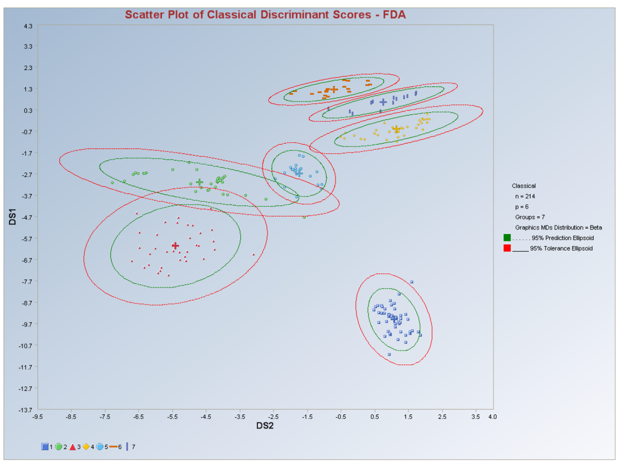
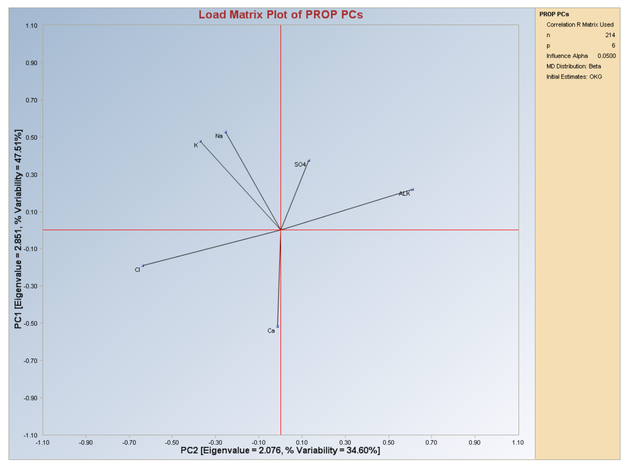
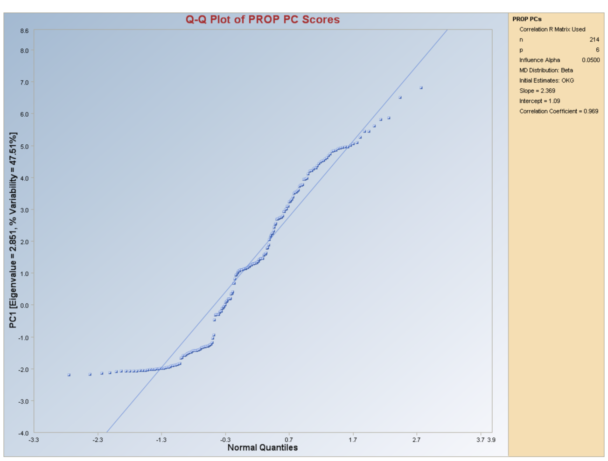
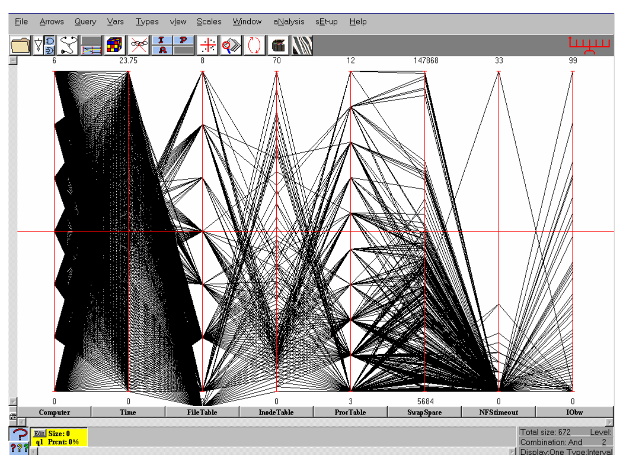
United States
Environmental Protection
Agency

Scout 2008 Version 1.0
User Guide
Part IV









RESEARCH AND DEVELOPMENT

EPA/600/R-08/038
February 2009
www.epa.gov

Scout 2008 Version 1.0 User Guide

(Second Edition, December 2008)

John Nocerino

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Environmental Sciences Division
Technology Support Center
Characterization and Monitoring Branch
944 E. Harmon Ave.
Las Vegas, NV 89119

Anita Singh, Ph.D.[1]
Robert Maichle[1]
Narain Armbya[1]
Ashok K. Singh, Ph.D.[2]

[1]Lockheed Martin Environmental Services
1050 E. Flamingo Road, Suite N240
Las Vegas, NV 89119

[2]Department of Hotel Management
University of Nevada, Las Vegas
Las Vegas, NV 89154

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC 20460

7663cmb09

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.

The Scout 2008 software was developed by Lockheed-Martin under a contract with the USEPA. Use of any portion of Scout 2008 that does not comply with the Scout 2008 User Guide is not recommended.

Scout 2008 contains embedded licensed software. Any modification of the Scout 2008 source code may violate the embedded licensed software agreements and is expressly forbidden.

The Scout 2008 software provided by the USEPA was scanned with McAfee VirusScan and is certified free of viruses.

With respect to the Scout 2008 distributed software and documentation, neither the USEPA, nor any of their employees, assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed. Furthermore, the Scout 2008 software and documentation are supplied "as-is" without guarantee or warranty, expressed or implied, including without limitation, any warranty of merchantability or fitness for a specific purpose.

Acronyms and Abbreviations

% NDs	Percentage of Non-detect observations
ACL	alternative concentration limit
A-D, AD	Anderson-Darling test
AM	arithmetic mean
ANOVA	Analysis of Variance
AOC	area(s) of concern
B*	Between groups matrix
BC	Box-Cox-type transformation
BCA	bias-corrected accelerated bootstrap method
BD	break down point
BDL	below detection limit
BTV	background threshold value
BW	Black and White (for printing)
CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
CL	compliance limit, confidence limits, control limits
CLT	central limit theorem
CMLE	Cohen's maximum likelihood estimate
COPC	contaminant(s) of potential concern
CV	Coefficient of Variation, cross validation
D-D	distance-distance
DA	discriminant analysis
DL	detection limit
DL/2 (t)	UCL based upon DL/2 method using Student's t-distribution cutoff value
DL/2 Estimates	estimates based upon data set with non-detects replaced by half of the respective detection limits
DQO	data quality objective
DS	discriminant scores
EA	exposure area
EDF	empirical distribution function
EM	expectation maximization
EPA	Environmental Protection Agency
EPC	exposure point concentration
FP-ROS (Land)	UCL based upon fully parametric ROS method using Land's H-statistic

Gamma ROS (Approx.)	UCL based upon Gamma ROS method using the bias-corrected accelerated bootstrap method
Gamma ROS (BCA)	UCL based upon Gamma ROS method using the gamma approximate-UCL method
GOF, G.O.F.	goodness-of-fit
H-UCL	UCL based upon Land's H-statistic
HBK	Hawkins Bradu Kaas
HUBER	Huber estimation method
ID	identification code
IQR	interquartile range
K	Next K, Other K, Future K
KG	Kettenring Gnanadesikan
KM (%)	UCL based upon Kaplan-Meier estimates using the percentile bootstrap method
KM (Chebyshev)	UCL based upon Kaplan-Meier estimates using the Chebyshev inequality
KM (t)	UCL based upon Kaplan-Meier estimates using the Student's t-distribution cutoff value
KM (z)	UCL based upon Kaplan-Meier estimates using standard normal distribution cutoff value
K-M, KM	Kaplan-Meier
K-S, KS	Kolmogorov-Smirnov
LMS	least median squares
LN	lognormal distribution
Log-ROS Estimates	estimates based upon data set with extrapolated non-detect values obtained using robust ROS method
LPS	least percentile squares
MAD	Median Absolute Deviation
Maximum	Maximum value
MC	minimization criterion
MCD	minimum covariance determinant
MCL	maximum concentration limit
MD	Mahalanobis distance
Mean	classical average value
Median	Median value
Minimum	Minimum value
MLE	maximum likelihood estimate
MLE (t)	UCL based upon maximum likelihood estimates using Student's t-distribution cutoff value

MLE (Tiku)	UCL based upon maximum likelihood estimates using the Tiku's method
Multi Q-Q	multiple quantile-quantile plot
MVT	multivariate trimming
MVUE	minimum variance unbiased estimate
ND	non-detect or non-detects
NERL	National Exposure Research Laboratory
NumNDs	Number of Non-detects
NumObs	Number of Observations
OKG	Orthogonalized Kettenring Gnanadesikan
OLS	ordinary least squares
ORD	Office of Research and Development
PCA	principal component analysis
PCs	principal components
PCS	principal component scores
PLs	prediction limits
PRG	preliminary remediation goals
PROP	proposed estimation method
Q-Q	quantile-quantile
RBC	risk-based cleanup
RCRA	Resource Conservation and Recovery Act
ROS	regression on order statistics
RU	remediation unit
S	substantial difference
SD, *Sd, sd*	standard deviation
SLs	simultaneous limits
SSL	soil screening levels
S-W, SW	Shapiro-Wilk
TLs	tolerance limits
UCL	upper confidence limit
UCL95, 95% UCL	95% upper confidence limit
UPL	upper prediction limit
UPL95, 95% UPL	95% upper prediction limit
USEPA	United States Environmental Protection Agency
UTL	upper tolerance limit
Variance	classical variance
W*	Within groups matrix

WiB matrix	Inverse of W* cross-product B* matrix
WMW	Wilcoxon-Mann-Whitney
WRS	Wilcoxon Rank Sum
WSR	Wilcoxon Signed Rank
Wsum	Sum of weights
Wsum2	Sum of squared weights

Table of Contents

Chapter 10

Multivariate EDA

The Multivariate Exploratory Data Analysis (EDA) module of Scout performs principal component analysis (PCA) and discriminant analysis (DA). The data should have a minimum of two variables. In order to perform a DA, a group variable (column) should be included in the data set. The values (alphanumeric) of the group variable represent the various group categories.

10.1 Principal Component Analysis

Principal component analysis is one of the well recognized data dimension reduction techniques. While the first few high variance principal components (PCs) represent most of the systematic variation in the data, the last few low variance PCs provide useful information about the random variation that might be present in the experimental results. Graphical displays of the first few PCs are routinely used as unsupervised pattern recognition and classification techniques. The normal probability Q-Q plots and scatter plots of the PCs are also used for the detection of multivariate outliers.

Since the MLE of the dispersion matrix and the correlation matrix get distorted by outliers, the classical PCs (obtained using the covariance or correlation matrix) also get distorted by outliers. The robust PCs give more precise estimates of the systematic and random variation in the data by assigning reduced weights to the outlying observations.

Let $p = (p_1, p_2, ..., p_p)$ represent the matrix of eigen vectors corresponding to the eigen values (λ_1, λ_2, …, λ_p) of the sample dispersion (correlation) matrix (classical or robust). The eigen vector, p_1, corresponds to the largest eigen value, λ_1,…, and the eigen vector, p_p, corresponds to the smallest eigen value, λ_p. The equation, $y = px$, represents the p principal components, with $y_i = p'_i x$ representing the ith principal component.

Q-Q plots of the principal components are sometimes used to reveal suspect observations and also to provide checks on the normality assumption. Scatter plots of the first few high-variance PCs reveal outliers which may inappropriately inflate the variances and covariances. Plots of the last few low-variance PCs typically identify observations that violate the correlation structure imposed by the main stream of the data, but that are not necessarily outlying with respect to any of the individual variables.

Scout can compute the PCs for both the classical dispersion (correlation) matrix and the robust dispersion (correlation) matrix. The iterative or robust procedures available in Scout are: the sequential classical, PROP, Huber, MVT, and MCD procedures.

Few rules have been incorporated into Scout for the ease of graphing in the Multivariate EDA module.

- A rule, called the proportion rule, exists where only the scores and loadings corresponding to the proportion of eigen values greater than 0.0001 will be plotted.

- If any of the final matrix used to compute the eigen values and the loadings are singular, then the graphing is based on the proportions rule.

- If the any of the eigen values of the final matrix is less than 10^{-20} or greater than 10^{+20} then those loadings and the scores based on those eigen values will not be plotted.

- If the classical initial matrix used for generating the scores in any of the robust method is singular, then a message will be displayed and further calculations will be stopped.

- If the standard deviation of any of the scores is less than 10^{-7} or greater 10^{+7}, then contours will not be plotted on their respective scatter plots.

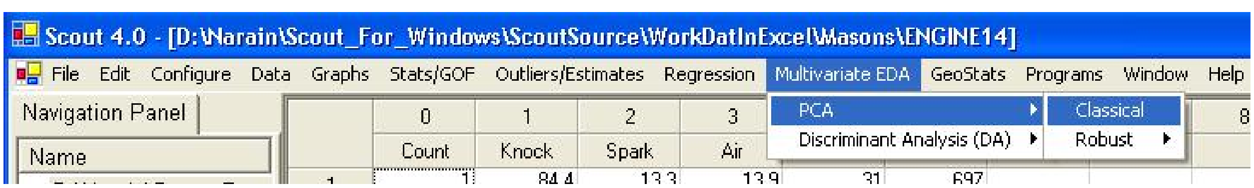
- If the coefficient variation of any of the scores is less than 10^{-7} or greater 10^{+7}, then contours will not be plotted on their respective scatter plots.

- If the absolute value of the correlation between the two variables used in scatter plots is greater than 0.99, then the contours will not be plotted.

- If the absolute difference between the standard deviations of the two variables used in the scatter plot is less than 10^{-20}, then contours will not be plotted.

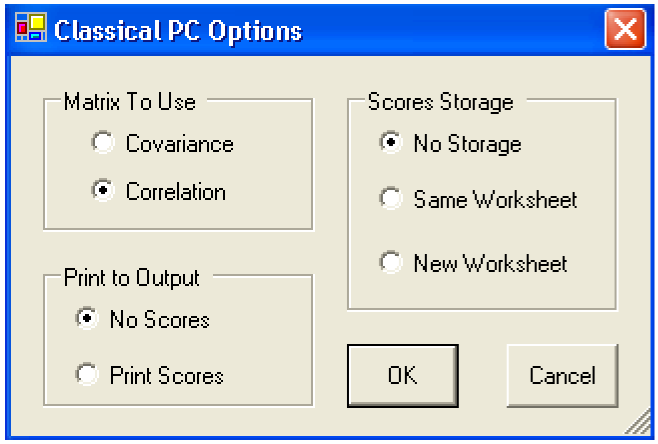
10.1.1 Classical Principal Component Analysis

1. Click on **Multivariate EDA ► PCA ► Classical.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

    - Click on the "**Options**" button for the options window.



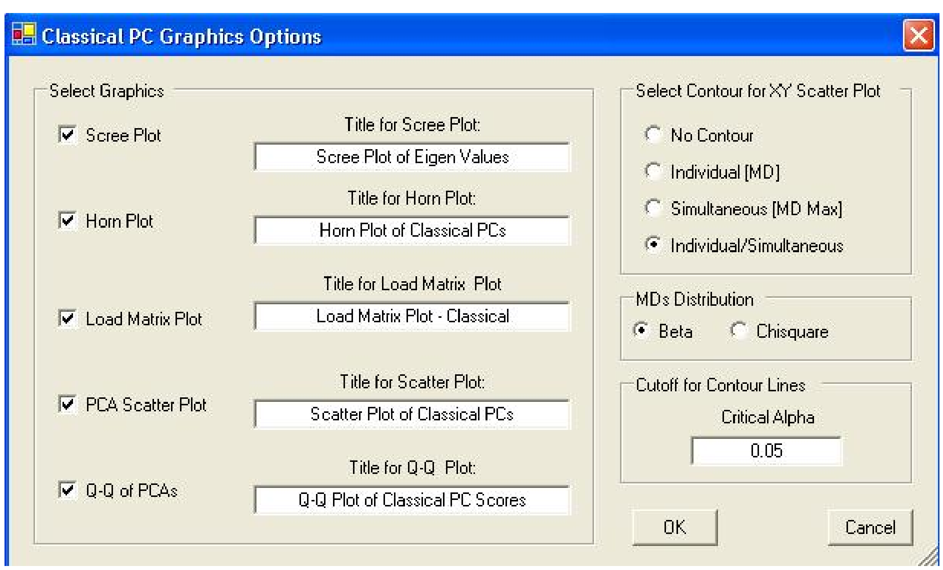
- Specify the storage of principal component scores. No scores will be stored when "**No Storage**" is selected. Scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. Scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Specify the printing of scores in the output in the "**Print to Output**" option. The default is "**No Scores**."

- Specify the "**Matrix To Use**" to compute the principal components. The default is "**Correlation**."

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The “**Scree Plot**” provides a scree plot of the eigen values.

- The “**Horn Plo**t” provides a comparison of the computed eigen values to the multi-normal generated eigen values.

- The “**Load Matrix Plot**” provides the scatter plot of the columns of the load matrix.

- The “**PCA Scatter Plot**” provides the scatter plot of the principal components scores and also the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is “**No Contour**.” Specify the distribution for the distances and the “**Critical Alpha**” value for the cutoff to compute the ellipses. The defaults are “**Beta**” and “**0.05**.”

- The “**Q-Q Plot of PCA**” provides the Q-Q plots of the component scores.

- Click on “**OK**” to continue or “**Cancel**” to cancel the graphics options.

- Click on “**OK**” to continue or “**Cancel**” to cancel the PCA computations.

Output example: The data set “**BUSHFIRE.xls**” was used for the classical PCA. It has 38 observations and five groups. The initial estimate of scale matrix was the classical covariance matrix. The classical correlation matrix was obtained from this covariance matrix and the principal components (eigen values) and the principal component loadings (a matrix of eigen vectors) were obtained from the correlation matrix.

Output for the Classical Principal Component Analysis.
Data Set used: Bushfire.

	Principal Components Analysis using the Classical Method
Date/Time of Computation	1/29/2008 10:40:15 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BushFire
Full Precision	OFF
Display Scores Option	Do not Display PC Scores in Output
PC Scores Storage	Do Not Store Scores to Worksheet
Matrix Used to Compute PCs	Correlation
Graphics	Scree Plot Selected
Scree Plot Title	Scree Plot of Eigen Values
Graphics	Horn Plot Selected
Horn Plot Title	Horn Plot of Classical PCs
Graphics	Load Matrix Plot Selected
Load Matrix Plot Title	Load Matrix Plot - Classical
Graphics	XY Scatter Plot Selected
XY Scatter Plot Title	Scatter Plot of Classical PCs
Contour	No Contour Lines will be Displayed
Graphics	Scores Plot Selected
Scores Plot Title	Q-Q Plot of Classical PC Scores

Summary Statistics	
Number of Observations	38
Number of Selected Variables	5

Mean

Case 1	Case 2	Case 3	Case 4	Case 5
103.6	129.1	288.6	227.9	286.6

Standard Deviation

Case 1	Case 2	Case 3	Case 4	Case 5
20.15	35	177.2	64.06	52.17

Output for the Classical Principal Component Analysis (continued).

	Determinant	1.195E+12		
	Log of Determinant	27.81		
Eigenvalues of Classical Covariance S Matrix				
Eval 1	Eval 2	Eval 3	Eval 4	Eval 5
1.825	48.18	341.6	1035	38435
	Sum of Eigenvalues	39862		

Classical Correlation R Matrix

	Case 1	Case 2	Case 3	Case 4	Case 5
Case 1	1	0.802	-0.585	-0.495	-0.49
Case 2	0.802	1	-0.525	-0.528	-0.516
Case 3	-0.585	-0.525	1	0.974	0.976
Case 4	-0.495	-0.528	0.974	1	0.999
Case 5	-0.49	-0.516	0.976	0.999	1
		Determinant	6.8489E-6		

Eigenvalues of Classical Correlation R Matrix

Eval 1	Eval 2	Eval 3	Eval 4	Eval 5
5.5901E-4	0.0155	0.213	0.979	3.792
	Sum of Eigenvalues	5		

Summary Table (Eigenvalues)

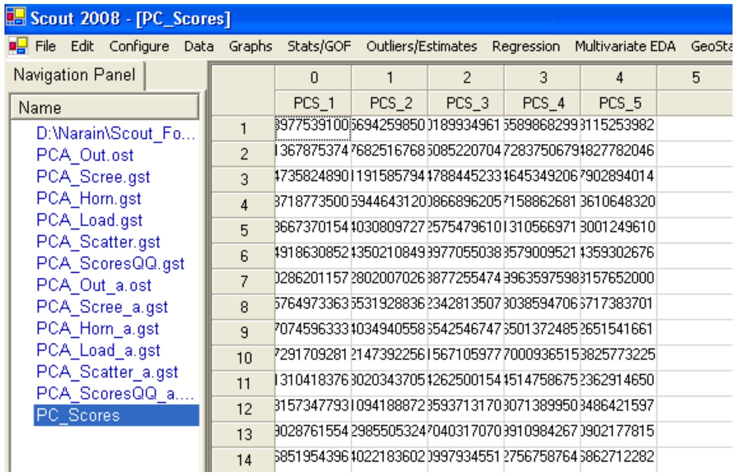
	Eigen Value	Difference	Proportion	Cumulative
PC1	3.792	2.813	0.758	75.84
PC2	0.979	0.766	0.196	95.42
PC3	0.213	0.198	0.0426	99.68
PC4	0.0155	0.0149	0.0031	99.99
PC5	5.5901E-4	N/A	1.1180E-4	100

PC Loadings (Eigen Vectors)

	PC1	PC2	PC3	PC4	PC5
Case 1	-0.383	0.596	0.669	-0.226	0.00614
Case 2	-0.383	0.591	-0.692	0.159	-0.0165
Case 3	0.49	0.267	-0.227	-0.798	-0.0115
Case 4	0.484	0.33	0.119	0.383	-0.704
Case 5	0.482	0.34	0.0927	0.373	0.71

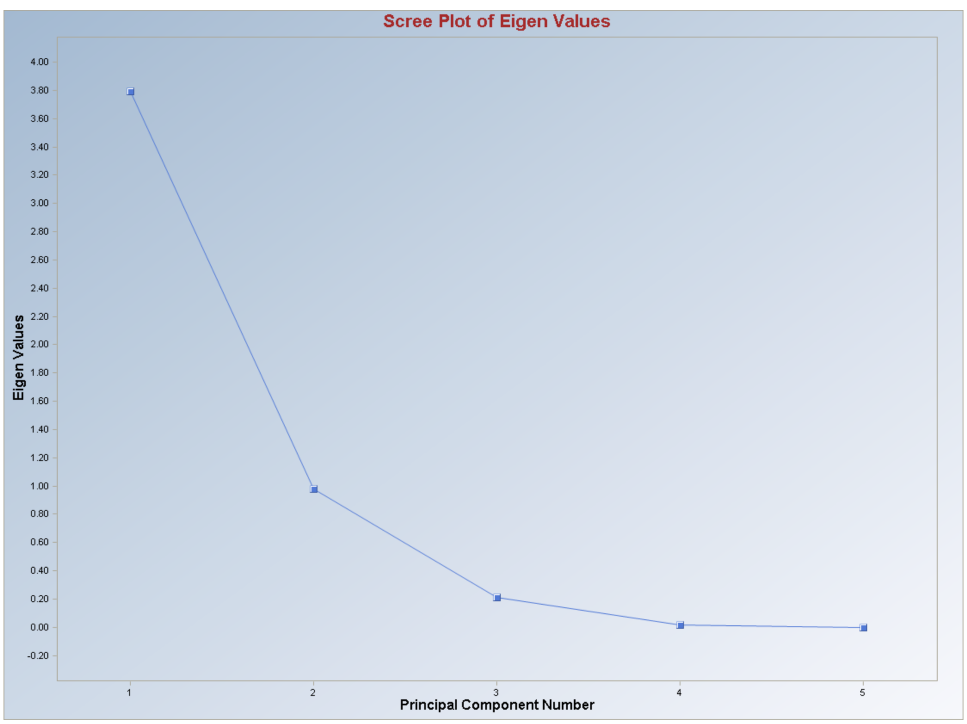
Note: *If the proportion of a principal component is less than 0.01, then that principal component will not be used in the graphing of the load matrix plot, scatter plot of the scores and the Q-Q plots of the scores.*

Output for the Classical Principal Component Analysis (continued).

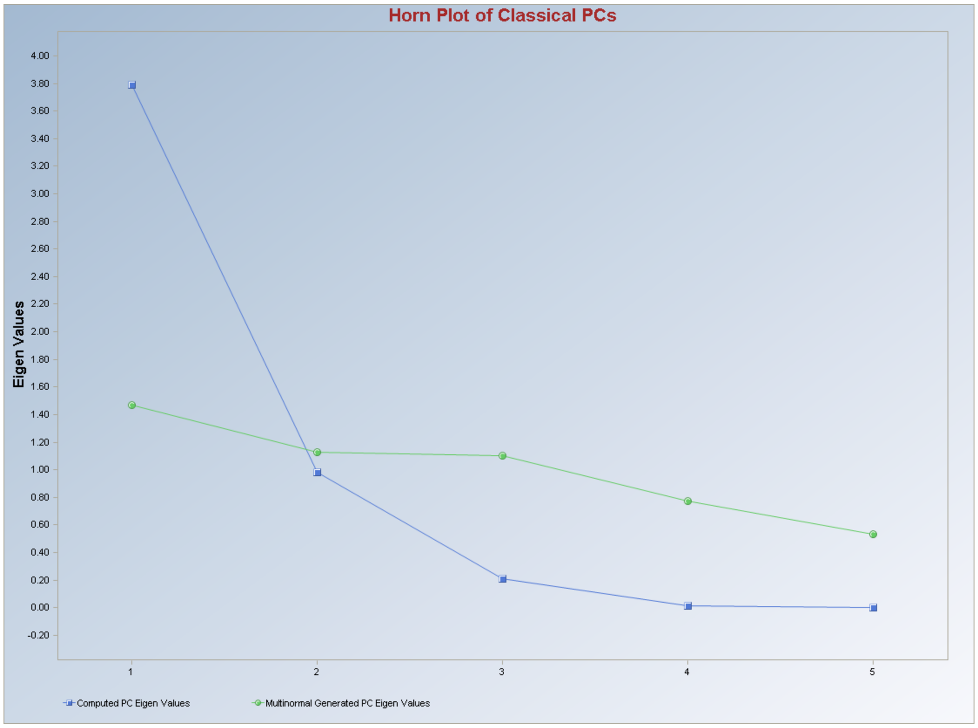


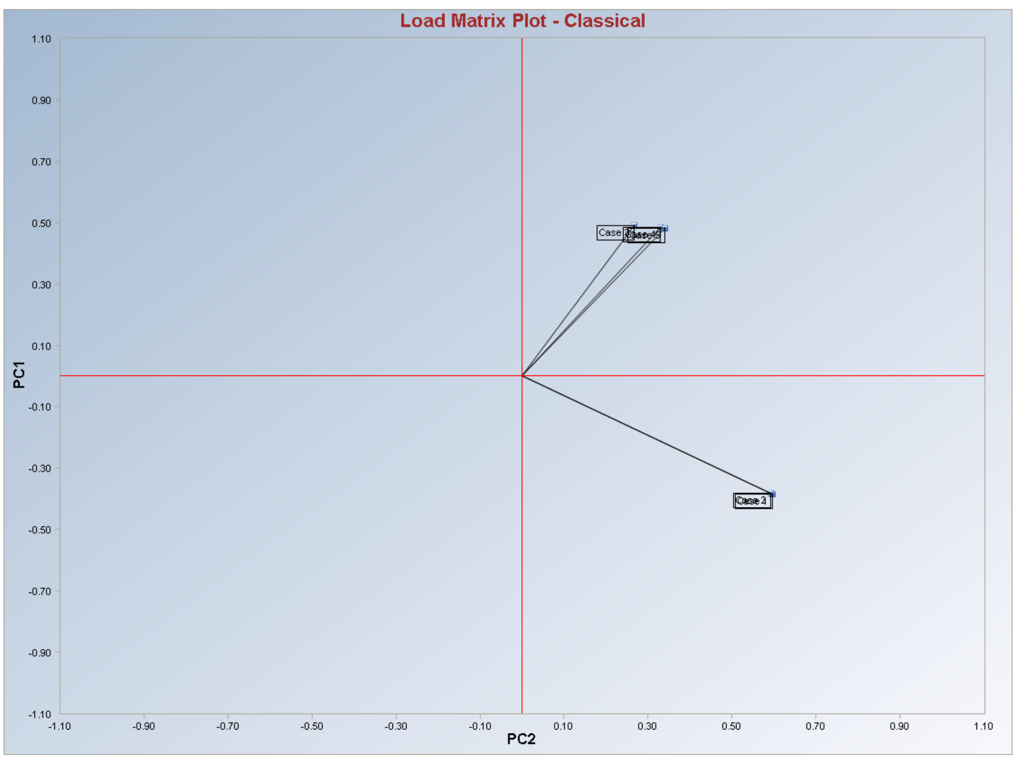
Note: *The scores storage in the "**New Worksheet**" option was chosen in the "**Classical PC Options**" window. This resulted in a new worksheet named **PC_Scores** being generated and the principal component scores being stored in that worksheet. Those scores are available to the user for further computations. The score storage option of PCA remains the same for all of the other PCA procedures incorporated in the principal component module of Scout.*

Output for the Classical Principal Component Analysis.

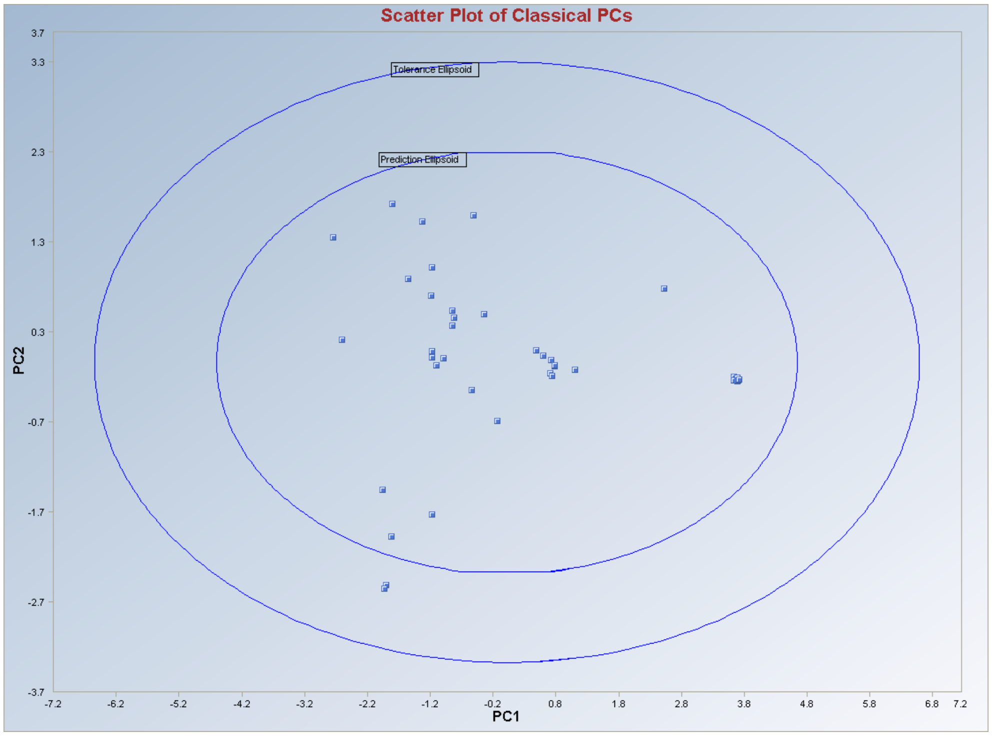


Output for the Classical Principal Component Analysis (continued).

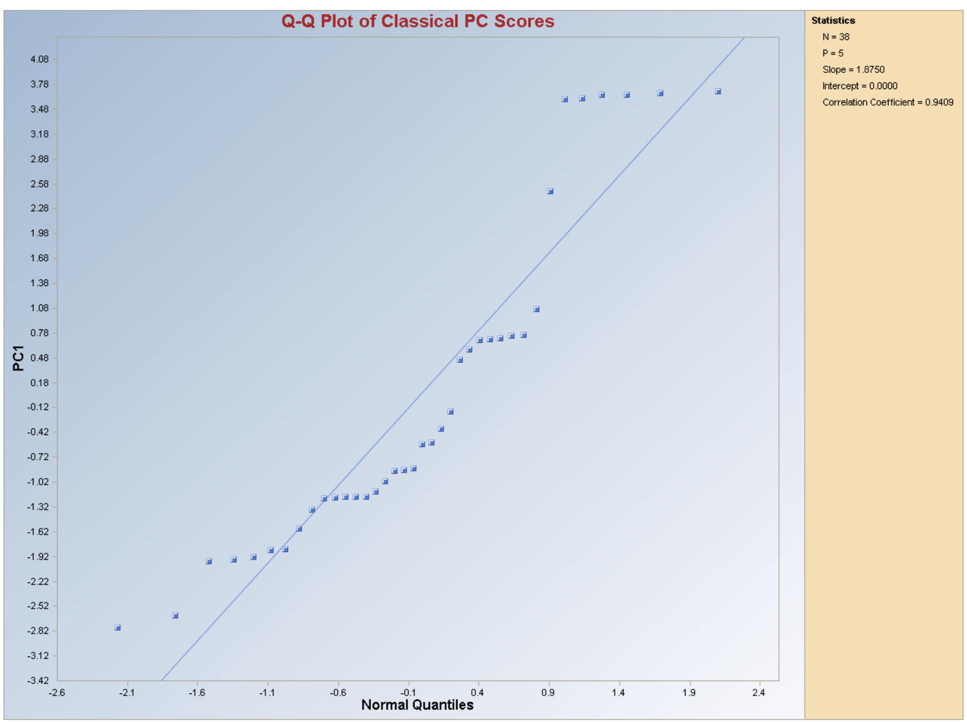




Output for the Classical Principal Component Analysis (continued).



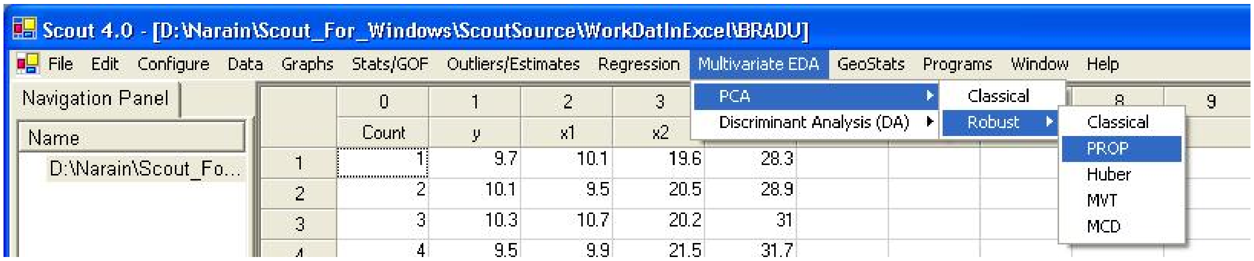
Observations outside of the simultaneous ellipse (tolerance ellipsoid) are considered to be anomalous. Observations between the individual (prediction ellipsoid – inner ellipse) and the simultaneous (tolerance ellipsoid – outer ellipse) ellipses may also represent outliers.



Note: *The drop-down bars in the graphics toolbar can be used to obtain different load matrix plots, scatter plots of the components scores and the selected variables, and the Q-Q plots of the component scores, as explained in Chapter 2.*

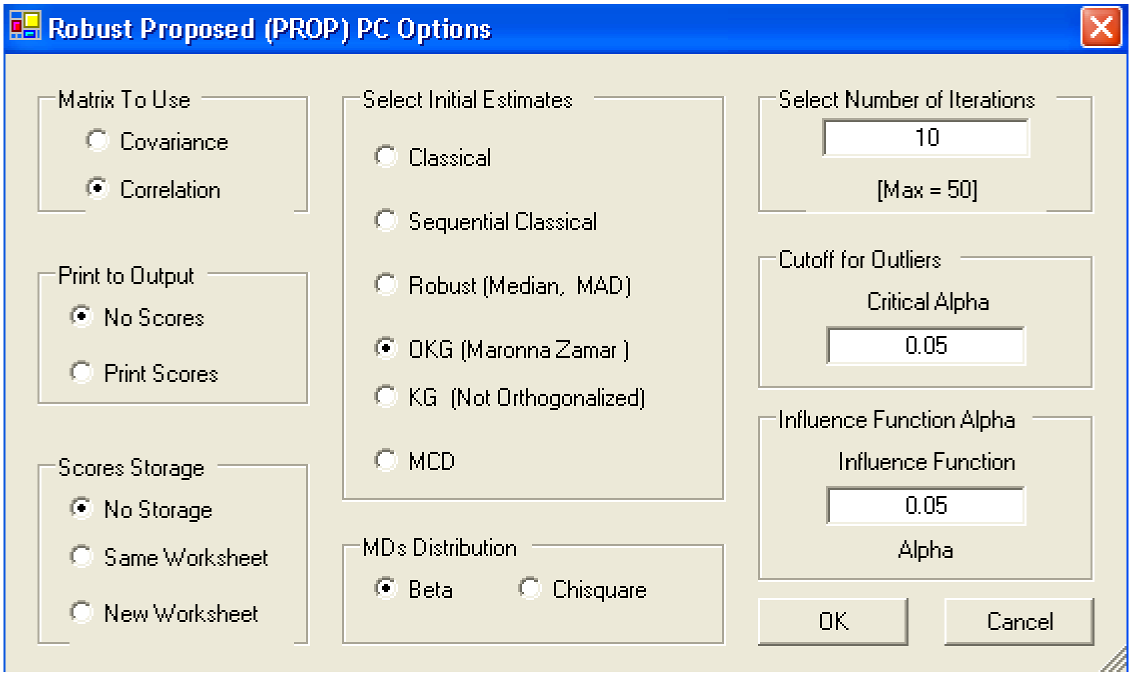
10.1.2 Iterative and Robust Principal Component Analysis

1. Click on **Multivariate EDA ► PCA ► Robust► Sequential Classical, Huber, MVT or PROP.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

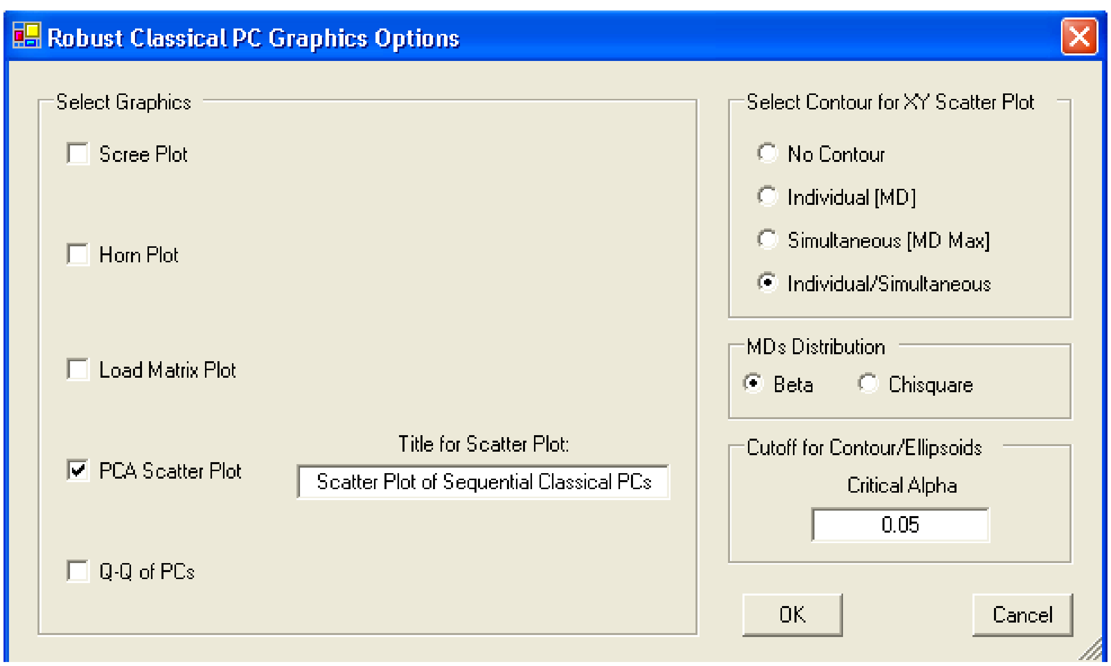
- Click on the "**Options**" button for the options window.



  - Specify the storage of principal component scores. No scores will be stored when "**No Storage**" is selected. Scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. Scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Specify the printing of scores in the output in the "**Print to Output**" option. The default is "**No Scores**."
- Specify the "**Matrix To Use**" to compute the principal components. The default is "**Correlation**."
- Specify the initial estimates. The default is "**OKG (Maronna Zamar)**."
- Specify the distribution for MDs. The default is "**Beta**."
- Specify the number of iterations. The default is "**10**."
- Specify the cutoff for the outliers and the influence function alpha (or trim percentage for MVT). The defaults are "**0.05**" and "**0.05 (0.1 for MVT)**."
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scree Plot**" provides a scree plot of the eigen values.

- The "**Horn Plot**" provides a comparison of the computed eigen values to the multi-normal generated eigen values.

- The "**Load Matrix Plot**" provides the scatter plot of the columns of the load matrix.

- The "**PCA Scatter Plot**" provides the scatter plot of the principal components scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for the distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- The "**Q-Q Plot of PCA**" provides the Q-Q plots of the component scores.

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the robust PCA computations.

10.1.2.1 Sequential Classical PCA

Output example: The data set "**BUSHFIRE.xls**" was used for the sequential classical PCA. It has 38 observations and five groups. The initial estimate of scale matrix was the classical covariance matrix. The outliers were found iteratively and the observations were given weights accordingly. The weighted covariance matrix was calculated. The correlation matrix was obtained from this weighted covariance matrix and the principal components (eigen values) and the principal component loadings (a matrix of eigen vectors) were obtained from the correlation matrix.

Output for the Iterative Sequential Classical Principal Component Analysis.
Data Set used: Bushfire.

	Robust Principal Components Analysis using the Classical Iterative Method
Date/Time of Computation	1/29/2008 11:39:12 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BushFire
Full Precision	OFF
Display Scores Option	Do not Display PC Scores in Output
PC Scores Storage	Do Not Store Scores to Worksheet
Matrix Used to Compute PCs	Correlation
Critical Alpha to Determine Outliers	0.05
Initial Estimates	Robust OKG (Maronna Zamar) Matrix
Number of Iterations	10
Graphics	XY Scatter Plot Selected
XY Scatter Plot Title	Scatter Plot of Sequential Classical PCs
Contour	Contour Ellipses drawn at Individual Beta MD(0.05) and at Max MD(0.05)

Summary Statistics	
Number of Observations	38
Number of Selected Variables	5

Mean

Case 1	Case 2	Case 3	Case 4	Case 5
103.6	129.1	288.6	227.9	286.6

Standard Deviation

Case 1	Case 2	Case 3	Case 4	Case 5
20.15	35	177.2	64.06	52.17

Classical Covariance S Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
406.1	565.4	-2091	-638.7	-515.6
565.4	1225	-3258	-1184	-942.5
-2091	-3258	31405	11060	9021
-638.7	-1184	11060	4103	3340
-515.6	-942.5	9021	3340	2722
		Determinant	1.195E+12	
		Log of Determinant	27.81	

Output for the Sequential Classical Principal Component Analysis (continued).

Initial Robust OKG (Maronna Zamar) Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
427	652.6	1014	344.6	177.4
652.6	1826	3306	802.7	585.5
1014	3306	20637	3455	3206
344.6	802.7	3455	1597	857.6
177.4	585.5	3206	857.6	735.7
		Determinant	6.282E+14	
		Log of Determinant	34.07	

genvalues of Initial Robust OKG (Maronna Zamar) Covariance S Ma				
Case 1	Case 2	Case 3	Case 4	Case 5
104.6	177.6	954	1581	22405

Initial Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
1	0.739	0.342	0.417	0.316
0.739	1	0.539	0.47	0.505
0.342	0.539	1	0.602	0.823
0.417	0.47	0.602	1	0.791
0.316	0.505	0.823	0.791	1
		Determinant	0.0332	

Eigen Values of Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
0.111	0.216	0.425	1.012	3.236

Final Mean Vector				
Case 1	Case 2	Case 3	Case 4	Case 5
107.5	141.9	221.7	201.4	265.3

Final Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
337.8	315.1	-961	-140.2	-115.4
315.1	510.8	713.4	410.9	346
-961	713.4	16189	4712	3922
-140.2	410.9	4712	1529	1271
-115.4	346	3922	1271	1060
		Determinant	2.038E+10	

Output for the Sequential Classical Principal Component Analysis (continued).

Final Correlation R Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
1	0.759	-0.411	-0.195	-0.193
0.759	1	0.248	0.465	0.47
-0.411	0.248	1	0.947	0.947
-0.195	0.465	0.947	1	0.998
-0.193	0.47	0.947	0.998	1
		Determinant	4.5043E-6	

Eigenvalues for Final Correlation R Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
0.00153	0.0156	0.0334	1.779	3.17

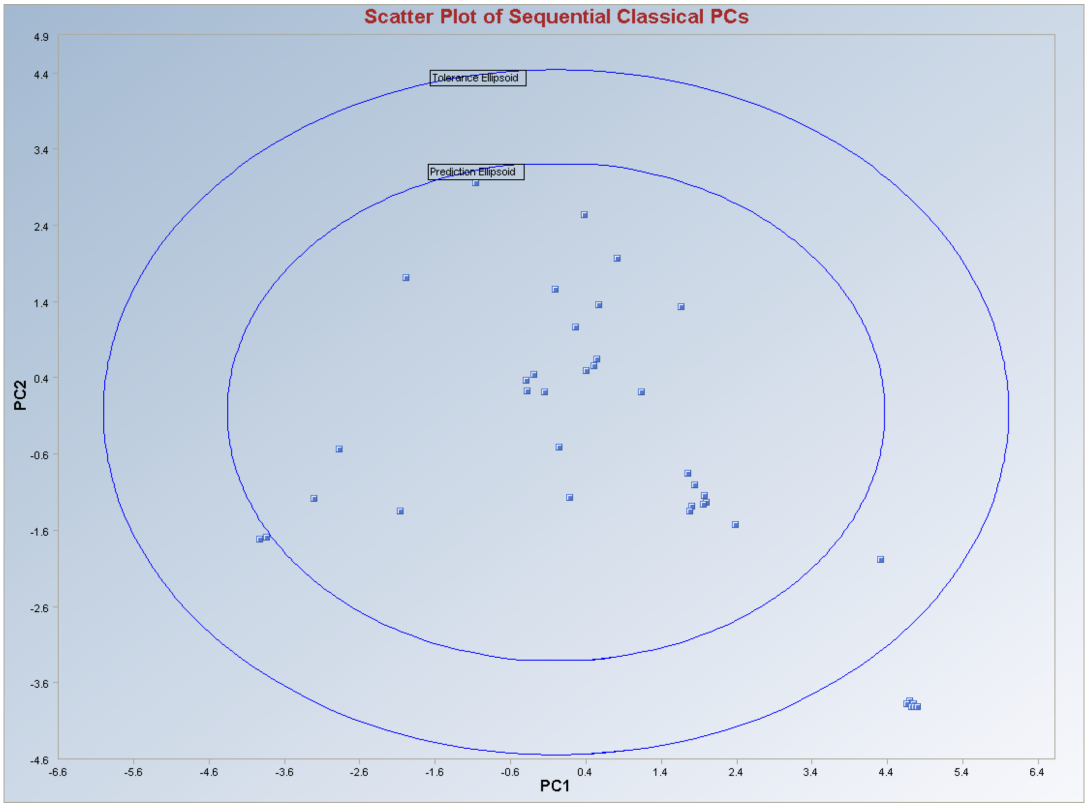
Summary Table (Eigen Values)

	Eigen Value	Difference	Proportion	Cumulative
PC1	3.17	1.391	0.634	63.4
PC2	1.779	1.746	0.356	98.99
PC3	0.0334	0.0178	0.00668	99.66
PC4	0.0156	0.014	0.00311	99.97
PC5	0.00153	N/A	3.0684E-4	100

Load Matrix (Eigen Vectors)

	PC1	PC2	PC3	PC4	PC5
Case 1	-0.11	0.732	-0.141	0.653	-0.0691
Case 2	0.265	0.658	-0.0606	-0.698	0.0786
Case 3	0.54	-0.175	-0.816	0.11	-0.00554
Case 4	0.56	-7.677E-4	0.4	0.253	0.68
Case 5	0.56	0.00216	0.388	0.0989	-0.725

Output for the Sequential Classical Principal Component Analysis (continued).



Observations outside the tolerance ellipse are considered to be anomalous. Observations between the prediction and the tolerance ellipses are observations with reduced (but > 0) weights. Those observations may represent potential outliers needing further investigation.

Note*: The drop-down bars in the graphics toolbar can be used to obtain different load matrix plots, scatter plots of components scores and selected variables, and Q-Q plots of the component scores, as explained in Chapter 2.*

10.1.2.2 Huber PCA

Output example: The data set "**BUSHFIRE.xls**" was used for the Huber PCA. It has 38 observations and five groups. The initial estimate of scale matrix was the classical covariance matrix. The outliers were found iteratively using the Huber influence function and the observations were given weights accordingly. The weighted covariance matrix was calculated. The correlation matrix was obtained from this weighted covariance matrix and the principal components (eigen values) and the principal component loadings (a matrix of eigen vectors) were obtained from the correlation matrix.

Output for the Principal Component Analysis Based Upon the Huber Influence Function.
Data Set used: Bushfire.

	Robust Principal Components Analysis using the Huber Influence Function
Date/Time of Computation	1/29/2008 11:48:33 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BushFire
Full Precision	OFF
Display Scores Option	Do not Display PC Scores in Output
PC Scores Storage	Do Not Store Scores to Worksheet
Matrix Used to Compute PCs	Correlation
Distributional Squared MDs	Beta Distribution
Influence Function Alpha	0.05
Initial Estimates	Robust OKG (Maronna Zamar) Matrix
Number of Iterations	10
Graphics	XY Scatter Plot Selected
XY Scatter Plot Title	Scatter Plot of Huber PCs
Contour	Contour Ellipses drawn at Individual Beta MD(0.05) and at Max MD(0.05)

Summary Statistics	
Number of Observations	38
Number of Selected Variables	5

Mean

Case 1	Case 2	Case 3	Case 4	Case 5
103.6	129.1	288.6	227.9	286.6

Standard Deviation

Case 1	Case 2	Case 3	Case 4	Case 5
20.15	35	177.2	64.06	52.17

Classical Covariance S Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
406.1	565.4	-2091	-638.7	-515.6
565.4	1225	-3258	-1184	-942.5
-2091	-3258	31405	11060	9021
-638.7	-1184	11060	4103	3340
-515.6	-942.5	9021	3340	2722

Determinant	1.195E+12
Log of Determinant	27.81

Output for the Principal Component Analysis Based Upon the Huber Influence Function (continued).

Initial Robust OKG (Maronna Zamar) Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
427	652.6	1014	344.6	177.4
652.6	1826	3306	802.7	585.5
1014	3306	20637	3455	3206
344.6	802.7	3455	1597	857.6
177.4	585.5	3206	857.6	735.7
		Determinant	6.282E+14	
		Log of Determinant	34.07	

genvalues of Initial Robust OKG (Maronna Zamar) Covariance S Ma				
Case 1	Case 2	Case 3	Case 4	Case 5
104.6	177.6	954	1581	22405

Initial Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
1	0.739	0.342	0.417	0.316
0.739	1	0.539	0.47	0.505
0.342	0.539	1	0.602	0.823
0.417	0.47	0.602	1	0.791
0.316	0.505	0.823	0.791	1
		Determinant	0.0332	

Eigen Values of Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
0.111	0.216	0.425	1.012	3.236

Final Mean Vector				
Case 1	Case 2	Case 3	Case 4	Case 5
103.8	129.8	294.1	230.1	288.5

Final Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
417.9	575.1	-2274	-704.5	-569.9
575.1	1232	-3704	-1365	-1092
-2274	-3704	30006	10416	8473
-704.5	-1365	10416	3808	3089
-569.9	-1092	8473	3089	2509
		Determinant	7.753E+11	

Output for the Principal Component Analysis Based Upon the Huber Influence Function (continued).

