HOW TO USE SURVIVAL FORESTS (SFPDV1)

This program in f77 is an unorthodox approach to survival analysis. it allows the analyst to view the importance of the covariates as the experiment evolves over time. It is model-free. Its basis is an accurate method to estimate (nonparametrically) individual survival curves $\hat{S}(t,\mathbf{x})$ for any set of covariates $\mathbf{x}$. in the data base. This is done using a survival forest, details of which will appear in a later paper in Statistical Science.

It has been tested on a variety of simulated data sets where the true $S(t,\mathbf{x})$ are known and the accuracy of the estimates were validated. some of these had time varying covariate strengths and the output from SF tracked these in a reasonable approximation.

Here, the purpose is to instruct the user on how to set up and get output from SF. First of all, categorical values must be declared. This is done by specifying cat(m), for every covariate m. (see the read in of vet-lung data) Cat(m)=1 indicates that the mth covariate is numerical, cat(m)=J>1 indicates that the mth variable is categorical with J values. If the mth variable is categorical with J values, these must be coded as 1,2, ..., J. This recoding can be done at read time (see the read for the vet-lung data set).

The user must also supply names for the output files. This is done immediately following the array dimensionality specifications.

Now look at the parameter statement:

```
      parameter(ns=136,mdim=6,jbt=100,
     1  ndsize=1,nrnodes=2*ns+1,ntsm=100,
     1  look=10,itime=1,
     1   isurv=1,indsurv=0,
     1  icor=0, ireg=0,
     1  iscale=0,mdimsc=3)
```

ns:: the sample size of the data set.

mdim: the number of covariates.

jbt: the number of trees to be used in the forest. Use at least 100, for accuracy, 200 is better.

ntsm: he number of time points at which the survival curves are evaluated. These are equally spaced order statistics from the non-censored death times

look: there is a method in SF that gives test set estimates $\hat{t}(n)$of times of death. $t(n)$. The accuracy of these estimates is measured by $err=av(|t(n)-\hat{t}(n)|/\hat{t}(n))$ where the average is over non-censored times of death. If look=10, then at every tenth tree, err is computed and what is outputted is 100*(1-err), the larger this number, the smaller the error. It is a measure of the information in the data set.

itime: there are two parameter settings in SF. Setting itime=1 will give the most accuracy to estimates of the survival curves over the time domain,. Setting itime=1 with give the most accurate estimate of the cross-section of the survival curves at a given time point. Use itime=1 only if isurv=1 and possibly if indsurv=1.

isurv: isurv=1 gives a two column output of length ntsm. The first column at the selected ntsm time points. The second is the average survival curve at that time point.

indsurv: indsurv=1 sends a lot of information to a file (name specified by user). The file has ns rows--one for each case in the data. In each row, the first ntsm entries are the values of the survival curve for that individual. The next variable is the censoring indicator--zero if censored, one otherwise. Next is the time of death or censoring. The next mdim variables are the covariate values.

The next two switches are at the heart of the analysis. The Cox model assumes that the effects of the covariates are constant over time. Other models assume a fixed form for the time dependence. Given the estimates $\hat{S}(t,\mathbf{x})$we can explore the evolution over time of the effects of the covariates.

