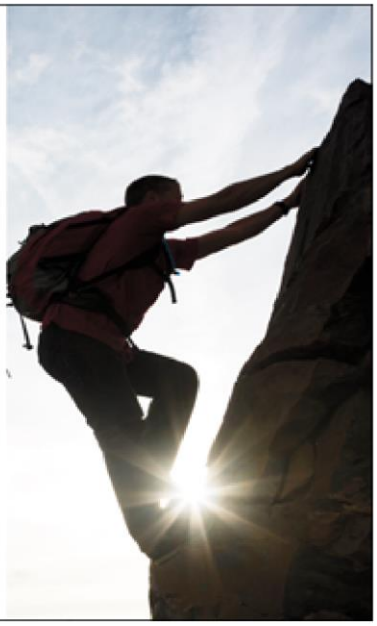
Distribution Fitting (Bivariate Mixture Distributions)

statgraphics® centurion



Revised: 12/4/2019

Summary

The **Distribution Fitting (Bivariate Mixture Distributions)** procedure fits a distribution to continuous numeric data that consists of a mixture of 2 or more bivariate Gaussian distributions. The components of the mixture may represent different groups in the sample used to fit the overall distribution, or the mixture model may approximate some distribution with a complicated shape.

The procedure fits the distribution and creates graphs of the fitted model. Tools are also provided for determining how many components are needed to represent a data sample.

The calculations are performed by the "EMCluster" package in R. To run the procedure, R must be installed on your computer together with those packages. For information on downloading and installing R, refer to the document titled "R – Installation and Configuration".

Sample StatFolio: *bivariate mixture.sgp*

Sample Data

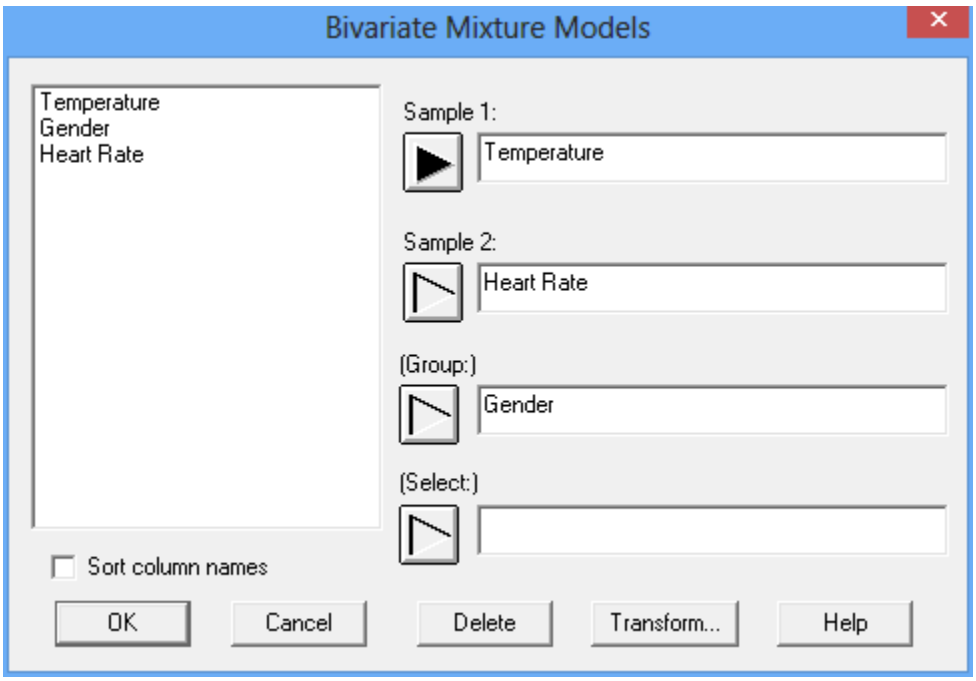
The file *bodytemp.sgd* contains measurements of the body temperature and heart rate of 130 individuals. The first several rows of that file are shown below:

Temperature	*Gender*	*Heart rate*
98.4	Male	84
98.4	Male	82
98.2	Female	65
97.8	Female	71
98	Male	78
97.9	Male	72
99	Female	79
98.5	Male	68
98.8	Female	64
98	Male	67
…	…	…

Half of the subjects were male and half were female.

