(revised 4/2016)

How to Use EXTREE: A Program for Fitting the Extended Tree Model

James E. Corter

Teachers College, Columbia University

CONTENTS:

I. Description of input file format

II. Example input and output files

III. Installing EXTREE on your system

IV. References

AUTHOR'S ADDRESS:

Box 41, Teachers College

Columbia University

New York, NY 10027

(BITNET: jecorter@cutcv2)

Please note that the ADDTREE/P and EXTREE programs are the

property of James Corter and Amos Tversky. The programs are freely

available to be used for research or academic purposes. However, the

programs should not be modified and re-distributed, or resold, or used

for any business or commercial purpose without the express written

consent of the author, James Corter.

(EXTREE copyright (c) 1982, 1991 by James E. Corter.)

=================================================================

I. Description of Input File Format

Input files should contain the following lines:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

(1) first line: a list of analysis option commands, zero or more

of the following keywords. Note that lowercase letters may be

used, and that all keywords may be abbreviated to 3 characters.

DATA Requests printing of the raw data matrix.

TRANSFORMED Requests printing of the data, after transformation

into distance-like numbers.

MODELDISTANCES Requests printing of the model (tree) distances.

RESIDUALS Requests printing of the residuals matrix.

HEIGHTS Requests printing of the distance of each node from the

root.

NOTREE Suppresses printing of the tree graph.

NONUMBER Suppresses numbering of objects in the tree graph.

SPECIFYORDER If this is included, line 1 of the input file must

be followed by a line containing a list of object numbers. The

leaves of the tree graph are then ordered to conform as closely as

possible to this list. (NOTE: Care must be taken not to omit or

duplicate numbers in this list. Also, note that not all orderings

of the leaves are possible, since the tree structure imposes

constraints on the possible orderings.)

NOSUBTRACTCONSTANT The default assumption is interval-scale data;

this implies complete freedom in choosing an additive constant.

Therefore the primary approach is to either add OR subtract an

additive constant to exactly satisfy the triangle inequality:

d(i,j) + d(i,k) <= d(j,k) for all i,j,k, AND d(i,j) + d(i,k) =

d(j,k) for some i,j,k. But if "nosubtractconstant" is specified,

strict inequality is allowed, i.e. if d(i,j) + d(i,k) > d(j,k)

holds for all i,j,k in the data, no constant is subtracted.

MINVARROOT The default rooting is to combine the last few

remaining clusters into the root node; if "minvarroot" is

specified, the program will search for the root that minimizes the

variance of the distances from the root to the leaves.

ROOTAT Specify where the root is to be placed. If this option is

requested, this keyword must be followed by two node numbers, e.g.

"rootat 12 13 ". Note that the two nodes specified must be

contiguous in the tree structure. A previous run will generally be

necessary to determine the correct node numbers.

LINESIZE Sets the length of lines in the output file. The current

default value is 80 characters (so that output may be conveniently

read on a video terminal). Example: "linesize 66 ".

NOPATTERNMATRIX Suppresses the marked feature pattern matrix

which otherwise is printed with the extended tree.

TREE Allows the user to specify the tree structure. This is a

tricky option to use, since it requires knowledge of how the

program assigns numbers to the higher nodes. The higher nodes are

numbered starting with n+1, where n is the number of objects. The

numbers specifying the tree structure should follow the list of

object names (which aren't optional in this case). "Example Two"

below illustrates this option.

MARKED Allows the user to specify the marked features to be added

to the tree structure, rather than allowing the program to choose

them. The features are specified by listing the host nodes for

each of the features, with each feature on a different line. E.g.

a line reading "1 3 5" specifies a marked feature common to objects

1, 3, and 5. These lines must come at the very end of the file,

following the object names and the tree structure lines if they are

also specified. They must be preceded by a line of asterisks

'\*\*\*\*\*' to set them off from the tree features (this helps to

prevent mistakes). Again, see "Example Two" below.

