Gaussian Mixtures

The `galaxies` data in the MASS package (Venables and Ripley, 2002) is a frequently used example for Gaussian mixture models. It contains the velocities of 82 galaxies from a redshift survey in the Corona Borealis region. Clustering of galaxy velocities reveals information about the large scale structure of the universe.

```
library(MASS)
data(galaxies)
X = galaxies / 1000
```

The `Mclust` function from the `mclust` package (Fraley et al, 2012) is used to fit Gaussian mixture models. The code below fits a model with `G=4` components to the `galaxies` data, allowing the variances to be unequal (`model="V"`).

```
library(mclust, quietly=TRUE)
```

```
## Warning: package 'mclust' was built under R version 3.2.4
```

```
## Package 'mclust' version 5.2
```

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

```
fit = Mclust(X, G=4, model="V")
summary(fit)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 4 components:
##
##  log.likelihood  n df       BIC      ICL
##       -199.2545 82 11 -446.9829 -466.264
##
## Clustering table:
##  1  2  3  4
##  7 35 32  8
```

Figure 1 shows the resulting density plot.

```
plot(fit, what="density", main="", xlab="Velocity (Mm/s)")
rug(X)
```

Section 6.2 of Drton and Plummer (2017) considers singular BIC for Gaussian mixture models using the `galaxies` data set as an example. Singularities occur when two mixture components coincide (i.e. they have the same mean and variance) or on the boundary of the parameter space where the prior probability of a mixture component is zero.

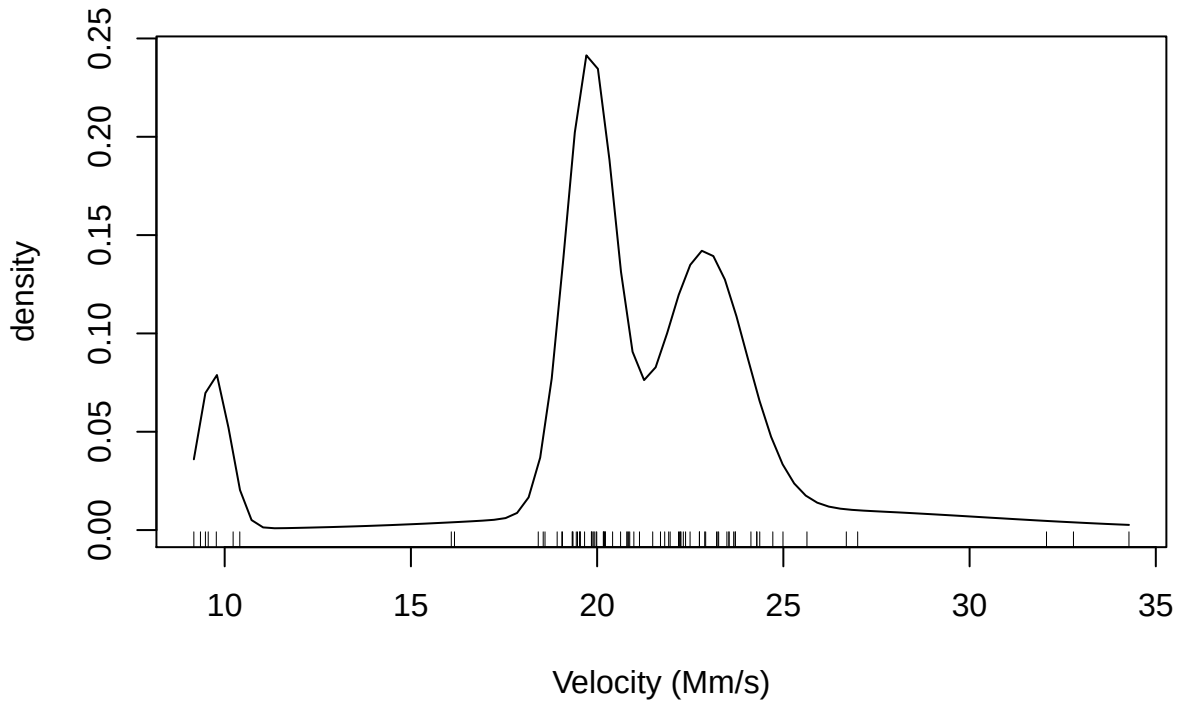

Figure 1: Density estimate for `galaxies` data from a 4-component mixture model

The `GaussianMixtures()` function creates an object representing a family of mixture models up to a specified maximum number of components (`maxNumComponents=10` in this example). The `phi` parameter controls the penalty to be used for `sBIC` (See below) and the `restarts` parameter determines the number of times each model is fitted starting from randomly chosen starting points. Due to the multi-modal likelihood surface for mixture models, multiple restarts are used to find a good (local) maximum.

```
library(sBIC)
gMix = GaussianMixtures(maxNumComponents=10, phi=1, restarts=100)
```