Data Input

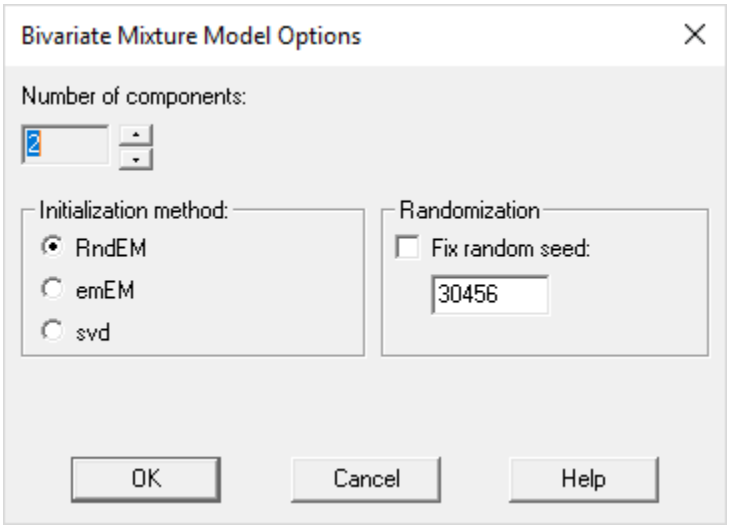
When the procedure is first selected, a data input dialog box is displayed requesting the names of the columns containing the data:



- **Sample 1:** name of first data column to be used to fit the distribution.
- **Sample 2:** name of second data column to be used to fit the distribution.
- **(Group:)** optional numeric or character column identifying group membership for each observation. This entry has no effect on the fitted model. It is only used to summarize membership percentages in each component of the model.
- **(Select:)** optional subset selection.

Analysis Options

After the data have been specified, the *Analysis Options* dialog box is displayed:



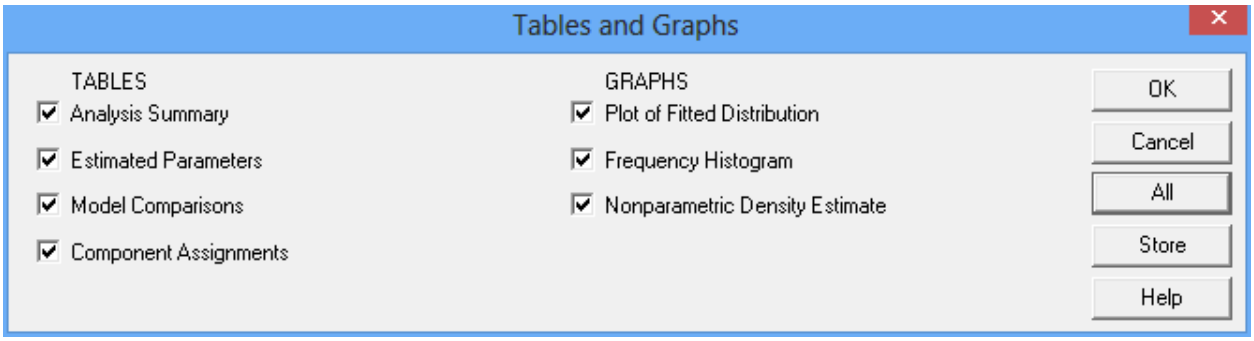
- **Number of components:** *K*, the number of separate bivariate Gaussian distributions contained in the model.

- **Initialization method:** method used to initialize the model-fitting algorithm. *RndEm* (Maitra 2009) randomly selects K centers and groups all other data to closest center. It then repeats the process and selects the best initialization based on log likelihood. *emEM* (Biernacki et al. 2003) consists of both short-EM and long-EM steps. *svd* (Maitra 2001) selects centers from major component space and singular value decomposition of data. For more details, see Wei-Chen Chen and Ranjan Maitra (2015).

- **Randomization:** whether to fix the seed of the random number generator using the value indicated. If the seed is fixed, the same results will be obtained each time the procedure is run (assuming no other options are changed).

Tables and Graphs

The following tables and graphs may be created:



Statistical Model

The statistical model fit by this procedure is a mixture of *K* bivariate Gaussian distributions. Each component distribution is parameterized by a vector of means

$$\mu_j = \begin{pmatrix} \mu_{j,1} \\ \mu_{j,2} \end{pmatrix} \quad (1)$$

a vector of standard deviations

$$\sigma_j = \begin{pmatrix} \sigma_{j,1} \\ \sigma_{j,2} \end{pmatrix} \quad (2)$$

and a correlation coefficient ρ_j. The density function is the weighted sum of *K* such component distributions and includes *K* mixing parameters $p_j > 0$ which sum to 1.

Analysis Summary

The *Analysis Summary* displays the R commands that were executed.

