United States Environmental Protection Agency

EPA/600/F-08/016
December 2008
www.epa.gov

Quick Overview

Scout 2008 Version 1.0

A Statistical Software Package for the Classical and Robust Analysis of Univariate and Multivariate Data Sets with and without Non-detect Observations

RESEARCH AND DEVELOPMENT

EPA/600/F-08/016
December 2008
www.epa.gov

Quick Overview

Scout 2008 Version 1.0

A Statistical Software Package for the Classical and Robust Analysis of Univariate and Multivariate Data Sets with and without Non-detect Observations

Notice

The United States Environmental Protection Agency (EPA) through its Office of Research and Development (ORD) funded and managed the research described here. It has been peer reviewed by the EPA and approved for publication. Mention of trade names and commercial products does not constitute endorsement or recommendation by the EPA for use.

Introduction

The Scout 2008 version 1.0 statistical software package has been updated from past DOS and Windows versions to provide classical and robust univariate and multivariate graphical and statistical methods that are not typically available in commercial or freeware statistical software packages. Scout 2008 version 1.0 runs under the Microsoft Windows XP operating system (see the section on "Software used to develop Scout 2008" for details). Scout 2008 includes the latest state-of-the science classical, robust, and resistant univariate and multivariate outlier identification and robust estimation methods available, including: iterative classical, iterative influence function (e.g., biweight, Huber, PROP) based M-estimates, multivariate trimming (MVT), least median-of-squared residuals (LMS) regression, and minimum covariance determinant (MCD). Scout 2008 offers classical and robust methods to estimate: multivariate location and scale, univariate intervals, multiple linear regression parameters, principal components (PCs), and discriminant (Fisher, linear, and quadratic) functions (DFs). The discriminant analysis module of Scout 2008 can perform cross validation using several methods, including: leave-one-out (LOO), split samples, and bootstrap methods. Some initial choices for the iterative estimation of location and scale include: the orthogonalized Kettenring and Gnanadesikan (OKG) method; median, median absolute deviation (MAD), or inter-quartile range (IQR) based estimates; and the MCD method.

Below detection limit (BDL) or non-detect (ND) observations are inevitable in many applications, including those from the environmental, ecological, medical, pharmaceutical, and chemometric fields. For left-censored data sets with BDL, Scout 2008 offers both univariate and some multivariate graphical and statistical methods to estimate location and scale, interval estimates, and tolerance and prediction ellipsoids. Statistical methods incorporated in the Quality Assurance/Quality Control (QA/QC) module of Scout 2008 can be used to address univariate and multivariate industrial and environmental quality control applications. The QA/QC module of Scout 2008 can be

used to perform background versus site comparisons based upon univariate, as well as multivariate, data sets with and without BDL observations.

Two stand-alone software packages, ParallAX and ProUCL version 4.00.03, have been incorporated into the Scout 2008 statistical package. The ParallAX software offers graphical tools and unique classification algorithms to analyze multivariate data using a parallel coordinates system (details and references may be found in the Scout 2008 version 1.0 User Guide). ProUCL version 4.00.03 is a statistical software package that addresses specific environmental applications (e.g., in risk analysis, computing upper interval limits, and comparing site and background populations). ProUCL version 4.00.03 also has many statistical inference methods for data sets with and without ND observations. The details of those methods may be found in the reference by Singh et al. (2006), and in the ProUCL Technical Guide (USEPA, 2007). Those and other related references, the ProUCL version 4.0 software package, the Scout 2008 statistical software package, and the corresponding Scout 2008 Version 1.0 User Guide, are available at are available at the following web site:

http://www.epa.gov/nerlesd1/databases/datahome.htm.

It is strongly recommended that the Scout 2008 Version 1.0 User Guide be reviewed prior to using the Scout 2008 software. Some facts about the modules and the statistical methods available in the Scout 2008 software are presented in that document; more detail for those statistical methods may be found in the comprehensive bibliography given in the Scout 2008 Version 1.0 User Guide.

Please note that a "List of acronyms" used in this facts sheet is given at the end (before the "References").

Software used to develop Scout 2008

Several embedded commercial software programs were evaluated to shorten development time and to ensure the longevity of Super 2008. Scout 2008 has been developed in the Microsoft .NET Framework using the C# programming language to run under the Microsoft Windows XP operating systems. As such, to properly run Scout 2008, the computer using the program must have the .NET Framework pre-installed. Most computers running under Windows XP has this framework installed; however, if there are problems encountered when trying to run Scout 2008, then check to make sure that NET Framework version 1.1 (or later) is installed. The .NET files can installed from one of the following two Web sites:

- http://msdn2.microsoft.com/en-us/netframework/default.aspx
  Note: download .Net version 1.1.