THRESHOLD Allows the user to set the minimum size of marked

segment to be used. I.e. if the estimate of a marked segment is

less than this value, the feature will be eliminated and the other

parameters of the model re-estimated without it. The default value

is 1/50th of the maximium dissimilarity. Example: "thresh 0.23 ".

TTHRESHOLD Sets the minimum size of tree arcs to be used; i.e.

arcs of length less than this value are eliminated from the model.

The default value is the value of threshold. Ex: "tthresh 0.13 ".

RHO Sets a parameter that controls how pairwise marked segments

are collapsed into higher-order "cliques" of marked segments (i.e.

overlapping features corresponding to marked segments on more than

2 branches). The value of this parameter is defined in relative

terms; that is, it can vary from 0.0 to 1.0 (default value is 0.5).

A value of 0.5 means that a clique of pairwise marked segments

(e.g. a pairwise marked feature shared by objects i and j, another

shared by objects i and k, and a third shared by objects j and k)

will be collapsed into a single feature represented by marked

segments on the three arcs corresponding to objects i,j, and k, as

long as the marked segment with minimum initial length estimate is

at least .50 of the maximium-length segment. Example: "rho 0.67 ".

PARMS Sets the number of marked features to be tried. The actual

number of marked features used in the final solution may be

somewhat less than this if some features are eliminated as being

less than the threshold. It is not recommended to use more than n

marked features, where n is the number of objects. The default

value of this parameter is one-half n.

Note that any of the keywords mentioned above may be abbreviated to

three (or more) characters; also, capital or lower-case letters are

ok. Keywords may be separated by either blanks or commas.

However, when a keyword above specifies that a parameter is to be

read in following the keyword (e.g. ("rootat 12 13 "), it is best

to follow the numerical parameter with a <space> (" ") instead of

a <comma> (",") since some PASCALs will object to finding a comma

while attempting to read a number.

If the first line is blank, the standard analysis is done

and the default output is obtained (estimates, tree graph, and

statistics).

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

(1.5) (optional line: see SPECIFYORDER above)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

(2) second line: a comment line for labeling of the output (up to

80 characters in length).

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

(3) third line: the following parameters (separated by commas or

blanks):

<number of objects> (integer)

SIMILARITIES or DISSIMILARITIES (specifies similarity or

distance (dissimilarities) data)

FULL or LOWERHALF (specifies shape of matrix)

<fieldsize> (integer; width of OUTPUT data fields)

<number of digits after the decimal point> (integer; number of

digits after the decimal place in the OUTPUT matrices)

NOTE: the last two parameters here refer only to the OUTPUT. Also

note that you should try to leave at least two spaces between

numbers: e.g. if your data is numbers between 0 and 99, and you

want to see one digit after the decimal place, you should specify

6 and 1 for these parameters. This is to leave room for possible

minus signs and the decimal point itself.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

(4) fourth, etc. lines: the data matrix itself

NOTE: some Pascals require real format, that is, the data must have

a decimal point, and digits both before and after (e.g. 45.0, not

45 or 45. or .45) NOTE: the data matrix can be in a variety of

formats, as long as the order of entries corresponds to the order

of entries in a (full or lowerhalf) matrix. So for example, the

input data file might contain only one number to a line, as long as

the order of the lines corresponds to the order of entries in a

lowerhalf (or full) matrix. However, one thing that must NOT occur

is trailing blanks on a line, since the program then expects

another number on that line.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

(5) (optional) lines immediately following the data: a list of

object names, one to a line, maximum length 20 characters.

