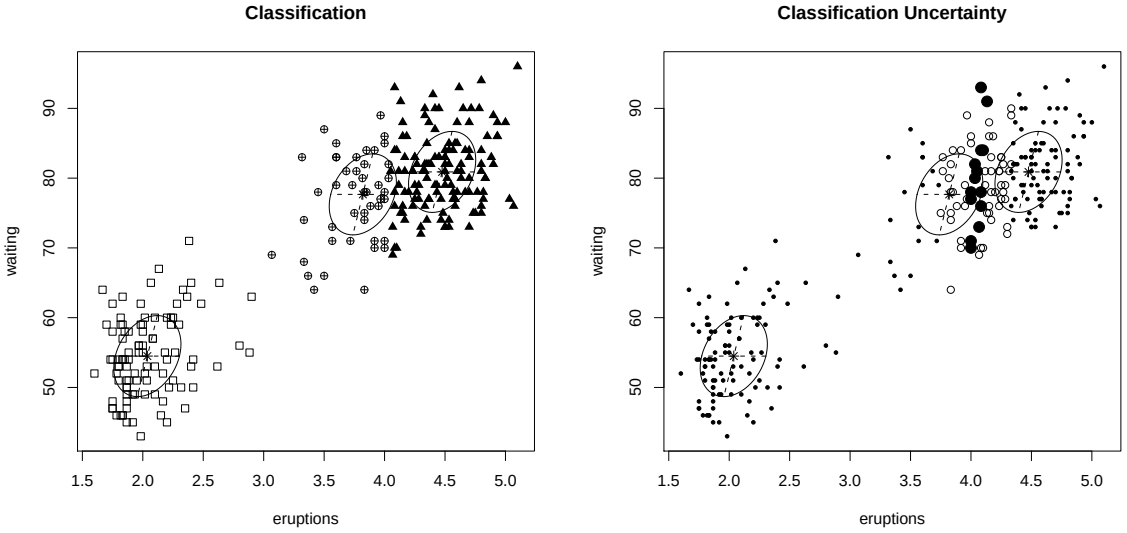


Figure 18: Classification (left) and uncertainty (right) plots created with `mclust2Dplot` for the `Mclust` model of the `faithful` dataset. The ellipses shown are the multivariate analogs of the standard deviations for each mixture component. In the classification plot, points in different classes are indicated by different symbols. In the uncertainty plot, the symbols have the following meaning: large filled symbols, 95% quantile of uncertainty; smaller open symbols, 75–95% quantile; small dots, first three quartiles of uncertainty.

9 Displays for Multidimensional Data

Once parameter values of a mixture model fit are available, projections of the data showing the means and standard deviations of the corresponding components or clusters may be plotted. In the two-dimensional case, density and uncertainty surfaces may also be plotted.

9.1 Displays for Two-Dimensional Data

The function `mclust2Dplot` may be used for displaying the classification, uncertainty or classification errors for `MCLUST` models of two-dimensional data. In the following example, classification and uncertainty plots are produced for the `faithful` dataset in Figure 1.

```
> faithfulMclust <- Mclust(faithful)

> mclust2Dplot(data = faithful, what = "classification", identify = TRUE,
     parameters = faithfulMclust$parameters, z = faithfulMclust$z)

> mclust2Dplot(data = faithful, what = "uncertainty", identify = TRUE,
    parameters = faithfulMclust$parameters, z = faithfulMclust$z)
```

The resulting plots are displayed in Figure 18.

The function `surfacePlot` may be used for displaying the density or uncertainty for `MCLUST` models of two-dimensional data. It also returns the grid coordinates and corre-

