SLAC-168
December 1975

FUNDIM

Program for Determination of Nonlinear Relationships in Multivariate Data

J. H. Friedman
Stanford Linear Accelerator Center
Stanford, California 94305

This note documents in minimum detail the information necessary to invoke a program that implements the Friedman-Meisel algorithm for determination of nonlinear relationships in multivariate data. The user is presumed to be completely familiar with the algorithm, as described in J. H. Friedman and W. S. Meisel, "Determination of Nonlinear Relationships in Multivariate Data," (to appear). This note is a preliminary version of a more complete one to follow, and as such, it is subject to considerable revision.

```
CALL FUNDIM(D,DP2,N,X,TREE,PERM,PARTN,CUTDIR,FITS,PRINT)
```

Input

```
INTEGER D:      dimensionality of input data
INTEGER DP2:    D + 2
INTEGER N:      cardinality of input data
Real X(D,N):    array storing input data set
LOGICAL PRINT:  print flag
  = .FALSE.:    minimum printout
  =.TRUE.:      the entire contents of the arrays TREE,PARTN,CUTDIR
                are printed.  Also controls printing of individual
                bucket information and data histograms.  (See below).
```

Output

INTEGER*2 TREE(3,TREMAX):
Binary tree representing the partitioning of the data space.

Nonterminal nodes

TREE(1,I) points to low son of Ith node
TREE(2,I) points to high son of ith node
TREE(3,I) points to vector in CUTDIR array containing partitioning direction for Ith node

Terminal nodes

-TREE(1,I) points to location in PERM array containing first point in terminal bucket

-TREE(2,I) points to location in PERM array containing last point in terminal bucket

That is, the consecutive locations

-TREE(1,I) through - TREE(2,I) in the PERM array contain the locations in the X(D,N) array of the data points in the bucket represented by the Ith terminal node.

The number of data points in the Ith terminal bucket is
= TREE(1,I) - TREE(2,I) + 1

NOTE: TREE(1,I) < 0 and TREE(2,I) < 0
is the flag for a terminal node.

TREE(3,I) points to the location in the FITS array containing the fitted information for the Ith terminal bucket.

INTEGER*2 PERM(N):
permutation array for the data vectors. Pointers to data points in the same terminal bucket are stored in consecutive locations in the PERM array.

REAL PARTN(TREMAX):
PARTN(I) contains the split point for the Ith nonterminal node. Locations corresponding to terminal nodes are not used.

REAL CUTDIR(D,TREMAX):
CUTDIR(,I) contains the partitioning direction for the Ith nonterminal node.

REAL FITS(D,DP2,TREMAX):
FITS(, , I) contains the fitted information for the Ith terminal node.

FITS(,1-D,I) contain the principal component vectors in descending order of variance.

FITS(,D+1,I) contain the eigenvalues (variances) corresponding to the principal components.

FITS(,D+2,I) contains the mean value of the data points in the bucket.

Labled Commons (parameters)

```
COMMON/TREBND/TREMAX
INTEGER TREMAX      (input)
```

Maximum number of nodes (nonterminal and terminal) allowed in partitioning tree. Equal to twice the maximum number of terminal nodes (buckets). TREMAX must be less than or equal to the corresponding dimensioned quantity in the arrays TREE, PARTN, CUTDIR, and FITS. If TREMAX is not large enough, the program will terminate with the message "KDTREE STORAGE EXCEEDED".

```
COMMON/SCRTCH/Z(N)
```

Z(N) is a scratch array used for intermediate storage in FUNDIM.

```
COMMON/POISPT/CUTOFF,POISON(N)
      Real CUTOFF      (input)
      LOGICAL*1 POISON      (output)
```

CUTOFF = generalized (Molhalanobis) distance squared from sample center at which a data point is considered to be inconsistent. During measurement space partitioning these points are poisoned so that they will not be used in further partitioning calculations.

POISON(N)=logical array that flags points that are poisoned.

POISON(I) = .TRUE.: X(,I) was poisoned
POISON(I) = .FALSE.: X(,I) was <u>not</u> poisoned during data space partitioning.

```
COMMON/CUTOFF/MINBUC
INTEGER MINBUC      (input)
```

MINBUC = minimum number of data points comprising a terminal bucket. (Default value = 3*D).

```
COMMON/INTERV/NVAL
INTEGER NVAL      (Input)
```

NVAL = number of evaluation points in search of best split point along best splitting direction.
NVAL = 0: split point = median (Default value = 0).

```
COMMON/EITMIN/EFACT
REAL EFACT      (input)
```

EFACT = ratio of eigenvalue to <u>largest</u> eigenvalue of covariance matrix at which the covariance matrix is considered to be singular and its rank reduced. (Default value = 1.0E-05)

```
COMMON/DIMTER/VARFAC
REAL VARFAC     (input)
```

1-VARFAC = fraction of trace of covariance matrix used for determination of local intrinsic dimensionality. The intrinsic dimensionality of a collection of data points is taken to be the number of the largest eigenvalues of their covariance matrix required to exceed 1-VARFAC times the trace. (Default value, VARFAC=.01).

```
COMMON/TRMINF/TRMFAC,MAXTRM
REAL TRMFAC     (input)
INTEGER MAXTRM     (input)
```

MAXTRM = maximum number of data points trimmed at any one time during the "robust" covariance estimation. MAXTRIM must be $\leq$ 200. (Default = 100).

TRMFAC = fraction of data trimmed at any one time during "robust" covariance estimation. (Default = 0.1).

The actual number of data points trimmed is the minimum of TRMFAC * number of data points in sample, and MAXTRM.

```
COMMON/DUMP/DODUMP
LOGICAL DODUMP     (input)
```

DODUMP = logical flag that controls printing of complete history of data measurement space partitioning.

DODUMP = .TRUE.: print complete history of data space partitioning

DODUMP = .FALSE.: no output of partitioning history. (Default value = .FALSE.).

NOTE: This output flag operates independently of the print flag (PRINT) that appears in the calling sequence of FUNDIM.

```
COMMON/HSTFLG/HFLAG
INTEGER HFLAG     (input)
```

HFLAG = flag that controls histogramming of data projections.

HFLAG = 0: no projection histograms

HFLAG = 1: histograms of data as projected on each of the coordinate axes.

a) total data sample
b) each terminal bucket subsample individually

HFLAG = 2: histograms of data as projected on each principal axis of data sample.

a) total data sample
b) each terminal bucket subsample individually

In all histograms, unpoisoned points are indicated by the symbol 'x' while poisoned points are indicated by the symbols '*' superimposed.

NOTE: HFLAG operates in conjunction with the print flag (PRINT) that appears in the calling sequence of FUNDIM. If PRINT = .FALSE., then no histograms will be printed regardless of the value of HFLAG.

Default values

The default values indicated for some of the input parameters appearing in the labeled commons are set in a BLOCK DATA subprogram internal to FUNDIM. They may be overridden with a user provided value, by declaring the appropriate labeled common and entering the value via an executable statement before calling FUNDIM. Those input parameters and storage arrays for which there is no default provided, must be entered by the user before invoking FUNDIM.

DIST = VALUE (D,DP2,Y,TREE,PARTN,CUTDIR,FITS)

Input

Y(D) = coordinates of data point to be tested for inconsistency

All other parameters are the resulting output arrays from the invocation of FUNDIM and are described above.

Output

DIST = generalized (Malhalanobis) distance squared of data point from the center ("robustized" sample mean) of the terminal bucket in which it lies.

The decision to flag the data point as inconsistent is made as follows:

If DIST > DSTMAX then <inconsistent>
else <consistent>

where the value of DSTMAX is chosen by the user.