Final Correlation R Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
1	0.802	-0.642	-0.558	-0.557
0.802	1	-0.609	-0.63	-0.621
-0.642	-0.609	1	0.974	0.977
-0.558	-0.63	0.974	1	0.999
-0.557	-0.621	0.977	0.999	1
		Determinant	5.2523E-6	

Eigenvalues for Final Correlation R Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
6.0815E-4	0.0127	0.215	0.8	3.972

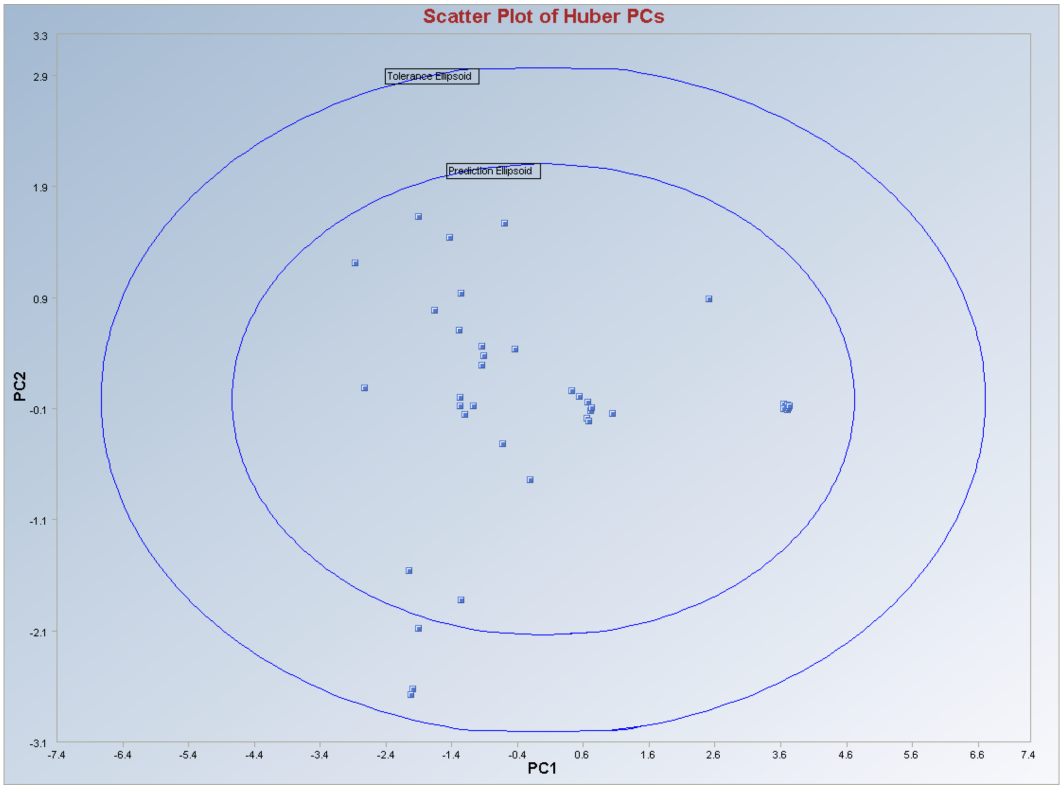
Summary Table (Eigen Values)

	Eigen Value	Difference	Proportion	Cumulative
PC1	3.972	3.173	0.794	79.45
PC2	0.8	0.585	0.16	95.44
PC3	0.215	0.202	0.043	99.73
PC4	0.0127	0.012	0.00253	99.99
PC5	6.0815E-4	N/A	1.2163E-4	100

Load Matrix (Eigen Vectors)

	PC1	PC2	PC3	PC4	PC5
Case 1	-0.391	0.615	0.643	-0.234	0.00221
Case 2	-0.404	0.552	-0.705	0.185	-0.012
Case 3	0.48	0.28	-0.263	-0.788	-0.026
Case 4	0.476	0.342	0.11	0.397	-0.697
Case 5	0.476	0.35	0.0842	0.362	0.716

Output for the Principal Component Analysis Based Upon the Huber Influence Function (continued).



Observations outside of the simultaneous tolerance ellipse are considered to be anomalous. Observations between the individual prediction ellipsoid and the simultaneous tolerance ellipsoid received reduced weights (< 1) and may also represent potential outliers.

***Note**: The drop-down bars in the graphics toolbar can be used to obtain the different load matrix plots, scatter plots of components scores and the variables and the Q-Q plots of the component scores, as explained in Chapter 2.*

10.1.2.3 Multivariate Trimming PCA

Output example: The data set “**BUSHFIRE.xls**” was used for the MVT PCA. It has 38 observations and five groups. The initial estimate of scale matrix was the classical covariance matrix. The outliers were found iteratively using the trimming percentage and a critical alpha and the observations were given weights accordingly. The weighted covariance matrix was calculated. The correlation matrix was obtained from this weighted covariance matrix and the principal components (eigen values) and the principal component loadings (a matrix of eigen vectors) were obtained from the correlation matrix.

Output for the Principal Component Analysis Based Upon the MVT Method.
Data Set used: Bushfire.

	Robust Principal Components Analysis using the MVT Method
Date/Time of Computation	1/29/2008 11:54:09 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BushFire
Full Precision	OFF
Display Scores Option	Do not Display PC Scores in Output
PC Scores Storage	Do Not Store Scores to Worksheet
Matrix Used to Compute PCs	Correlation
Trimming Percentage	10%
Critical Alpha to Determine Outliers	0.05 (planned to be used for verification of trimming non-outliers
Initial Estimates	Robust OKG (Maronna Zamar) Matrix
Number of Iterations	10
Graphics	XY Scatter Plot Selected
XY Scatter Plot Title	Scatter Plot of MVT PCs
Contour	Contour Ellipses drawn at Individual Beta MD(0.05) and at Max MD(0.05)

Summary Statistics				
Number of Observations			38	
Number of Selected Variables			5	
Mean				
Case 1	Case 2	Case 3	Case 4	Case 5
103.6	129.1	288.6	227.9	286.6
Standard Deviation				
Case 1	Case 2	Case 3	Case 4	Case 5
20.15	35	177.2	64.06	52.17
Classical Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
406.1	565.4	-2091	-638.7	-515.6
565.4	1225	-3258	-1184	-942.5
-2091	-3258	31405	11060	9021
-638.7	-1184	11060	4103	3340
-515.6	-942.5	9021	3340	2722
		Determinant	1.195E+12	
		Log of Determinant	27.81	

Output for the Principal Component Analysis Based Upon the MVT Method (continued).

Initial Robust OKG (Maronna Zamar) Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
427	652.6	1014	344.6	177.4
652.6	1826	3306	802.7	585.5
1014	3306	20637	3455	3206
344.6	802.7	3455	1597	857.6
177.4	585.5	3206	857.6	735.7
		Determinant	6.282E+14	
		Log of Determinant	34.07	

genvalues of Initial Robust OKG (Maronna Zamar) Covariance S Ma				
Case 1	Case 2	Case 3	Case 4	Case 5
104.6	177.6	954	1581	22405

Initial Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
1	0.739	0.342	0.417	0.316
0.739	1	0.539	0.47	0.505
0.342	0.539	1	0.602	0.823
0.417	0.47	0.602	1	0.791
0.316	0.505	0.823	0.791	1
		Determinant	0.0332	

Eigen Values of Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
0.111	0.216	0.425	1.012	3.236

Final Mean Vector				
Case 1	Case 2	Case 3	Case 4	Case 5
104.4	131.6	310.3	236.3	293.7

Final Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
431.9	587.1	-2523	-789.4	-639.8
587.1	1245	-4266	-1582	-1272
-2523	-4266	27995	9621	7800
-789.4	-1582	9621	3479	2810
-639.8	-1272	7800	2810	2272
		Determinant	2.729E+11	

Output for the Principal Component Analysis Based Upon the MVT Method (continued).

Final Correlation R Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
1	0.801	-0.726	-0.644	-0.646
0.801	1	-0.722	-0.76	-0.756
-0.726	-0.722	1	0.975	0.978
-0.644	-0.76	0.975	1	0.999
-0.646	-0.756	0.978	0.999	1
		Determinant	2.2922E-6	

Eigenvalues for Final Correlation R Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
6.1666E-4	0.0074	0.212	0.563	4.218

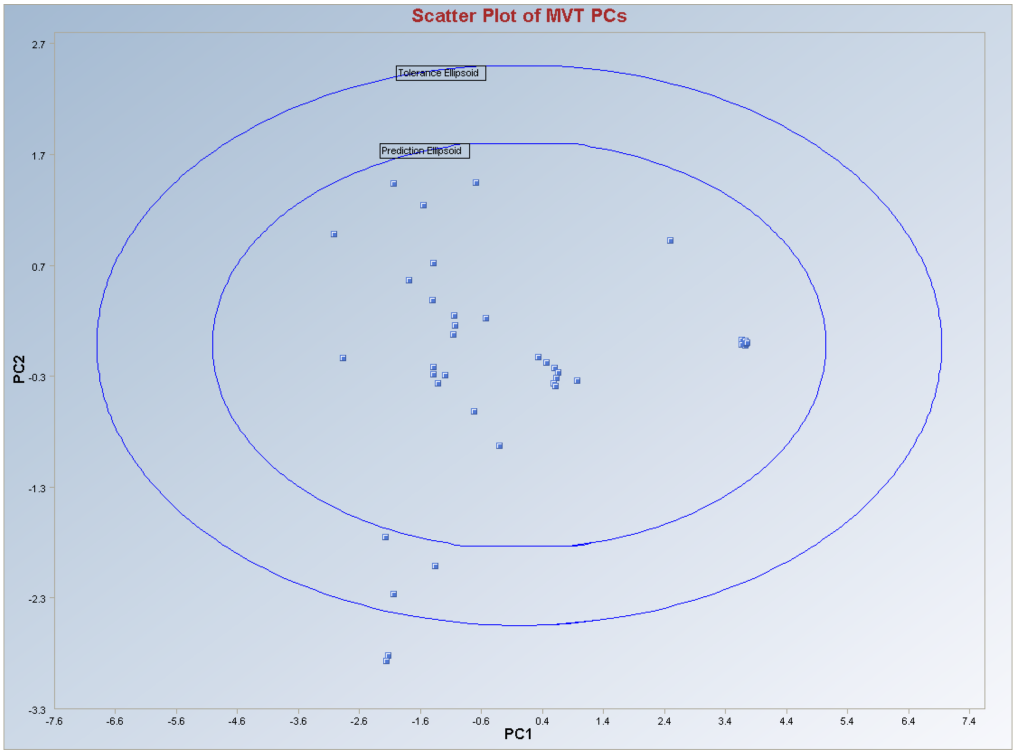
Summary Table (Eigen Values)

	Eigen Value	Difference	Proportion	Cumulative
PC1	4.218	3.655	0.844	84.36
PC2	0.563	0.351	0.113	95.61
PC3	0.212	0.204	0.0423	99.84
PC4	0.0074	0.00679	0.00148	99.99
PC5	6.1666E-4	N/A	1.2333E-4	100

Load Matrix (Eigen Vectors)

	PC1	PC2	PC3	PC4	PC5
Case 1	-0.4	0.678	0.567	-0.244	-0.0152
Case 2	-0.426	0.456	-0.75	0.221	0.0075
Case 3	0.47	0.273	-0.328	-0.769	-0.0822
Case 4	0.468	0.358	0.0782	0.451	-0.665
Case 5	0.468	0.361	0.0531	0.312	0.742

Output for the Principal Component Analysis Based Upon the MVT Methods (continued).



Observations outside of the simultaneous ellipse are considered to be outlying. Observations between the individual and the simultaneous ellipses receiving reduced weights may also be considered to be discordant.

Note: *The drop-down bars in the graphics toolbar can be used to obtain different load matrix plots, scatter plots of components scores and selected variables, and the Q-Q plots of the component scores, as explained in Chapter 2.*

10.1.2.4 PROP PCA

Output example: The data set "**BUSHFIRE.xls**" was used for the PROP PCA. It has 38 observations and five groups. The initial estimate of scale matrix was the classical covariance matrix. The outliers were found iteratively using the PROP influence function and the observations were given weights accordingly. The weighted covariance matrix was calculated. The correlation matrix was obtained from this weighted covariance matrix and the principal components (eigen values) and the principal component loadings (a matrix of eigen vectors) were obtained from the correlation matrix.

Output for the Principal Component Analysis Based Upon the PROP Influence Function.
Data Set used: Bushfire.

	Robust Principal Components Analysis using the PROP Influence Function
Date/Time of Computation	1/29/2008 12:12:42 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BushFire
Full Precision	OFF
Display Scores Option	Do not Display PC Scores in Output
PC Scores Storage	Do Not Store Scores to Worksheet
Matrix Used to Compute PCs	Correlation
Distributional Squared MDs	Beta Distribution
Influence Function Alpha	0.05
Initial Estimates	Robust OKG (Maronna Zamar) Matrix
Number of Iterations	10
Graphics	XY Scatter Plot Selected
XY Scatter Plot Title	Scatter Plot of PROP PCs
Contour	Contour Ellipses drawn at Individual Beta MD(0.05) and at Max MD(0.05)

Summary Statistics	
Number of Observations	38
Number of Selected Variables	5

Mean

Case 1	Case 2	Case 3	Case 4	Case 5
103.6	129.1	288.6	227.9	286.6

Standard Deviation

Case 1	Case 2	Case 3	Case 4	Case 5
20.15	35	177.2	64.06	52.17

Classical Covariance S Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
406.1	565.4	-2091	-638.7	-515.6
565.4	1225	-3258	-1184	-942.5
-2091	-3258	31405	11060	9021
-638.7	-1184	11060	4103	3340
-515.6	-942.5	9021	3340	2722

Determinant	1.195E+12
Log of Determinant	27.81

Output for the Principal Component Analysis Based Upon the PROP Influence Function (continued).

Initial Robust OKG (Maronna Zamar) Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
427	652.6	1014	344.6	177.4
652.6	1826	3306	802.7	585.5
1014	3306	20637	3455	3206
344.6	802.7	3455	1597	857.6
177.4	585.5	3206	857.6	735.7
		Determinant	6.282E+14	
		Log of Determinant	34.07	

genvalues of Initial Robust OKG (Maronna Zamar) Covariance S Ma				
Case 1	Case 2	Case 3	Case 4	Case 5
104.6	177.6	954	1581	22405

Initial Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
1	0.739	0.342	0.417	0.316
0.739	1	0.539	0.47	0.505
0.342	0.539	1	0.602	0.823
0.417	0.47	0.602	1	0.791
0.316	0.505	0.823	0.791	1
		Determinant	0.0332	

Eigen Values of Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
0.111	0.216	0.425	1.012	3.236

Final Mean Vector				
Case 1	Case 2	Case 3	Case 4	Case 5
104.6	146.1	275.2	217.7	279.2

Final Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
280.4	213.6	-1449	-326.5	-264.7
213.6	187.5	-956.1	-195.2	-163.6
-1449	-956.1	8688	2136	1695
-326.5	-195.2	2136	563	439.2
-264.7	-163.6	1695	439.2	345.4
		Determinant	33022620	

Output for the Principal Component Analysis Based Upon the PROP Influence Function (continued).

Final Correlation R Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
1	0.931	-0.929	-0.822	-0.851
0.931	1	-0.749	-0.601	-0.643
-0.929	-0.749	1	0.966	0.979
-0.822	-0.601	0.966	1	0.996
-0.851	-0.643	0.979	0.996	1
		Determinant	3.7184E-7	

Eigenvalues for Final Correlation R Matix

Case 1	Case 2	Case 3	Case 4	Case 5
0.00156	0.00427	0.0221	0.571	4.401

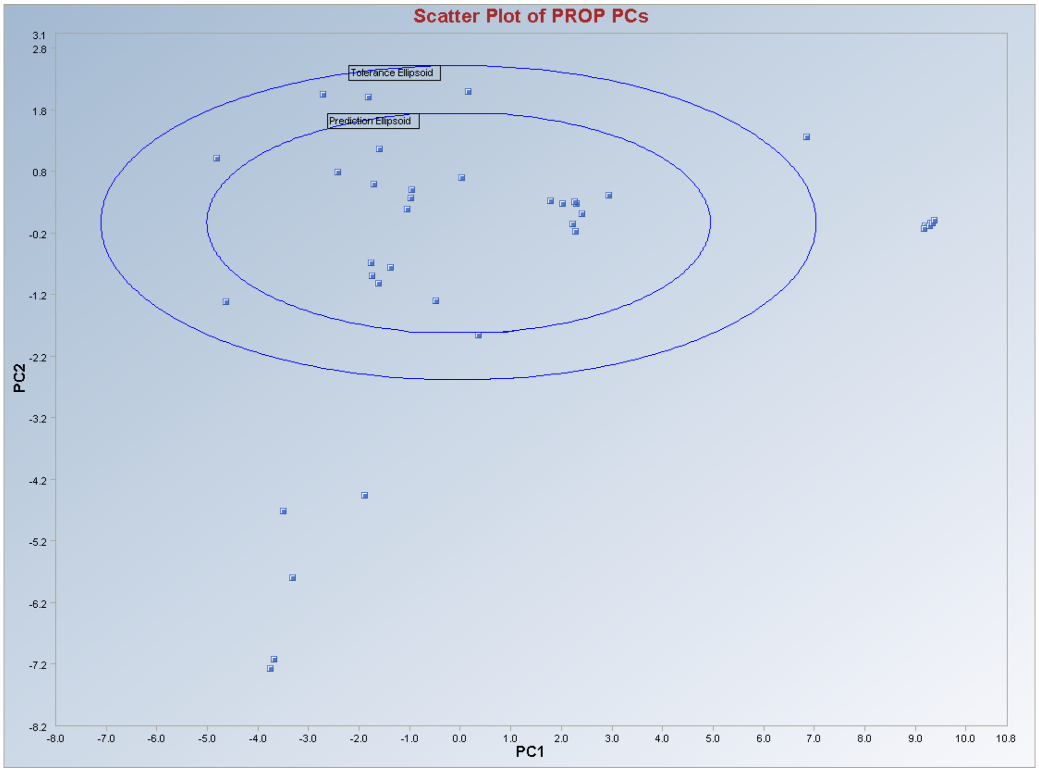
Summary Table (Eigen Values)

	Eigen Value	Difference	Proportion	Cumulative
PC1	4.401	3.829	0.88	88.01
PC2	0.571	0.549	0.114	99.44
PC3	0.0221	0.0179	0.00443	99.88
PC4	0.00427	0.00271	8.5466E-4	99.97
PC5	0.00156	N/A	3.1278E-4	100

Load Matrix (Eigen Vectors)

	PC1	PC2	PC3	PC4	PC5
Case 1	-0.46	0.33	0.54	-0.531	-0.326
Case 2	-0.395	0.732	-0.493	0.197	0.16
Case 3	0.472	0.159	-0.505	-0.564	-0.423
Case 4	0.449	0.439	0.354	0.523	-0.455
Case 5	0.457	0.371	0.291	-0.296	0.694

Output for the Principal Component Analysis Based Upon the PROP Influence Function (continued).

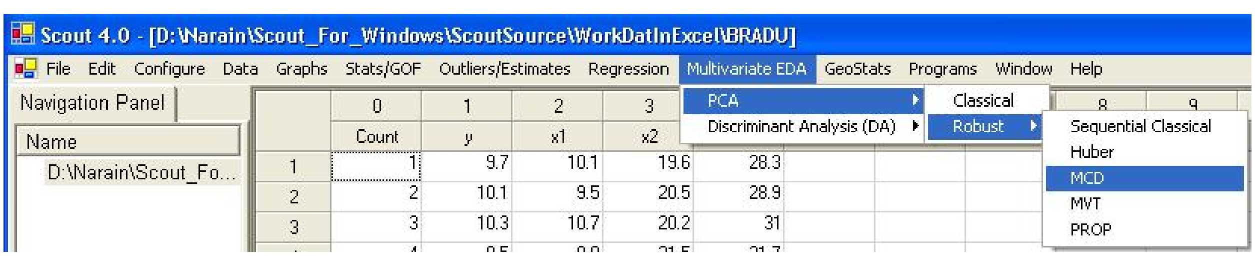


Observations outside of the simultaneous (tolerance) ellipsoid are considered to be outliers. Observations (if any) between the individual (prediction ellipsoid) and the simultaneous (tolerance) ellipses received reduced (< 1) weights and may represent potential intermediate outliers.

Note: *The drop-down bars in the graphics toolbar can be used to obtain different load matrix plots, scatter plots of principal components scores and selected variables, and the Q-Q plots of the component scores, as explained in Chapter 2.*

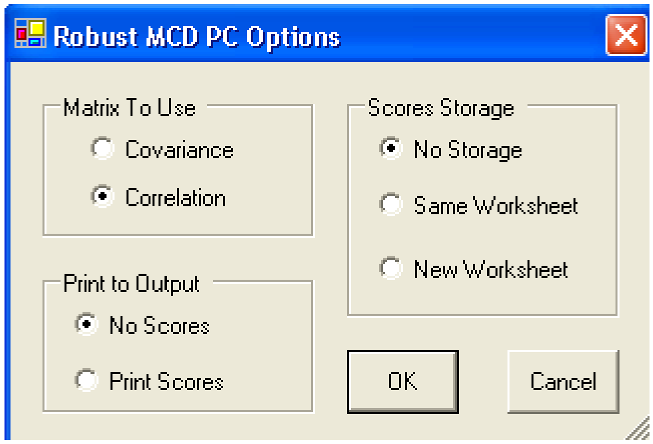
10.1.2.5 Minimum Covariance Determinant PCA

1. Click on **Multivariate EDA ► PCA ► Robust ► MCD.**



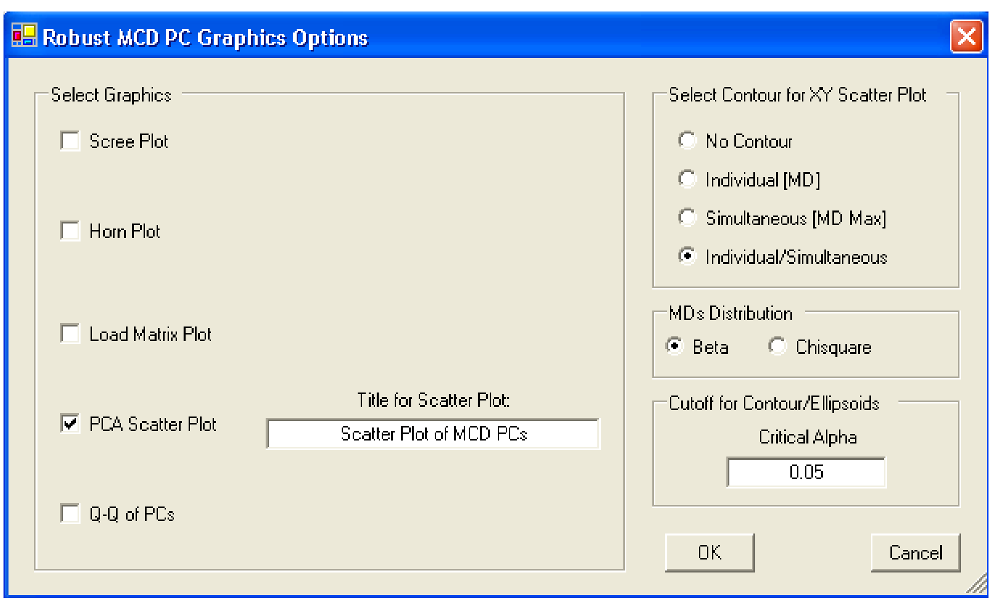
2. The "**Select Variables**" screen (Section 3.4) will appear.

- Click on the "**Options**" button for the options window.



  - Specify storage of the principal component scores. The default is "**No Storage**."
  - Specify the "**Matrix To Use**" to compute the principal components. The default is "**Correlation**."
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scree Plot**" provides a scree plot of the eigen values.

- The "**Horn Plot**" provides a comparison of computed eigen values to the multi-normal generated eigen values.

- The "**Load Matrix Plot**" provides the scatter plot of the columns of the load matrix.

- The "**PCA Scatter Plot**" provides the scatter plot of the principal components scores and also the selected variables. The user has the option of drawing contours on the scatter plot to identify outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- The "**Q-Q Plot of PCA**" provides the Q-Q plots of the component scores.

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

* Click on "**OK**" to continue or "**Cancel**" to cancel the robust PCA computations.

Output example: The data set "**BUSHFIRE.xls**" was used for the MCD PCA. It has 38 observations and five groups. The MCD estimate of scale was calculated. The correlation matrix was obtained from this MCD covariance matrix and the principal components (eigen values) and the principal component loadings (a matrix of eigen vectors) were obtained from the correlation matrix.

Output for the MCD Principal Component Analysis.
Data Set used: Bushfire.

	Principal Components Analysis using the MCD Method
Date/Time of Computation	1/29/2008 12:19:48 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BushFire
Full Precision	OFF
Display Scores Option	Do not Display PC Scores in Output
PC Scores Storage	Do Not Store Scores to Worksheet
Matrix Used to Compute PCs	Correlation
Graphics	XY Scatter Plot Selected
XY Scatter Plot Title	Scatter Plot of MCD PCs
Contour	Contour Ellipses drawn at Individual Beta MD(0.05) and at Max MD(0.05)

Summary Statistics	
Number of Observations	38
Number of Selected Variables	5

Mean

Case 1	Case 2	Case 3	Case 4	Case 5
103.6	129.1	288.6	227.9	286.6

Standard Deviation

Case 1	Case 2	Case 3	Case 4	Case 5
20.15	35	177.2	64.06	52.17

Covariance S Matrix

Case 1	Case 2	Case 3	Case 4	Case 5
406.1	565.4	-2091	-638.7	-515.6
565.4	1225	-3258	-1184	-942.5
-2091	-3258	31405	11060	9021
-638.7	-1184	11060	4103	3340
-515.6	-942.5	9021	3340	2722
		Determinant	1.195E+12	
		Log of Determinant	27.81	

MCD Mean

Case 1	Case 2	Case 3	Case 4	Case 5
105.5	146.9	274.4	217.5	279

Output for the MCD Principal Component Analysis (continued).

MCD Covariance S Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
287.9	222.8	-1408	-316.7	-258.4
222.8	196.6	-936	-191.2	-161.6
-1408	-936	8314	2043	1623
-316.7	-191.2	2043	538.1	420.3
-258.4	-161.6	1623	420.3	331
		Determinant	75211116	
		Log of Determinant	18.14	

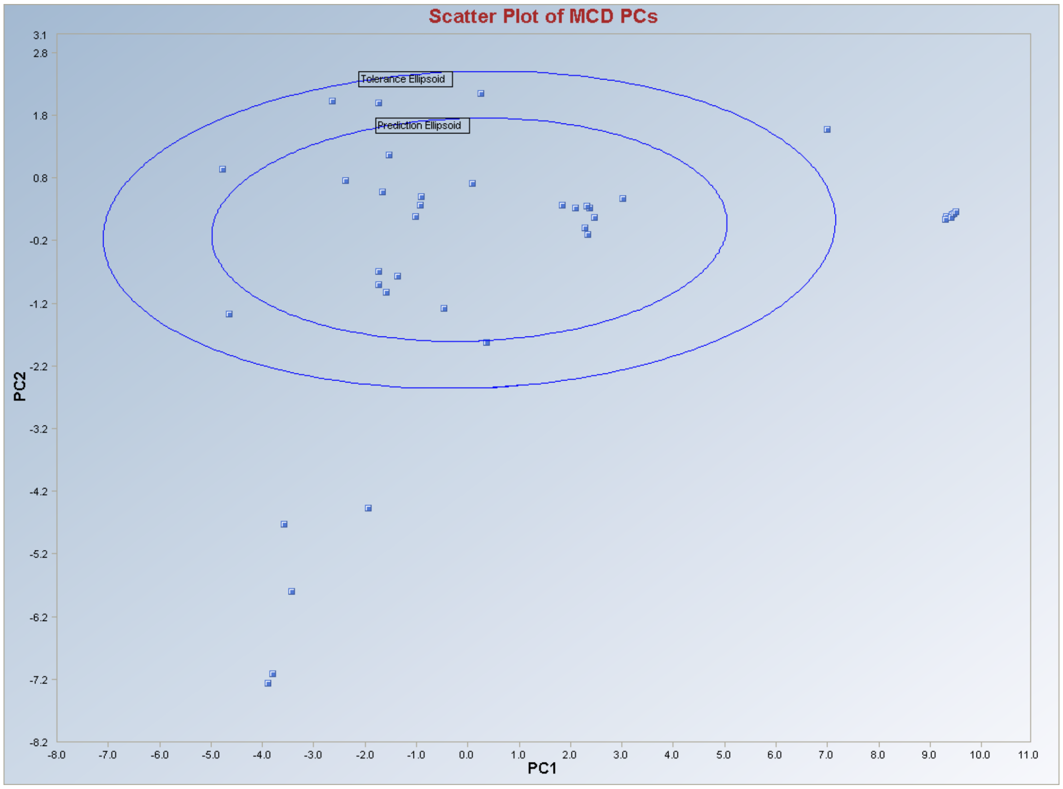
MCD Correlation R Matrix				
Case 1	Case 2	Case 3	Case 4	Case 5
1	0.936	-0.91	-0.805	-0.837
0.936	1	-0.732	-0.588	-0.634
-0.91	-0.732	1	0.966	0.979
-0.805	-0.588	0.966	1	0.996
-0.837	-0.634	0.979	0.996	1
		Determinant	8.9759E-7	

Eigenvalues for MCD Correlation R Matrix				
Eval 1	Eval 2	Eval 3	Eval 4	Eval 5
0.00217	0.00735	0.0214	0.602	4.367

Summary Table (Eigen Values)				
	Eigen Value	Difference	Proportion	Cumulative
PC1	4.367	3.766	0.873	87.35
PC2	0.602	0.58	0.12	99.38
PC3	0.0214	0.0141	0.00428	99.81
PC4	0.00735	0.00518	0.00147	99.96
PC5	0.00217	N/A	4.3397E-4	100

PC Load Matrix (Eigen Vectors)					
	PC1	PC2	PC3	PC4	PC5
Case 1	-0.458	0.351	0.482	0.65	0.111
Case 2	-0.395	0.723	-0.47	-0.305	-0.089
Case 3	0.472	0.176	-0.567	0.628	0.176
Case 4	0.449	0.436	0.37	-0.299	0.618
Case 5	0.458	0.365	0.298	0.0339	-0.753

Output for the MCD Principal Component Analysis (continued).



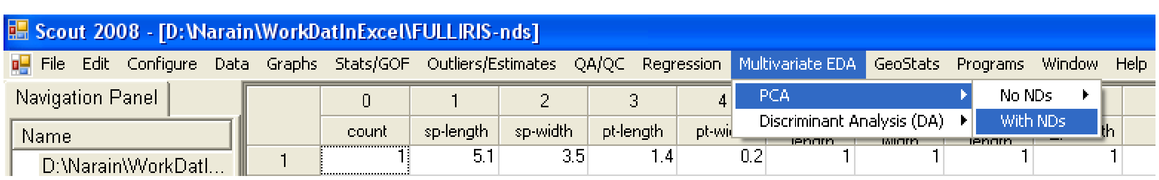
Observations outside of the simultaneous (Tolerance) ellipse are considered to be anomalous. Observations (if any) between the individual and the simultaneous ellipses may represent potential outliers.

Note: *The drop-down bars in the graphics toolbar can be used to obtain different load matrix plots, scatter plots of the components scores and the selected variables, and the Q-Q plots of the component scores, as explained in Chapter 2.*

10.1.3 Kaplan-Meier Principal Component Analysis

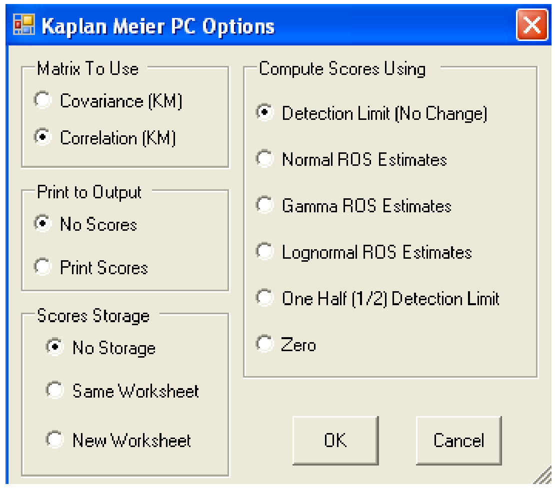
Principal component analysis of data with non-detects can be conducted in Scout. The Kaplan-Meier estimates of the covariance matrix and the correlation matrix is used for this analysis.

1. Click on **Multivariate EDA ► PCA ► With NDs.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

- Click on the "**Options**" button for the options window.



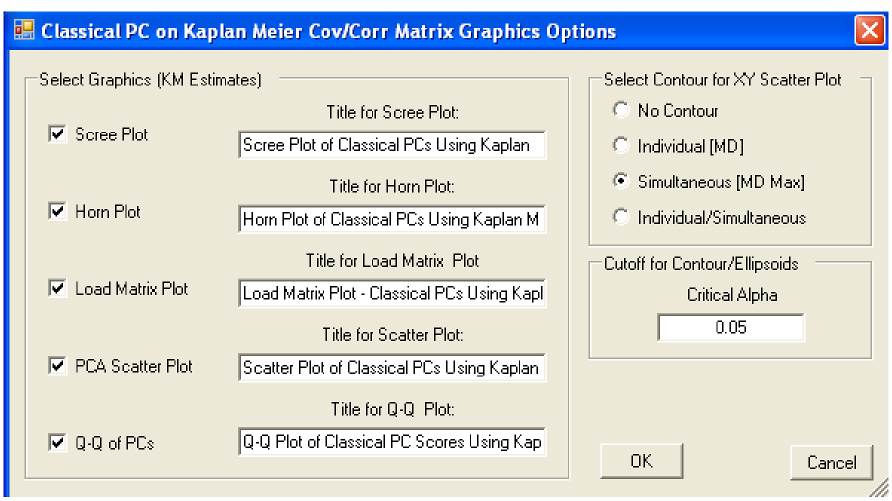
- Specify storage of the principal component scores. The default is "**No Storage**."

- Specify the "**Matrix To Use**" to compute the principal components. The default is "**Correlation (KM)**."

- Specify the estimates of the data to compute scores. Default is "**Detection Limit**."

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scree Plot**" provides a scree plot of the eigen values.

    - The “**Horn Plot**” provides a comparison of computed eigen values to the multi-normal generated eigen values.

    - The “**Load Matrix Plot**” provides the scatter plot of the columns of the load matrix.

    - The “**PCA Scatter Plot**” provides the scatter plot of the principal components scores and also the selected variables. The user has the option of drawing contours on the scatter plot to identify outliers. The default is “**No Contour**.” Specify the distribution for distances and the “**Critical Alpha**” value for the cutoff to compute the ellipses. The defaults are “**Beta**” and “**0.05**.”

    - The “**Q-Q Plot of PCA**” provides the Q-Q plots of the component scores.

    - Click on “**OK**” to continue or “**Cancel**” to cancel the graphics options.

- Click on “**OK**” to continue or “**Cancel**” to cancel the KM PCA computations.

Output example: The data set "**FullIris.xls**" was used for the KM PCA.

	Principal Components Analysis using the Classical Method
Date/Time of Computation	10/30/2008 7:43:49 AM
User Selected Options	
From File	D:\Narain\WorkDatInExcel\FULLIRIS-nds
Full Precision	OFF
Display Scores Option	Do not Display PC Scores in Output
PC Scores Storage	Do Not Store Scores to Worksheet
Matrix Used to Compute PCs	Correlation
Graphics	Load Matrix Plot Selected
Load Matrix Plot Title	Load Matrix Plot - Classical PCs Using Kaplan Meier Estimates
Graphics	XY Scatter Plot Selected
XY Scatter Plot Title	Scatter Plot of Classical PCs Using Kaplan Meier Estimates
Non-Detect Values Displayed As	Detection Limit (No Change to Original Data)
Contour	Contour Ellipses drawn at Individual Beta MD(0.05) and at Max MD(0.05)

Summary Statistics	
Number of Observations	150
Number of Selected Variables	4

KM Mean

sp-length	sp-width	pt-length	pt-width
5.845	3.037	3.754	1.175

KM Variance

sp-length	sp-width	pt-length	pt-width
0.675	0.199	3.117	0.604

KM Standard Deviation

sp-length	sp-width	pt-length	pt-width
0.822	0.446	1.765	0.777

KM Covariance S Matrix

sp-length	sp-width	pt-length	pt-width
0.675	-0.0763	1.245	0.522
-0.0763	0.199	-0.428	-0.152
1.245	-0.428	3.117	1.288
0.522	-0.152	1.288	0.604
		Determinant	0.00327

Output for the KM Principal Component Analysis (continued).

Eigenvalues of Classical Covariance S Matrix

Eval 1	Eval 2	Eval 3	Eval 4
4.23	0.244	0.0803	0.0395
		Sum of Eigenvalues	4.594

Classical Correlation R Matrix

	sp-length	sp-width	pt-length	pt-width
sp-length	1	-0.208	0.858	0.818
sp-width	-0.208	1	-0.543	-0.438
pt-length	0.858	-0.543	1	0.939
pt-width	0.818	-0.438	0.939	1
		Determinant	0.013	
		Log of Determinant	-4.345	

Eigenvalues of Classical Correlation R Matrix

Eval 1	Eval 2	Eval 3	Eval 4
2.987	0.83	0.147	0.0355
		Sum of Eigenvalues	4

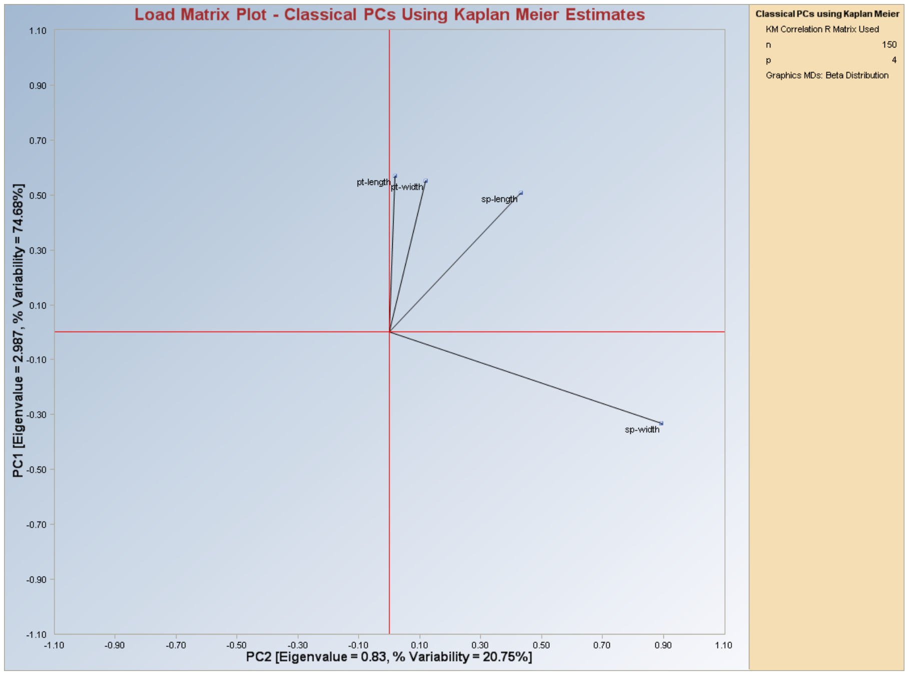
Summary Table (Eigenvalues)

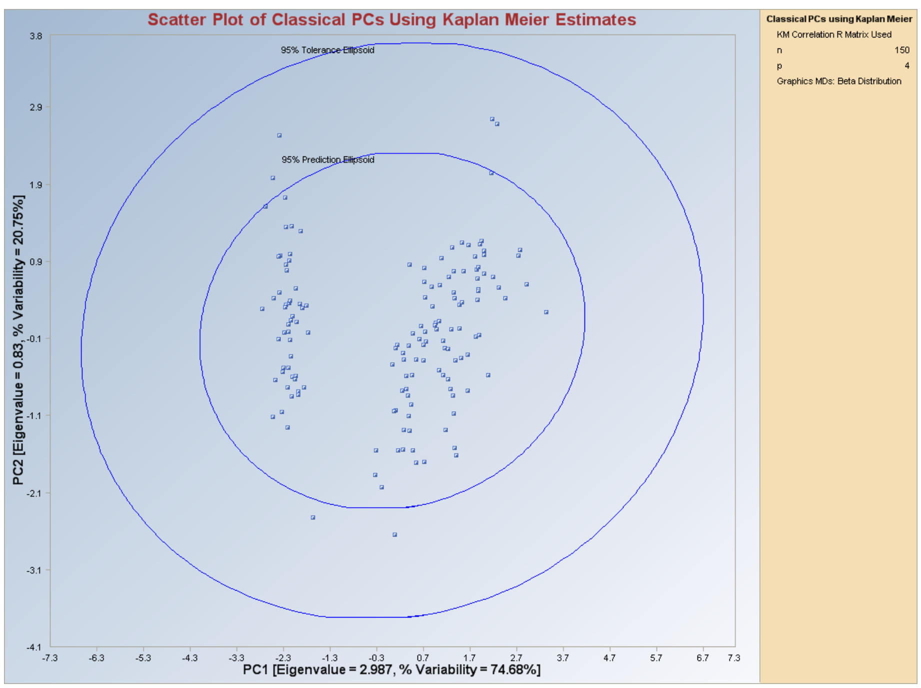
	Eigen Value	Difference	Proportion	Cumulative
PC1	2.987	2.158	0.747	74.68
PC2	0.83	0.683	0.207	95.43
PC3	0.147	0.112	0.0368	99.11
PC4	0.0355	N/A	0.00888	100

PC Loadings (Eigen Vectors)

	PC1	PC2	PC3	PC4
sp-length	0.509	0.433	-0.681	-0.301
sp-width	-0.331	0.894	0.237	0.189
pt-length	0.571	0.0187	0.078	0.817
pt-width	0.552	0.118	0.689	-0.455

Output for the KM Principal Component Analysis (continued).





Observations outside of the simultaneous (Tolerance) ellipse are considered to be anomalous. Observations (if any) between the individual and the simultaneous ellipses may represent potential outliers.

***Note**: The drop-down bars in the graphics toolbar can be used to obtain different load matrix plots, scatter plots of the components scores and the selected variables, and the Q-Q plots of the component scores, as explained in Chapter 2.*

10.2 Discriminant Analysis (DA)

Discriminant and classification analyses are multivariate techniques concerned with separating distinct groups of observations (Johnson and Wichern, 2002) and with allocating new observations (classification analysis) to previously defined groups (populations). The separation procedure is rather exploratory. In practice, the investigator has some knowledge about the nature and the number of groups. The study might be about **k** known groups (e.g., parts of a polluted site, type of species, geographic regions of a country). Some of those groups may be similar in nature and can be merged together.

The objective here is to establish **g ≤ k** significantly different groups. Let **s** = min (**g-1, p**). Then, **s** linear (Fisher) discriminant functions (also known as classification rules) can be computed for those **g** multivariate **p**-dimensional groups. Those functions (rules) are then used in all of the subsequent classifications.

Classification procedures are less exploratory. Discriminant functions (rules) obtained in the separation procedures are used to assign current and new observations into previously defined groups. The correct classification of the current observations with known group membership is the basis for the validity of discriminant functions. Scout outputs the classification, the misclassification matrices (confusion matrix), and the apparent error rates. The apparent error rate is the percent of misclassified observations. This number tends to be biased because the data being classified are the same data used to calculate the classification rules. The validity of the discriminant rules can be judged by performing cross validation. Several cross validation rules, including bootstrap cross validation methods, have been incorporated into Scout.

Outliers can distort the discriminant functions and the corresponding scores significantly. This can result in several misclassifications. Scout incorporates the robust procedures to minimize the distortion of various estimates and classification rules.

Three commonly used discriminant analysis methods are available in Scout. For Fisher Discriminant Analysis (FDA), one can also plot the scatter plots of discriminant scores. Moreover, simultaneous (tolerance) and individual (prediction) ellipsoids can be drawn on the scatter plots of the discriminant scores. The methods included in Scout are briefly described as follows. The details of the robustified methods (especially based upon the PROP influence function) can be found in Singh and Nocerino (1995).

- **Fisher Discriminant Analysis**

Assign x_0 to π_i, i = 1, 2, …, g, if:

$$\sum_{i=1}^{n}[l_i'(x_0 - \bar{x}_h^*]^2 = \min[\sum_{i=1}^{n}[l_i'(x_0 - \bar{x}_1^*]^2]; i = 1,2,...,g$$

and the Fisher discriminant score, y_i, is given by

$$y_i = l_i'x \qquad\qquad i = 1, 2, \ldots, s$$

where l_i are called the scaled (normalized) eigen vectors and are obtained from the eigen vectors of the $W^{*-1}\hat{B}^*$ matrix and are given by

$$l_i = \frac{e_i}{\sqrt{e_i' S^*_{pooled} e_i}}$$

- **Linear Discriminant Analysis**

  Assign x_0 to π_i, i = 1, 2, …, g, if:

  $$d_k^*(x_0) = \max\left[d_1^*(x_0), d_2^*(x_0), \ldots, d_g^*(x_0)\right]$$

  where the linear discriminant scores, $d_i^*(x)$, are given by

  $$d_i^*(x) = \mu_i' \Sigma^{-1} x - \tfrac{1}{2}\left[\mu_i' \Sigma^{-1} \mu_i\right] + \ln p_i$$

  where i = 1, 2, …, g.

- **Quadratic Discriminant Analysis**

  Assign x_0 to π_i, i = 1, 2, …, g, if:

  $$d_k^Q(x_0) = \max[d_1^Q(x_0), d_2^Q(x_0), \ldots, d_g^Q(x_0)]$$

  where the linear discriminant scores, $d_i^*(x)$, are given by

  $$d_i^Q(x) = -\tfrac{1}{2}\ln\left|\Sigma_i\right| - \tfrac{1}{2}[(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)] + \ln p_i$$

  where i = 1, 2, …, g.

As mentioned before, cross validation can be used to verify the validity and effectiveness of discriminant or classification rules. Various cross validation techniques have been provided in Scout. The user can select any of those techniques and compare their performances.

- **Leave One Out** (LOO) cross validation, where the classification rules are obtained using (n – 1) observations (training data or set) and testing is done on the classification test data with the left out observation. This is the most commonly used cross validation method employed in statistical software. Details can be found in Lachenbruch and Mickey (1968).

- **Split** cross validation, where the data is split to form two sets: the training set and test set. The training set is used to compute the classification rules, and the test set is used to validate those rules.

- **M-Fold** cross validation, where the data is divided into **M** equal (roughly) subsets. For each of the M subsets, combined data for the (**M** – 1) subsets are used as the training set and the remaining subset is used as the test set. This process is repeated **M** times for each of the M subsets.

- **Simple Bootstrap**

- **Standard Bootstrap**

- **Bias Adjusted Bootstrap**

The details of the bootstrap methods can be found in the referenced provided with the Scout software package.

Note*: The training sets and the test sets used in the various cross validation methods are obtained randomly. This random selection of the training sets (e.g., in robust methods) may result in some singular matrices needed to obtain the discriminant rules. Scout provides appropriate error or warning messages whenever such a condition occurs. Many times, in practice, matrices used to derive discriminant functions (e.g., in robust methods) become singular. This is especially true when not enough observations are available in each of the groups. When this happens, Scout gives an error message and further computations are stopped.*

Scout also provides an option to classify new observations or unknown observations into existing groups. There are certain logistical rules that need to be followed when using the classification of unknown or new observations.

- The first three letters of the group name of the new or unknown observations should be "**UNK**" or "**unk**" only.

- The set of unknown or new observations should be the last subset of observations in a data set. Otherwise an error message is obtained.

There are a few rules in the DA module of Scout which will not allow the contours to be plotted on the scatter plots. These rules are:

- If the standard deviation of any of the scores is less than 10^{-7} or greater 10^{+7}, then contours will not be plotted on their respective scatter plots.

- If the coefficient variation of any of the scores is less than 10^{-7} or greater 10^{+7}, then contours will not be plotted on their respective scatter plots.

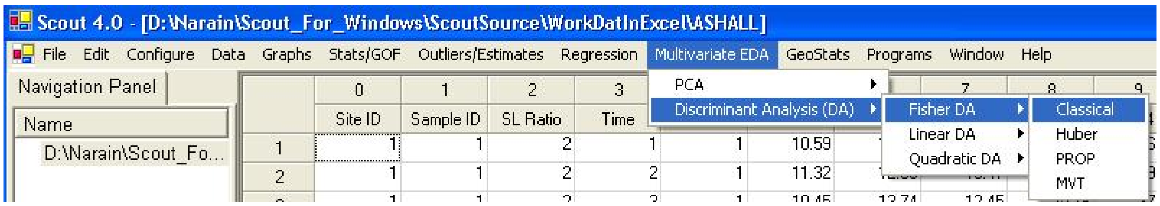
- If the absolute value of the correlation between the two variables used in scatter plots is greater than 0.99, then the contours will not be plotted.

- If the absolute difference between the standard deviations of the two variables used in the scatter plot is less than 10^{-20}, then contours will not be plotted.

10.2.1 Fisher Discriminant Analysis

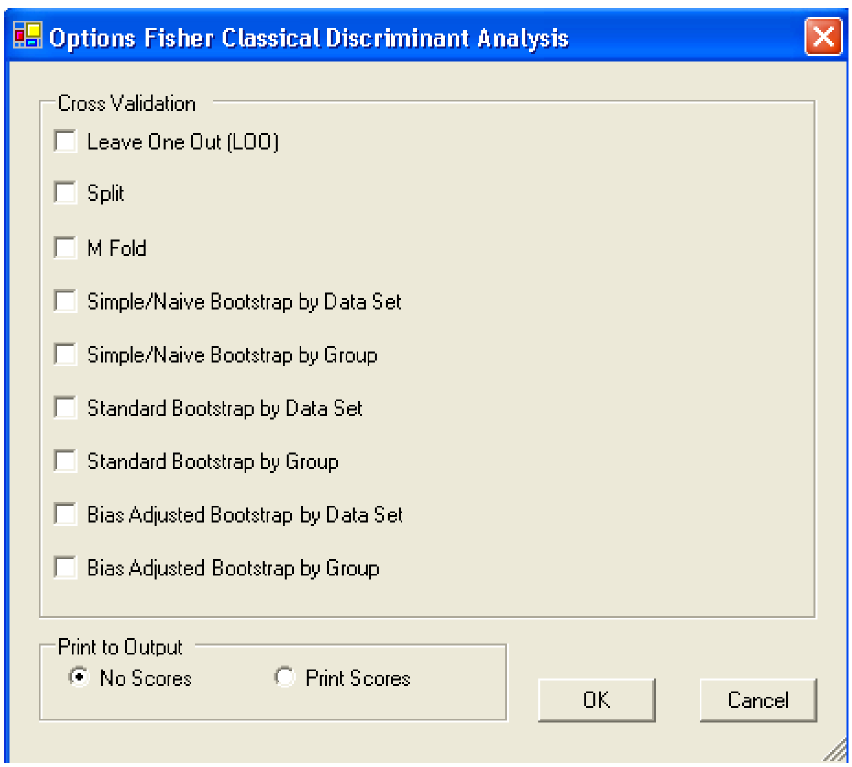
10.2.1.1 Classical Fisher DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Fisher DA ► Classical.**



2. A "**Select Variables**" screen (Section 3.5) appears.

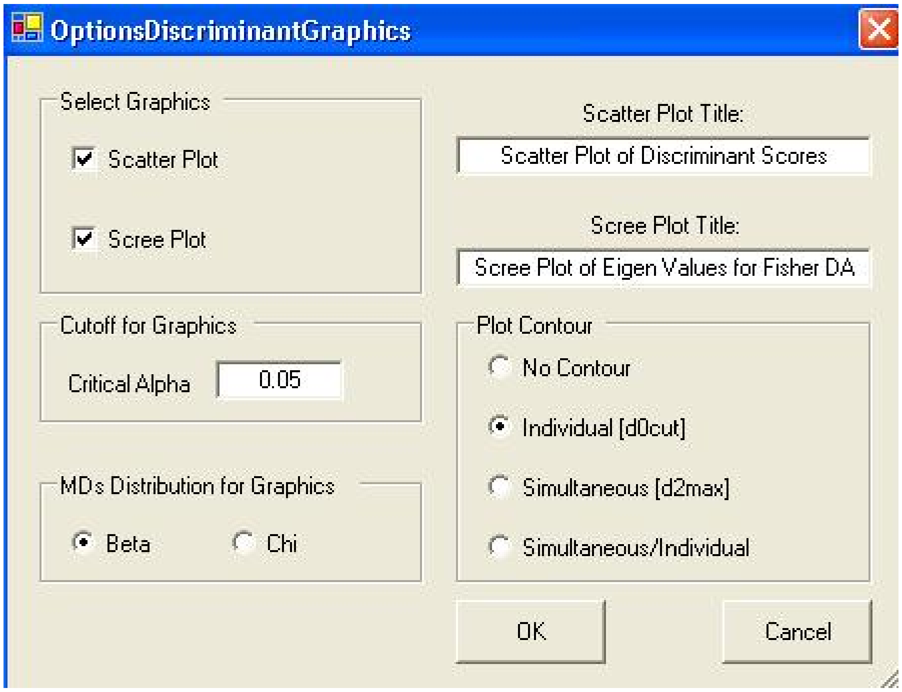
   - Click on the "**Options**" button for the options window.