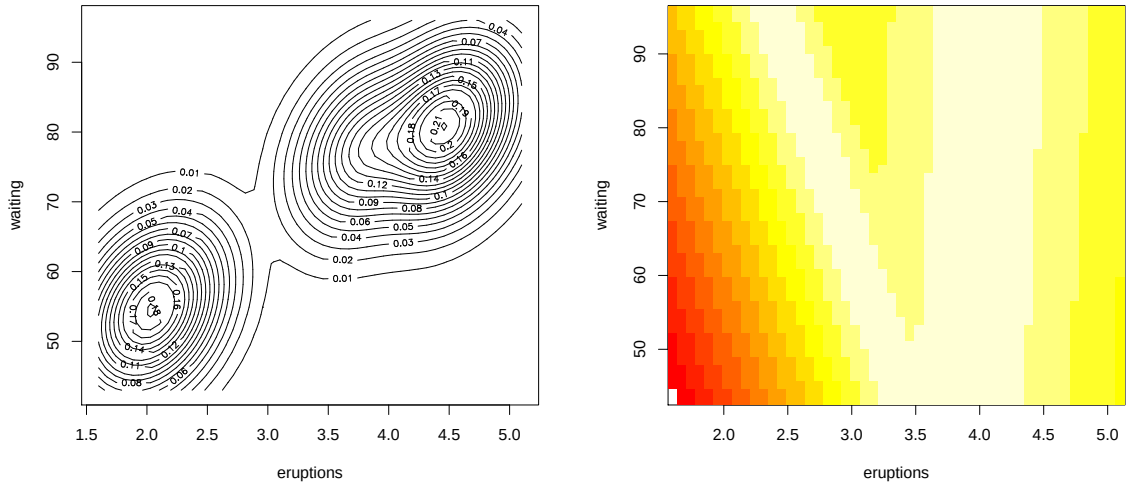


Figure 19: Density (left column) and uncertainty (right column) surfaces for the `faithful` dataset. A square root transformation was used for the density plot, which is plotted as a contour surface. A logarithmic transformation was used for the uncertainty plot, which is plotted as an image surface.

sponding surface values. The following example shows how to display density and uncertainty surfaces for the `Mclust` model fit to the `faithful` dataset.

```
> surfacePlot(data = faithful, what = "density", type = "contour",
     parameters = faithfulMclust$parameters, transformation = "sqrt")

> surfacePlot(data = faithful, what = "uncertainty", type = "image",
    parameters = faithfulMclust$parameters,  transformation = "log")
```

The resulting plots are displayed in Figure 19.

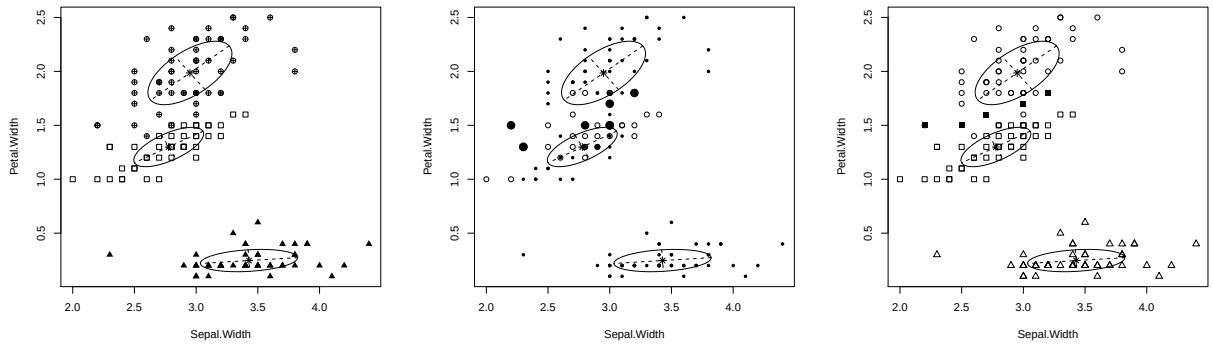


Figure 20: A coordinate projection of the `iris` dataset created with `coordProj`. Plots show the 3-group model-based classification (left) with associated uncertainty (middle) and classification errors (right).

9.2 Displays for Higher Dimensional Data

9.2.1 Coordinate Projections

To plot coordinate projections in `MCLUST`, use the function `coordProj`. The example we consider is a 3-group model for the `iris` dataset:

```
> irisBIC <- mclustBIC(iris[,-5])
> irisSummary3 <- summary(irisBIC, data = iris[,-5], G = 3)

> coordProj( data = iris[,-5], dimens = c(2,4), what = "classification",
   parameters = irisSummary3$parameters, z = irisSummary3$z)

> coordProj( data = iris[,-5], dimens = c(2,4), what = "uncertainty",
   parameters = irisSummary3$parameters, z = irisSummary3$z)

> coordProj( data = iris[,-5], dimens = c(2,4), what = "errors",
   parameters = irisSummary3$parameters, z = irisSummary3$z, truth = iris[,5])
```

These plots are displayed in Figure 20.

