AroniSmartLytics™
Manual

By *AroniSoft LLC*

Web: http://www.aroni.us
e-mail: aronisoft@aroni.us
or aroni@aroni.us

I. Welcome to *AroniSmartLytics™*

Thank you for selecting ***AroniSmartLytics,*** the leading applied and research statistical, data mining and Big Data tool built for the Mac OS X platforms. **AroniSmartLytics™** is built on the premises that Statistics and Data Mining techniques need to be user friendly tools for Researchers, Students, Business Analysts, Practitioners and others to focus on the problem of interest not on the complexities of applications or writing computer code. With the help of Apple®'s OS X®, ***AroniSmartLytics*** has accomplished that. We recommend that you read this manual to familiarize yourself with the installation and basic operation of ***AroniSmartLytics.*** You may also wish to read or skim through and skip to any chapters that cover features you frequently use. Softwares, especially this, are most useful when used. In fact, as and because you use it, you get more insights into the world of Statistics and Data Mining.

AroniSoft LLC.
P. O. Box 1104
Fair Lawn, NJ 07410-1104
United States of America
Web:

- http://www.aroni.us
- http://www.smartlytics.tech

Sales information: aroni@aroni.us
Technical support: **aroni@aroni.us**

Copyrights and Trademarks

License and Disclaimer : AroniStat™ , AroniSmartStat™, AroniSmartIntelligence™, AroniSmartLytics™, AroniStatMobile™, AroniSmartInvest™ and AroniSmartInvestMobile™ Terms and Disclaimer

To use AroniStat™ , AroniSmartStat™, AroniSmartIntelligence™, AroniSmartLytics™, AroniStatMobile™, AroniSmartInvest™ or AroniSmartInvestMobile™ (the Software) you must first agree to the following terms. If you do not agree, stop using the software and uninstall the software and any of its parts. Your use of the Software shall constitute your acceptance of this Agreement and its terms.

ARONISOFT LLC PROVIDES THE SOFTWARE "AS-IS" AND PROVIDED WITH ALL FAULTS. NEITHER ARONISOFT LLC NOR ANY OF ITS SUPPLIERS OR RESELLERS MAKE ANY WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. ARONISOFT LLC AND ITS SUPPLIERS SPECIFICALLY DISCLAIM THE IMPLIED WARRANTIES OF TITLE, NON-INFRINGEMENT, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, SYSTEM INTEGRATION, AND DATA ACCURACY. THERE IS NO WARRANTY OR GUARANTEE THAT THE OPERATION OF THE SOFTWARE WILL BE UNINTERRUPTED, ERROR-FREE, OR VIRUS-FREE, OR THAT THE SOFTWARE WILL MEET ANY PARTICULAR CRITERIA OF PERFORMANCE, QUALITY, ACCURACY, PURPOSE, OR NEED. YOU ASSUME THE ENTIRE RISK OF SELECTION, INSTALLATION, AND USE OF THE SOFTWARE. THIS DISCLAIMER OF WARRANTY CONSTITUTES AN ESSENTIAL PART OF THIS AGREEMENT. NO USE OF THE SOFTWARE IS AUTHORIZED HEREUNDER EXCEPT UNDER THIS DISCLAIMER.

THIS SOFTWARE DEPENDS ON CERTAIN DATA TRANSFER PROTOCOLS THAT HAVE BEEN DEVELOPED AND ARE MAINTAINED BY THIRD PARTY DATA PROVIDERS. THESE PROVIDERS MAY DECIDE TO CHANGE THESE PROTOCOLS AT ANY TIME WITH OR WITHOUT NOTICE. ARONISOFT LLC WILL MAKE ANY REASONABLE EFFORT TO KEEP THE SOFTWARE UP TO DATE IN RESPECT TO THESE CHANGES, BUT DOES NOT GUARANTEE THE TIMELINESS OF THESE UPDATES, OR THAT IT WILL EVEN BE POSSIBLE TO INCORPORATE SOME OR ALL OF THESE CHANGES. SOME OF THE CONSEQUENCES OF SUCH CHANGES BY THIRD PARTY DATA PROVIDERS MAY BE, BUT NOT LIMITED TO: INACCURATE OR UNAVAILABLE MARKET DATA, POOR PERFORMANCE AND OTHERS. YOU ASSUME THE ENTIRE RISK OF ALL SUCH OCCURRENCES, AS WELL AS THE RISK THAT THE SOFTWARE MAY BE RENDERED TEMPORARILY OR PERMANENTLY, IN PART OR COMPLETELY NONFUNCTIONAL BY SUCH CHANGES.

ARONISOFT LLC DOES NOT MAKE ANY GUARANTEE THAT THIS SOFTWARE COMPLIES OR WILL CONTINUE TO COMPLY WITH THE THIRD PARTY DATA PROVIDER TERMS OF SERVICE OR ANY OTHER AGREEMENTS BETWEEN THE THIRD PARTY DATA PROVIDER AND YOU. YOU ARE SOLELY AND FULLY RESPONSIBLE FOR COMPLYING WITH THE THIRD PARTY DATA PROVIDER TERMS OF SERVICE AND ANY OTHER AGREEMENTS BETWEEN THIRD PARTY DATA PROVIDER AND YOU.

Limitation of Liability. Neither AroniSoft LLC nor anyone else who has been involved in the creation, production or delivery, of this software or its accompanying documentation shall be liable for any direct,

indirect, incidental, special, exemplary or consequential damages whatsoever, including but not limited to any loss of actual or anticipated profits or benefits, resulting from the use of the software or its documentation.

Complete Agreement. You acknowledge that you have read this license agreement and that it is complete and exclusive agreement between you and AroniSoft LLC, regarding the software and its documentation. This License shall be governed by and construed in accordance with the laws of the State of New Jersey and Delaware, United States of America, as if performed wholly within the states and without giving effect to the principles of conflict of law. If any portion hereof is found to be void or unenforceable, the remaining provisions of this License shall remain in full force and effect. This License constitutes the entire License between the parties with respect to the use of the Software

General Provisions. AroniSoft LLC reserves all rights not expressly granted herein. AroniSoft LLC may modify this Agreement at any time by providing such revised Agreement to you or posting the revised Agreement on the product website presently located at http://www.aroni.us. Your continued use of the Software shall constitute your acceptance of such revised Agreement. Nothing in this Agreement shall constitute a partnership or joint venture between you and AroniSoft LLC. Should any term or provision hereof be deemed invalid, void or unenforceable either in its entirety or in a particular application, the remainder of this Agreement shall nonetheless remain in full force and effect.

Termination. You may terminate this agreement by returning all your copies of the software and documentation to AroniSoft LLC. This license terminates automatically and without notice to you if you fail to comply with any provisions of this agreement. You agree to return all copies of the software and documentation to AroniSoft LLC upon termination.

YOU EXPRESSLY ACKNOWLEDGE THAT YOU HAVE READ THIS AGREEMENT AND UNDERSTAND THE RIGHTS, OBLIGATIONS, TERMS AND CONDITIONS SET FORTH HEREIN. BY CONTINUING TO INSTALL THE SOFTWARE, YOU EXPRESSLY CONSENT TO BE BOUND BY ITS TERMS AND CONDITIONS AND GRANT TO ARONISOFT LLC THE RIGHTS SET FORTH HEREIN.

YOU AGREE TO USE THE SOFTWARE AT YOUR OWN RISK.

For more information, Contact us at:

e-mail: aronisoft@aroni.us or aroni@aroni.us

web: http://www.aroni.us

I. WELCOME TO *ARONISMARTLYTICS™* 2

II. WHY ARONISMARTLYTICS™ 8

III. THE WORLD OF STATISTICS AND ANALYTICS WITH ARONISMARTLYTICS™ 9

IV. KEY ARONISMARTLYTICS™ FEATURES 13

8) CALCULATE THE PROBABILITIES: SELECT A DISCRETE OR CONTINUOUS DISTRIBUTION FROM THE BROWSER, ENTER THE INPUT PARAMETERS AND CALCULATE THE PROBABILITIES BASED ON THE SELECTED DISTRIBUTION. CHANGE THE INPUT PARAMETERS AND SEE HOW THE PROBABILITY DISTRIBUTIONS CHANGE. 14

V. MODULE 1: PARAMETRIC AND NON PARAMETRIC STATISTICS. 16

A. REFERENCE ABOUT THE STATISTICAL CONCEPTS, THE MOST USED PROBABILITY DISTRIBUTIONS, THE RELATIONSHIP AMONG PROBABILITY DISTRIBUTIONS, KEY STATISTICAL TESTS, AND HOW TO SELECT A STATISTICAL TEST. 22

B. ANNOTATION 23

C. RELATIONSHIPS AMONG COMMON DISTRIBUTIONS ERROR! BOOKMARK NOT DEFINED.

D. ACCESSING OTHER RESOURCES ON INTERNET WEB PAGES 23

VI. MODULE 2: ARONISMARTLYTICS™ PROBABILITY DISTRIBUTIONS ERROR! BOOKMARK NOT DEFINED.

A. GRAPHING CAPABILITY 24

B. CALCULATE THE PROBABILITIES 24

C. PROBABILITY FUNCTION STATISTICS 24

VII. MODULE 3: ARONISMARTLYTICS™ DESCRIPTIVE STATISTICS 26

VIII. MODULE 4: ARONISMARTLYTICS™ ANALYSIS AND TESTING 26

IX. MODULE 5: ARONISMARTLYTICS™ REGRESSION ANALYSIS 28

X. MODULE 6: SEGMENTATION AND CLASSIFICATION 32

XI. FINITE MIXTURE GAUSSIAN MODELS WITH ARONISMARTLYTICS™ 34

SETTING UP THE SEGMENTATION MODEL IN ARONISMARTLYTICS 35

SETTING UP THE FINITE MIXTURE GAUSSIAN MODELS IN ARONISMARTLYTICS 35

CHOOSING THE MAIN ANALYSIS MODEL 35

INPUT DATA AND DATA PRESENTATION 36

SIMULATING A SAMPLE FROM A NORMAL DISTRIBUTION 37
BOOTSTRAP ASSESSMENT OF NUMBER OF COMPONENTS 39
FIT A FIXED G-COMPONENT MODEL NORMAL MIXTURE MODEL 40
FIT A RANGE OF G-COMPONENT MODEL NORMAL MIXTURE MODEL 42
DISCRIMINANT ANALYSIS 43
PREDICTION OF NEW DATA 44
FINAL CONSIDERATIONS ON EXTRA-OPTIONS FOR FINITE MIXTURE MODELS 46

XII. GENERAL MIXED MODELS WITH ARONISMARTLYTICS™ 46

INTRODUCTION 47

SETTING UP THE GENERAL MIXTURE MODEL IN ARONISMARTLYTICS 48

CHOOSING THE MAIN ANALYSIS MODEL 48

A. WEIGHT AND TRAINING AND PARTITION DATA FORMAT. 49
B. INITIALIZATION DATA FORMAT. 49
1. GAUSSIAN 50
2. MULTINOMIAL 50
3. GAUSSIAN HIGH DIMENSIONAL 51
C. GENERAL MIXED MODEL INPUTS 52
1. REQUIRED INPUTS 52
2. OPTIONAL INPUTS 52
3. STRATEGY AND ALGORITHMS 54

