WALD III

SOFTWARE FOR THE MASSES (AND AN EXAMPLE)

Leo Breiman
UCB Statistics
leo@stat.berkeley.edu

IS THERE AN OBLIGATION?

Tens of thousands of statisticians around the world are using statistical packages to arrive at conclusions-SAS, SPSS, STATA , etc.

These packages are filled with a lot of obsolete crank-it-out procedures, and very few up-to-date methods (except Matlab which is expensive and not in common use).

The statisticians using these packages are the very large majority in the statistical profession.

WHAT OBLIGATION DOES ACADEMIC STATISTICS HAVE TO PROVIDE THE FIELD WITH USEFUL, MODERN SOFTWARE?

(PREFERABLY FREE--KUDOS TO R)

A PERSONAL VIEW

My answer is personal--I think we have to close the great gap in statistics today between theory and practice.

We need to develop good software tools that can help the practice of statistics out in the field: the technology transfer from academic statistics to the field needs to be greatly increased.

An excellent example of the merging of theory and practice is wavelets, but there are few such examples.

The pivotal tool is developing software is the ability to do it.

Therefore, I have a radical proposal:

That every Ph.D. graduate in statistics be required to have a good working knowledge of art least one higher level language--
C, ,JAVA or Fortran 90.

My plan for myself is to put more working software on my web site for free. The following discusses my next effort.

A LESSON FROM THE PAST

Three decades ago many statisticians and quantitative social scientists were enamored of multilinear regression and its theory of hypothesis testing on the coefficients.

Every statistical package had a regression program variable selection program based on F-to delete and F to enter.

It was almost impossible to get a paper published unless you showed that a certain coefficient was signifigant at the 5% level.

This was regardless of how well the linear model fit the data and little effort was made to find out.

Many conclusions were undoubtedly wrong, and I don't think statisticians now-a-days dispute the error of these ways.

ENTER THE COX MODEL

In the last decade or two, the Cox model for that analysis of survival data has come to occupy the place in the medical field that multilinear regression once had in the social science.

A method for fitting the Cox model appears in every statistical package under the sun.

My friend and biostatistician Richard Olshen tells me that use of the Cox model is the requirement for publication in some medical journals.

When I voiced my concerns to well-known biostatisticians over the last few years, the response was "well, the data is too weak to do anything else".

I considered this a challenge to create a better method.

SURVIVAL ANALYSIS

Data from a survival experiment are of the form:

$$(c_n, t_n, \mathbf{x}_n), n = 1, \ldots, N\}$$

Here, c_n=1 if the nth case died during the duration of the experiment and 0 if it was censored by dropping out during the course of the experiment or lasting until the end.

The t_n is the time of death or censoring, and $\mathbf{x}_n$ is a vector of covariates.

The goal of survival analysis is to trace the effects of the covariates on the times of death.

HAZARD AND SURVIVAL

Given a random death time $T(\mathbf{x})$ depending on the covariate vector, define the hazard function by:

$$h(t,\mathbf{x}) = P(T(\mathbf{x}) \in (t+dt,t) | T(\mathbf{x}) \geq t) / dt$$

and the survival function:

$$S(t,\mathbf{x}) = P(T(\mathbf{x}) \geq t) = \exp(-\int_0^t h(\tau,\mathbf{x}) d\tau)$$

The Cox model makes the assumption that

$$h(t,\mathbf{x}) = r(t)\exp(\beta \cdot \mathbf{x})$$

and then makes a clever partial likelihood jump that allows the $r(t)$ to be canceled out in the estimation of β. One result is that all survival curves have the same shape.

Can this possibly represent the complex action of nature even as a first approximation?

THREE SIMULATED DATA SETS

Its useful when testing new methods to generate simulated data where truth is known.

Sim1. This is 300 case data with five uniformly distributed covariates, 19% censoring, generated from a Cox model where the hazard function is:

$$h(t,\mathbf{x}) = \exp(x_1 - 2x_4 + 2x_5)$$

Sim2 This 300 case data with three uniformly distributed covariates, 15% censoring, has hazard function::

$$\textit{if } \mathbf{x}_1 \leq .5,\ h(t,\mathbf{x}) = 0 \textit{ if } .5 \leq t \leq 2.5,\ \textit{else } \exp(x_2)$$

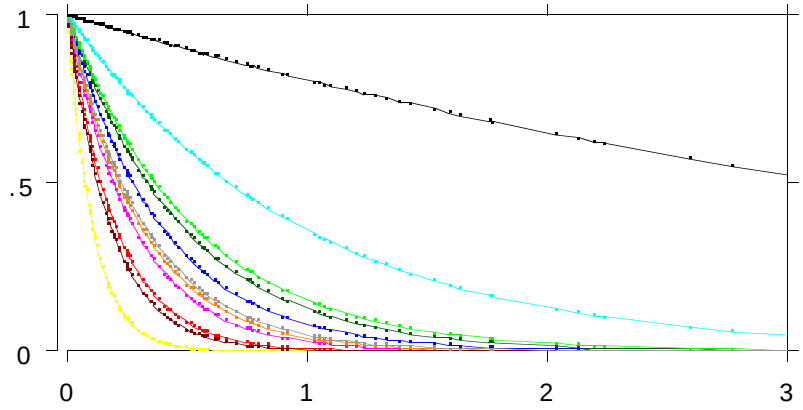
$$\textit{if } \mathbf{x}_1 > .5,\ h(t,\mathbf{x}) = 0 \textit{ if } 2.5 \leq t \leq 4.5,\ \textit{else } \exp(x_3)$$

Sim3 This 300 case data set with six uniformly distributed covariates, 29% censoring, has hazard function:

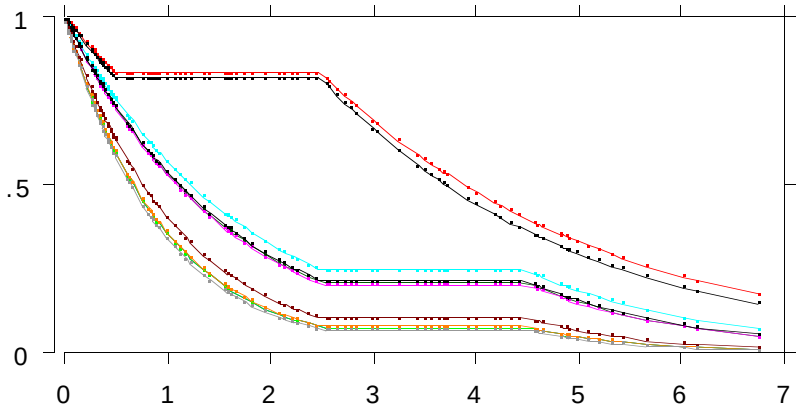
$$h(t,\mathbf{x}) = (1 + z_2 t)\exp^*(z_1 + z_2 t)$$

where $z_1 = .5x_1,\ z_2 = x_4 + x_5$

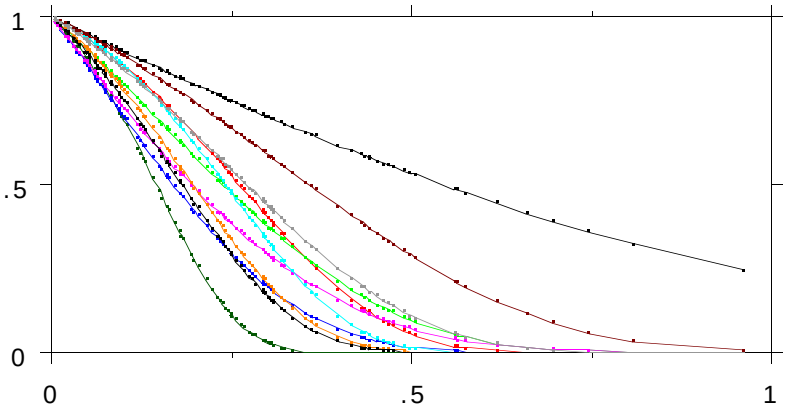
SIM1 SURVIVAL CURVES



SIM2 SURVIVAL CURVES



SIM3 SURVIVAL CURVES



INTRODUCING SURVIVAL FORESTS (SF)

I am developing a method (*survival forests (SF))* for estimating the survival functions $S(t,\mathbf{x})$ At present it is still in the experimental stage

These functions can also be estimated using the Cox model by first estimating coefficients and then doing an ml estimate of $r(t)$.

With simulated data the true survival functions $S^*(t,\mathbf{x})$ are known.

Define the error in the estimates as:

$$error = av_{k,n}|S(t_k,\mathbf{x}_n) - S^*(t_k,\mathbf{x}_n)|$$

where the $\{t_k\}$ are the uncensored death times

Errors

data set	Cox	SF
Sim1	.054	.066
Sim2	.135	.100
Sim3	.572	.068

FOUR REAL DATA SETS

Simulated data may not reflect the vagaries of real data. I have been working with four real data sets.

Bcan: this is a breast cancer data set sent to me from England. It has 272 cases, 6 covariates and 17% censoring.

Vcan: a veterans cancer data set with 136 cases, 6 covariates and 7% censoring.

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

Gcan: the German cancer study with 686 cases, 7 covariates, and 56% censoring.

ARE SIMULATED AND REAL COMPARABLE

Do the real data sets have as much information in them as the simulated?

In SF there is a test set method for estimating the times of death--estimating $td(n)$ by $\hat{t}d(n)$.

Measure the error in this estimate by:

$$error = av_n(|td(n) - \hat{t}d(n)|/\hat{t}d(n))$$

and the strength of the data by 100(1-error). If the error is high the strength is low.

Data Set	Strength
Sim1	19.5
Sim2	13.5
Sim3	32.0
Bcan	27.4
Vcan	18.9
UIS	35.5
Gcan	45.2

The real data sets have, on average, more strength than the simulated. Strength is an indicator of how well the survival functions are estimated. Note that the lowest strength simulated data set Sim2 also had the highest error in estimating these functions.

CORRELATIONS OVER TIME

There are three methods (so far) for using SF to look at the time process in the data. The first method is time-varying correlations. We work with

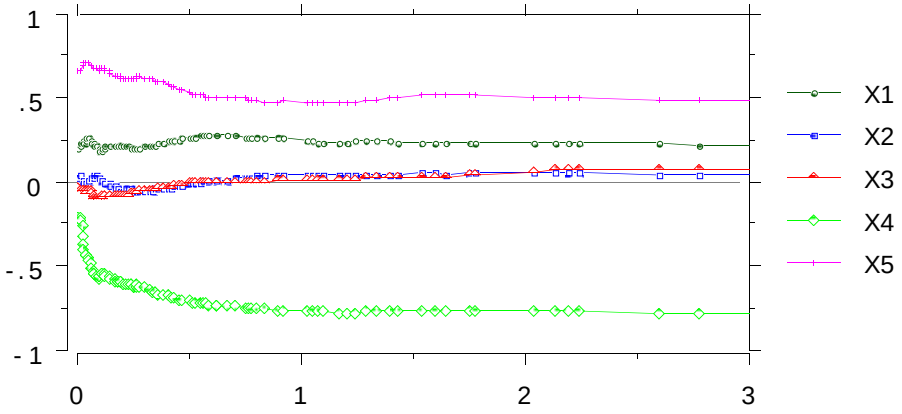
$$L(t,\mathbf{x}) = -\log(S(t,\mathbf{x}))$$

Take the time points $\{t_k\}$ to be 100 order statistics from the uncensored death times.