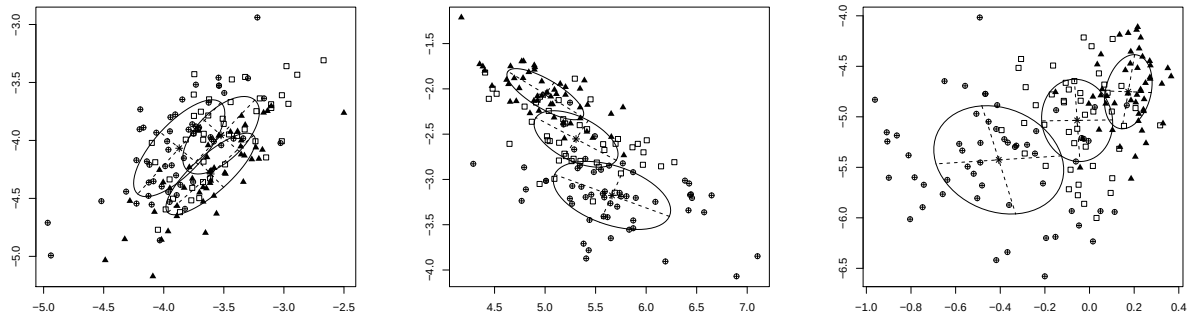


Figure 21: Some random projections of the `iris` dataset created with `coordProj`. Plots show the 3-group classification from model-based clustering with three different seeds.

9.2.2 Random Projections

To plot random projections in `MCLUST`, use the function `randProj`. Again we consider is a 3-group model for the `iris` dataset:

```
> randProj( data = iris[,-5], seed = 43, what = "classification",
   parameters = irisSummary3$parameters, z = irisSummary3$z)

> randProj( data = iris[,-5], seed = 79, what = "classification",
   parameters = irisSummary3$parameters, z = irisSummary3$z)

> randProj( data = iris[,-5], seed = 201, what = "classification",
   parameters = irisSummary3$parameters, z = irisSummary3$z)
```

These plots are displayed in Figure 21.

10 Simulation from Mixture Densities

Given the parameters for a mixture model, data can be simulated from that model for evaluation and verification. The function `sim` allows simulation from mixture models generated by `MCLUST` functions. Besides the model, `sim` allows a seed as input for reproducibility. As an example, below we simulate two different datasets of the same size as the `faithful` dataset from the model produced by `Mclust` for the `faithful` dataset:

```
> faithfulMclust <- Mclust(faithful)

> sim0 <- sim( modelName = faithfulMclust$modelName,
               parameters = faithfulMclust$parameters,
               n = nrow(faithful), seed = 0)
> sim1 <- sim( modelName = faithfulMclust$modelName,
               parameters = faithfulMclust$parameters,
               n = nrow(faithful), seed = 1)
```

The results can be plotted as follows:

```
> mclust2Dplot(data=faithful, parameters = faithfulMclust$parameters,
                              classification = faithfulMclust$classification)

> mclust2Dplot(data=sim0[,-1], parameters = faithfulMclust$parameters,
                               classification = sim0[,1])

> mclust2Dplot(data=sim1[,-1], parameters = faithfulMclust$parameters,
                               classification = sim1[,1])
```

The plots are shown in Figure 22. Note that `sim` produces a dataset in which the first column is the classification.