SETTING UP THE MODEL 57

CLUSTER ANALYSIS 57
DISCRIMINANT ANALYSIS 58

XIII. BAYESIAN MODELS, BIG DATA, AND TEXT MINING WITH ARONISMARTLYTICS™ 59

BIG DATA, TEXT MINING , UNSTRUCTURED DATA AND BAYESIAN MODELS 59

BIG DATA, UNSTRUCTURED DATA AND BAYESIAN NETWORK MODELS 59

PROCESSING UNSTRUCTURED TEXT IN IN ARONISMARTLYTICS™ 60

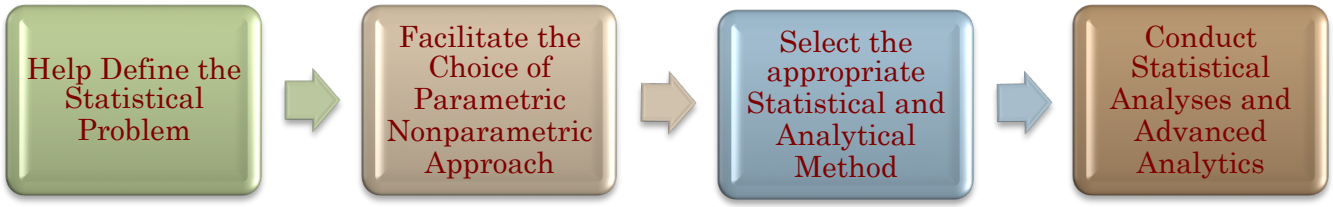
SETTING UP THE BAYESIAN NETWORK MODELS IN ARONISMARTLYTICS™ 64

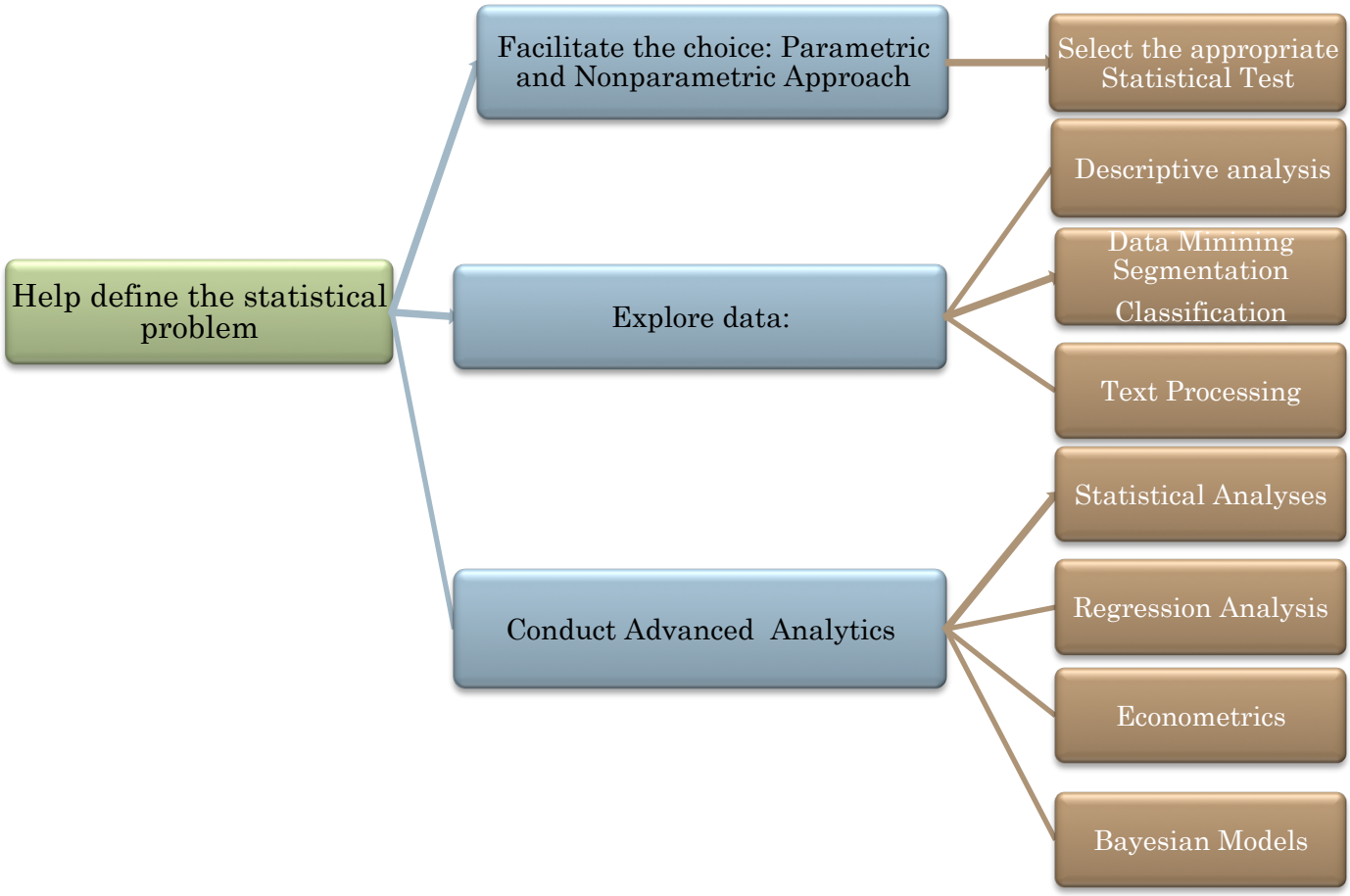
XIV. ARONISMARTLYTICS™ ADDITIONAL HELP 65

II. Why AroniSmartLytics™

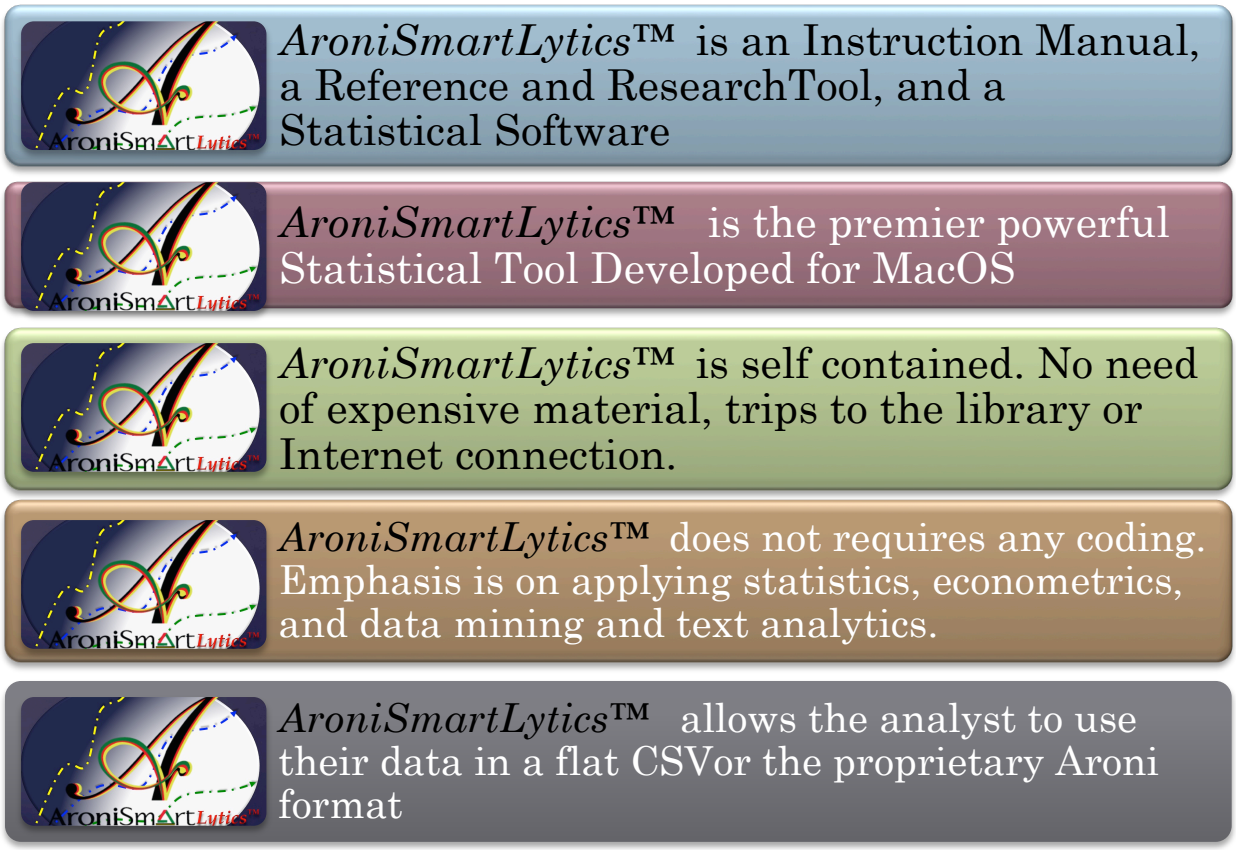
roniSmartLytics ™ is an applied statistics and data mining reference, electronic handbook, and powerful research tools. The tool is intended for beginners, students of statistics, casual and regular users and advanced statisticians, data miners and researchers.

The uniqueness and benefits of ***AroniSmartLytics™*** can be described in four steps:





Several statistical tools exist. AroniSmartLytics™ is the first in the kind:

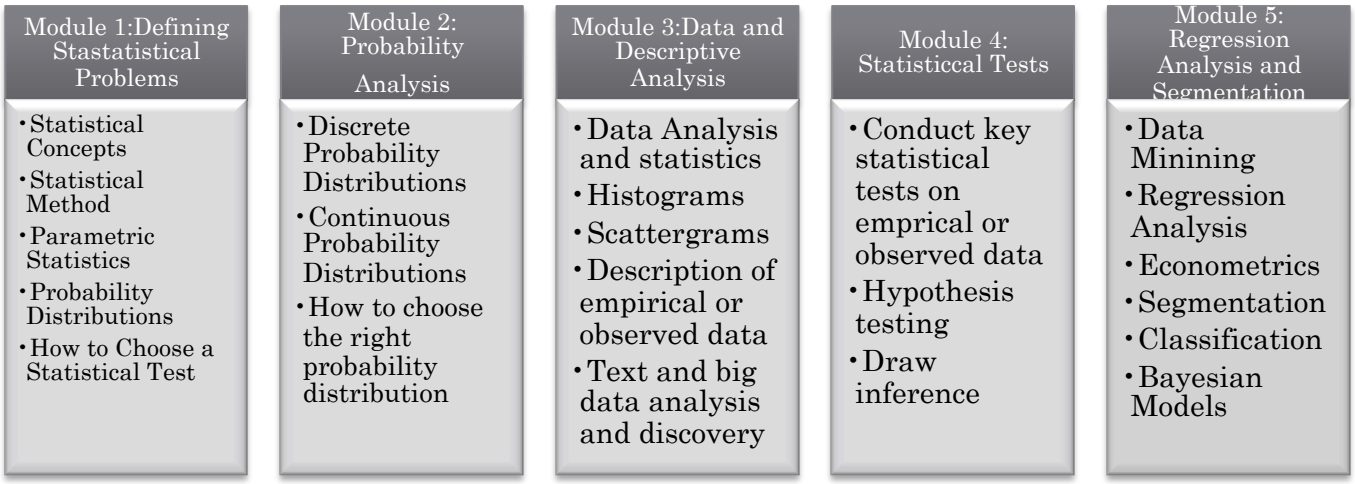


III. The World of Statistics and Analytics with AroniSmartLytics™

Researchers, Scientists, Engineers, Students, Politicians, Economists, Political Analysts, Business Analysts, Financial Analysts, Statisticians, and most of us the **practitioners** use statistics, econometrics, data mining, text analytics and other analytical techniques to answer questions or to support decision making that require data analysis. The analytical techniques help to describe the results of, among others, investigations, experiments, and observations. When data about the subject or object of interest is somehow comprehensive, data analysis using ***descriptive statistics*** may be what is needed.

Unfortunately, in the world of statistics, econometrics, data mining, text analytics, and discovery, things may not be that straightforward. The population of interest may not be fully observable or accessible. Hence, the statistician will have to make inference or, in statistics terms, appeal to ***inferential statistics***. Inferential statistics involve statistical tests. One of the major issues facing statisticians and researchers is to choose which test to use. The choice involves first a decision between two families of test statistics: ***parametric*** and ***nonparametric***. Then, a choice of specific statistical test must be made.

AroniSmartLytics™ intends to help beginners, practitioners and most advanced statisticians, analysts and researchers to navigate this tricky step in learning, research, and interpretation and application of or action on results. To achieve these goals, ***AroniSmartLytics™*** is built around five modules that provide the five main benefits:



1. The specific module of ***AroniSmartLytics™*** dedicated to Statistical tests offers an in-depth, user friendly and intuitive reference and selection tool that will facilitate the task, save time, while being rigorously thought out. In this module, ***AroniSmartLytics™*** shows in a unique graphical template the key relationships and the connectivity among statistical variables. This visual and intuitive tool offers to statisticians, econometricians and researchers an invaluable reference for linking and deriving key probability distributions, and hence arriving quickly to solutions.

   Beginners and casual or regular users will quickly find this tool easy to use and handy. Advanced users and researchers will have at their fingertip a powerful research and reference tool.

2. The module dedicated to Probability Distribution is a user-friendly interface with most of the usual statistical distributions, their properties and their graphical representations.
3. For researchers, advanced statisticians, econometricians, data analysts, students, and other users, the three remaining modules: Descriptive analysis, Statistical testing, and Regression Analysis and Segmentation help conduct descriptive analyses, conduct statistical analyses, test hypotheses, conduct regression analysis and perform segmentation and classification on real or empirical data provided by the users themselves.
4. With the advance in Big Data, Text Mining, and Discovery, Bayesian Models and Bayesian Networks have become an indispensable tool, in the hands of more savvy analysts. AroniSmartLytics® has a module dedicated to Bayesian Models and Network.
5. The graphing capability included in the tool will help to easily simulate distributions and to create histograms and scattergrams.

Statistics may seem, at sometimes, overwhelming for beginners or may require remembering several mathematical formulas. Hence Statisticians and users of statistics need a tool that can help refresh their memories and eventually provide the link needed to connect statistical concepts, while using it to solve real life problems. AroniSmartLytics™ implements an approach to quickly understand statistical methods and at the same time conduct rigorous statistical researches, without writing any code.