   - Specify the preferred "**Cross Validation**" methods and their respective parameters.
   - Specify the "**Print to Output**." The default is "**No Scores**."

- Click “**OK**” to continue or “**Cancel**” to cancel the options.

- Click on the “**Graphics**” button for the graphics options window and check all of the check boxes.



  - The “**Scree Plot**” provides a scree plot of the eigen values.

  - The “**Scatter Plot**” provides the scatter plot of the discriminant analysis scores and also the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is “**No Contour**.” Specify the distribution for distances and the “**Critical Alpha**” value for the cutoff to compute the ellipses. The defaults are “**Beta**” and “**0.05**.”

  - Click on “**OK**” to continue or “**Cancel**” to cancel the graphics options.

- Specify the storage of the discriminant scores. No scores will be stored when “**No Storage**” is selected. Scores will be stored in the data worksheet starting from the first available empty column when the “**Same Worksheet**” is selected. Scores will be stored in a new worksheet if the “**New Worksheet**” is selected. The default is “**No Storage**.”

- Click on “**OK**” to continue or “**Cancel**” to cancel the DA computations.

Output example: The data set "**BEETLES.xls**" was used for the classical Fisher DA. It has 74 observations and two variables in three groups. The initial estimates of location and scale for each group were the classical mean and the covariance matrix. The classification rules were obtained using those estimates. The output shows that one observation was misclassified.

Output for the Classical Fisher Discriminant Analysis.
Data Set: Beetles (2 variables 3 groups).

	Classical Fisher Linear Discriminant Analysis
User Selected Options	
Date/Time of Computation	1/18/2008 10:22:23 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BEETLES
Full Precision	OFF
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors Assumed
Graphics Options	Both Scree Plot and Scatter Plots are Selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics
Total Number of Observations	74
Number of Selected Variables	2

Number of Data Rows per Group

1	2	3
21	31	22

Mean Vector for Group 1

x1-1	x2-1
146.2	14.1

Covariance S Matrix for Group 1

x1-1	x2-1
31.66	-0.969
-0.969	0.79

Mean Vector for Group 2

x1-2	x2-2
124.6	14.29

Covariance S Matrix for Group 2

x1-2	x2-2
21.37	-0.327
-0.327	1.213

Output for the Classical Fisher Discriminant Analysis (continued).

Mean Vector for Group 3

x1-3	x2-3
138.3	10.09

Covariance S Matrix for Group 3

x1-3	x2-3
17.16	-0.502
-0.502	0.944

Grand Mean Vector for Data

x1	x2
134.8	12.99

Pooled Covariance Matrix

x1	x2
23.02	-0.56
-0.56	1.014

Between Groups Matrix B

x1	x2
6187	-366.5
-366.5	263

Within Groups Matrix W

x1	x2
1635	-39.73
-39.73	72.01

W Inverse B Matrix (WiB)

x1	x2
3.711	-0.137
-3.041	3.576

Unordered Eigenvalues of WiB

Eval 1	Eval 2
4.293	2.994

Output for the Classical Fisher Discriminant Analysis (continued).

Associated Matrix of Eigen Vectors of WiB			
Eval 1	Eval 2		
0.0287	0.0235		
-0.973	0.982		
Ordered Eigen Values of WiB			
d1	d2		
4.293	2.994		
Normalized Eigen Vectors for Ordered Eigen Values			
Normalized Eigen Vector 1			
Eval 1	Eval 2		
0.0284	-0.963		
Normalized Eigen Vector 2			
Eval 1	Eval 2		
0.0243	1.017		
Classification Summary			
	Predicted Membership		
Actual	1	2	3
1	20	1	0
2	0	31	0
3	0	0	22
# Correct	20	31	22
Prop Correct	95.24%	100%	100%
		Total Observations	74
		Correctly Classified	73
		Incorrectly Classified	1
Misclassification Summary			
Obs No.	Actual	Predicted	
17	1	2	

Apparent Error Rate 0.0135

Output for the Classical Fisher Discriminant Analysis (continued).

Cross Validation Results

Leave One Out (LOO) Cross Validation Results

LOO Classification Summary

	Predicted Membership		
Actual	1	2	3
1	17	4	0
2	7	23	1
3	0	0	22
# Correct	17	23	22
Prop Correct	80.95%	74.19%	100%

Total Observations	74
Correctly Classified	62
Incorrectly Classified	12

LOO Misclassification Summary

Obs No.	Actual	Predicted
4	1	2
6	1	2
10	1	2
17	1	2
31	2	1
32	2	1
39	2	1
40	2	1
41	2	3
44	2	1
47	2	1
51	2	1

LOO Error Rate 0.162

Split (50/50) Cross Validation Results

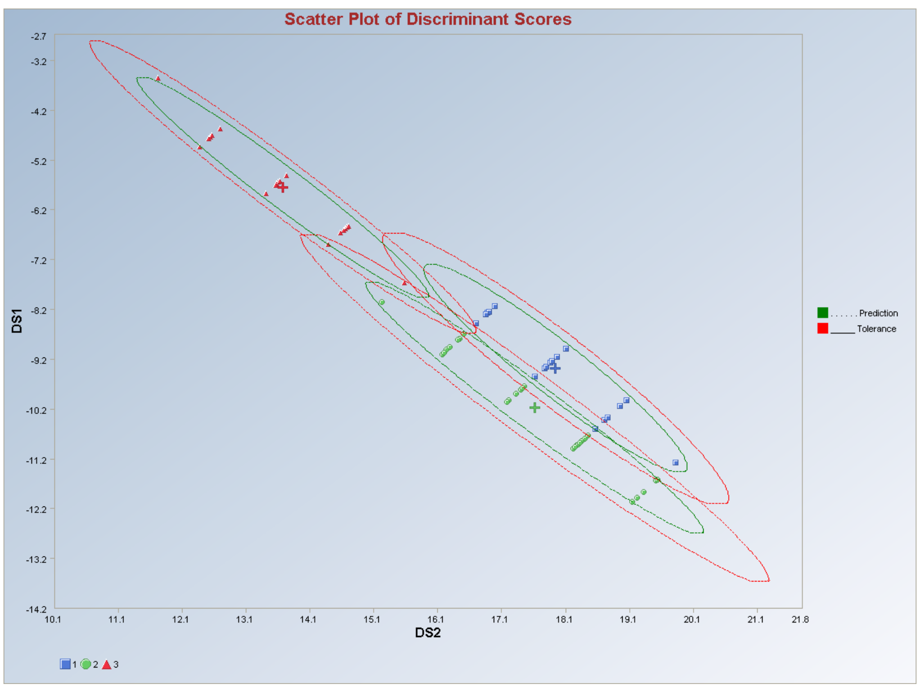
Error Rate for Training Set: 0.0245

Error Rate for Test Set: 0.0878

Output for the Classical Fisher Discriminant Analysis (continued).

3 Fold Cross Validation Results	
Average Error Rate: 0.2158	
Simple/Naive Bootstrap (for whole dataset) Cross Validation Results	
Average Error Rate from Bootstrap: 0.0408	
Simple/Naive Bootstrap (Groupwise) Cross Validation Results	
Average Error Rate from Bootstrap: 0.0447	
Standard Bootstrap (for whole dataset) Cross Validation Results	
Error Rate from Bootstrap Training Set: 0.0436	
Error Rate from Bootstrap Test Set: 0.0636	
Standard Bootstrap (Groupwise) Cross Validation Results	
Error Rate from Bootstrap Training Set: 0.0377	
Error Rate from Bootstrap Test Set: 0.0570	
Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results	
Average Correct Training Set: 70.1700	
Average Incorrect Training Set: 3.8300	
Average Correct Test Set: 63.5100	
Average Incorrect Test Set: 10.4900	
Error Rate Bias: -0.0900	
Bias Adjusted Error Rate: 0.1035	
Bias Adjusted Bootstrap (Groupwise) Cross Validation Results	
Average Correct Training Set: 70.8000	
Average Incorrect Training Set: 3.2000	
Average Correct Test Set: 62.0600	
Average Incorrect Test Set: 11.9400	
Error Rate Bias: -0.1181	
Bias Adjusted Error Rate: 0.1316	

Output for the Classical Fisher Discriminant Analysis (continued).

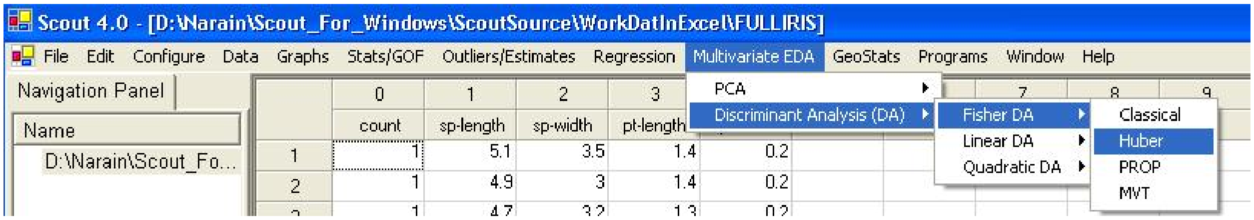


The color-coded big "+" represents the mean of the respective group, as shown in the above figure. Observations outside of the simultaneous (Tolerance) ellipse (if specified by the user) of a group category (e.g., #2) are considered to be anomalous for that particular group.

Note: *The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of discriminant scores and selected variables, as explained in Chapter 2.*

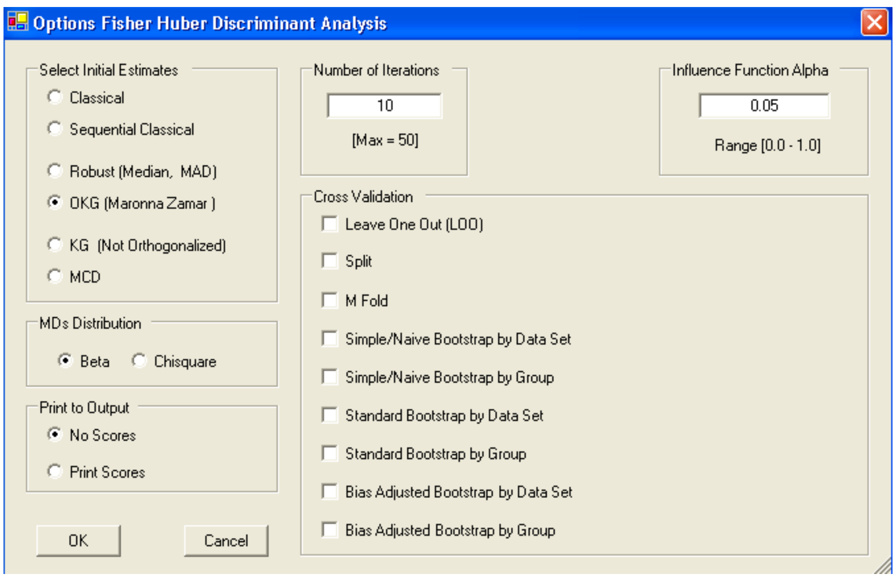
10.2.1.2 Huber Fisher DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Fisher DA ► Huber.**



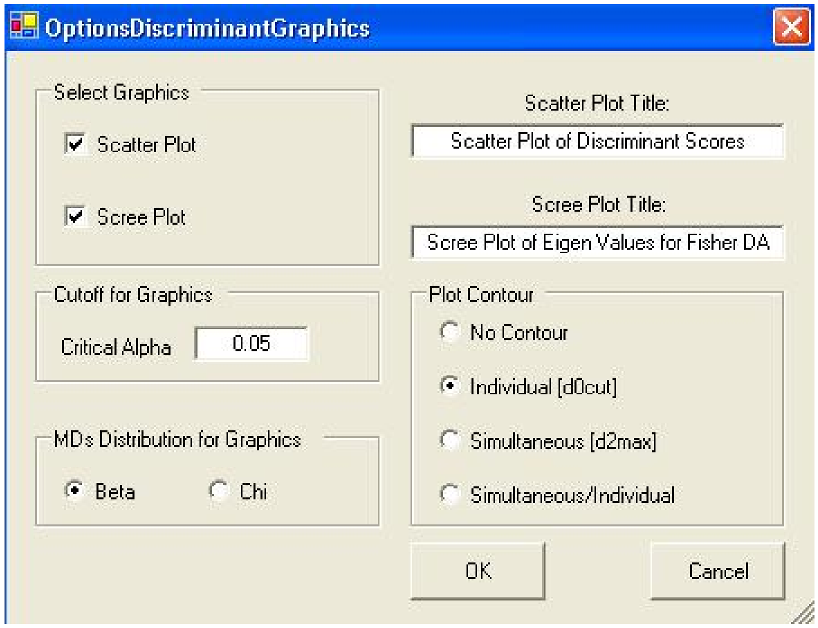
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



  - Specify the options to calculate the robust estimates of location and scatter (scale).
  - Specify the "**Print to Output**." The default is "**No Scores**."
  - Specify the preferred cross validation methods and their respective parameters.
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scree Plot**" provides a scree plot of the eigen values.

- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the storage of discriminant scores. No scores will be stored when "**No Storage**" is selected. Scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the Huber Fisher DA computations.

Output example: The data set "**IRIS.xls**" was used for the Huber Fisher DA. It has 150 observations and four variables in three groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the Huber influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that three observations were misclassified. The cross validation results suggest the same.

Output for the Huber Fisher Discriminant Analysis.
Data Set: IRIS (4 variables 3 groups).

	Robust Fisher Linear Discriminant Analysis using Huber Influence Function
User Selected Options	
Date/Time of Computation	1/18/2008 10:54:42 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\FULLIRIS
Full Precision	OFF
Influence Function Alpha	0.05
Squared MDs	Beta Distribution
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors Assumed
Graphics Options	Both Scree Plot and Scatter Plots are Selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05) snd Max MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics

Total Number of Observations	150
Number of Selected Variables	4

Number of Data Rows per Group

1	2	3
50	50	50

Mean Vector for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
5.006	3.428	1.462	0.246

Covariance S Matrix for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.124	0.0992	0.0164	0.0103
0.0992	0.144	0.0117	0.0093
0.0164	0.0117	0.0302	0.00607
0.0103	0.0093	0.00607	0.0111

IQR Fix!

Output for the Huber Fisher Discriminant Analysis (continued).

Final Robust Mean Vector for Group 1			
sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
5.008	3.431	1.463	0.245

Final Robust Covariance S Matrix for Group 1			
sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.123	0.0965	0.0162	0.0108
0.0965	0.137	0.0115	0.00989
0.0162	0.0115	0.0289	0.00585
0.0108	0.00989	0.00585	0.0105

Mean Vector for Group 2			
sp-le~th-2	sp-width-2	pt-le~th-2	pt-width-2
5.936	2.77	4.26	1.326

Covariance S Matrix for Group 2			
sp-le~th-2	sp-width-2	pt-le~th-2	pt-width-2
0.266	0.0852	0.183	0.0558
0.0852	0.0985	0.0827	0.0412
0.183	0.0827	0.221	0.0731
0.0558	0.0412	0.0731	0.0391

Final Robust Mean Vector for Group 2			
sp-le~th-2	sp-width-2	pt-le~th-2	pt-width-2
5.936	2.773	4.261	1.326

Final Robust Covariance S Matrix for Group 2			
sp-le~th-2	sp-width-2	pt-le~th-2	pt-width-2
0.266	0.0864	0.181	0.0554
0.0864	0.0969	0.0834	0.0421
0.181	0.0834	0.218	0.0727
0.0554	0.0421	0.0727	0.0391

Mean Vector for Group 3			
sp-le~th-3	sp-width-3	pt-le~th-3	pt-width-3
6.588	2.974	5.552	2.026

Output for the Huber Fisher Discriminant Analysis (continued).

Covariance S Matrix for Group 3

sp-le~th-3	sp-width-3	pt-le~th-3	pt-width-3
0.404	0.0938	0.303	0.0491
0.0938	0.104	0.0714	0.0476
0.303	0.0714	0.305	0.0488
0.0491	0.0476	0.0488	0.0754

Final Robust Mean Vector for Group 3

sp-le~th-3	sp-width-3	pt-le~th-3	pt-width-3
6.578	2.973	5.542	2.025

Final Robust Covariance S Matrix for Group 3

sp-le~th-3	sp-width-3	pt-le~th-3	pt-width-3
0.389	0.0918	0.287	0.0469
0.0918	0.0997	0.0716	0.0491
0.287	0.0716	0.287	0.046
0.0469	0.0491	0.046	0.0759

Robust Grand Mean Vector for Data

sp-length	sp-width	pt-length	pt-width
5.843	3.057	3.758	1.199

Robust Pooled Covariance Matrix

sp-length	sp-width	pt-length	pt-width
0.26	0.0915	0.162	0.0378
0.0915	0.111	0.0557	0.0338
0.162	0.0557	0.178	0.0417
0.0378	0.0338	0.0417	0.0419

Between Groups Matrix B

sp-length	sp-width	pt-length	pt-width
61.68	-19.79	162	70.04
-19.79	11.26	-56.89	-22.84
162	-56.89	430.5	184.3
70.04	-22.84	184.3	79.56

Output for the Huber Fisher Discriminant Analysis (continued).

Within Groups Matrix W			
sp-length	sp-width	pt-length	pt-width
37.55	13.24	23.39	5.468
13.24	16.07	8.047	4.884
23.39	8.047	25.79	6.023
5.468	4.884	6.023	6.059

W Inverse B Matrix (WiB)			
sp-length	sp-width	pt-length	pt-width
-2.912	1.04	-7.755	-3.315
-6.357	2.497	-17.15	-7.252
8.332	-3.073	22.29	9.491
11.03	-3.666	29.1	12.53

Unordered Eigenvalues of WiB			
Eval 1	Eval 2	Eval 3	Eval 4
34.11	0.29	-4.08E-15	-3.04E-16

Associated Matrix of Eigen Vectors of WiB			
Eval 1	Eval 2	Eval 3	Eval 4
-0.188	-0.0056	0.624	-0.479
-0.418	0.599	-0.445	-0.136
0.542	-0.243	-0.478	-0.199
0.705	0.763	0.43	0.844

Ordered Eigen Values of WiB	
d1	d2
34.11	0.29

Normalized Eigen Vectors for Ordered Eigen Values

Normalized Eigen Vector 1			
Eval 1	Eval 2	Eval 3	Eval 4
-3.147	-6.981	9.051	11.78

Normalized Eigen Vector 2			
Eval 1	Eval 2	Eval 3	Eval 4
-0.0762	8.148	-3.312	10.38

Output for the Huber Fisher Discriminant Analysis (continued).

Classification Summary

	Predicted Membership		
Actual	1	2	3
1	50	0	0
2	0	48	2
3	0	1	49
# Correct	50	48	49
Prop Correct	100%	96%	98%

Total Observations	150
Correctly Classified	147
Incorrectly Classified	3

Misclassification Summary

Obs No.	Actual	Predicted
71	2	3
84	2	3
134	3	2

Apparent Error Rate 0.02

Cross Validation Results

Leave One Out (LOO) Cross Validation Results

LOO Classification Summary

	Predicted Membership		
Actual	1	2	3
1	50	0	0
2	0	48	2
3	0	1	49
# Correct	50	48	49
Prop Correct	100%	96%	98%

Total Observations	150
Correctly Classified	147
Incorrectly Classified	3

Output for the Huber Fisher Discriminant Analysis (continued).

LOO Misclassification Summary				
Obs No.	Actual	Predicted		
71	2	3		
84	2	3		
134	3	2		
			LOO Error Rate	0.02

Split (50/50) Cross Validation Results

Error Rate for Training Set: 0.0093

Error Rate for Test Set: 0.0107

Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results

Validation Failed becuase of not enough Non-Outliers in Grouyp 1 times.

Average Correct Training Set: 147.5556

Average Incorrect Training Set: 2.4444

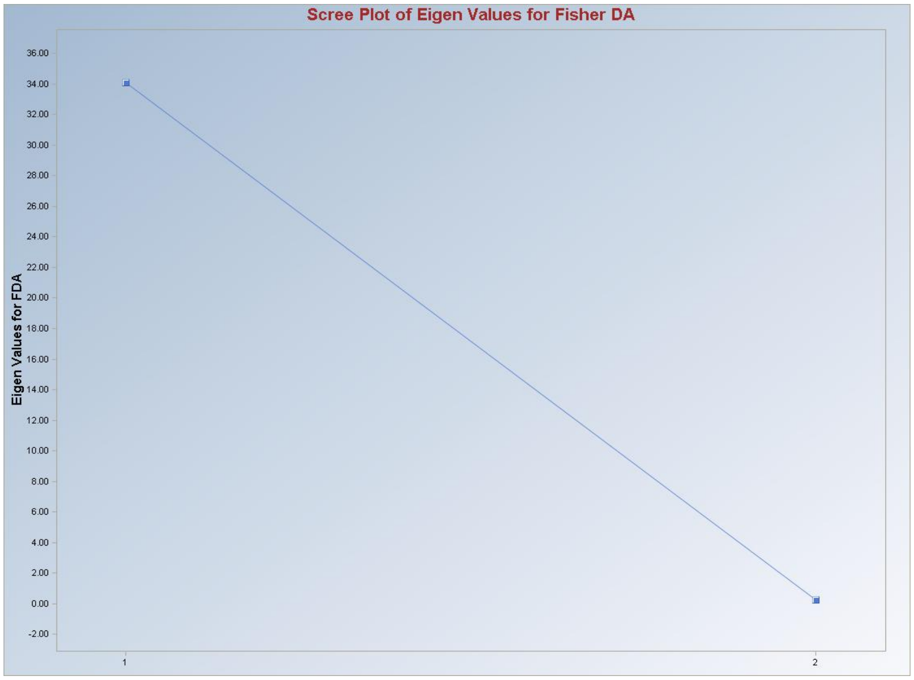
Average Correct Test Set: 147.1111

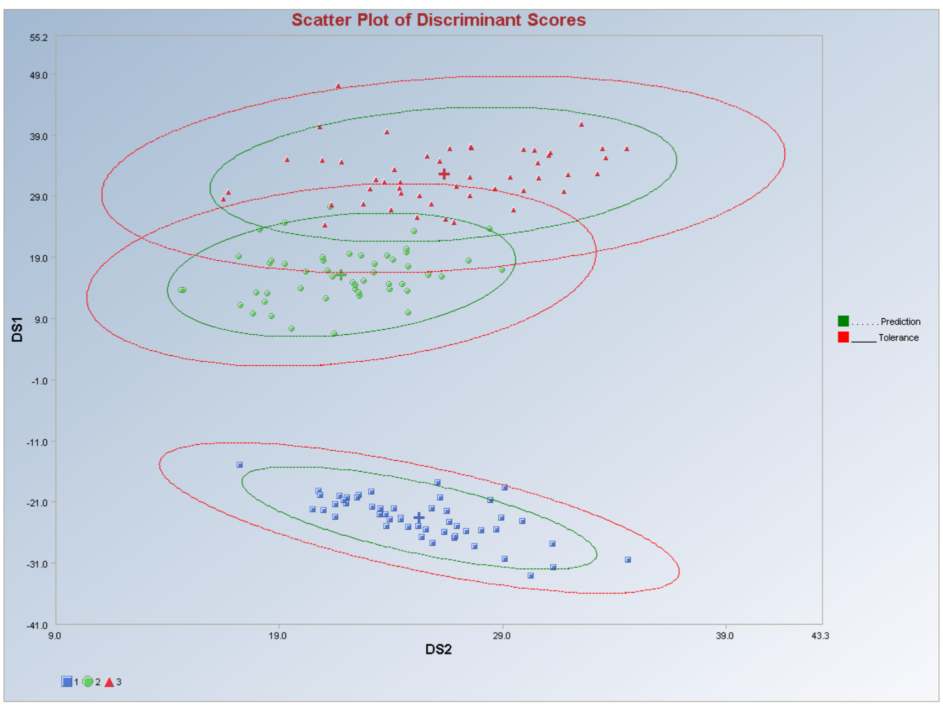
Average Incorrect Test Set: 2.8889

Error Rate Bias: -0.0030

Bias Adjusted Error Rate: 0.0230

Output for the Huber Fisher Discriminant Analysis (continued).



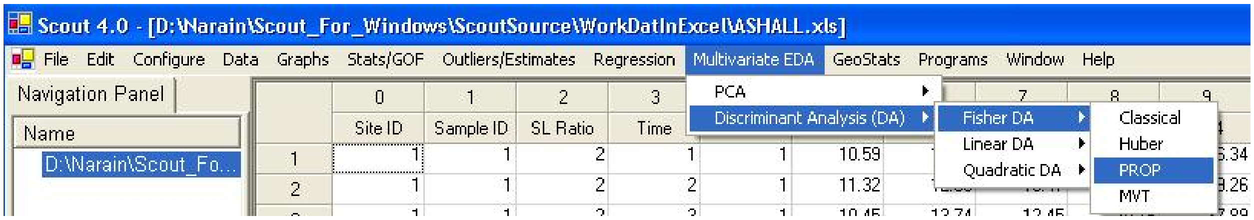


On a scatter plot of discriminant scores, it is desirable to use only one ellipsoid (e.g., prediction ellipsoid) for each group. That will reduce the clutter on a graph.

Note*: The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of discriminant scores and selected variables, as explained in Chapter 2.*

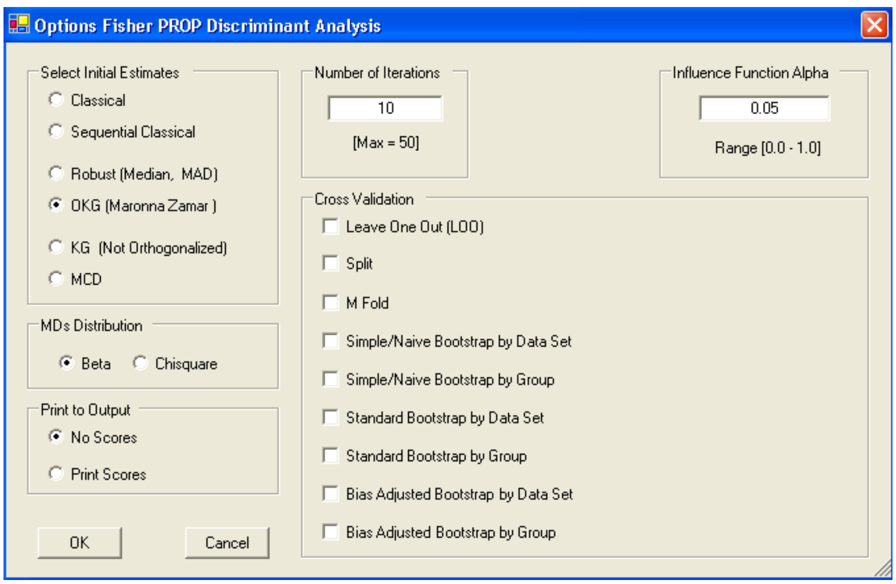
10.2.1.3 PROP Fisher DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Fisher DA ► PROP.**



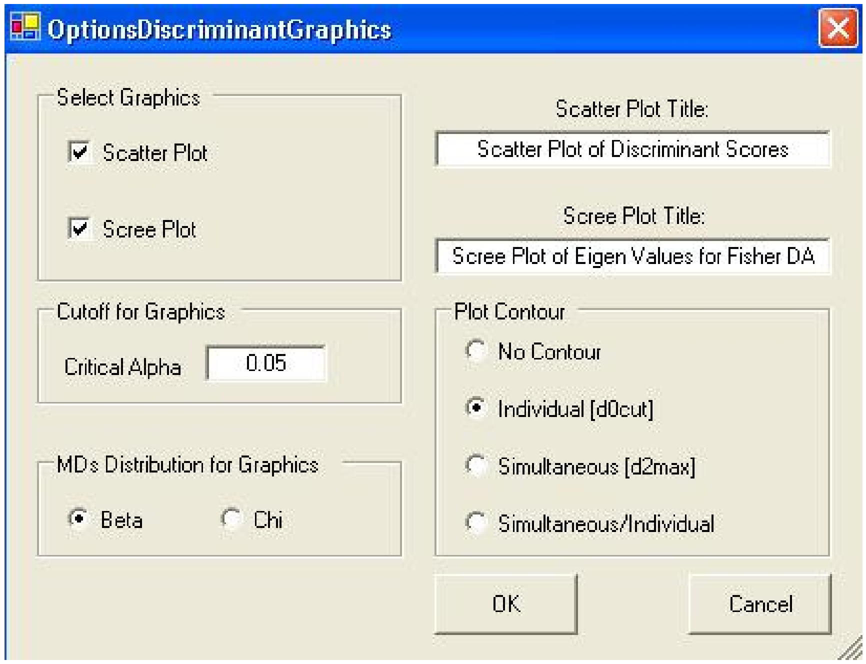
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



  - Specify the options to calculate the robust estimates of location and scatter (scale).
  - Specify the "**Print to Output**." The default is "**No Scores**."
  - Specify the preferred cross validation methods and their respective parameters.
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scree Plot**" provides a scree plot of the eigen values.

- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the storage of discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the computations.

Output example: The data set "**IRIS.xls**" was used for the PROP Fisher DA. It has 150 observations and four variables in three groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the PROP influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that three observations were misclassified. The cross validation results suggest the same.

Output for the PROP Fisher Discriminant Analysis.
Data Set: Iris (4 variables 3 groups).

	Robust Fisher Linear Discriminant Analysis using PROP Influence Function
User Selected Options	
Date/Time of Computation	1/18/2008 11:59:51 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\FULLIRIS
Full Precision	OFF
Influence Function Alpha	0.05
Squared MDs	Beta Distribution
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors Assumed
Graphics Options	Both Scree Plot and Scatter Plots are Selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05) snd Max MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics

Total Number of Observations	150
Number of Selected Variables	4

Number of Data Rows per Group

1	2	3
50	50	50

Mean Vector for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
5.006	3.428	1.462	0.246

Covariance S Matrix for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.124	0.0992	0.0164	0.0103
0.0992	0.144	0.0117	0.0093
0.0164	0.0117	0.0302	0.00607
0.0103	0.0093	0.00607	0.0111

IQR Fix!

(Complete results are not shown.)

Output for the PROP Fisher Discriminant Analysis (continued).

Associated Matrix of Eigen Vectors of WiB

Eval 1	Eval 2	Eval 3	Eval 4
-0.163	-0.0206	-0.53	-0.322
-0.477	0.607	-0.172	0.454
0.511	-0.237	-0.178	0.475
0.696	0.758	0.811	-0.682

Ordered Eigen Values of WiB

d1	d2
39.09	0.288

Normalized Eigen Vectors for Ordered Eigen Values

Normalized Eigen Vector 1

Eval 1	Eval 2	Eval 3	Eval 4
-3.305	-9.675	10.37	14.11

Normalized Eigen Vector 2

Eval 1	Eval 2	Eval 3	Eval 4
-0.283	8.358	-3.266	10.45

Classification Summary

	Predicted Membership		
Actual	1	2	3
1	50	0	0
2	0	49	1
3	0	1	49
# Correct	50	49	49
Prop Correct	100%	98%	98%

Total Observations	150
Correctly Classified	148
Incorrectly Classified	2

Misclassification Summary

Obs No.	Actual	Predicted
84	2	3
134	3	2

Apparent Error Rate	0.0133

Output for the PROP Fisher Discriminant Analysis (continued).

Cross Validation Results

Leave One Out (LOO) Cross Validation Results

LOO Classification Summary

	Predicted Membership		
Actual	1	2	3
1	50	0	0
2	0	48	2
3	0	1	49
# Correct	50	48	49
Prop Correct	100%	96%	98%

Total Observations	150
Correctly Classified	147
Incorrectly Classified	3

LOO Misclassification Summary

Obs No.	Actual	Predicted
71	2	3
84	2	3
134	3	2

LOO Error Rate 0.02

Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results

Validation Failed becuase of not enough Non-Outliers in Grouyp 1 times.

Average Correct Training Set: 146.6667

Average Incorrect Training Set: 3.3333

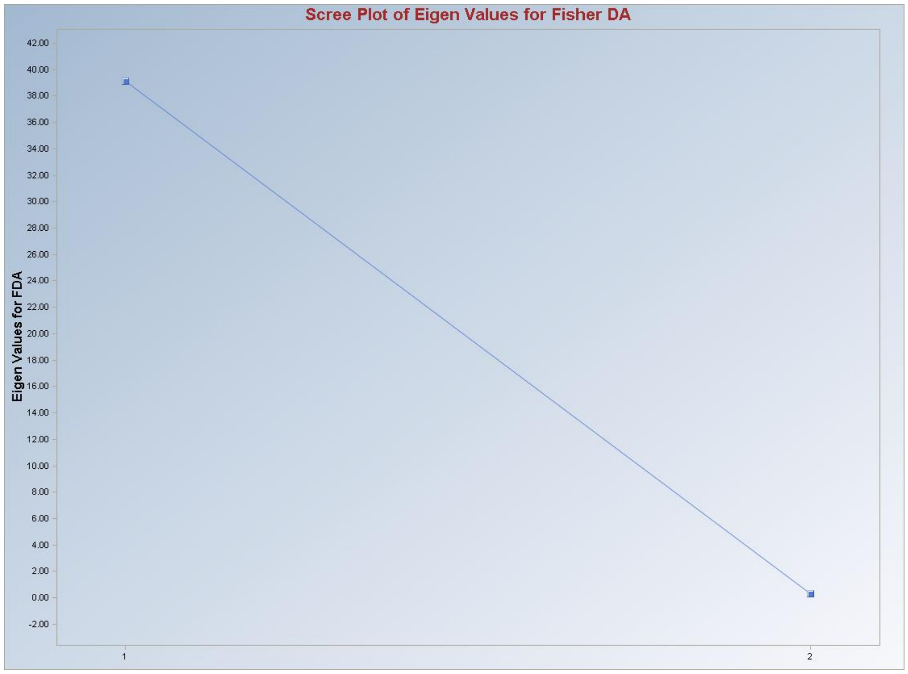
Average Correct Test Set: 139.5556

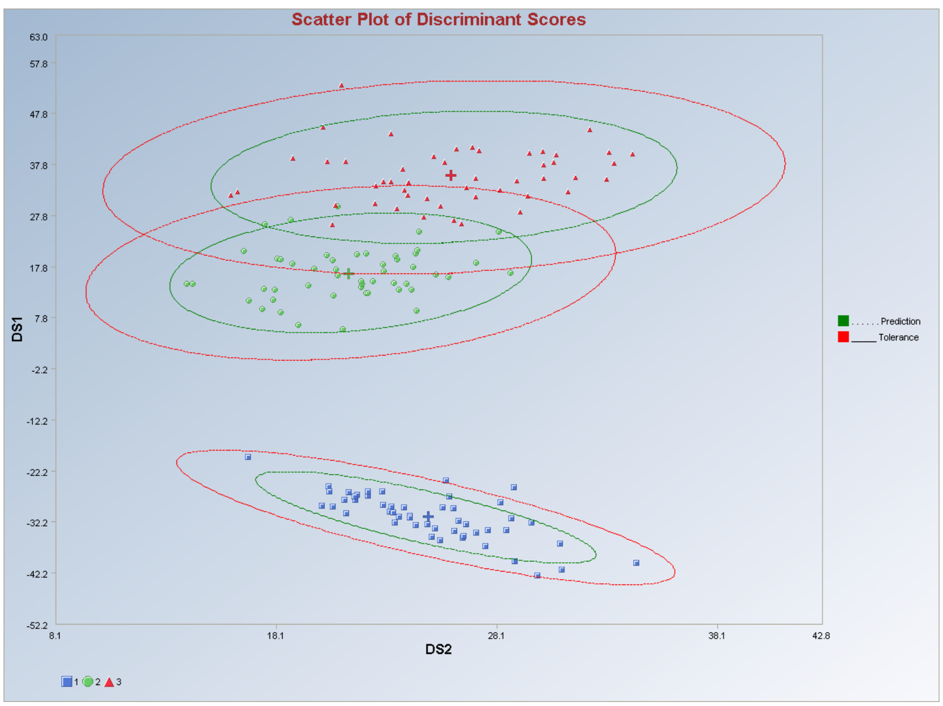
Average Incorrect Test Set: 10.4444

Error Rate Bias: -0.0474

Bias Adjusted Error Rate: 0.0607

Output for the PROP Fisher Discriminant Analysis (continued).



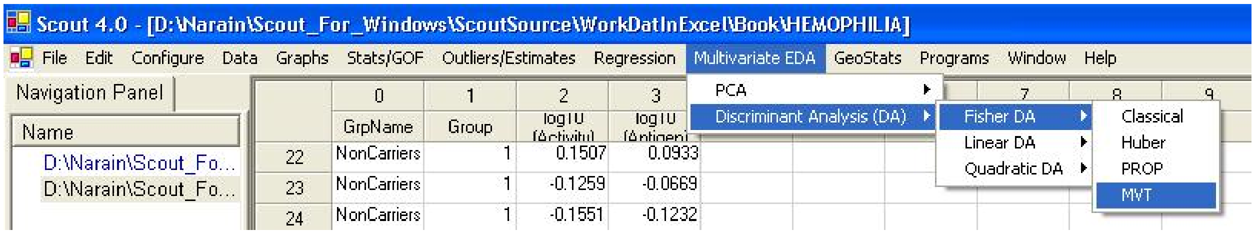


Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

Note*: The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

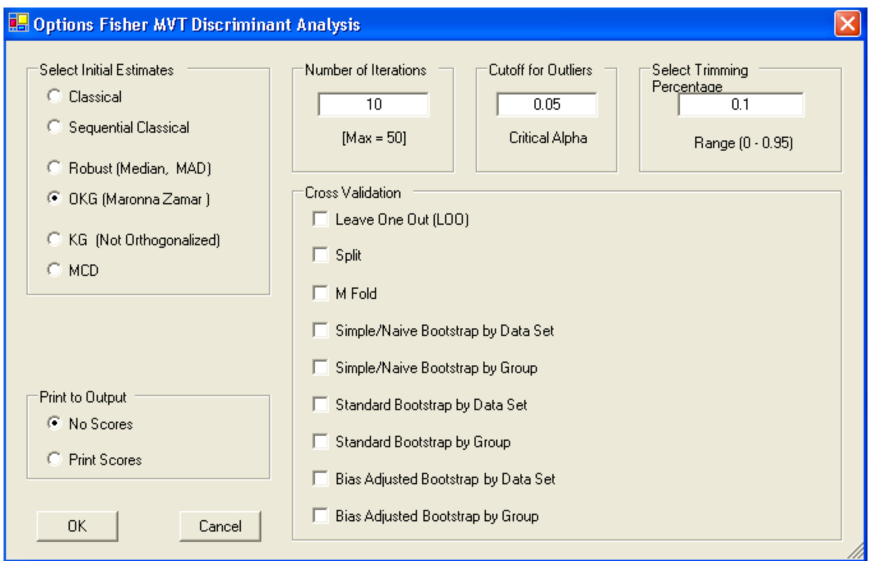
10.2.1.4 MVT Fisher DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Fisher DA ► MVT.**



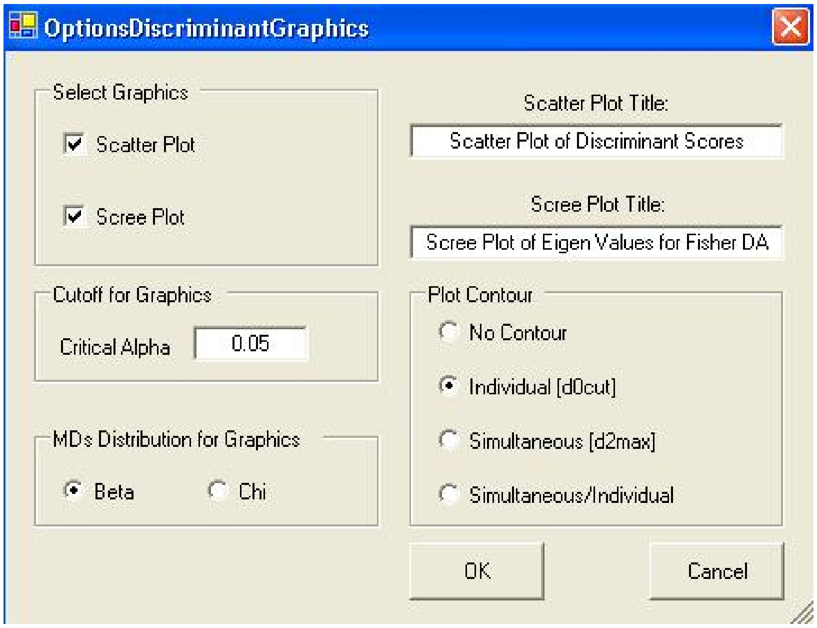
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



  - Specify the options to calculate the robust estimates of location and scatter (scale or dispersion).
  - Specify the "**Print to Output**." The default is "No Scores."
  - Specify the preferred cross validation methods and their respective parameters.
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scree Plot**" provides a scree plot of the eigen values.

- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the storage of discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the DA computations.

Output example: The data set "**Salmon.xls**" was used for the MVT Fisher DA. It has 102 variables in two groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the trimming percentage and critical alpha and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The $W^{-1}B$ matrix used for computing the classification rules was singular and the calculations were stopped.

Output for the MVT Fisher Discriminant Analysis.
Data Set: Salmon (2 variables 2 groups).

	Robust Fisher Linear Discriminant Analysis using MVT Method
User Selected Options	
Date/Time of Computation	1/18/2008 2:01:48 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Book\SALMON.xls
Full Precision	OFF
Trimming Percentage	10%
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors Assumed
Graphics Options	Both Scree Plot and Scatter Plots are Selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics

Total Number of Observations	100
Number of Selected Variables	2

Number of Data Rows per Group

alaskan	canadian
50	50

Mean Vector for Group alaskan

Fresh~skan	Marin~skan
98.38	429.7

Covariance S Matrix for Group alaskan

Fresh~skan	Marin~skan
260.6	-188.1
-188.1	1399

Final Robust Mean Vector for Group alaskan

Fresh~skan	Marin~skan
98.42	429.8

(Complete results are not shown.)

Output for the MVT Fisher Discriminant Analysis (continued).

Final Robust Mean Vector for Group canadian	
Fresh~dian	Marin~dian
138.1	366.4

Final Robust Covariance S Matrix for Group canadian	
Fresh~dian	Marin~dian
300.3	224.7
224.7	610.7

Robust Grand Mean Vector for Data	
FreshWater	Marine
117.9	398.1

Robust Pooled Covariance Matrix	
FreshWater	Marine
241.8	0.425
0.425	946.5

Between Groups Matrix B	
FreshWater	Marine
35403	-56624
-56624	90567

Within Groups Matrix W	
FreshWater	Marine
21281	37.38
37.38	83292

W Inverse B Matrix (WiB)	
FreshWater	Marine
1.665	-2.663
-0.681	1.089

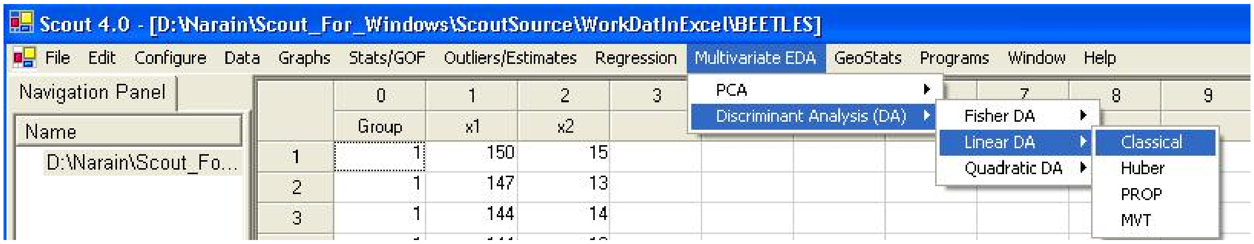
Failed in calculating Eigen Values - WiB produce Singular Condition

***Note**: When a matrix obtained during the calculations of discriminant scores is singular, an appropriate message is displayed and the computations are stopped.*

10.2.2 Linear Discriminant Analysis

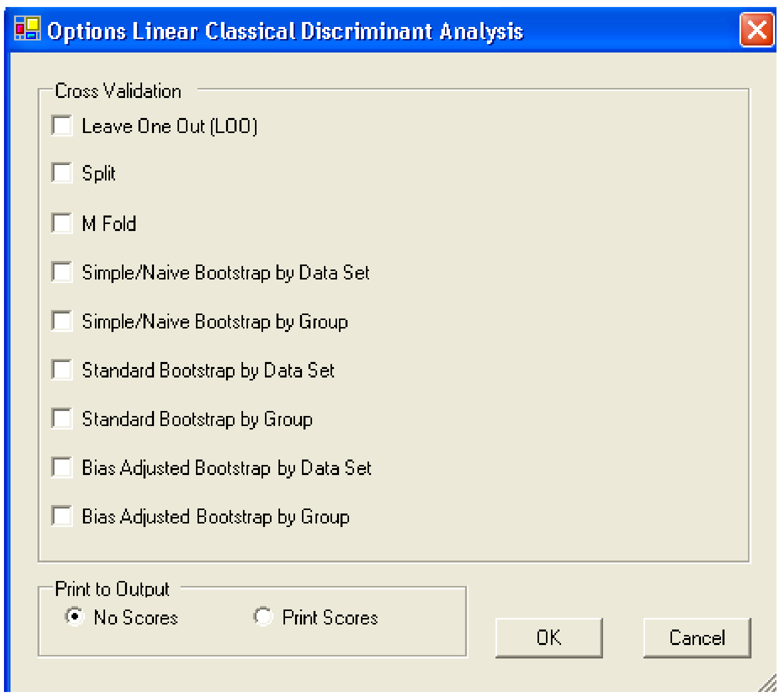
<u>10.2.2.1 Classical Linear DA</u>

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Linear DA ► Classical.**



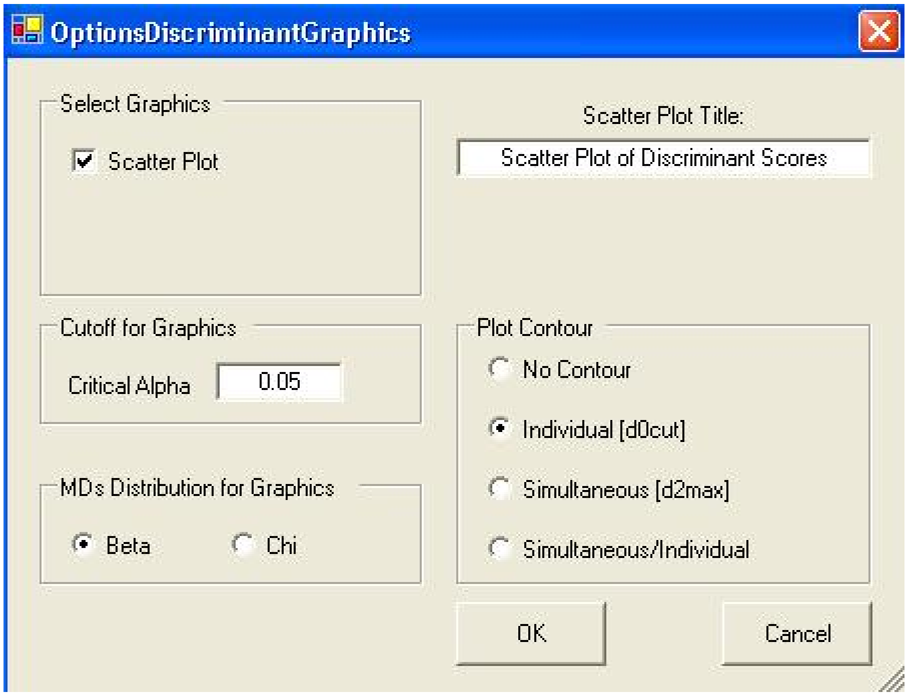
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



  - Specify the preferred cross validation methods and their respective parameters.
  - Specify the "**Print to Output**." The default is "**No Scores**."
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the prior probabilities. The prior probabilities can be: "**Equal**" for all of the groups; "**Estimated**," based on the number of observations in each group; or "**User Supplied**," where a column of priors can be obtained from "**Select Group Priors Column**." The default is "**Equal**" priors.

- Specify the storage for the discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the DA computations.

Output example: The data set "**BEETLES.xls**" was used for the classical linear DA. It has 74 observations and two variables in three groups. The initial estimates of location and scale for each group were the classical mean and the covariance matrix. The classification rules were obtained using those estimates. The output shows that one observation was misclassified.

Output for the Classical Linear Discriminant Analysis.
Data Set: Beetles (2 variables 3 groups).

	Classical Linear Discriminant Analysis
User Selected Options	
Date/Time of Computation	1/18/2008 2:09:58 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BEETLES
Full Precision	OFF
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics

Total Number of Observations	74
Number of Selected Variables	2

Number of Data Rows per Group

1	2	3
21	31	22

Mean Vector for Group 1

x1-1	x2-1
146.2	14.1

Covariance S Matrix for Group 1

x1-1	x2-1
31.66	-0.969
-0.969	0.79

Mean Vector for Group 2

x1-2	x2-2
124.6	14.29

Covariance S Matrix for Group 2

x1-2	x2-2
21.37	-0.327
-0.327	1.213

(Complete results are not shown.)

Output for the Classical Linear Discriminant Analysis (continued).

Classification Summary

	Predicted Membership		
Actual	1	2	3
1	20	1	0
2	0	31	0
3	0	0	22
# Correct	20	31	22
Prop Correct	95.24%	100%	100%

Total Observations	74
Correctly Classified	73
Incorrectly Classified	1

Misclassification Summary

Obs No.	Actual	Predicted
17	1	2

Apparent Error Rate 0.0135

Linear Discriminant Function Constants and Coefficients

	1	**2**	**3**
Constant	-620.8	-488.4	-506.7
x1	6.778	5.834	6.332
x2	17.64	17.31	13.44

Output for the Classical Linear Discriminant Analysis (continued).

Cross Validation Results

Leave One Out (LOO) Cross Validation Results

LOO Classification Summary

	Predicted Membership		
Actual	1	2	3
1	20	1	0
2	0	31	0
3	0	0	22
# Correct	20	31	22
Prop Correct	95.24%	100%	100%

Total Observations	74
Correctly Classified	73
Incorrectly Classified	1

LOO Misclassification Summary

Obs No.	Actual	Predicted
17	1	2

LOO Error Rate 0.0135

Split (50/50) Cross Validation Results

Error Rate for Training Set: 0.0051

Error Rate for Test Set: 0.0078

3 Fold Cross Validation Results

Average Error Rate: 0.0139

Simple/Naive Bootstrap (for whole dataset) Cross Validation Results

Average Error Rate from Bootstrap: 0.0099

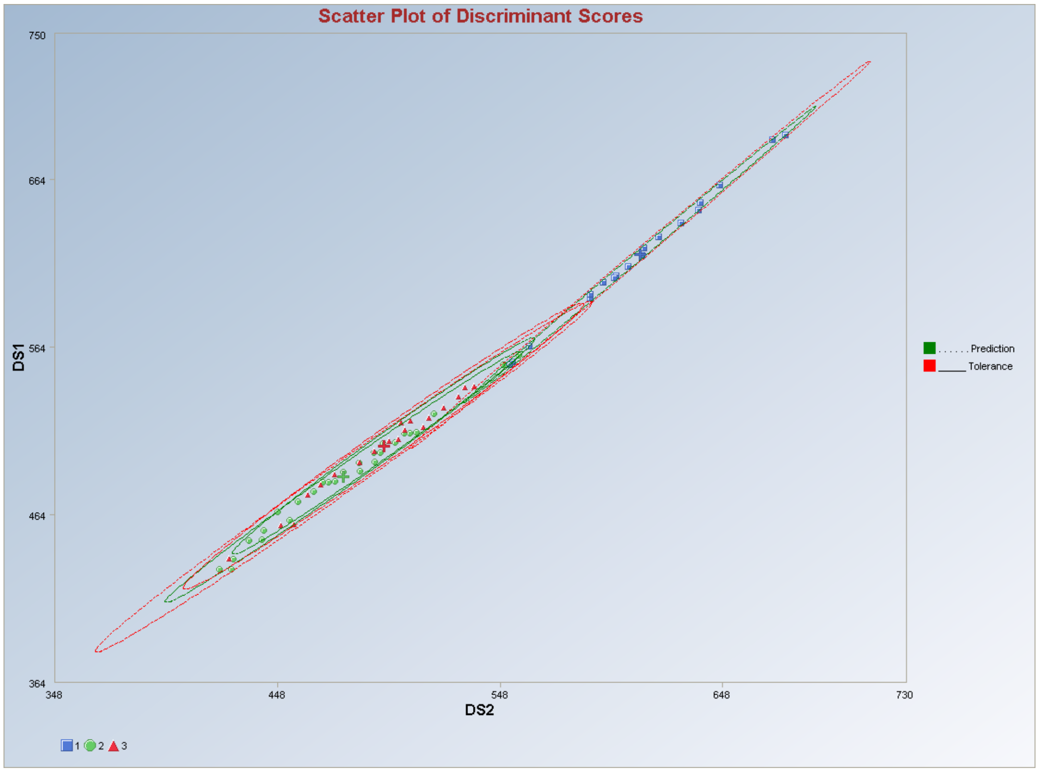
Simple/Naive Bootstrap (Groupwise) Cross Validation Results

Average Error Rate from Bootstrap: 0.0107

Output for the Classical Linear Discriminant Analysis (continued).

