

Using the DATAMINE Program

This chapter serves as a user's manual for the DATAMINE program, which demonstrates the algorithms presented in this book. Each menu selection is discussed in its own section.

File / Read Data File

A text file in standard database format is read. In particular, standard-format Excel™ CSV files may be read, as well as databases produced by many common statistical and data analysis programs. The first line must specify the names of the variables in the database. The maximum length of each variable name is 15 characters. The name must start with a letter and may contain only letters, numbers, and the underscore (_) character.

Subsequent lines contain the data, one case per line. Missing data is not allowed.

Spaces, tabs, and commas may be used as delimiters for the first (variable names) and subsequent (data) lines.

Here are the first few lines from a typical data file. Six variables are present, and three cases are shown.

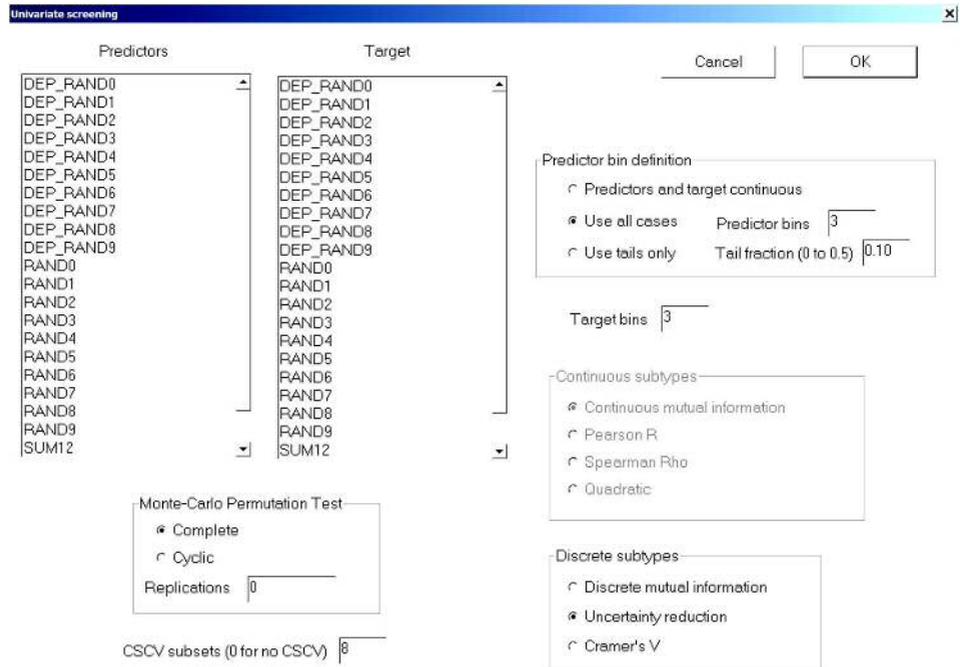
```
RAND0 RAND1 RAND2 RAND3 RAND4 RAND5
-0.82449359 0.25341070 0.30325535 -0.40908301 -0.10667177 0.73517430
-0.47731471 -0.13823473 -0.03947150 0.34984449 0.31303233 0.66533709
0.12963752 -0.42903802 0.71724504 0.97796118 -0.23133837 0.81885117
```

File / Exit

The program is terminated.

Screen / Univariate Screen

The algorithm described starting on Page 133 is used to screen a set of predictor candidates for a relationship with a single target. The following menu will appear:



The user must make the following selections and specifications:

Predictors - Select a set of predictor candidates to be tested for a relationship with a single target.

Target - Select a single target.

Predictor bin definition - Specify the nature of the predictors (and by extension, the target). The choices are:

Predictors and target continuous - All variables are to be treated as continuous.

Use all cases - All variables are treated as discrete. Continuous variables are converted to discrete bins. The user must specify the number of bins to use for the predictors.

Use tails only - The predictors are split into two bins: the tails (extreme values). The user must specify the fraction of extreme values to keep in each tail.

Target bins - If the user selected either of the discrete options above (*Use all cases* or *Use tails only*) then this specifies the number of bins into which the target variable is categorized.

Continuous subtypes - If the user selected *Predictors and target continuous* above, you specify the relationship criterion to be used. See the section beginning on Page 96.