AroniSmartLytics™ is self-contained. Some, if not most, concepts contained in the tool may be found on Internet and library resources. AroniSmartLytics™ will give that on the user's computer, at the fingertips. No need for internet connection, multiple searches of internet search engines and portals, or time consuming visits to libraries. AroniSmartLytics™ user does not need pages to seep through or manuals to memorize. The tool will always be at computer length. At the same time, **AroniSmartLytics™** is **a statistical and analytical software destined to be used by researchers, students, and practitioners.**

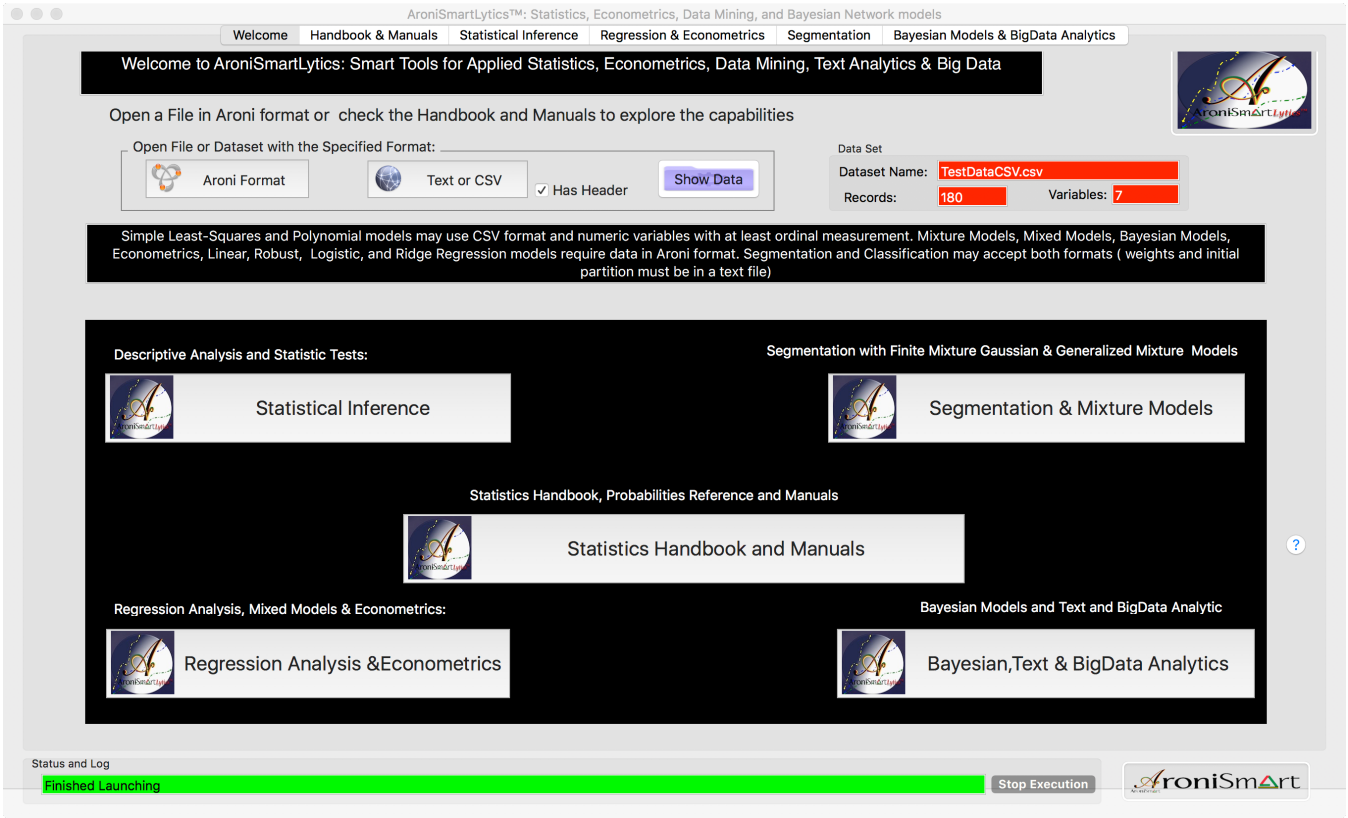
AroniSmartLytics™ is a cost effective tool, either in time, ease of use, access, or price. It is the best statistical reference, research and analytics tool for the busiest and savvy statistician and statistical practitioner and researcher and data mining expert. The following sections detail the features of AroniSmartLytics™

Enjoy the world of Statistics, Analytics, Econometrics, Data Mining, and Text Analytics with AroniSmartLytics™

IV. Key AroniSmartLytics™ Features

AroniSmartLytics™ is built on the premises that Statistics, Analytics, and Data Mining need to be user friendly tools for Researchers, Students, Practitioners and others to focus on the problem of interest not on the complexities of applications or writing computer code.

AroniSmartLytics™ assumes that the user is familiar with basic statistical concepts and needs a reference and research tool. While the emphasis in on research and productivity, it is assumed that the user would like also to have a tool that can help cement the statistical, analytical, and data mining concepts and can be used as a reference for more in depth study, analysis, or research. The tool is designed as on-road statistical learning and research companion. It is a versatile and powerful research, statistical, analytical and data mining tool with many and innovative features.



Key main features unique to AroniSmartLytics™:

1) **Probability function statistics:** Formulas for probability function, cumulative function, mean, variance, median, mode, moment generating function, skewness and kurtosis are given in the embedded pdf document.
2) **Relationships among common distributions.** This visual representation maps on one page the relationships among the key statistical distributions
3) **How to select parametric test.** This visual representation maps on one page steps on how to select an appropriate parametric test and how to calculate the test.
4) **How to select nonparametric test.** This visual representation shows the path from defining a statistical problem, choosing the type of data, selecting the non parametric test and describing the test itself, its assumptions and its mathematical and statistical properties.
5) **Annotation:** the main pdf document in which probability or tests are described can be annotated through the Tools menu and saved for further reference. The annotation gives the ability to add notes and references in the document for further enriching the statistical concepts.
6) **Descriptive statistics on observed or empirical data.** AroniSmartLytics™ has a statistical engine to compute all the statistics from the data provided by the user. All the statistics described in 1) are produced for each variable or sample of data.
7) **Graphing Capability:** Select a discrete or continuous distribution from the browser, and observe the graph of that distribution. Change the input parameters and see how the distribution changes its form and shape. For empirical data, select a variable to plot its histogram, or two variables to plot a scattergram. The plots will help in making the decision on which probability distribution better describes the data, choosing between parametric and non parametric testing and selecting a parametric test.
8) **Calculate the probabilities:** Select a discrete or continuous distribution from the browser, enter the input parameters and calculate the probabilities based on the selected distribution. Change the input parameters and see how the probability distributions change.
9) **Conduct the statistical testing.** This is ultimately the first of the two main goals of **AroniSmartLytics™**. **AroniSmartLytics™** is a powerful statistical, analytical and research tool. The engine for conducting the statistical tests is at the core of the

software application and drives all other statistical analyses. The software application gives the ability to the users to enter their data, conduct descriptive and advanced analyses, chose a test, and conduct the test.

10) **Regression Analysis, Segmentation and Classification.** This is ultimately the second of the two main goals of **AroniSmartLytics™**. **AroniSmartLytics™** is a powerful statistical, analytical and research tool. Regression analysis, segmentation and classification are at the heart of quantitative analyses critical to running businesses, conduct researches and studies and making informed decisions. The software application gives the ability to the users to enter their data in CSV format or, in a more flexible proprietary Aroni format that can handle numeric interval (real or integer), nominal, binary and text formats.

11) **Bayesian Models and Network and Text Mining.** Big Data has been here and exploding. Businesses, individual and the society have been slowly adjusting to the new trend in technology, data mining and analytics adapted to the needs of Big Data. AroniSoft, with its AroniSmart™ line of technology and analytical tools and methodologies, has been working to make sure our customers will be able to harness the benefits offered by Big Data. **AroniSmartLytics™** has a module dedicated to Big Data: Bayesian Models. The module is dedicated to analyzing text data and discovering network relationships among different variables and attributes. The module include several Bayesian Network models, estimators and algorithms to deal with unstructured data. **AroniSmartLytics™** models have been optimized for the needs of Big Data analytics.

To support text and unstructured data mining, a Text processing capability has been added in the Bayesian and BigData Analytics module. AroniSmartLytics™ allows the analyst to load plain text from various sources, organized in folders and sub-folders and create a dataset in "Aroni" format, to be processed in the Bayesian Models, Regression Analytics, or Segmentation modules

V. Welcome: Selecting the Right Module

The first tab of AroniSmartLytics™ presents the user with an intuitive user graphics interface to help in selecting the right module and loading the dataset to analyze:

1) **Load and view the Dataset to Analyze.** From the top of the module, open the dataset may be opened or explored. Two data formats are available: the proprietary "Aroni" and Text or CSV. Two data formats are allowed: CSV and Aroni. The data is loaded from the folders selected by clicking the Aroni Format or Text CSV button. If a dataset in Text and CSV format has a header, make sure the "Has Header" checkbox is selected.

   CSV dataset must meet the following requirements:

   a. Be a text file in comma separated vector(csv) format.

   b. Only **numerical variables are available for further analysis.** Hence, class or categorical variables, in string or other format, need to be transformed into numerical coding if they will be used in further analysis.

   c. Preferably one column must be an **identifier and labeled Observation** or Identifier. If the Observation or Identifier column is missing, the system assigns a default observation column, with default observation numbers.

   d. **Empty or missing values** are expected to be replaced by a period "." in the data set test file and in table viewer, although non consecutive blanks are also acceptable.

   e. **The first column** must be a header column. Otherwise, unselect the check box "Column With Headers" to assign default Column headers.

   Follow the steps to open the file. The file is loaded in the table in the middle left , each column corresponding to a variable, with Observation as the first variable.

   Aroni dataset format must be as follows:

   Aroni formatted data has three sections:

   1. An option **comments section** at the top. The comments lines must start with /*, *, */, % or //

2. The next section, which is required, is the **declaration section**. That is where the **attributes of the data** and the **variables** are defined.
3. The third section is the **data section.** This is where the data, in columns is entered. Columns may be whitespace or comma separated.

The data in Aroni format may look like the following:

A) Comment Section

```
/*****************************************/
/*1. Title: Aroni Data Example
 *This is the comment section
 *Be free to comment as much as you like
 * This may end a comment */
 // This is also a comment
 %As is this
```

B) Declaration Section

```
@DATASET datasetName
@VARIABLE variable1 NUMERIC
@VARIABLE  variable2 NUMERIC
@VARIABLE  variable3 NUMERIC
@VARIABLE variable4 NUMERIC
@VARIABLE  dependentVariable  {Iris-setosa,Iris-versicolor,Iris-virginica}
```

C) **The Data section of Aroni format** is as follows:

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
```

5.0,3.6,1.4,0.2,Iris-setosa

5.4,3.9,1.7,0.4,Iris-setosa

The @DATASET, @VARIABLE and @DATA declarations are protected words and must be written as such. However they are case insensitive and can be @dataset, @variable or @data.

The @DATASET Declaration

The dataset name is defined as the first line after the potential comments in the file. The format is:

@dataset <dataset-name> where <dataset-name> is a string. The string must be quoted if the name includes spaces. Spaces in the name are discouraged.

The @VARIABLE Declaration

Each variable declaration is on its own line, from the first to the last variable. Variable declarations are hence an ordered sequence of @variable declarations.

Each variable in the dataset must have its own @variable statement that defines the name and the attributes of the variable, including the data type and when relevant, the values.

The declaration of the variables follows the order they appear in the data section, with the first variable corresponding to the first column, the second to the second column, etc. The order of the variables indicates the column position in the data section.

The format for the @variable statement is:

@variable <variable-name> <datatype> where the <variable-name> must start with an alphabetic character. If spaces are to be included in the name then the entire name must be quoted. Spaces in the variable names are discouraged.

The <datatype> can be any of the four types:

- numeric
  - integer may replace “numeric” and is treated as numeric
  - real may replace “numeric” and is treated as numeric
- <nominal-specification>

- string
- date [<date-format>]

The protected keywords numeric, real, integer, string and date are case insensitive.

Numeric attributes can be real or integer numbers.

Nominal values are defined by providing a <nominal-specification> listing the possible values: <nominal-name1>, <nominal-name2>, <nominal-name3>, ...

For example, the class value of the Iris dataset can be defined as follows:

@variable class {Iris-setosa,Iris-versicolor,Iris-virginica}

Values that contain spaces must be quoted.

Nominal values are treated as categorical values and assigned categorical values during analysis. It is possible to have a file that is entirely composed of categorical values, such as in the following case:

```
@dataset aroninominaldata
@variable continent {college, highschool, other}
@variable culture {west, other}
@variable gender {male, female}
@variable income {low, middle, high}
@variable choices {affirm, confirm, reaffirm}