Standard Bootstrap (for whole dataset) Cross Validation Results	
Error Rate from Bootstrap Training Set: 0.0119	
Error Rate from Bootstrap Test Set: 0.0051	
Standard Bootstrap (Groupwise) Cross Validation Results	
Error Rate from Bootstrap Training Set: 0.0103	
Error Rate from Bootstrap Test Set: 0.0059	
Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results	
Average Correct Training Set: 73.3300	
Average Incorrect Training Set: 0.6700	
Average Correct Test Set: 73.1100	
Average Incorrect Test Set: 0.8900	
Error Rate Bias: -0.0030	
Bias Adjusted Error Rate: 0.0165	
Bias Adjusted Bootstrap (Groupwise) Cross Validation Results	
Average Correct Training Set: 73.2600	
Average Incorrect Training Set: 0.7400	
Average Correct Test Set: 73.0800	
Average Incorrect Test Set: 0.9200	
Error Rate Bias: -0.0024	
Bias Adjusted Error Rate: 0.0159	

Output for the Classical Linear Discriminant Analysis (continued).

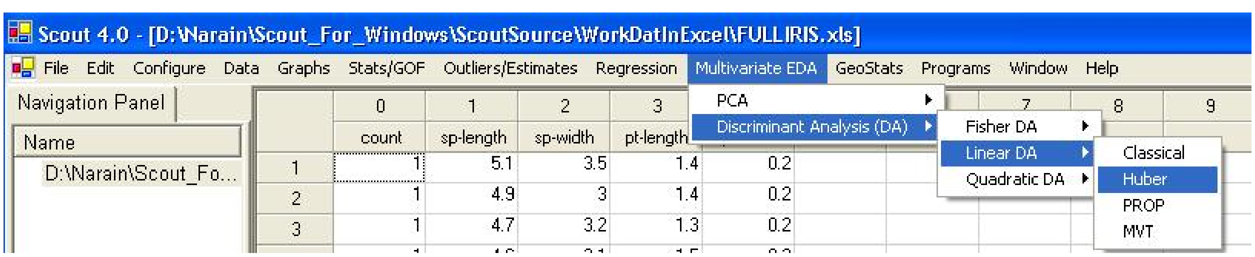


Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

***Note**: The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

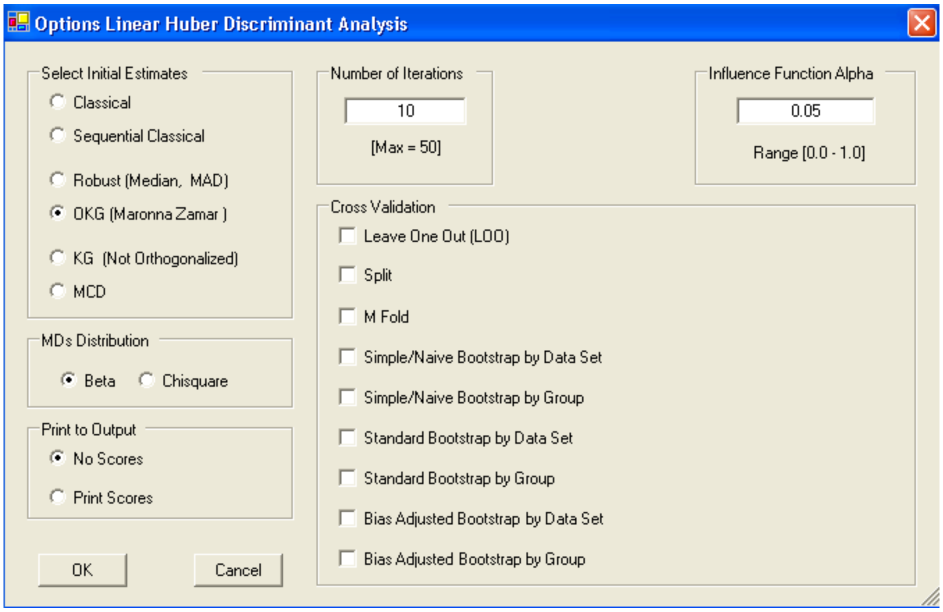
10.2.2.2 Huber Linear DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Linear DA ► Huber.**



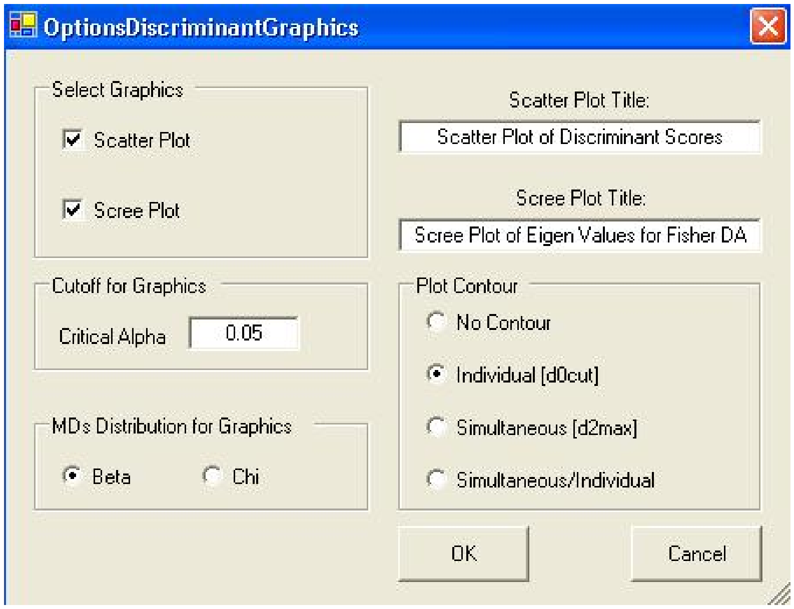
3. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



- Specify the options to calculate the robust estimates of the location and the scatter (scale or dispersion).
- Specify the "**Print to Output**." The default is "**No Scores**."
- Specify the preferred cross validation methods and their respective parameters.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the prior probabilities. The prior probabilities can be: "**Equal**" for all of the groups; "**Estimated**," based on number of observations in each group; or "**User Supplied**," where a column of priors can be obtained from the "**Select Group Priors Column**." The default is "**Equal**" priors.

- Specify the storage for the discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the DA computations.

Output example: The data set "**IRIS.xls**" was used for the Huber linear DA. It has 150 observations and four variables in three groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the Huber influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that three observations were misclassified. The cross validation results suggest the same.

Output for the Huber Linear Discriminant Analysis.
Data Set: IRIS (4 variables 3 groups).

	Linear Discriminant Analysis with Huber
User Selected Options	
Date/Time of Computation	1/18/2008 2:35:20 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\FULLIRIS
Full Precision	OFF
Influence Function Alpha	0.05
Squared MDs	Beta Distribution
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics
Total Number of Observations	150
Number of Selected Variables	4

Number of Data Rows per Group

1	2	3
50	50	50

Mean Vector for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
5.006	3.428	1.462	0.246

Covariance S Matrix for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.124	0.0992	0.0164	0.0103
0.0992	0.144	0.0117	0.0093
0.0164	0.0117	0.0302	0.00607
0.0103	0.0093	0.00607	0.0111

IQR Fix!

(Complete results are not shown.)

Output for the Huber Linear Discriminant Analysis (continued).

Classification Summary

	Predicted Membership		
Actual	1	2	3
1	50	0	0
2	0	48	2
3	0	1	49
# Correct	50	48	49
Prop Correct	100%	96%	98%

Total Observations	150
Correctly Classified	147
Incorrectly Classified	3

Misclassification Summary

Obs No.	Actual	Predicted
71	2	3
84	2	3
134	3	2

Apparent Error Rate 0.02

Linear Discriminant Function Constants and Coefficients

	1	2	3
Constant	-89.15	-74.4	-106.8
sp-length	23.15	15.7	12.59
sp-width	25.92	7.246	3.16
pt-length	-16.28	6.078	13.92
pt-width	-19.74	5.586	20.6

Output for the Huber Linear Discriminant Analysis (continued).

Cross Validation Results

Leave One Out (LOO) Cross Validation Results

LOO Classification Summary

	Predicted Membership		
Actual	1	2	3
1	50	0	0
2	0	48	2
3	0	1	49
# Correct	50	48	49
Prop Correct	100%	96%	98%

Total Observations	150
Correctly Classified	147
Incorrectly Classified	3

LOO Misclassification Summary

Obs No.	Actual	Predicted
71	2	3
84	2	3
134	3	2

LOO Error Rate 0.02

3 Fold Cross Validation Results

Average Error Rate: 0.2667

Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results

Validation Failed becuase of not enough Non-Outliers in Grouyp 9 times.

Average Correct Training Set: 147.2857

Average Incorrect Training Set: 2.7143

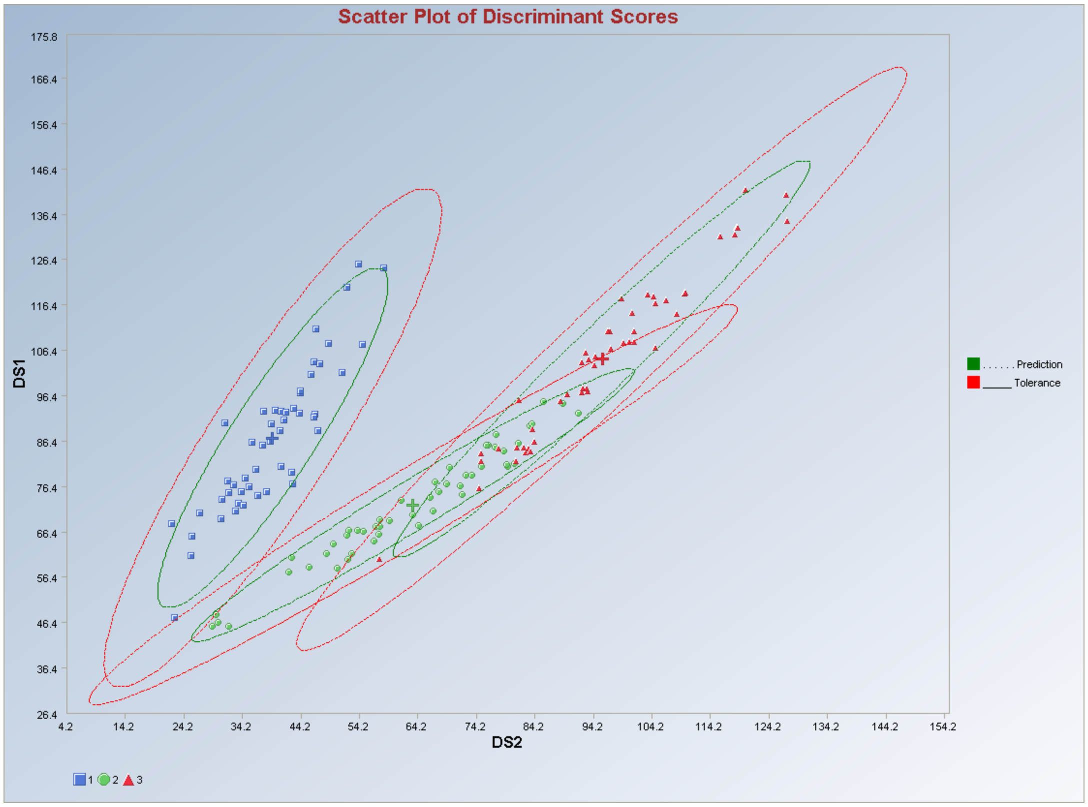
Average Correct Test Set: 146.8132

Average Incorrect Test Set: 3.1868

Error Rate Bias: -0.0032

Bias Adjusted Error Rate: 0.0232

Output for the Huber Linear Discriminant Analysis (continued).

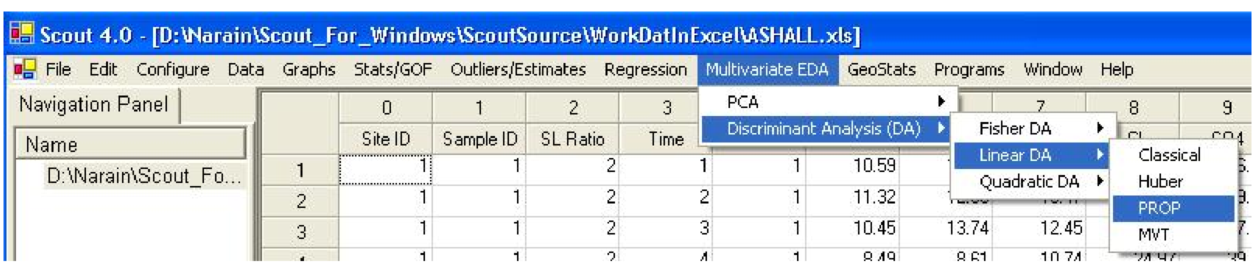


Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

***Note**: The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

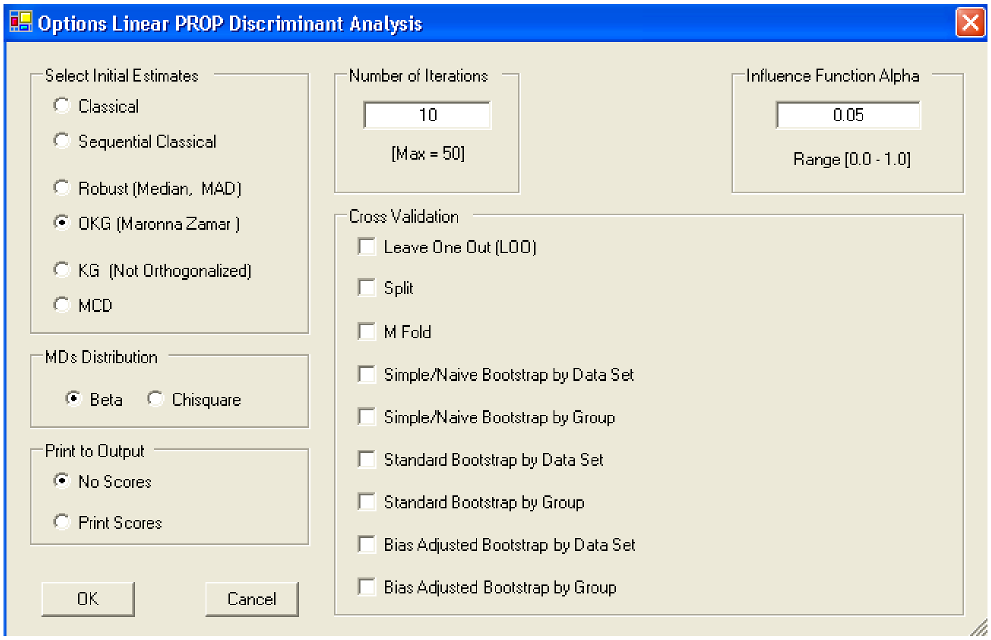
10.2.2.3 PROP Linear DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Linear DA ► PROP.**



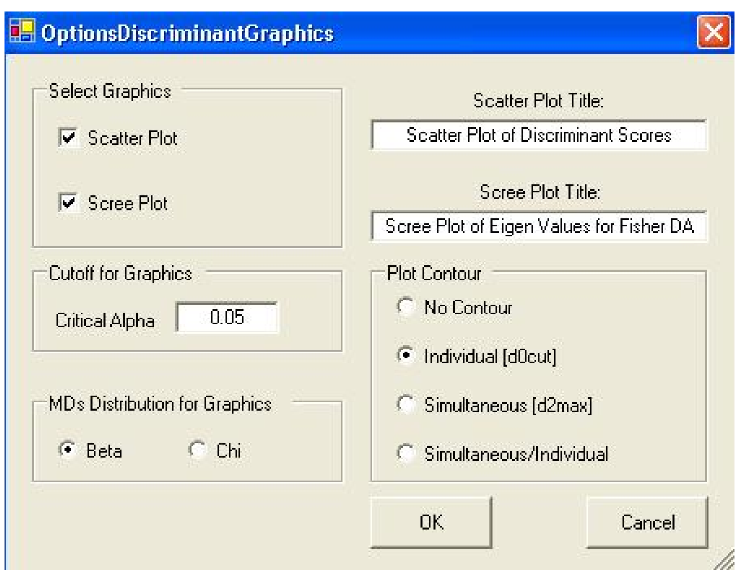
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



- Specify the options to calculate the robust estimates of the location and the scatter (scale or dispersion).
- Specify the "**Print to Output**." The default is "**No Scores**."
- Specify the preferred cross validation methods and their respective parameters.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The “**Scatter Plot**” provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is “**No Contour**.” Specify the distribution for distances and the “**Critical Alpha**” value for the cutoff to compute the ellipses. The defaults are “**Beta**” and “**0.05**.”

- Click on “**OK**” to continue or “**Cancel**” to cancel the graphics options.

* Specify the prior probabilities. The prior probabilities can be: “**Equal**” for all of the groups; “**Estimated**,” based on number of observations in each group; or “**User Supplied**,” where a column of priors can be obtained from the “**Select Group Priors Column**.” The default is “**Equal**” priors.

* Specify the storage for the discriminant scores. No scores will be stored when “**No Storage**” is selected. The scores will be stored in the data worksheet starting from the first available empty column when the “**Same Worksheet**” is selected. The scores will be stored in a new worksheet if the “**New Worksheet**” is selected. The default is “**No Storage**.”

* Click on “**OK**” to continue or “**Cancel**” to cancel the DA computations.

Output example: The data set “**ASHALL7grp.xls**” was used for the PROP linear DA. It has 214 observations and six variables in seven groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the PROP influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that six observations were misclassified. The cross validation results suggest the same.

Output for the PROP Linear Discriminant Analysis.
Data Set: Ashall (6 variables 7 groups).

	Linear Discriminant Analysis with PROP
User Selected Options	
Date/Time of Computation	1/18/2008 3:07:47 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\ASHALL7grp
Full Precision	OFF
Influence Function Alpha	0.05
Squared MDs	Beta Distribution
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics
Total Number of Observations	214
Number of Selected Variables	6

Number of Data Rows per Group

1	2	3	4	5	6	7
51	35	37	35	23	20	13

Mean Vector for Group 1

Ca-1	Na-1	K-1	Cl-1	SO4-1	ALK-1
10.02	16.81	17.22	32.35	34.86	0.508

Covariance S Matrix for Group 1

Ca-1	Na-1	K-1	Cl-1	SO4-1	ALK-1
7.599	-5.274	-5.41	-11.89	13.04	0.33
-5.274	8.901	8.475	14.42	-10.28	-0.309
-5.41	8.475	8.575	13.97	-10.47	-0.306
-11.89	14.42	13.97	29.6	-21.27	-0.555
13.04	-10.28	-10.47	-21.27	26.83	0.586
0.33	-0.309	-0.306	-0.555	0.586	0.0394

(Complete results are not shown.)

Output for the PROP Linear Discriminant Analysis (continued).

Classification Summary							
	Predicted Membership						
Actual	1	2	3	4	5	6	7
1	51	0	0	0	0	0	0
2	0	32	0	0	3	0	0
3	0	0	37	0	0	0	0
4	0	0	0	35	0	0	0
5	0	0	0	0	23	0	0
6	0	0	0	0	0	18	2
7	0	0	0	0	0	1	12
# Correct	51	32	37	35	23	18	12
Prop Correct	100%	91.43%	100%	100%	100%	90%	92.31%

Total Observations	214
Correctly Classified	208
Incorrectly Classified	6

Misclassification Summary		
Obs No.	Actual	Predicted
42	2	5
43	2	5
44	2	5
154	6	7
155	6	7
160	7	6

Apparent Error Rate 0.028

Linear Discriminant Function Constants and Coefficients

	1	**2**	**3**	**4**	**5**	**6**	**7**
Constant	-385.2	-181.4	-270.1	-179	-137	-134.9	-155.8
Ca	-0.455	-1.697	-1.708	2.892	0.46	2.198	3.595
Na	-1.252	4.025	5.277	0.42	0.413	0.573	0.238
K	20.89	-1.94	2.423	1.696	6.038	-1.306	1.907
Cl	2.01	5.015	4.279	4.729	3.067	4.518	4.019
SO4	10.39	5.206	7.884	3.468	4.722	1.626	2.135
ALK	10.04	12.74	14.11	8.793	10.05	9.101	8.284

Output for the PROP Linear Discriminant Analysis (continued).

Cross Validation Results
Split (50/50) Cross Validation Results
Error Rate for Training Set: 0.0827
Error Rate for Test Set: 0.0523
5 Fold Cross Validation Results
Average Error Rate: 0.0476
Standard Bootstrap (for whole dataset) for whole dataset
Error Rate from Bootstrap Training Set: 0.0234
Error Rate from Bootstrap Test Set: 0.0154
Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results
Average Correct Training Set: 209.6000
Average Incorrect Training Set: 4.4000
Average Correct Test Set: 207.8000
Average Incorrect Test Set: 6.2000
Error Rate Bias: -0.0084
Bias Adjusted Error Rate: 0.0364

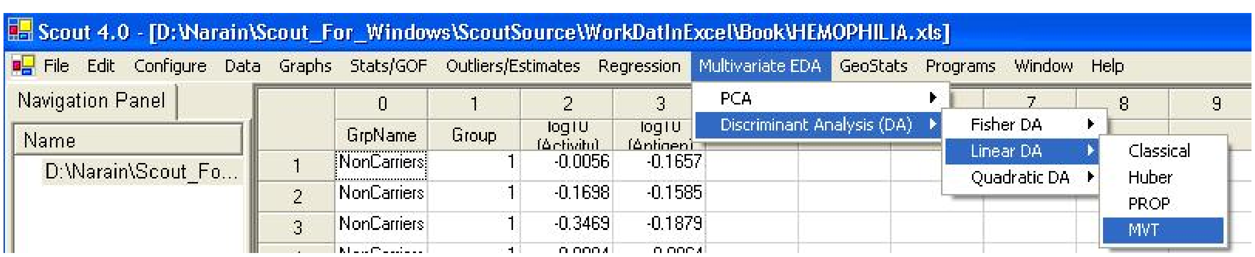
Output for the PROP Linear Discriminant Analysis (continued).

Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

***Note**: The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

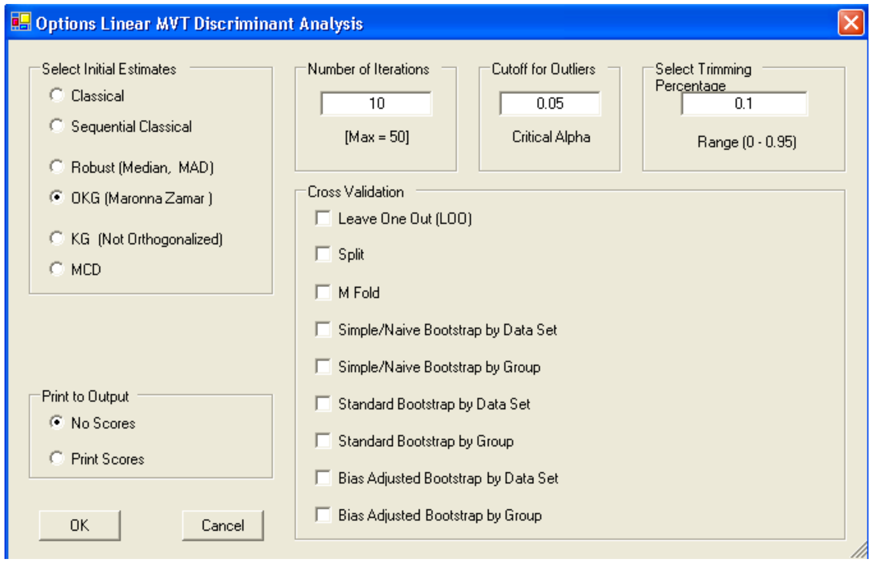
10.2.2.4 MVT Linear DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Linear DA ► MVT.**



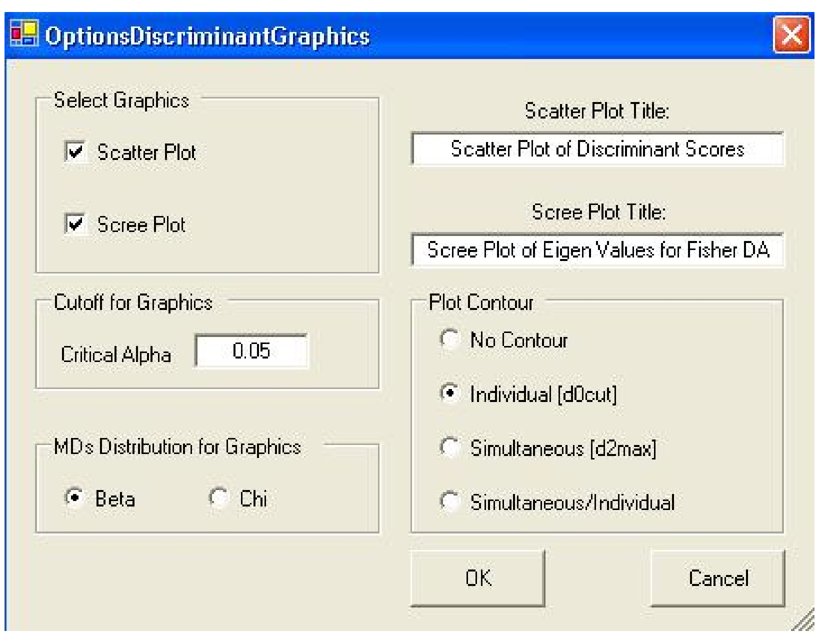
2. A "**Select Variables**" screen (Section 3.5) appears.

   - Click on the "**Options**" button for the options window.



- Specify the options to calculate the robust estimates of the location and the scatter (scale or dispersion).
- Specify the "**Print to Output**." The default is "**No Scores**."
- Specify the preferred cross validation methods and their respective parameters.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

• Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



  - The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

  - Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the prior probabilities. The prior probabilities can be: "**Equal**" for all of the groups; "**Estimated**," based on number of observations in each group; or "**User Supplied**," where a column of priors can be obtained from the "**Select Group Priors Column**." The default is "**Equal**" priors.

- Specify the storage of the discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the DA computations.

Output example: The data set "**Salmon.xls**" was used for the MVT linear DA. It has one 102 variables in two groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the trimming percentage and critical alpha and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that 13 observations were misclassified.

Output for the MVT Linear Discriminant Analysis.
Data Set: Salmon (2 variables 2 groups).

	Linear Discriminant Analysis Using MVT Method
User Selected Options	
Date/Time of Computation	1/18/2008 3:16:35 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Book\HEMOPHILIA
Full Precision	OFF
Trimming Percentage	10%
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics
Total Number of Observations	75
Number of Selected Variables	2

Number of Data Rows per Group

carriers	nonca~iers
46	29

Mean Vector for Group carriers

log10~iers	log10~iers
-0.303	-0.00708

Covariance S Matrix for Group carriers

log10~iers	log10~iers
0.0243	0.0148
0.0148	0.0236

Final Robust Mean Vector for Group carriers

log10~iers	log10~iers
-0.3	-0.00157

(Complete results are not shown.)

Output for the MVT Linear Discriminant Analysis (continued).

Classification Summary

	Predicted Membership	
Actual	carriers	noncarriers
carriers	37	9
noncarriers	4	25
# Correct	37	25
Prop Correct	80.43%	86.21%

Total Observations	75
Correctly Classified	62
Incorrectly Classified	13

Misclassification Summary

Obs No.	Actual	Predicted
3	noncarriers	carriers
5	noncarriers	carriers
7	noncarriers	carriers
17	noncarriers	carriers
30	carriers	noncarriers
35	carriers	noncarriers
58	carriers	noncarriers
60	carriers	noncarriers
62	carriers	noncarriers
63	carriers	noncarriers
64	carriers	noncarriers
67	carriers	noncarriers
69	carriers	noncarriers

Apparent Error Rate 0.173

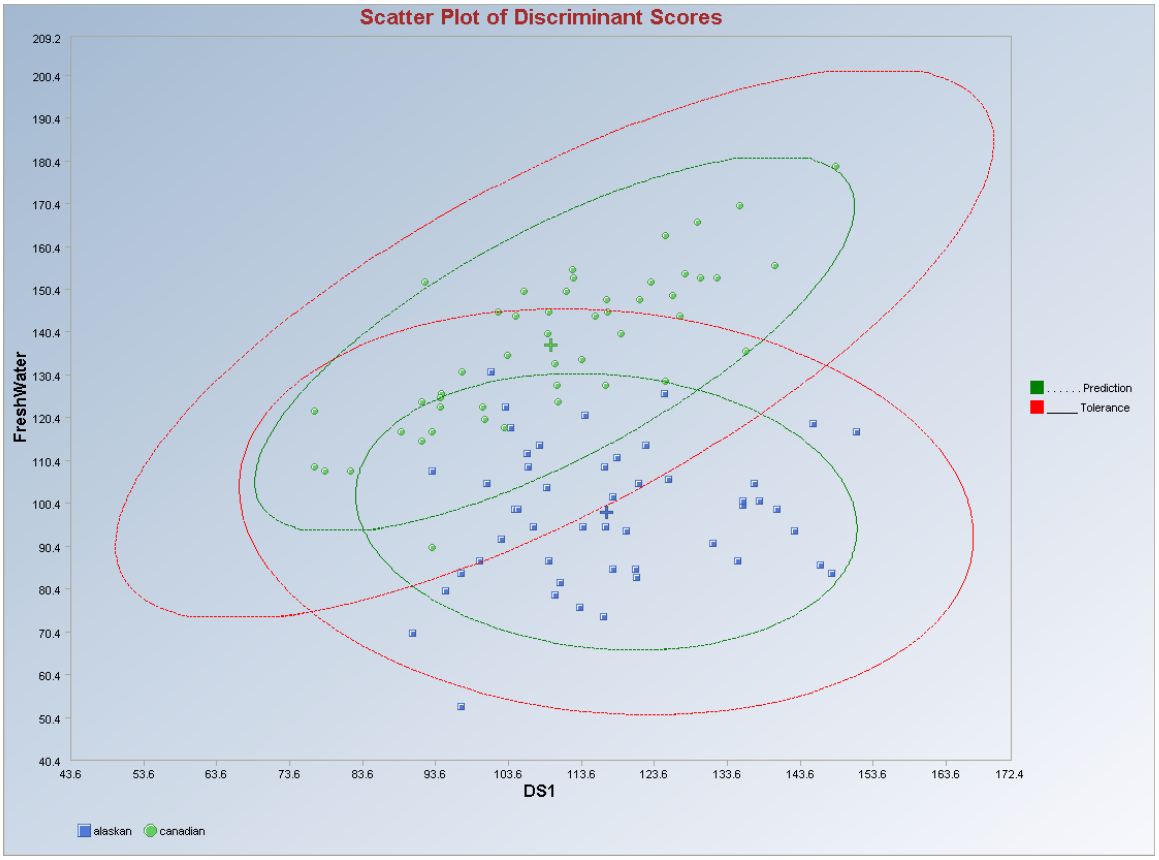
r Discriminant Function Constants and Coeffi

	carriers	noncarriers
Constant	-5.435	-1.285
log10(Activity)	-31.72	-9.478
log10(Antigen)	18.68	1.402

Output for the MVT Linear Discriminant Analysis (continued).

Cross Validation Results	
Simple/Naive Bootstrap (for whole dataset) Cross Validation Results	
Average Error Rate from Bootstrap: 0.0760	
Standard Bootstrap (for whole dataset) for whole dataset	
Error Rate from Bootstrap Training Set: 0.0730	
Error Rate from Bootstrap Test Set: 0.0330	
Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results	
Average Correct Training Set: 92.9000	
Average Incorrect Training Set: 7.1000	
Average Correct Test Set: 92.9000	
Average Incorrect Test Set: 7.1000	
Error Rate Bias: 0.0000	
Bias Adjusted Error Rate: 0.0700	

Output for the MVT Linear Discriminant Analysis (continued).



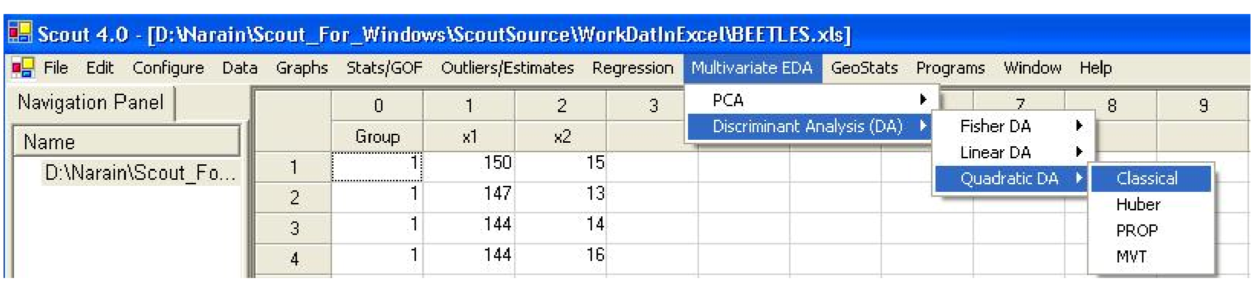
Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

Note: *The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

10.2.3 Quadratic Discriminant Analysis

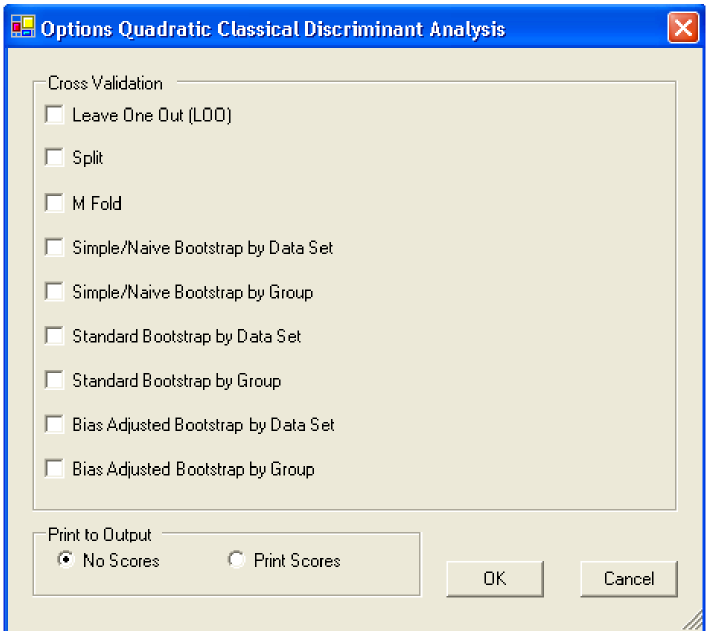
10.2.3.1 Classical Quadratic DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Quadratic DA ► Classical.**



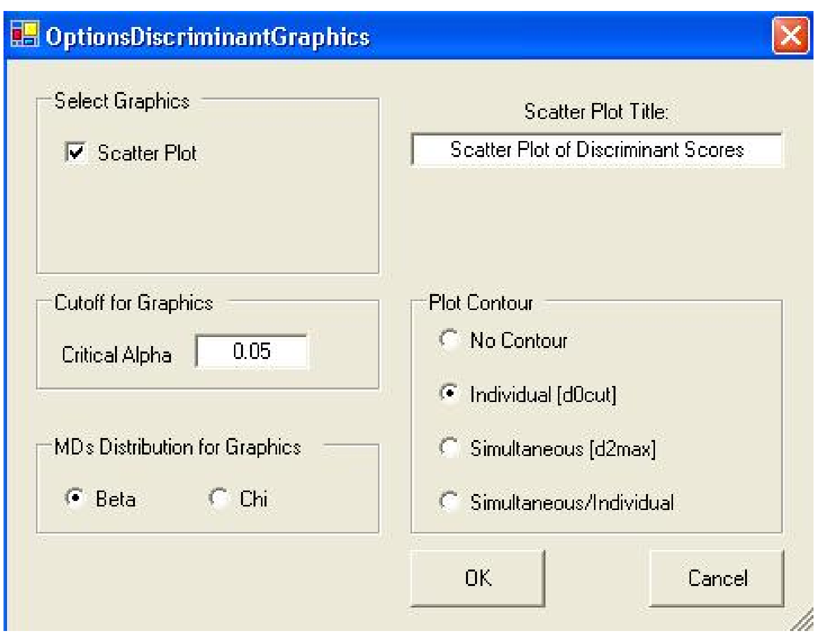
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



  - Specify the preferred cross validation methods and their respective parameters.
  - Specify the "**Print to Output**." The default is "**No Scores**."
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The “**Scatter Plot**” provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is “**No Contour**.” Specify the distribution for distances and the “**Critical Alpha**” value for the cutoff to compute the ellipses. The defaults are “**Beta**” and “**0.05**.”

- Click on “**OK**” to continue or “**Cancel**” to cancel the graphics options.

* Specify the prior probabilities. The prior probabilities can be: “**Equal**” for all of the groups; “**Estimated**,” based on the number of observations in each group; or “**User Supplied**,” where a column of priors can be obtained from the “**Select Group Priors Column**.” The default is “**Equal**” priors.

* Specify the storage of discriminant scores. No scores will be stored when “**No Storage**” is selected. The scores will be stored in the data worksheet starting from the first available empty column when the “**Same Worksheet**” is selected. The scores will be stored in a new worksheet if the “**New Worksheet**” is selected. The default is “**No Storage**.”

* Click on “**OK**” to continue or “**Cancel**” to cancel the DA computations.

Output example: The data set “**BEETLES.xls**” was used for the quadratic linear DA. It has 74 observations and two variables in three groups. The initial estimates of location and scale for each group were the classical mean and the covariance matrix. The classification rules were obtained using those estimates. The output shows that one observation was misclassified.

Output for the Classical Quadratic Discriminant Analysis.
Data Set: Beetles (2 variables 3 groups).

	Classical Quadratic Discriminant Analysis
User Selected Options	
Date/Time of Computation	1/18/2008 3:23:37 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BEETLES
Full Precision	OFF
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics

Total Number of Observations	74
Number of Selected Variables	2

Number of Data Rows per Group

1	2	3
21	31	22

Mean Vector for Group 1

x1-1	x2-1
146.2	14.1

Covariance S Matrix for Group 1

x1-1	x2-1
31.66	-0.969
-0.969	0.79

Mean Vector for Group 2

x1-2	x2-2
124.6	14.29

Covariance S Matrix for Group 2

x1-2	x2-2
21.37	-0.327
-0.327	1.213

(Complete results are not shown.)

Output for the Classical Quadratic Discriminant Analysis (continued).

Classification Summary				
	Predicted Membership			
Actual	1	2	3	
1	20	1	0	
2	0	31	0	
3	0	0	22	
# Correct	20	31	22	
Prop Correct	95.24%	100%	100%	
		Total Observations	74	
		Correctly Classified	73	
		Incorrectly Classified	1	
Misclassification Summary				
Obs No.	Actual	Predicted		
17	1	2		
			Apparent Error Rate	0.0135
				Cross Validation Results
Leave One Out (LOO) Cross Validation Results				
LOO Classification Summary				
	Predicted Membership			
Actual	1	2	3	
1	20	1	0	
2	0	31	0	
3	0	0	22	
# Correct	20	31	22	
Prop Correct	95.24%	100%	100%	
		Total Observations	74	
		Correctly Classified	73	
		Incorrectly Classified	1	

Output for the Classical Quadratic Discriminant Analysis (continued).

LOO Misclassification Summary				
Obs No.	Actual	Predicted		
17	1	2		
			LOO Error Rate	0.0135

Split (50/50) Cross Validation Results

Error Rate for Training Set: 0.0000

Error Rate for Test Set: 0.0081

3 Fold Cross Validation Results

Average Error Rate: 0.0267

Simple/Naive Bootstrap (for whole dataset) Cross Validation Results

Average Error Rate from Bootstrap: 0.0068

Standard Bootstrap (for whole dataset) Cross Validation Results

Error Rate from Bootstrap Training Set: 0.0041

Error Rate from Bootstrap Test Set: 0.0081

Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results

Average Correct Training Set: 73.8000

Average Incorrect Training Set: 0.2000

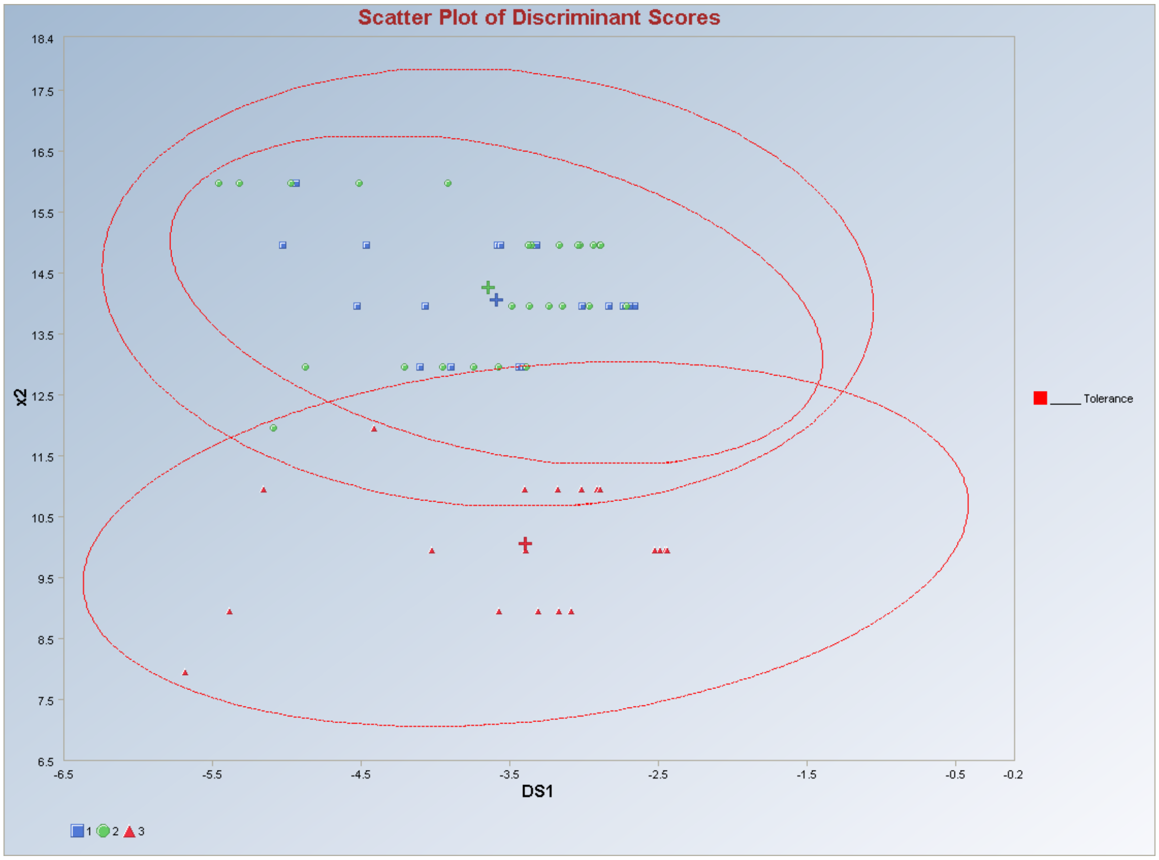
Average Correct Test Set: 72.7000

Average Incorrect Test Set: 1.3000

Error Rate Bias: -0.0149

Bias Adjusted Error Rate: 0.0284

Output for the Classical Quadratic Discriminant Analysis (continued).

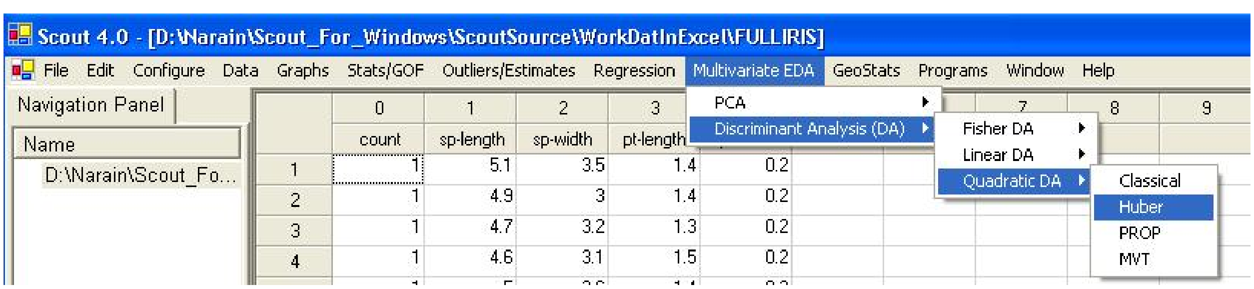


Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

Note: *The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

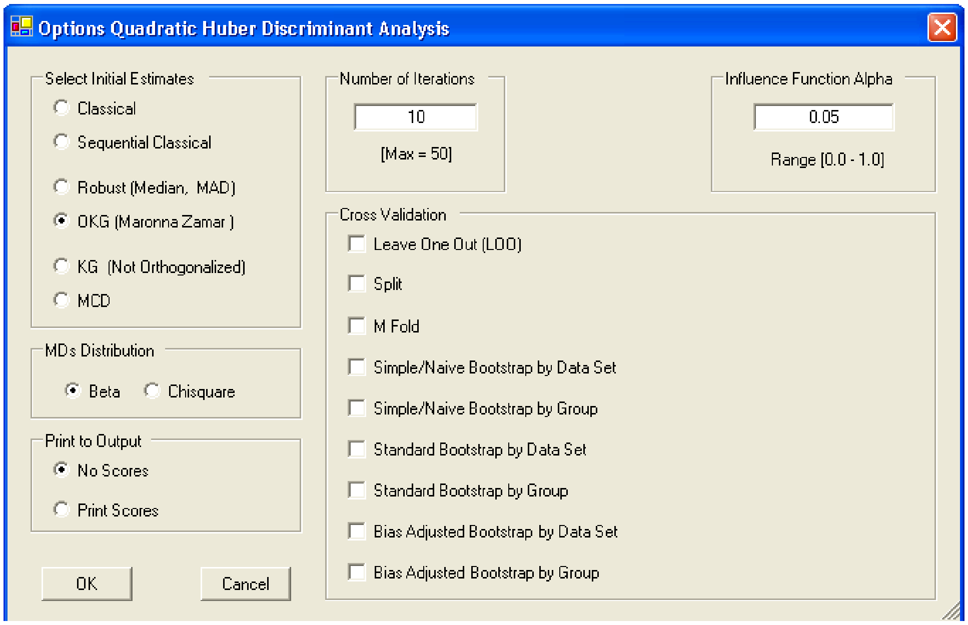
10.2.3.2 Huber Quadratic DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Quadratic DA ► Huber.**



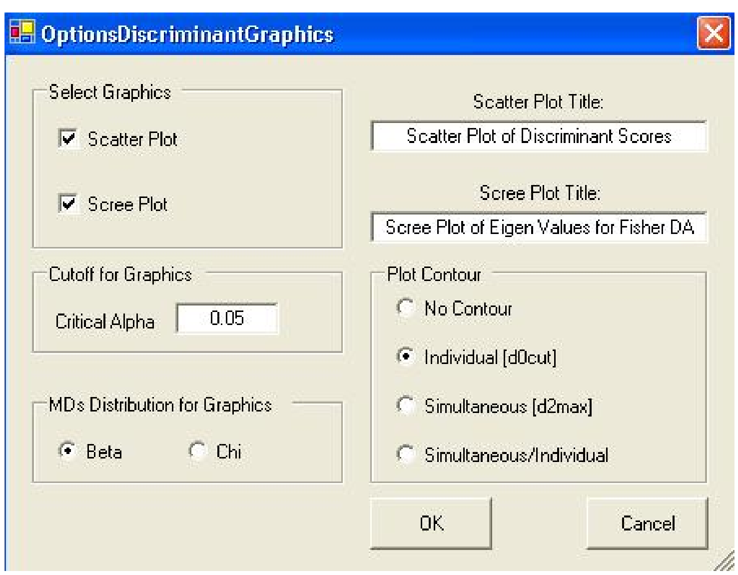
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



  - Specify the options to calculate the robust estimates of the location and the scatter (scale or dispersion).
  - Specify the "**Print to Output**." The default is "**No Scores**."
  - Specify the preferred cross validation methods and their respective parameters.
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the prior probabilities. The prior probabilities can be: "**Equal**" for all of the groups; "**Estimated**," based on number of observations in each group; or "**User Supplied**," where a column of priors can be obtained from the "**Select Group Priors Column**." The default is "**Equal**" priors.

- Specify the storage of discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the DA computations.

Output example: The data set "**IRIS.xls**" was used for the Huber quadratic DA. It has 150 observations and four variables in three groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the Huber influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that three observations were misclassified. The cross validation results suggest the same.

Output for the Huber Quadratic Discriminant Analysis.
Data Set: IRIS (4 variables 3 groups).

	Quadratic Discriminant Analysis with Huber
User Selected Options	
Date/Time of Computation	1/18/2008 3:30:55 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\FULLIRIS
Full Precision	OFF
Influence Function Alpha	0.05
Squared MDs	Beta Distribution
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05) snd Max MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics
Total Number of Observations	150
Number of Selected Variables	4

Number of Data Rows per Group

1	2	3
50	50	50

Mean Vector for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
5.006	3.428	1.462	0.246

Covariance S Matrix for Group 1

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.124	0.0992	0.0164	0.0103
0.0992	0.144	0.0117	0.0093
0.0164	0.0117	0.0302	0.00607
0.0103	0.0093	0.00607	0.0111

IQR Fix!

(Complete results are not shown.)

Output for the Huber Quadratic Discriminant Analysis (continued).

Classification Summary

	Predicted Membership		
Actual	1	2	3
1	50	0	0
2	0	48	2
3	0	1	49
# Correct	50	48	49
Prop Correct	100%	96%	98%

Total Observations	150
Correctly Classified	147
Incorrectly Classified	3

Misclassification Summary

Obs No.	Actual	Predicted
71	2	3
84	2	3
134	3	2

Apparent Error Rate 0.02

Cross Validation Results

Split (50/50) Cross Validation Results

Error Rate for Training Set: 0.0053

Error Rate for Test Set: 0.0493

3 Fold Cross Validation Results

Average Error Rate: 0.2667

Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results

Average Correct Training Set: 133.6000

Average Incorrect Training Set: 1.4000

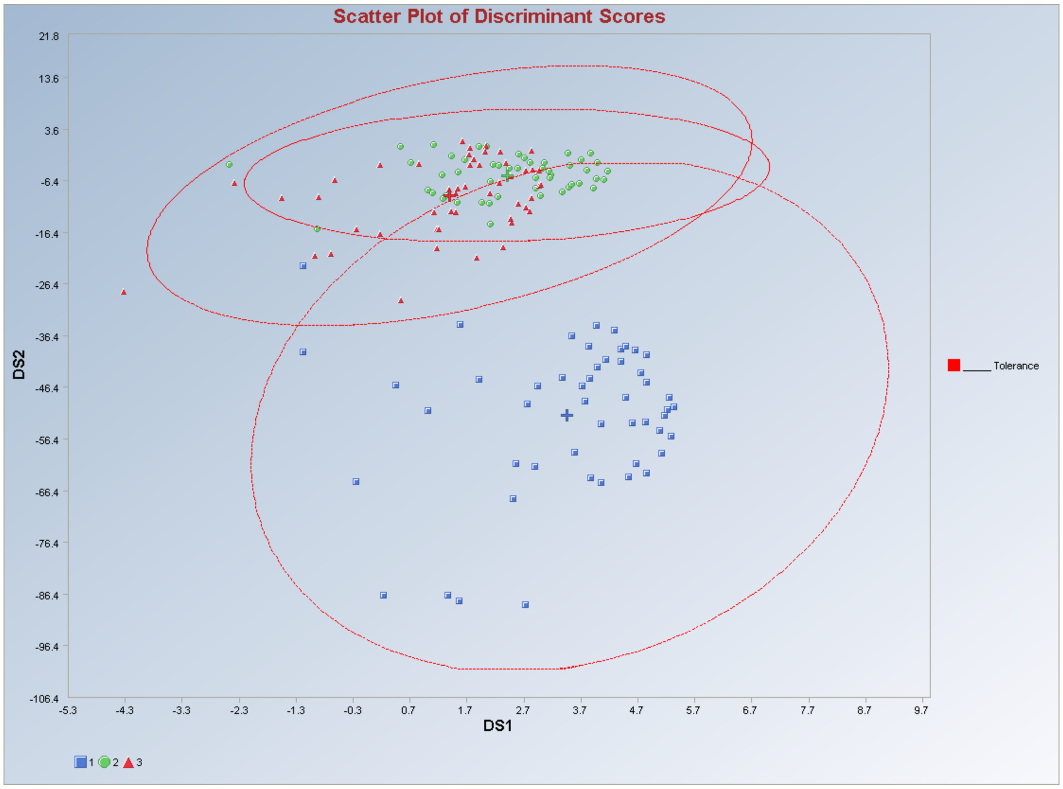
Average Correct Test Set: 137.6000

Average Incorrect Test Set: 12.4000

Error Rate Bias: -0.0733

Bias Adjusted Error Rate: 0.0933

Output for the Huber Quadratic Discriminant Analysis (continued).