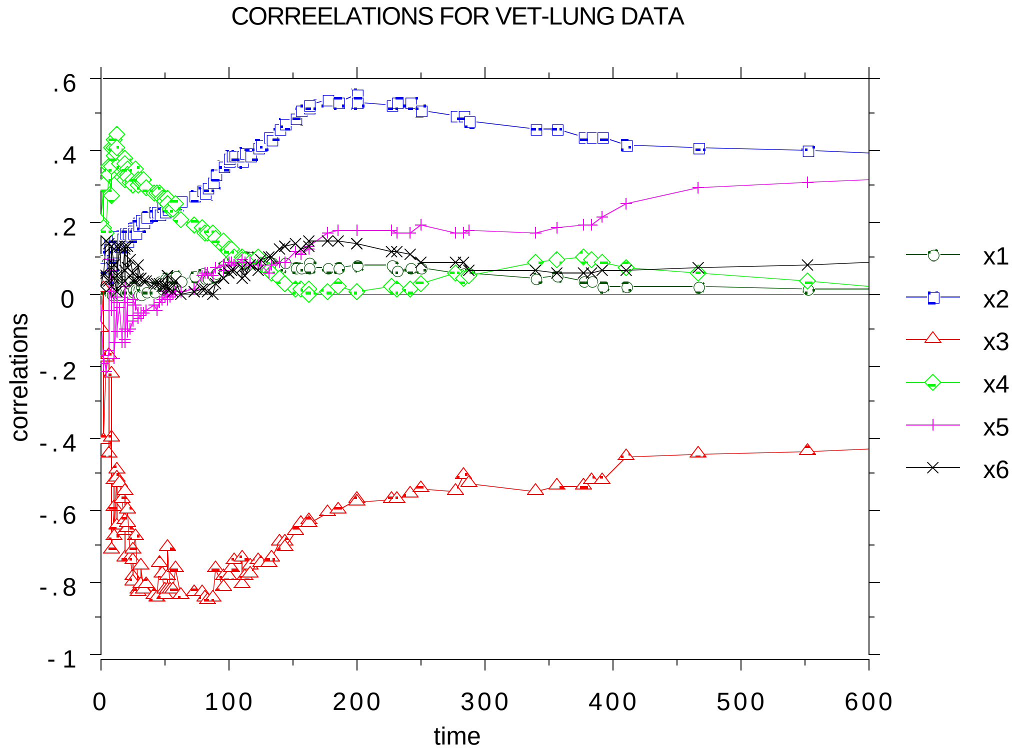
The data set used for illustration is the vet-lung data with 136 cases, 6 covariates and 7% censoring, The variables are are:

```
#          Treatment   1=standard,   2=test
#         Celltype    1=squamous,   2=smallcell,   3=adeno,   4=large
#         Survival in days
#         Status      1=dead, 0=censored
#         Karnofsky score
#         Months from Diagnosis
#        Age in years
#         Prior therapy   0=no, 10=yes
```

icor: this switch has the values 0,1,2. If 1, then at each of the ntsm time points it computes the correlation between log($\hat{S}(t,\mathbf{x}_n)$) and each of the covariates. Setting icor =2 does the following: at each time point, that monotone transformation of each covariate is found which maximises the correlation with log($\hat{S}(t,\mathbf{x}_n)$)

)

The output from icor consists first of a line which gives the average values of the correlations over time with the covariates. Then there is a line space followed by a matrix with mdim+1 columns and ntsm deep. The first column are the values of the ntsm time points in ascending order. The other mdim columns are the values of the correlations corresponding to the time point. Here is a graph of the correlations vs. time for the data set vet-lung for icor=1.



There are two dominant variables with some change in time. The effect of x3 is weakening, and x2 has a bulge around time 200.;

NOTE: The noise near t=0 should not be taken seriously. This is the region where the estimates $\log(\hat{S}(t,\mathbf{x}_n))$ are the noisiest.

The time average of the correlations over the 6 covariates are:

.043 -.001 -.637 .316 -.044 .071

The graph for icor=2 looks similar. The time averaged correlations are

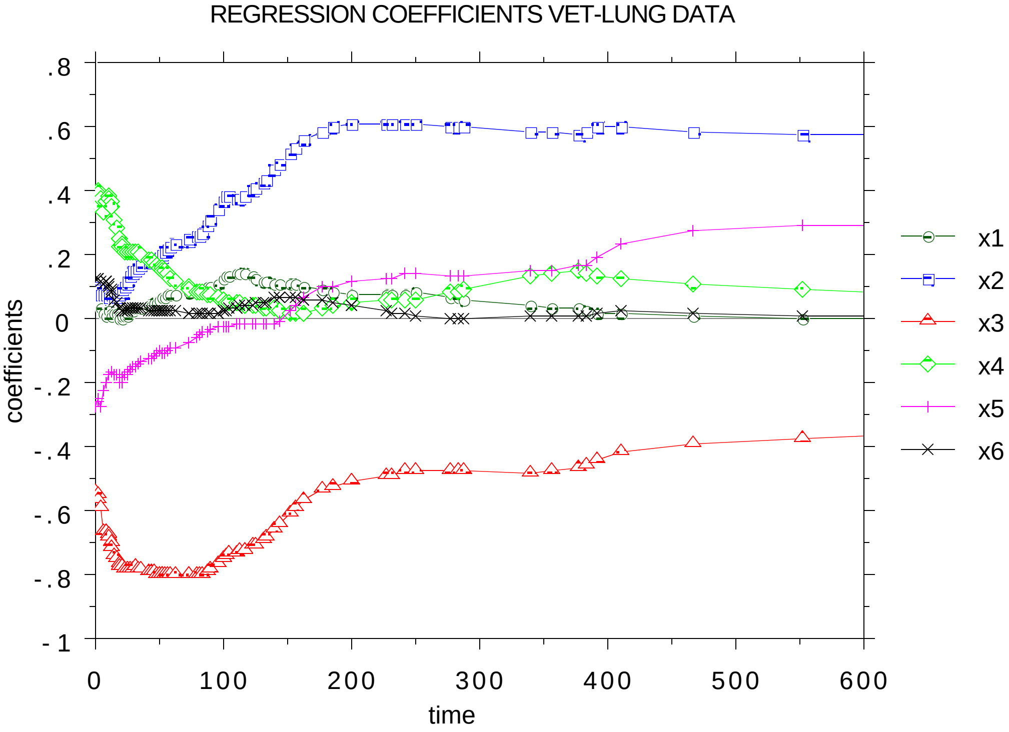
.146 .038 -.648 .406 -.106 .170

Note that the average correlations of x1,x4,x5,x6 have significantly increased--suggesting a nonlinear relationship with $\log(\hat{S}(t,\mathbf{x}_n))$.

ireg This switch also has the values 0,1,2. At ireg=1, ordinary LS linear regression is done using the covariates with response $\log(\hat{S}(t,\mathbf{x}_n))$ normalized to mean zero and sd one. With ireg=2 a monotone transformation is found for each covariate at each time point so as the minimize the RSS in fitting the normalized $\log(\hat{S}(t,\mathbf{x}_n))$.

The output file consists of a header line which gives the average value of rsq and the average coefficient for each covariate, Then there is a blank line and a matrix with mdim+2 columns and ntsm rows. The first column are the time points in ascending order. The next column is the value of rsq for the regression at hat time point. The next mdim columns are the regression coefficients for each covariate.

Here is a graph of the regression coefficients as a function of time using ireg=1.:



There appears to be a definite time trend in the dominant covariates x2 and x3.

The average rsq and the average values of the coefficients are:

.750 . 060 . 296 -.680 157 -.054 .042

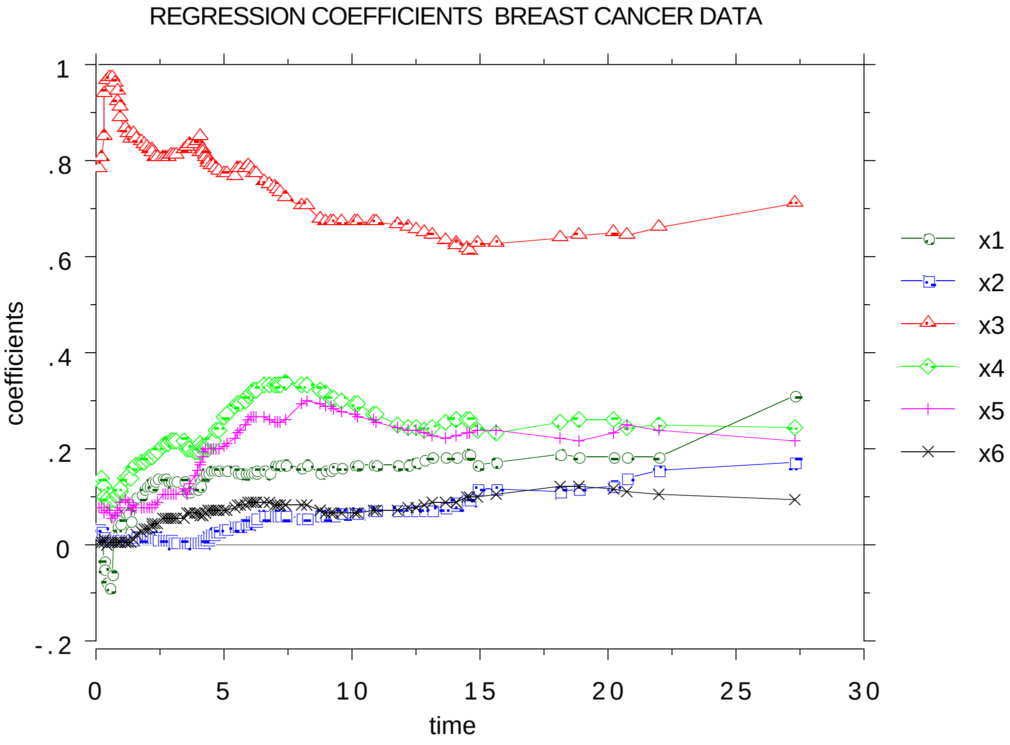
Using ireg=2 leads to similar results.