@data
college,west,no,low,affirm
college,west,no,high,reaffirm
college,west,yes,low,affirm
college,west,yes,middle,,confirm
college,other,no,low,affirm
college,other,no,middle,reaffirm
college,other,yes,middle, affirm
college,other,yes,high,confirm
other,west,no,low,affirm
other,west,no,high,reaffirm
```

other,west,yes,middle, affirm
other,west,yes,middle,confirm
other,other,no,low,affirm
other,other,no,high,reaffirm
other,other,yes,low,affirm
other,other,yes,high,affirm
other,other,no,high,affirm
highschool,west,no,low,affirm
highschool,west,no,high,affirm
highschool,west,yes,low,affirm
highschool,west,yes,high,confirm
highschool,other,no,low,affirm
highschool,other,no,high,reaffirm
highschool,other,yes,low,affirm
highschool,other,yes,high,affirm
highschool,other,yes,low,affirm

String variables correspond to arbitrary textual values. String attributes are declared as follows: @VARIABLE <variable-name> string

Date attribute declarations take the form: @VARIABLE <variable-name> date [<date-format>]

where <variable-name> is the name for the variable and <date-format> is an optional string specifying how date values should be parsed and printed

The default format string accepts the ISO-8601 combined date and time format: yyyy-MM-dd'T'HH:mm:ss. Dates must be specified in the data section as the corresponding string representations of the date/time.

2) **Select the analytical module.** Once the data is loaded, choose the right analytical module by clicking the button with the name corresponding to the methodologies. For example, click the button labeled "Statistical Inference" for Parametric and Non-Parametric tests. The following capabilities are available in AroniSmartLytics™. Each module is explained in details in the subsequent dedicated chapters and sections.

   a. Statistics Handbook and Manuals: to get more information on Statistical Concepts and Methods. From the module learn about.

b. Statistical Inference: descriptive statistics, including graphing capabilities, and statistical inference.

c. Segmentation and Mixture Models: for segmenting and clustering observations into groups.

d. Regression Analysis and Econometrics: univariate and multivariate regression models, mixed effects models, and econometrics.

e. Bayesian, Big Data and Text Analytics: for Bayesian models and advanced big data mining and text and sentiment analytics, including network graphing

VI. Module 2: Handbook and Manual about Parametric and Non Parametric Statistics.

Select Handbook and Manuals to get more information on Statistical Concepts and Methods.

A. Statistics Handbook: Reference about the statistical concepts, the most used probability distributions, the relationship among probability distributions, key statistical tests, and how to select a statistical test.

Statistics Handbook module of AroniSmartLytics™ presents the user with a true statistical encyclopedia that covers the following:

1. Statistical concepts: Statistical Method, Parametric and Nonparametric statistics.
2. Most used probability distributions: Discrete and continuous, their parameters, assumptions and properties.
3. Relationship among probability distributions: the link among the key probability distributions.
4. How to choose between parametric and nonparametric tests, which parametric test is appropriate in what situation.
5. Key parametric statistical test, including Z-test, t-test, F-test and Analysis of Variance (ANOVA) for fixed effects with random subjects: One-way, factorial, and nested design.
6. Key non parametric statistical tests
7. How to select an appropriate nonparametric statistical test.

The materials on this module are organized within user friendly outline viewers. In Statistics Handbook tab, each section corresponds to a path terminated by a document icon with a description of each subject in a pdf or html format. Click on the pdf or html file icon to open it in the viewer on the right hand side.

Note: *To get a better read, the pdf document may be adjusted by expanding the left side handles.*

For the probability distributions and the key statistical tests, key parameters, statistics, and equations are described and formulas clearly documented. The user may search the reference document by using **Find menu** item in the menu bar. The navigation through the reference document is also possible using the menu bar. Sometimes, it may be suitable to select pages in the outline on the right or scroll down the main pdf viewer window. Help is available by clicking on the menu bar.

B. Annotation

The main pdf document can be annotated through the Tools menu and saved for further reference. The annotation gives the ability to add notes and references in the document for further enriching the statistical concepts. The annotation function follows the usual pdf document annotation. Html and other formats will not be annotated.

Select Tools from main menu and the annotation tool. By default the annotation will be placed at the bottom left of the current page. Use the Annotation preference document to add features to the annotation including formatting, adding a link, text, color, etc. The annotated document may be saved and printed as needed. The Save As menu allows to create a copy, whereas the Save menu allows to save in the manual itself.

C. Accessing other resources on Internet web pages

AroniSmartLytics™ is flexible and allows the user to create a link to Internet web pages while the program is running. The user may then access the linked source as needed.

To create a web link navigate to the bottom left. Click the button with "+" to create a link and "-" to remove the link. Once the link is created open the Popup menu to the right to edit and add web link. The link must follow the standard: "http://" for web pages. In order to access the sources internet link, AroniSmartLytics™ must be running while there is connection to internet.

D. AroniSmartLytics™ Probability Distributions

As AroniSmartLytics™ module on probability distributions opens, the needed features and functionalities are readily available and should be referred to as often as possible, when exploring parametric statistics.

(1) Graphing Capability

Select a discrete or continuous distribution from the browser, and observe the graph of that distribution. Change the input parameters and see how the distribution changes its form and shape.

(2) Calculate the probabilities

Select a discrete or continuous distribution from the browser, enter the input parameters and calculate the probabilities based on the selected distribution. Change the input parameters and see how the probability distributions change.

(3) Probability Function Statistics

Select a probability function from the browser and find the formulas in the embedded document. It is possible to navigate the pdf documents using the regular Adobe® Acrobat® reader functionalities.

(a) **Discrete**: Bernoulli, Binomial, Discrete Uniform, Geometric, Hypergeometric, Negative Binomial, and Poisson.

(b) **Continuous**: Beta, Cauchy, Chi Squared, Double Exponential, Exponential, Fisher-Snedecor's F, Gamma, Logistic, Lognormal, Normal, Pareto, Student's t, Uniform, and Weibull.

(c) **Probability function**: Calculate probability densities or mass based on input parameters. The input parameters may be changed in the Input Parameters form. Clicking the "Graph" button updates the charts and the probability estimates. The input parameters may be changed in the Input Parameters form. Clicking the "Graph" button updates the charts and the probability estimates.

(d) **Cumulative function:** Calculate cumulative density based on input parameters. The input parameters may be changed in the Input Parameters form. Clicking the "Graph" button updates the charts and the probability estimates. The input parameters may be changed in the Input Parameters form. Clicking the **"Refresh Graphs"** button updates the charts and the probability estimates.

(e) **Confidence Intervals** and **Probability between low and upper values.** By varying the upper and lower values and graphing, new probabilities are estimated. The lower and upper values are changed in the "Variable Interval Values" entry form. Click "Refresh Graphs" button to update the charts and the probability estimates.

(f) Graph Analysis: Graph probability and cumulative density functions based on input parameters. Vary input parameters and observe the changes in graphs. Graph bounds may be modified in the "Graph Bounds" form. Click "Refresh Graphs" button to update charts.

Both the probability and cumulative density functions, in the case of continuous distributions and the probability and cumulative mass functions, for discrete distributions are simultaneously updated.

The graphs shows in red the interval of interest. The area or mass outside the interval is colored in green.

E. Relationships among common distributions and Selecting a Statistical Text

Probability distributions are connected and form one network with interesting relationships. AroniSmartLytics™ provides a graphical map of common distributions and their relationships, either as transformations or special cases. This feature guides statisticians in choosing the right distribution, understand the connections, and simulate distributions.

VII. Module 3: AroniSmartLytics™ Descriptive Statistics and Statistical Inference

By now, if the statistician, researcher, student or practitioner has gone through the material in Module 1 and explored the probability distributions in Module 2, Module 3 and Module 4 are where the rubber meets the road. The user has the data collected for a research and/or study and is ready to descriptively analyze the data, conduct tests and advanced analyses and draw inference. **AroniSmartLytics ™** Module 3 is dedicated to Descriptive Analysis and Statistical Inference. The module has three parts: Selecting Parameters, Plotting and Descriptive statistics.

A. Descriptive Analytics

i) **Graphing Capability:** Once the data is loaded, a Histogram of the fist numerical variable is plotted in the lower left corner and its statistics are given in the top right corner, in the box "Sample Statistics for Ordinal/Continuous Variable". If the data has more than two numerical variables, a scatter plot of the first two variables is displayed in the lower right corner.

ii) To refresh the graphs and the statistics, choose the variable for the Histogram and the descriptive statistics in "Select Variable for Histogram and Descriptive Statistics" combo box or two variables in the two popup menu buttons corresponding to the scattergram in the lower left corner. Click "Refresh Graphs & Statistics" button to refresh.

iii) Repeat the steps for any new data or new analysis.

B. Inference and Statistical Testing

By now, the statistician, researcher, business analyst, data miner, student or practitioner has gone through the material in Module 1, explored the probability distributions in Module 2, and loaded the research data in Module 3, is ready to do heavy lifting by conducting statistical testing. Using the previous modules, especially Module 1, the user knows what statistical test is the right one or maybe a couple of statistical tests is a candidate.

AroniSmartLytics Module 4 is an advanced study and research board. AroniSmartLytics Module 4 is dedicated to Statistics Testing and Inference. The module has three parts:

1) **Load the Variables (Columns) to Analyze.** The data is loaded from Module 3. The list of numerical variables is shown in the browser in the top middle part. There are four browsers:

   f. Main Browser in the middle shows the list of all the variables. Variables are selected from and sent back to it, using the buttons to the right or the left.

      i. **Sample 1** and **Sample 2** buttons are clicked to send selected variables to Sample 1 or Sample 2 browsers.

      ii. Click the middle button with arrow pointing to the main browser for removing a selected variable from Sample1 and Sample 2 browsers and sending it back to the main browser.

      iii. Click the left side button to send to or from "Select Grouping Variables and the Type" browser.

   g. Sample 1 Browser: this is where the variables used in Statistical tests that require one set of variables are loaded. It is also where the first set of variables are loaded for tests that required two sets of variables.

   h. Sample 2 Browser: this is where the second set of variables used in Statistical tests that require two sets of variables are loaded.

   i. Select 2 Grouping Variables and the Type browser: this is where two variables used to create a contingency table, factorial and nested designs or strata variable for Cox regression are loaded.

2) Contingency Table: Before loading the data, the user is encouraged to create the contingency table using other tools, such as Apple's Numbers or Microsoft Excel. AroniSmartLytics™ offers the option to create a contingency table with three types:

   a. Proportion: proportion of cases within each cell;

   b. Frequency: frequency count of cases within each cell;

c. Average: average by cell of the values of the first variable loaded in Sample 1 browser.

To create a contingency table, select the checkbox: Two-Way contingency table in top left corner. Once the checkbox is selected, the contingency table will be used in subsequent analyses. Uncheck the box to use the variables in Sample 1 and Sample 2 browsers.

3) **Factorial, Nested, or Cox Regression:** The grouping or class variables for Factorial and Nested designs and the strata variable for Cox Regression need to be loaded into the appropriate browser: Select Grouping Variables and the Type browser. For these analyses the Type option relevant for Contingency table is ignored.

4) **Output Viewer.** The text viewer that occupies the center is where all the statistics produced in the Module are output. The data shown in the output may include the contingency table and the statistical tests, including covariance, p-values, etc. The Output Viewer is editable and can be cleaned by highlighting and deleting a small or large portion. The users may also add their own comments,

5) Selecting Statistical Problem/Test: The top right corner is dedicated to statistical test selection. A popup menu allows the user to select a statistical test from the extensive list provided. All the tests included on the extensive list provided are implemented in AroniSmartLytics™. By clicking the checkbox: Parametric Statistic Test, the user may toggle between Parametric and Nonparametric tests. Use the popup menu button to select a specific test. Some tests require additional values, such as known mean, median, variance, or proportion. For such tests, the user is required to provide the known value in the appropriate text box in the top right corner.

Repeat the steps for any new data or analysis.

VIII. Module 4: AroniSmartLytics™ Regression Analysis and Econometrics

With Module 3, the statistician, researcher, business analyst, data miner, student or practitioner has gone through all possible parametric and nonparametric tests and performed Statistical Testing and Inference. The tests have confirmed or disproved hypotheses about the population when evaluated on certain attributes.

AroniSmartLytics™ Module 5 is an advanced study and research board that focuses on Regression Analysis:

1. **Regression Analysis** consists in predicting values of a response or dependent or classification variable from a collection of predictor or independent variable values. With the output from the regression analysis, the intelligent analyst is able to assess the effects of predictor variables on the dependent variable and its values. The regression analysis functions are in a dedicated tab.
2. The **Regression Analysis** sub-module is accessed by clicking on Regression Analysis tab. The sub-module has three parts:
    a. **Load the Variables (Columns) to Analyze.** The data is loaded from Module 3, either as a CSV or Aroni format. The list of variables is shown in the browser in the middle left side of the browser. There are three sections.
    b. **Left section** is where the Regression Analysis Model is setup and the variables used in the model are selected.
        i. **Regression Analysis model:** The regression model is selected from the top left corner. Depending on the data format (CSV or Aroni format) various models are presented. Aroni format offers more flexible Regression analysis capabilities, with models that may handle categorical or nominal variables. CSV data formats are restricted to numerical variables with interval scale measurement.
        ii. **Variable selection:** Select one dependent variable and a number of predictor (independent) variables to be used in the regression analysis.
        iii. **In case of Interaction:** Click on interaction check box and then enter the model in the text field to the right of the interaction check box. Check the spelling of the variables and enter the variable "Intercept" if the intercept is needed. Only two-way interactions are allowed.

            Example: Intercept+X1+ X2+X1*X2+X3|X4 is equivalent to Intercept+X1+X2+X1*X2+X3+X4+X3*X4.

iv. **Regression Analysis Model Options selection:** Depending on the model, multiple options are available.

1. **Some Advanced Regression Models Require more options:** Especially, the validation options may need to be set up for more advanced models. The analyst, student, data miner, business analyst or researcher is expected to be familiar with the common model validation techniques, including percent split, cross-validation or use of validation data. Default options are used in case the validation options are not properly set up.

2. **Generalized Linear Models with Spline:** AroniSmartLytics™ supports Generalized Linear Models with Spline. Options of the model are set up in the middle: Family (Binomial, Cox, Gamma, Gaussian, Poisson and Cox) and Link (Identify, Inverse, Logit, Logarithmic, Cox). When the family is Binomial and one of the dependent variables includes the number of successes, the denominator variable that captures the number of counts needs to be selected from the denominator combo box.

   A spline order: linear, quadratic, cubic, and quartic may also be set. By default, a linear spline is used, but works in very special cases. If a spline order fails, select a higher order may be set.

   Generalized Linear Models also accept interaction models.

3. **Hierarchical Mixed Effects and Econometrics Models:** Selecting Mixed Effects or Econometrics Models open popups a controller with various options. Options to control the model are also available and needs to be set, unless the user prefers the defaults. When the model contains interaction terms, the form needs to be entered in the indicated field.

   - **Hierarchical Mixed Effects Models:** AroniSmartLytics™ supports Hierarchical Mixed Effects Models. The selection of Hierarchical Mixed Effects Models

opens up the "Mixed Effects Model Specifications" window, in which the options are set. Depending on the model structure, **fixed effects, random and grouping effects** need to be set.

If only fixed effects are selected, the model becomes a general linear regression model.

Selecting both random and fixed effects make a regular mixed effects model. In order to setup a hierarchical mixed effects or a Linear mixed effects model for repeated measures, grouping variables (level 1 or 2 micro units, or subjects) need to be selected. Selecting multiple grouping variables implies nested units.
Hierarchical Mixed Effects Models accept interaction models.
Fixed Effects and Random Effects interaction models are set in the bottom of the window. Grouping effects still need to be selected in the grouping browser.

Example: Intercept+X1+ X2+X1*X2+X3|X4 is equivalent to Intercept+X1+X2+X1*X2+X3+X4+X3*X4.

- **Econometrics Models.**

Econometrics Models controller contains a series of models used in Econometrics. The econometrics models included are: Count Data (Poisson, Negative Binomial), Logistic Regression, Logit, Probit, Tobit, Interval, Weighted Least Squares, Ordinary Least Squares (OLS). Models. Some Econometrics Models, such as Logistic Regression, and OLS may be run outside Econometrics Models. Each Econometrics models may have its own options that should be set before running the model. For example, choosing Count Data Models requires selecting he distribution: Poisson, Negative Binomial, etc.

Same for Probit where the user needs to choose among the following distributions: Binary, Ordered, and Random Effects, or Logit where the choices are: **Binary, Ordered, and Multinomial.**

j. **Run the model:** Once the model is setup, click "Run Model" button to run the model. For Hierarchical Mixed Effects Models, click Ok button in the Mixed Effects Model Specifications window.

k. **Output Viewer.** The text viewer that occupies the right center is where all the statistics produced in the sub-module are output. The data shown in the output may include ANOVA tables, data, predictions, the statistical tests, including covariance, F-test, AIC/BIC values, p-values, etc. The Output Viewer is editable and can be cleaned by highlighting and deleting a small or large portion. The users may also add their own comments. The controls on the right allow easy navigation.

4. **Providing test and validation data set:** The top middle section is dedicated to data saving and retrieving. Data and output in Output viewer may be saved by clicking on "Save Model in File" path control. A saved model output can be loaded into the Output Viewer by clicking on "Open Saved Model" path control. The path to the Test/Validation data used in validation may be supplied through "Test/Validation Dataset" path control.

   Repeat the steps for any new data or analysis. It is recommended to perform an analysis on one dataset at a time in a given session so as not to get the analysis output mixed.