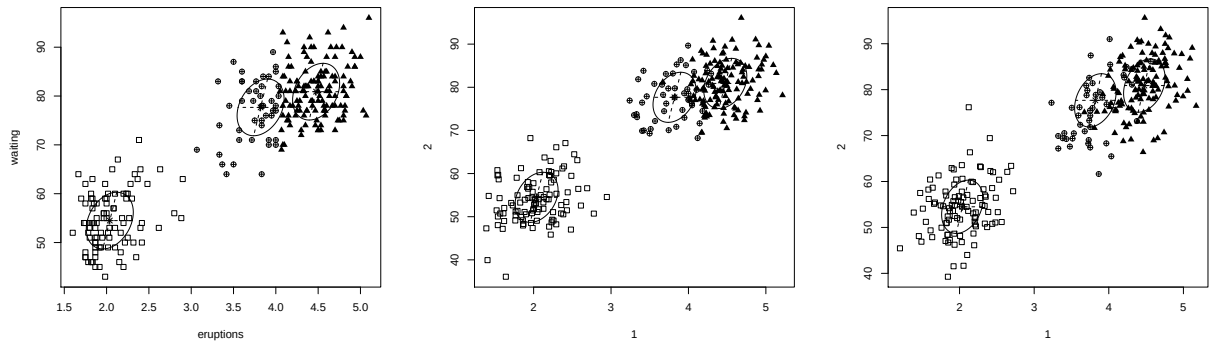


Figure 22: Data simulated from a model of the `faithful` dataset. The left hand figure is the `faithful` dataset, and the other two figures are datasets of the same size simulated from the `Mclust` model for the `faithful` dataset. The ellipse defined by the covariance matrices of the model is shown on all of the plots.

11 Extensions

11.1 Large Datasets

`Mclust` and `mclustBIC` include a provision for using a subsample of the data in the hierarchical clustering phase before applying EM to the full data set, in order to extend the method to larger datasets. Some other methods for handling such cases are discussed in [24, 14]. The following example uses a random sample of size 100 in the initial hierarchical clustering phase of `EMclust` applied to the iris data:

```
> nrow(iris)
[1] 150
> S <- sample(1:nrow(iris), size = 100)
> Mclust(iris[,-5], initialization = list(subset = S))
```

For very large data sets, the discrimination capabilities of `MCLUST` can be used for classification. First, cluster analysis with the methodolgy of `Mclust`/`mclustBIC` can be performed on a subset of the data. Then the remaining data points can then be classified (in reasonable sized blocks) using one of the discriminant analysis techniques described in section 7.

11.2 High Dimensional Data

Models in which the orientation is allowed to vary between clusters (`EEV`, `VEV`, `EVV`, `VVV`), have $\mathcal{O}(d^2)$ parameters per cluster, where d is the dimension of the data. For this reason, `MCLUST` may not work well or may otherwise be inefficient for these models when applied to high-dimensional data. It may still be possible to analyze such data with `MCLUST` by restriction to models with fewer parameters (e.g. spherical or diagonal models), or else by applying a dimension-reduction technique such as principal components.

Some of the more parsimonious models (e.g. spherical, diagonal, or fixed covariance) can be applied to datasets in which the number of observations is smaller than the data dimension.

12 Function Summary

12.1 Hierarchical Clustering

`hc`	Merge sequences for model-based hierarchical clustering.
`hclass`	Classifications corresponding to `hc` results.

12.2 Parameterized Gaussian Mixture Models

`em`	EM algorithm (starting with E-step).
`me`	EM algorithm (starting with M-step).
`estep`	E-step of the EM algorithm.
`mstep`	M-step of the EM algorithm.
`mvn`	One-component fit.

12.3 Density Computation for Parameterized Gaussian Mixtures

`cdens`	Component density (without mixing proportions).
`dens`	Mixture density.

12.4 Model-based Clustering / Density Estimation

`mclustBIC`	BIC computation; clusters and models through `summary`.
`Mclust`	Combines `mclustBIC` and its `summary` (fewer options).

12.5 Discriminant Analysis

Class Densities as Mixture Components

`cv1EMtrain`	Training via leave-one-out crossvalidation.
`bicEMtrain`	Training via BIC.
`estep`	E-step of the EM algorithm.
`mstep`	M-step of the EM algorithm.

Parameterized Gaussian Mixture for Class Densities (MclustDA)

`mclustDAtrain`	MclustDA training.
`mclustDAtest`	MclustDA density; classification via `summary`.
`mclustDA`	Combines `mclustDAtrain` and `mclustDAtest` (fewer options).