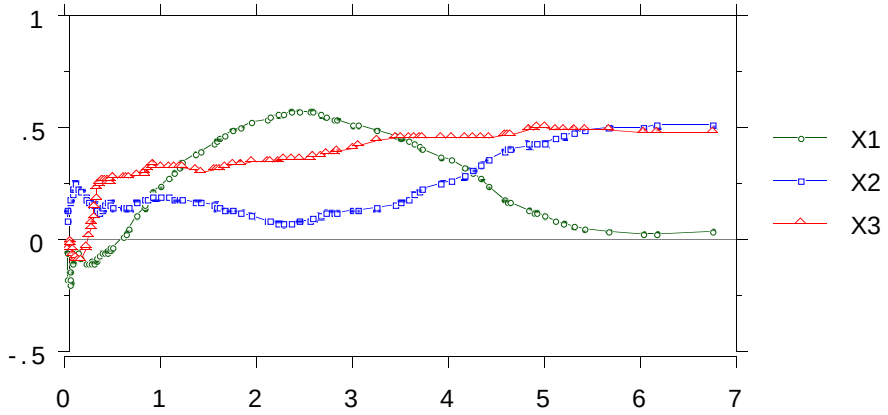
At each t_k compute the correlation between $L(t_k,\mathbf{x}_n)$ and each of the covariates. If a Cox model fits the data, these correlations should be constant in time.

Graphs of the corelations follow.