   Output results are cleaned up at each session.

IX. Module 5: Segmentation and Classification

With Module 4, the statistician, researcher, business analyst, data miner, student or practitioner has gone through all possible parametric and nonparametric tests and performed Statistical Testing and Inference. The tests have confirmed or disproved hypotheses about the population when evaluated on certain attributes.

AroniSmartLytics™ Module 6 is an advanced study and research board that focuses on **Segmentation, Clustering, Discrimination and Classification:**

1. **Segmentation, Clustering, Discrimination and Classification:** Separating distinct sets of observations into defined groups. AroniSmartLytics™ allows to conduct both Discrimination and Classification and Clustering.

The purpose of segmentation, clustering and discrimination analysis is to discover, or explain, group structures in multivariate data sets with unknown (cluster analysis or clustering) or known class (discriminant analysis or classification). This module of AroniSmartLytics™ is an exploratory data analysis tool for solving clustering and classification problems. But it can also be regarded as a semi-parametric tool to estimate densities with Gaussian mixture distributions and multinomial distributions.

With Discrimination and Classification, the groups are previously defined and the intelligent analyst or researcher is investigating observed differences and seeking to understand the causal relationships and ultimately assign new objects to the set of groups identified. Before a set of groups used in classification is known, an exploratory search based on similarity or distances (dissimilarities) measures is performed. AroniSmartLytics ™ provides exploratory techniques, in which no assumptions on the number or the structure of the groups are made. The end result is a set of known groups and their profiles, that may be used in the Discrimination and Classification.

2. The **Segmentation, Discrimination, Classification, and Clustering Models** sub-module is accessed by clicking on **Segmentation and Mixture Models** tab. The sub-module has three parts:

    c. **Load the Variables (Columns) to Analyze.** The data is loaded from Module 3, either as a CSV or Aroni format. The list of variables is shown in the browser in the middle left side of the browser. There are three sections

    d. **Left section** is where the Segmentation Analysis Model is setup and the variables used in the model are selected..

        i. **Segmentation model:** The segmentation model is selected from the bottom left corner. Two options are available: Finite Mixture Gaussian Models and General Mixture Models.

        ii. **Variable selection:** select predictor (independent) variables to be used in the segmentation analysis. For the segmentation models, it is highly recommended to avoid the selection of classification or dependent variables to avoid degeneracy and singularity that may lead to crashes. The Finite Mixture Gaussian Models try to account for singularity and are successful in most instances.

iii. **Setting up Model Options. Click "Model Setup" button to setup the segmentation model.** Depending on the model, multiple options are available. The options may be complex in each case and require a knowledge of segmentation techniques. Those not familiar with these techniques may choose to only select the number of clusters and use the default options. The following chapters will cover these options in detail.

l. **Run the model:** Once the model is setup, click **"Run Model"** button to run the model.

m. **Output Viewer.** The text viewer that occupies the right center is where all the statistics produced in the sub-module are output. The data shown in the output may include analytical results, segment profiles, data, classifications, statistical tests, including covariance, F-test, log-likelihood, AIC/BIC values, p-values, etc. The Output Viewer is editable and can be cleaned by highlighting and deleting a small or large portion. The users may also add their own comments. The controls on the right allow easy navigation.

5. **Graphs:** The bottom middle section is dedicated to graphics of segment profiles. Only the **Finite Mixture Gaussian Models** offers the graphic capabilities, as they produce only one classification model. There are three graphs:

   a. Selected variables averages for all segments.

   b. A scatterplot of two selected variables for a selected segment and

   c. the mixture proportion of a selected segment.

Repeat the steps for any new data or analysis. It is recommended to perform an analysis on one dataset at a time in a given session so as not to get the analysis output mixed.

Output results are cleaned up at each session.

X. Module 6: Finite Mixture Gaussian Models with AroniSmartLytics™

Setting Up the Segmentation Model in AroniSmartLytics

One of the major drivers of AroniSmartLytics™ is an advanced and complete segmentation package. The segmentation models are configured through an intuitive, user friendly interface. There are two segmentation approaches, each with its own setup interface. Both are launched by clicking on "Model Setup" button. Depending on the choice among the following in "Choose Segmentation Model" box, a different panel will be presented:

- by selecting "**Finite Mixture Gaussian Models**" option the panel for finite mixture Gaussian models will appear.
- by selecting: "**General Mixture Model**" option the panel for general mixture models will appear.

Both actions open a panel, populated with defaults values. The user may opt to run the model with defaults values or customize the options before running the model.

This chapter will focus on Finite Mixture Gaussian Models

Setting Up the Finite Mixture Gaussian Models in AroniSmartLytics

The Finite Mixture Gaussian Model panel, has five tabs:

- Tab1: **Setup:** that is where a main segmentation/ clustering/ discriminant analysis model is selected.
- Tab 2: **Starting Methods:** for setting initial options for analysis and segmentation.
- Tab 3: **Bootstrap and error**: for setting bootstrap, error, and simulation parameters.
- Tab 4: **Covariance and Output**: for setting up covariance structure and output options.
- Tab 5: **Extra Options**: for advanced options to fine tune and optimize the model.

Choosing the Main Analysis Model

The user must first choose the main analysis model from "**Setup**" tab. Several options are offered, but only one option can be selected and executed at a time.

The available options shown in "**Choice of Analysis and Setup"** box are:

- **Simulation - Simulate a Sample from Normal Distribution**
- **(Bootstrap) Assessment on Number of Segments**
- **Fit a fixed number of segments - g-segment or component Normal Mixture Model**
- **Fit a range of segments - g-segment or components of Normal Mixture Model**
- **Perform Discriminant Analysis**
- **Prediction of New Data**
- **Parameter estimate from Data and Allocation**

Data consists of entities in rows and variables in columns. One variable, named Identifier or observation uniquely identifies the entity. Other numerical variables may be used in the model.

By default the user is given the option "Fit a fixed number of segments" which corresponds to fitting fixed g-component Normal Mixture Model.

In the same tab, within the "**Entities, Variables & Groups**" box, the system automatically displays the number of entities and corresponding numeric variables used in the analysis. The user may also select the **Fixed or range of segments** to consider and the **percentage of data** used in the model.

The user has the option to use defaults values, which are:

- Number of segments: Fixed, 3 segments
- Percentage of data: 67%

Depending on the choice of analysis selected in this step, the path for further options selection varies. The following sections describe the available options for each analysis selection. Some choices may involve restrictions on the size of the data points (entities) and the variables.

The outputs are displayed in Output viewer and may be used in further analyses, in plotting tools, or exported into other applications for further analysis.

Input Data and Data Presentation

The segmentation is based on a dataset as entered by the user. For most of the analysis options, the input file mainly contains the complete data set to be analyzed. For others, additional datasets containing parameters may be required. In general, depending on which options are utilized when

running the program, extra information may be required as input in the datasets or through the user interface

The dataset consists of sets experimental unit **individuals entities or observations** described by several variables. Individuals are represented in a standard way: a row vector for the data entered by the user and the row from the database table. Hence, the data sets are represented by a matrix with ***m*** rows and ***n*** columns, the rows representing observations or entities, and the columns representing variables.

For the ad-hoc modeling data in CSV format entered by the user, the variables must be separated by a comma or semi-column and terminated by an end-of-line character. The first row must be a list of variables, each per column. Otherwise the system add default variable names. The variables names must also be separated by a comma or semi-column and terminated by an end-of-line character. One of the variables, preferably the first column, must have the name “observation” and contain unique identifiers for the entities or observations. Variables used in segmentation must be numerical.

The Aroni format is described in Chapter VII.

Example of dataset entered by the user

For a sample consisting of 4 entities each with 3 variables

Example 10.1

Observation, high, low, size, close

ABC1, 65.3, 45.6, 2765.7,1.542

ABC2, 68.5, 76.8, 3687.6 ,1.345

ABC3, 67.3, 56.7, 7986.0,932

ABC4, 43.6, 43.1, 6235.32, 2.012

ABC5, 32.0, 42.3, 9741,1.034

Simulating a Sample from a Normal Distribution

To simulate a sample from a Normal distribution, the user must provide an input file containing the following elements, depending on how many segments to generate, as selected in Fixed option:

Row containing the initial means for Segment/Component 1

Lower diagonal form of initial covariance for Segment/Component 1

Row containing the initial means for Segment/Component 2

Lower diagonal form of initial covariance matrix for Segment/Component 2

Row containing the mean for Segment/Component 3

Lower diagonal form of initial covariance matrix for Segment/Component 3

etc.

Row of the mixing proportions of Segment1 Segment 2 Segment 3, etc.

For example, below is a input parameters file layout for 3 segments/components and two segmentation variables:

25.2 12.33

1.96
0 0.83

0.52 67.3

.9
.1 .9

2.09 54.7

17.89
5.9 19.28

0.15 0.50 0.35

Example 10.2 This example would give the following initial parameters:

Means: $\mu_1 \approx \begin{bmatrix} 25.2 \\ 12.33 \end{bmatrix} \mu 2 \approx \begin{bmatrix} 0.52 \\ 67.3 \end{bmatrix} \mu 3 \approx \begin{bmatrix} 2.09 \\ 54.7 \end{bmatrix}$

Covariance: $\Sigma_1 \approx \begin{bmatrix} 1.96 & 0 \\ 0.0 & 0.83 \end{bmatrix} \Sigma_2 \approx \begin{bmatrix} 0.9 & 0 \\ 0.1 & 0.9 \end{bmatrix} \Sigma_3 \approx \begin{bmatrix} 17.89 & 5.9 \\ 5.9 & 19.28 \end{bmatrix}$

And the mixing proportions $\pi_1 \approx 0.15 \; \pi_2 \approx 0.50 \; \pi_3 \approx 0.35$

Note

The number of segments to model is selected through the user interface while setting up the model. Select Fixed and then the number of segments. The **input parameters file** must be entered from the "**Covariance & Output Options**" tab. The input file must adhere to the layout described in the examples above.

Input file

Output file

Parameters to set in User Interface

Covariance Structure

Bootstrap Assessment of Number of Components

Sometimes the user may not have a good idea on how many segments to fit to the data. One approach to assess the number is to conduct the likelihood ratio test, by using the Log-Likelihood statistic -2log(λ) and to utilize a bootstrap procedure to estimate it's corresponding P-value McLachlan (1987).

In that case, selecting the analysis "Bootstrap Assessment of Number of Components achieves the results. A resampling approach is used in this case to assess the null distribution, hence the P-value) of the log likelihood ratio test (−2 log λ) to test the number of segments H_0: $g = g_0$ vs H_1: $g = g_0+1$.

The input parameters file should contain the parameters under the null hypothesis for the original sample ONLY. The format of the input parameters file is the same as specified in the Example 10.1.

The output file contains the sorted values of -2 log λ and their corresponding likelihood under the null and composite hypotheses.

The user must set the following parameters in the user interface:

- Number of segments to test: enter the number of Fixed components in Setup tab
- Percent of data to be used: enter in Setup tab
- Calculating standard errors: From "**Bootstrap and Standard Error**" tab select either "**Bootstrap(Must Set K-Means and Random Starts**" button in Bootstrap box or "**Standard Error/Simple Analysis in Standard Error Analysis**" button, or both buttons. Selecting "**Standard Error/Simple Analysis in Standard Error Analysis**" requires other options in Standard Error Analysis box:
  Method of estimation: Select a method in Estimation Method radio boxes

  Number of replications: select or enter number of replications

  Enter or select the number of bootstrap replications required in the same tab

  - Enter or select the number of K-means starts
  - Enter or select the number of random starts
  - Covariance structure: select covariance structure in "**Covariance & Output Options**" tab.

The user may also choose to select extra options by selecting "**Modify Extra Options (Except known allocation)"** button in "**Covariance & Output Options**" tab. The extra options are then setup in "Extra Options" tab.

Note: Extra Options

Setting up extra options requires advanced statistical expertise. Only a brief description of each option will be provided in this manual. The user is encouraged to refer to McLachlan (1987), McLachlan and Peel (2000) and other advanced statistical manuals.

Choosing an appropriate covariance structure is critical. The user may constrain the covariance matrices to be either equal for all components, arbitrary, or diagonal (equal or unequal). Generally unless the user has some prior knowledge of the covariance structure, arbitrary covariance should be used. This is the default option.

AroniSmartLytics™ is designed to handle some ill formed covariance structures. However, the user may try equal covariance in the case no solution can be found due to singular covariance matrices. Singularity is usually caused by any of the following situations:

Input file

1. Two or more of the variables are highly correlated

2. There are too many variables and not enough points

3. One of the variables is discrete and a cluster is being fitted to a single point of high

density

Fit a Fixed g-component Model Normal Mixture Model

Sometimes the user has a good idea on how many segments (g) to fit to the data. This option corresponds to this situation. The user must specify the known number of components (g) in Setup tab, by selecting "Fixed Groups" button and selecting or entering the number of groups.

Initial Posterior Probabilities

This option is the default with g=3.

For the default stock segmentation, the data set comes from the database. Otherwise the user must supply input file containing the data. Furthermore depending on the options, the user must in either case, supply an "**input parameters file**" as described in example 4.1, plus any other information appended at the end of the file depending on what options are chosen. These options are selected in "**Starting Methods**" tab.

Following are the potential options:

The user determines the initial classification of the entities by selecting "**Outright Initial Classification**" radio button. Hence this classification must be appended at the end of the input parameters file in the layout:

0 3 3 3

1 1 1 0

Automatic Initial Grouping

This example would give the starting partition with the first entities belonging to component 0, the next 3 to 3, the next 3 to 1 , the last entity to segment 0

The user determines the initial parameters for the model by selecting "**Initial Parameters Estimates**" radio button. This option starts the segmentation model, specifically the EM algorithm, from the specified initial values of the unknown mixture model parameters, ie. the elements of the component means and the covariance matrices. When this option is selected, the input parameters file must be in the format given in Example 4.1.

The user determines the initial parameters for the model by selecting "**Initial Probability Grouping**" radio button. This option starts the segmentation model, specifically the EM algorithm, from the specified initial values of the classification probabilities. This is a general case of specified initial classification, where for initial classification, the posterior probabilities are 0 or 1.

Example 10.4 The user defined posterior probabilities (or weights) must be appended to the end of the input file. Each entity has a row (line) of probabilities assigned to each segment.

Specified initial classification

For example, if three segments are assumed for 4 entities, then the probabilities may be in the format:

0.2 0.6 0.2

0.1 0.3 0.6

07 0.2 0.1

0.3 0.3 0.4

With this option, which is the default, the user does not make any assumption on the groupings and hence does not supply any the initial parameters. However, the user needs to select which approach the model will start from. The model then starts from the chosen approach to perform the segmentation. The option is chosen by selecting "**Automatic Initial Grouping**" radio button and various clustering techniques available in the box "Data: Unstandardized<<Both>>Standardized.

Example 10.3

The various clustering methods available in AroniSmartLytics are:

- Hierarchical clustering (on standardized and unstandardized data):
    - Nearest Neighbor (Single Linkage)

- Furthest Neighbor (Complete Linkage)
- Group Average (Average Linkage)
  - Median
  - Centroid
- Flexible Sorting
- Incremental Sum of Squares (Ward's Method)

- Random partitions of the data
- K-means clustering algorithm

Standardizing data

Specified initial Parameters

The choice of these methods is controlled in two ways. The random and k-means clustering methods require the number of replications. Random starts method also requires the percentage of data to be used, otherwise the percentage will be set to 100%.

The user may select the option of automatically standardizing data before segmentation. **Standardization** is selected by using the slider before each method: To the left means unstandardized data, to the right means Standardized, and in the center means both standardized and unstandardized data.

If the user chooses flexible sorting, then an extra parameter BETA is required and must be entered in the text field: Beta-Flexible Sorting.

Explanation of BETA values is given in the text entry below the text field: BETA equal to zero corresponds to the Furthest Neighbor method, as BETA tends to 1 the method generally produces long shaped clusters, and for BETA smaller than zero the method produces small compact clusters.

Fit a Range of g-component Model Normal Mixture Model

In general, the user does not have a good idea on how many segments to fit to the data or just wants to test a range of segments. Selecting this option accomplishes the purposes.

The user starts with a reasonable range with the assumption that the optimal number of the segments is within the range. Hence, the user must specify a range for the number of segments from "**Setup**" tab by selecting "**Select Range"** radio button and setting the minimum and the maximum.

BETA and Clusters shapes

For this specified range, the program fits the segmentation model for each value of g, in turn, in the specified range. Finally, various test statistics are applied to evaluate the final statistics.

There is no need for the input parameters file.

The default output file contains the fits obtained sequentially for the range specified, plus a summary of the fits.

The user must set the following parameters in the user interface:

- Search Range of segments to test: enter the minimum and maximum number of components in Setup tab
- Enter or select the number of K-means starts
- Enter or select the number of random starts
- Percent of data to be used: enter in Setup tab
- Covariance structure: select covariance structure in "**Covariance & Output Options**" tab

Note: Extra Options

Parameters to set in User Interface

- If the users would like to assess the number of segments with the likelihood ratio test, by using the Log-Likelihood statistic -2log(λ) and utilizing a bootstrap procedure to estimate it's corresponding P-value (McLachlan (1987)), then the "**Bootstrap (must set K-means and Random Starts)**" button must be selected and the number of bootstrap replications set in "**Bootstrap & Standard Error Analysis**" tab

Example 10.5

- If the users wants to stop the analysis when the p-value of Log-Likelihood statistic -2log(λ) reaches a fixed threshold, i.e. becomes significant, then the "**Set Stopping Rule for p-value**" button must be checked and the "**Level of Significance"** set in "**Bootstrap & Standard Error Analysis**" tab. For the level of significance, enter integer values: 15 = 15%

Discriminant Analysis

The uses provides the initial classification of data in the input parameters file. For example, the input parameters file may be:

```
1       0
2       1
5       0
14      2
52      2
-1      -1
```

with the first column being the entity number in the input file and the second column giving the entity known segment. Two -1 (negative 1) are appended at the end of the file in the last row.

The user must supply the name of the output file in the "**Covariance & Output Options**" tab.

The user must set the following parameters in the user interface:

- How many segments to fit, by checking "**Fixed Groups**" radio button and setting the number in "**Setup**" tab
- Covariance structure: select covariance structure in "**Covariance & Output Options**" tab

Prediction of New Data

This option assumes that the user has estimated the parameters from an existing model and is only interested in sample validation or classification. Hence, based on these model parameters, this option predicts the posterior probabilities and allocation for a new sample:

The user must supply an "**input parameters file**" in the format given in example 10.2.

The user must supply the name of the output file in the "**Covariance & Output Options**" tab.

Advanced Segmentation Options in AroniSmartLytics™

AroniSmartLytics™ is a powerful, yet intuitive segmentation tool. It is also flexible , allowing a variety of users from beginners to advanced, to more sophisticated. Beginners may accomplish most of their requirements with common options. Advanced users, with statistical expertise may want to fine tune the model by setting up options.

These options are grouped in "**Extra Options**" tab. To access them, the user must first check the "**Modify Extra Options (Except Known Allocation)**" in "**Covariance & Output Options**" tab. Then the user needs to accessing the "**Extra Options**" tab, where the options are presented. To activate the options, check the "radio" button associated with the option of

interest. Some options have sub-options. These are indicated by a **>>** at the end of the option name, which points to the sub-options menu.

Following are the Extra Options (McLachlan, 2000, chapter 2)

- Stochastic EM option
- Modify EM stopping criteria
- Space efficiency
- Add extra output files
- Partial classification
- Estimate standard error
- Bootstrap test
- Display discriminant density values
- Change component distribution

The Stochastic EM is an extension of the EM algorithm which may be specified. The basic principle of the Stochastic EM is similar in spirit to simulated annealing, in that randomness is added to the iterative process to give the algorithm a chance to escape local maxima especially in the late stage of convergence.

EM Stopping Criteria: The purpose of the option is to adjust the stopping criteria by examining the changes in the log likelihood from the previous iteration to the current iteration. If this change, whether in improvement or deterioration, differs by less than a specified tolerance multiplied by the current log likelihood then the algorithm stops and assumed to have converged. The convergence must happen within a predetermined number of steps, otherwise the system stops and warns of no convergence.

The option provides for sub options, that allow the user to set the tolerance and the maximum number of steps, both for the initialization and the final convergence. These parameters can be set in "**Modify EM Stopping Criteria**" box. By default the values are set to: 1.0E-06 for tolerance and 500 steps.

Stochastic EM Algorithm

The system is designed to easily handle moderate to large data sets. However, in some instances, users may try to model extremely large data sets, leading to inefficiencies and potential crashes. In these particular instances, the user may only focus on needed data and apply space saving and efficiency techniques. Three space efficiency levels are implemented and may be selected in "**Space Efficiency**" box:

- None
- Moderate
- Extreme

This option creates user friendly output files used for plotting some of the results. Arbitrary names are given. The files include the entity allocation and the bootstrap distributions.

In some instances, the user has information regarding the fixed classification of some entities and does not intend to change the classification. With this option, the user fixes the classification. The classification is included in the "**input parameters file**", the format given in example 10.1. The specified entities retain their segments throughout the fitting process.

The standard errors of the estimates may be calculated in any model by choosing this option and setting the options for Standard Error Analysis in " Bootstrap and Standard Analysis" tab.

One of the major causes of model failure or abnormal results in fitting normal mixtures models is the presence of outliers. To mitigate the impact of the outliers or small samples, multivariate t-distributions may be more appropriate.

The t-distributions option has two sub options: the estimation method and the degree of freedom that must be set for each segment. The available estimation methods are:

- Fixed user-defined degrees of freedom v for each segment
- Degrees of freedom v estimated for each segment
  - The user must supply the initial value of v
- Common degrees of freedom v estimated for the segments
  - The user must supply initial common value v
- Degrees of freedom v estimated for each segment
  - The user must supply initial moments estimates

This option increases the efficiency and the convergence speed when Bootstrap options are selected.

Final considerations on extra-options for Finite Mixture Models

The user is encouraged to read McLachlan (1997) and McLachlan and Peele (1998) to learn the advantages of each of the extra-options. Other statistical manuals, especially those explaining EM algorithm, bootstrapping and clustering should be used for users interested in fine tuning the model.

XI. Module 6: General Mixed Models with AroniSmartLytics™

Introduction

General Mixture Models sub-module fits either a clustering, a density estimation or a discriminant analysis mixture models of multivariate Gaussian or multinomial components to a given dataset. Because of the complexity of General Mixture Models, it is important to give a brief explanation on the approach in detail.

As with **Finite Mixture Gaussian Model** the purpose is to discover, or explain, group structures in multivariate data sets with unknown (cluster analysis or clustering) or known class (discriminant analysis or classification). This sub-module is an exploratory data analysis tool for solving clustering and classification problems. But it can also be regarded as a semi-parametric tool to estimate densities with Gaussian mixture distributions and multinomial distributions.

Unlike Finite Mixture Gaussian Models that rely on initial estimates and the maximum likelihood via the EM (*Expectation Maximization*, Dempster et al. 1977), General Mixture models uses a chained approach.

Estimation of the mixture parameters is performed either through maximum likelihood via the EM (*Expectation Maximization*, Dempster et al. 1977), the SEM (*Stochastic EM*, Celeux and Diebolt 1985) algorithm or through classification maximum likelihood via the CEM algorithm (*Clustering EM* , Celeux and Govaert 1992). These three algorithms can be chained to obtain original fitting strategies (e.g. CEM then EM with results of CEM) to use advantages of each of them in the estimation process.

Mixture problems usually have multiple relative maxima. AroniSmartLytics will produce different results, depending on the initial estimates supplied by the user. If the user does not input own initial estimates, some initial estimates procedures are proposed (random centers for instance). Finite Mixture Models use several initial estimation methods, such as random starts, K-means, etc.

It is possible to constrain some input parameters. For example, dispersions or covariance may be assumed to be equal among classes, etc. This is similar to setting the covariance structure for Finite Mixture Models.

In the Gaussian case, twenty two models are implemented. Among them, fourteen models, based on the eigenvalue decomposition, are most generally used. They depend on constraints on the variance matrix such as same variance matrix between clusters, spherical variance matrix ... and they are suitable for data sets in any dimension.

The eight remaining Gaussian models have to be used when the dimension of the data set is high and only in a discriminant analysis situation.

In the qualitative (multinomial)case, five multinomial models are available. They are based on a reparametrization of the multinomial probabilities.

In both cases, the models and the number of clusters can be chosen by different criteria:

- BIC (Bayesian Information Criterion),
- ICL (Integrated Completed Likelihood, a classification version of BIC),
- NEC (Entropy Criterion),
- Cross-Validation (CV) or,
- Double Cross-Validation (DCV).

Setting up the General Mixture Model in AroniSmartLytics

The General Mixture Model panel, shown in the Figure 4.1, has five tabs:

- Tab1: **Setup:** that is where a main segmentation/ clustering/ discriminant analysis model is selected.
- Tab 2: **Group/Strategies:** for setting initial options for analysis and segmentation and starting strategies.
- Tab 3: **Algorithm Options**: for selecting expectation or maximization algorithms and setting their parameters.
- Tab 4: **Criterion/Model Selection**: for setting up the final model selection criteria and statistics.
- Tab 5: **Output/Help**: a quick reference on the models, strategies, algorithms, stopping criteria and potential outputs.

Choosing the Main Analysis Model

The user must first choose the main analysis model from "**Setup**" tab. Two options are offered, but only one option can be selected and executed at a time.

The available options shown in "Choice of Analysis and Setup" box are:

- **Cluster Analysis**
- **Discriminant Analysis**

By default the user is given the option "Cluster Analysis" which corresponds to fitting a g-component Mixture Model in Finite Mixture Models.

In the same tab, to the top left corner, the system automatically displays the number of entities and corresponding numeric variables contained in the data used in the analysis.

Depending on the choice of analysis selected in this step, the path for further options selection varies. The following sections describe the available options for each analysis selection. Some choices may involve restrictions on the size of the data points (entities) and the variables. The user may supply additional input data: in weight file, in case of weighted data and training data and partition data used in Discriminant Analysis.

A. Weight and training and partition data format.

Weight file contains weights stored in a vector of real or integer numbers, with n (number of individuals, entities or observations) rows. A weight represent the repetition of individuals, entities or observations. By default, a weight of 1 is assigned to each observation.

Partition data is used in case the user supplies the initialization partition (USER_PARTITION). A partition gives a classification of the observation into a specific unique group or mixture component.

The partition is a matrix of 0 and 1 with n rows and k columns, each row corresponding to an observation, entity or individual and each column indicating the group membership (0 if the individual does not belong to the group/component/segment associated to this column and 1 otherwise).

Some observations, entities, or individuals can have no group assignment. Such observations are represented by a row of 0.

The data is organized as follows:

0 1 0 [Observation 1 in segment/component 2]

1 0 0 [Observation 2 in segment/component 1]

0 1 0 [Observation 3 in segment/component 2]

0 0 1 [Observation 4 in segment/component 3]

...

0 0 1 [Observation n-3 in segment/component 3]

0 0 1 [Observation n-2 in segment/component 3]

1 0 0 [Observation n-1 in segment/component 1]

0 2 1 [Observation n in segment/component 2]

B. Initialization data format.

Initialization data is supplied in the case of USER initialization type. There are three types of initialization files:

1. Gaussian

It gives the values of the unknown mixture model parameters in the quantitative case:

0.67333 [Initial proportion of component/Segment 1]

29.44 45.220000 [Initial mean of component/Segment 1]

28.710889 0.000000 [Initial covariance matrix of component 1]

0.000000 89.288900

0.33667 [Initial proportion component/Segment 2]

97.9322 8.824777 [Initial mean of component/Segment 2]

8.3666 0.000000 [Initial covariance matrix of component/Segment 2]

0.000000 8.3666

2. Multinomial

The following example gives the values of the unknown mixture model parameters in the qualitative case (with modalities: [2;3;4;5]):

0.2 [Initial proportion of component/Segment 1]

1 2 3 4 [Initial mean of component/Segment 1]

0.1 0.1 [Initial covariance matrix of component/Segment 1]

0.1 0.2 0.1

0.1 0.1 0.3 0.1

0.1 0.1 0.1 0.4 0.1

0.8 [Initial proportion of component/Segment 2]

2 3 4 5 [Initial mean of component/Segment 2]

0.5 0.5 [Initial covariance matrix of component/Segment 2]

0.25 0.25 0.5

0.1667 0.1667 0.1667 0.5

0.125 0.125 0.125 0.125 0.5

3. Gaussian High Dimensional

The following example gives the values of the unknown mixture parameters in the gaussian high dimensional (HD) case :

0.**75 [Initial proportion of component/Segment 1]**

14.842 11.718 32.014 36.81 13.35 [Initial mean of component/Segment 1]

3 [Initial subDimension of component/Segment 1]

144.744604 0.214614 0.101925 [Initial parameter Akj of component/Segment 1]

0.063887 [Initial parameter Bk of component/Segment 1]

-0.262423 0.355093 0.004478 [Initial parameter Bk of orientation array of component/Segment 1]