12.6 Support for Modeling and Classification

`.Mclust`	vector of default values.
`mclustOptions`	set `MCLUST` options.
`map`	Convert conditional probabilities to a classification.
`unmap`	Convert a classification to indicator variables.
`bic`	BIC for parameterized Gaussian mixture models.
`sim`	Simulate data from a parameterized Gaussian mixture model.
`mapClass`	Mapping between two classifications.
`classError`	Classification error.
`adjustedRandIndex`	Adjusted Rand Index.
`sigma2decomp`	Convert mixture covariances to decomposition form.
`decomp2sigma`	Convert decomposition form to mixture covariances.
`nVarParams`	Number of variance parameters.

12.7 Plotting Functions

12.7.1 One-Dimensional Data

`mclust1Dplot` Classification, uncertainty, density and/or classification errors.

12.7.2 Two-Dimensional Data

`mclust2Dplot` Classification, uncertainty, and/or classification errors.

`surfacePlot` Contour, image, or perspective plot of either density or uncertainty.

12.7.3 More than Two Dimensions

Classification, uncertainty, and/or classification errors.

`coordProj`	coordinate projections
`randProj`	random projections

12.7.4 Other Plotting Functions

`clPairs` pairs plot showing classification

`uncerPlot` relative uncertainty of misclassified observations

`plot.Mclust` plots associated with `Mclust` results

`plot.mclustBIC` BIC plot associated with `mclustBIC` results

`plot.mclustDA` plots associated with `mclustDA` results

`plot.mclustDAtrain` plots associated with `mclustDAtrain` results

A Appendix: Clustering Models

MCLUST usually assumes a normal or Gaussian mixture model

$$\prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \ \phi_k(\mathbf{x}_i \mid \mu_k, \Sigma_k),$$

where where $\mathbf{x}$ represents the data, G is the number of components, and τ_k is the probabilty that an observation belongs to the kth component ($\tau_k \geq 0; \quad \sum_{k=1}^{G} \tau_k = 1$) , and

$$\phi_k(\mathbf{x} \mid \mu_k, \Sigma_k) = (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) \right\}. \tag{1}$$

The exception is for model-based hierarchical clustering, for which the model used is the classification likelihood with a parameterized normal distribution assumed for each class:

$$\prod_{i=1}^{n} \phi_{\ell_i}(\mathbf{x}_i \mid \mu_{\ell_i}, \Sigma_{\ell_i}),$$

where the ℓ_i are labels indicating a unique classification of each observation: $\ell_i = k$ if $\mathbf{x}_i$ belongs to the kth component.

The components or clusters in both these models are ellipsoidal, centered at the means μ_k. The covariances Σ_k determine their other geometric features. Each covariance matrix is parameterized by eigenvalue decomposition in the form

$$\Sigma_k = \lambda_k D_k A_k D_k^T,$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k, and λ_k is a scalar. The orientation of the principal components of Σ_k is determined by D_k, while A_k determines the shape of the density contours; λ_k specifies the volume of the corresponding ellipsoid, which is proportional to $\lambda_k^d \, |A_k|$, where d is the data dimension. Characteristics (orientation, volume and shape) of distributions are usually estimated from the data, and can be allowed to vary between clusters, or constrained to be the same for all clusters [19, 2, 6]. This parameterization includes but is not restricted to well-known models such as equal-volume spherical variance ($\Sigma_k = \lambda I$) which gives the sum of squares criterion [23], constant variance [15], and unconstrained variance [22].

In one dimension, there are just two models: `E` for equal variance and `V` for varying variance. In more than one dimension, the model identifiers code geometric characteristics of the model. For example, `EVI` denotes a model in which the volumes of all clusters are equal (`E`), the shapes of the clusters may vary (`V`), and the orientation is the identity (`I`). Clusters in this model have diagonal covariances with orientation parallel to the coordinate axes. Parameters associated with characteristics designated by `E` or `V` are determined from the data. Table 1 shows the various multivariate model options currently available in MCLUST for hierarchical clustering (denoted HC) and EM. These are a subset of the parameterizations discussed in [6], which gives details of the EM algorithm for maximum likelihood estimation for these models.

A.1 Modeling Noise and Outliers

`MCLUST` uses a mixture model which has a single term representing noise as a first order Poisson process to handle noisy data:

$$\prod_{i=1}^{n} \left[\frac{\tau_0}{V} + \sum_{k=1}^{G} \tau_k \phi_k(\mathbf{x}_i \mid \theta_k) \right], \tag{2}$$

in which V is the hypervolume of the data region, and $\tau_k \geq 0$; $\sum_{k=0}^{G} \tau_k = 1$. This model has been used successfully in a number of applications [2, 7, 4, 5].