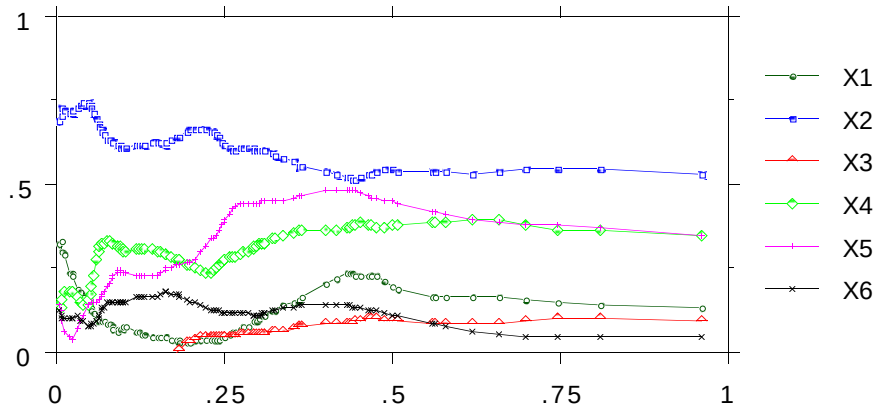
SIM 1 CORRELATIONS



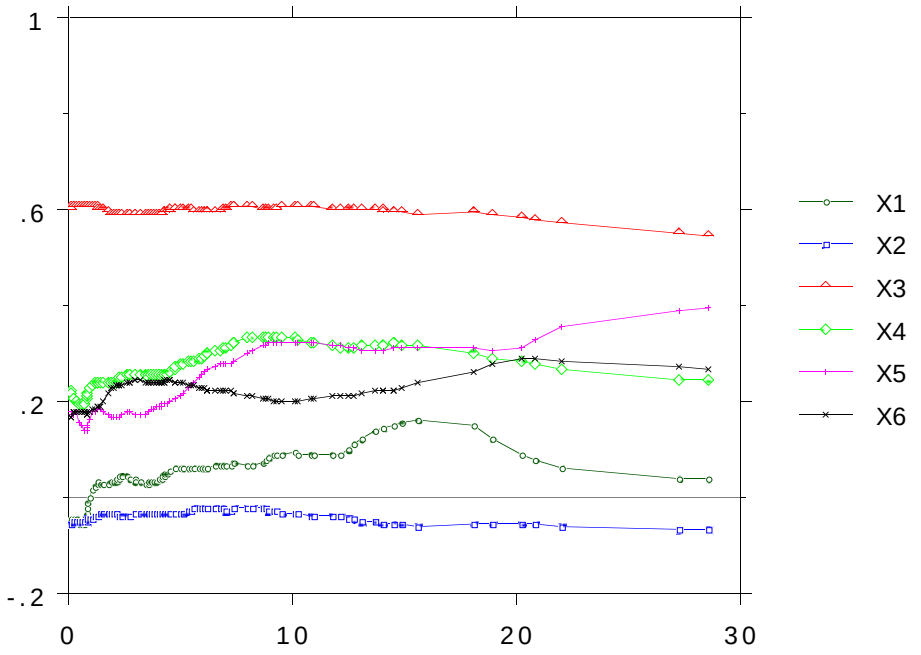
SIM 2 CORRELATIONS



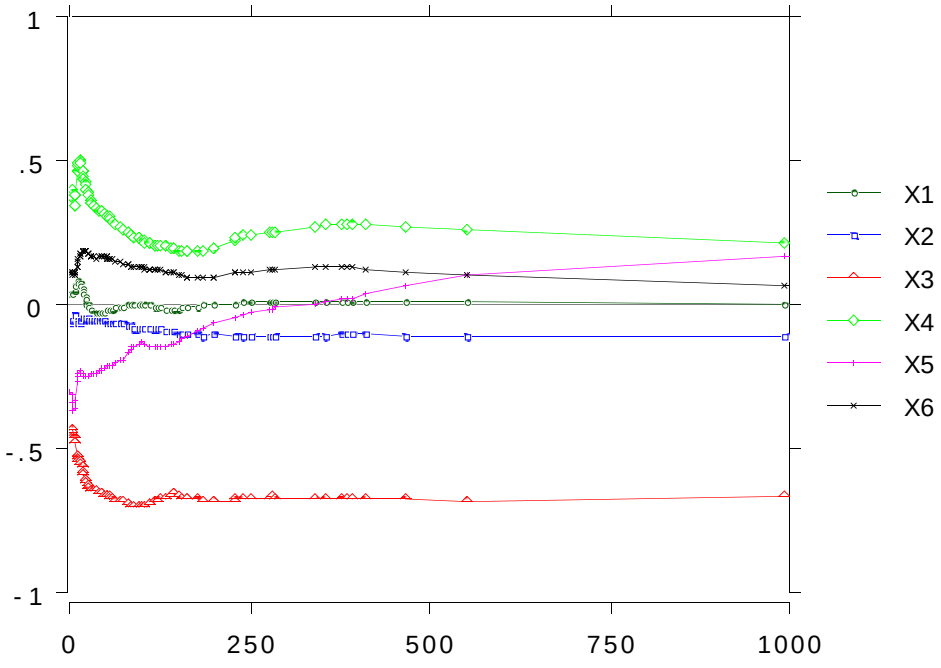
SIM 3 CORRELATIONS



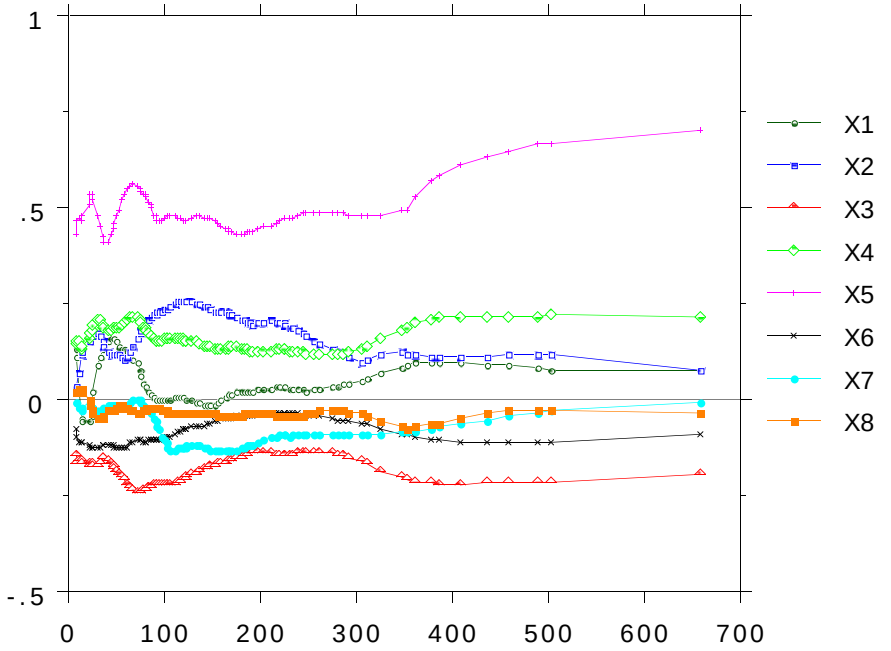
BCAN CORRELATIONS



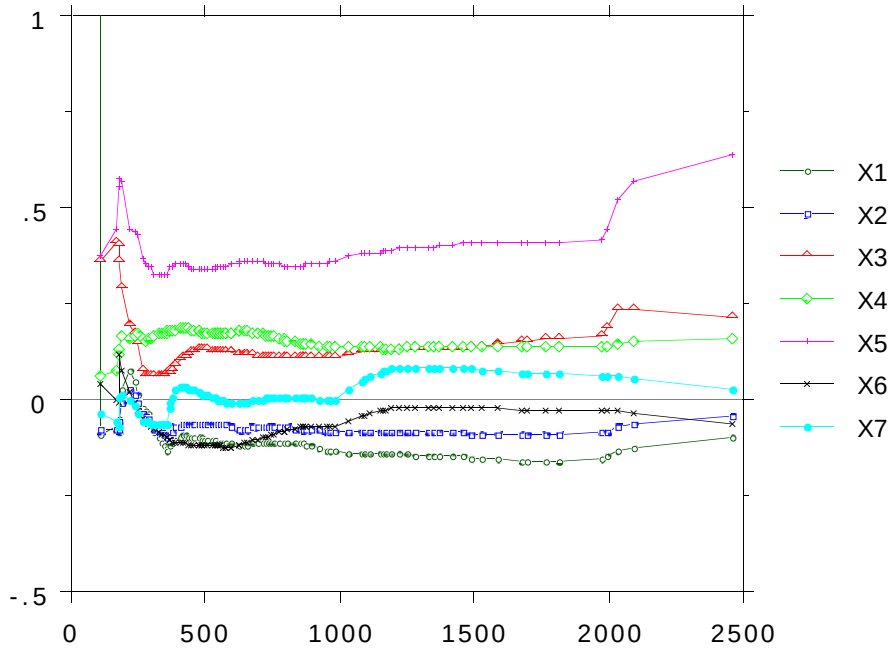
VCAN CORRELATIONS



UIS CORRELATIONS



GCAN CORRELATIONS



TIME FITTING COX MODELS

Another tool that is useful is fitting a Cox model at each time t_k to the $L(t_k, \mathbf{x}_n)$.