Bivariate Mixture Models

```
d<-
read.csv("C:\\Users\\Neil\\AppData\\Local\\Temp\\data.csv",dec=".",sep=",",stringsAsFactors=TRUE
)
setwd("C:\\Users\\Neil\\AppData\\Local\\Temp\\")
library("EMCluster")

## Warning: package 'EMCluster' was built under R version 3.2.5

## Loading required package: MASS

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 3.2.5

set.seed(8604)
AIC<-vector(mode="numeric",length=0)
BIC<-vector(mode="numeric",length=0)
CLC<-vector(mode="numeric",length=0)
logL<-vector(mode="numeric",length=0)
for (clusters in 1:2) {
ret<-init.EM(d,nclass=clusters,method="Rnd.EM")
em<-em.ic(d,ret)
AIC<-c(AIC,em$AIC)
BIC<-c(BIC,em$BIC)
CLC<-c(CLC,em$CLC)
logL<-c(logL,ret$llhdval)
}
summary(ret)

## Method: Rnd.EM
##  n = 130, p = 2, nclass = 2, flag = 0, total parameters = 11,
##  logL = -566.8589, AIC = 1155.7179, BIC = 1187.2607.
## nc:
## [1] 109  21
## pi:
## [1] 0.8624 0.1376

write.table(ret$pi,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\props.csv",sep=",")
write.table(ret$Mu,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\mus.csv",sep=",")
write.table(ret$LTSigma,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\sigmas.csv",sep=",")
write.table(AIC,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\aic.csv",sep=",")
write.table(BIC,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\bic.csv",sep=",")
write.table(logL,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\logl.csv",sep=",")
write.table(CLC,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\clc.csv",sep=",")
```

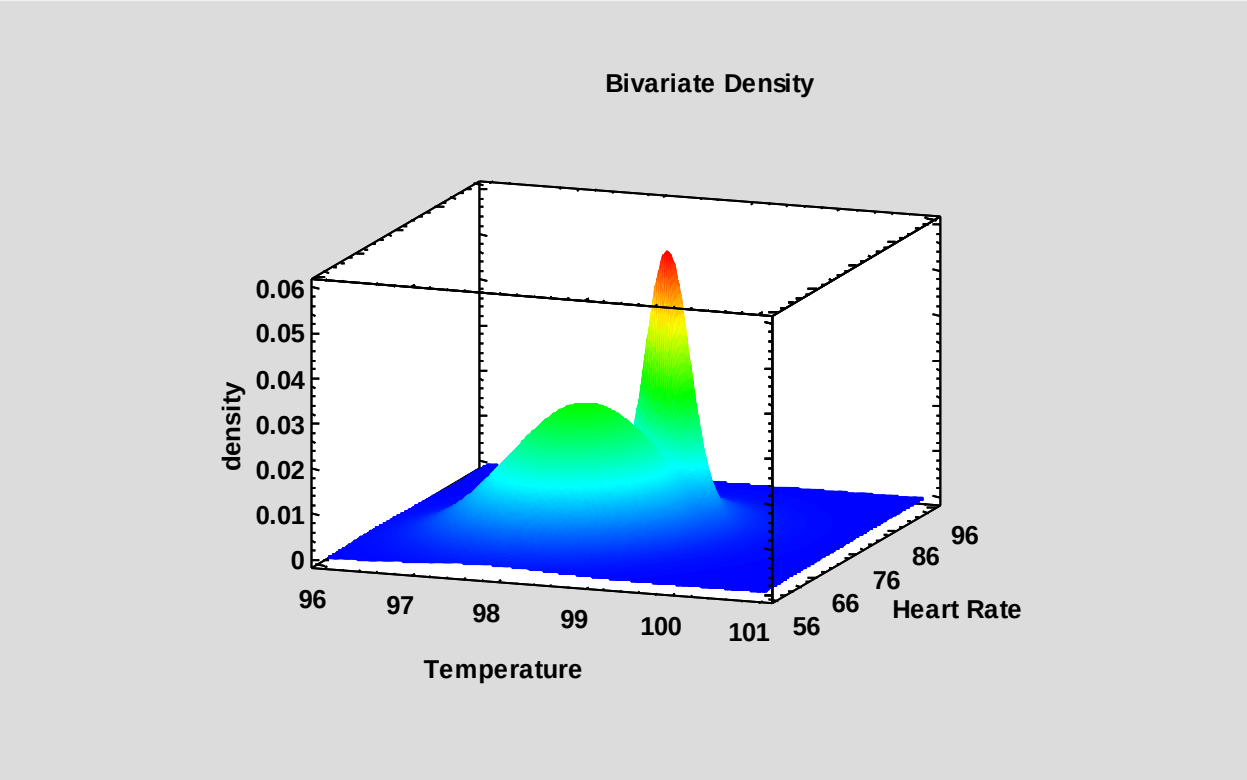
In the lines below *summary(ret)*, it summarizes the results of the model fitting process. Of particular interest are:

1. *n*: the number of observations used to fit the model.
2. *nclass*: the number of components in the fitted model.
3. *total parameters*: the number of estimated parameters.
4. *logL:* the final value of the log likelihood function.
5. *AIC:* the value of the Akaike Information Criterion.
6. *BIC:* the value of the Bayesian Information Criterion.
7. *nc*: the number of observations in each component of the model.

8. *pi*: the estimated proportion of the distribution for each component.

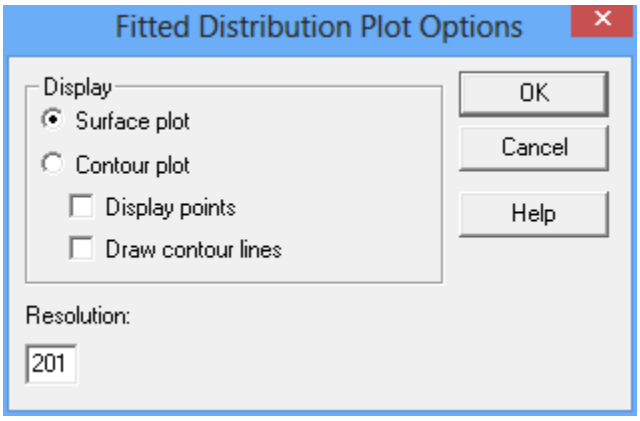
Plot of Fitted Distribution

This plot shows the fitted bivariate density function.



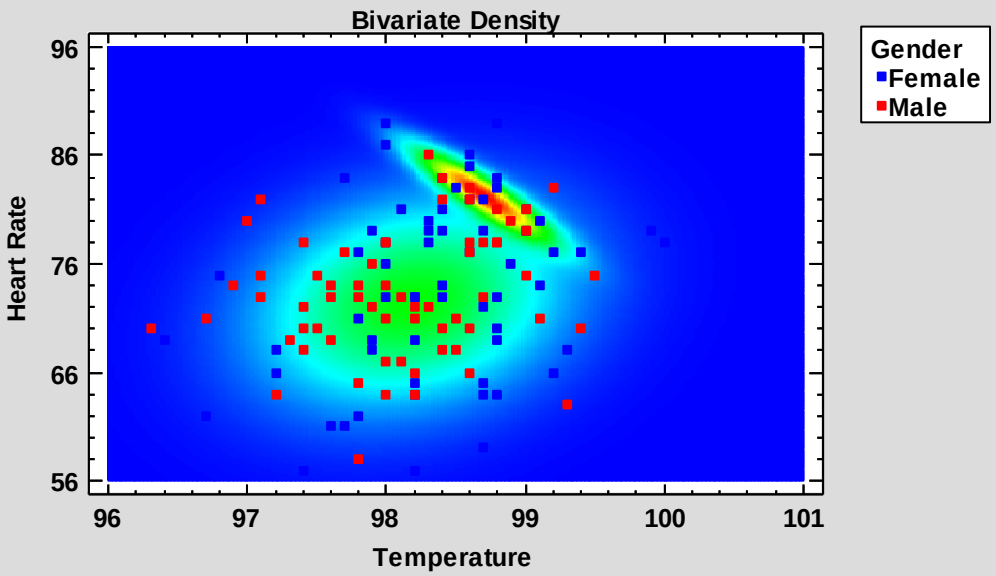
It is a mixture of 2 bivariate normal distributions.

Pane Options



- **Display:** select either a surface or contour plot.
- **Display points:** if creating a contour plot, whether individual observations should be displayed.
- **Draw contour lines:** if creating a contour plot, whether lines should be drawn rather than using a palette of colors.
- **Resolution:** the number of locations along each axis at which the density function is evaluated.

The plot below shows a contour plot.



The color of each point indicates whether that point corresponds to a male or a female. The plot seems to indicate a primary component at low temperature and low heart rate with little correlation between the 2 variables, and a secondary component at higher temperatures and heart rates with a strong negative correlation.