- http://www.microsoft.com/downloads/details.aspx?FamilyId=262D25E3-F589-4842-8157-034D1E7CF3A3&displaylang=en

The Scout 2008 source code uses the following embedded licensed software:

Chart FX 6.2 (for graphics), http://www.softwarefx.com

Quinn-Curtis QCChart 3D Charting Tools for .Net (for graphics), http://www.quinn-curtis.com

NMath (for mathematical and statistical libraries), http://www.centerspace.net/

FarPoint (for spreadsheet applications), http://www.fpoint.com/

Scout 2008 Version 1.0 Modules

The Scout 2008 statistical software package has been updated from past Scout versions, that were either DOS or Windows based platforms, to a modern software platform, .NET, that runs under Microsoft Windows XP. The new Scout 2008 operates using a modular building block architectural approach and the program has elements to group similar functionalities into modules. All of the blocks have the potential to be modified as future modules are added. The modules are basically divided into pull-down menus, found at the top of the Scout 2008 main window. Those modules are briefly described below.

Data Module

The Data module can be used to generate univariate data sets from normal, lognormal, gamma, and uniform distributions, and multivariate data sets from multivariate normal distributions. In addition to performing typical univariate and bivariate data transformation operations on uncensored data sets without NDs, Scout 2008 has ND imputation methods (e.g., substitution and regression on order statistics (ROS) methods) for left-censored univariate data sets with multiple detection limits (MDLs). Basic tools to estimate missing observations are also available in this module. Transformation and imputation methods available in this module can be performed on individual data sets and also on grouped data sets consisting of data from two or more populations. The analysis of data from distributions that follow Benford's Law (data that have more frequently occurring smaller values than larger values, such as lake areas) is also available.

Graphs Module

For uncensored and left-censored data sets with multiple BDL observations, several univariate graphs can be generated for a single data set or a grouped data set consisting of data from two or more populations. Graphical displays in this module include:

- histograms,

- single and side-by-side multiple box plots (with options to draw horizontal lines at the threshold levels, such as not-to-exceed action levels and background screening levels),
- single and multiple normal quantile-quantile (Q-Q) plots, and
- index plots for single and grouped data sets.

Simple two-dimensional and three-dimensional scatter graphs can also be generated using this module. Observations on these graphs can be labeled by observation numbers or by their group labels (whenever applicable). For grouped data sets, one can interactively change and save the group memberships of the various observations on univariate index plots (with and without NDs) and on bivariate and three-dimensional scatter plots. Several formal and informal graphical displays are also available in other modules of Scout 2008, which are described in the following sections.

Stats/Goodness of Fit (GOF) Module

For both uncensored and left-censored univariate data sets, the Stats/GOF module can compute a variety of descriptive statistics, and classical parametric and nonparametric interval estimates, including Q-Q plots). This module has parametric and nonparametric univariate methods, including Kaplan-Meier (1958), ROS (Helsel, 2005), and bootstrap (Singh, Maichle, and Lee, 2006, Efron and Tibshirani,1997) methods, that can be used on left-censored data sets with ND observations potentially consisting of MDLs. Several univariate single sample (e.g., the t-test, the sign test, and the proportion test) and two sample (e.g., the t-test, the Wilcoxon-Mann-Whitney test, the quantile test, and Gehan's test) hypotheses tests for uncensored and left-censored data sets with MDLs are also available in this module (technical information is provided in the ProUCL Version 4.0 Technical Guide). Simple classical multivariate descriptive statistics, including the mean vector, and the covariance and correlation matrices, can also be computed. Various robust multivariate estimates of location (mean vector) and scale (covariance matrix) are computed by the Outliers/Estimates module. For multivariate data sets with ND observations, Scout 2008 computes estimates of the mean vector and the covariance matrix using the Kaplan-Meier (KM) method (still under investigation). The KM covariance and correlation matrices thus obtained can be used to perform principal component analysis (PCA) and other multivariate analyses. It should be noted that for exploratory graphical purposes, one may want to perform PCA on multivariate data sets with NDs imputed by substitution (e.g., DL/2) or ROS methods available in the Data module.

Univariate Goodness-of-Fit (GOF) Tests

The Stats/GOF module has univariate GOF (normal, lognormal, and gamma) tests for uncensored and left-censored data sets potentially consisting of MDLs. GOF tests for data sets with NDs also include ROS methods. All of the GOF tests in Scout 2008 are supplemented with respective hypothesized Q-Q plots with the appropriate test statistics

and the associated critical values or p-values displayed on those graphs (refer to the ProUCL Version 4.0 Technical Guide for details).