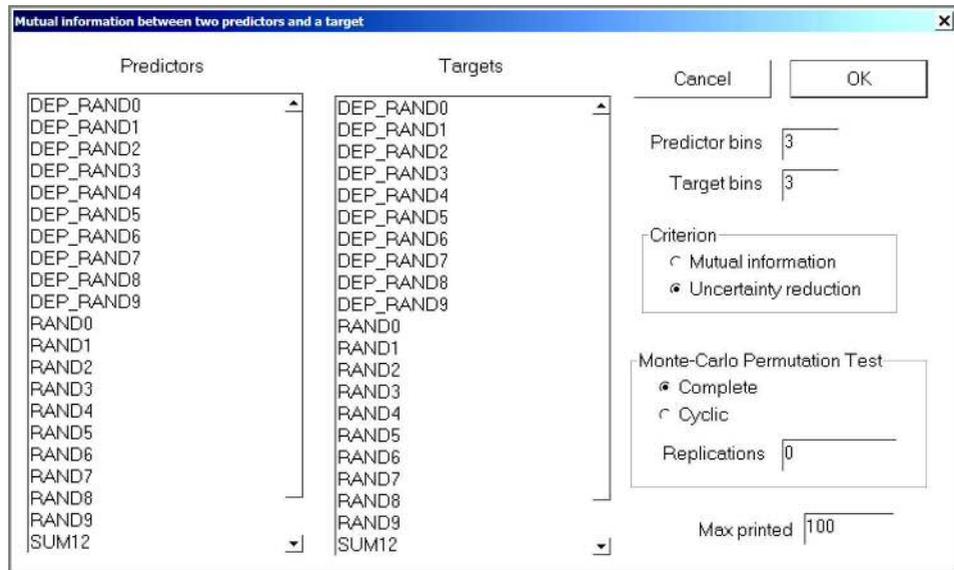
Discrete subtypes - If the user selected either of the discrete options above (*Use all cases* or *Use tails only*) then this specifies the relationship criterion to be used. See the section beginning on Page 96.

Monte-Carlo Permutation Test - A value of *Replications* greater than 1 will cause a Monte-Carlo permutation test to be performed, with this many tests run, one of which is unpermuted. The user also specifies the type of permutation, *complete* or *cyclic*. This topic is discussed starting on Page 109.

CSCV subsets - This controls performance of the CSCV test, discussed starting on Page 119.

Screen / Bivariate Screen

This section discusses bivariate screening, in which we search for relationships between one or more predictor candidates and one or more target candidates. The following menu will appear:



The user must make the following selections and specifications:

Predictors - Select a set of predictor candidates to be tested for pairwise relationships with one or more targets.

Target - Select a set of targets to be tested for a relationship with pairs of predictors.

Predictor bins - This specifies the number of bins into which the predictor variables are categorized.

Target bins - This specifies the number of bins into which the target variables are categorized.

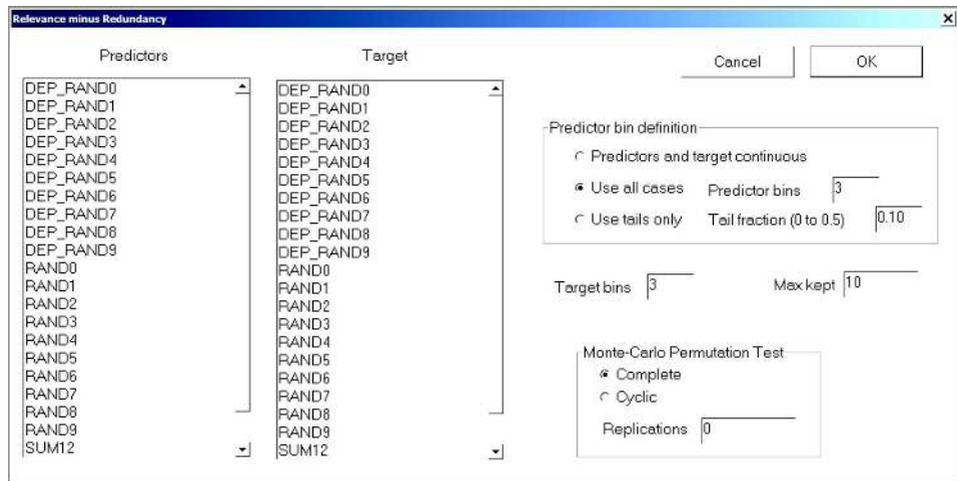
Criterion - The user chooses whether the relationship criterion is mutual information (Page 32) or uncertainty reduction (Page 77).

Monte-Carlo Permutation Test - A value of *Replications* greater than 1 will cause a Monte-Carlo permutation test to be performed, with this many tests run, one of which is unpermuted. The user also specifies the type of permutation, *complete* or *cyclic*. This topic is discussed starting on Page 109.