The vet-lung data set is small with little censoring. Graphs for icor=2 for three other data sets show the diversity of results.

i) a breast cancer data set sent to me from England. It has 272 cases, 6 covariates and 17% censoring.

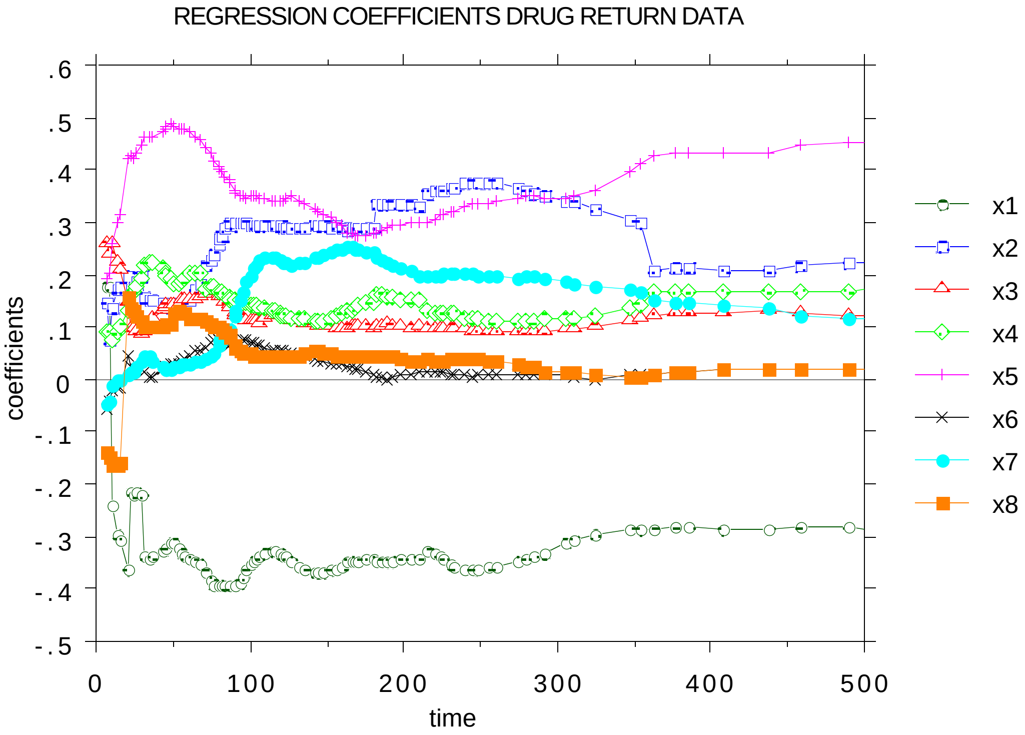
ii) a return to drugs data set with 575 cases, 8 covariates and 19% censoring. (see the book "Applied Survival Analysis" by Hosmer and Lemeshow)