Estimated Parameters

This table shows the estimated model parameters:

Estimated Parameters

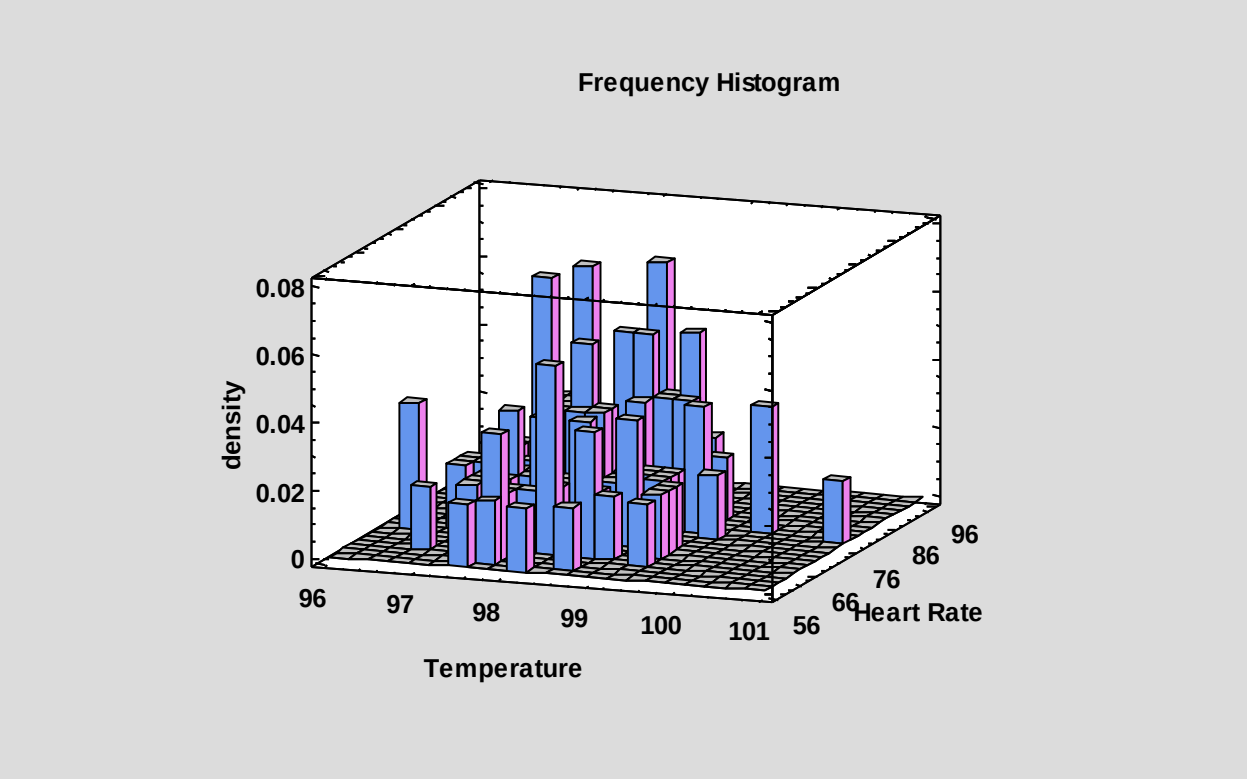
Sample size: 130

Component	*Proportion*	*Mean 1*	*Mean 2*	*Sigma 1*	*Sigma 2*	*Correlation*
1	0.862423	98.1814	72.3688	0.752765	6.47316	0.198505
2	0.137577	98.6747	82.4923	0.339429	2.94411	-0.898109
Combined		98.2492	73.7615	0.730358	7.03486	0.143541