GOF Tests to Assess Multi-Normality

It is not easy to theoretically test and verify multivariate normality of a data set since even minor changes in data observations (or the presence of mild variance inflating values, or those values not complying with the covariance structure outliers) may significantly influence the test statistics, such as: multivariate kurtosis (MK) and multivariate skewness (MS), approximate skewness and approximate standardized kurtosis (Mardia,1970, 1974), and the Q-Q plot of Mahalanobis distances (MDs) often used to test multi-normality (and also to identify outliers). In addition to exact and approximate tests based upon MK, MS, and omnibus tests based upon them, a linear pattern displayed by data pairs (the theoretical quantiles of the distribution of MDs, and the ordered observed MDs) on the Q-Q plot of MDs may be used to assess approximate multi-normality (Singh, 1993), provided that there are no outliers present in the data set. Due to the reasons listed above, various GOF test statistics and graphical displays may lead to different conclusions regarding the multi-normality of a data set. Therefore, the multi-normality of a data set should be cautiously determined using GOF test statistics as well as graphical displays. Other measures, such as Q-Q plots and scatter plots of principal components based upon the correlation matrix, may also be used to assess the approximate multi-normality (Singh and Nocerino, 1995) of a data set. The GOF Q-Q plot of the MDs is formalized by displaying the exact test statistics for the: MS, MK, and the correlation coefficient along with their simulated critical values for a specified level of significance, α.

Classical Interval Estimates

For uncensored and left-censored data sets of various sizes and skewness levels, the Intervals option of the Stats/GOF module can estimate the mean, the variance, and other population parameters, as well as compute several parametric (e.g., normal, lognormal, and gamma distributions) and nonparametric (e.g., KM, bootstrap) upper limits (refer to the ProUCL Version 4.0 Technical Guide for details), such as the upper confidence limit (UCL) of the mean, the upper prediction limit (UPL), and the upper tolerance limit (UTL). The Intervals option of the Stats/GOF module can compute two-sided classical intervals for both uncensored and left-censored data sets. Several graphical displays are available for classical intervals, such as the graphical comparisons of the interval estimates for various groups (e.g., monitoring wells, or background versus site areas).

Robust Interval Estimates with Graphical Displays

For data sets without non-detect observations, the Stats/GOF module has several robust estimation methods to compute univariate estimates of location and scale, and robust interval estimates. The univariate iterative robust estimation methods in Scout 2008 include: Tukey's bisquare influence function (Tukey (1977), Kafadar (1982)), Huber (1981) and PROP (Singh, 1993; Singh and Nocerino, 1997) influence functions, and the

trimming method (Devlin et al., 1981). Two choices (the classical mean and the standard deviation (*sd*), or the median and 1.48MAD or IQR/1.345) of the initial estimates (Hoaglin, Mosteller, and Tukey, 1983) are available for all of the iterative univariate estimation methods. The Robust Intervals option can be used to compute robust confidence intervals of the mean, prediction intervals for k (≥1) observations, tolerance intervals, simultaneous (with Bonferroni critical values from the distribution of the Max (MDs)), and the individual (with critical values from the distribution of MDs) intervals.

The Robust Interval option provides a graphical comparison of the various robust and classical interval estimation methods. Depending upon the selected options (e.g., trimming % or tuning constants (TCs)) and methods, some relevant robust statistics, such as the mean, the *sd*, the influence function, α, the trimming percentage (%), or the location and scale TCs used with the biweight method, are also displayed on those interval method comparison graphs. Both of the classical and robust control-chart-type interval index plots, exhibiting the associated limits for the selected variables, are also available. The Group Analysis option of the Robust Interval option can be used to formally compare the interval estimates of a characteristic of interest for various groups (e.g., lead concentrations in various areas of a polluted site, arsenic concentrations in monitoring wells, or the effectiveness of two or more treatment drugs) under study.

Outliers/Estimates Module

This module offers both univariate and multivariate classical and robust outlier identification and estimation methods. For univariate uncensored and left-censored data sets, Scout 2008 has classical outlier tests, such as the Dixon test, the Rosner test, and the Grubbs test (refer to the ProUCL Version 4.0 Technical Guide for details). For univariate data sets, this module also has Tukey's robust (1977) biweight (and its variation suggested by Kafadar (1982)) outlier identification and estimation method. Several other univariate robust methods are available as special cases of the multivariate iterative, robust and resistant methods. Multivariate (also used on univariate data) outlier identification and estimation methods included in Scout 2008 are:

- the sequential classical methods based upon the Max MD and the multivariate kurtosis (Stapanian et al., 1991, 1993),
- the iterative robust and resistant M-estimation methods (Maronna, 1976) based upon the Huber (Huber, 1981) and PROP (Singh, 1993, 1996)) influence functions,
- the MVT method (Devlin, Gnanadesikan and Kettenring, 1981), and
- the re-weighted fast MCD (Rousseeuw and Van Driessen, 1999) method.