The basic model-based clustering method needs to be modified when the data contains noise. First, a good initial noise estimate must be obtained. Some possible methods for denoising include a Voronoï method [1] and a nearest-neighbor method [3]. The function `NNclean` in the contributed `R` package `prabclus` is an implementation of the nearest-neighbor method. Next, hierarchical clustering is applied to the denoised data. Finally, EM based on the Gaussian model with the added noise term (2) is applied to the entire data set, with the data removed in the denoising process as the initial noise estimate.

A.2 Model Selection via BIC

Several measures have been proposed for choosing the clustering model (parameterization and number of clusters); see, e.g., Chapter 6 of McLachlan and Peel (2000). We use the Bayesian Information Criterion (BIC) approximation to the Bayes factor (Schwarz 1978), which adds a penalty to the loglikelihood based on the number of parameters, and has performed well in a number of applications (e.g. Fraley and Raftery 1998, 2002). The BIC has the form

$$\text{BIC} \equiv 2 \ \text{loglik}_{\mathcal{M}}(\mathbf{x}, \theta_k^*) - (\#\ \text{params})_{\mathcal{M}} \ \log(n), \tag{3}$$

where $\text{loglik}_{\mathcal{M}}(\mathbf{x}, \theta_k^*)$ is the maximized loglikelihood for the model and data, $(\#\ \text{params})_{\mathcal{M}}$ is the number of independent parameters to be estimated in the model $\mathcal{M}$, and n is the number of observations in the data.

The symbols used in the BIC plots to represent the various models in `MCLUST` are shown in Table 2.

A.3 Adding a Prior to the Model

By default, `MCLUST` does not use prior for modeling. However, users can optionally specify a conjugate prior of the type described inthis section. For univariate data, we use a normal prior on the mean (conditional on the variance):

$$\begin{aligned} \mu \mid \sigma^2 &\sim \mathcal{N}(\mu_{\mathcal{P}}, \sigma^2/\kappa_{\mathcal{P}}) \\ &\propto \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{ -\frac{\kappa_{\mathcal{P}}}{2\sigma^2} \left(\mu - \mu_{\mathcal{P}}\right)^2 \right\} \end{aligned} \tag{4}$$

and an inverse gamma prior on the variance:

$$\begin{aligned} \sigma^2 &\sim \text{inverseGamma}(\nu_{\mathcal{P}}/2, \varsigma^2_{\mathcal{P}}/2) \\ &\propto \left(\sigma^2\right)^{-\frac{\nu_{\mathcal{P}}+2}{2}} \exp\left\{-\frac{\varsigma^2_{\mathcal{P}}}{2\sigma^2}\right\}. \end{aligned} \tag{5}$$

For multivariate data, we use a normal prior on the mean (conditional on the covariance matrix):

$$\begin{aligned} \mu \mid \Sigma &\sim \mathcal{N}(\mu_{\mathcal{P}}, \Sigma/\kappa_{\mathcal{P}}) \\ &\propto |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{\kappa_{\mathcal{P}}}{2}\text{trace}\left[(\mu-\mu_{\mathcal{P}})^T \Sigma^{-1} (\mu-\mu_{\mathcal{P}})\right]\right\}, \end{aligned} \tag{6}$$

and an inverse Wishart prior on the covariance matrix:

$$\begin{aligned} \Sigma &\sim \text{inverseWishart}(\nu_{\mathcal{P}}, \Lambda_{\mathcal{P}}) \\ &\propto |\Sigma|^{-\frac{\nu_{\mathcal{P}}+d+1}{2}} \exp\left\{-\frac{1}{2}\text{trace}\left[\Sigma^{-1}\Lambda_{\mathcal{P}}^{-1}\right]\right\}. \end{aligned} \tag{7}$$

The hyperparameters $\mu_{\mathcal{P}}$, $\kappa_{\mathcal{P}}$, and $\nu_{\mathcal{P}}$ are called the *mean*, *shrinkage* and *degrees of freedom*, respectively. Parameters $\varsigma^2_{\mathcal{P}}$ (a scalar) and $\Lambda_{\mathcal{P}}$ (a matrix) are the *scale* of the prior distribution in the univariate and multivariate cases, respectively. These priors are called *conjugate priors* for the normal distribution because the posterior can be expressed as the product of a normal distribution and an inverse gamma or Wishart distribution.