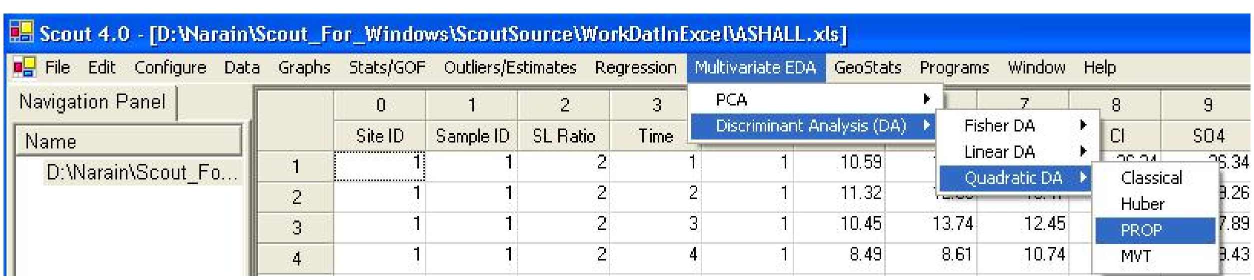


Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

***Note**: The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

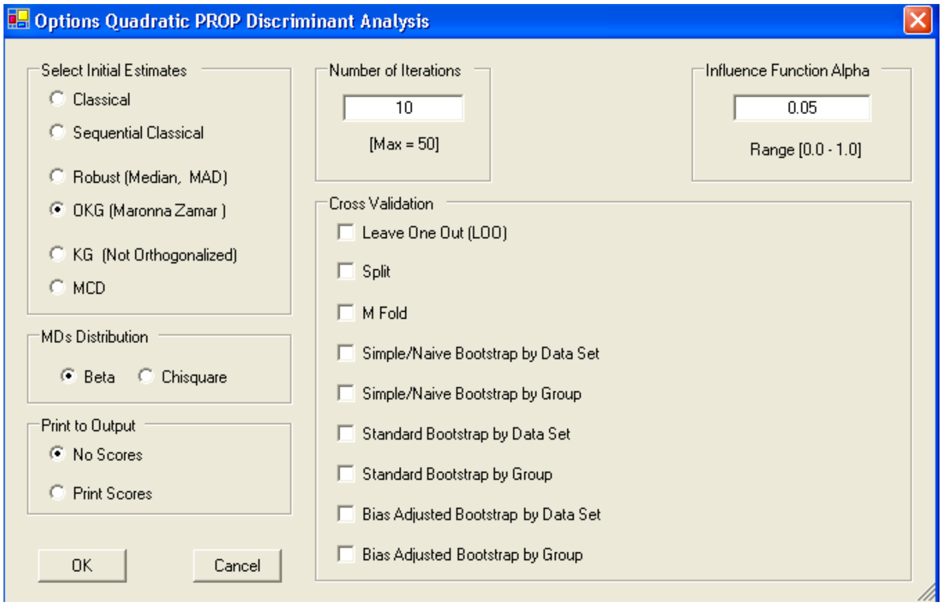
10.2.3.3 PROP Quadratic DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Quadratic DA ► PROP.**



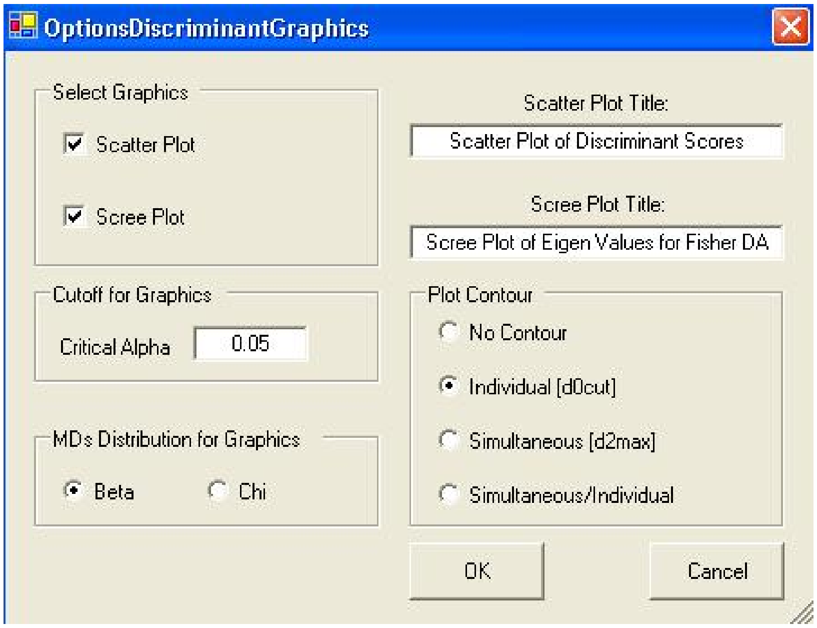
2. A "**Select Variables**" screen (Section 3.5) appears.

   - Click on the "**Options**" button for the options window.



- Specify the options to calculate the robust estimates of the location and the scatter (scale or dispersion).
- Specify the “**Print to Output**.” The default is “**No Scores**.”
- Specify the preferred cross validation methods and their respective parameters.
- Click “**OK**” to continue or “**Cancel**” to cancel the options.

- Click on the “**Graphics**” button for the graphics options window and check all of the preferred check boxes.



- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the prior probabilities. The prior probabilities can be: "**Equal**" for all of the groups; "**Estimated**," based on number of observations in each group; or "**User Supplied**," where a column of priors can be obtained from the "**Select Group Priors Column**." The default is "**Equal**" priors.

- Specify the storage of discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the DA computations.

Output example: The data set "**ASHALL7grp.xls**" was used for the PROP quadratic DA. It has 214 observations and six variables in seven groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the PROP influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that seven observations were misclassified. The cross validation results suggest the same.

Output for the PROP Quadratic Discriminant Analysis.
Data Set: Ashall (6 variables 7 groups).

	Quadratic Discriminant Analysis with PROP
User Selected Options	
Date/Time of Computation	1/18/2008 3:39:25 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\ASHALL7grp
Full Precision	OFF
Influence Function Alpha	0.05
Squared MDs	Beta Distribution
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics
Total Number of Observations	214
Number of Selected Variables	6

Number of Data Rows per Group

1	2	3	4	5	6	7
51	35	37	35	23	20	13

Mean Vector for Group 1

Ca-1	Na-1	K-1	Cl-1	SO4-1	ALK-1
10.02	16.81	17.22	32.35	34.86	0.508

Covariance S Matrix for Group 1

Ca-1	Na-1	K-1	Cl-1	SO4-1	ALK-1
7.599	-5.274	-5.41	-11.89	13.04	0.33
-5.274	8.901	8.475	14.42	-10.28	-0.309
-5.41	8.475	8.575	13.97	-10.47	-0.306
-11.89	14.42	13.97	29.6	-21.27	-0.555
13.04	-10.28	-10.47	-21.27	26.83	0.586
0.33	-0.309	-0.306	-0.555	0.586	0.0394

(Complete output is not shown.)

Output for the PROP Quadratic Discriminant Analysis (continued).

Classification Summary

	Predicted Membership						
Actual	1	2	3	4	5	6	7
1	51	0	0	0	0	0	0
2	0	31	4	0	0	0	0
3	0	0	37	0	0	0	0
4	0	0	1	34	0	0	0
5	0	0	1	0	22	0	0
6	0	0	1	0	0	19	0
7	0	0	0	0	0	0	13
# Correct	51	31	37	34	22	19	13
Prop Correct	100%	88.57%	100%	97.14%	95.65%	95%	100%

Total Observations	214
Correctly Classified	207
Incorrectly Classified	7

Misclassification Summary

Obs No.	Actual	Predicted
42	2	3
43	2	3
66	2	3
67	2	3
143	5	3
195	4	3
211	6	3

Apparent Error Rate 0.0327

Output for the PROP Quadratic Discriminant Analysis (continued).

Cross Validation Results

Leave One Out (LOO) Cross Validation Resul

LOO Classification Summary

	Predicted Membership						
Actual	1	2	3	4	5	6	7
1	51	0	0	0	0	0	0
2	0	30	5	0	0	0	0
3	0	0	37	0	0	0	0
4	0	0	0	35	0	0	0
5	0	0	1	0	22	0	0
6	0	0	3	0	0	17	0
7	0	0	3	0	0	0	10
# Correct	51	30	37	35	22	17	10
Prop Correct	100%	85.71%	100%	100%	95.65%	85%	76.92%

Total Observations	214
Correctly Classified	202
Incorrectly Classified	12

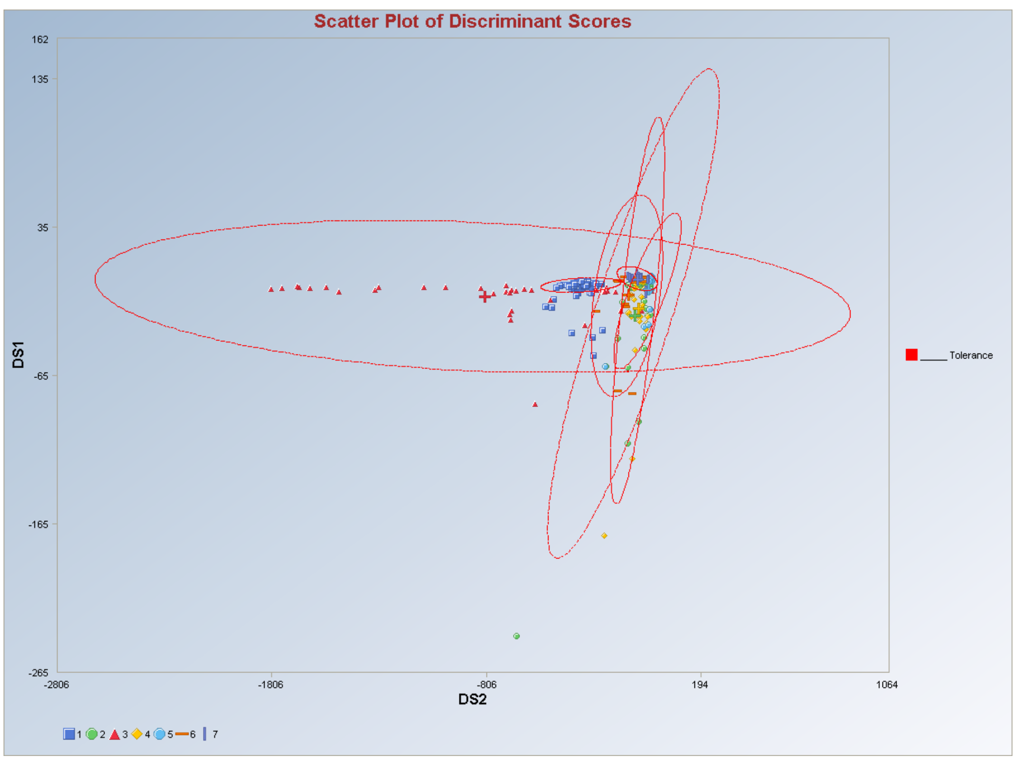
LOO Misclassification Summary

Obs No.	Actual	Predicted
42	2	3
43	2	3
66	2	3
67	2	3
68	2	3
143	5	3
145	6	3
152	6	3
158	6	3
163	7	3
164	7	3
170	7	3

LOO Error Rate 0.0561

Output for the PROP Quadratic Discriminant Analysis (continued).

Split (50/50) Cross Validation Results	
Validation Failed Not Enough Non-Outliers 9 times.	
Error Rate for Training Set: 0.0561	
Error Rate for Test Set: 0.0327	
Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results	
Average Correct Training Set: 177.7000	
Average Incorrect Training Set: 36.3000	
Average Correct Test Set: 184.3000	
Average Incorrect Test Set: 29.7000	
Error Rate Bias: 0.0308	
Bias Adjusted Error Rate: 0.0636	

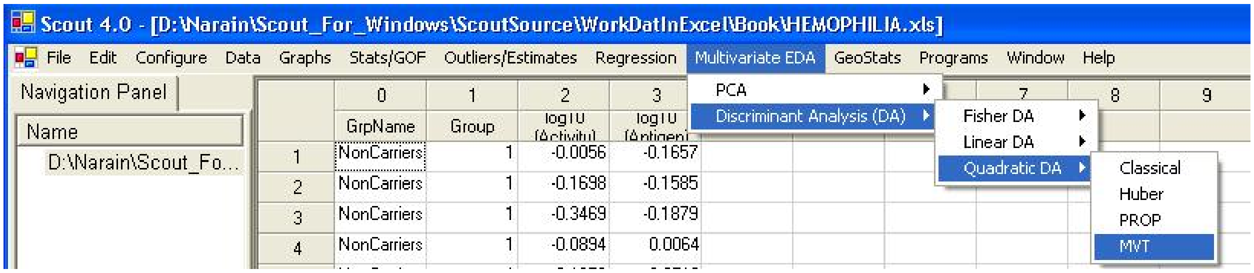


Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

Note: *The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

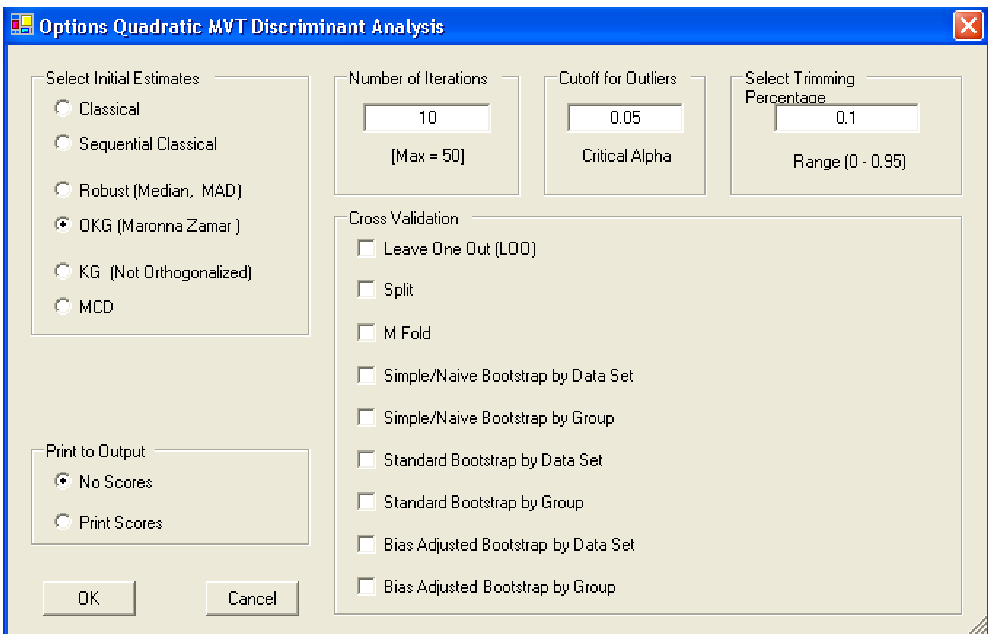
10.2.3.4 MVT Quadratic DA

1. Click on **Multivariate EDA ► Discriminant Analysis (DA) ► Quadratic DA ► MVT.**



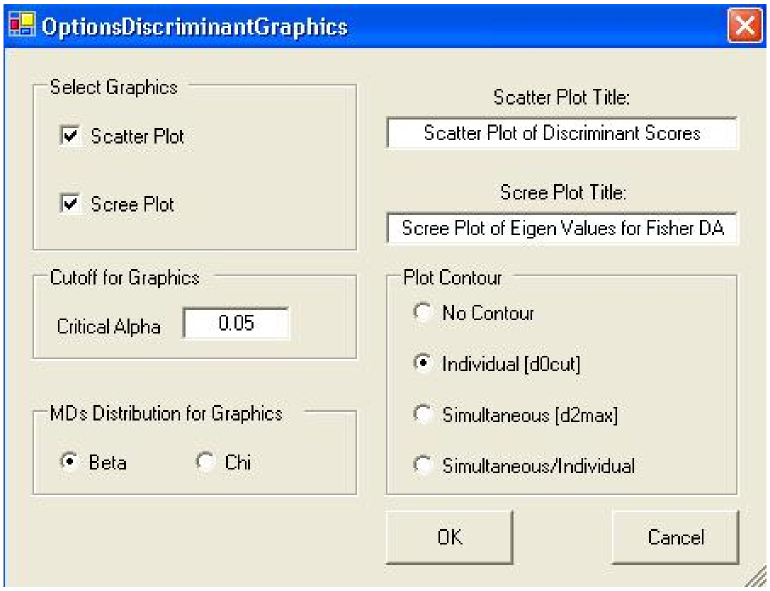
2. A "**Select Variables**" screen (Section 3.5) appears.

- Click on the "**Options**" button for the options window.



  - Specify the options to calculate the robust estimates of the location and the scatter (scale or dispersion).
  - Specify the "**Print to Output**." The default is "**No Scores**."
  - Specify the preferred cross validation methods and their respective parameters.
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**Graphics**" button for the graphics options window and check all of the preferred check boxes.



- The "**Scatter Plot**" provides the scatter plot of the discriminant analysis scores and also of the selected variables. The user has the option of drawing contours on the scatter plot to identify any outliers. The default is "**No Contour**." Specify the distribution for distances and the "**Critical Alpha**" value for the cutoff to compute the ellipses. The defaults are "**Beta**" and "**0.05**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the graphics options.

- Specify the prior probabilities. The prior probabilities can be: "**Equal**" for all of the groups; "**Estimated**," based on number of observations in each group, or "**User Supplied**," where a column of priors can be obtained from the "**Select Group Priors Column**." The default is "**Equal**" priors.

- Specify the storage of the discriminant scores. No scores will be stored when "**No Storage**" is selected. The scores will be stored in the data worksheet starting from the first available empty column when the "**Same Worksheet**" is selected. The scores will be stored in a new worksheet if the "**New Worksheet**" is selected. The default is "**No Storage**."

- Click on "**OK**" to continue or "**Cancel**" to cancel the DA computations.

Output example: The data set "**Salmon.xls**" was used for the MVT quadratic DA. It has one 102 variables in two groups. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the trimming percentage and critical alpha and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were calculated. The classification rules were obtained using those weighted estimates. The output shows that six observations were misclassified. The cross validation results suggest the same.

Output for the MVT Quadratic Discriminant Analysis.
Data Set: Salmon (2 variables 2 groups).

	Quadratic Discriminant Analysis Using MVT Method
User Selected Options	
Date/Time of Computation	1/18/2008 3:48:10 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Book\SALMON
Full Precision	OFF
Trimming Percentage	10%
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors will be used
Graphics Options	Scatter Plots selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05) snd Max MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics

Total Number of Observations	100
Number of Selected Variables	2

Number of Data Rows per Group

alaskan	canadian
50	50

Mean Vector for Group alaskan

Fresh~skan	Marin~skan
98.38	429.7

Covariance S Matrix for Group alaskan

Fresh~skan	Marin~skan
260.6	-188.1
-188.1	1399

Final Robust Mean Vector for Group alaskan

Fresh~skan	Marin~skan
98.42	429.8

(Complete output is not shown.)

Output for the MVT Quadratic Discriminant Analysis (continued).

Classification Summary		
	Predicted Membership	
Actual	alaskan	canadian
alaskan	47	3
canadian	3	47
# Correct	47	47
Prop Correct	94%	94%

Total Observations	100
Correctly Classified	94
Incorrectly Classified	6

Misclassification Summary		
Obs No.	Actual	Predicted
2	alaskan	canadian
12	alaskan	canadian
13	alaskan	canadian
51	canadian	alaskan
68	canadian	alaskan
71	canadian	alaskan

Apparent Error Rate 0.06

Cross Validation Results

Leave One Out (LOO) Cross Validation Results

LOO Classification Summary		
	Predicted Membership	
Actual	alaskan	canadian
alaskan	46	4
canadian	3	47
# Correct	46	47
Prop Correct	92%	94%

Total Observations	100
Correctly Classified	93
Incorrectly Classified	7

Output for the MVT Quadratic Discriminant Analysis (continued).

LOO Misclassification Summary				
Obs No.	Actual	Predicted		
2	alaskan	canadian		
12	alaskan	canadian		
13	alaskan	canadian		
30	alaskan	canadian		
51	canadian	alaskan		
68	canadian	alaskan		
71	canadian	alaskan		
			LOO Error Rate	0.07

Bias Adjusted Bootstrap (for whole dataset) Cross Validation Results

Average Correct Training Set: 90.9000

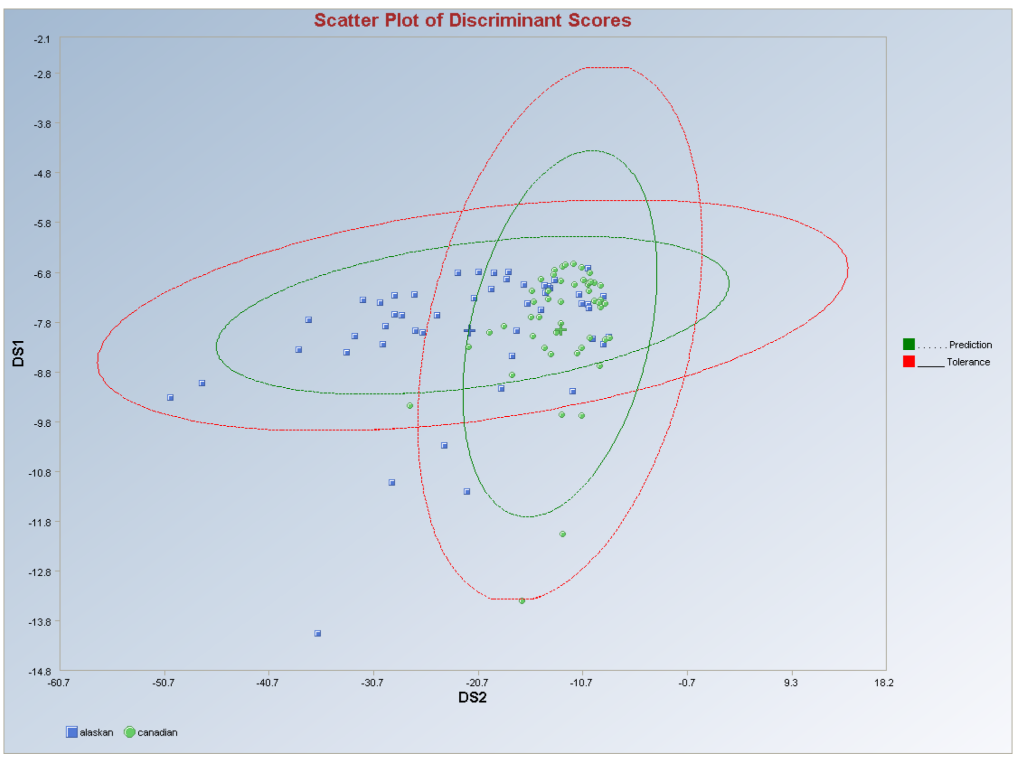
Average Incorrect Training Set: 9.1000

Average Correct Test Set: 92.6000

Average Incorrect Test Set: 7.4000

Error Rate Bias: 0.0170

Bias Adjusted Error Rate: 0.0770



Observations outside of the simultaneous (Tolerance) ellipses are considered to be anomalous. Observations between the individual and the simultaneous ellipses are considered to be discordant.

***Note**: The drop-down bars in the graphics toolbar can be used to obtain different scatter plots of the discriminant scores and the variables, as explained in Chapter 2.*

10.2.4 Classification of Unknown Observations

Unknown or new observations can be classified into existing groups. There are certain rules that need to be followed when using the unknown or new observations.

- The first three letters of the group name of the new or unknown observations should be “**UNK**” or “**unk**” only.

- The set of unknown or new observations should be the last set of observations in a data set; otherwise, an error message is obtained.

- Unknown or new observations will not be used in the cross validation.

- Unknown or new observations will not be used in the graphs.

- The results of the classification of the unknown observations are printed at the end of the output sheet.

Last set of observations.

	0	1	2	3	4	5	6	7	8	9	10	11
	Site ID	Sample ID	SL Ratio	Time	Id5	Ca	Na	K	Cl	SO4	ALK	
188	3	1	2	2	6	15.11	12.81	6.01	17.52	19.56	18.34	
189	3	1	4	2	9	5.35	18.57	7.91	18.07	21.55	13.97	
190	3	1	4	2	1	10.08	21.09	10.74	27.15	22.06	10.73	
191	3	1	4	2	3	9.48	18.88	8.96	22.14	23.49	8.78	
192	3	1	4	2	4	10.3	17.32	8.09	24.39	23.66	8.49	
193	3	1	4	2	5	10.2	17.29	8.06	23.62	19.18	10.47	
194	3	1	4	2	6	9.11	19.03	8.98	25.41	21.32	11.87	
195	4	1	2	2	9	34.34	7.62	6.02	48.78	17.27	5.7	
196	4	1	2	2	1	23.62	5.48	4.27	35.18	13.13	5.07	
197	4	1	2	2	2	22.65	5.03	4.03	37.46	12.41	4.36	
198	4	1	2	2	3	21.95	5.07	3.84	32.3	11.89	5.86	
199	4	1	2	2	4	23.99	5.53	4.24	33.26	12.35	10.33	
200	4	1	4	2	9	25.56	6.82	5.21	38.87	12.37	4.38	
201	4	1	4	2	1	22.29	7.11	5.45	39.54	11.65	3.24	
202	4	1	4	2	2	26.39	7.49	5.87	42.33	10.72	1.63	
203	4	1	4	2	3	23.24	6.87	5.26	39.98	12.36	3.35	
204	4	1	4	2	4	24.76	6.78	5.28	40.83	12.59	2.23	
205	5	1	2	2	9	15.47	4.29	3.96	9.65	12.83	13.76	
206	5	1	2	2	1	13.23	4.76	4.22	10.48	13.22	13.63	
207	5	1	2	2	2	12.52	5.94	5.12	12.76	15.39	12.78	
208	5	1	4	2	9	14.06	6.12	5.44	13.58	12.69	12.62	
209	5	1	4	2	1	11.96	6.19	5.49	13.28	12.52	13.99	
210	5	1	4	2	2	10.52	8.13	7.4	17.99	14.63	10.79	
211	6	1	2	2	9	18.51	2.43	1.62	7.29	1.04	12.19	
212	6	1	2	2	1	18.45	2.41	1.67	19.62	0.43	14.99	
213	6	1	4	2	9	21.25	4.27	2.84	28.12	1.27	11.61	
214	6	1	4	2	1	22.85	4.45	3.13	31.94	0.46	10.18	
215	UNK	1	5	4	1	22.59	6.9	7.35	44.05	2.27	3.59	
216	unk	1	6	2	1	23.83	7.59	8.04	47.71	2.05	2.66	
217	UNK	1	6	4	1	25.49	7.78	8.21	49.96	2.19	2.29	
218												
219												
220												
221												
222												
223												
224												

Unknown observations in-between data.

	0	1	2	3	4	5	6	7	8	9	10	11
	Site ID	Sample ID	SL Ratio	Time	Id5	Ca	Na	K	Cl	SO4	ALK	
188	3	1	2	2	6	15.11	12.81	6.01	17.52	19.56	18.34	
189	3	1	4	2	9	5.35	18.57	7.91	18.07	21.55	13.97	
190	3	1	4	2	1	10.08	21.09	10.74	27.15	22.06	10.73	
191	3	1	4	2	3	9.48	18.88	8.96	22.14	23.49	8.78	
192	UNK	1	6	4	1	25.49	7.78	8.21	49.96	2.19	2.29	
193	3	1	4	2	5	10.2	17.29	8.06	23.62	19.18	10.47	
194	3	1	4	2	6	9.11	19.03	8.98	25.41	21.32	11.87	
195	4	1	2	2	9	34.34	7.62	6.02	48.78	17.27	5.7	
196	4	1	2	2	1	23.62	5.48	4.27	35.18	13.13	5.07	
197	4	1	2	2	2	22.65	5.03	4.03	37.46	12.41	4.36	
198	4	1	2	2	3	21.95	5.07	3.84	32.3	11.89	5.86	
199	4	1	2	2	4	23.99	5.53	4.24	33.26	12.35	10.33	
200	UNK	1	5	4	1	22.59	6.9	7.35	44.05	2.27	3.59	
201	4	1	4	2	1	22.29	7.11	5.45	39.54	11.65	3.24	
202	4	1	4	2	2	26.39	7.49	5.87	42.33	10.72	1.63	
203	4	1	4	2	3	23.24	6.87	5.26	39.98	12.36	3.35	
204	4	1	4	2	4	24.76	6.78	5.28	40.83	12.59	2.23	
205	5	1	2	2	9	15.47	4.29	3.96	9.65	12.83	13.76	
206	5	1	2	2	1	13.23	4.76	4.22	10.48	13.22	13.63	
207	unk	1	6	2	1	23.83	7.59	8.04	47.71	2.05	2.66	
208	5	1	4	2	9	14.06	6.12	5.44	13.58	12.69	12.62	
209	5	1	4	2	1	11.96	6.19	5.49	13.28	12.52	13.99	
210	5	1	4	2	2	10.52	8.13	7.4	17.99	14.63	10.79	
211	6	1	2	2	9	18.51	2.43	1.62	7.29	1.04	12.19	
212	6	1	2	2	1	18.45	2.41	1.67	19.62	0.43	14.99	
213	6	1	4	2	9	21.25	4.27	2.84	28.12	1.27	11.61	
214	6	1	4	2	1	22.85	4.45	3.13	31.94	0.46	10.18	
215												
216												
217												
218												
219												
220												
221												
222												

Error Message.

	Robust Fisher Linear Discriminant Analysis using Huber Influence Function
User Selected Options	
Date/Time of Computation	1/16/2008 10:34:14 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\ASHALL.xls
Full Precision	OFF
Influence Function Alpha	0.05
Squared MDs	Beta Distribution
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Number of Iterations	10
Storage Options	No Discriminant Scores will be stored to Worksheet
Group Probabilities:	Equal Priors Assumed
Graphics Options	Both Scree Plot and Scatter Plots are Selected
Contour Options	Contour Ellipses drawn using Individual MD(0.05)
Alpha for Graphics	0.05
Distribution of MDs	Beta Distribution used in Graphics

Unknown Group data not inserted at end of dataset

Please reorder your data to place 'unknowns' Last!

Results of the Classification of Unknown Observations.

7	0	0	0	0	0	0	13
# Correct	51	31	37	34	22	19	13
Prop Correct	100%	88.57%	100%	97.14%	95.65%	95%	100%

Total Observations 214

Correctly Classified 207

Incorrectly Classified 7

Misclassification Summary

Obs No.	Actual	Predicted
42	2	3
43	2	3
66	2	3
67	2	3
143	5	3
195	4	3
211	6	3

Apparent Error Rate 0.0327

Cross Validation Results

as Adjusted Bootstrap (Groupwise) Cross Validation Res

Average Correct Training Set: 186.5000

Average Incorrect Training Set: 27.5000

Average Correct Test Set: 176.3000

Average Incorrect Test Set: 37.7000

Error Rate Bias: -0.0477

Bias Adjusted Error Rate: 0.0804

Unknown Observation Results

215	3
216	3
217	3

References

Ammann, L. P. (1989). "Robust Principal Components," Communications in Statistics Simulation and Computation, 18, 857–874.

Croux, C., Filzmoser, P., and Oliveira, M.R. (2007). "Algorithms for Projection-Pursuit Robust Principal Component Analysis," Chemometrics and Intelligent Laboratory Systems.

Davison, A. and Hall, P. (1992). "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.

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

He, X., and Fung, W.K. (2000). "High Break Down Estimation for Multiple Populations with Applications to Discriminant Analysis," Journal of Multivariate Analysis, 72, 151-162.

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

Johnson, R.A, and Wichern, D.W. (2002). Applied Multivariate Statistical Analysis, Prentice Hall, Upper Saddle River, New Jersey.

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

Scout. 2002. A Data Analysis Program, Technology Support Project, USEPA, NERL-LV, Las Vegas, Nevada.

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

Snapinn, S. and Knoke, J. (1989). "Estimation of Error Rates in Discriminant Analysis with Selection of Variables," Biometrics, Vol. 45, No. 1, March 1989, pp. 289-299.

Todorov, V. (2007). Robust Selection of Variables in Linear Discriminant Analysis, Stat. Meth. & Appl., 15:395-407.

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

Xie, Y., Wang, J., Liang, Y., Sun, L., Song, X. and Yu, R. (1993). "Robust Principal Component Analysis by Projection Pursuit," Journal of Chemometrics, Vol. 7, pp. 527-541.

Chapter 11

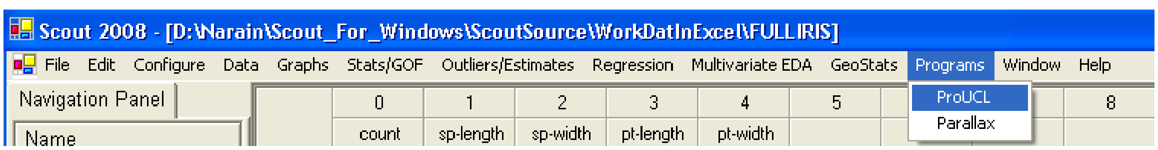
Programs

Access to two additional standalone statistical packages is provided through Scout. Those additional packages are ProUCL 4.00.04 and ParallAX.

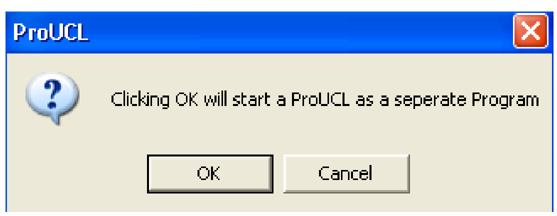
11.1 ProUCL

ProUCL 4.00.04 is a statistical software package developed to address environmental applications.

More information on ProUCL 4.00.04 and the ProUCL Technical and the User Guide can be downloaded from the following web site: http://www.epa.gov/esd/tsc/software.htm.



Clicking on the "**ProUCL**" option in the "**Programs**" drop-down menu will bring up a prompt.

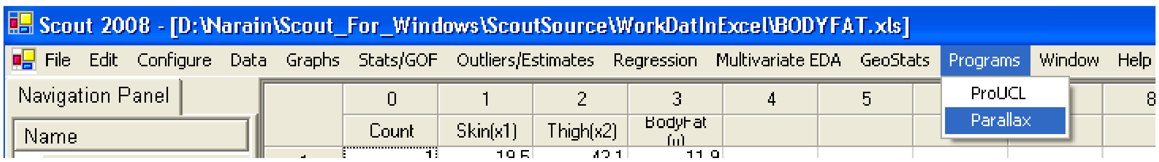


When the "**OK**" button is clicked on, ProUCL 4.00.04 is opened in a new window.

11.2 ParallAX

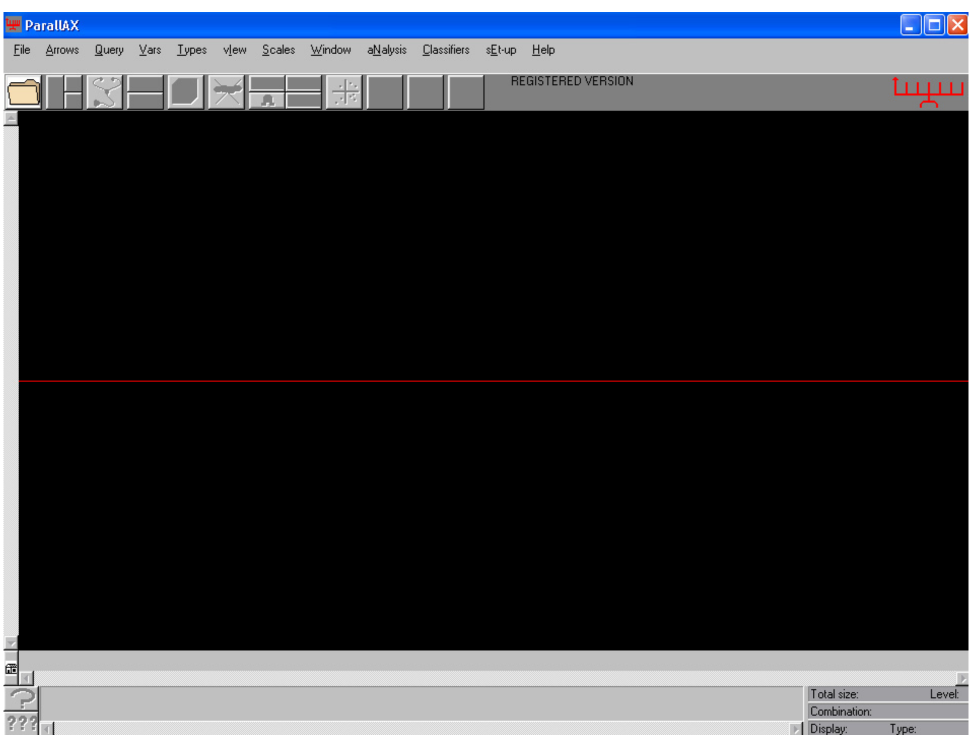
ParallAX software offers graphical tools to analyze multivariate data using a parallel coordinates system. This is a standalone program developed in 1997 by MDG Corporation, Israel.

ParallAX is started in Scout by default whenever the user starts the Scout program. A message in green text appears in the log panel with the successful starting of ParallAX. The screen of the ParallAX (see below) will be running in the background. The user can access ParallAX by minimizing Scout. If Scout failed to start ParallAX, then a message in red text appears in the log panel stating the unsuccessful starting of ParallAX. The user can then start ParallAX by either restarting Scout or by going to the directory where the file, "**Scout.exe,**" is installed on the computer and then by clicking on the "**ParallAX.exe**" file twice.



Clicking on the "**ParallAX**" option in the "**Programs**" drop-down menu will bring up a prompt.

When the "**OK**" button is clicked on, ParallAX is opened in a new window.



Note to the User

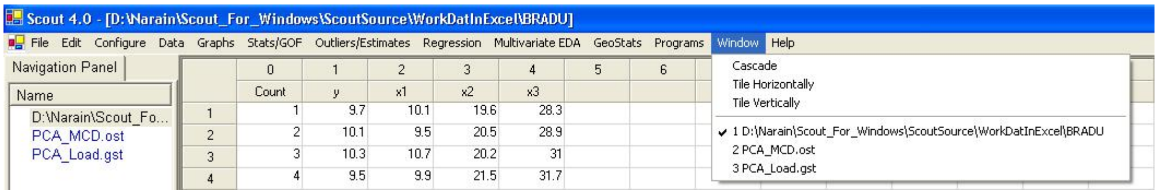
When the user wants to work with the software, ParallAX, an Excel file named "**ParallAX-Fix.xls**," provided along with the Scout package, should be opened first. Then, the ParallAX software can be opened using the drop-down menu. This happens because the standalone program ParallAX looks for its initializing files in the folder from which the data file (*.xls or *.dat) was last accessed.

If the ParallAX software is opened immediately after opening the Scout program, then the process explained above does not need to be done.

The ParallAX User's Manual along with classification examples are provided in the appendices that follow.

Chapter 12

Windows



Click on the Window menu to reveal the drop-down options as shown above.

The following Window drop-down menu options are available:

- Cascade option: arranges windows in a cascade format. This is similar to a typical Windows program option.

- Tile option: resizes each window vertically or horizontally and then displays all of the open windows. This is similar to a typical Windows program option.

The drop-down options list also includes a list of all of the open windows with a check mark in front of the active window. Click on any of the windows listed to make that window active. This is especially useful if you have more than 20 windows open, as the navigation panel only holds the first 20 windows.

Appendix A

ParallAX

User's Manual

MDG

גרמ

Table of Contents

Table of Contents (Cont.)

1.0 Introduction

ParallAX is a novel, some say revolutionary, tool for effectively analyzing multivariate data sets, i.e., software, discovering patterns, properties, and relations. There are two main parts for the ParallAX: the Visual Analysis portion (for doing what sometimes is called Visual Data Mining or Exploratory Data Analysis), and the ***Automatic Classifiers*** that find rules to distinguish elements from a given category or set of categories. The software is based on the *Parallel Coordinates* (abbreviated ||-coords) *methodology,* which transforms the search for relations in a data set to a *pattern recognition problem.* Intuitive interactive commands enable the user to work with data sets having many (i.e., hundreds or more) variables that are displayed *without* the loss of information. Of course, to really understand and appreciate this statement, one needs familiarity with the ||-coords methodology. However, such familiarity is not necessary in order to become an expert user of ParallAX and have lots of fun in the process. Everything needed is described below using as an example a *real data set*.

The main window of ParallAX, shown in Figure 1, has the familiar structure of GUI's in popular Windows applications. Starting from the top, it is composed of the: *Operational*, *Graph*, *Queries* and *Summary* areas.

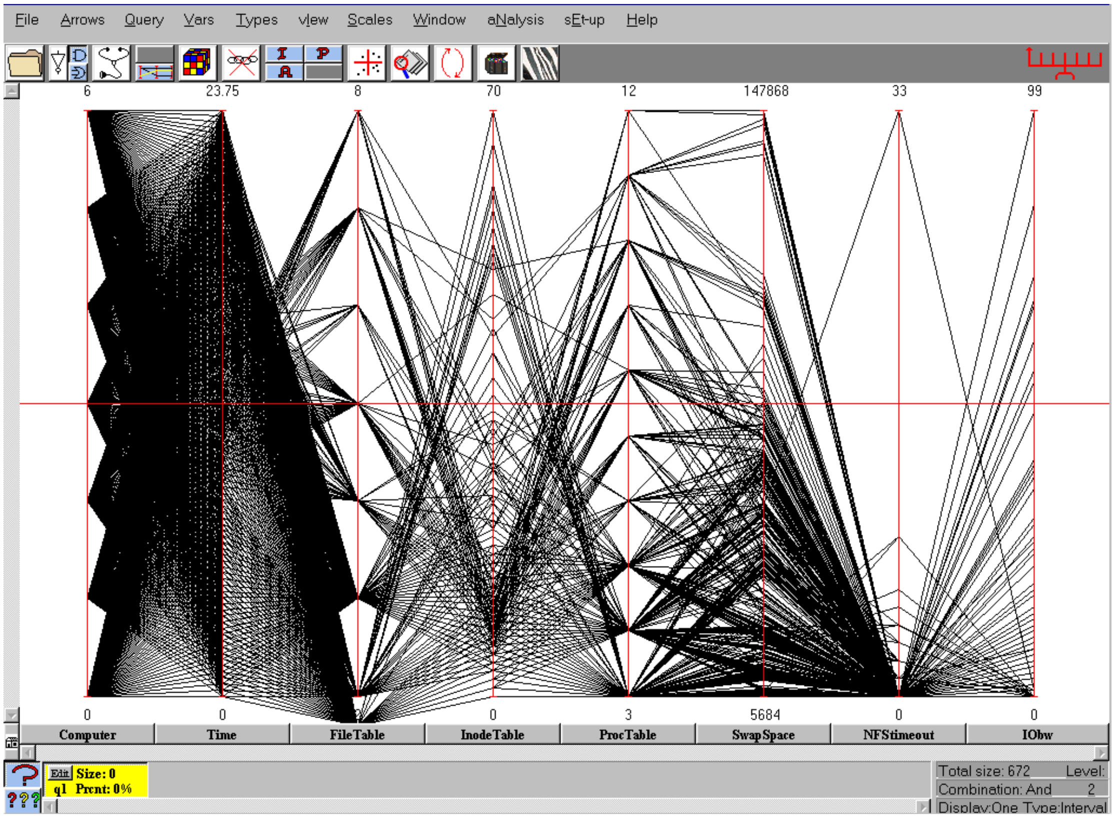


Figure 1. The ParallAX main window or Graph area.

- The "*Operational*" area consists of a main menu with the related pull-down menus, and a toolbar including the most frequently used operations for one touch access. The toolbar is self-explanatory and the names of the buttons are displayed when the mouse icon is pointed at them.
- The data set input is a table; the precise format is given below, where each column consists of values of a single variable. In ||-coords each variable has its own vertical axis. Typically, the scale ranges from the minimum to the maximum value occurring in the data set for that variable (see, for example, the 2nd axis labeled "Time" in Figure 1). A data record is on a single row of the table with the values for each variable separated by a blank. It is represented in ||-coords by a *polygonal line* whose vertices are at the position on each axis corresponding to its value for that variable. For example, the data item (3, -2, 0, 1.5, -4) is represented by the polygonal line having a vertex at a value of 3 on the first axis, a value of -2 on the second axis, 0 on the 3rd, 1.5 on the 4th and –4 on the 5th (last) axis. The "*Graph*" area of the

ParallAX's main window includes the axes, with their minima and maxima, the variable's label button on each axis, and the polygonal lines representing the data. The user may choose, using the ***sEt-up*** pull-down menu (second from the right), either a white or a black (which is the default) background for this area. A particular axis may be selected by pressing its button. A large number of variables may generate a very dense display. In such a case, the user may choose either to see the entire graph or to scroll through enlarged portions of the graph (these options are found using the ***sEt-up*** menu). **Note: Very important** - in the last line of the ***sEt-up*** menu make sure that the "***sort points at graph loading***" on the last option is chosen. This is especially important for improving the performance with large data sets. In real data sets some of the variable values may be missing. In ***ParallAX****,* a point below the actual minimum value on the variable's axis indicates missing values for some data items. In the example data set shown in Figure 1, the variable, "*FileTable*," has several missing values, which are displayed by the lowest point on the third from the left axis.

- Below the Graph is the "*Query*" area and contains a rectangular button for each query. The button's color is the same as the color of the polygonal lines selected by the query (see Figure 4 for an example). The rectangle contains the query label ("q" and the number in the sequence of invoked queries), size, and percent (% of the total data set captured by the query). As the analysis progresses many query boxes may accumulate. They may be moved with the horizontal slider under the query rectangles. Clicking on the small "Edit" button, in the query rectangle, produces a list of other color choices.
- In the "*Summary*" area*,* in the bottom right, general information is displayed. It includes the total number of polygonal lines *currently* appearing, the level of isolation (how many queries have been sequentially isolated to produce this state), the active query type, and the active query logical (Boolean operator) combination. These terms are defined below.

Scatter plot windows (see Figure 2 for example) are opened by selecting a pair of axes buttons (they do not have to be adjacent) and then clicking on the iconized button fourth from the right. The representative points of the polygonal lines selected in the main window are also highlighted by the same color. Several scatter plot windows may be opened simultaneously.

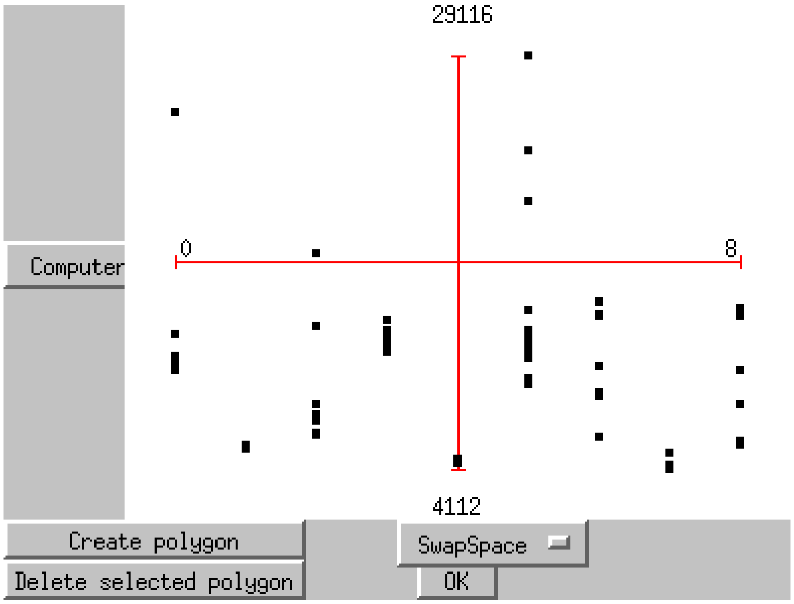


Figure 2. ParallAX scatter plot of the "Computer" number versus the "SwapSpace" variable of the example data set.

2.0 Visual Data Exploration

2.1 Getting Started

This is a good time to install ***ParallAX*** with all four of its directories: *Bmp*, *Dat*, *Ini* and *ParallAX,* into a separate directory. It may be helpful to prepare a data set for practice as we go through the paces. Call your data set any name you like and use the extension .dat, e.g., *testdata.dat.* The data set format is:

# Comment – Write something about the data set to help your recall later on
nvars = # Here write the number of variables
ids = # Here write the labels (as short as possible) for the variables separated by blanks
undefined_data = M # You can define any symbol here and use it consistently below
data =

Data table is placed here. Each data item is in a row with blank (not tab) separated values. Missing data values are marked with M (or any other symbol to the right of the relation, "undefined_data =")

For example,
# This is a small data set with 5 variables, 2 data items, and 1 missing value marked by M
nvars = 5
ids = A B C D E
undefined_data = M
data =

1	4.4	M	17.5	.333
3	3.1	9	9.11	8.2

Input the data set into the "*Dat*" directory of *ParallAX.* From there double-click on the *ParallAX* icon and the *Main Window* should appear on the screen. Click "open" in the "*File*" menu and the list of the data sets in the *Dat* directory appears. Select a data set and press OK; a bunch of polygonal lines appear. *Do not let the picture intimidate.* Very soon you'll learn to discover quite a bit from it. This is done by means of queries which are commands selecting subsets of the data set. The simplest queries are defined by two arrowheads which may be placed anywhere in the main window (on the axes or between axes, depending on the query type). The colored polygonal lines lying between the arrows are those included in the query. From the ***sEt-up*** menu, the background may be changed to white (black is default), and the distance between the axes may also be changed. The default is "*Viewing the whole graph*." If there are many variables, the distance between the axes may be increased and then the graph may be "*scrolled*" using the slider under the axes labels. The *permutation* of the axes may be changed using the "*Permutation Editor*," whose button is iconized by a Rubik's Cube discussed later.

A query may be combined with other queries using set (Boolean) operators (union, intersection, and complement). Many complex queries can be constructed and displayed, either *one at a time* using the single "?" button (default) or *all at a time* with the "???" button on the lower left corner. From the *Query* menu above the button iconized by a stethoscope some or all of the queries may be deleted. To concentrate on the selected query, *isolate* it using the upper-half of the fourth button from the left. The *previous* state can be recovered with the lower-half button. Besides the queries, there are other features in addition to the *Automatic Classification Algorithms.*

2.2 Queries

2.2.1 The Basics

ParallAX's three *basic queries* are:

- The ***Interval*** denoted by ***I*** – defines an *interval* range on a specific variable axis. The end-points are selected delimiting the variable's values within the interval, and, in turn, the polygonal lines (data items) having these values.

- The ***Angle*** denoted by ***A*** – defines an *angle* range between two variable axes, and, in turn, selects the polygonal lines having segments within this angle range.
- The ***Pinch*** denoted by ***P*** – selects a subset of the polygonal lines *between* a pair of axes.