For all of the iterative robust methods in Scout 2008, several robust choices for the initial estimates of location and scale are available, including the OKG (Devlin, Gnanadesikan and Kettenring, 1975), Maronna and Zamar (2002)) method. The success of a robust method in identifying multiple outliers depends upon the coverage factor (e.g., h in the MCD and LMS methods) or the cutoff levels used (e.g., for the influence function, α in the PROP M-estimation method), and the behavior of the influence function used (e.g.,

nondecreasing (Huber, 1981), redescending (Hampel, 1974), or smooth redescending (PROP, Singh, 1993)) to identify those outliers. This module generates Excel-type output spreadsheets summarizing the initial and the final estimates of location, scale, MDs, and the associated weights. Several graphical displays, including method comparison plots, are also generated by this module. The method comparison option can be used to graphically assess and compare the performances of the various outlier identification methods.

Graphical Displays in Scout 2008

Scout 2008 is equipped with univariate graphs, including:

- side-by-side box plots,
- histograms,
- index plots,
- multiple Q-Q plots,
- interval graphs,
- control charts,
- bivariate scatter plots of raw data, PC scores, and DF scores, with options to draw prediction or tolerance ellipsoids superimposed on those scatter plots,
- bivariate regression line plots,
- Y versus $\hat{Y}$ (Y-hat) plots,
- Residual versus Y (or $\hat{Y}$) plots,
- multiple residuals versus residuals (R-R) plots,
- multivariate Q-Q plots and index plots of Mahalanobis distances (MDs),
- multiple distance-distance (D-D) plots of MDs,
- PC loadings plots,
- PC Scree and Horn's test plots,
- Q-Q plots and scatter plots of PCs, and
- scatter plots for Fisher linear discriminant analyses.

Emphasis is given to the interactive graphical displays of univariate and multivariate data sets, which becomes quite useful when dealing with large data sets, perhaps consisting of (or representing) multiple populations. On most of the graphical displays generated by Scout 2008, data can be labeled by their respective observation numbers, numerical values, or by their group assignments (whenever available). On graphical displays with two or more groups, group assignments of various observations can be interactively changed and saved. The interactive group re-assignment option can be quite useful in:

- exploratory discriminant and classification analyses,
- geostatistical applications, such as site characterization and remediation decision making processes,

- extracting a background (e.g., lowest set of values) data set from a potentially mixed data set representing multiple populations (e.g., various areas of a large site consisting of clean, contaminated, and highly contaminated hot areas).

The user can label all of the observations simultaneously or label selected observations of interest (e.g., outliers, regression outliers, polluted site locations, or NDs).

In practice, the number of outliers present in a data set is not known in advance. It is desirable to use the graphical displays listed above with more than one value of the coverage factor, h, or the influence function, α, on the same data set. The graphical displays offer additional information about the patterns and outliers present in a data set. This kind of information cannot be obtained simply by reviewing the estimated parameters computed by statistical procedures. Most computed statistics (e.g., the mean vector, MDs, kurtosis, or PCs) get distorted by the presence of outliers. This extra step of using graphical displays can be helpful in determining the appropriate value of h or α that may be useful in controlling the masking or swamping of multiple outliers. Specifically, this can help the user to pick an appropriate value of h (MCD) or the influence function alpha (e.g., PROP), which in turn will help in obtaining more reliable and accurate estimates of the population parameters (e.g., location, scale, or regression parameters).

Several modules in Scout 2008 (Graphs, Stats/GOF, Outliers/Estimates, QA/QC, Regression, and Multivariate EDA) generate graphical displays to:

- determine univariate and multivariate data distributions,
- graphically compare grouped data,
- identify outliers,
- compare the performances of robust and classical outlier identification and estimation methods,
- identify regression outliers and leverage points,
- distinguish between good and bad leverage points,
- determine if test (site) data are in compliance with the training (reference) data (QA/QC module),
- perform discriminant and classification analyses.

Whenever applicable, those graphical displays have been formalized by displaying or drawing limits at appropriate critical values associated with the various GOF tests, individual MDs, Max (MDs), kurtosis, skewness, residuals, and leverage distances.

The Method Comparison option is available in both the Outliers/Estimates and Regression modules. Bivariate prediction or tolerance ellipsoids, and multivariate multiple D-D plots options available in the Outliers/Estimates module can be used to graphically compare the performances (e.g., in terms of masking and swamping effects) of various outlier identification methods available in Scout 2008. Those graphical displays provide useful information about the effectiveness (identifying all of the outliers without masking) and efficiency (not identifying inliers as outliers) of the various outlier methods in properly identifying potential outliers that might be present in the data set.