Functions `priorControl` and `defaultPrior` are provided in `MCLUST` for specifying a prior. When called with defaults, the following choices are made for the prior hyperparameters:

$\mu_{\mathcal{P}}$: the mean of the data.

$\kappa_{\mathcal{P}}$: .01
The posterior mean $\frac{n_k\bar{y}_k + \kappa_{\mathcal{P}}\mu_{\mathcal{P}}}{\kappa_{\mathcal{P}} + n_k}$ can be viewed as adding $\kappa_{\mathcal{P}}$ observations with value $\mu_{\mathcal{P}}$ to each group in the data. The value we used was determined by experimentation; values close to and bigger than 1 caused large perturbations in the modeling in cases where there were no missing BIC values without the prior. The value .01 resulted in BIC curves that appeared to be smooth extensions of their counterparts without the prior.

$\nu_{\mathcal{P}}$: $d+2$
Analogously to the univariate case, the marginal prior distribution of μ is a t distribution centered at $\mu_{\mathcal{P}}$ with $\nu_{\mathcal{P}} - d + 1$ degrees of freedom. The mean of this distribution is $\mu_{\mathcal{P}}$ provided that $\nu_{\mathcal{P}} > d$, and it has a finite covariance matrix provided $\nu_{\mathcal{P}} > d+1$ (see, e. g. Schafer 1997). We chose the smallest integer value for the degrees of freedom that gives a finite covariance matrix.

$\varsigma^2_{\mathcal{P}}$: $\frac{\text{sum}(\text{diag}(\text{var}(\text{data})))/d}{G^{2/d}}$ (For univariate models, and multivariate spherical or diagonal models.) The average of the diagonal elements of the empirical covariance matrix of the data divided by the square of the number of components to the $1/d$ power. This

is roughly equivalent to partitioning the range of the data into G intervals of fairly equal size.

$\Lambda_{\mathcal{P}}$: $\frac{\text{var(data)}}{G^{2/d}}$ (For multivariate ellipsoidal models.) The empirical covariance matrix of the data divided by the square of the number of components to the $1/d$ power.

References

[1] D. Allard and C. Fraley. Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoï tessellation. *Journal of the American Statistical Association*, 92:1485–1493, 1997.

[2] J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[3] S. D. Byers and A. E. Raftery. Nearest neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 93:577–584, 1998.

[4] J. G. Campbell, C. Fraley, F. Murtagh, and A. E. Raftery. Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, 18:1539–1548, 1997.

[5] J. G. Campbell, C. Fraley, D. Stanford, F. Murtagh, and A. E. Raftery. Model-based methods for real-time textile fault detection. *International Journal of Imaging Systems and Technology*, 10:339–346, 1999.

[6] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

[7] A. Dasgupta and A. E. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93:294–302, 1998.

[8] C. Fraley. Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20:270–281, 1998.

[9] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal*, 41:578–588, 1998.

[10] C. Fraley and A. E. Raftery. MCLUST: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306, 1999.

[11] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[12] C. Fraley and A. E. Raftery. Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification*, 20:263–286, 2003.

[13] C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. Technical Report 486, University of Washington, Department of Statistics, August 2005.

[14] C. Fraley, A. E. Raftery, and R. Wehrens. Incremental model-based clustering for large datasets with small clusters. *Journal of Computational and Graphical Statistics*, 14:1–18, 2005.

[15] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62:1159–1178, 1967.

[16] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[17] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[18] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[19] F. Murtagh and A. E. Raftery. Fitting straight lines to point patterns. *Pattern Recognition*, 17:479–483, 1984.

[20] J. L. Schafer. *Analysis of Incomplete Multivariate Data by Simulation*. Chapman and Hall, 1997.

[21] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[22] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[23] J. H. Ward. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58:234–244, 1963.

[24] R. Wehrens, L. Buydens, C. Fraley, and A. Raftery. Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification*, 21:231–253, 2004.