Max printed - If the user specifies numerous predictors and targets, the number of combinations of pairs of predictors with individual targets can be enormous. A line in the DATAMINE.LOG file is printed for each such combination, sorted from best to worst. This option lets the user limit the number of lines printed, beginning with the best.

Screen / Relevance Minus Redundancy

This section discusses relevance-minus-redundancy screening, in which we use a forward stepwise search for relationships between a set of predictor candidates and a single target variable. This algorithm was discussed on Page 148. The following menu will appear:



The user must make the following selections and specifications:

Predictors - Select a set of predictor candidates to be stepwise tested for inclusion in the set of predictors having maximum relationship with the target.

Target - Select a single target to be tested for a relationship with a set of predictors.

Predictor bin definition - Specify the nature of the predictors (and by extension, the target). The choices are:

Predictors and target continuous - All variables are to be treated as continuous.

Use all cases - All variables are treated as discrete. Continuous variables are converted to discrete bins. The user must specify the number of bins to use for the predictors.

Use tails only - The predictors are split into two bins: the tails (extreme values). The user must specify the fraction of extreme values to keep in each tail.

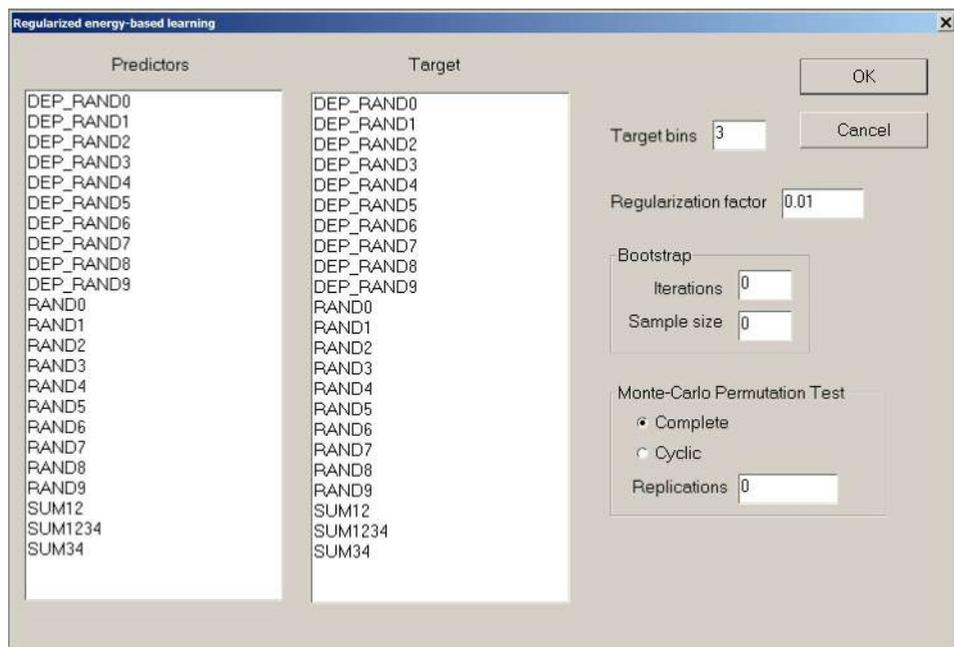
Target bins - If the user selected either of the discrete options above (*Use all cases* or *Use tails only*) then this specifies the number of bins into which the target variable is categorized.

Max kept - This is the maximum number of variables in the predictor set.

Monte-Carlo Permutation Test - A value of *Replications* greater than 1 will cause a Monte-Carlo permutation test to be performed, with this many tests run, one of which is unpermuted. The user also specifies the type of permutation, *complete* or *cyclic*. This topic is discussed starting on Page 109.

Screen / FREL

The FREL (*Feature Weighting as Regularized Energy-Based Learning*) algorithm presented starting on Page 166 is used to rank predictor candidates in terms of their relationship with a single target variable. This method is particularly useful when the data is fairly clean (noise-free) but has relatively few cases compared to the number of predictor candidates. The following menu screen appears:



The user must make the following selections and specifications:

Predictors - Select a set of predictor candidates to be ranked in terms of their relationship with the target.

Target - Select a single target to be tested for a relationship with a set of predictors.

Target bins - This specifies the number of bins into which the target variable is categorized.

Regularization factor - This controls penalization for excessively large weights in the ranking scores. It is legal and computationally harmless to set this to zero. A general discussion of this parameter appears on Page 171. Also see a more specific example of its use on Page 193.

Bootstrap iterations and **Sample size** - This is the number of bootstrap iterations to use, as well as the sample size for each. Bootstrapping is nearly always beneficial. See the discussion on Page 173 for details.