Similarly, the Regression module has bivariate regression fits, multivariate R-R and Y-Ŷ plots options to compare the performances of the various robust methods available in Scout 2008. Those graphical displays provide useful information about the effectiveness and efficiency of the various regression methods in producing proper regression fits that are not distorted by outliers and leverage points.

QA/QC Module

The QA/QC module provides univariate and multivariate classical, as well as robust, methods that may be used in quality assurance and quality control (QA/QC) applications. The QA/QC module has univariate control-chart-type interval graphs, multivariate control-chart-type index plots, and prediction and tolerance ellipsoids (Singh and Nocerino, 1995). Those graphs can be generated using all of the observations in a data set or just by using the observations in a specified training (e.g., background data, or placebo data) subset of the data. This training/test data option can be specifically useful in determining whether or not the observations from one test group (e.g., a polluted site, a test group, or a new treatment drug) can be considered to be from the training group (e.g., a reference group, background, or a placebo), perhaps with known, well-established and acceptable behavior of the characteristics of interest (e.g., concentrations of the contaminants of potential concern, or a placebo drug effect).

Univariate QA/QC

The Univariate option of the QA/QC module offers univariate control-chart-type interval graphs (e.g., Singh and Nocerino, 1997). Those graphs are used to compare test (e.g., site, project) data values with control limits (e.g., prediction, tolerance, or simultaneous limits) computed based upon some training (e.g., background, reference, or controlled) data set. The QA/QC module can be used to compare training (e.g., background, reference, or up-gradient wells) and test (e.g., polluted site, groundwater monitoring wells, dredged sediments) data sets. For such graphical displays, relevant statistics and limits are computed using a training (e.g., controlled, background, reference) data set, and all points in the training and test data sets are plotted on those graphical displays. Test data points (e.g., site observations) lying outside the limits (e.g., tolerance or simultaneous limits) may represent out-of-control observations, perhaps representing observations not belonging to the training data set population. The Univariate QA/QC option can also handle left-censored data sets consisting of ND observations. For data sets with NDs, the estimates of all of the relevant statistics (the mean, the *sd*, the standard error of the mean, or the upper and lower limits) are computed using the KM method. The individual ND data points displayed on those interval graphs are shown (in red color) based upon the user selected option (e.g., 0, DL, DL/2, or ROS estimates).

Multivariate QA/QC

Classical methods included in the Multivariate QA/QC module can handle data sets with ND observations. For data sets with NDs, the KM method is used to compute relevant statistics (e.g., the mean vector and the covariance matrix, or prediction and tolerance

ellipsoids) based upon the training data set. Those KM statistics are used to generate multivariate control chart-type graphs and prediction or tolerance ellipsoids. All of the data (raw or processed), including the imputed data (for NDs) from both the training and test data sets, are plotted on those control chart-type graphs. Processed data may represent Mahalanobis distances (used in control chart-type index plots) or principal component scores (used in prediction or tolerance ellipsoids). For uncensored data sets, classical estimates of the location and scale should be in agreement with the respective KM estimates. For uncensored full data sets, all of the robust methods available in the Outliers/Estimates module are also available in the QA/QC module. Relevant statistics needed to generate Q-Q plots, index plots, or contour ellipsoids are computed based upon the training data set. Observations from both training and test data sets are plotted on those control-chart-type graphs. Test data set observations lying above the Max-MD (computed using the Bonferroni inequality) limit on the Q-Q and index plots of the MDs, or lying outside of the tolerance ellipsoids, potentially represent observations that may not belong (nonconforming observations) to the population of the training data set.

Regression Module

The Regression module can perform multiple linear classical and robust regressions using several methods available in the literature. Specifically, Scout 2008 can perform the least median of squared (LMS) regression (Rousseeuw, 1984, Rousseeuw and Leroy (1987)), as well as the least percentile of squared (LPS) regression. Scout 2008 also performs robust regression based upon the M-estimation (without leverage option) and the generalized M-estimation (with leverage option) procedures for the MVT method, and the Huber, biweight, and PROP influence functions (Singh and Nocerino (1995)). This module generates several formalized graphical displays, including: the Q-Q and index plots of residuals, with the appropriate limits drawn at the critical values of the residuals univariate unsquared MDs; scatter plots of the residuals versus the unsquared leverage distances (Rousseeuw and van Zomeren, 1990) for the generalized M-estimation regression method; residual versus residual (R-R) plots; Y versus $\hat{Y}$ plots; and Y versus standardized residuals scatter plots. Residuals are not standardized when the scale estimate (standard deviation of the residuals) is very small (such as less than $1e^{-10}$).