-0.170051 -0.888586 -0.207320

-0.601104 0.057918 0.532327

-0.687179 -0.071419 -0.105660

-0.262061 0.275444 -0.813918

0.25 [Initial proportion of component 2]

13.27 12.138 28.102 32.624 11.816 [Initial mean of component/Segment 2]

3 [Initial subDimension of component/Segment 2]

99.333875 0.155693 0.138530 [Initial parameter Akj of component/Segment 2]

0.049261 [Initial parameter Bk of component/Segment 2]

-0.259855 0.127732 -0.473221 **[Initial parameter Bk oforientation array of component/Segment 2]**

-0.239538 0.555917 0.750006

-0.587639 -0.078075 -0.049268

-0.675526 0.144512 -0.205855

-0.271003 -0.804774 0.410791

C. General Mixed Model inputs

1. Required inputs

General Mixed Models requires two inputs :

data : the data entered in module 3, in CSV or Aroni format

number of Segments of Clusters : a vector of 3 integers representing the number of clusters to be tested.

2. Optional inputs

Optional inputs depend on the model selected.

Criterion: This option permits to select the criterion giving the best configuration of an execution (model, number of cluster and strategy) :

- BIC : Bayesian Information Criterion;
- ICL : Integrated Completed Likelihood;
- NEC : Entropy Criterion;
- CV : Cross-Validation;
- DCV : Double Cross-Validation

Default value is 'BIC'.

Model : Specifying a model different from the default one is possible when enough information on the data is available (for example there is the same number of individuals in each class).

Default value is 'Binary pk Ekjh' for qualitative or 'Gaussian pk Lk C' for quantitative data

When Gaussian HD models are choosen, subDimensionFree and/or subDimensionEqual parameters are required.

The 28 Gaussian Models (Covariance Structure).	
Category	Model
Spherical	Gaussian_p_L_I Gaussian_p_Lk_I Gaussian pk L I Gaussian pk Lk I
Diagonal	Gaussian p L B Gaussian p Lk B Gaussian p L Bk Gaussian p Lk Bk Gaussian pk L B Gaussian pk Lk B Gaussian pk L Bk Gaussian pk Lk Bk
General	Gaussian p L C Gaussian p Lk C Gaussian p L D Ak D Gaussian pk L C Gaussian pk Lk C Gaussian pk L D Ak D Gaussian pk Lk D Ak D Gaussian pk L Dk A Dk Gaussian pk Lk Dk A Dk Gaussian pk L Ck Gaussian pk Lk Ck Gaussian p Lk D Ak D Gaussian p L Dk A Dk Gaussian p Lk Dk A Dk Gaussian p L Ck Gaussian p Lk Ck

The 10 Binary Models.
Binary p E Binary p Ej Binary p Ek Binary p Ekj Binary p Ekjh

Binary pk E Binary pk Ej Binary pk Ek Binary pk Ekj Binary pk Ekjh

The 16 High Dimensional (HD) Models.
Gaussian HD p Akj Bk Qk Dk Gaussian HD p Ak Bk Qk Dk Gaussian HD p Akj Bk Qk D Gaussian HD p Akj B Qk D Gaussian HD p Ak Bk Qk D Gaussian HD p Ak B Qk D Gaussian HD p Aj Bk Qk D Gaussian HD p Aj B Qk D Gaussian HD pk Akj Bk Qk Dk Gaussian HD pk Ak Bk Qk Dk Gaussian HD pk Akj Bk Qk D Gaussian HD pk Akj B Qk D Gaussian HD pk Ak Bk Qk D Gaussian HD pk Ak B Qk D Gaussian HD pk Aj Bk Qk D Gaussian HD pk Aj B Qk D

Weight: A weight file is required for weighted observations. Each observation corresponds to a weight

Default value is : 1.

Partition: A partition file is required when the partition of the data is already known and the user wants to input the partition in the modeling.

Default value is no known Partition.

3. Strategy and Algorithms

Strategy: Strategy provides various options to execute the algorithms

1) Strategy initialization

There are different ways to initialize an algorithm:

– RANDOM: Initialization from a random position. This is the default way to initialize an algorithm. This random initial position is obtained by choosing, at random, centers in the data set. This simple strategy is repeated 5 times by default or as many times as the user can chooses, from different random positions. The position that maximizes the likelihood is selected.

– USER PARTITION: This option initializes the strategy from a specified classification (full or partial) of the observations. This option provides the possibility to use AroniSmartLytics for Discriminant Analysis and in this case, the partition must be full.

– USER: With this option, the user specifies initial values of the unknown mixture model parameters, i.e. the mixing proportions and the parameters of the distribution.

– SMALL EM: A maximum of 50 iterations of the EM algorithm according to the process: n numbers of iterations of EM are done (with random initialization) until the SMALL EM stop criterion value has been reached. The runs are repeated until the number of iterations reaches 50 (or if in one run 50 iterations are reached before the stop criterion value). Usually repeating runs of EM generally leads to more accurate results; one or very few runs of EM can often lead to suboptimal solutions.

– CEM INIT: 10 repetitions of 50 iterations of the CEM algorithm are done. One advantage of initializing an algorithm with CEM lies in the fact that CEM converges generally in a small number of iterations. Thus, without consuming a large amount of CPU times, several runs of CEM are performed. Then EM is run with the best solution among the 10 repetitions.

– SEM MAX: A run of 500 iterations of SEM. The idea is that an SEM sequence is expected to enter rapidly in the neighborhood of the global maximum of the likelihood function.

Default value is RANDOM.

2) Algorithm: defining the algorithms used in the strategy, the stopping rule and when to stop.

a. Algorithms:

- EM : Expectation Maximization
- CEM : Classification EM
- SEM : Stochastic EM
- MAP : Maximum a Posteriori,
- M : Only M (maximization) step .

b. Stopping rules for the algorithm:

- NBITERATION : Sets the maximum number of iterations,
- EPSILON : Sets relative increase of the log-likelihood criterion.
- NBITERATION EPSILON : Sets the maximum number of iterations and the epsilon value.

Default values are 200 NBITERATION of EM with an EPSILON value of 10e-4.

Summary of Model options (blue – first row is the default				
Criterion	***Initialization***	***Algorithm***	***Stopping Rules***	***Models: Quantitatif / Qualitatif***
BIC	RANDOM	EM	NBITERATION EPSILON	'Gaussian pk Lk Ck'/ 'Binary pk Ekjh'
CV	USER	CEM	EPSILON	
ICL	USER PARTITION	SEM	NBITERATION	
NEC	SMALL EM	MAP		
DCV	CEM INIT	M		
	SEM MAX			

Note 1: with HD models, only USER PARTITION + M and USER + MAP (the two steps of Discriminant Analysis) are available.

Note 2: with CEM algorithm, EPSILON = 0 is allowed.

Note 3: with SEM algorithm, only NBITERATION is allowed.

Setting Up the Model

Cluster analysis

Select *Cluster Analysis* from *Setup* Menu. Other tabs will be accessible for setting the required inputs for the analysis.

• the number of variables and the number of observations are already given, based on the data set opened in Descriptive Analysis module

• The number of clusters/segments/components to analyze can be entered in Groups and Strategies tab. One or several values, up to three can be entered. When multiple values are entered, AroniSmartLytics selects the best number of clusters/segments based on the selected options.

Key options for Cluster Analysis include:

• ***Model :***

- *Gaussian Model* (Gaussian pk Lk C by default),
- *Qualitative Model* (Binary pk Ekjh by default),

Note: Gaussian (High Dimension) HD models are not available in Cluster Analysis.

• *Criteria* (BIC by default),

• *Strategy* (RANDOM initialization and 100 iterations of EM by default),

• *Weight for data* (1 by default),

• *Partition* (none by default, if given, it won't change during the execution).

The run the Segmentation, click OK and follow instructions, and then click on Run the Model button from the main AroniSmartLytics™ window.

After running the program, a selection of outputs to be displayed is proposed in the Output window in the top left corner. Several outputs are listed. The selected output displays in the Output viewer, in the middle of the window. The last selected output is displayed at the bottom. Use the buttons on the left to scroll down and up or clear the Output Viewer.

The outputs include key statistics, diagnostic statistics, and error. Key statistics include the partition, classification, variable averages per segment and covariance matrices

Discriminant analysis

To perform Discriminant Analysis, the user must supply classified observations (training observations). AroniSmartLytics™ will then classify optional observations (remaining or test observations). The optional observations must be given in another separate file with the same number of columns.

Observations in the training file are classified in K groups, with each observation belonging to one known group among the K groups. The group information is then used to design a classification rule that assigns any set of observation to one of the K groups.

To run Discriminant analysis, select *Discriminant Analysis* in from *Setup* tab of General Mixture Models.

Most of the input required to run Discriminant Analysis are already given, as default or from the data set such as the number of variables and observations that read from the data and displayed in Number of Variables and Number of Entities text fields.

- Provide the number of clusters to test in Groups/Strategies tab,
- the array of modalities in qualitative case
- selecting a *full* partition file of training observations from the setup tab
- The training dataset must be already loaded
- Select the reclassification rule among the following:

- *MAP: When MAP is selected the following output is produced:*
    - Reclassifying an array of samples displays an matrix of dimensions K by K, for K groups. Each value (i, j) represents the percentage of observations in group i classified in the group j after the M step.
    - List of samples misclassified displays the list of observations misclassified by MAP.
- *CV or DCV information : This produces the following output:*

  – CV : Reclassifying an array of observations by Cross Validation rule and list of observation misclassified.

– DCV : Double Cross Validation rate.

• *Model Parameter :* displays the parameters of the best model.

For Gaussian HD models: proportions, means, subDimension, parameters Akj, parameters Bk and orientation.

For the other models: parameters, means and dispersion.

AroniSmartLytics™ runs the second step of discriminant analysis (MAP step with classification rule computed in the first step). The aim of this step is to assign remaining observations to one of the groups. At the end of the analysis, an output can be displayed to see numerical or graphical results.

XII. Module 7: Bayesian Models, Big Data, and Text Mining with AroniSmartLytics™

Big Data, Text Mining , Unstructured Data and Bayesian Models

One of the major reasons behind the uniqueness of AroniSmartLytics™ is a set of advanced cutting-edge tools. Bayesian Network Models and unstructured data mining capabilities constitute some of the latest advances in data mining and statistical models included in AroniSmartLytics™. These capabilities were driven by the increasing importance of BigData and unstructured data from various sources.

Big Data, Unstructured Data and Bayesian Network Models

There are several authoritative sources on the definition, origins, current status, challenges, and future of Big Data. Here, the definition will be limited to the general understanding.

Big Data has entered into the public lingo to describe the unprecedented exponential growth and availability of data, both structured and unstructured, especially with the development of internet and electronic social media, portals, groups, forums, and

communities. This situation has led to a fundamental change in how businesses and the society store, use, and conduct analytics. The implied consequence and the ultimate result are that more, unstructured but timely data may lead to better and more accurate analytics and actionable insights.

When Big Data was becoming a new trend, in 2001, industry analyst Doug Laney reasoned that Big Data must meet the three Vs conditions: volume, velocity and variety.

- **Volume:** all the transactions, communications, e-mails, searches, tweets, RSS, news, texts, feeds, etc generate massive data, mostly unstructured.
- **Velocity:** data is being generated at unprecedented speed. This requires an even faster reaction time if the society, businesses, and individual need to continuously harness the value from the data and leverage insights generated.
- **Variety:** the data generated by billions and billions of sources today comes in all types of formats. The traditional databases used to contain structured and numeric data. Today, this structured information is a tiny portion of available data. Data and information created today are mostly composed of unstructured text documents, RSS feeds, email, video, audio, stock ticker data, financial transactions, ams, etc. Hence, the tools and resources to gather, manage, merge, structure, update, and maintain these structured and unstructured varieties of data have become the focus of several businesses.

However, data without insights remains useless and a waste of effort. Decision makers do not need data to make strategic or tactical decisions. They need information and insight. Hence, the tools to analyze structured and unstructured data become very important. The Text Mining capability in Descriptive Analysis module and Bayesian Network Models implemented in AroniSmartLytics ™ are dedicated to process, mine, and generate insights from Big Data, including unstructured and text files. Bayesian Models may also be used on structured data that meet specific requirements. The data requirements conditions for each model are documented through messages, tool tips, and documentation. FAQs and other relavant documents can be found on the AroniSoft (http://www.aroni.us) Website

Processing unstructured text in in AroniSmartLytics™

Unstructured text stored in flat text files can be opened and processed through the Descriptive Analysis Module.