2.2.1.1 Interval Query

The ***Interval*** is the most frequently used query. It is activated by selecting its icon, *I*, on the tool bar and also selecting the desired variable axis. Placing the cursor on the axis and clicking the left mouse button causes down and up pointing arrowheads to appear. Each arrowhead is then dragged in the desired directions to specify the upper and lower end-points of the required interval. The polygonal lines, which are positioned within the specified interval, are selected. On each arrowhead the variable's value at that position is displayed next to it. This feature may be switched off using the ***sEt-up*** button (Hide Interval Limits). An example is shown on the second axis in Figure 3. To move a particular arrowhead, it is first selected by pointing at it with the cursor and pressing the left mouse button. When one arrowhead is selected, it is enlarged and the other becomes deselected. On occasion, it is useful to select *both* arrowheads. Pointing at the deselected arrowhead and pressing the right mouse button selects it. Once both arrowheads are selected, dragging on any of the arrowheads moves the whole interval while preserving its length. When a specific value is wanted for an interval end-point, the particular arrowhead is pointed at and the left mouse button is double-clicked. A dialogue box appears and the desired value is entered.

Within the query rectangle appear the query number (q#), and the percentage (% of the total) of the selected polygonal lines. The color of the query rectangle is the same as that appearing on the selected polygonal lines.

The "Query" pull-down menu (third position from the left) offers choices for query deletion and new query creation. New queries may also be added with the button iconized by a stethoscope. Having generated one or more queries, one may want to delete some of them. Clicking on the "*New query*" produces a new *current* query and an associated differently colored query rectangle. All the subsequent query commands will act on this and *not* on the previous queries.

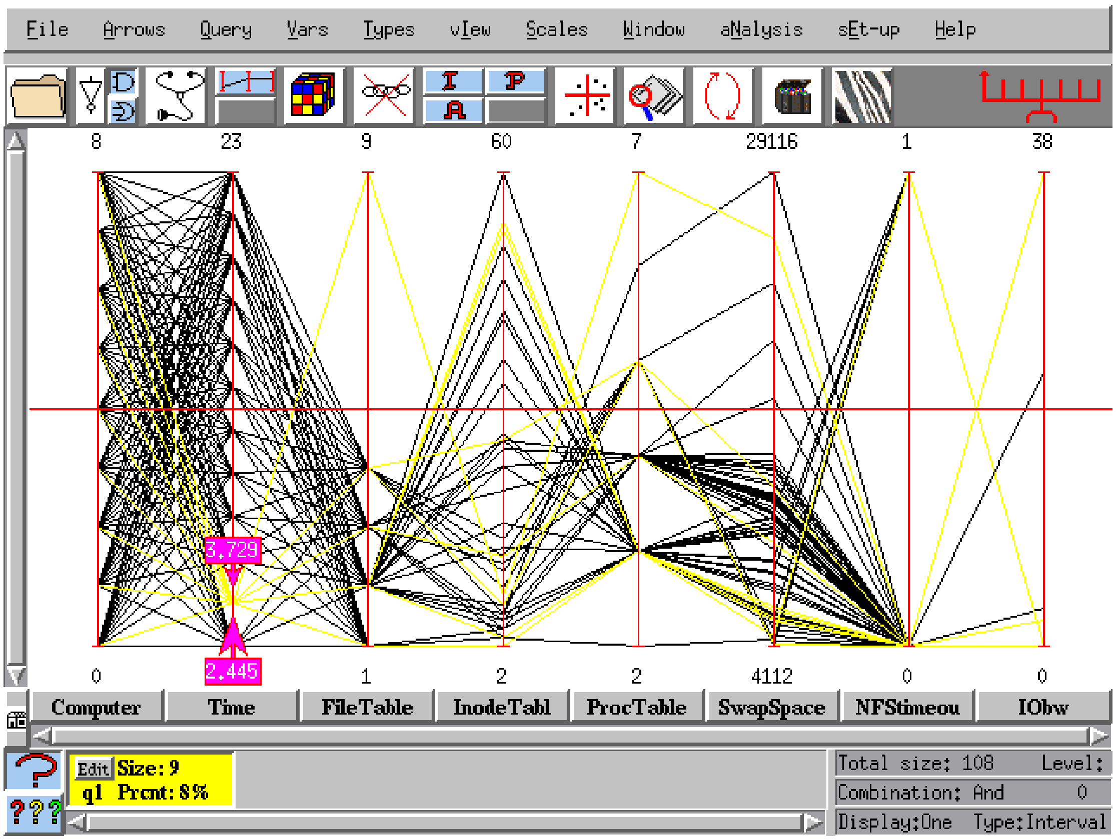


Figure 3. The Interval query applied on the second (Time) axis. Note the arrowheads with the indicated variable values. Here, the bottom arrow (enlarged) is selected.

2.2.1.2 Angle Query

One of the most valuable relations (correlations) among an adjacent pair of variables occurs when the corresponding portion (between the adjacent axes) of the polygonal lines are parallel (or almost parallel) segments; or those lines intersect (if at all) *outside the pair of adjacent parallel axes.* This, of course, is something that the user learns to "extrapolate" with practice.

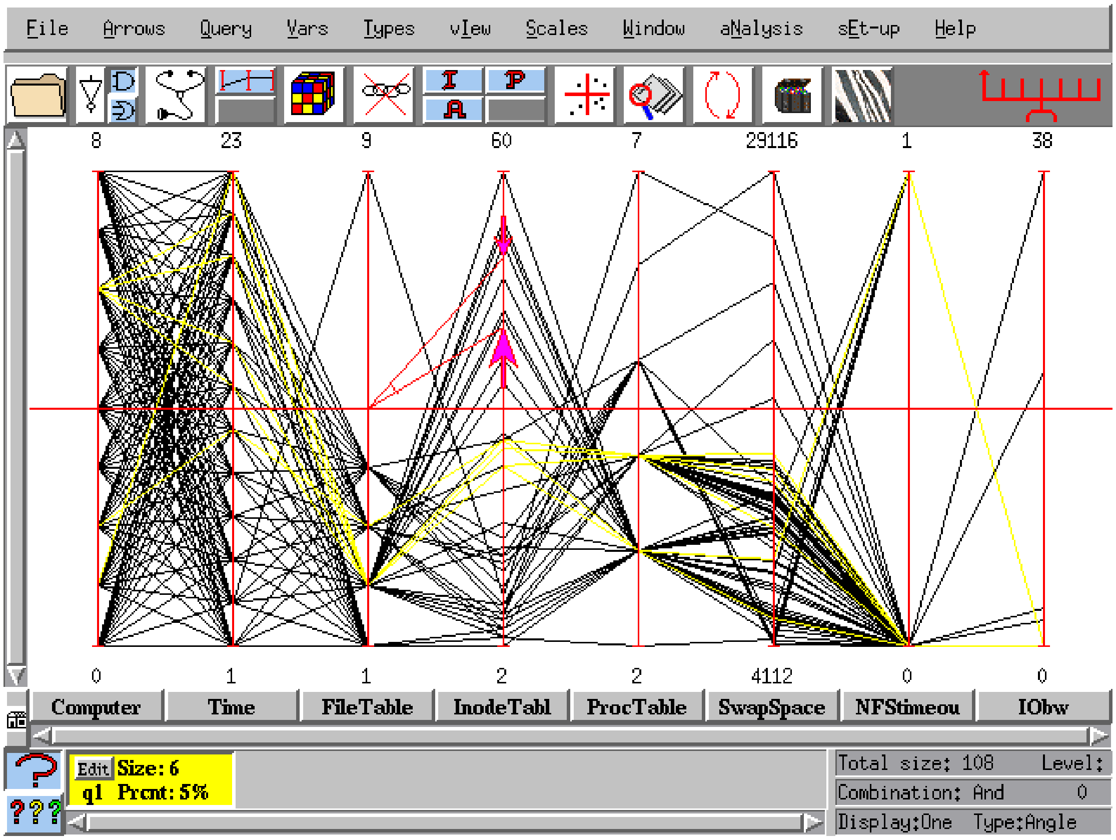


Figure 4. The Angle query shown between the third and fourth axes. Note the selected polygonal lines (colored yellow) whose segments between those axes have the specified angle range.

From a basic result of the parallel coordinates methodology, it is known that this *pattern* corresponds to a ***positive correlation*** between the two variables. Among other reasons, the ***Angle*** query is provided in order to search for such parallel or nearly parallel lines. To activate it, the icon ***A*** is selected on the toolbar. Place the cursor on the centerline of the right axis, say X_i, and click the left mouse button. Two arrowheads connected to the centerline of the left axis, X_{i-1}, appear and an example is shown between the third and the fourth axes in Figure 4. The selected arrowhead is moved to the desired angle. The same can be done, after selecting it, with the second arrowhead. This results in the coloring (i.e., selecting) of the polygonal lines whose segments between these two axes are within the specified *angle* range.

2.2.1.3 Pinch Query

The ***Pinch*** query is complementary to the ***Angle*** type, in the sense that it looks for the intersection points *between a pair of adjacent axes*. Reasoning geometrically, this *pattern* corresponds to ***negative correlation*** between the adjacent variables.

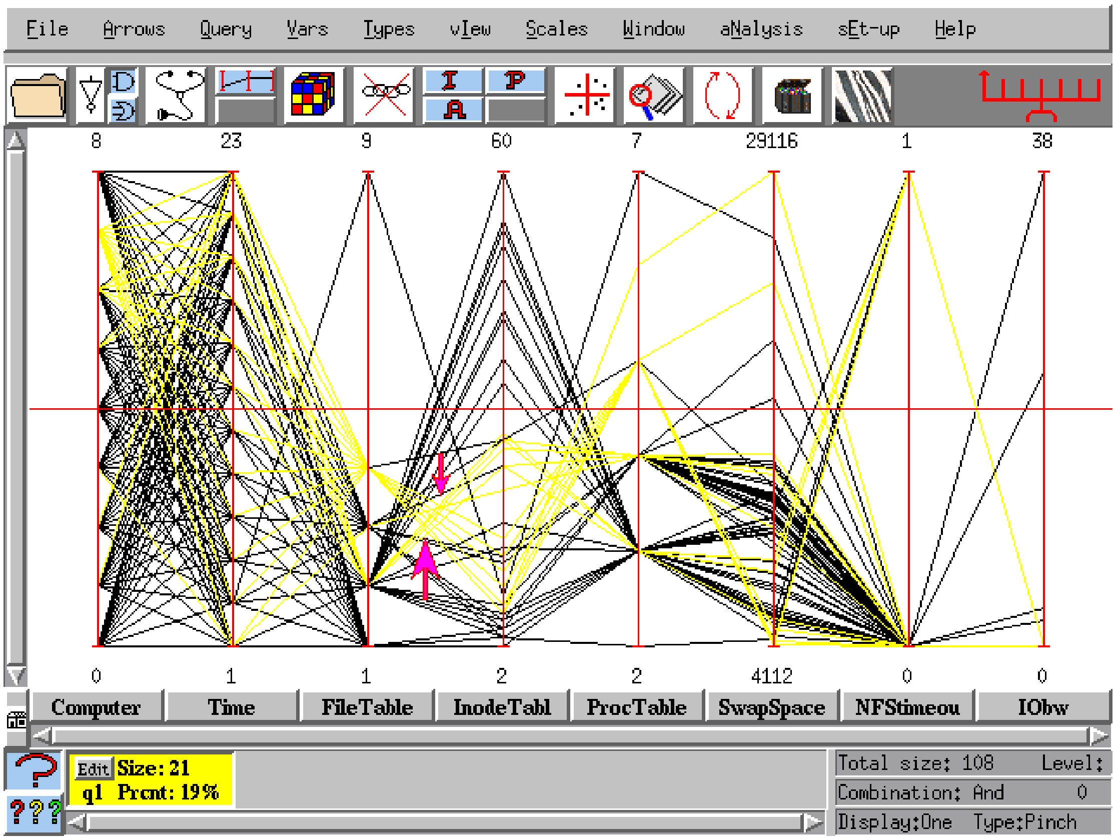


Figure 5. The Pinch query shown here between the third and the fourth axes.

As with the other queries, the ***Pinch*** is defined by two arrowheads that can, in principle, be located anywhere on the graph. Typically, the arrowheads are located between the adjacent axes, X_i and X_{i+1}. All of the polygonal lines whose segments between those axes (or the extension of the segments outside of those axes) that pass between the arrowheads will be included in the query, as in the example shown in Figure 5.

Although those queries may be activated (started) from the main window, they also appear on the corresponding scatter plots and may be manipulated from there by dragging a red square in the scatter plot. The arrowheads are represented in the scatter plots by lines (there is a basic point-to-line duality, or correspondence, between orthogonal and parallel coordinates). It is instructive to view those queries also in the scatter plot window. As an example, in Figures 6, 7, and 8, the scatter plot counterparts of the query types shown in the relative Figures 3, 4, and 5, are displayed (for different axes). Note that the axes labels have a button from which a different axis may be selected, thus changing the scatter plot.

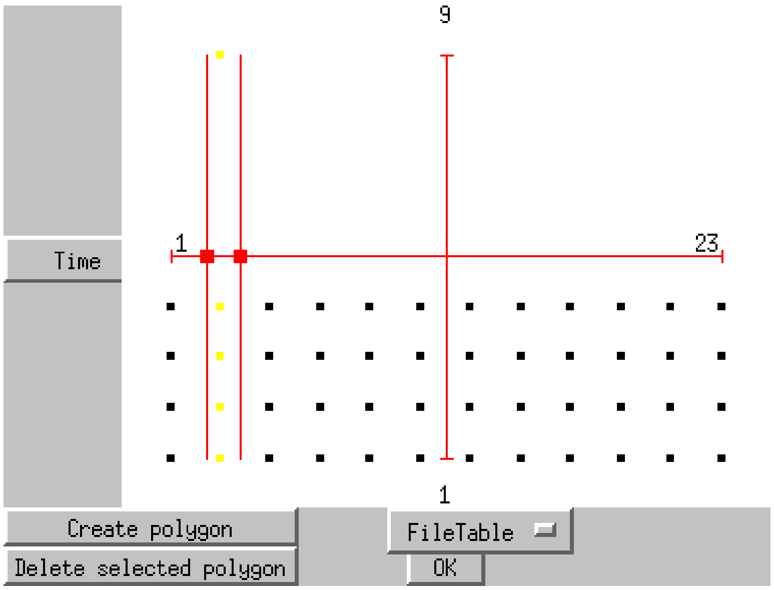


Figure 6. The Interval query on the scatter plot of FileTable vs. Time. Compare with Figure 3.

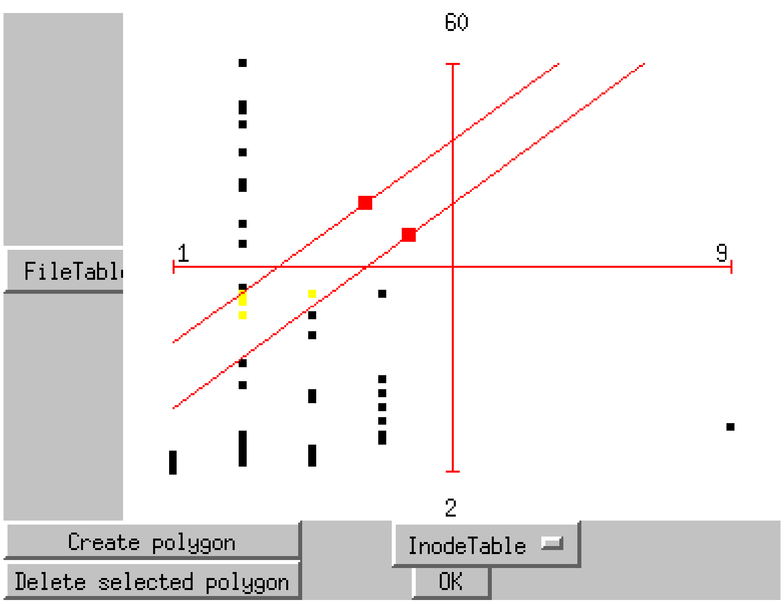


Figure 7. The Angle query on the scatter plot of InodeTable vs. FileTable. Compare with Figure 4.

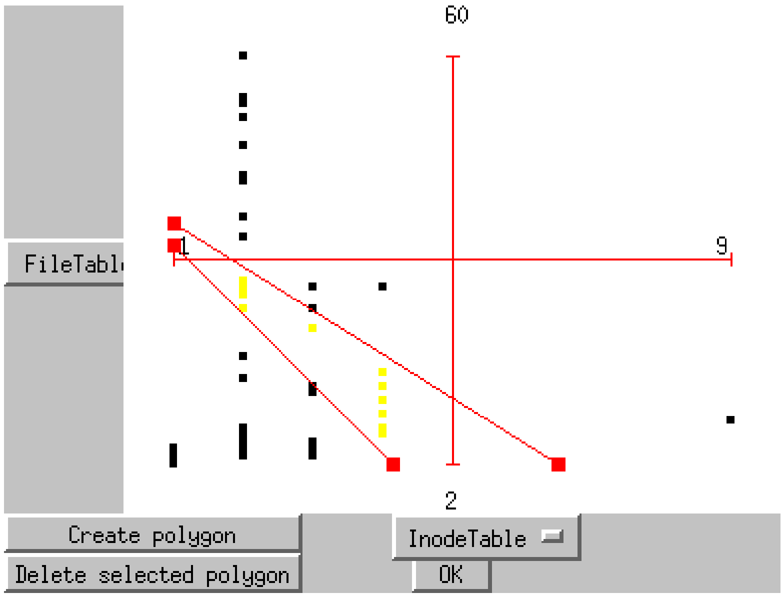


Figure 8. The Pinch query on the scatter plot of InodeTable vs. FileTable. Compare with Figure 5.

2.2.2 More Queries

2.2.2.1 Polygon

Another very useful query is the ***Polygon*** that is activated and operated only on a scatter plot. The polygon is specified by sequentially marking (clicking) with the cursor the vertices in the scatter plot (there are no restrictions and the polygon may have as many vertices as needed and may be convex or not). The construction of the polygon commences after the *"Create Polygon"* button is selected. All the points inside the polygon are included in the query, and the polygon may be moved after its creation, either all of it or a particular vertex (chosen by the user), by selecting and dragging any of the vertices. This query is especially useful when there are points which cannot be picked conveniently by means of the other query types (see the example in Figure 9). The polygon may be deselected with the lower button and deleted with the "*Delete Query*" option of the Query menu.

2.2.2.2 Complex Queries

A single query defines a subset of the data elements. A complex query is the result of combining a set of queries by means of the set (Boolean) operations: union ($\cup$), intersection ($\cap$), and complement. The corresponding operator buttons, appropriately iconized, (as digital electronic Boolean operators), appear in the second position from the left on the toolbar. The complement (or negation) is relative to the data elements displayed when the query atom is defined; i.e., if the set of data elements included in the original query is denoted by *A,* and the

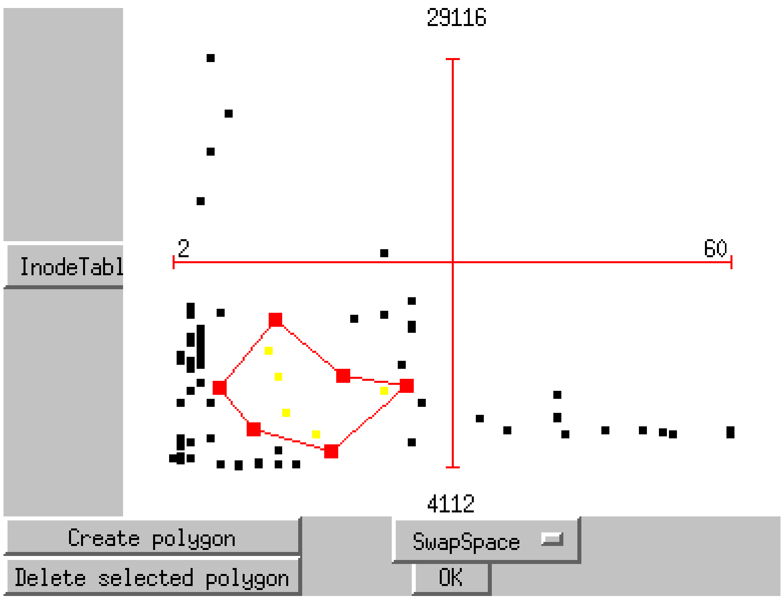


Figure 9. The Polygon query.

set of displayed data elements is denoted by *P*, then the complemented query, $\bar{A},$ will be defined as:

$$A = P \setminus A = \{ a_i \mid a_i \in P , a_i \notin A \} \qquad (11)$$

To define a complex query, the desired set operation must first be selected (the and, $\cap$, operation is the default). To construct the *complement* of a query, the negation operation is selected *before* the query is constructed. For the next query, ***ParallAX*** will apply the existing combination of the selected buttons (union, union + negation, intersection, or intersection + negation). ***So be careful with this; it requires care.*** A very useful option is the construction of multidimensional intervals or a "multidimensional box." Select the appropriate axes buttons and also the interval, *I,* button. Place the cursor at any of the selected axes and click the left mouse button; pairs of arrowheads will appear on *all* of the selected axes. Dragging any one of the arrowheads causes all of the arrowheads pointing in the same direction to move simultaneously.

2.3 Supplementary Operations

ParallAX has additional operations to help the exploratory data and analysis which act on the axes, the display, or portions of the Graph**.**

2.3.1 Inverting Axes

This operation is complementary to the Angle query that searches for groups of polygonal lines that (nearly) intersect *outside* a pair of axes (i.e., clusters having a positive correlation for a particular pair of variables). The intersections may be quite distant and difficult to spot. By contrast intersections *in between* a pair of axes are much easier to notice. *Inverting* one of the adjacent axes (i.e., interchanging the minimum and maximum of the variable) reverses the situation, that is, the distant intersections now appear as intersections between the axes and vice versa. Such clusters of polygonal lines can now by picked with the ***Pinch*** operation. To carry out this operation, the axis to be inverted is selected and the "*Flip axes*" button (iconized third from the right) is clicked and has its minimum and maximum values marked in red (see Figure 10).

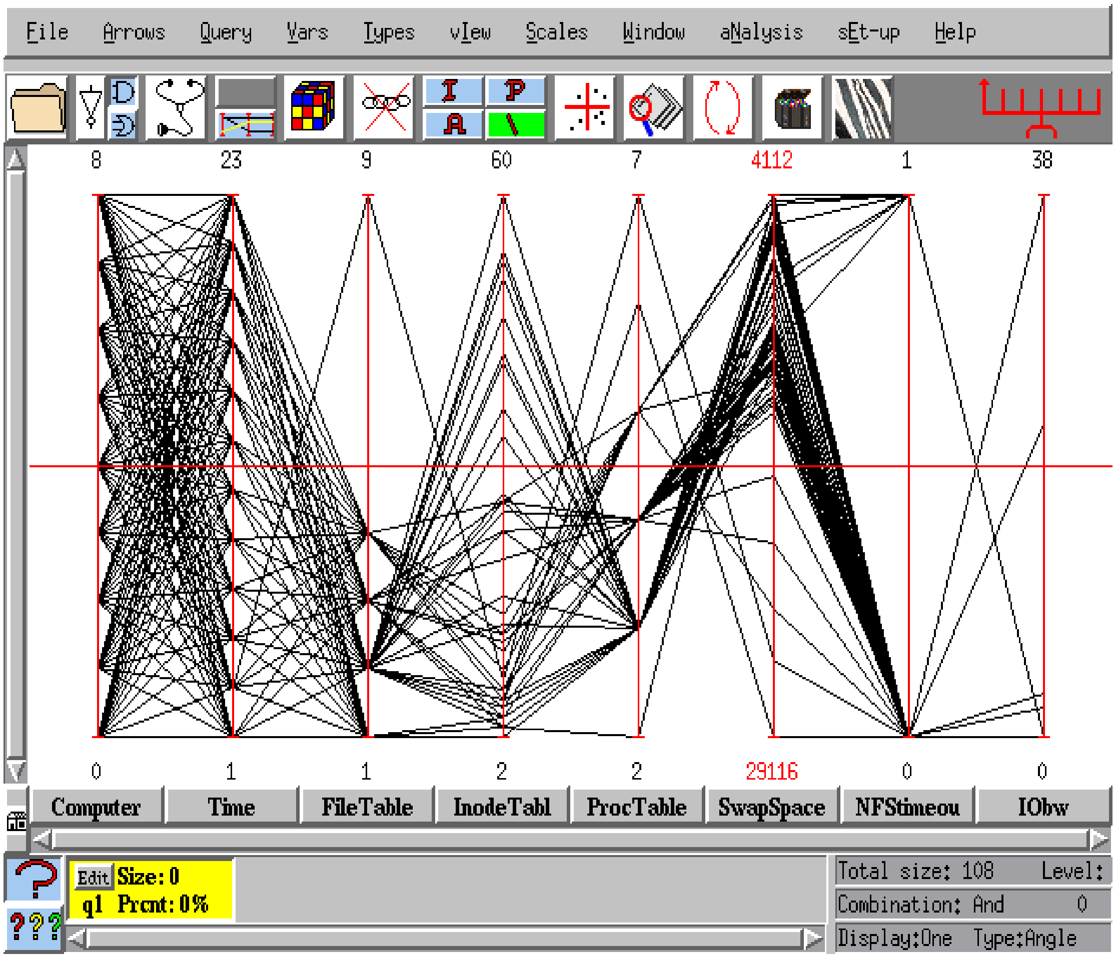


Figure 10. The ||-coords graph with one inverted axis (SwapSpace).

2.3.2 Permutations

Even though mathematical relations have clear patterns (see Bibliography) which are easily recognized by their regularity (see any elementary paper on ||-coords), the graph of most data sets do not look terribly "regular." However, patterns between adjacent axes are the easiest to discover. In order to discover all possible pair-wise patterns, it is not enough to look at the ||-coords graph in the form that it first appeared. Rather all of the possible adjacencies need to be inspected. It is possible to change the order of variables in a very efficient way. ***ParallAX*** allows the user to chose about N/2 (actually $\lceil N/2 \rceil$), where N is the number of variables, cleverly constructed permutations which *contain all possible adjacencies,* and these are automatically provided. Click the Rubik's cube button, the fourth from the left icon, and those permutations are listed on the upper right window. It is a good idea to view the data with each one listed, and then construct, by means of the permutations editor there, a *customized* permutation containing the axes adjacencies of choice. Of course, a particular axis can be included more than once and in any position. If it is desired to view as adjacent a particular pair of variables, then enter that pair in the lower left editor window and a permutation is displayed where the required adjacency appears and the remaining variables are randomly ordered.

2.3.3 Isolate/Previous/Scale

After defining a query (or a set of queries), the user may wish to concentrate on the selected data items (i.e., polygonal lines). As already mentioned, in order to do that, clicking the top half of the fourth button from the left may isolate the current query. This yields a new graph containing only the data selected by the previous query. The graph is displayed with the values of the minima and maxima of the variables in the previous graph (before isolation). In order to update the minima and maxima of the new graph, which enlarges the space used by the graph, the user may choose ***Scales*** from the menu. Clicking on the button below *Isolate* returns to the *Previous* state.

2.3.4 Relative Complement

A query defines a subset of the data elements. When two or more queries have been defined, two or more subsets of elements have been specified. The user may wish to use set operations, such as the union ($\cup$), intersection ($\cap$), or relative complement (\), to operate on the queries (sets). The use of the union and intersection operations has already been described (see "*Complex Queries*"). The "*Relative Complement*," iconized by \, is a specialized and advanced query. When choosing this function, ***ParallAX*** displays the list of all of the possible combinations ($2\binom{n}{2}$ possible combinations). The user chooses one of them, and a new query is defined which is the set difference of the 2 queries chosen; i.e., if the first query is denoted by Q_A and the second query is denoted by Q_B, the resulting query, denoted by Q_R, is:

$$Q_R = Q_A \setminus Q_B = \{ a_i \mid a_i \in Q_A , a_i \notin Q_B \} \tag{12}$$

The new query is not directly composed of basic queries or polygons and it depends on the two other queries.

2.3.5 Zooming

When we want to view a portion of the graph in greater detail, a rectangular portion of the graph can be isolated and enlarged by means of the "*Zoom*" button, iconized by a magnifying glass. An example is shown in Figure 11.

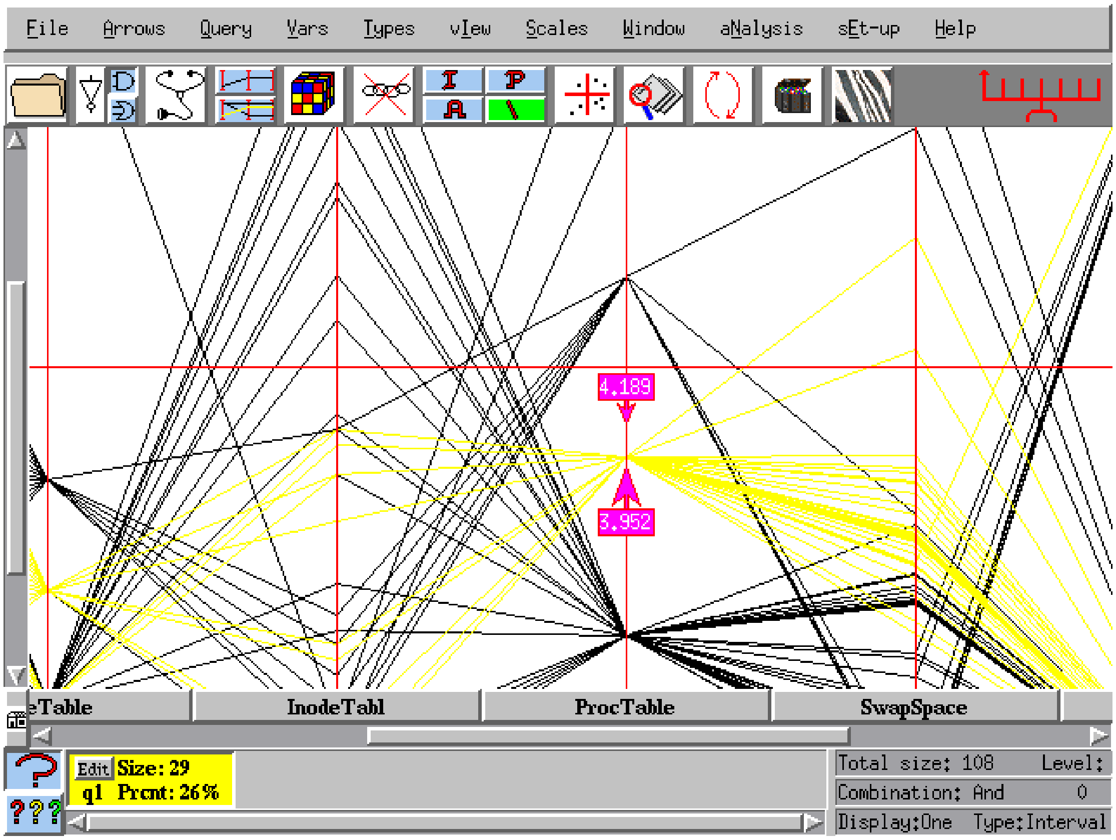


Figure 11. The Zoom function.

2.3.6 More Supplementary Operations

- *Save as (from the* "*File*" *menu).* It is possible to save, in the *Dat* directory, a subset of the data set by a separate name. This can be done by isolating the data set and using the "*Save as*" option from the *File* button. A dialogue box appears. Enter a file name with the .dat extension and the file is saved.
- *Select off screen arrows (from the* "*Arrows*" *menu).* Pointing at it and clicking the left mouse button selects an arrowhead. At times, arrowheads get off the screen. In order to delete them, they need to be selected first by means of this function.

- *Delete selected arrows (from the* "*Arrows*" *menu).* One may select, or delete, as many arrowheads as desired. If both of the arrows of a query are deleted, then the whole query is deleted. If only one arrow is deleted, then the query remains unbounded on that side, and all of the data elements found lower or higher than the remaining arrow are included in the query. This is a good way to delete a query, when many queries are operating on the data, without destroying other queries that may be present.

- *New query (from* "*Query*" *menu)* - A new query rectangle is added and becomes the current query.

- *Clear current query (from* "*Query*" *menu)* - All of the displayed queries are cleared: all arrowheads are deleted and the polygonal lines receive their original color. So, make sure that this is what you want before using.

- *Delete variable (from the* "*Vars*" *menu)* - If the user presses some variable(s) button(s), and then chooses this function, the selected variable(s) are deleted from the display. This is equivalent to choosing the current permutation *without the chosen variables.* This can be very useful when there are many variables.

- *Find variable (from the* "*Vars*" *menu)* - In a data set with a large number of variables, it is hard to find variables by their names. ***ParallAX*** comes to the rescue. Choose this from the "*Vars*" menu and a list of variables in alphabetical order appears. Choose the desired variable, and on the *Graph* the corresponding axis button is shown selected (i.e., depressed).

- *Show one query* / *Show many queries* - The user may choose to see a single query or many queries simultaneously by selecting "?" or "???" respectively in the lower left hand corner. When "?" is selected, and there are several queries, the active query is chosen by selecting the appropriate query rectangle. Viewing many queries in large data sets still may cause some problems with the query colors; hopefully it will be fixed soon, so some care should be exercised.

The *Vars* menu contains a number of useful functions.

1. When there are a large number of variables, it is tedious searching for individual variables. Clicking on *"Find Variable"* produces the list of variables alphabetically. Selecting the desired variable in the list selects the axes button of this variable. By the way, this renders that variable axis ready to operate on with the *Interval Query*.
2. At times it is useful to know the *order* in which the data appears in the data table. Clicking on the "*Add Index Variable*" produces a dialog box where the name of the new variable can be specified. The variable then appears at the right end of the graph and has as the value of each data item its position (rank) on the data table at input.
3. On occasion the user wants to designate a subset of the data set into a separate category. In such a case, the "*Add Categorical Variable*" 3rd entry on the menu is invoked and given whatever name is desired. The new variable then appears on the right hand end of the graph with the designated subset assigned the category value 1 while it's complement takes the value 0. Further subdivisions of the data set can be assigned other category values using the "*Set Category*" option on the menu.
4. One or more variables can be omitted from the graph by selecting the variable buttons and then invoking the "*Delete variable(s)*" options.

2.4 Preprocessing

Some operations may be used for *preprocessing* to provide the user with insights on the structure of a data set easily and early in the analysis process. Then, the data items or variables that seem superfluous, and whose presence may obscure the information, can be eliminated. In fact, such elimination plays an important part in focusing on the desired information.

2.4.1 Zebra

Zebra (*banding)* is a multidimensional contouring operation. It is designed to portray easily variations in *all* of the variables due to variations in one variable. To operate this function, select the axis of the desired variable and the "*Zebra*" button iconized in the last (most right) position of the toolbar. In the dialogue box that appears, enter the number of intervals. The selected axis is then divided into equal length intervals. It is a good idea to start with 2, view the result and then increase the number. The polygonal lines ranging in each interval are colored by a different color. The result of this operation is a contoured view of the data, highlighting different aspects, especially dependencies, intersection points, data clusters and extreme points and others. It can also point out areas with high density and reveal periodic events. An example of Zebra results is shown in Figure 12.

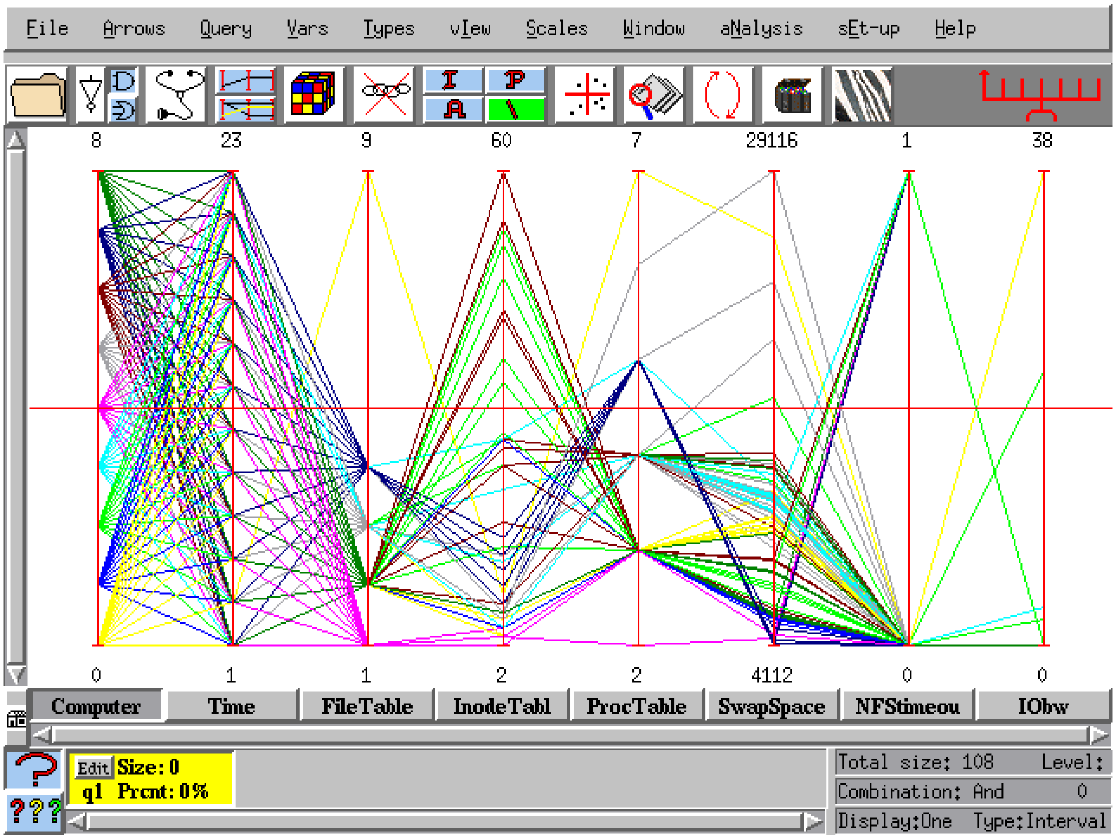


Figure 12. An Example of the "Zebra" function applied with 7 subdivisions on the Computer Axis (1st from the left).

2.4.2 Outliers

This is an automated algorithm suited to large data sets having a number of outliers. In general, application of this algorithm is recommended only for expert users (which, of course, you will soon be). It is a good idea to study the outliers of a data set and try to determine the reason that they are outliers. On the other hand, outliers determine the display scale and removing them enlarges the scale for the remaining data. This allows for the observation of patterns that may be hidden by the high density of data. It is really best to manually remove the outliers after examining each one of them. A convenient place to start eliminating data is close to the limits of the axes. Points near the limits and far from the large mass of data are good candidates for elimination.

The *Outliers* function starts an iterative algorithm that performs this task. The user may supply some parameters to the algorithm, or leave their default values. The parameters are:

- The maximum (relative) number of outliers (the default is 5%). If the algorithm reaches this value, it will stop searching fore more outliers.
- A factor, whose default value is 6, which influences the distances between elements on an axis; considered by the algorithm as a starting point for the outliers search.
- A divider (whose default value is 10) indicating the length of a segment on the axis. If we denote the divider by *d* and the axis length by *l*, the algorithm will ignore outliers whose distance to the closest element (non-outlier) is less than *l* / *d*.

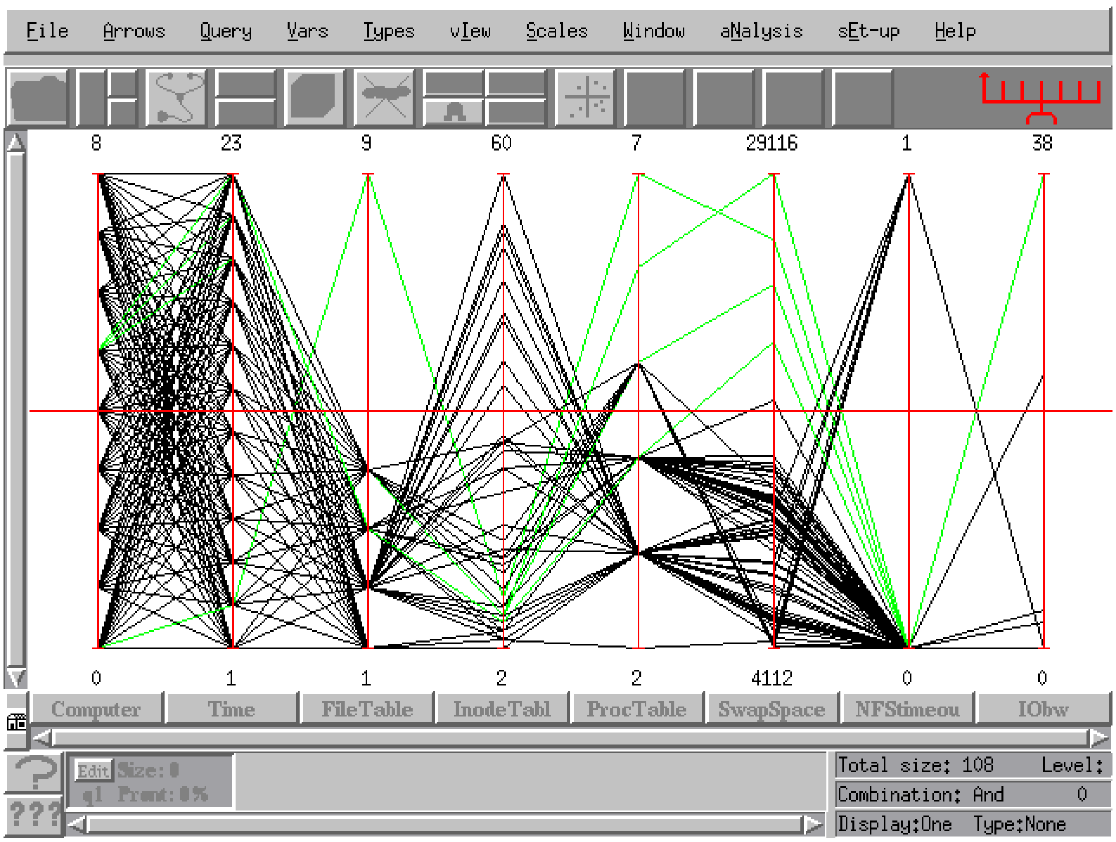


Figure 13. The result of the Outliers operation (before user approval).

The algorithm starts looking for outliers from the leftmost variable in the displayed permutation to the right. After finding all of the outliers on an axis, it passes to next axis, until the last one in the permutation is reached. Then, it starts again from the first axis, and so on. The algorithm stops when the maximum relative number of outliers is reached, or, if that does not happen, when it does not find any more outliers after passing on all of the variables in the permutation.

After that, it displays all of the outliers found highlighted (colored in green) and waits for the user to approve this. The user may not approve of the choice, retaining the current graph. Otherwise, the algorithm issues an Isolate operation and displays the graph without the outliers. Even in this stage, there is a possibility to return to the previous graph, by performing the previous operation. The example shown in Figure 13 is the result of the Outliers function applied to the demo data set, with the default parameters, before the actual removal of the outliers (i.e., before the user approved it).

3.0 Automated Classification

Even though the Visual Exploration is fun and effective, it requires time and skill. Hence, the most frequent and insistent requests have been for automation of at least some of the discovery process. Some of the functions we have already presented have, of course, elements of automation. It was recently discovered that it is possible to do ***automatic classification*** (patent pending) effectively based on ||- coords. Given a data set, ***P***, and a subset, ***S***, a rule is sought that distinguishes elements of ***S*** from the others. Obviously, we would like this to be as accurate and efficient as possible. This is the basic classification problem and it can be directly generalized to the case where there are a number of subsets (also called *categories*) that need to be distinguished from each other. There are important trade-offs between the rule's complexity and precision. In our case, we are able to state the rule precisely (unlike the "learning" of "black boxes") as well as visually. This as we will see, turns out to be very helpful. In addition, our algorithms find the minimal subset of the variables needed to state the rule and order these variables according to their information content. The basic idea of our algorithms is geometrical and it entails the construction of a (hyper) surface that contains as many of the points of ***S*** and as few of the points of ***P-S*** (the complement of ***S***). This brings up the important matter of measuring the precision of the rules obtained by our classifiers. We discuss this later on. There are three classifiers and they are found by clicking the "*Classifier*" menu's first line.

3.1 Wrapping

The simplest approach to geometrical classification is to *wrap,* in some efficient way, the points of ***S*** and then state, in as simple a way as possible the rule (which is actually the description of the wrap – an approximation of a convex surface). The algorithm, even at the expense of some precision, further simplifies the description of the wrap. The rule is stated in terms of conditions on the variables needed to *fully* state the rule. Also these variables are optimally ordered (in terms of their information content). To apply this and any of the other classifier algorithms, the subset ***S*** needs to be specified and used as the input. In many data sets, there are one or more variables that specify various categories or classes. In that case, using the interval query isolates a specific category. Otherwise ***S*** is defined by means of the queries. When this is done, choose "*Wrapping*" from the Classifiers menu. The "*Select axes*" dialog box appears and provides an important choice; namely, to choose the variables in terms of which we would like to have the rule stated (think of the many applications where this is essential). We can "*Select all*" with the button and then skip the ones we want to skip. *If the subset **S** is specified in terms of **interval** queries only, be sure to deselect those variables at this stage or the rule is likely to be a trivial restatement of the defining conditions.* Click the OK button and the "*Classifier summary*" appears with the expression with the *approximate* conditions for the rule as well as the percentages of the misclassification for the "*Training phase*" (see below). That is, "*False positives*" refer to those data items in ***P-S*** that were misclassified as *belonging* to ***S***, while "*False negatives*" are data items in ***S*** that were misclassified as *belonging* to ***S.*** If those errors are small, then this rule may suffice. Still, look in the Graph where the last query displayed contains all of the elements of ***S*** and the "*False positives.*" The variables needed to state the rule are displayed first with arrowheads in the suggested order of their importance. It is possible to save the rule and to apply it to another data set. To do so, select the "*Save classifier*" option and give the rule a name in the dialog box that appears; click OK and the rule is saved in the Data directory. To apply it again on another set of data ***S'***, which is already displayed in the graph, select the category variable on which the rule is to be applied and also select the "*Apply classifier*" to chose the rule from the list. The result has the format already described.

As an example, we can see in Figure 14 an Interval query on the axis INodeTable. After performing the wrapping algorithm on all of the axes except for the INodeTable, the resulting query and permutation are shown in Figure 15 and the difference in Figure 16.

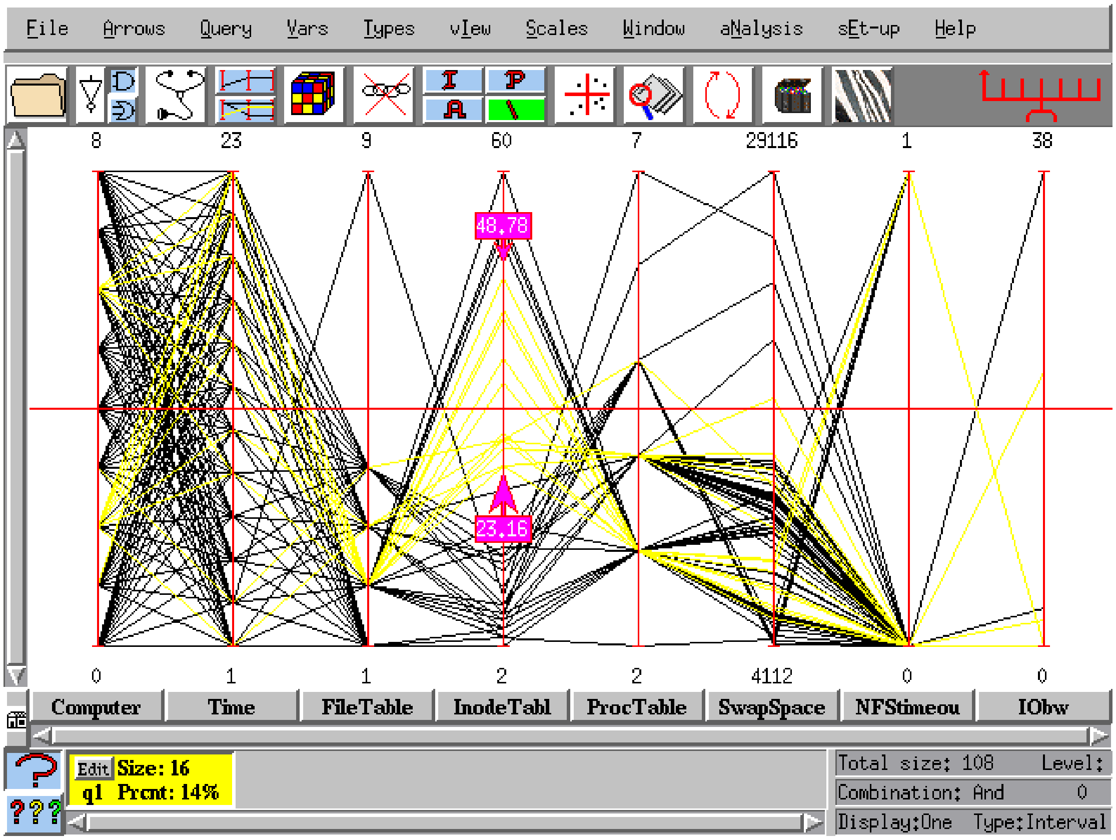


Figure 14. An Interval query defining the input set in the Wrapping operation.

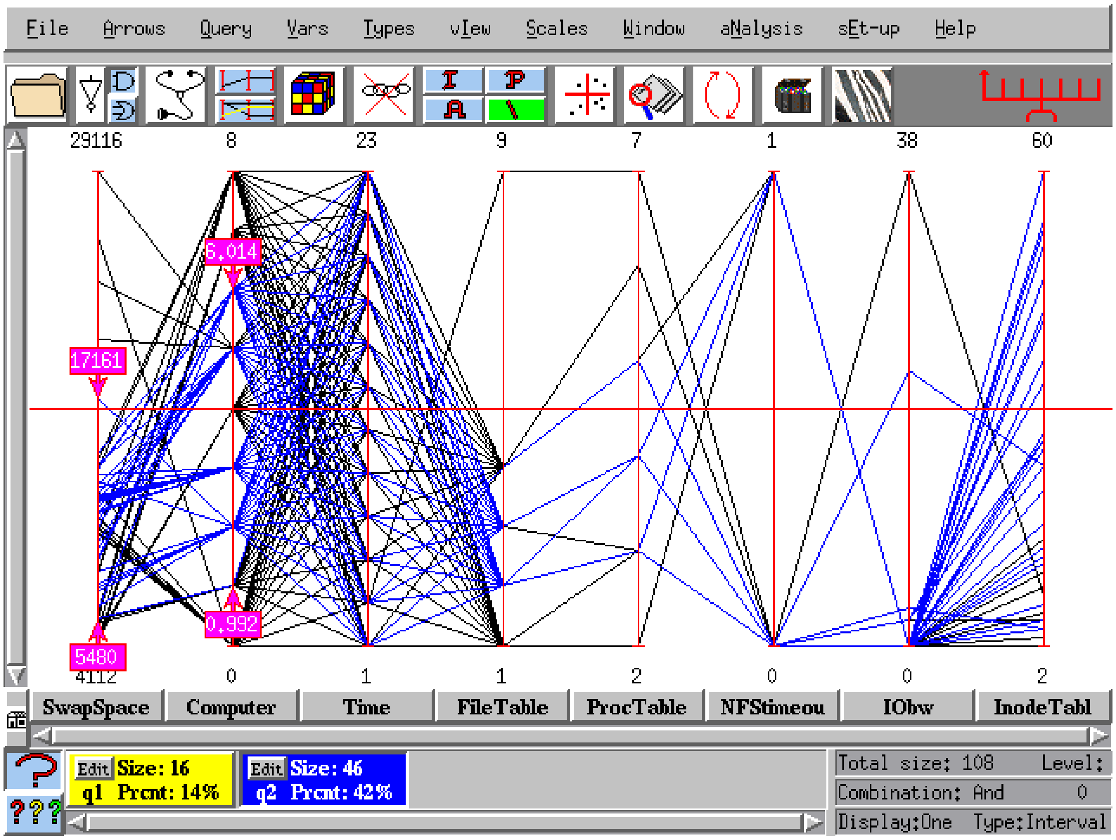


Figure 15. The result of the Wrapping operation.