The graphical displays included in the Regression module are useful to identify regression outliers, inconsistent (bad) leverage points, and to distinguish between good (consistent) and bad (inconsistent) leverage points. The Method Comparison option available in this module can be used to graphically assess and compare the performances of the various regression methods. For most of the graphical displays, Scout 2008 collects and employs user-selected critical levels to compute the appropriate critical values of the statistics plotted (e.g., the critical values of the MDs, or the critical value of the Max MD) in the graphical displays. Scout 2008 also generates confidence or prediction bands around fitted regression models, including classical first order, quadratic, and cubic, linear, as well as robust linear models. In addition to the robust regression methods, Scout 2008 also performs regression diagnostics.

The LPS regression estimates, obtained by minimizing the k^{th} (k> 50%) percentile of the squared residuals, will have a lower breakdown (BD) point than the LMS estimates. For example, the BD point of the LPS regression estimates obtained by minimizing the 75^{th} (k = 75%) percentile of the squared residuals is (n - [0.75n] – p + 2)/n, where n is the number of observations, p is the number of regression variables, and [x] is (in standard mathematical notation) the truncation function representing the largest integer contained in x. For example, if n = 100, then [n*0.75] = 75; however, if n = 50, then [n*0.75] = [37.5] = 37, which is truncated down to an integer.

Since the number of outliers (both regression and leverage) are not known in advance, it is suggested to use graphical displays to get some idea about the influence function alpha, α (for influence function based methods), or the percentage, k, of the outliers (for LPS method) that may be present in the data set. Based upon the outlier information thus obtained, one may perform an appropriate LMS, LPS, or M-estimation based regression on a given data set.

Multivariate EDA

The Multivariate EDA module can perform classical, as well as robust, principal component and discriminant analyses. The details of the robust PCA and DA based upon the MVT method, and the PROP and Huber influence functions, are given by Singh and Nocerino (1995). More details about the robust PCA and robust discriminant analyses based upon the MCD method can be found in Hubert, Rousseeuw, and Branden (2005), and Todorov and Pires (2007).

Principal Component Analysis (PCA)

For uncensored data sets without non-detect observations, Scout 2008 can perform classical PCA and robust PCA analyses based upon M-estimation (Campbell, 1980) methods (e.g., PROP, Huber, or MVT), and the MCD method. PCA is often used on large dimensional data sets as a dimension reduction technique so that statistical analyses can be performed on a much smaller number, k (≤ p), of PCs. PCA is also used to gain a better understanding of any relationships among variables, to reveal any grouping of observations, to find outliers, to estimate non-systematic variance, or to reveal any hidden data structure. PCA can be performed by using a covariance matrix or by first pre-processing the data matrix to be variable-scaled to unit variance (that is, by using the correlation matrix). PCA is usually performed using the correlation matrix since an un-scaled variable (that is, when using the covariance matrix) with a relatively much larger scale metric (e.g., a variable scaled in inches versus a variable scaled in miles) can dominate the variance, making the PCA results rather meaningless. However, if the variables are of the same type and measured in the same units, auto-scaling (i.e., using the correlation matrix) could exaggerate minor variations; thus, the covariance matrix should be used in such cases. Typically, the first few PCs explain most of the systematic variation present in a data set, and the last few PCs are useful in identifying observations which may not follow the variance structure displayed by the main (dominant) body of

data. Q-Q plots and scatter plots of first few PCs can be used to identify variance inflating outliers or to identify the presence of a mixture of different populations in a data set. Prediction and tolerance ellipsoids can be drawn on the scatter plot of PC scores. Significant jumps or turns in the Q-Q plots of PC scores suggest the presence of a mixture of different populations in the data set. Such graphs can also be used to assess the approximate multi-normality of a data set. Q-Q plots and scatter plots of the PC scores obtained using the covariance matrix may be used to identify potential outliers; whereas, Q-Q plots and scatter plots of PC scores based upon the correlation matrix may be cautiously used to assess approximate multi-normality.

Based upon the PC statistics and scores thus obtained, the PCA module generates scree and Horn plots for the eigenvalues, a loadings matrix plot, scatter plots of PC scores with options to draw prediction or tolerance ellipsoids, and normal Q-Q plots of the PC scores. The PC scores can be stored in the same or a different worksheet for future analyses. For multivariate data sets with NDs, not much guidance is available in the statistical literature on how to perform PCA. This topic is still under investigation. Scout 2008 can be used to perform PCA based upon based upon Kaplan-Meier (1958) method (still being investigated). Using the KM covariance (or correlation) matrix, Scout 2008 can generate scree and Horn Plots. For exploratory purposes, one can compute and plot the PC scores obtained using the KM covariance matrix and imputed values of the NDs (e.g., using substitution or ROS methods) from the original data set. For exploratory purposes, one may use the Data module of Scout 2008 to impute ND observations before using PCA module. This step will yield a full data set without any ND observations (the NDs are replaced by the imputed or substituted values). One can then use any of the classical and robust PCA methods available in Scout 2008.