II. Example Input and Output Files

INPUT FILE EXAMPLE ONE:

nonumber,residuals,parms 6

numbers: abstract concept (Shepard, Kilpatric, & Cunningham; 1975)

10 dis low 7 1

421

584 284

709 346 354

684 646 059 413

804 588 671 429 409

788 758 421 300 388 396

909 630 796 592 742 400 417

821 791 367 804 246 671 350 400

850 625 808 263 683 592 296 459 392

zero

one

two

three

four

five

six

seven

eight

nine

OUTPUT FILE EXAMPLE ONE:

extree analysis (EXTREE version 1.5):

numbers: abstract concept (Shepard, Kilpatric, & Cunningham; 1975)

( 0.0 needed for positivity of distances )

303.0 added to exactly satisfy triangle inequality

node length children label

1 454.6 zero

2 269.4 one

3 185.6 two

4 235.6 three

5 176.4 four

6 327.4 five

7 264.1 six

8 375.6 seven

9 325.3 eight

10 330.4 nine

11 180.6 1 2

12 103.2 3 5

13 53.9 4 10

14 81.8 6 8

15 97.5 12 9

16 37.9 13 7

17 3.4 11 15

18 74.0 16 14

19 0.0 17 18

------------------------------------------- zero

-----------------|

| -------------------------- one

|

| ------------------ two

| ----------|

|--------| ----------------- four

| |

| -------------------------------- eight

|

| ----------------------- three

| -----|

| ----| -------------------------------- nine

| | |

------| -------------------------- six

|

| -------------------------------- five

--------|

------------------------------------ seven

stress formula 1 = 0.0975

stress formula 2 = 0.5220

r(monotonic) squared=0.7275

r-squared (p.v.a.f.)=0.6249

extree analysis:

6 marked features will be tried

features smaller than 24.2 will be eliminated

i j estimate [ set(i) ] [ set(j) ]

- - -------- -----------------------

3 11 191.6 [ 3 ] [ 1 2 ]

8 9 185.4 [ 8 ] [ 9 ]

2 3 183.2 [ 2 ] [ 3 ]

10 9 172.4 [ 10 ] [ 9 ]

12 11 150.8 [ 3 5 ] [ 1 2 ]

2 4 134.8 [ 2 ] [ 4 ]

5 6 118.2 [ 5 ] [ 6 ]

3 4 113.3 [ 3 ] [ 4 ]

7 15 112.7 [ 7 ] [ 3 5 9 ]

4 12 107.5 [ 4 ] [ 3 5 ]

2 13 103.8 [ 2 ] [ 4 10 ]

4 11 101.6 [ 4 ] [ 1 2 ]

10 8 94.9 [ 10 ] [ 8 ]

9 7 90.4 [ 9 ] [ 7 ]

13 11 84.5 [ 4 10 ] [ 1 2 ]

2 12 78.5 [ 2 ] [ 3 5 ]

9 14 76.3 [ 9 ] [ 6 8 ]

5 9 63.9 [ 5 ] [ 9 ]

10 7 56.4 [ 10 ] [ 7 ]

1 12 58.6 [ 1 ] [ 3 5 ]

12 7 52.4 [ 3 5 ] [ 7 ]

1 3 51.0 [ 1 ] [ 3 ]

1 15 50.0 [ 1 ] [ 3 5 9 ]

6 12 44.2 [ 6 ] [ 3 5 ]

5 4 43.8 [ 5 ] [ 4 ]

5 7 39.3 [ 5 ] [ 7 ]

2 14 37.5 [ 2 ] [ 6 8 ]

3 7 36.2 [ 3 ] [ 7 ]

2 8 33.1 [ 2 ] [ 8 ]

10 14 31.6 [ 10 ] [ 6 8 ]

checking for cliques & redundant patterns of marked features

clique = 4 11 3

clique = 4 11 12

clique = 8 9 10

clique 1 ave. weight = 135.5

clique 2 ave. weight = 120.0

clique 3 ave. weight = 150.9

feature C

4 ( 4 )

11 ( 1 2 )

3 ( 3 )

feature D

4 ( 4 )

11 ( 1 2 )

12 ( 3 5 )

feature E

8 ( 8 )

9 ( 9 )

10 ( 10 )

feature H

2 ( 2 )

