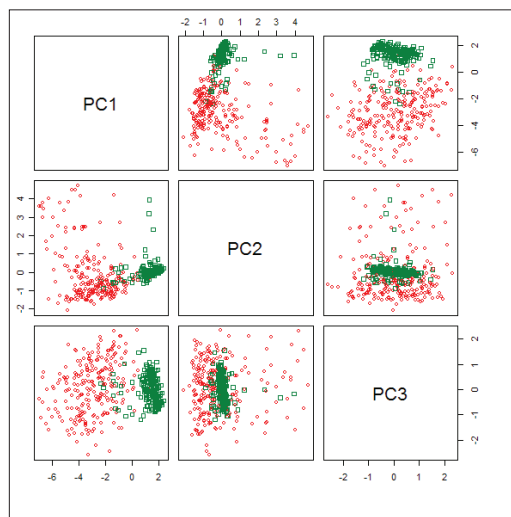


R STATISTICS COMPONENT COLLECTION

The R Statistics Component Collection for Pipeline Pilot allows you to perform exploratory analyses, create informative graphics, and make educated decisions. It includes components that implement statistical methods for data manipulation, clustering, model-building, and data analysis. The underlying statistical engine is the widely used open source package R. This collection lets you apply R statistical analysis and data manipulation methods to Pipeline Pilot data streams. You can incorporate output results from R directly into your pipeline for further analysis using other components in the Pipeline Pilot framework. You can use your existing R scripts in custom Pipeline Pilot components, enabling you to reuse them in different protocols or share them across the organization.

WITH THE R STATISTICS COLLECTION YOU CAN:

- Correlate multiple properties in a heat map display to see which ones are most relevant
- View distributions among population subgroups using box plots
- Perform an ANOVA to determine differences between means of multiple data sets
- Model data with logistic regression, a support vector machine (SVM), a neural network, or any of 10 other learning methods
- Apply the models you build to make predictions for new data sets, including model applicability domain (MAD) support to ensure the models are applied properly
- Save training data with any model, allowing the model to be extended as more experimental data become available.



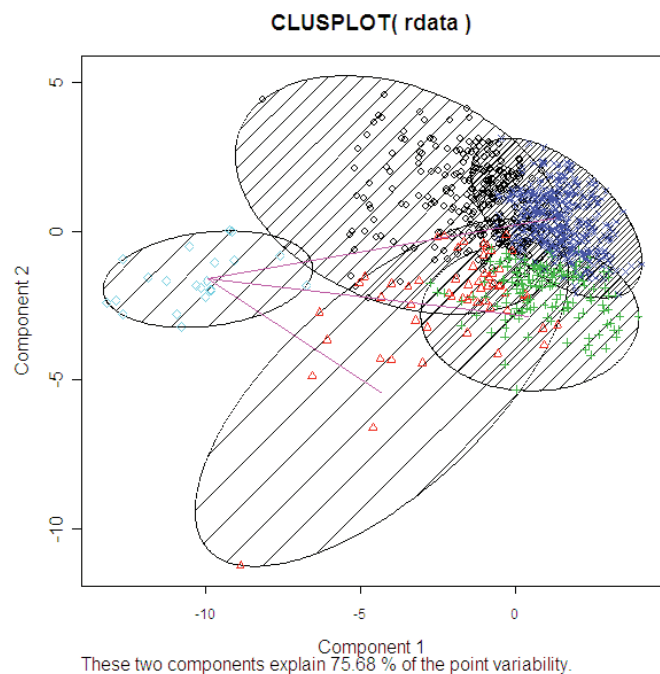
Pairs plot of breast cancer data based on principal component analysis

- Apply numerous different clustering methods
- Apply your own R script to each Pipeline Pilot data record or to the data stream as a whole

ANALYSIS

To compare multiple sets of measurements, you can perform a significance analysis using a t-test or ANOVA to determine whether the means for the different sets are the same. The R Correlation Matrix component generates a matrix for a set of descriptors that indicates the degree of correlation between them and a heat map plot that helps you visually find patterns in descriptor space. Components include:

- R ANOVA
- R K-Nearest Neighbors
- R Correlation Matrix
- R Principal Components Analysis
- R Probability Distributions
- R One-variable Tests
- R Factor Analysis
- R Two-variables Tests