Discriminant and Classification Analysis

Scout 2008 can be used to perform classical and robust (based upon the MVT method, and the PROP and Huber influence functions) Fisher linear discriminant analysis (FDA), and linear and quadratic discriminant analyses. The classical and robust FDA methods are supplemented with graphical displays. The available graphical displays include scree plots of eigenvalues and scatter plots of discriminant scores (for Fisher Discriminant Analysis) with options to draw prediction or tolerance ellipsoids by groups. As with all other displays with group assignment options, on scatter plots of discriminant scores, one can reclassify an observation from one group into another group interactively by using the "Change Group" and "Save Changes" options. That option can be quite useful for properly classifying borderline observations. Based upon the discriminant functions (classical or robust), Scout 2008 can be used to classify and plot observations with unknown (new observations) group memberships into one of the groups used in deriving those discriminant functions. Several cross validation (CV) methods for DA are also available in Scout 2008. The CV methods in Scout 2008 include: leave-one-out (Lachenbruch and Mickey (1968)), split samples (training and test sets), M-fold CV, and bootstrap methods (e.g., Davison and Hall (1992), Bradley and Efron (1997)). In order to use the CV methods properly, the user should make sure that enough data are available in each of the various groups included in the data set.

Output Generated by Scout 2008

All of the modules of Scout generate graphical output displays (*.gst file), Excel-type-spreadsheets (*.ost file), or both graphical displays and Excel-type-spreadsheets. The "ost" output file generated by Scout 2008 can be saved as an Excel file; and the "gst" graphical display can be saved as Microsoft Word or WordPerfect files for future documentation. All of the relevant information, statistics, and the classical and robust estimates of the parameters of interest are displayed on output sheets. All of the classical estimates, initial robust estimates, final robust estimates, and associated weights are displayed on output sheets generated by Scout 2008. The user can save intermediate results in a separate spreadsheet by choosing the Intermediate Iterations option. In addition to graphs, most graphical displays also exhibit relevant estimates, test statistics, and associated critical levels and p-values.

A Note on Data Set Size and Dimension

It is suggested that for the appropriate identification of outliers and for reliable robust estimates of various parameters of interest, the number of observations, n, in a data set should be at least 5p, where p represents the dimensionality (number of variables) of the data set. This is especially true when the dimension of p exceeds 5. From a theoretical point of view, Scout 2008 can compute various robust statistics and estimates for values of $n > (p+2)$. However, the results (estimate, graphs, and outliers) obtained using high p dimensional (often called "the curse of dimensionality") data sets of small n sizes may not always be reliable or defensible. It should also be noted that multivariate methods require that the number of observations be the same for each of the variables included in the multivariate data set to be used for statistical analyses.

List of Acronyms

α	level of significance
BDL	below detection limit
CV	cross validation
DA	discriminant analysis
D-D	multiple distance-distance
DF	discriminant function
DL	detection limit
FDA	Fisher linear discriminant analysis
h	coverage factor
IQR	inter-quartile range
KM	Kaplan-Meier
LMS	least median-of-squared
LOO	leave-one-out (LOO)
LPS	least percentile-of-squared
MAD	median absolute deviation
MCD	minimum covariance determinant
MD	Mahalanobis distance
MDL	multiple detection limit
MK	multivariate kurtosis
MS	multivariate skewness
MVT	multivariate trimming
n	number of observations

ND	non-detect observation
OKG	orthogonalized Kettenring and Gnanadesika
p	number of variables (dimensions)
PCA	principal component analysis
PC	principal component
PROP	proposed influence function
Q-Q	quantile-quantile
QA/QC	Quality Assurance/Quality Control
ROS	regression on order statistics
R-R	residual-residual
sd	standard deviation
TCs	tuning constants
UCL	upper confidence limit
UPL	upper prediction limit
UTL	upper tolerance limit

References

Campbell, N. A., "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation," Applied Statistics, 29, 1980, pp. 231–237.

Davison, A. and Hall, P., "On the Bias and Variability of Bootstrap and Cross-Validation Estimates of Error Rate in Discrimination Problems," Biometrika, Vol. 79, No. 2, June, 1992, pp. 279-284.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R., "Robust Estimation and Outlier Detection with Correlation Coefficients," Biometrika, 62, 1975, pp. 531-545.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R., "Robust Estimation of Dispersion Matrices and Principal Components," Journal of the American Statistical Association, 76, 1981, pp. 354-362.

Efron, B. and Tibshirani, R., "Improvements on Cross-Validation: The .632+ Bootstrap Method," Journal of the American Statistical Association, Vol. 92, No. 438, June, 1997, pp. 548-560.

Hampel, Frank R., "The Influence Curve and its Role in Robust Estimation," Journal of the American Statistical Association, 69, 1974, pp. 383–393.