4 ( 4 )

feature I

5 ( 5 )

6 ( 6 )

feature N

7 ( 7 )

15 ( 3 5 9 )

iteration: 1

spillover of marked features on arcs:

features smaller than threshold: 17 23

maximum leaf spillover= 0.0

iteration: 2

spillover of marked features on arcs:

features smaller than threshold:

maximum leaf spillover= 0.0

node length children label

1 454.6 zero

2 269.4 one

3 205.2 two

4 283.3 three

5 156.8 four

6 308.2 five

7 253.3 six

8 394.8 seven

9 295.5 eight

10 282.7 nine

11 241.7 1 2

12 133.0 3 5

13 64.7 4 10

14 50.4 6 8

15 210.1 12 9

16 108.3 13 7

17 2.5 11 15

18 54.3 16 14

19 0.0 17 18

20 86.0 4 11 3 "C"

21 94.7 4 11 12 "D"

22 127.8 8 9 10 "E"

23 0.0 2 4 "H"

24 101.6 5 6 "I"

25 123.5 7 15 "N"

marked feature

pattern matrix

--------------

-------------------------------- zero C D . . .

CCCCCCDDDDDDD---| . . . . .

| -------------------- one C D . . .

| . . . . .

| CCCCCC--------- two C D . . N

| DDDDDDD--| . . . . .

|NNNNNNNN-----| IIIIIII----- four . D . I N

| | . . . . .

| EEEEEEEEE------------ eight . . E . N

| . . . . .

| CCCCCCDDDDDDD-------- three C D . . .

| ----| . . . . .

| --------| EEEEEEEEE------------ nine . . E . .

| | | . . . . .

---| NNNNNNNNN--------- six . . . . N

| . . . . .

| IIIIIII--------------- five . . . I .

----| . . . . .

EEEEEEEEE------------------- seven . . E . .

stress formula 1 = 0.0603

stress formula 2 = 0.2708

r(monotonic) squared=0.9267

r-squared (p.v.a.f.)=0.9013

final set of marked features:

feature objects sharing feature

------- -----------------------

C [ three, zero, one, two, ]

D [ three, zero, one, two, four, ]

E [ seven, eight, nine, ]

I [ four, five, ]

N [ six, two, four, eight, ]

residual distances:

-0.0

3.9 -111.0

164.0 -13.9 -43.0

-19.7 127.4 -0.0 -107.6

-4.6 -35.5 10.4 -82.9 0.0

-23.6 131.5 4.4 1.7 19.7 -21.1

13.8 -80.1 48.8 -6.5 43.1 -0.0 -86.8

-77.9 77.3 36.3 88.2 -36.3 53.0 -24.0 -49.0

-55.9 -95.7 50.2 0.0 -26.5 80.6 -1.7 116.6 -67.6

INPUT FILE EXAMPLE TWO:

nonumber,threshold 0.0 ,tree,marked

numbers: abstract concept (Shepard, Kilpatric, & Cunningham; 1975)

10 dis low 7 1

421

584 284

709 346 354

684 646 059 413

804 588 671 429 409

788 758 421 300 388 396

909 630 796 592 742 400 417

821 791 367 804 246 671 350 400

850 625 808 263 683 592 296 459 392

zero

one

two

three

four

five

six

seven

eight

nine

1 2

11 3

12 4

13 5

14 6

15 7

16 8

17 9

18 10

\*\*\*\*\*\*\*\*

2 4 6 8 10

3 5 7 9

OUTPUT FILE EXAMPLE TWO:

extree analysis (EXTREE version 1.5):

numbers: abstract concept (Shepard, Kilpatric, & Cunningham; 1975)

( 0.0 needed for positivity of distances )