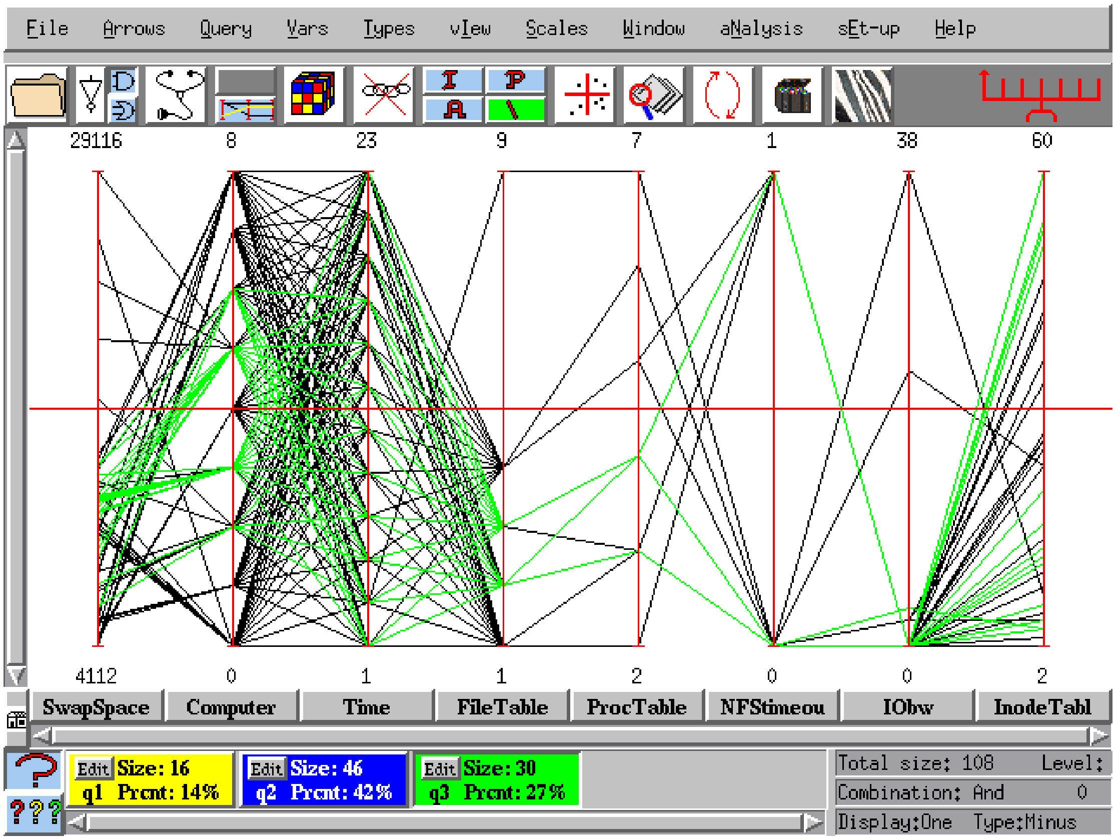
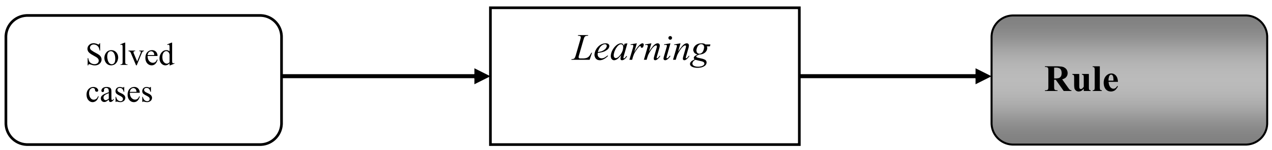


Figure 16. Set of "unwanted" elements by the Wrapping operation (obtained using the relative complement, "\"*).*

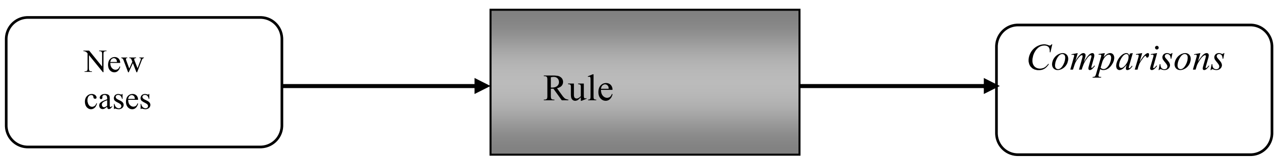
3.2 The Classification Process

ParallAX includes two very advanced classifiers: the "*Nested Cavities" NC* and "*Enclosed Cavities*" *EC*. Compared with 23 other well-accepted classifiers, as applied to some benchmark data sets, *in all cases,* they were the most accurate. Also, they are computationally very efficient. The classifiers exploit the inherent property of this tool, visualization, as well as the computational advantages of the ||-coords methodology. The classification results are displayed graphically on the screen giving the analyst the ability to *understand* the results. The ability to visualize the rules is lacking in many other classifiers.

The classification problem arises in a variety of fields and can be divided into two phases. In the *training phase,* the classifier "*learns*" to discriminate between classes using a data set called the training data, consisting of solved cases having samples associated with correct classification. The output of the classifier in our case is a *rule*, which is based on the solved cases. Then, there is the *testing phase,* where the rule is applied to a new data set and the results it provides are compared to the known correct cases. Figure 17 illustrates the classification process in general.



a) The training phase.



b) The testing phase.

Figure 17. The classification process.

3.2.1 Analyzing the Errors

For the classes designated as "positive" and "negative," the error committed when predicting a positive sample as negative is called a "*false negative*" and the error committed when a negative sample is predicted positive is called a "*false positive*." The error rate of these two types of misclassification is calculated based on the following equations:

$$\textbf{False positive error rate} = \frac{\textbf{number misclassified positive cases}}{\textbf{number of negative cases}}$$

$$\textbf{False negative error rate} = \frac{\textbf{number misclassified negative cases}}{\textbf{number of positive cases}}$$

Keep these formulae in mind when examining the error rates given by the classifier.

3.3 Nested Cavities Classifier – NC

This new classifier is based on an iterative top-down process of creating a (hyper)surface containing as many points of the designated subset, ***S***, and as few points of its complement, ***P-S***. The algorithm involves creating an exterior wrap, then constructing and removing a wrap containing all the unwanted points (and some of the wanted ones), then returning a smaller wrap with the wanted points (and some of the unwanted ones) creating a fine nesting of cavities which provide an increasingly more precise approximation for the desired subset, ***S.*** If this process converges, and it does NOT always converge, then the result (i.e., the approximate description of the (hyper) surface) is the rule, which can be quite complex. Again it is stated as conditions on the variables needed for the classification. The queries that add points have an even number while those that remove points have an odd number (except for the first one which contains the class elements). To apply the ***NC***, select the class on which the rule is to be defined, choose "*Nested Cavities*" from the Classifiers menu, select the variables as for *Wrapping*, limit the number of iterations allowed (100 is default) and then press OK. In the beginning, especially for large sets, it is worth picking a smaller number of iterations, and if convergence looks likely, then remove the iteration restriction. A great deal can be learned from studying the classification rule. Notice the leading list of variables occurring in the successive iterations. Those who tend to occur consistently or most frequently are the most important and there are other clues that come with experience. An example of the spectacular results that may be obtained is shown in Figures 18 and 19. The classifier was applied to a data set with 32 variables and 2 classes shown in Figure 18. It is sought to find a rule to distinguish elements of class 1 from its complement class 2 whose elements are colored black. Notice how interwoven the two classes are as shown in the scatter plot of the first 2 variables shown in Figure 18. The result is displayed in Figure 19. The ***NC*** is the one used most frequently, as it tends to be more successful.

3.4 Enclosed Cavities Classifier – EC

On occasion, when the NC does not give satisfactory results, it is worth applying the next classifier EC. Basically, classification using the EC is based on obtaining an exterior wrap of the wanted data points. Then, removing the unwanted points with cavities that *do not contain any of the wanted points.* The result is something akin to "Swiss cheese." The operation is the same as for **NC** with the **EC** tending to be slower especially for large data sets. It is advised to use the default settings of the 2nd dialog box until enough experience has been obtained to make judicious choices.

3.5 Error Analysis

Once a rule is obtained, it is possible and desirable to assess its precision. Two ways are provided and they are accessed from the "*Check Classifier*" option of the Classifier menu.

3.5.1 Train-and-Test

This is the most frequently used method. The data is randomly split in two. The usual proportions are either 2/3 or 1/2 for training, i.e., deriving the rule, and applying the rule (i.e., testing) on the remainder. The actual portion chosen for training is prescribed in the dialog box. Then the classifier used is chosen (Note: *Extended Cavities* and *Wrapping with Cavities* are synonyms for ***NC*** and ***EC*** respectively). Make sure to use the same list of variables and iterations as used in the derivation of the rule.

3.5.2 Cross Validation

Here all of the data set is partitioned in a number of subsets and split randomly for training and testing. This gives a better error estimate than Train-and-test but also takes much longer.

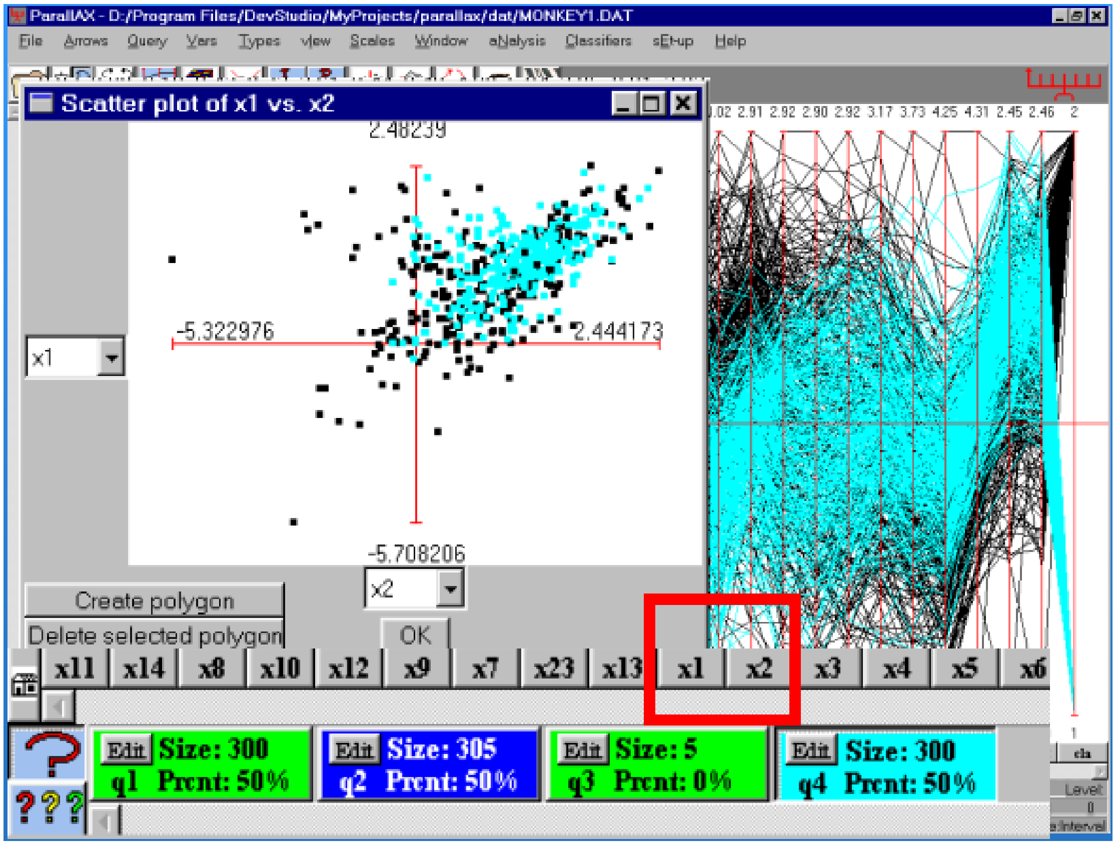


Figure 18. A real data set with 32 variables and 2 classes (categories) – the rule is sought for class 1 shown in color. The complement class 2 is shown in black. In the insert is the scatter plot of the first 2 variables in the permutation on input. An effective classification should lead to a physical separation of the 2 classes.

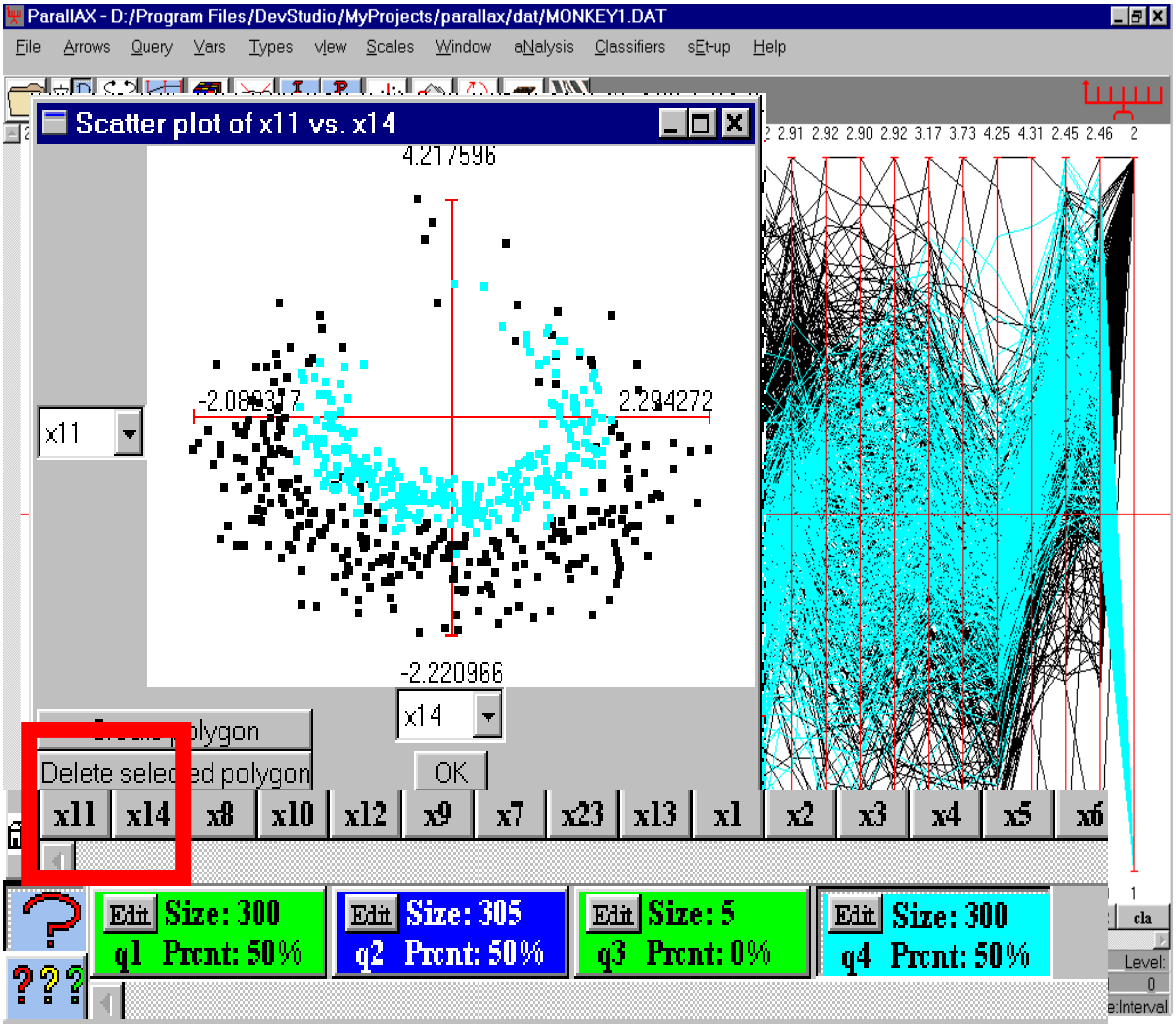


Figure 19. Above are seen some of the results obtained by the NC classifier. It turns out that only 9 of the variables are needed to specify the rule. They are placed up front sorted according to their information content. In the insert is the scatter plot of the first two variables showing a remarkable separation. Viewing the remaining scatter plots of the variables shown in the list provides a "road map" to actually seeing the RULE as represented by a 9-dimensional hypersurface embedded in the 32-dimensional space of the original data set.

===

The reader is requested to send any questions or comments to

A. Inselberg aiisreal@math.tau.ac.il

or mail to:

MDG Ltd

36A Yehuda Halevy Street

Raanana 43556, ISRAEL

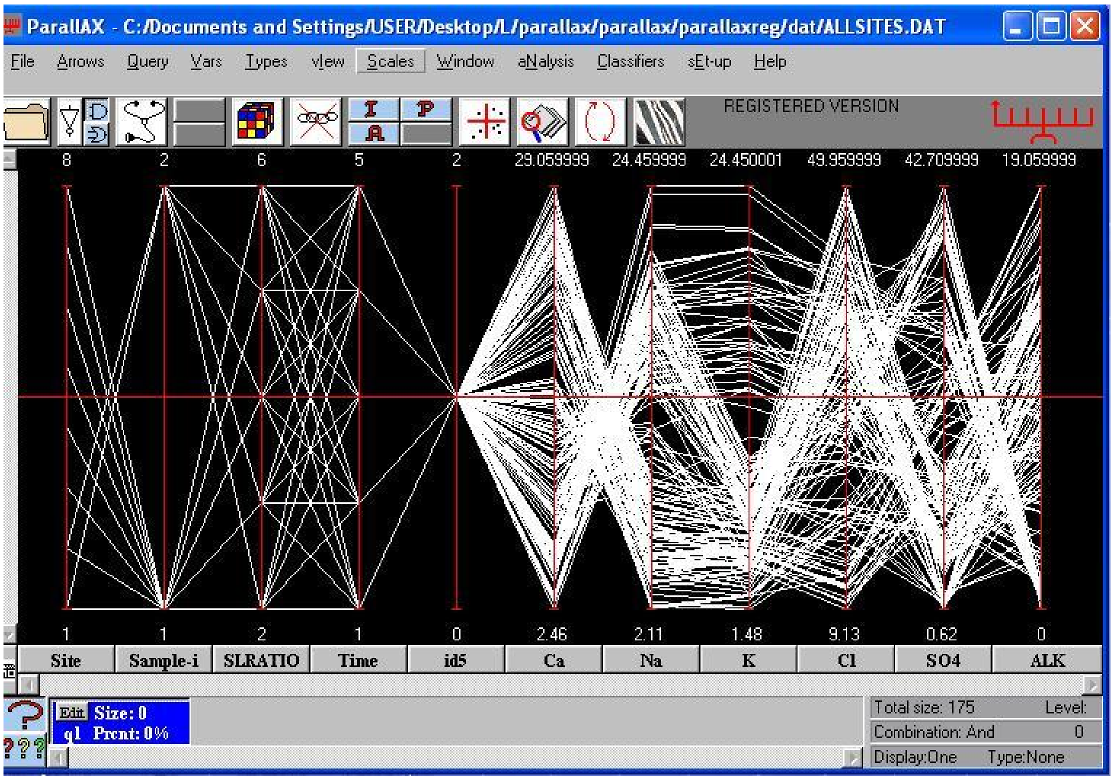
Tel/FAX: 972 – 9 – 771 - 9726

Thank you for using ParallAX!

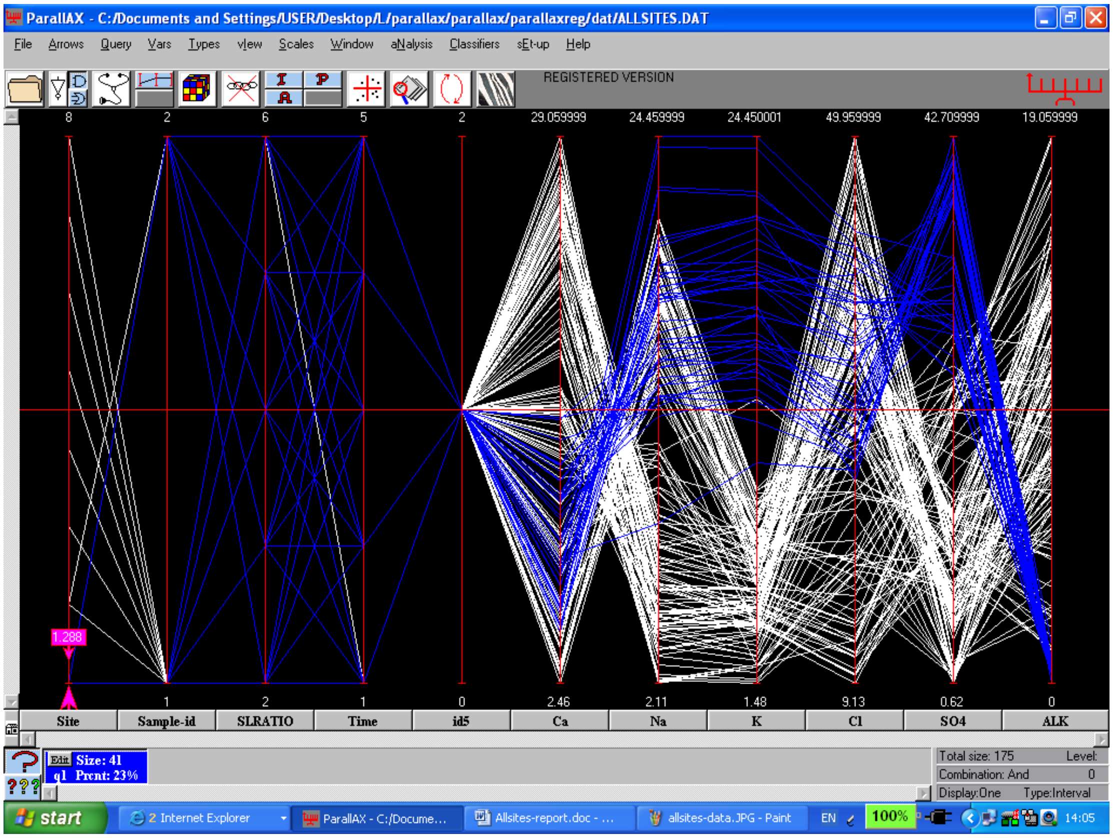
Appendix B

Classification Examples

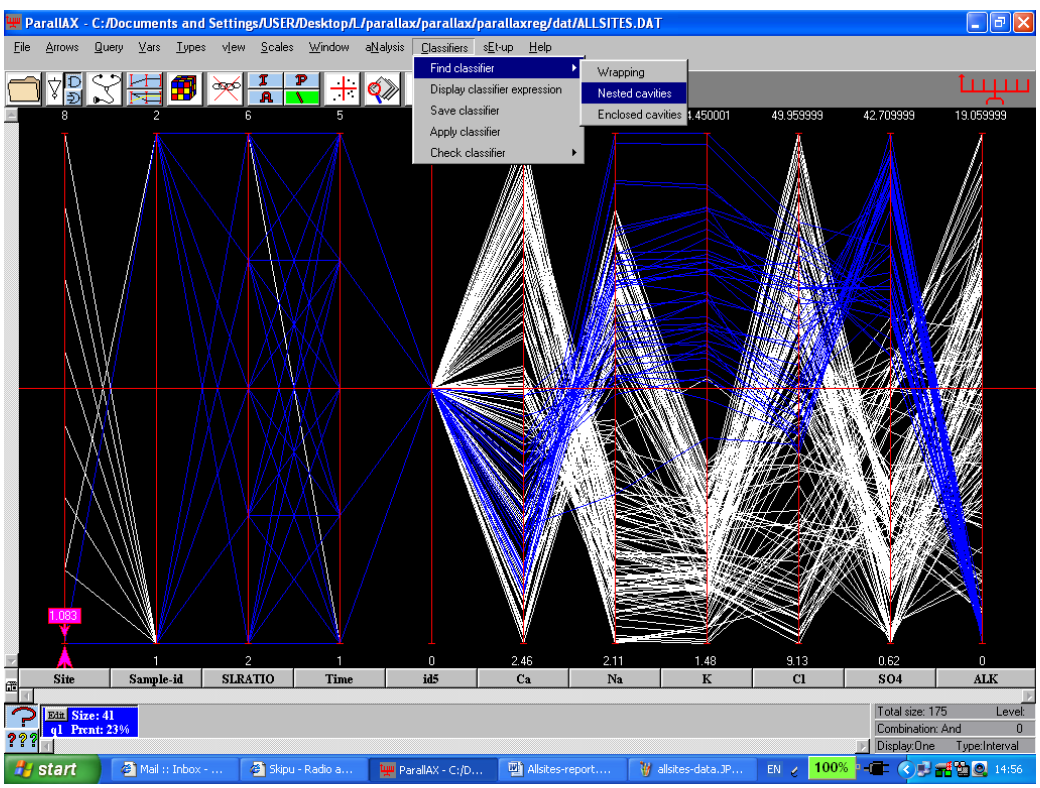
The following is an example using the data set, Allsites.dat.



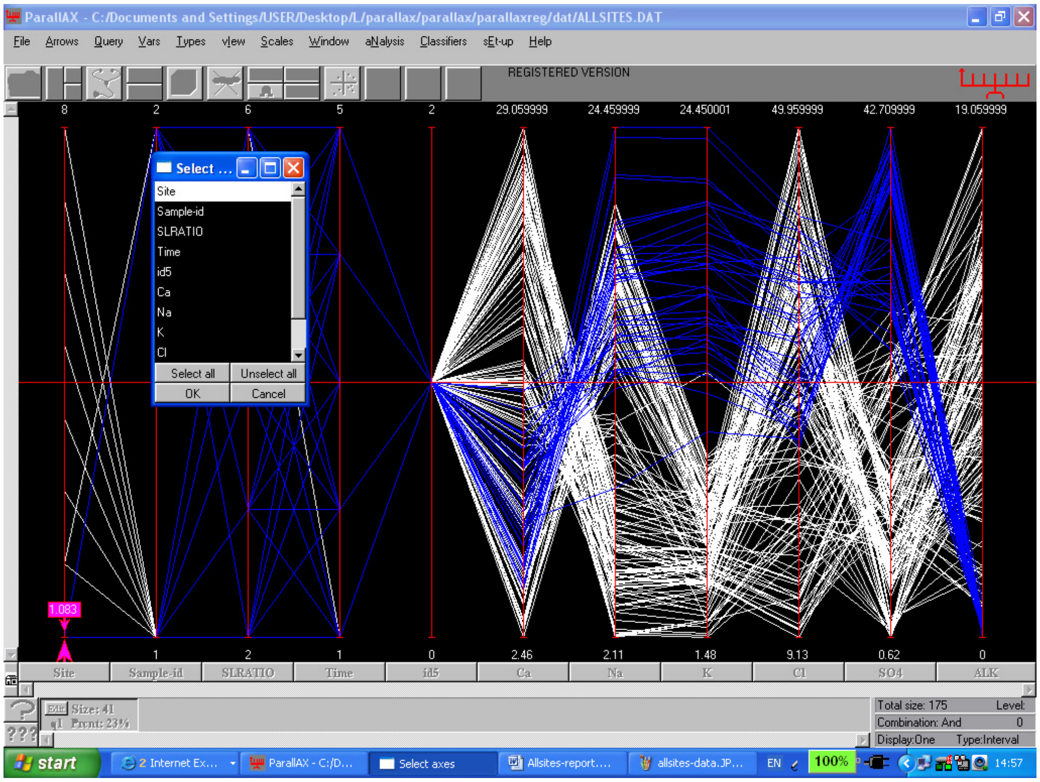
Above is the full data set; there are eight sites considered as the “classes” for classification.



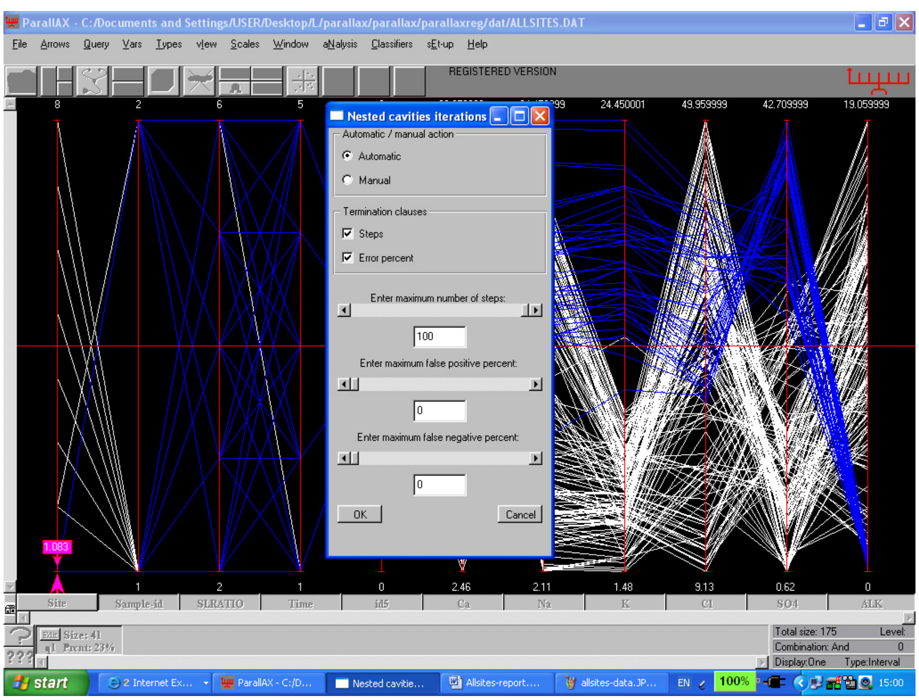
Site one is selected and is the input to the classifier.



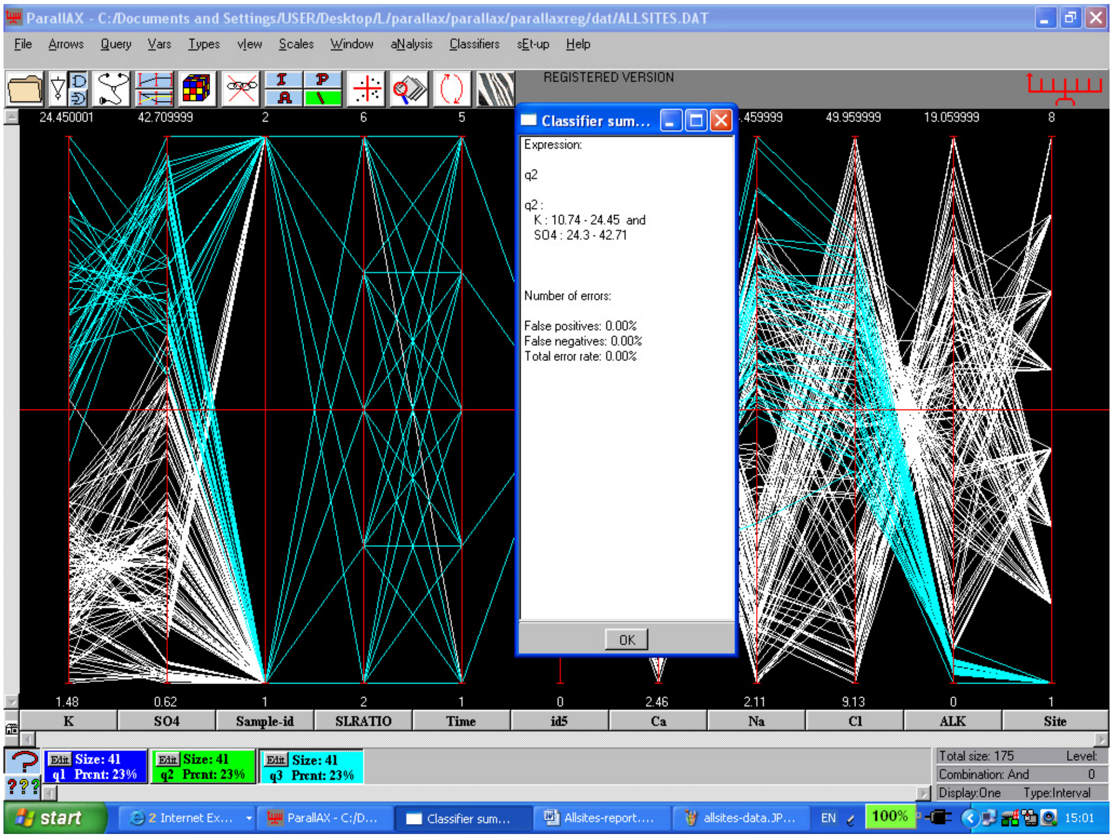
The “Classifiers” button is selected by the cursor and then the “Nested Cavities” is chosen, which is the most powerful algorithm (there are 3).



This window appears. Click on “Select All” and deselect “Sites,” which is the class variable. Then click OK.



The next box appears; click OK (accept the default).

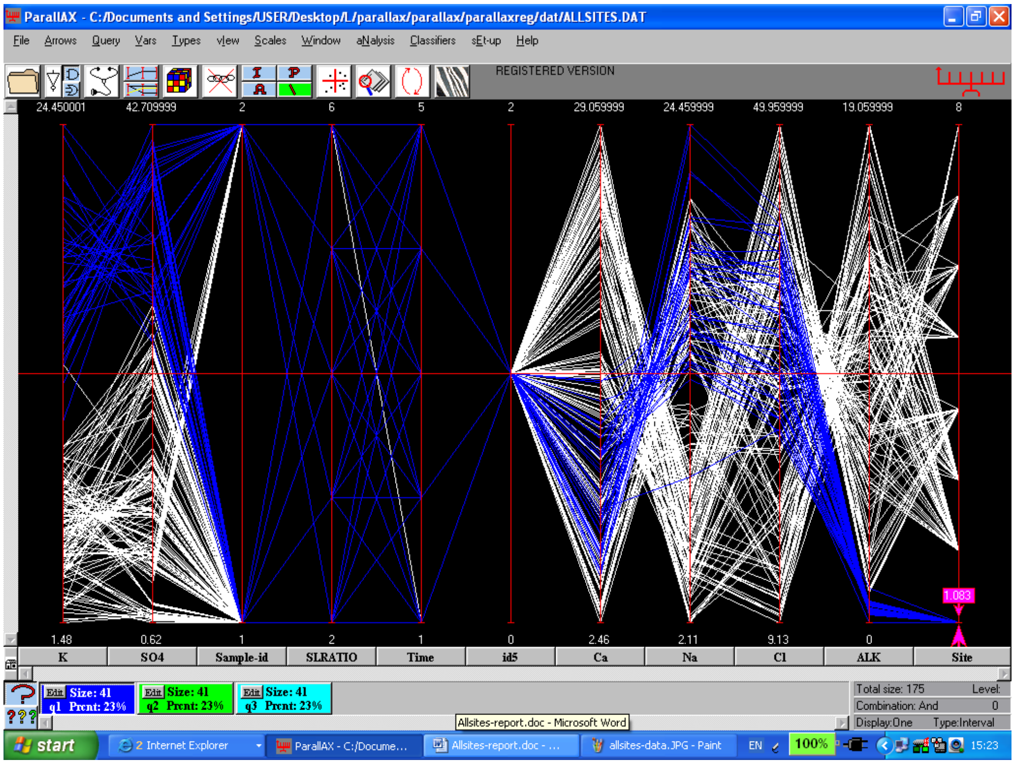


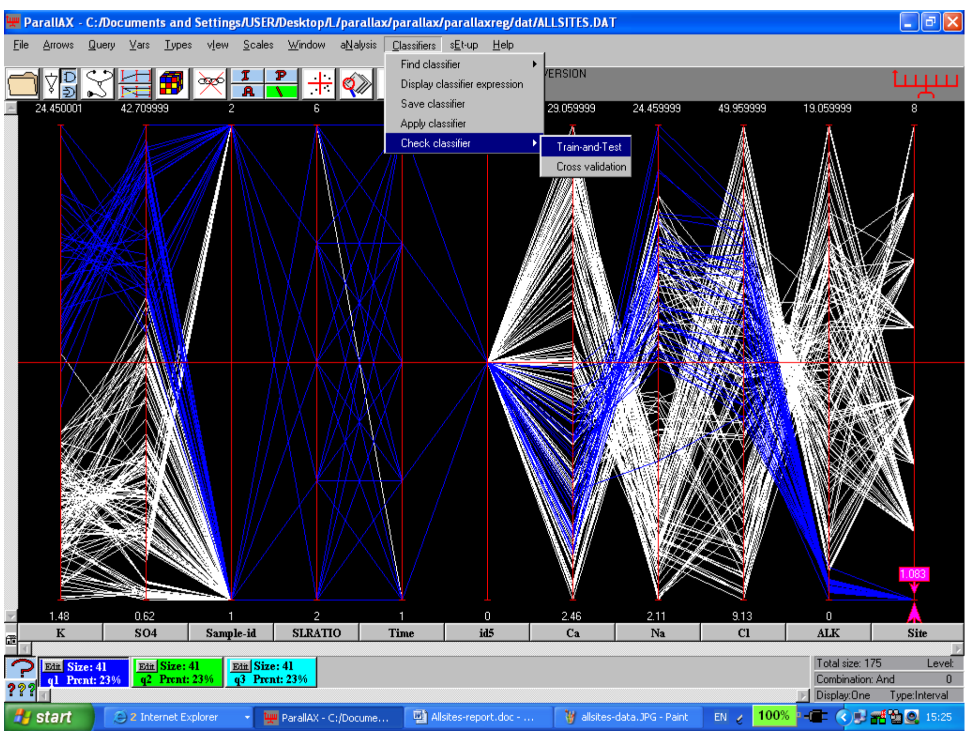
The classification result is in the above window.
The rule distinguishing Site 1 from the rest is:

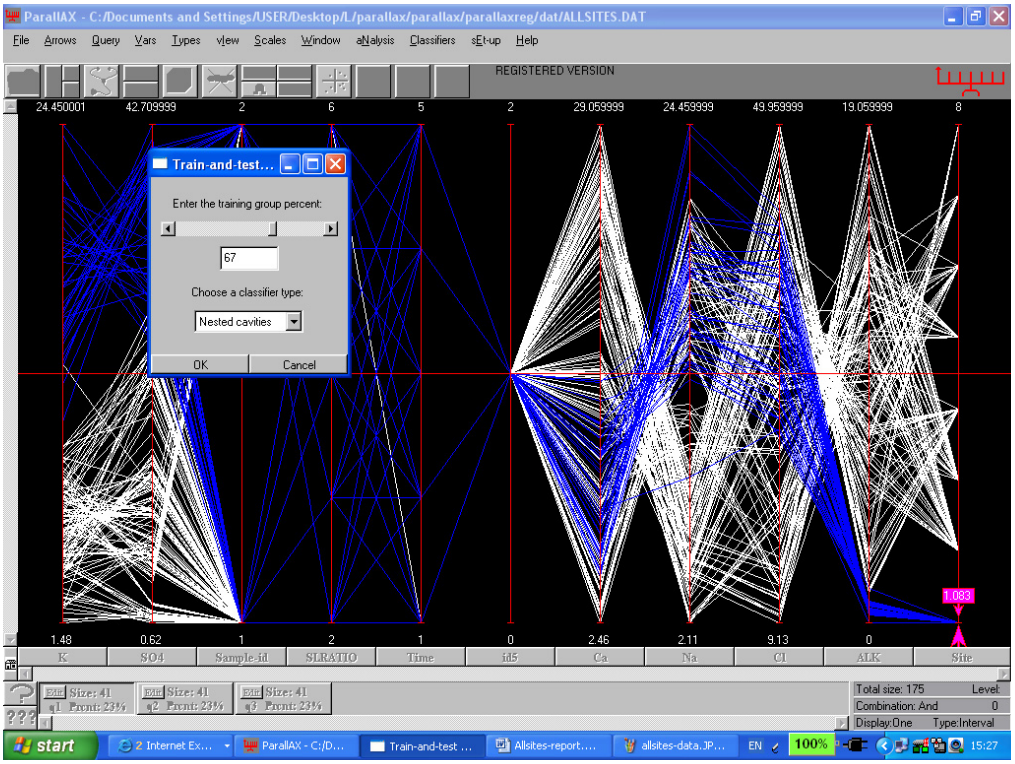
K: 10.74 - 24.45 and SO_4: 24.3 - 42.71**.**

Those are the ranges for K and SO_4. Note that the axes order is changed, with K being first (K is the best single predictor), SO_4 being second and Site (the class variable) being last. Next, the rule's precision is tested.

From the boxes on the bottom left, select the BLUE (leftmost) box.

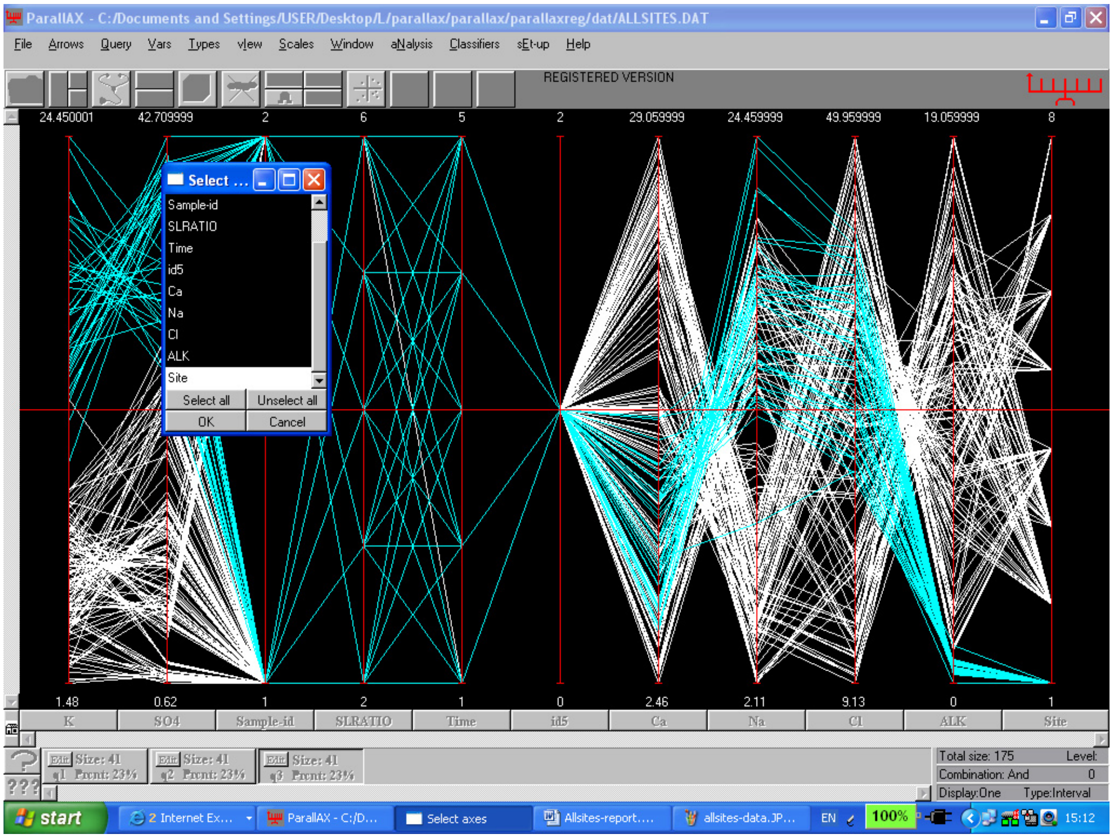

ParallAX - C:/Documents and Settings/USER/Desktop/L/parallax/parallax/parallaxreg/dat/ALLSITES.DAT
File Arrows Query Vars Types vIew Scales Window aNalysis Classifiers sEt-up Help
REGISTERED VERSION
K
SO4
Sample-id
SLRATIO
Time
id5
Ca
Na
Cl
ALK
Site
Total size: 175
Combination: And
Display:One Type:Interval


ParallAX - C:/Documents and Settings/USER/Desktop/L/parallax/parallax/parallaxreg/dat/ALLSITES.DAT
Find classifier
Display classifier expression
Save classifier
Apply classifier
Check classifier
Train-and-Test
Cross validation
K
SO4
Sample-id
SLRATIO
Time
id5
Ca
Na
Cl
ALK
Site

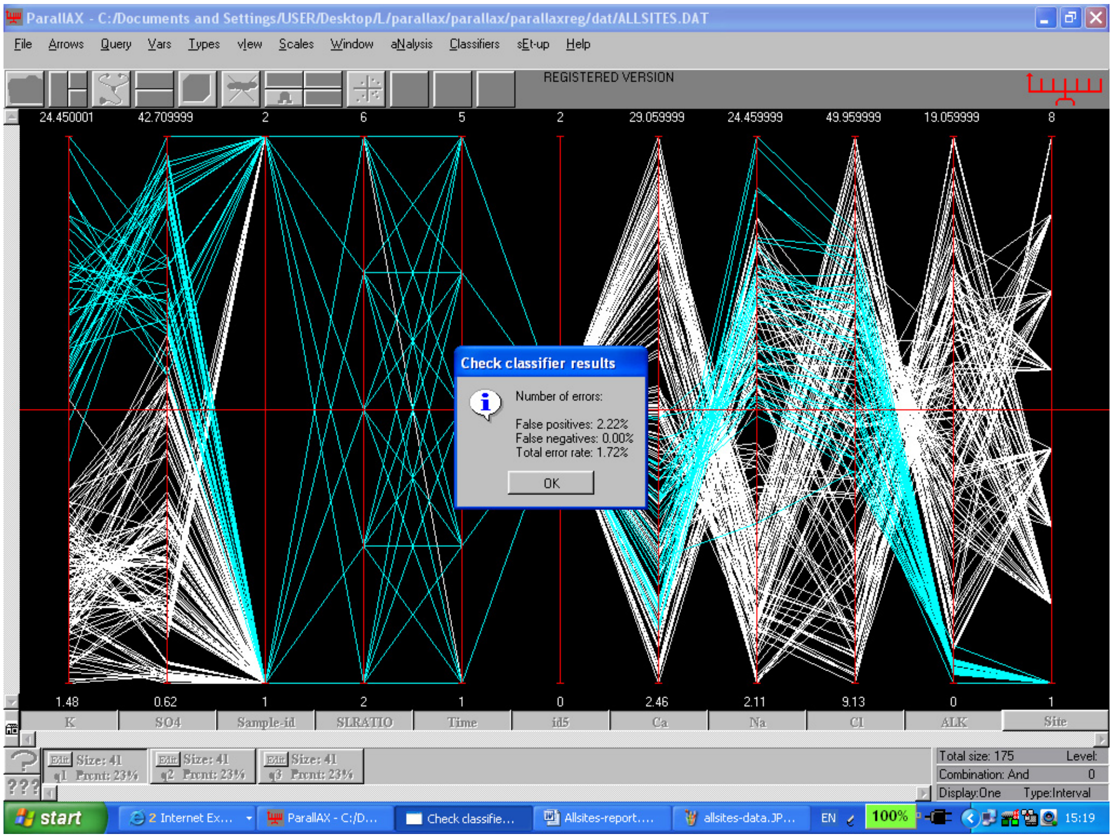


Click on "Classifiers," then (at the bottom) "Check Classifier" and then choose "Train-and-Test."

In the box which appears next, input 67 (chooses at random 67% of the data) and pick "Nested Cavities" (for the classification algorithm). A rule is then constructed based on 67% of the data, which is then tested on the remaining 33% of the data; click OK.

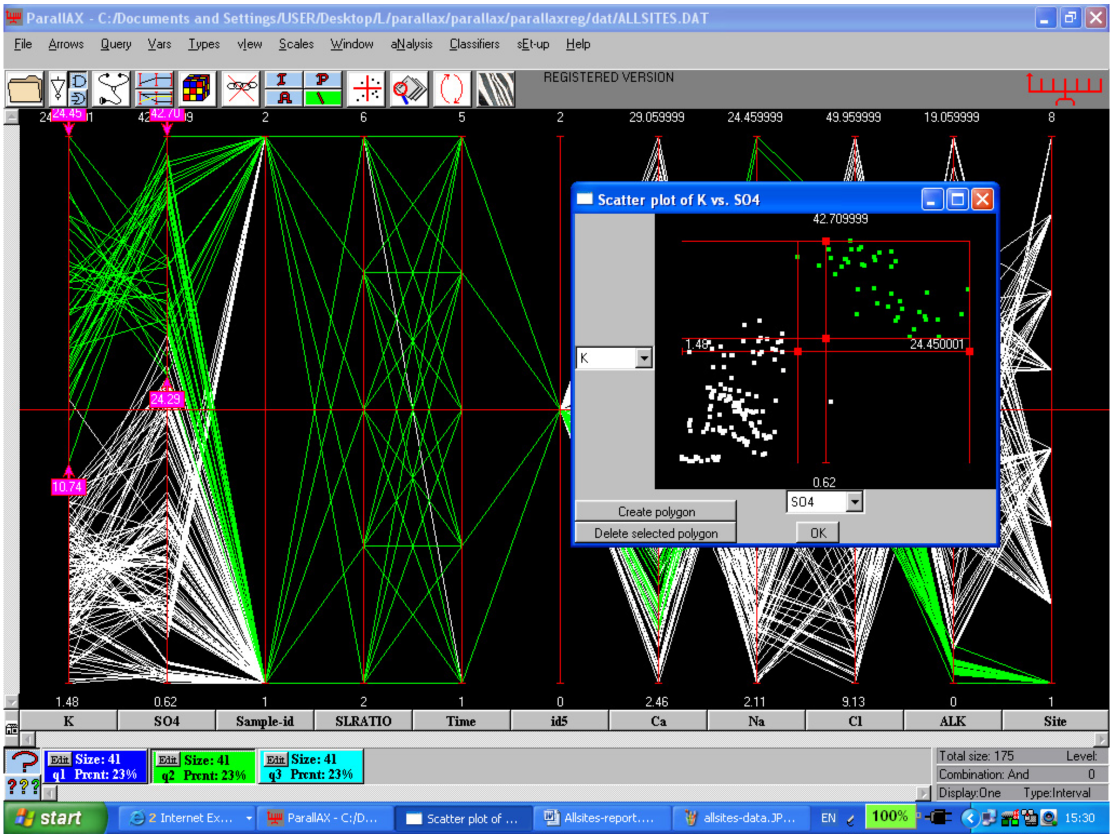


Again, "Select All" and deselect "Site," which is now at the end of the list; click OK.

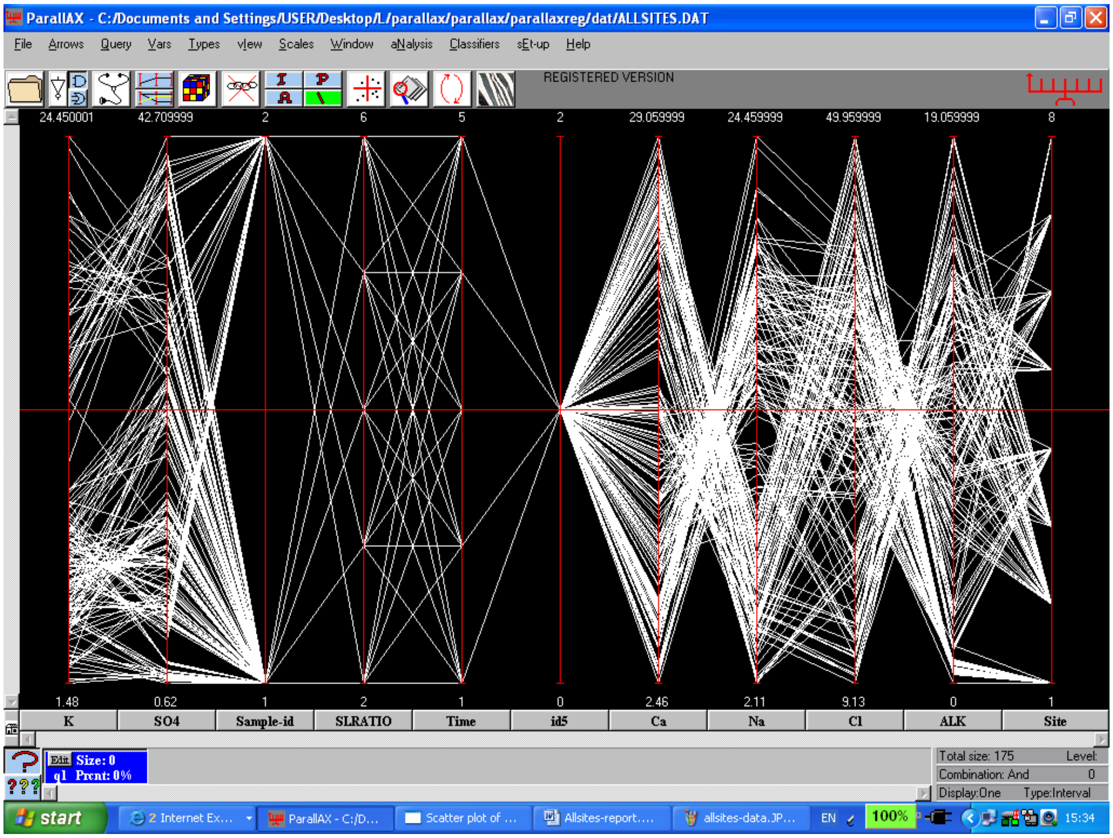


In the above window is the answer in percent of false positives, false negatives and the (weighted) average error. A high false negatives indicates that the sample is too small for a reliable rule.