Helsel, D.R. 2005. *Nondetects and Data Analysis*. Statistics for Censored Environmental Data. John Wiley and Sons, NY.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W., Understanding Robust and Exploratory Data Analysis, John Wiley and Sons, NY, 1983.

Horn, P.S., Pesce, A.J., and Copeland, B.E., "A Robust Approach to Reference Interval Estimation and Evaluation," Clinical Chemistry, 44:3, 1998, pp. 622-631.

Huber, P.J., Robust Statistics, John Wiley and Sons, NY, 1981.

Hubert, M., Rousseeuw, P.J., and Vanden Branden, K., "ROBPCA: A New Approach to Robust Principal Component Analysis," Technometrics, 47, 2005, pp. 64-79.

Kafadar, K., "A Biweight Approach to the One-Sample Problem," Journal of the American Statistical Association, 77, 1982, pp. 416-424.

Kaplan, E.L. and Meier, O. 1958. *Nonparametric Estimation from Incomplete Observations*. Journal of the American Statistical Association, Vol. 53. 457-481.

Lachenbruch, P.A., and Mickey, M.R., "Estimation of Error Rates in Discriminant Analysis," Technometrics, Vol. 10, No. 1, February, 1968, pp. 1-11.

Mardia, K.V., "Measures of Multivariate Skewness and Kurtosis with Applications," Biometrika, 57, 1970, pp. 519-530.

Mardia, K.V., "Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies," Sankhya, B 36, 1974, pp. 15-128.

Maronna, R.A., "Robust M-Estimators of Multivariate Location and Scatter," The Annals of Statistics, Vol. 4, No. 1, 1976, pp. 51-67.

Maronna, R.A., and Zamar, R.H., "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," Technometrics, 44, 2002, pp. 307-317.

Rousseeuw, P.J., "Least Median of Squares Regression," Journal of the American Statistical Association, 79, 1984, pp. 871-880.

Rousseeuw, P.J., and Leroy, A.M., Robust Regression and Outlier Detection, John Wiley and Sons, NY, 1987.

Rousseeuw, P.J., and van Zomeren, B.C., "Unmasking Multivariate Outliers and Leverage Points," Journal of the American Statistical Association, 85, 1990, pp. 633-651.

Rousseeuw, P.J., and Van Driessen, K., "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technometrics, 41, 1999, pp. 212-223.

Singh, A., Omnibus Robust Procedures for Assessment of Multivariate Normality and Detection of Multivariate Outliers, In Multivariate Environmental Statistics, Elsevier Science Publishers, Patil G.P. and Rao, C.R., Editors, 1993, pp. 445-488.

Singh, A., "Outliers and Robust Procedures in Some Chemometrics Applications," Chemometrics and Intelligent Laboratory Systems, 33, 1996, pp. 75-100.

Singh, A., Maichle, R., and Lee, S., On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations, EPA/600/R-06/022, March 2006.

Singh, A. and Nocerino, J.M., Robust Procedures for the Identification of Multiple Outliers, Handbook of Environmental Chemistry, Statistical Methods, Vol. 2. G, Springer Verlag, Germany, 1995, pp. 229-277.

Singh, A. and Nocerino, J.M., "Robust Intervals in Some Chemometric Applications," Chemometrics and Intelligent Laboratory Systems, 37, 1997, pp. 55-69.

Stapanian, M.A., Garner, F.C., Fitzgerald, K.E., Flatman, G.T., and Englund, E.J., "Properties of Two Multivariate Outlier Tests," Comm. Statist. Simula Computa, 20, 1991, pp. 667-687.

Stapanian, M.A., F.C. Garner, K.E. Fitzgerald, G.T. Flatman, and J.M. Nocerino. "Finding suspected causes of measurement error in multivariate environmental data." Journal of Chemometrics, 1993, 7:165-176.

Tukey, J.W., Exploratory Data Analysis, Addison-Wesley Publishing Company, Reading, MA, 1977.

U.S. Environmental Protection Agency (EPA). 2007. *ProUCL Version 4.0, A Statistical Software*. The software ProUCL 4.0 can be freely downloaded from the U.S. EPA Web site at: http://www.epa.gov/nerlesd1/tsc/software.htm

U.S. Environmental Protection Agency (EPA). 2007. *ProUCL 4.0. Technical Guide* Publication EPA/600/R-07/041.

U.S. Environmental Protection Agency (EPA). 2007. *ProUCL 4.0. User Guide* Publication EPA/600/R-07/038.

Valentin, T. and Pires, A., "Comparative Performance of Several Robust Linear Discriminant Analysis Methods," REVSTAT – Statistical Journal, Vol. 5, Number 1, March, 2007, pp. 63-83.