Cluster Plot

CLUSTERING

With R, Pipeline Pilot offers a variety of different clustering methods that you can use in combination with all types of Pipeline Pilot data. For example, you can use fingerprints for a molecular data set as the descriptors for hierarchical or k-means clustering methods in R. Components include:

- R Cluster Agnes
- R Cluster Fanny
- R Cluster Clara
- R Cluster PAM
- R Cluster Diana
- R Cluster K-Means

DATA MANIPULATION

If a data set is incomplete, contains noise, or is irregular for other reasons, you can replace missing values or smooth the data using one of the R data manipulation components. Components include:

- R Missing Values
- R Loess Smoother
- R Remove Zero-Variance Properties
- R Spline Smoother
- R Smoother
- R Friedman SuperSmoother

CHARTS

Charts play an essential role in analyzing and reporting statistical results. These components produce PNG images that can be displayed in an HTML viewer or embedded in a report. Components include:

- R 2D Plot
- R Histogram
- R 3D Plot
- R Parallel Coordinates Plot
- R Pairs Plot
- R XY Plot
- R Conditional Plot

LEARN MODELS

To complement Pipeline Pilot's learning methods in the Data Modeling and Advanced Data Modeling Collections, the R Statistics Collection offers neural networks, support vector machine and several other model types for statistical learning. A variety of methods for both classification and regression problems are supported. Components include:

- Learn R Linear Model
- Learn R Linear Discriminant Analysis Model
- Learn R Generalized Linear Model
- Learn R Neural Net Model
- Learn R Non-Linear Model
- Learn R Support Vector Machine Model
- Learn R Logistic Regression Model
- Learn R Partial Least Squares Model

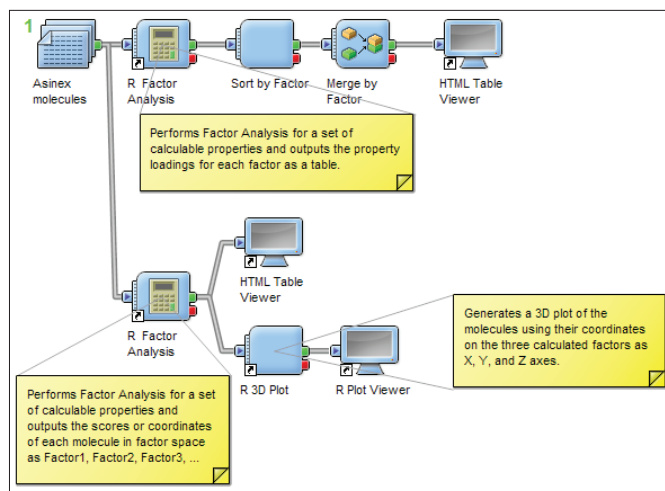
MULTI-DIMENSIONAL SCALING

Data sets can be characterized by a matrix containing the pairwise distances between the data items. Components in the R Statistics Collection provide multi-dimensional scaling to place the data in a low dimensional space, preserving as closely as possible the original distances. Components include:

- R Classical MDS
- R Sammon
- R Nonmetric MDS
- R Self Organizing Map

CUSTOMIZATION

All components in the R Statistics Collection are implemented as subprotocols. If you are familiar with R scripting, this means that you can modify and customize them to incorporate additional R capabilities into Pipeline Pilot. In addition, the following two



Protocol built with the R Statistics Component Collection showing Factor Analysis of Asinex data

components allow you to apply an R script to the Pipeline Pilot data stream as a whole or separately to each data record entering the component:

- R Custom Script
- R Custom Script for Each Data

ABOUT PIPELINE PILOT

Pipeline Pilot is an enterprise-scalable scientific informatics platform that enhances research and development organizations' ability to innovate by uncovering scientific value locked in disparate data silos, automating scientific workflows, and facilitating collaboration throughout the wider scientific community. Pipeline Pilot's Component Collections are the "scientific building blocks" of the platform and are grouped by category of science or function. By graphically combining components, you can construct workflows for data retrieval, filtering, analysis, and reporting.

To learn more about Pipeline Pilot, go to accelrys.com/pipeline-pilot