At each t_k let

$$y_n = L(t_k, \mathbf{x}_n)$$

and

$$f(n, t(k), \beta(k)) = t(k) \exp(\beta(k) \cdot \mathbf{x}_n)$$

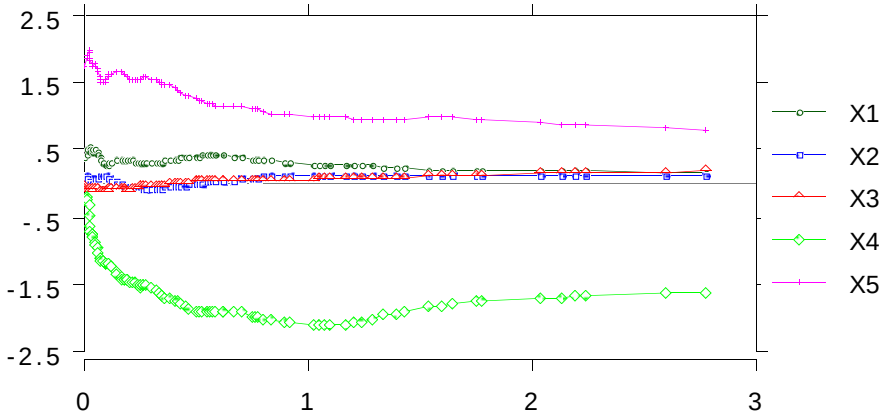
The right hand side is the negative log of the Cox expression with the $t(k), \beta(k)$ parameters to be determined by minimizing

$$\sum_n (y_n - f(n, t(k), \beta(k)))^2$$

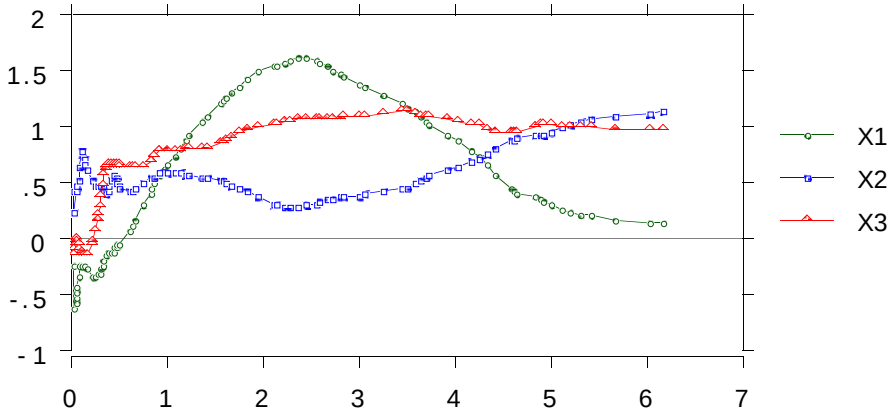
If the Cox model is a good fit, the $\beta(k)$ will ce constant in time and the $t(k)\}$ are an estimate of the integral of the baseline risk.

Again, we show some graphs:

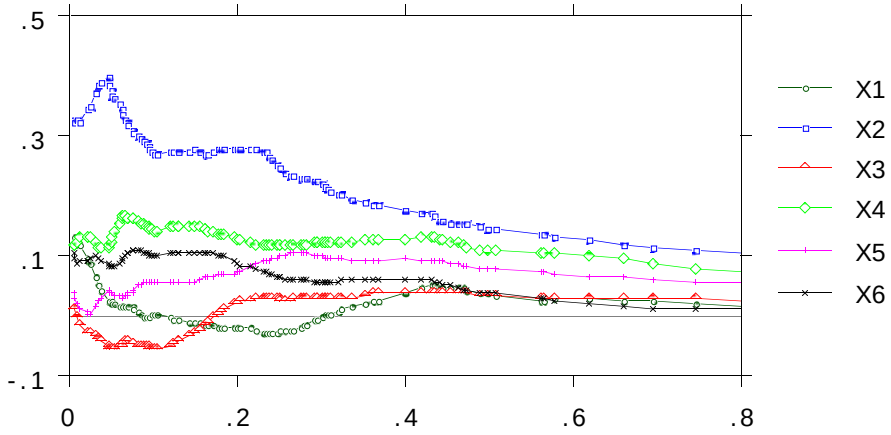
SIM 1 COEFFICIENT TRACES AV RSQ=.84



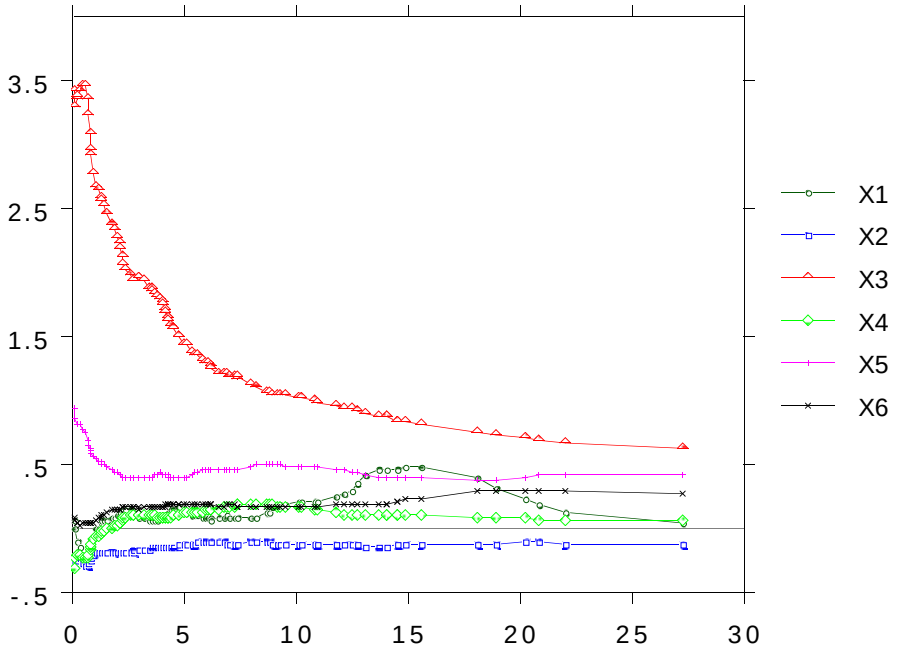
SIM 2 COEFFICIENT TRACES AV RSQ=.33



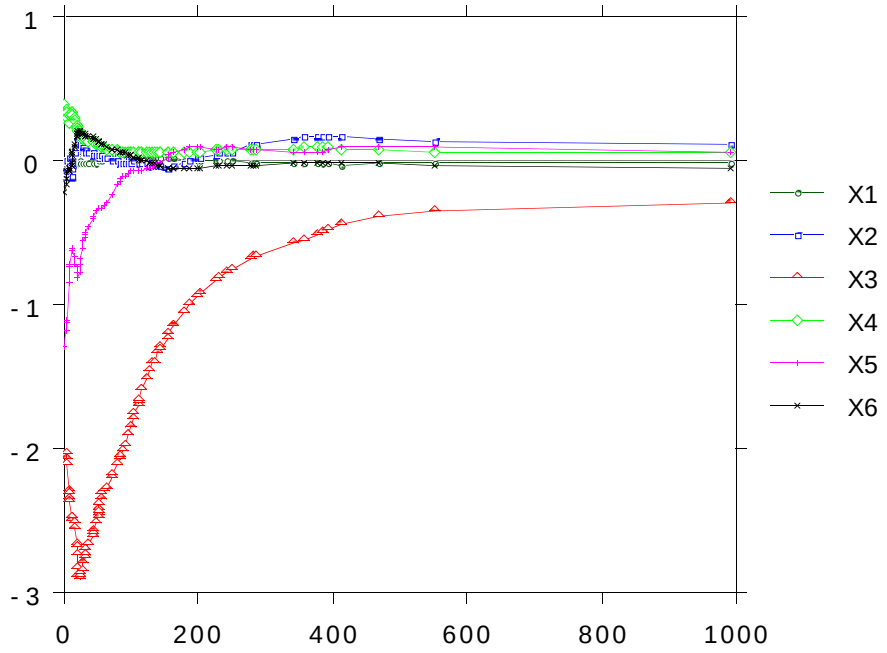
SIM 3 COEFFICIENT TRACES AV RSQ=.62



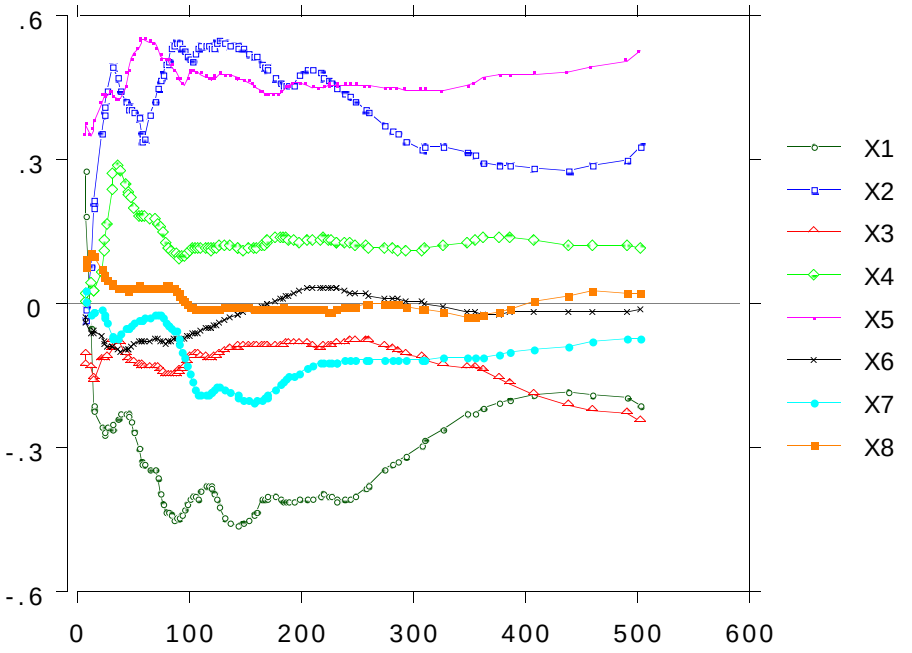
BCAN COEFFICIENT TRACES AV RSQ=.70



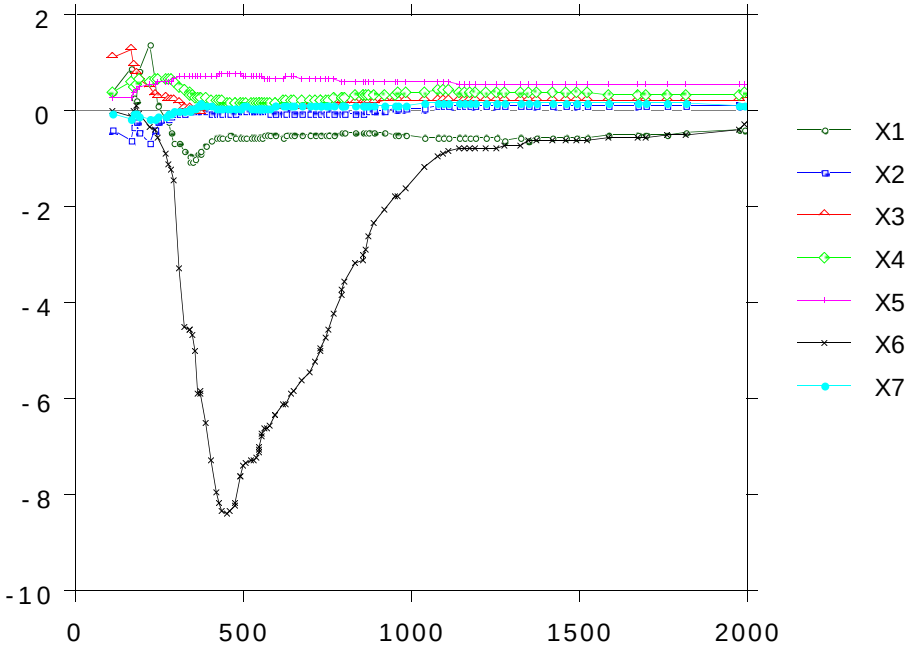
VCAN COEFFICIENT TRACES AV RSQ=.78



UIS COEFFICIENT TRACES AV RSQ=.48



GCAN COEFFICIENT TRACES AV RSQ=.44



END NOTE ON APPLICATIONS

When there are changes in time over the duration of the experiment, the Cox model gives dubious results.

For instance, in the UIS data the Cox model singles out the significant variables (in order) as X1>X4>X7 and no others significant at the .05 level.

Examination of the correlations in time and the coefficient traces for the UIS data shows that there are three important variables XI,X2,X5.

The Cox model may be dangerous to the medical experimental profession.

HOW SV WORKS

Building a Survival Tree

With fixed covariate effects and random censoring, the log likelihood can be expressed in terms of the hazard function $h(t,\mathbf{x})$ as

$$LL = \sum c_n \log[h(t_n, \mathbf{x}_n)] - \sum \int_0^{t_n} h(\tau, \mathbf{x}_n) d\tau$$

This expression is our starting point. The tree will be grown to maximize it.

Divide the time-covariate space into a union of L disjoint rectangles indexed by l. Call these nodes.

Use the notation $l = I_l \otimes R_l$ where I_l is a time interval and R_l is a rectangle in covariate space.

The essential step is to express $h(t,\mathbf{x})$ as