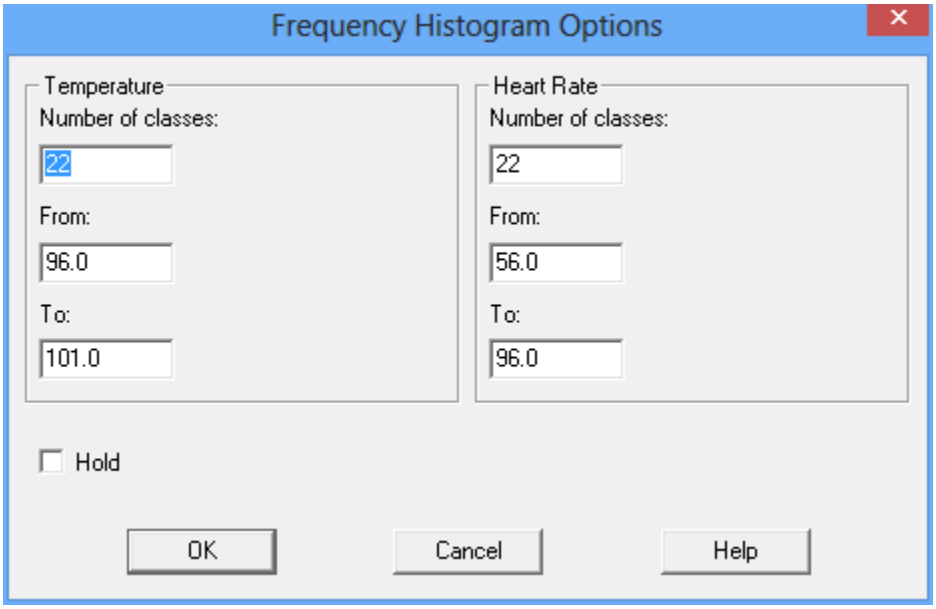
It includes estimates of the 2 means for each component, the 2 component standard deviations, the correlation coefficients and the component proportions p_j. The means, standard deviations and correlation coefficient of the combined distribution is also displayed.

Frequency Histogram

This plot shows a frequency histogram for the data. The height of each bar is proportional to the number of observations observed in a small rectangular area defined by a range of *Temperature* and *Heart Rate*.



Pane Options



For each variable, specify:

- **Number of classes:** number of intervals into which the range of the data should be divided.
- **From:** lower limit of the first class.
- **To:** upper limit of the last class.

Also specify:

- **Hold:** if checked, the scaling of the classes will remain constant even if the data change.

Model Comparisons

This table compares the results of fitting mixtures of Gaussian distributions with components varying between 1 and *K*, the number of components indicated on the *Analysis Options* dialog box.

Model Comparisons

Number of components	*Parameters*	*Log likelihood*	*AIC*	*BIC*	*CLC*
1	5	-577.367	1164.73	1179.07	1154.73
2	11	-566.859	1155.72	1187.26	1163.14
3	17	-560.885	1155.77	1204.52	1183.9
4	23	-559.125	1164.25	1230.2	1180.2

The table includes:

1. **Parameters** - the number of estimated parameters *m* in the model. For a model with K components, $m = 6K\text{-}1$.

2. **Log likelihood** - the value of the log likelihood function $ln(\hat{L})$. Models with more parameters will always have larger values of the log likelihood function.

3. **AIC** – the value of the Akaike Information Criterion. The AIC is a widely used criterion for model selection and penalizes the likelihood function based on the number of estimated parameters according to

$$AIC = 2m - 2ln(\hat{L}) \tag{3}$$

4. **BIC** – the value of the Bayesian Information Criterion. It is similar to the AIC except that it uses a different equation to penalize the likelihood function:

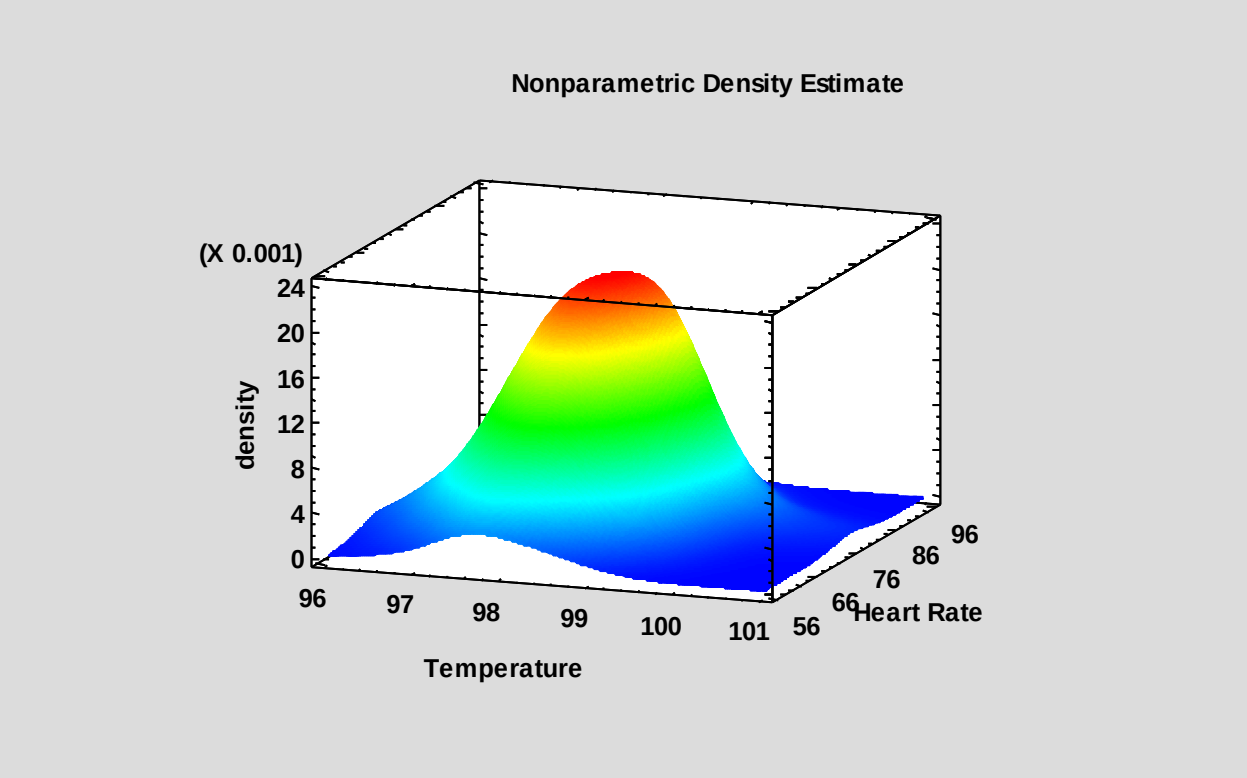
$$BIC = \ln(n)m - 2ln(\hat{L}) \tag{4}$$