Unstructured Data Requirement.

The unstructured data must meet the following requirements:

- The flat files must be in text format
- The flats files must be saved in dedicated sub-folders. No other unrelated file or sub-sub-folder must be in the sub-folders
- Each sub-folder corresponds to a category
- The sub-folders must be within a dedicated folder. This folder will be the one to be selected for text processing.

Organization of the output from processing unstructured text files.

- Each flat file will be transformed into an observation.
- The text in each flat file will be stored as the value of the observation, under the variable (or attribute) called "text"
- The name of each sub-folder will be stored as the value of the observation, under the variable (or attribute) called "__class__". Hence, all the flat files within a subfolder, will have the __class_ value equal to the name of the sub-folder. There is an option to save the path of the sub-folder as a variable (attribute)
- The folder path becomes the name of the data set.
- The dataset will be stored in "Aroni" format, as describe above, with the header, organized as follows:

  @dataset <path to the folder>
  @variable __class_ {<sub-folder1>, <sub-folder 2>, …}
  @variable text string
  @variable filename string

How to process unstructured text files.

Warning: please be aware that depending on the size and the complexity of the unstructured text files, processing may take a long time and may require large memory and CPU power.

The purpose of the "Processing Text Files" capability is to translate unstructured text files into "Aroni" format, where the text is transformed into word vectors. The processed words become the new attributes. To process text files stored in flat files, saved within sub-folders , under a main folder, follow these steps:

- Launch AroniSmartLytics
- From Descriptive Analysis Module, there are three options for the data format: CSV (with or without header), Aroni, and Process Text Files.
- Choose "Process Text Files". A new window opens up and the current open dataset is displayed in the table and the viewer.
- Select "Open Directory" and choose the main "folder" described above. This is the folder that contains the sub-folders where the flat files are stored.
    - Available options for processing directory:
        - Output Filename: to output the flat-file name into the dataset
        - Retain Attribute: to retain all the text in the files
        - Character set: default: is UTF-8
- Select "Create Data". Once the unprocessed data is loaded, the controls to set options become enabled.
- Choose and select the Text files processing options. The options selected depend on how the text should be translated into the word vector. A number of options are available:
    - All text strings, except the __class__ variable will be transformed into word vectors.
    - Word count: get the word counts rather than word presence flag (0 or 1).
    - Minimum Frequency: required minimum term frequency for word counts (default = 1)."

- TF Transform, or Term Frequency: transform the word frequencies into $\log(1+f_{ij})$, where f_{ij} is the frequency of word i in the jth text file
- IDF Transform, or Inverse Document Frequency: transform the word frequencies into f_{ij} *log(number of flat text files/number of text files containing the word i), where f_{ij} is the frequency of word i in the jth text file
- Lower case: transform all words into lower case format
- Ignore stop list: ignore the AroniSmartLytics™ internal stop list.
- Use own stop list: provide own customized stop list file. The stop list file must be in a text format, with one word per line and comments sratung with # or %
- Prefix: provide a prefix for the created variables (word attributes) names.
- Optimize count: do not enforce the maximum number of words and the minimum term frequency per __class__ (sub-folder), but based on the total number of text files in all the sub-folders.
- Number of words to keep: approximate number of word fields to create. The words beyond the specified number are discarded. Settings the number of words to keep waits untill the full dictionary is built. This requires huge memory and should be used with caution.
- Prune rate: the rate (e.g., 20 for every 20% of the input dataset) at which to periodically prune the words dictionary generated from all the relevant words after applying stemming algorithms. This option is usually preferred over "Number of words to keep" for large text files or when memory size is of concern.
- Stemming algorithm: The stemming algorithm to use. Three algorithms are implemented: Lovins and Iterated Lovins

- Click "**Process Text File**" button. Depending on the size and complexity of the text in the file and the computer resources, the processing may take long time. At the end the processed data is displayed in the table and in the text viewer.

- Click: "**Save Copy**" to save "Aroni" format version in a selected folder on the computer.
- There are options to remove some unwanted words attributes, by using the selection matrix at the top left of variable browser (ALL, Selected, Invert, and Reset) or pattern. After selecting the variables from the variables list, and selecting an option from the option matrix, click "**Refresh Table**" button to keep only the selected variables (only the first 20 variables will be displayed in the table). Once satisfied, click "**Save Copy**", to save the dataset with the selected word attributes. Do not forget to include __class__ variable.

Setting Up the Bayesian Network Models in AroniSmartLytics™

Bayesian Network Models are accessed through the Bayesian Models tab. Bayesian Network Models are available for "Aroni" format files only, since the class variable must be either nominal (including binary). Aroni format files can be either supplied and opened from the local folder or a result of processing unstructured text files.

Setting up the Bayesian Models model is intuitive. However, the analyst, student, researchers are expected to be familiar with the Bayesian statistics and Bayesian Network models, estimators and search algorithms. Follow the steps below to setup a Bayesian Model:

- Step 1**: Model :** Choose a Bayesian Model.
- Step 2: If the model is **Bayesian Network,** select an Estimator, then chose a Search Algorithm.
- Step 3: **Select the nominal dependent (class) variable (attribute)**
- Step 4: **Select independent variables (attributes).** Depending on the model, the variables must be nominal or numeric. For Bayesian Network, numerical variables will be discretized into categories. To avoid discretization, declare categorical variables as nominal, in Aroni format file, before opening the file.
- Step 5: click **Model Setup** button. A panel opens with specific options controls for the selected model enabled. Default values are already filled in for some options. The user may opt to run the model with defaults values or customize the options before running the model.
- Step 6: Choose appropriate options for the model selected in Step 1 and 2.

- Step 7: click "Ok/Continue" to apply the options or "Cancel" to cancel the operation. "Reset" button reapply the default options. Clicking "Ok/Continue" and "Cancel" closes the options panel.
- Step 8: Click "Run Model"

After the model finishes to run, the output results are displayed in the text viewer. For Bayesian Network models, the network structure is also shown.

As the models are run, a list of the models is shown in the browser on the right.

Use Top, Clear, End buttons to scroll into the output and select the model to view from the browser.

XIII. AroniSmartLytics™ Additional Help

The team at AroniSmartLytics™ truly believes that AroniSmartLytics ™ is unique and at the cutting edge, and that the user will enjoy leveraging the power of this intuitive, user friendly, yet powerful and rigorous tool to increase their knowledge and productivity and advance research and informed decision making.

Welcome to AroniSmartLytics™ World of Statistics, Research, Data Mining, Big Data, and Business Intelligence.

AroniSmartLytics™ is the most advanced tool in the AroniSmart ™ series that includes AroniStat™, AroniSmartStat™. For a focus to probability distributions, AroniSoft LLC has a simplified solution: AroniStat™. For Applied Statistics for Research, AroniSmartStat™ is the best option. AroniSmartLytics ™ addresses the needs of the most advanced researchers, business analysts, students, teachers, political analysts, statisticians, and data miners and decision makers.

For more help:

e-mail: aronisoft@optonline.net or aroni@aroni.us

web: http://www.aroni.us;

or http://www.aronismart.com

Selected References

Aitchison, J. and Aitken, C. G. G. (1976), "*Multivariate Binary Discrimination by the Kernel Method*," Biometrika, 63, 413-420.

Banfield, J. D. and Raftery, A. E. (1993), "Model-Based Gaussian and non Gaussian Clustering," Biometrics, 49, 803-821.

Biernacki, C. Celeux, G. and Govaert, G. (1999), "An improvement of the NEC criterion for assessing the number of components arising from a mixture," Pattern Recognition letters, No 20, 267-272.

Biernacki, C. and Govaert, G. (1999). Choosing Models in Model-based Clustering and Discriminant Analysis. Journal of Statistical Computation and Simulation, 64, 49-71.

Biernacki, C. Celeux, G. and Govaert, G. (2000), "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 22, No 7, 719-725.

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, 41, 561-575.

R.R. Bouckaert., R.R (1995) "Bayesian Belief Networks: from Construction to Inference." Ph.D. thesis, University of Utrecht, 1995.

Bouveyron C., Girard S. and Schmid C., High Dimensional Discriminant Analysis, Communications in Statistics: Theory and Methods, 36, 2607-2623.

Bozdogan, H. (1993), "Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix," in Information and Classification,O. Opitz, B. Lausen, and R. Klar (eds.), Heidelberg: Springer-Verlag, pp. 40-54.

Buntine, W.L. (1996). A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering, 8:195–210.

Celeux, G. and Govaert, G. (1995) "Parsimonious Gaussian models in cluster analysis". Pattern Recognition, 28, 781-793.

Celeux, G. and Soromenho, G. (1996) "An entropy criterion for assessing the number of clusters in a mixture model". Journal of Classification, 13, 195-212.
Cheng, J and Greiner, R (1999). Comparing bayesian network classifiers. Proceedings UAI, 101–107.

Chow, C.K. and Liu C.N.(1968). "Approximating discrete probability distributions with dependence trees." IEEE Trans. on Info. Theory, IT-14: 426–467.

Cooper, G. and E. Herskovits. "A Bayesian method for the induction of probabilistic networks from data." Machine Learning, 9: 309–347, 1992.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Statis. Soc. B, 39, 1-38.

Everitt, B. (1984). An Introduction to Latent Variable Models. London, Chapman and Hall.

Fraley, C. and Raftery, A. E. (1998): How Many Clusters ? Answers via Model-based Cluster Analysis. The Computer Journal, 41, 578-588.
215-231.

Flury, B. W., Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. SIAM J. Scientific Statist. Comput., 7, 169-184.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. JASA, 62, 1159-1178.

Friedman, N., Geiger, D, and Goldszmidt, M. (1997). Bayesian Network Classifiers. Machine Learning, 29: 131–163.

George, E.I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. Journal Of the American Statistical Association, 88 (423), pp. 881-889.

Goodman, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," Biometrika, 61,

Heckerman, D., Geiger, D. and Chickering, D.M (1995). "Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning, 20(3): 197–243.

Kanyamibwa, F. (1997). "Management of Quality Related Costs in Manufacturing Systems. A Thesis In Business Administration. The Smeal College of Business Administration. The Pennsylvania State University.

Kanyamibwa, F., Christy, D.P, and Fong, D. K. H, "Variable Selection in Product Design," Quality Management Journal, vol. 8, no. 1: 62-79.

Kanyamibwa, F. and J. K. Ord (2000). "Economic Process Control Under Uncertainty. Production and Operations Management. Volume 9, Issue 2, pages 184-202.

Keribin, C. (2000). Consistent estimation of the order of mixture. Sankhya, 62, 49-66.

Lauritzen, S.L. and Spiegelhalter, D.J. (1998). "Local Computations with Probabilities on graphical structures and their applications to expert systems (with discussion)". Journal of the Royal Statistical Society B. 1988, 50, 157-224

Maronna, R. and Jacovkis, P. M. (1974). Multivariate procedures with variable metrics. Biometrics, 30, 499-505.

McLachlan, G. J. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. in Handbook of Statistics (Vol. 2),

Moore, A. and Lee, M.S. (1998). Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets, JAIR, Volume 8, pages 67-91

P. R. Krishnaiah and L. N. Kanal (Eds.). Amsterdam: North-Holland, pp. 199-208.

McLachlan, G. J. and Peel D. (2000). Finite Mixture Models. New York, Wiley.

Press, Teukolsky, Vetterling and Flannery(1995)
Press, W.H., Teukolsky S.A, Vetterling, W. T, and Flannery, B. (2007). Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Third edition.

Schwarz, G. (1978), "Estimating the Dimension of a Model," Annals of Statistics, 6, 461-464.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. Biometrics, 27, 387-397.

Tipping, M. E. and C.M. Bishop (1999). Mixtures of probabilistic principal component analysers. Neural Computation 11, 443-482.

Verma, T. and Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. Proc. of the Eighth Conference on Uncertainty in Artificial Intelligence, 323-330.

Ward, J.H. (1963) Hierarchical groupings to optimize an objective function. JASA, 58, 236-244.

Witten, I.H. and E. Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.