303.0 added to exactly satisfy triangle inequality

user-specified tree structure

user-specified marked features

node length children label

1 454.6 zero

2 269.4 one

3 261.5 two

4 286.0 three

5 298.2 four

6 384.4 five

7 261.0 six

8 395.0 seven

9 339.6 eight

10 355.4 nine

11 113.5 1 2

12 82.5 11 3

13 18.8 12 4

14 74.9 13 5

15 70.8 14 6

16 39.7 15 7

17 0.0 16 8

18 0.0 9 10

19 0.0 17 18

-------------------------------- zero

--------|

------| -------------------- one

-| |

------|| ------------------- two

-----| ||

--| | |--------------------- three

| | | |

| | | --------------------- four

| | |

| | ---------------------------- five

| |

| -------------------- six

|

|--------------------------- seven

|

|----------------------- eight

|

------------------------- nine

stress formula 1 = 0.0892

stress formula 2 = 0.4834

r(monotonic) squared=0.7663

r-squared (p.v.a.f.)=0.6093

extree analysis:

features smaller than 0.0 will be eliminated

feature C

2 ( 2 )

4 ( 4 )

6 ( 6 )

8 ( 8 )

10 ( 10 )

feature D

3 ( 3 )

5 ( 5 )

7 ( 7 )

9 ( 9 )

iteration: 1

spillover of marked features on arcs:

features smaller than threshold:

maximum leaf spillover= 0.0

node length children label

1 421.7 zero

2 302.3 one

3 298.0 two

4 308.3 three

5 349.4 four

6 404.6 five

7 310.7 six

8 414.3 seven

9 355.1 eight

10 339.9 nine

11 77.0 1 2

12 116.2 11 3

13 10.4 12 4

14 86.4 13 5

15 76.7 14 6

16 38.6 15 7

17 0.0 16 8

18 36.4 9 10

19 0.0 17 18

20 65.7 2 4 6 8 10 "C"

21 129.0 3 5 7 9 "D"

marked feature

pattern matrix

--------------

---------------------------- zero . .

-----| . .

--------| CCCCC---------------- one C .

| | . .

------| DDDDDDDDD----------- two . D

-----| | . .

--| | |CCCC---------------- three C .

| | | | . .

| | | DDDDDDDDD-------------- four . D

| | | . .

| | CCCCC---------------------- five C .

| | . .

| DDDDDDDDD------------ six . D

| . .

|CCCC----------------------- seven C .

| . .

| DDDDDDDDD--------------- eight . D

--| . .

CCCCC------------------ nine C .

stress formula 1 = 0.0745

stress formula 2 = 0.3590

r(monotonic) squared=0.8711

r-squared (p.v.a.f.)=0.7769

final set of marked features:

feature objects sharing feature

------- -----------------------

C [ one, three, five, seven, nine, ]

D [ two, four, six, eight, ]

III. Installing EXTREE on your System

The ADDTREE/P and EXTREE programs included have been

successfully complied and tested using the DEC VMS operating system

PASCAL compiler, and on an IBM-PC using Turbo PASCAL 6.0. Minor

modifications (noted in the source code) are necessary for use with

Turbo PASCAL. Previous versions of the programs have been

successfully compiled on UNIX systems and on IBM mainframes (using

the IBM interactive PASCAL compiler). Since the PASCAL language as

originally defined does not contain specifications for certain I/O

and file definition operations, statements pertaining to these

operations may vary from compiler to compiler.

For IBM compilers, the PARM='MARGINS(1,80)' option should be

used, since source code appears in 80 columns in the supplied

source code file.

Please notify James Corter of problems or comments regarding

portability.

IV. References

Corter, J.E. (1982). ADDTREE/P: A PASCAL program for fitting

additive trees based on Sattath & Tversky's ADDTREE program.

Behavior Research Methods and Instrumentation, 14, 353-354.

Corter, J.E., & Tversky, A. (1986). Extended similarity trees.

Psychometrika, 51, 429-451.

Sattath, S., & Tversky, A. (1977). Additive similarity trees.

Psychometrika, 42, 319-345.