Learning coefficients are known exactly for Gaussian mixtures with known and equal variances, but this model is rarely applied in practice. For unequal variances, the learning coefficients are unknown, but upper bounds are given by Drton and Plummer (2017, equation 6.11). These bounds are implemented by setting the penalty parameter `phi=1` in the `GaussianMixtures()` function. We refer to the singular BIC using these approximate penalties as $\overline{\text{sBIC}}_1$. It is calculated by supplying the data `X` and the model set `gMix` to the `sBIC()` function. The RNG seed is set for reproducibility, due to the random restarts.

```
set.seed(1234)
m = sBIC(X, gMix)
print(m)
```

```
## $logLike
##  [1] -240.3379 -220.2445 -203.1792 -197.4621 -190.0724 -186.8674 -185.8390
##  [8] -186.7764 -184.0937 -185.8294
##
## $sBIC
##  [1] -244.7446 -231.2612 -220.8038 -219.0979 -216.4564 -216.2255 -217.5358
##  [8] -220.6820 -220.2114 -224.1505
##
```

```
## $BIC
##  [1] -244.7446 -231.2613 -220.8061 -221.6990 -220.9195 -224.3245 -229.9061
##  [8] -237.4537 -241.3811 -249.7268
##
## $modelPoset
## [1] "GaussianMixtures: 0x7fde15f23e68"
```

Figure 2 compares BIC with $\overline{\text{sBIC}}_1$. Both criteria have been standardized so that the value for the 1-component model is 0. This figures reproduces Figure 7 of Drton and Plummer (2017). The reproduction is not exact because, in the interests of speed, we have reduced the number of restarts from 5000 to 100. This mainly affects the models with larger number of components.

```
matplot(
  cbind(m$BIC - m$BIC[1], m$sBIC - m$sBIC[1]),
  pch = c(1, 3),
  col = "black",
  xlab = "Number of components",
  ylab = expression(BIC - BIC(M[1])),
  las=1, xaxt="n"
)
axis(1, at = 1:10)
legend("topleft",
       c(expression(BIC), expression(bar(sBIC)[1])),
       pch = c(1, 3),
       y.intersp = 1.2)
```

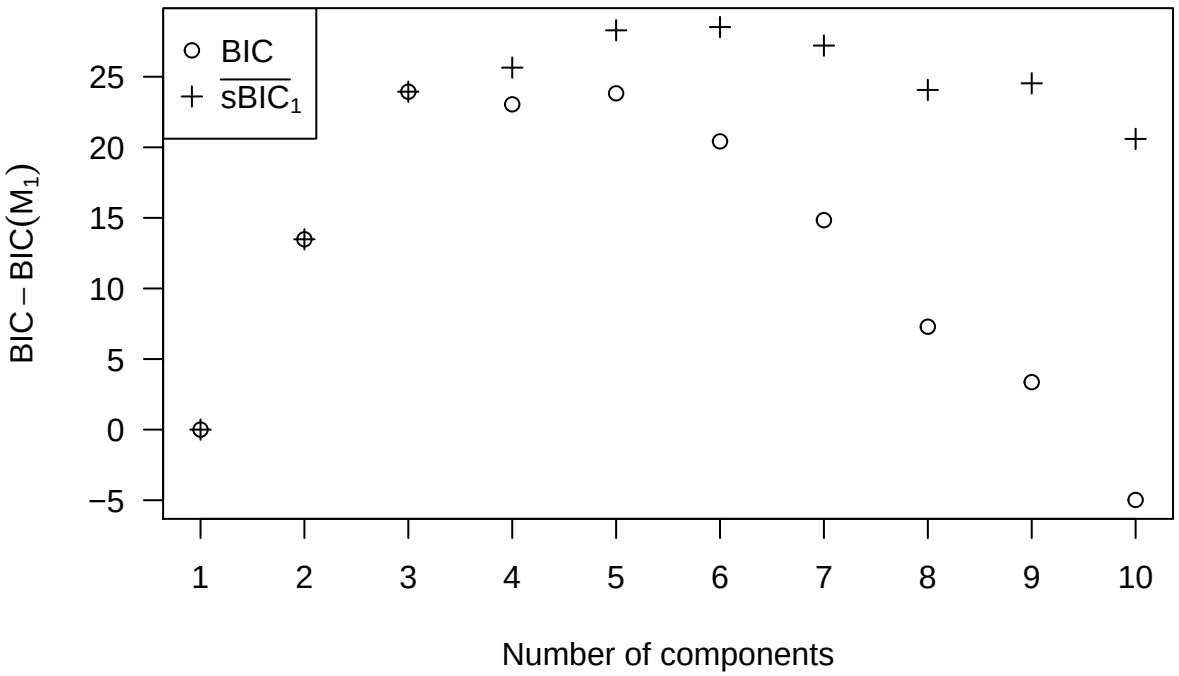

Figure 2: Comparison of singular BIC with BIC for choosing the number of components in the galaxies data