5. **CLC** – the value of the Classification Likelihood Criterion. It is similar to AIC and BIC but penalizes the likelihood function based on an entropy measure (see Biernacki and Govaert 1997).

Models with the smallest values of AIC, BIC and CLC are preferable. In the table above, the AIC selects a model with 2 components while the other criteria select a model with only 1 component.

Nonparametric Density Estimate

An alternative estimate of the bivariate density function may be obtained by counting the number of observations which fall within a window of fixed size moved across the range of the data.



The estimated density function is given by:

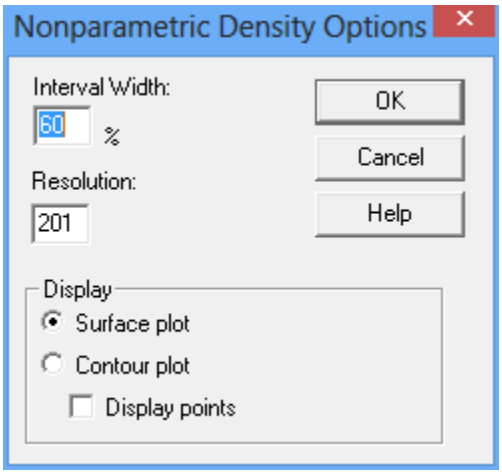
$$f(x) = \frac{(\det S)^{-1/2}}{h^2 n} \sum_{i=1}^{n} W\left(\frac{1}{h^2} \left(X_{1,i} - X_1\right)^T S^{-1} \left(X_{2,i} - X_2\right)^T \right) \quad (5)$$

where S is the sample covariance matrix of the 2 variables, h is the width of the window and $W(u)$ is the weighting function defined by

$$W(u) = \frac{1}{2\pi} \exp(-u/2) \quad (6)$$

A width of 50% is not unreasonable for a small sample but may not give as much detail as a smaller value in larger samples.

Pane Options



- **Interval Width:** the width of the window *h* within which observations affect the estimated density, as a percentage of the range covered by the x-axis. *h* = 60% is not unreasonable for a small sample but may not give as much detail as a smaller value in larger samples.

- **Resolution**: the number of points along each axis at which the density is estimated.

- **Display:** select either a surface or contour plot.

- **Display points:** if creating a contour plot, whether individual observations should be displayed.

Component Assignments

To determine which component each of the observed data values is most likely to belong to, the program compares the height of the weighted component probability density functions at each value. It then assigns an observation to that value for which the weighted pdf is greatest. A table is given showing the component assignments:

Component Assignments

Group Percentages

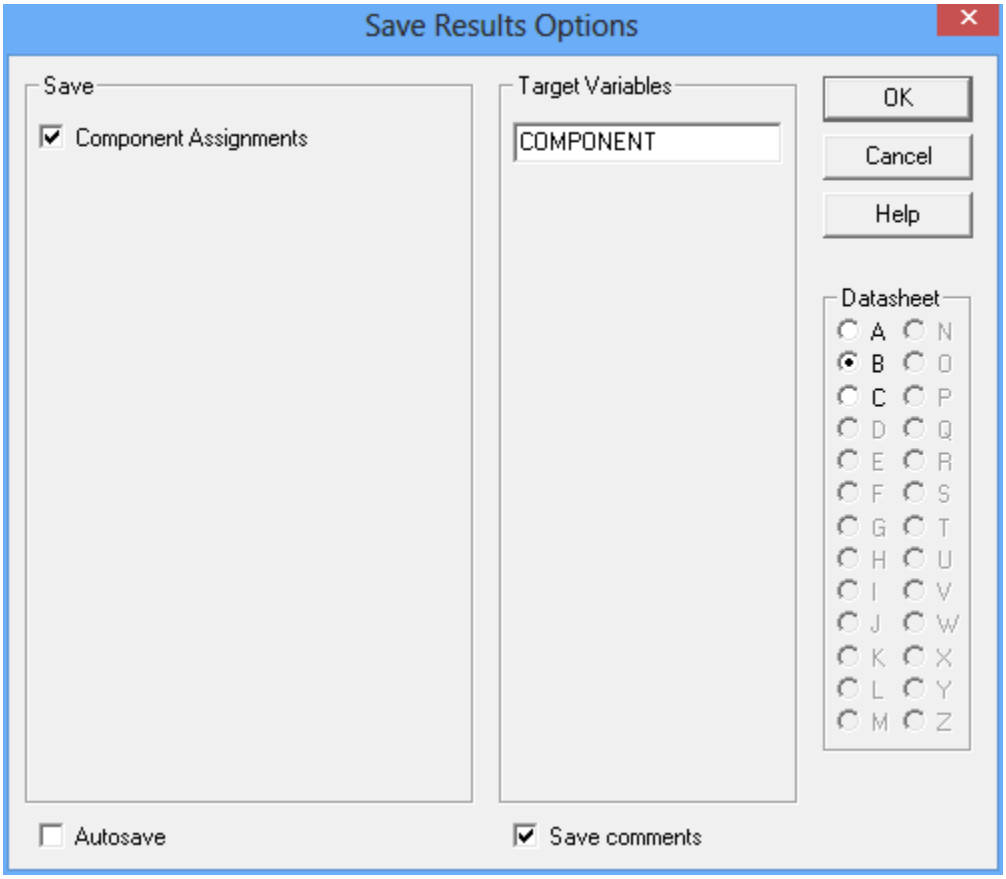
Gender	*Count*	*Component 1*	*Component 2*
Female	65	80.00%	20.00%
Male	65	87.69%	12.31%
All	130	83.85%	16.15%

Row	*Temperature*	*Heart Rate*	*Component*
1	98.4	84.0	2
2	98.4	82.0	1
3	98.2	65.0	1
4	97.8	71.0	1
5	98.0	78.0	1
6	97.9	72.0	1
7	99.0	79.0	2
8	98.5	68.0	1
9	98.8	64.0	1
10	98.0	67.0	1
11	97.4	78.0	1
...	...	...	...

For the sample data, about 84% of the observations have been assigned to Component #1 and 16% to Component #2. If an entry was made in the *Group* field on the data input dialog box, the assignment percentages will also be displayed by group.

Save Results

The component assignments may be saved in a Statgraphics datasheet by pressing the *Save Results* button on the analysis toolbar. The following dialog box will be presented:



- **Component Assignments** – most likely component associated with each observation.
- **Target Variables:** enter names for the columns to be created.
- **Datasheet:** the datasheet into which the results will be saved.
- **Autosave:** if checked, the results will be saved automatically each time a saved StatFolio is loaded.
- **Save comments:** if checked, comments for each column will be saved in the second line of the datasheet header.

References

Biernacki, C., Celeux, G, and Govaert, G. (2003). "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models." Computational Statistics and Data Analysis, **413**, 561-575.

Biernacki C and Govaert G (1997). "Using the classification likelihood to choose the number of clusters." Computing Science and Statistics **29**, 451–457.

Maitra R. (2001). "Clustering massive datasets with applications to software metrics and tomography." Technometrics, **43**(3), 336-346.

Maitra R. (2009). "Initializing Partition-Optimization Algorithms." IEEE/ACM Transactions on Computational Biology and Bioinformatics, **6**, 144-157.

R Package "EMCluster" (2018) - https://cran.r-project.org/web/packages/EMCluster/EMCluster.pdf

Wei-Chen Chen and Ranjan Maitra (2015) – A Quick Guide for the EMCluster Package.