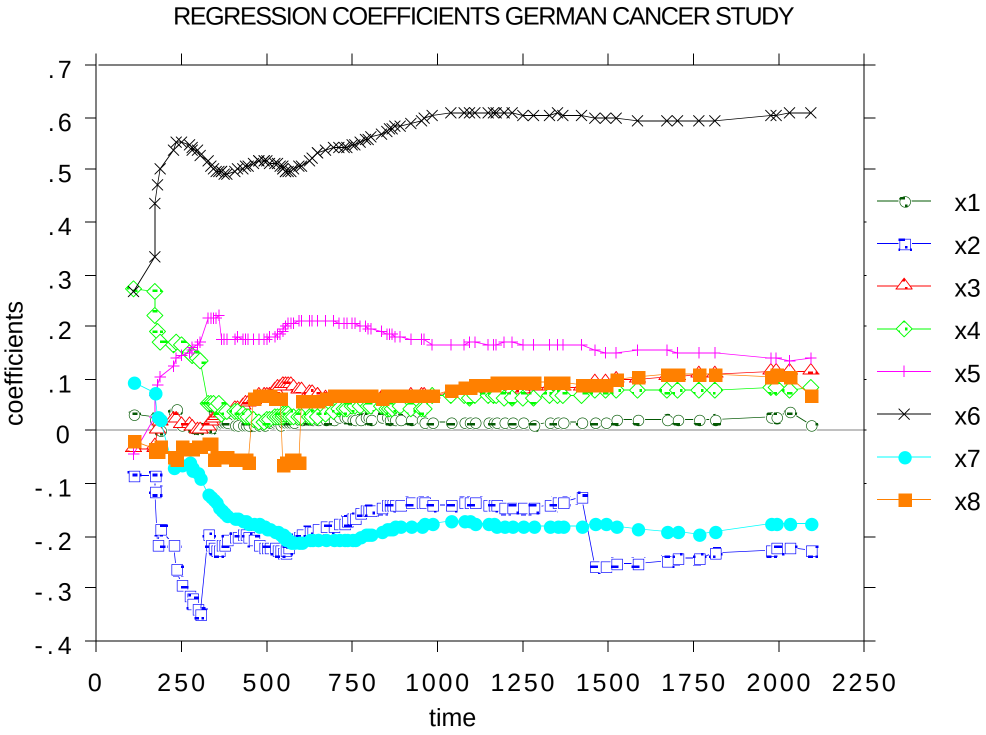
iii) the German cancer study with 686 cases, 7 covariates, and 56% censoring.



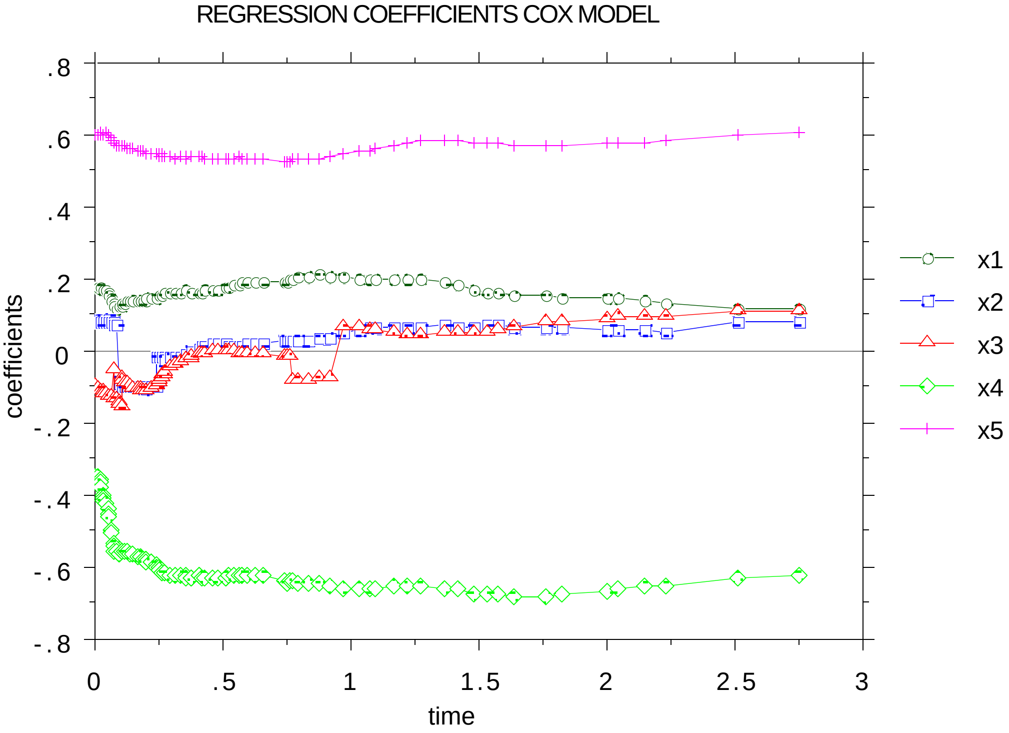
The summary results are:

.905 . 129 .042 . .772 .228 . 181 .062


REGRESSION COEFFICIENTS DRUG RETURN DATA
coefficients
.6
.5
.4
.3
.2
.1
0
-.1
-.2
-.3
-.4
-.5
0
100
200
300
400
500
time
x1
x2
x3
x4
x5
x6
x7
x8


REGRESSION COEFFICIENTS GERMAN CANCER STUDY
coefficients
.7
.6
.5
.4
.3
.2
.1
0
-.1
-.2
-.3
-.4
0
250
500
750
1000
1250
1500
1750
2000
2250
time
x1
x2
x3
x4
x5
x6
x7
x8

Are the variabilities in time artifacts of the estimation process? One of the simulated data sets used was based on a Cox model with five covariates and x(1)-2.0*x(4)+2*x(5) in the exponential. The covariates were all uniform [0,1}. The ireg=2 graph is given below.



The summary statistics are:

.807 .164 .016 -.041 -.584 .561

On average the error rates for the simulated data sets were higher than the real data sets analyzed. Therefore, the expectation is that the accuracy seen in the output from the simulated data sets should carry over to the real data sets.

NOTE Because icor=2 and ireg=2 try to find optimum monotone transformations, they are more sensitive to noise than icor=1 and ireg=1.

i<u>scale, mdimsc</u> define the proximity between two cases in the data set to be the total number of times that they occupy the same terminal node summed over all trees in the forest. These proximities, appropriated normalized, form Euclidean distances in the space of dimension ns. Extracting the first few eigenvectors from a modification of the proximity matrix, allows us to project the proximity structure of the data into low dimensions. The projection vectors are called scaling coordinates. The number of such coordinates computed is mdimsc. Usually mdimsc=3 gives most of the relevant information.

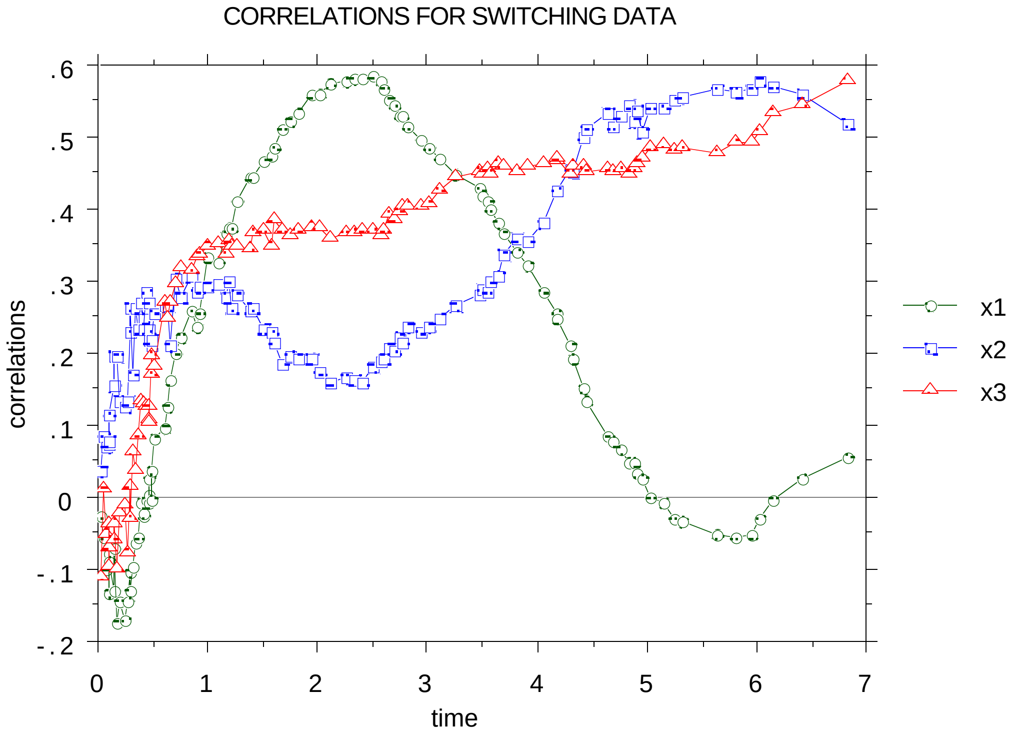
The format of the output is a matrix with ns rows. The columns are:

1) case number
2) censoring variable
3) time of death or censoring
4) to(4+mdimsc-1). the mdimsc scaling coordinates
4+mdimsc) to (4 + mdimsc+mdim-1) covariate values

The use of the scaling diagrams has not been fully developed. But there is one illustration we can give of its usefulness. One of the simulated data sets has a switching coordinate. It has 300 cases, 5 covariates, 19% censoring, and its hazard function $h(t,\mathbf{x})$ is given by

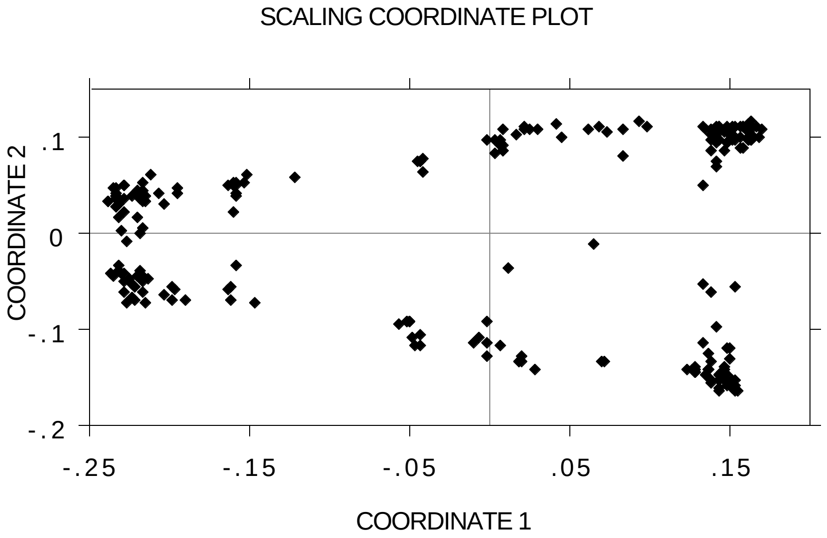
$$\mathit{if}\ \mathbf{x}_1 \le .5,\ h(t,\mathbf{x})=0\ \mathit{if}\ .5\le t\le 2.5,\ \ \mathit{else}\ \exp(x_2)$$
$$\mathit{if}\ \mathbf{x}_1 > .5,\ h(t,\mathbf{x})=0\ \mathit{if}\ 2.5\le t\le 4.5,\ \ \mathit{else}\ \exp(x_3)$$

The covariates are uniform [0,1]. Running icor=1 on this data gives the following graph:



This shows that something odd is going on with x1, but doesn't reveal its switching character.

Here is a diagram of the second scaling coordinate vs. the first.



Noting that it looked symmetric above and below the zero of 2nd coordinate, I constructed the histograms of x1 above and below this zero point The result was that the data above the zero of the 2nd scaling coordinate contained only the values of x1<.5 and below contained only the values of x1<.5. This is a strong hint that x1 functions as a switching variable.