$$h(t, \mathbf{x}) = \exp[\sum_l I((t, \mathbf{x}) \in l) \alpha(l)]$$

where I is the 0,1, indicator function.

CONTINUED

Define (1)

$$N_D(l) = \sum_n c_n I((t_n, \mathbf{x}_n) \in l)$$

Then $N_D(l)$is the number of deaths in the node l.

Define also (2)

$$T(l) = \sum_n I(\mathbf{x}_n \in R_l) |(0, t_n) \cap I_l|$$

where | | denotes length of the enclosed interval.

Substituting the expression for $h(t,\mathbf{x})$, (1) and (2) into the expression for the LL, gives

$$LL = \sum_l \alpha(l) N_D(l) - \sum_l T(l) \exp(\alpha(l))$$

Maximizing this over the $\alpha(l)$ gives

$$\alpha(l) = \log(N_D(l) / T(l))$$

and

$$LL = \sum_l N_D(l) \log(N_D(l) / T(l)) - \sum_l N_D(l)$$

SPLITTING NODES

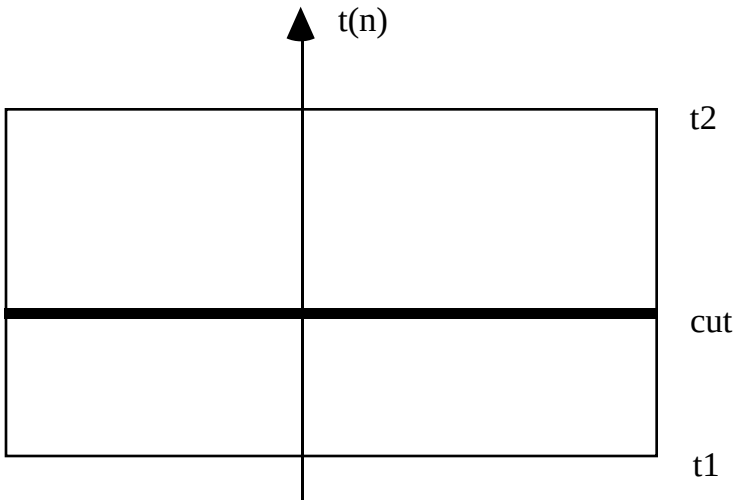
The second term above is constant.

At each step the split of l is found which most increases

$$N_D(l)\log(N_D(l)/T(l))$$

Univariate splits on a covariate are the ordinary splits as used in CART but using the above expression to maximize.

Splits on time are tricky.



A case with covariate $\mathbf{x}_n$ keeps traveling through nodes until it comes to a node such that $t_1 < t_n \leq t_2$.

In a time split of a node, all cases in the original node are in each child node.

CONTINUED

To work, there have to be many splits on time.

With probability .75 the split on each node is on the time variable only.

With probability ,25, the split is the best split on any of the covariates.

The tree is grown until each terminal node has exactly one uncensored death in it. But terminal nodes can contain many other cases.