Monte-Carlo Permutation Test - A value of *Replications* greater than 1 will cause a Monte-Carlo permutation test to be performed, with this many tests run, one of which is unpermuted. The user also specifies the type of permutation, *complete* or *cyclic*. This topic is discussed starting on Page 174.

Analyze / Eigen Analysis

An eigenvalue / eigenvector analysis as described starting on Page 221 is performed. The eigenvalues and their cumulative percent of total variance are printed, along with the factor structure. A graph of the cumulative percent is displayed on the screen.

The user specifies the variables that are to take part in the analysis. If the *Nonparametric* box is checked, Spearman rho (Page 98) is used to compute the correlation matrix instead of ordinary correlation. This is useful when the data may have outliers.

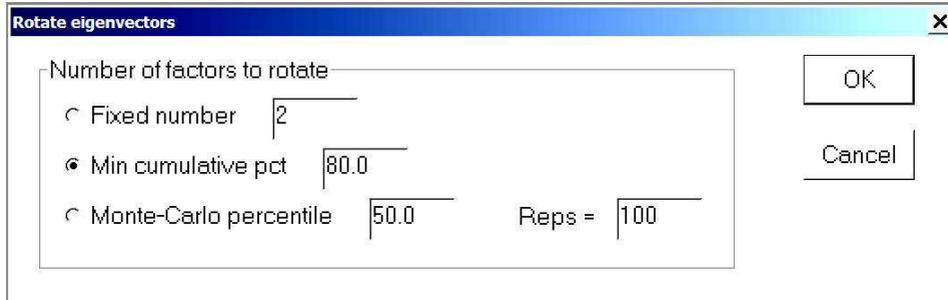
Analyze / Factor Analysis

A maximum-likelihood factor analysis as described starting on Page 255 is performed. The eigenvalues and their cumulative percent of total variance are printed first, along with the factor structure and initial Psi estimates (basic communalities). A graph of the cumulative percent is displayed on the screen. Then, the final factor analysis information is printed. Note that the *Squared length* printed at the top of each column of factor loadings is roughly analogous to the eigenvalues for an ordinary principal components analysis, but only roughly. This is because these factors are unique only up to rotation, so the natural ordering seen with the eigenvalues is no longer guaranteed.

The user specifies the variables that are to take part in the analysis. If the *Nonparametric* box is checked, Spearman rho (Page 98) is used to compute the correlation matrix instead of ordinary correlation. This is useful when the data may have outliers.

Analyze / Rotate

If the user has performed either an *Eigen analysis* or a *Factor analysis*, a varimax factor rotation (Page 231) may be performed. The following menu appears:



The user must specify the number of factors to rotate. If the starting factors are from an *Eigen analysis*, we rotate the factor loadings corresponding to the specified number of largest eigenvalues. If they are from a *Factor analysis*, fully sensible results are obtained only if the user specifies the fixed number of factors that were computed in the factor analysis.

There are three ways to specify the number of factors to be rotated:

- 1) A fixed number
- 2) Those (starting from the largest eigenvalue) that make up the specified minimum percent of total variance.
- 3) Horn's algorithm, described on Page 234, determines the number of factors to keep. In this case, the percentile and number of replications must be specified.

Analyze / Cluster Variables

The technique described starting on Page 247 is used to cluster variables. This operation may be invoked only if an *Eigen analysis* (most sensible) or *Factor analysis* (less sensible) has been performed. The user makes three specifications:

Centroid method (vs leader) - If this box is checked, the centroid method is used for updating group identifiers. Otherwise the leader method (keep the characteristics of one group) is used.

Number of factors to keep - This is the number of factors on which to base the clustering. If an *Eigen analysis* is used for this clustering (the usually recommendation), these will be the factors corresponding to the largest eigenvalues.

Start printing group membership when n reaches - The number of groups starts out at the number of variables. Each time a group is absorbed, the program can print group membership information. Obviously this can result in a huge printout if the number of variables is large. This option lets the user specify that group membership printing does not begin until this many groups remain.

Analyze / Coherence

A time-domain coherence analysis, as described on Page 295, is performed. The user specifies the variables that are to take part (which must be aligned in time) as well as the following parameters:

Connect - If this box is checked, the plotted coherence values are connected. Otherwise they are discrete vertical bars.

Nonparametric - If this box is checked, Spearman rho (Page 98) is used to compute the correlation matrix. Otherwise it is computed with ordinary correlation. This option is recommended if the data may have outliers.

Lookback window cases - This many of the most recent cases are used in the moving window for computation of coherence within the window. Longer windows result in more accurate measurements but poorer location in time.