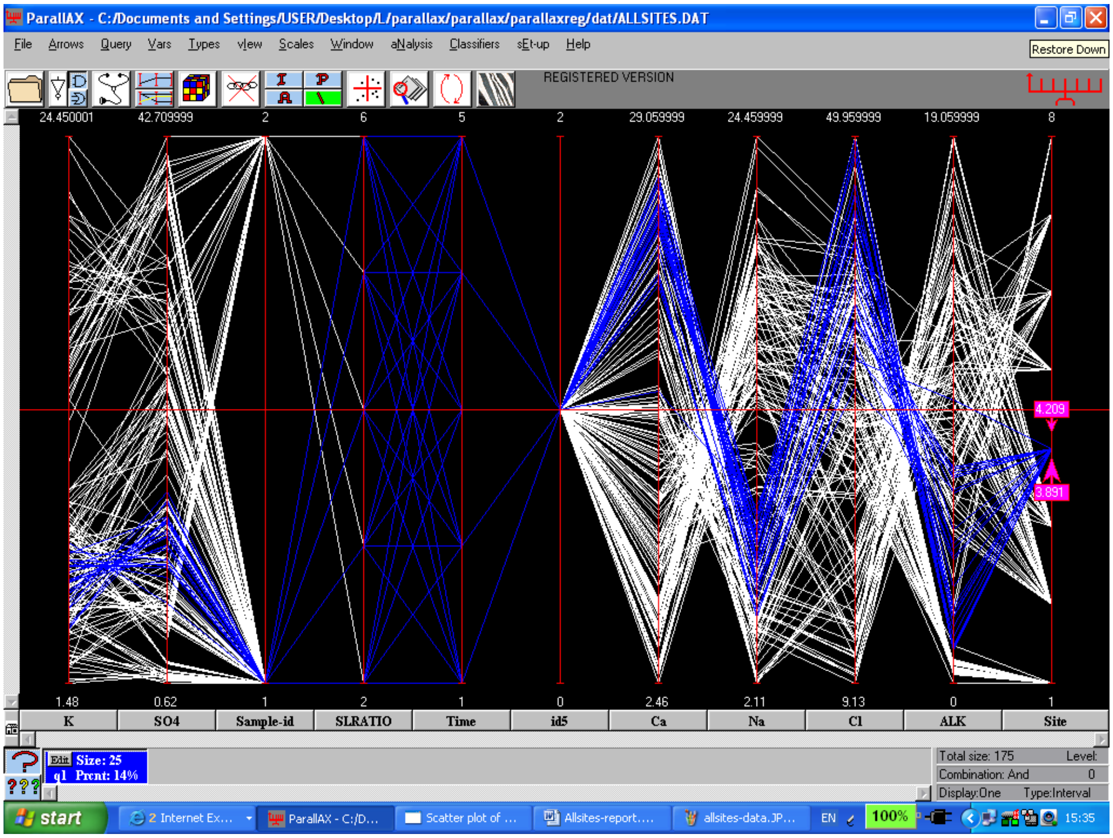
Click OK and then click on the second GREEN box at the bottom left. Then click the scatter plot button on top to obtain the K vs. SO4 plot and visually see the result of the classification. Data from Site 1 is colored GREEN and is separated from the rest of the data.



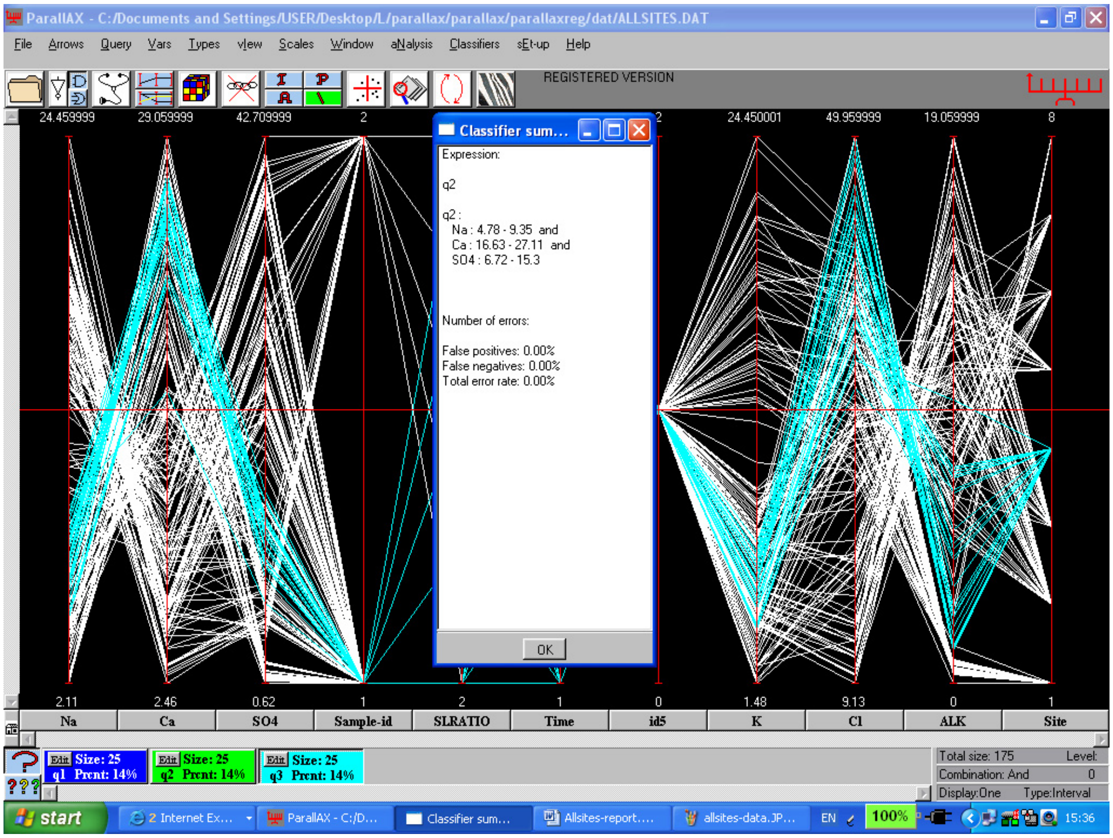
Go to the Query button on top and “Delete all queries”; the following display is next.



Repeat the classification for any other site. Here, Site 4 is chosen (the last axis).



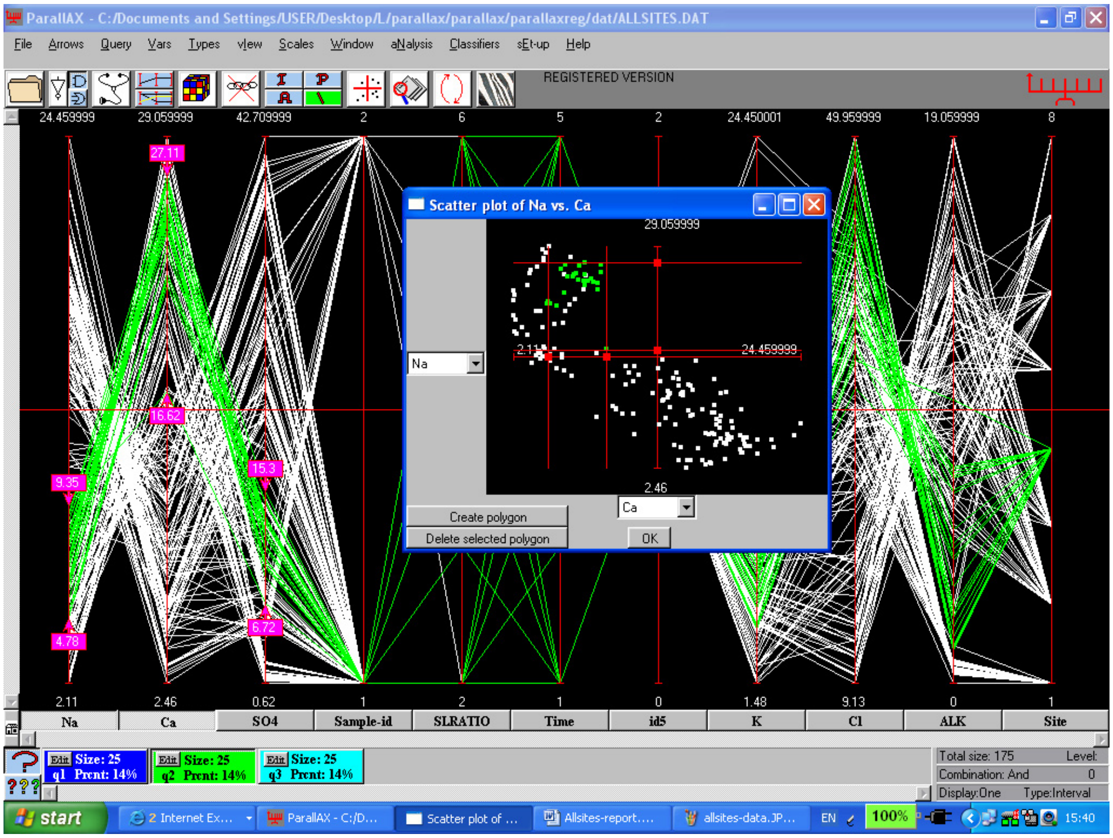
The above window is obtained.



The rule distinguishing Site 4 from the others is:

Na: 4.78 - 9.35 and Ca: 16.63 - 27.11 and SO4: 6.72 - 15.3.

The error is 0% and the plot of the first two variables is in the next window.


ParallAX - C:/Documents and Settings/USER/Desktop/L/parallax/parallax/parallaxreg/dat/ALLSITES.DAT
File Arrows Query Vars Types view Scales Window aNalysis Classifiers sEt-up Help
REGISTERED VERSION
Scatter plot of Na vs. Ca
Create polygon
Delete selected polygon
OK
Na Ca SO4 Sample-id SLRATIO Time id5 K Cl ALK Site
Total size: 175
Combination: And
Display:One
Type:Interval

Appendix C

Benford's Law

(Available in pdf version only)

Bibliography

Agullo, J., "Exact Algorithms to Compute the Least Median of Squares Estimate in Multiple Linear Regression," in L_1-Statistical Procedures and Related Topics, ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 1997, pp. 133-146.

Alqallaf, F.A. Konis, K.P., Martin, R.D., and Zamar, R.H., "Scalable Robust Covariance and Correlation Estimates for Data Mining," In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Edmonton, 2002.

Ammann, Larry P., "Robust Principal Components," Communications in Statistics — Simulation and Computation, 18, 1989, pp. 857–874.

Andersen, R., Modern Methods for Robust Regression, Sage Publications, Thousand Oaks, CA, 2007.

Anderson, T.W., An Introduction to Multivariate Statistical Analysis, Wiley-Interscience, Third Edition, July 11, 2003.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W., Robust Estimates of Location, Princeton University Press, Princeton, NJ, 1972.

Appa, G.M., and Land, A.H., "Comment on 'A Cautionary Note on the Method of Least Median of Squares' by Hettmansperger, T.P. and Sheather, S.J.," The American Statistician, 47, 1993, pp. 160-162.

Atkinson, A.C., "Fast Very Robust Methods for the Detection of Multiple Outliers," Journal of the American Statistical Association, Vol. 89, No. 428, December, 1994, pp. 1329-1339.

Atkinson, A.C. and Mulira, H.M., "The Stalactite Plot for the Detection of Multivariate Outliers," Statistics and Computing, 1993, (3), pp. 27-35.

Atkinson, A., and Riani, R., Robust Diagnostic Regression Analysis, Springer-Verlag, NY, 2000.

Atkinson, A.C., and Weisberg, S., "Simulated Annealing for the Detection of Multiple Outliers Using Least Squares and Least Median of Squares Fitting," in Directions in Robust Statistics and Diagnostics, Part 1, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 1991, pp. 7-20.

Balakrishnan, N., and Kannan N., "Variance of a Winsorized mean when the sample contains multiple outliers," Communications in Statistics — Theory and Methods, 32, 2003, pp. 139–149.

Barndorff-Nielsen, O., "Exponential Families," in Encyclopedia of Statistical Sciences, Vo1. 2, eds. Kotz, S., and Johnson, N.L., John Wiley and Sons, NY, 1982, pp. 587-596.

Barnett, V., and Lewis, T., Outliers in Statistical Data, 3rd ed., John Wiley and Sons, NY, 1994.
Beckman, R.J., and Cook, R.D., "Outliers," Technometrics, 25, 1983, pp. 119-114.

Belsley, D.A., Kuh, E., and Welsch, R.E., Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley and Sons, NY, 1980.

Bernholt, T., "Robust Estimators are Hard to Compute," 2006, Technical Report Available from (http://ls2-www.cs.uni-dortmund.de/bernholt/ps/tr52-05.pdf).

Bernholt, T., and Fischer, P. "The Complexity of Computing the MCD-Estimator," Theoretical Computer Science, 326, 2004, pp. 383-398.

Bickel, P.J., "On Some Robust Estimates of Location," The Annals of Mathematical Statistics, 36, 1965, pp. 847-858.

Bickel, P.J., "One-Step Huber Estimates in the Linear Model," Journal of the American Statistical Association, 70, 1975, pp. 428-434.

Butler, R.W., "Nonparametric Interval and Point Prediction Using Data Trimming by a Grubbs-Type Outlier Rule," The Annals of Statistics, 10, 1982, pp. 197-204.

Butler, R.W., Davies, P.L., and Jhun, M., "Asymptotics for the Minimum Covariance Determinant Estimator," The Annals of Statistics, 21, 1993, pp. 1385-1400.

Cambanis, S., Huang, S., and Simons, G., "On the Theory of Elliptically Contoured Distributions," Journal of Multivariate Analysis, 11, 1981, pp. 368-385.

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

Caroni, C., "Outlier detection by robust principal components analysis," Communications in Statistics — Simulation and Computation, 29, 2000, pp. 139–151.

Caroni, C., and Prescott, P., "Sequential Application of Wilks's Multivariate Outlier Test," Applied Statistics, 1992, 41, No. 2, pp. 355-364.

Carroll, R.J., and Welsh, A.H., "A Note on Asymmetry and Robustness in Linear Regression," The American Statistician, 42, 1988, pp. 285-287.

Cattell, R.B., "The Scree Test for the Number of Factors," Multivariate Behavioral Research, 1, 1966, pp. 245-276.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P., Graphical Methods for Data Analysis, Duxbury Press, Boston, 1983.

Chatterjee, S., and Hadi, A.S., Sensitivity Analysis in Linear Regression, John Wiley and Sons, NY, 1988.

Chatterjee, Samprit, and Martin Machler, "Robust regression: A weighted least squares approach," Communications in Statistics — Theory and Methods, 26, 1997, pp. 1381–1394.

Chen, C.H. and Hardie, W., Handbook of Data Visualization, Springer, Berlin, 2008, pp. 643-680.

Coakley, C.W., and Hettmansperger, T.P., “A Bounded Influence High Break Down Efficient Regression Estimator,” Journal of the American Statistical Association, 84, 1993, pp. 872-880.

Cook, R.D., “Deletion of Influential Observations in Linear Regression,” Technometrics, 19, 1977, pp. 15-18.

Cook, R.D., and Critchley, F., “Identifying Outliers and Regression Mixtures Graphically,” Journal of the American Statistical Association, 95, 2000, pp. 781-794.

Cook, R.D., and Hawkins, D.M., “Comment on ‘Unmasking Multivariate Outliers and Leverage Points’ by P.J. Rousseeuw and B.C. van Zomeren,” Journal of the American Statistical Association, 85, 1990, pp. 640-644.

Cook, R.D., Hawkins, D.M., and Weisberg, S., “Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator,” Statistics and Probability Letters, 16, 1993, pp. 213-218.

Cook, R.D., and Wang, P.C., “Transformations and Influential Cases in Regression,” Technometrics, 25, 1983, pp. 337-343.

Cook, R.D., and Weisberg, S., Residuals and Influence in Regression, Chapman & Hall, London, 1982.

Croux C, Filzmoser P, and Oliveira M.R., “Algorithms for Projection-Pursuit Robust Principal Component Analysis,” Chemometrics and Intelligent Laboratory Systems, 2007.

Czorgo, S., “Testing for Normality in Arbitrary Dimension,” The Annals of Statistics, 14, 1986, pp. 708-723.

Davies, L., and Gather, U., “The Identification of Multiple Outliers,” Journal of the American Statistical Association, 88, 1993, pp. 782-792.

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.

DeCarlo, L.T., “On the Meaning and Use of Kurtosis,” Psychological Methods, Vol. 2, No. 3, 1997, pp. 292-307.

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.

Dixon, W.J., and Tukey, J.W., “Approximate Behavior of Winsorized *t* (trimming/Winsorization 2),” Technometrics, 10, 1968, pp. 83-98.

Dollinger, M.B., and Staudte, R.G., “Influence Functions of Iteratively Reweighted Least Squares Estimators,” Journal of the American Statistical Association, 86, 1991, pp. 709-716.

Draper, N.R., and Smith, H., Applied Regression Analysis, 2nd ed., John Wiley and Sons, NY, 1984.

Dufour, J., Khalaf, L., and Beaulieu, M., “Exact Skewness-Kurtosis Tests for Multivariate Normality and Goodness-of-Fit in Multivariate Regressions with Application to Asset Pricing Models,” Oxford Bulletin of Economics and Statistics, 65, Supplement (2003), 0305-9049.

Du Mond, C.E. and Lenth, R.V., “A Robust Confidence Interval for Location,” Technometrics, May 1987, Vol. 29, No. 2, pp. 211-219.

Easton, G.S., and McCulloch, R.E., “A Multivariate Generalization of Quantile-Quantile Plots,” Journal of the American Statistical Association, 85, 1990, pp. 376-386.

Efron, B. 1981. *Censored Data and Bootstrap*. Journal of American Statistical Association, Vol. 76, pp. 312-319.

Efron, B., and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall. New York.

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.

Eye, A. V. and Bogat, G.A., “Testing the Assumption of Multivariate Normality,” Psychology Science, Vol. 46, 2004 (2), pp. 243-258.

Falk, M., “Asymptotic Independence of Median and MAD,” Statistics and Probability Letters, 34, 1997, pp. 341-345.

Farebrother, R.W., “Notes on the Early History of Elemental Set Methods,” in L1-Statistical Procedures and Related Topics, ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 1997, pp. 161-170.

Fisher, A. and Horn, P., “Robust Prediction Intervals in a Regression Setting,” Computational Statistics & Data Analysis, 17, 1994, pp. 129-140.

Fox, J., Regression Diagnostics, Sage, 1991, Newbury Park, CA.

Fung, W., “Unmasking Outliers and Leverage Points: A Confirmation,” Journal of the American Statistical Association, 88, 1993, pp. 515-519.

Garner, F.C., Stapanian, M.A., and Fitzgerald, K.E., “Finding Causes of Outliers in Multivariate Environmental Data,” Journal of Chemometrics, Vol. 5, 1991, pp. 241-248.

Gather, U., and Becker, C., “Outlier Identification and Robust Methods,” in Robust Inference, eds. Maddala, G.S., and Rao, C.R., Elsevier Science B.V., Amsterdam, 1997, pp. 123-144.

Giummol`e, F. and Ventura, L., “Robust Prediction Limits Based on M-estimators,” Statistics and Probability Letters, 76, 2006, pp. 1725-1740

Gnanadesikan, R., Methods for Statistical Data Analysis of Multivariate Observations, 2nd ed., John Wiley and Sons, NY, 1997.

Gnanadesikan, R., and Kettenring, J.R., “Robust Estimates, Residuals, and Outlier Detection with Multi-response Data,” Biometrics, 28, 1972, pp. 81-124.

Gray, J.B., “Graphics for Regression Diagnostics,” in the American Statistical Association 1985 Proceedings of the Statistical Computing Section, 1985, pp. 102-108.

Green, P. J., “Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives (with discussion),” Journal of the Royal Statistical Society, Series B 46, 1984, pp.149–192.

Gross, A.M., “Confidence Interval Robustness with Long-Tailed Symmetric Distributions,” Journal of the American Statistical Association, 71, 1976, pp. 409-417.

Guenther, W.C., “Shortest Confidence Intervals,” The American Statistician, 23, 1969, pp. 22-25.

Hadi, A.S., “Identifying Multiple Outliers in Multivariate Data,” J.R. Statist. Soc. B, 54, No. 3, 1992, pp. 761-771.

Hadi, A.S., and Simonoff, J.S., “Procedures for the Identification of Multiple Outliers in Linear Models,” Journal of the American Statistical Association, 88, 1993, pp. 1264-1272.

Hahn, G.J. and Meeker, W.Q., Statistical Intervals, John Wiley and Sons, 1991.

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

Hampel, F.R., “Beyond Location Parameters: Robust Concepts and Methods,” Bulletin of the International Statistical Institute, 46, 1975, pp. 375-382.

Hampel, F.R., “The Break Down Points of the Mean Combined with Some Rejection Rules,” Technometrics, 27, 1985, pp. 95-107.

Hampel, Frank R.; Elvezio M. Ronchetti; Peter J. Rousseeuw; and Werner A. Stahel, Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, New York, 1986.

Hawkins, D.M., Identification of Outliers, Chapman & Hall, London, 1980.

Hawkins, D.M., “The Accuracy of Elemental Set Approximations for Regression,” Journal of the American Statistical Association, 88, 1993, pp. 580-589.

Hawkins, Douglas M., “A Feasible Solution Algorithm for Minimum Volume Ellipsoid Estimator in Multivariate Data,” Computational Statistics, 8, 1993, pp. 95–107.

Hawkins, Douglas M., “The Feasible Set Algorithm for Least Median of Squares Regression,” Computational Statistics & Data Analysis, 16, 1993, pp. 81–101.

Hawkins, D.M., “The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data,” Computational Statistics and Data Analysis, 17, 1994, pp. 197-210.

Hawkins, D.M., Bradu, D., and Kass, G.V., “Location of Several Outliers in Multiple Regression Data Using Elemental Sets,” Technometrics, 26, 1984, pp. 197-208.

Hawkins, D.M., and Simonoff, J.S., “High Break Down Regression and Multivariate Estimation,” Applied Statistics, 42, 1993, pp. 423-432.

He, X., and Fung, W.K., “High Break Down Estimation for Multiple Populations with Applications to Discriminant Analysis,” Journal of Multivariate Analysis, 72, 2000, pp. 151-162.

He, X., and Wang, G., “Cross-Checking Using the Minimum Volume Ellipsoid Estimator,” Statistica Sinica, 6, 1996, pp. 367-374.

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

Hettmansperger, T.P., and Sheather, S.J., “A Cautionary Note on the Method of Least Median Squares,” The American Statistician, 46, 1992, pp. 79-83.

Hills, M., “Allocation Rules and their Error Rates,” Journal of the Royal Statistical Society, Series B, Vol. 28, No. 1, 1966, pp. 1-31.

Hinich, M.J., and Talwar, P.P., “A Simple Method for Robust Regression,” Journal of the American Statistical Association, 70, 1975, pp. 113-119.

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

Hoaglin, D.C., and Welsh, R., “The Hat Matrix in Regression and ANOVA,” The American Statistician, 32, 1978, pp. 17-22.

Horn, P.S., “Some Easy t-Statistics,” Journal of the American Statistical Association, 78, 1983, pp. 930-936.

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., “Discussion of ‘Multivariate Outlier Detection and Robust Covariance Matrix Estimation’ by D. Pena and F.J. Prieto,” Technometrics, 43, 2001, pp. 303-306.

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

Hubert, M., Rousseeuw, P.J., and van Aelst, S., “High Break Down Multivariate Methods,” Statistical Science, 2007.

Hung, C.K., and Inselberg, A., “Description of Surfaces in Parallel Coordinates by Linked Planar Regions,” in Mathematics of Surfaces, R. Martin, M. Sabin, and J. Winkler (Eds.), Springer-Verlag, Berlin, 2007, pp. 177-208.

Iglewicz, B., and Hoaglin, D.C., How to Detect and Handle Outliers, Quality Press, American Society for Quality, Milwaukee, Wisconsin, 1993.

Inselberg, A. Parallel Coordinates, Visual Multidimensional Geometry and its Applications, Springer, Berlin, (expected June 2009).

Insightful, S-Plus 6 Robust Library User’s Guide, Insightful Corporation, Seattle, WA, 2002. Available from (http://math.carleton.ca/ffhelp/Splus/robust.pdf).

Jaeckel, L.A., “Robust Estimates of Location: Symmetry and Asymmetric Contamination,” The Annals of Mathematical Statistics, 42, 1971, pp. 1020-1034.

Jennings, L.W. and Young, D.M., “Extended Critical Values of the Multivariate Extreme Deviate Test for Detecting a Single Spurious Observation,” Commun. Statist. –Simula., 1988, 17(4), 1359-1373.

Johnson, R.A., and Wichern, D.W., Applied Multivariate Statistical Analysis, 2nd ed., Prentice Hall, Englewood Cliffs, NJ, 1988.

Justel, A., Pena, D., and Zamar, R., “A Multivariate Kolmogorov-Smirnov Test of Goodness of Fit,” Statistical & Probability Letters, 35, 1997, pp. 251-259.

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

Koltchinskii, V.I., and Li, L., “Testing for Spherical Symmetry of a Multivariate Distribution,” Journal of Multivariate Analysis, 65, 1998, pp. 228-244.

Koziol, J.A., "Probability Plots for Assessing Multivariate Normality," The Statistician, 42, 1993, pp. 161-173.

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

Lax, D.A., "Robust Estimators of Scale: Finite Sample Performance in Long-Tailed Symmetric Distributions," Journal of the American Statistical Association, 80, 1985, pp. 736-741.

Li, R., Fang, K., and Zhu, L., "Some Q-Q Probability Plots to Test Spherical and Elliptical Symmetry," Journal of Computational and Graphical Statistics, 6, 1997, pp. 435-450.

Ma, Y., and Genton, M.G., "Highly Robust Estimation of Dispersion Matrices," Journal of Multivariate Analysis, 78, 2001, pp. 11-36.

Maddela, G.S., and Rao, C.R. (editors), Robust Inference, Handbook of Statistics 15, Elsevier Science B.V., Amsterdam, 1997.

Mallows, C., "Some Comments on Cp," Technometrics, 15, 1973, pp. 661-676.

Marazzi, A., Algorithms, Routines, and S Functions for Robust Statistics, Wadsworth and Brooks/Cole, Belmont, CA, 1993.

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.

Mardia, K.V., "Assessment of Multinormality and the Robustness of Hotelling's T^2," Applied Statistics, 24, 1975, pp. 163-171.

Mardia, K.V., Mardia's Test of Multinormality, Kotz L., Johnson, N.L. (eds), Encyclopedia of Statistical Sciences, Vol. 5, 1985, pp. 217-221.

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

Mardia, K.V. and Kanazawa, M., "The Null Distribution of Multivariate Kurtosis," Commun. Statist.-Simula. Computa., 12(5), 1983, pp.569-576.

Mardia, K.V., Kent, J.T., and Bibby, J.M., Multivariate Analysis, Academic Press, London, 1979.

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., Martin, R.D., and Yohai, V.J., Robust Statistics: Theory and Methods, John Wiley and Sons, Hoboken, NJ, 2006.

Maronna, R.A., Stahel, W.A., and Yohai, V.J., “Bias-Robust Estimators of Multivariate Scatter Based on Projections,” Journal of Multivariate Analysis, 42, 1992, pp. 141-161.

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

Mayo, M.S., and Gray, J.B., “Elemental Subsets: the Building Blocks of Regression,” The American Statistician, 51, 1997, pp. 122-129.

Mecklin, C.J., and Mundfrom, D.J., On Using Asymptotic Critical Values in Testing for Multivariate Normality, Department of Mathematics and Statistics, Murray State University and University of Northern Colorado.

Mehrotra, D.V., “Robust Elementwise Estimation of a Dispersion Matrix,” Biometrics, 51, 1995, pp. 1344-1351.

Meintanis, S. G., and Donatos G.S., “A Comparative Study of Some Robust Methods for Coefficient Estimation in Linear Regression,” Computational Statistics & Data Analysis, 23, 1997, pp. 525–540.

Mφller, S.F., von Frese, J., and Bro, R., “Robust Methods for Multivariate Data Analysis,” Journal of Chemometrics, 19, 2005, pp. 549-563.

Morgenthaler, S., “A Survey of Robust Statistics,” Stat. Meth. & Appl., 2007, 15:271-293.

Morgenthaler, S., “Robust Confidence Intervals for a Location Parameter: The Configural Approach,” Journal of the American Statistical Association, Vol. 81, No. 394, June 1986, pp. 518-523.

Morgenthaler, S., Ronchetti, E., and Stahel, W.A. (editors), New Directions in Statistical Data Analysis and Robustness, Birkhauser, Boston, 1993.

Mosteller, F., and Tukey, J.W., Data Analysis and Regression, Addison-Wesley, Reading, MA, 1977.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman W., Applied Linear Statistical Models, 4th ed., McGraw-Hill, Boston, 1996.

Olive, D.J., “Applications of Robust Distances for Regression,” Technometrics, 44, 2002, pp. 64-71.

Olive, D.J., “A Resistant Estimator of Multivariate Location and Dispersion,” Computational Statistics and Data Analysis, 46, 2004, pp. 99-102.

Olive, D.J., “Prediction Intervals for Regression Models,” Computational Statistics and Data Analysis, 51, 2007, pp. 3115-3122.

Olive, D.J., and Hawkins, D.M., "Robust Regression with High Coverage," Statistics and Probability Letters, 63, 2003, pp. 259-266.

Ozturk, Omer, and Thomas P. Hettmansperger, "Simultaneous robust estimation of location and scale parameters: A minimum distance approach," Canadian Journal of Statistics, 26, 1998, pp. 217–229 (Corrections, 1999, ibid.27, 667).

Pena, D., and Prieto, F.J., "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," Technometrics, 2001, pp. 286-299.

Penny, K.L., "Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance," Applied Statistics, Vol. 45, No. 1, 1996, pp. 73-81.

Portnoy, S., "Using Regression Quantiles to Identify Outliers," in Statistical Data Analysis Based on the L1 Norm and Related Methods, ed. Y. Dodge, North Holland, Amsterdam, 1987, pp. 345-356.

ProUCL 3.0, A Statistical Software, National Exposure Research Lab, EPA, Las Vegas Nevada, October 2004. The software ProUCL 3.0 can be freely downloaded from the EPA Web site: http://www.epa.gov/nerlesd1/tsc/tsc.htm

Rao, C.R., Linear Statistical Inference and Its Applications, John Wiley and Sons, NY, 1973.

Rocke, D.M., and Woodruff, D.L., "Identification of Outliers in Multivariate Data," Journal of the American Statistical Association, 91, 1996, pp. 1047-1061.

Rocke, D.M., and Woodruff, D.L., "Robust Estimation of Multivariate Location and Shape," Journal of Statistical Planning and Inference, 57, 1997, pp. 245-255.

Rocke, D.M., and Woodruff, D.L., "Discussion of 'Multivariate Outlier Detection and Robust Covariance Matrix Estimation' by D. Pena and F.J. Prieto," Technometrics, 43, 2001, pp. 300-303.

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

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.

Ruiz-Gazen, A., "A Very Simple Robust Estimator of a Dispersion Matrix," Computational Statistics and Data Analysis, 21, 1996, pp. 149-162.

Ruppert, D., "Computing S-Estimators for Regression and Multivariate Location/Dispersion," Journal of Computational and Graphical Statistics, 1, 1992, pp. 253-270.

Ruppert, D., and Carroll, R.J., "Trimmed Least Squares Estimation in the Linear Model," Journal of the American Statistical Association, 75, 1980, pp. 828-838.

Scout, A Data Analysis Program, Technology Support Project, 2002, USEPA, NERL-LV, Las Vegas, Nevada.

Seber, G.A.E., Multivariate Observations, John Wiley & Sons, 1984.

Simonoff, J.S., "The Break Down and Influence Properties of Outlier-Rejection-Plus-Mean Procedures," Communications in Statistics Theory and Methods, 16, 1987, pp. 1749-1769.

Simonoff, J.S., "Outlier Detection and Robust Estimation of Scale," Journal of Statistical Computation and Simulation, 27, 1987, pp. 79-92.

Simpson, D.G., Ruppert, D., and Carroll, R.J., "On One-Step GM Estimates and Stability of Inferences in Linear Regression," Journal of the American Statistical Association, 87, 1992, pp. 439-450.

Simpson, James R., and Douglas C. Montgomery, "The Development and Evaluation of Alternative Generalized M Estimation Techniques," Communications in Statistics — Simulation and Computation, 27, 1998, pp. 999–1018.

Simpson, James R., and Douglas C. Montgomery, "A Performance Based Assessment of Robust Regression Methods," Communications in Statistics — Simulation and Computation, 27, 1988, pp. 1031–1049.

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 Chemometric 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.

Singh, A. and Nocerino, J.M., "Robust Estimation of the Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations," Chemometrics and Intelligent Laboratory Systems, Vol. 60, 2002, pp. 69-86.

Singh, A. and Singh, A.K., Estimation of the Exposure Point Concentration Term (95% UCL), Using Bias-Corrected Accelerated (BCA) Bootstrap Method and Several Other Methods for Normal, Lognormal, and Gamma Distributions, Draft EPA Internal Report, 2003.

Singh, A., Singh, A.K., and Iaci, R.J., Estimation of the Exposure Point Concentration Term Using a Gamma Distribution, EPA/600/R-02/084, October, 2002.

Singh, A.K., Singh, A., and Engelhardt, M., The lognormal Distribution in Environmental Applications, Technology Support Center Issue Paper, 1997. 182CMB97, EPA/600/R-97/006.

Singh, A.K., Singh, A., and Engelhardt, M., Some Practical Aspects of Sample Size and Power Computations for Estimating the Mean of Positively Skewed Distributions in Environmental Applications, 1999, EPA/600/S-99/006.

Snapinn, S. and Knoke, J., "Estimation of Error Rates in Discriminant Analysis with Selection of Variables," Biometrics, Vol. 45, No. 1, March 1989, pp. 289-299.

Staudte, R.G., and Sheather, S.J., Robust Estimation and Testing, John Wiley and Sons, NY, 1990.

Stahel, W., and Weisberg, S., Directions in Robust Statistics and Diagnostics, Part 1, Springer-Verlag, NY, 1991.

Stahel, W., and Weisberg, S., Directions in Robust Statistics and Diagnostics, Part 2, Springer-Verlag, NY, 1991.

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.

Stefanski, L.A., "A Note on High-Break Down Estimators," Statistics and Probability Letters, 11, 1991, pp. 353-358.

Stefanski, L.A., and Boos, D.D., "The Calculus of M–estimators," The American Statistician, 56, 2002, pp. 29-38.

Stigler, S.M., “The Asymptotic Distribution of the Trimmed Mean,” The Annals of Mathematical Statistics, 1, 1973, pp. 472-477.

Stigler, S.M., “Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920,” Journal of the American Statistical Association, 68, 1973, pp. 872-878.

Stigler, S.M., “Do Robust Estimators Work with Real Data?” The Annals of Statistics, 5, 1977, pp. 1055-1098.

Street, J.O., Carroll, R.J., and Ruppert, D., “A Note on Computing Regression Estimates Via Iteratively Reweighted Least Squares,” The American Statistician, 42, 1988, pp. 152-154.

Stromberg, A.J., “Computing the Exact Least Median of Squares Estimate and Stability Diagnostics in Multiple Linear Regression,” SIAM Journal of Scientific and Statistical Computing, 14, 1993, pp. 12891299.

Tableman, M., “The Influence Functions for the Least Trimmed Squares and the Least Trimmed Absolute Deviations Estimators,” Statistics and Probability Letters, 19, 1994, pp. 329-337.

Todorov, V., “Robust Selection of Variables in Linear Discriminant Analysis,” Stat. Meth. & Appl., 2007, 15:395-407.

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

Tukey, J.W., “Graphical Displays for Alternative Regression Fits,” in Directions in Robust Statistics and Diagnostics, Part 2, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 1991, pp. 309-326.

U.S. Environmental Protection Agency (US EPA). 2009. *ProUCL Version 4.00.04, A Statistical Software*. The software ProUCL 4.00.04 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 (US EPA). 2009. *ProUCL 4.00.04. Technical Guide* Publication EPA/600/R-07/041.

U.S. Environmental Protection Agency (US EPA). 2009. *ProUCL 4.00.04. 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.

Velleman, P.F., and Welsch, R.E., “Efficient Computing of Regression Diagnostics,” The American Statistician, 35, 1981, pp. 234-242.

Vısek, J.A., “On High Break Down Point Estimation,” Computational Statistics, 11, 1996, pp. 137-146.

Welsh, A.H., “Bahadur Representations for Robust Scale Estimators Based on Regression Residuals,” The Annals of Statistics, 14, 1986, pp. 1246-1251.

Welsh, A.H., and Ronchetti, E., “A Journey in Single Steps: Robust One-Step M-estimation in Linear Regression,” Journal of Statistical Planning and Inference, 103, 2002, pp. 287-310.

Wilcox, R.R., Introduction to Robust Estimation and Hypothesis Testing, 2nd ed., Elsevier Academic Press, San Diego, CA, 2005.

Wilcox, Rand R., and Jan Muska, “Tests of Hypothesis About Regression Parameters When Using a Robust Estimator,” Communications in Statistics — Theory and Methods, 28, 1999, pp. 2201–2212.

Willems, G., Pison, G., Rousseeuw, P.J., and Van Aelst, S., “A Robust Hotelling Test,” Metrika, 55, 2002, pp. 125-138.

Wisnowski, J.W., Simpson J.R., and Montgomery D.C., “A Performance Study for Multivariate Location and Shape Estimators,” Quality and Reliability Engineering International, 18, 2002, pp. 117-129.

Woodruff, D.L., and Rocke, D.M., “Heuristic Search Algorithms for the Minimum Volume Ellipsoid,” Journal of Computational and Graphical Statistics, 2, 1993, pp. 69-95.

Woodruff, D.L., and Rocke, D.M., “Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators,” Journal of the American Statistical Association, 89, 1994, pp. 888-896.

Xie, Y., Wang, J., Liang, Y., Sun, L., Song, X. and Yu, R., “Robust Principal Component Analysis by Projection Pursuit,” Journal of Chemometrics, Vol. 7, 1993, pp. 527-541.

Yohai, V.J. and Maronna, R., “Location Estimators Based on Linear Combinations of Modified Order Statistics,” Communications in Statistics Theory and Methods, 5, 1976, pp. 481-486.

Yohai, Victor J., and Zamar R.H., “High break down point estimates of regression by means of the minimization of an efficient scale,” Journal of the American Statistical Association, 83, 1988, pp. 406–413. (See also ibid., 1989, 84, 636.)

Glossary

Anderson-Darling (AD) test: The Anderson-Darling test assesses whether known data come from a specified distribution.

Bias: The systematic or persistent distortion of a measured value from its true value (this can occur during sampling design, the sampling process, or laboratory analysis).

Biweight: An influence function based on Tukey's or LAX/Kafadar's methods.

Bootstrap Method: The bootstrap method is a computer-based method for assigning measures of accuracy to sample estimates. This technique allows estimation of the sample distribution of almost any statistic using only very simple methods. Bootstrap methods are generally superior to ANOVA for small data sets or where sample distributions are non-normal.

Break Down point: This point represents that fraction of observations which can be altered (e.g., can be made very large) arbitrarily without affecting (influencing, distorting, changing drastically) the values of the estimates.

Central Limit Theorem (CLT): The central limit theorem states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size, increases.

Coefficient of Variation (CV): A dimensionless quantity used to measure the spread of data relative to the size of the numbers. For a normal distribution, the coefficient of variation is given by s/xBar. Also known as the relative standard deviation (RSD).

Confidence Coefficient: The confidence coefficient (a number in the closed interval [0, 1]) associated with a confidence interval for a population parameter is the probability that the random interval constructed from a random sample (data set) contains the true value of the parameter. The confidence coefficient is related to the significance level of an associated hypothesis test by the equality: level of significance = 1 – confidence coefficient.

Confidence Interval: Based upon the sampled data set, a confidence interval for a parameter is a random interval within which the unknown population parameter, such as the mean, or a future observation, x0, falls.

Confidence Limit: The lower or an upper boundary of a confidence interval. For example, the 95% upper confidence limit (UCL) is given by the upper bound of the associated confidence interval.

Correlation: A measure of linear association between two ordered lists.

Coverage, Coverage Probability: The coverage probability (e.g., = 0.95) of an upper confidence limit (UCL) of the population mean represents the confidence coefficient associated with the UCL.

Critical Alpha: The cutoff level for finding outliers.

Cross validation: The method of checking if the classification of observations in discriminant analysis are valid or not.

Data Quality Objectives (DQOs): Qualitative and quantitative statements derived from the DQO process that clarify study technical and quality objectives, define the appropriate type of data, and specify tolerable levels of potential decision errors that will be used as the basis for establishing the quality and quantity of data needed to support decisions.

Detection Limit: A measure of the capability of an analytical method to distinguish samples that do not contain a specific analyte from samples that contain low concentrations of the analyte. The lowest concentration or amount of the target analyte that can be determined to be different from zero by a single measurement at a stated level of probability. Detection limits are analyte- and matrix-specific and may be laboratory-dependent.

Empirical Distribution Function (EDF): In statistics, an empirical distribution function is a cumulative probability distribution function that concentrates probability $1/n$ at each of the n numbers in a sample.

Estimate: A numerical value computed using a random data set (sample), and is used to guess (estimate) the population parameter of interest (e.g., mean). For example, a sample mean represents an estimate of the unknown population mean.

Expectation Maximization (EM): The EM algorithm is used to approximate a probability function (p.f. or p.d.f.). EM is typically used to compute maximum likelihood estimates given incomplete samples.

Exposure Point Concentration (EPC): The contaminant concentration within an exposure unit to which the receptors are exposed. Estimates of the EPC represent the concentration term used in exposure assessment.

Extreme Values: The minimum and the maximum values.

Goodness-of-Fit (GOF): In general, the level of agreement between an observed set of values and a set wholly or partly derived from a model of the data.

Graphics Alpha: The alpha values used for identifying outliers on the graphs. This is usually same as critical alpha.

Gray Region: A range of values of the population parameter of interest (such as mean contaminant concentration) within which the consequences of making a decision error are relatively minor. The gray region is bounded on one side by the action level. The width of the gray region is denoted by the Greek letter delta in this guidance.

H-Statistic: The unique symmetric unbiased estimator of the central moment of a distribution.

H-UCL: UCL based on Land's H-Statistic.

Hypothesis: Hypothesis is a statement about the population parameter(s) that may be supported or rejected by examining the data set collected for this purpose. There are two hypotheses: a null hypothesis, (H_0), representing a testable presumption (often set up to be rejected based upon the sampled data), and an alternative hypothesis (H_A), representing the logical opposite of the null hypothesis.

Individual MD(α): The $\alpha 100\%$ critical value from the distribution of the distances (also called d0cut).

Individual Contour/Ellipsoid: Contour at Individual MD(α). Also called a prediction ellipsoid.

Influence Function Alpha: The values used for minimizing in Huber and PROP methods.

Jackknife Method: A statistical procedure in which, in its simplest form, estimates are formed of a parameter based on a set of N observations by deleting each observation in turn to obtain, in addition to the usual estimate base d on N observations, N estimates each based on N-1 observations.

Kolmogorov-Smirnov (KS) test: The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution. The Kolmogorov-Smirnov test is based on the empirical distribution function (EDF).

Kurtosis: Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution.

Level of Significance: The error probability (also known as false positive error rate) tolerated of falsely rejecting the null hypothesis and accepting the alternative hypothesis.

Leverage Distances: The distances (robust or classical Mahalanobis) obtained using the independent variables in regression.

Leverage Outliers: The outliers among the independent variables in regression.

Lilliefors test: A test of normality for large data sets when the mean and variance are unknown.

M-Estimation: The process of obtaining an M-estimators.

M-Estimators: A class of statistics which are obtained as the solution to the problem of minimizing certain functions of the data.

Max MD: Largest Mahalanobis distance obtained from the dataset.

Max MD(α): The α100% critical value of the test statistic (also called d2max).

Maximum Likelihood Estimates (MLE): Maximum likelihood estimation (MLE) is a popular statistical method used to make inferences about parameters of the underlying probability distribution of a given data set.

Mean: The sum of all the values of a set of measurements divided by the number of values in the set; a measure of central tendency.

Median: The middle value for an ordered set of n values. Represented by the central value when n is odd or by the average of the two most central values when n is even. The median is the 50th percentile.

Minimization Criterion: The criterion used in minimizing the residuals of regression.

Minimum Detectable Difference (MDD): The minimum detectable difference (MDD) is the smallest difference in means that the statistical test can resolve. The MDD depends on sample-to-sample variability, the number of samples, and the power of the statistical test.

Minimum Variance Unbiased Estimates (MVUE): A minimum variance unbiased estimator (MVUE or MVU estimator) is an unbiased estimator of parameters, whose variance is minimized for all values of the parameters. If an estimator is unbiased, then its mean squared error is equal to its variance.

Non-detect (ND): Censored data values.

Nonparametric: A term describing statistical methods that do not assume a particular population probability distribution, and are therefore valid for data from any population with any probability distribution, which can remain unknown.

Optimum: An interval is optimum if it possesses optimal properties as defined in the statistical literature. This may mean that it is the shortest interval providing the specified coverage (e.g., 0.95) to the population mean. For example, for normally distributed data sets, the UCL of the population mean based upon Student's t distribution is optimum.

Outlier: Measurements (usually larger or smaller than the majority of the data values in a sample) that are not representative of the population from which they were drawn. The presence of outliers distorts most statistics if used in any calculations.

p-value: In statistical hypothesis testing, the p-value of an observed value t_{observed} of some random variable T used as a test statistic is the probability that, given that the null hypothesis is true, T will assume a value as or more unfavorable to the null hypothesis as the observed value t_{observed}.

Parameter: A parameter is an unknown constant associated with a population.

Parametric: A term describing statistical methods that assume a normal distribution.

PC Loadings: A matrix of eigen vectors for the covariance or correlation matrix.

Population: The total collection of N objects, media, or people to be studied and from which a sample is to be drawn. The totality of items or units under consideration.

Prediction Interval: The interval (based upon historical data, or a background well) within which a newly and independently obtained (often labeled as a future observation) site observation (from a compliance well) of the predicted variable (lead) falls with a given probability (or confidence coefficient).

Probability of Type 2 Error (=β): The probability, referred to as β (beta), that the null hypothesis will not be rejected when in fact it is false (false negative).

Probability of Type I Error = Level of Significance (= α): The probability, referred to as α (alpha), that the null hypothesis will be rejected when in fact it is true (false positive).

pth Percentile: The specific value, X_p of a distribution that partitions a data set of measurements in such a way that the p percent (a number between 0 and 100) of the measurements fall at or below this value, and (100-p) percent of the measurements exceed this value, X_p.)

pth Quantile: The specific value of a distribution that divides the set of measurements in such a way that the proportion, p, of the measurements falls below (or are equal to) this value, and the proportion (1-p) of the measurements exceed this value.

Quality Assurance: An integrated system of management activities involving planning, implementation, assessment, reporting, and quality improvement to ensure that a process, item, or service is of the type and quality needed and expected by the client.

Quality Assurance Project Plan: A formal document describing, in comprehensive detail, the necessary QA, QC, and other technical activities that must be implemented to ensure that the results of the work performed will satisfy the stated performance criteria.

Quantile Plot: A graph that displays the entire distribution of a data set, ranging from the lowest to the highest value. The vertical axis represents the measured concentrations, and the horizontal axis is used to plot the percentiles of the distribution.

Range: The numerical difference between the minimum and maximum of a set of values.

Regression on Order Statistics (ROS): A regression line is fit to the normal scores of the order statistics for the uncensored observations and then to fill in values extrapolated from the straight line for the observations below the detection limit.

Resampling: The repeated process of obtaining representative samples and/or measurements of a population of interest.

Reliable UCL: This is similar to a stable UCL.

Regression Outliers: The outliers in the dependent variable of regression.

Robustness: Robustness is used to compare statistical tests. A robust test is the one with good performance (that is not unduly affected by outliers) for a wide variety of data distributions.

Sample: A sample here represents a random sample (data set) obtained from the population of interest (e.g., a site area, a reference area, or a monitoring well). The sample is supposed to be a representative sample of the population under study. The sample is used to draw inferences about the population parameter(s).

Shapiro-Wilk (SW) test: In statistics, the Shapiro-Wilk test tests the null hypothesis that a sample $x_1, \ldots, x_n$ came from a normally distributed population.

Simultaneous Contour/Ellipsoid: Contour at Max MD(α). Also called a tolerance ellipsoid.

Skewness: A measure of asymmetry of the distribution of the characteristic under study (e.g., lead concentrations). It can also be measured in terms of the standard deviation of log-transformed data. The higher is the standard deviation, the higher is the skewness.

Stable UCL: The UCL of a population mean is a stable UCL if it represents a number of practical merits, which also has some physical meaning. That is, a stable UCL represents a realistic number (e.g., contaminant concentration) that can occur in practice. Also, a stable UCL provides the specified (at least approximately, as much as possible, as close as possible to the specified value) coverage (e.g., ~0.95) to the population mean.

Standard Deviation (sd): A measure of variation (or spread) from an average value of the sample data values.

Standard Error (SE): A measure of an estimate's variability (or precision). The greater the standard error in relation to the size of the estimate, the less reliable the estimate. Standard errors are needed to construct confidence intervals for the parameters of interests such as the population mean and population percentiles.

Trimming percentage: The percentage value used for trimming outliers in MVT method.

Tolerance Limit: A confidence limit on a percentile of the population rather than a confidence limit on the mean. For example, a 95 percent one-sided TL for 95 percent coverage represents the value below which 95 percent of the population values are expected to fall with 95 percent confidence. In other words, a 95% UTL with coverage coefficient 95% represents a 95% upper confidence limit for the 95th percentile.

Unreliable UCL, Unstable UCL, Unrealistic UCL: The UCL of a population mean is unstable, unrealistic, or unreliable if it is orders of magnitude higher than the other UCLs of population mean. It represents an impractically large value that cannot be achieved in practice. For example, the use of Land's H statistic often results in impractically large inflated UCL value. Some other UCLs, such as the bootstrap t UCL and Hall's UCL, can be inflated by outliers resulting in an impractically large and unstable value. All such impractically large UCL values are called unstable, unrealistic, unreliable, or inflated UCLs.

Upper Confidence Limit (UCL): The upper boundary (or limit) of a confidence interval of a parameter of interest such as the population mean.

Upper Prediction Limit (UPL): The upper boundary of a prediction interval for an independently obtained observation (or an independent future observation).

Upper Tolerance Limit (UTL): The upper boundary of a tolerance interval.

Winsorization method: The Winsorization method is a procedure that replaces the n extreme values with the preset cut-off value. This method is sensitive to the number of outliers, but not to their actual values.

About the CD

The CD accompanying the hard copy of this report, "Scout 2008 Version 1.0 User Guide," contains the following contents:

- Scout 2008 Version 1.00.01 statistical software.

- J.M. Nocerino (editor), A. Singh, R. Maichle, N. Armbya, and A.K. Singh, "Scout 2008 Version 1.0 User Guide." U.S. Environmental Protection Agency, February 2009. (Microsoft Word format and pdf)

- A. Singh and A.K. Singh; J.M. Nocerino (editor), "ProUCL Version 4.00.04 Technical Guide." U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-07/041 (NTIS PB2007-107919), February 2009. (Microsoft Word format and pdf)

- A. Singh, R. Maichle, A.K. Singh, and S.E. Lee; J.M. Nocerino (editor), "ProUCL Version 4.00.04 User Guide." U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-07/038 (NTIS PB2007-107918), February 2009. (Microsoft Word format and pdf)

- "Robust Procedures for the Identification of Multiple Outliers," A. Singh and J.M. Nocerino. A chapter in *Chemometrics in Environmental Chemistry*, J. Einay, ed., a volume (2.G, Volume 2, Part G) in *The Handbook of Environmental Chemistry*, O. Hutzinger, ed. (Heidelberg, Springer-Verlag), 1995, pp. 229-277. (pdf format)

- A. Singh; J.M. Nocerino (editor), "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. (Microsoft Word and pdf)

United States
Environmental Protection
Agency

Office of Research
and Development (8101R)
Washington, DC 20460

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE ☐; detach, or copy this cover, and return to the address in the upper left-hand corner.

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT No. G-35

Official Business
Penalty for Private Use
$300

EPA/600/R-08/038
February 2009
www.epa.gov