GROWING THE FOREST

Each tree in the forest is grown on a bootstrap sample from the original training set. Again, about one-third of the cases are oob.

Let **x** be the covariate vector of an oob case. Put it down the corresponding tree and estmate $S(t,\mathbf{x})$ by:

$$\log(S(t,\mathbf{x})) = -\sum_{l}[N_D(l)/T(l)]\,I(\mathbf{x}\in R_l)|(0,t)\,\text{I}\,I_l|$$

where the sum is over all terminal nodes.

This estimate is accumulated and averaged every time **x** is not in the tree growing sample.

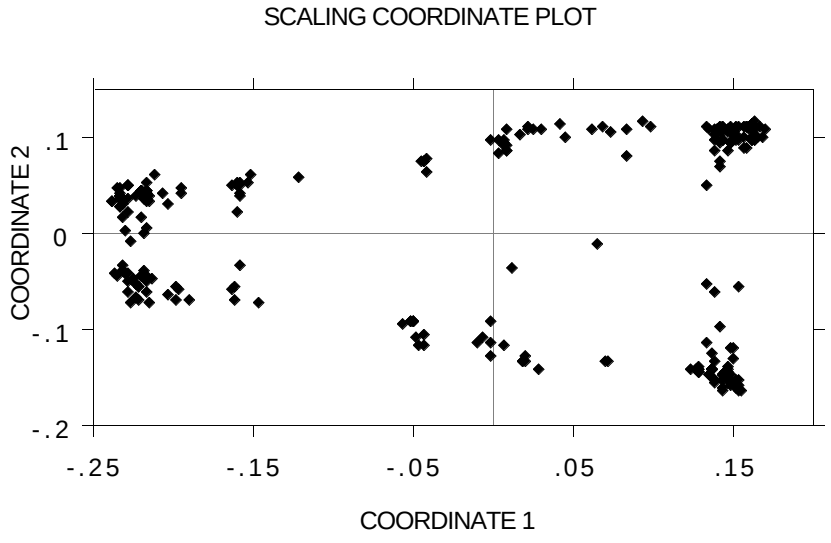
The averaged survival function, estimated for cases in the data not used in the corresponding tree, is automatically a test set estimate.

TYING UP A LOOSE END WITH SCALING

As in RF, proximities can be defined by of often two cases occupy the same terminal node, although there is a different definition for terminal nodes in SF.

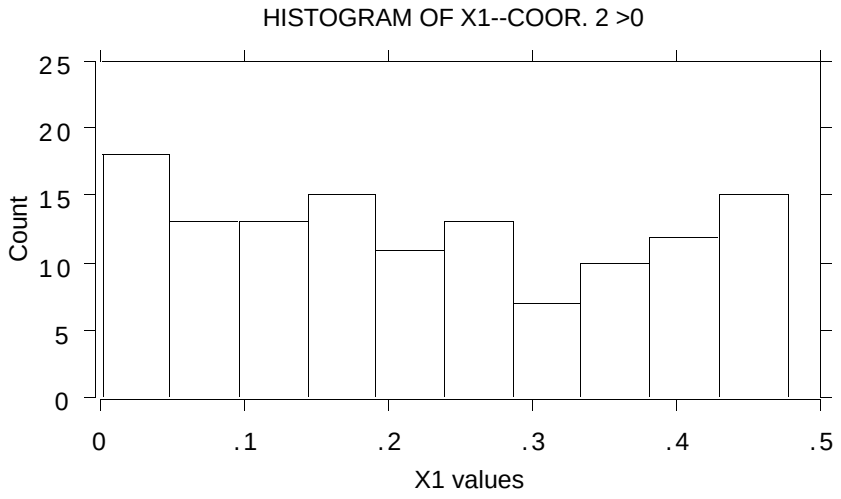
They can also be projected down into two dimensions. Recall, we have not yet solved the problem of how to recognize the action of the switch variable in Sim2.

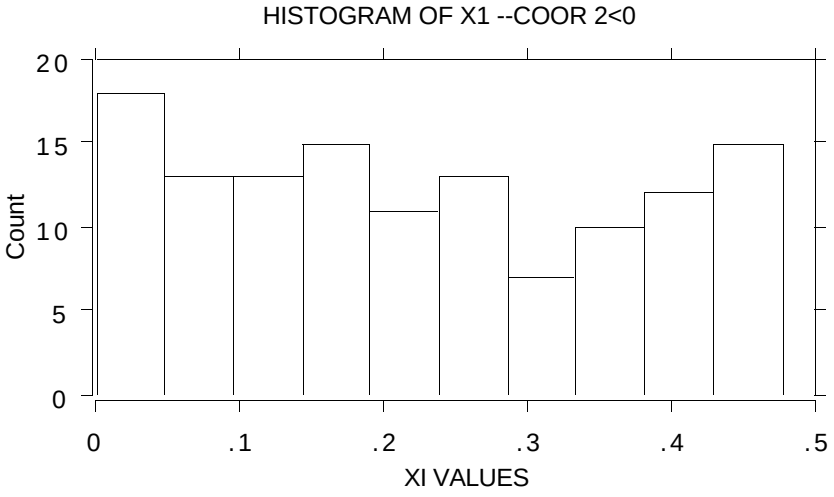
Here is a graph of the 2nd scaling coordinate versus the first for sim 2.



AHA!

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.





AHA! A CLUE

FINAL REMARKS

RF has been extensively tested in the field.

SV is still being born and needs more testing, working with, and extending.

But it may prove the point that algorithmic models can provide more (and more reliable) information than stochastic models.

As soon as it is in decent shape, SV will show up on my web site as free software for use by the masses.

THANK YOU