Plot / Series

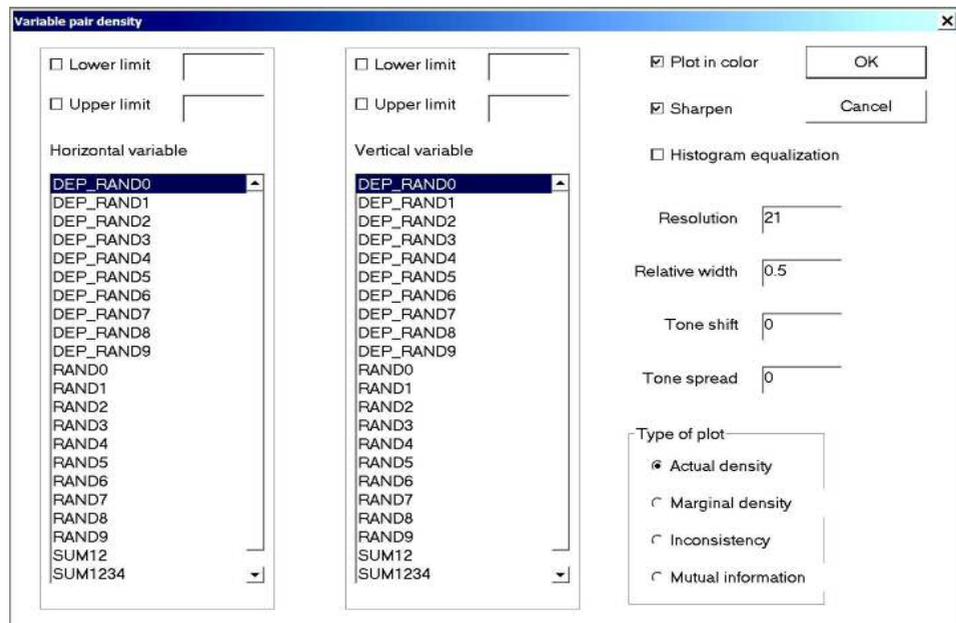
This just plots a time series of a single variable selected by the user. If the *Connected* box is checked, the plotted points are connected. Otherwise each point is represented by a discrete vertical line.

Plot / Histogram

This plots a histogram of a single variable selected by the user. The user may optionally request that the lower and/or upper bounds of the plot be limited to specified values. If this is not done, the actual plot limits are at or slightly outside the full range of the variable. The user also specifies the number of bins to use.

Plot / Density

A plot for revealing relationship anomalies, as discussed starting on Page 196, is done. The following menu appears:



The user specifies the following items:

Horizontal variable - This is the variable which will be represented by the horizontal axis. The user may optionally check the **Lower limit** and/or the **Upper limit** box above this list and specify a numeric value (values) for display limits. If a box is not checked, the corresponding limit is at or slightly outside the actual range of the variable.

Vertical variable - This specifies the variable for the vertical axis, as above.

Plot in color - If this box is checked, the plot will be in color, with yellow indicating large values of the plotted quantity, and blue indicating small values. Otherwise it is black-and-white, with black indicating large values and white indicating small values.

Sharpen - If this box is checked, areas of unusually large concentration are made to stand out from the background by accentuating them at the expense of contrast in other areas.

Histogram equalization - If this box is checked, the program applies a nonlinear transform to the data in such a way that every possible displayed tone or color occurs in the display in approximately equal quantity. The effect of this transformation is usually that small changes in the data are made more visible, while simultaneously reducing the prominence of large changes.

Resolution - This is the number of horizontal and vertical divisions at which the plot is computed. Computation time is roughly proportional to the square of this value. Larger values can reveal more detail about the relationship between the variables.

Relative width - This is the width of the Parzen smoothing window, relative to the standard deviation of each variable. Smaller values reveal more information but can also accentuate noise. If the data is noisy, large width values are appropriate to smooth out the noise.

Tone shift - This moves the overall display range. A positive value shifts the tones in the 'high' direction, and negative shifts tones toward the 'low' direction. The default of zero produces no change.

Tone spread - This expands or compresses the range of the display. The default of zero produces no change. Negative values are legal but rarely useful, as this compresses variation into a narrow range, making discrimination difficult. Positive values, rarely beyond five or so, expand the center of the display range while squashing the extremes. This emphasizes features in the interior of the grid range, at the expense of the extremes.

Actual density - This plots the actual density, as discussed on Page 200.

Marginal density - This plots the marginal density product, as discussed on Page 200.

Inconsistency - This plots the marginal inconsistency, as discussed on Page 201.

Mutual information - This plots the contribution of each region to the total mutual information, as discussed on Page 202.