The BIC and singular BIC results for the `galaxies` data can be compared with the posterior probabilities for the number of components derived by Richardson and Green (1997, Table 1) using reversible jump MCMC. Since Richardson and Green (1997) consider up to 14 components, we truncate the distribution up to 10 components and renormalize.

```
post.MCMC = c(0.000, 0.000, 0.061, 0.128, 0.182, 0.199, 0.160,
              0.109, 0.071, 0.040, 0.023, 0.013, 0.006, 0.003)[1:10]
post.MCMC = post.MCMC / sum(post.MCMC)
```

The posterior probabilities from BIC and $\overline{\text{sBIC}}_1$ are derived by exponentiating and then renormalizing using the helper function `postBIC()`.

```
postBIC <- function(BIC) {
  prob <- exp(BIC - max(BIC))
  prob/sum(prob)
}
normalizedProbs = rbind("BIC"=postBIC(m$BIC), "sBIC1"=postBIC(m$sBIC), "MCMC"=post.MCMC)
```

Figure 3 compares the posterior densities from the three approaches. This reproduces figure 8 from Drton and Plummer (2017).

```
barplot(
  normalizedProbs,
  beside = TRUE,
  col = c("white","grey","black"),
  legend = c(expression(BIC), expression(bar(sBIC)[1]), expression(MCMC)),
  xlab = "Number of components",
  ylab = "Probability",
  args.legend = list(y.intersp = 1.2),
  names.arg = 1:10
)
```

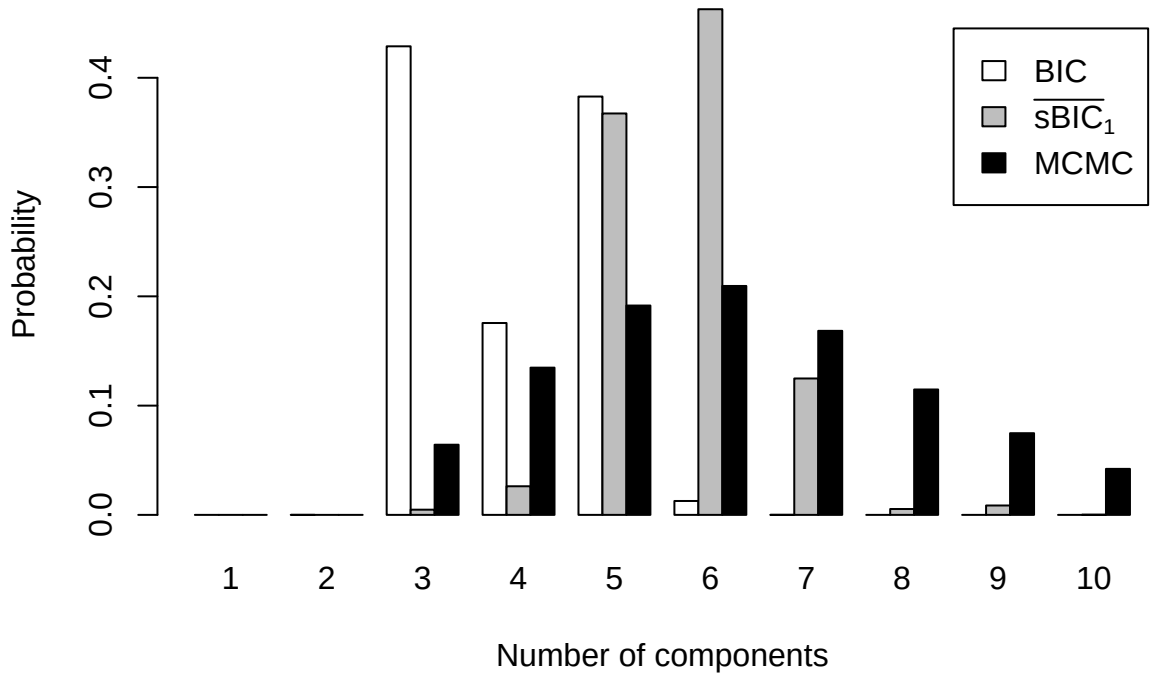

Figure 3: Posterior distribution of the number of components in a Gaussian mixture model with unequal variances applied to the galaxies data

Bibliography

- Drton M. and Plummer M. (2017), A Bayesian information criterion for singular models. J. R. Statist. Soc. B; 79: 1-38.
- Fraley C., Raftery A.E., Murphy T.B., and Scrucca L. (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597, Department of Statistics, University of Washington
- Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Statist. Soc. B; 59: 731-792.
- Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0