KM 41 (1992)
pg.5-18

FACTOR SCREENING BY SEQUENTIAL BIFURCATION

Bert Bettonvil

The purpose of screening is the reduction of a large set of explanatory variables or factors to the set of important variables, assuming that there are only a few important explanatory variables. As a screening method we propose a modification of Sequential Bifurcation, a method resembling the binary search technique.

This paper is limited to deterministic linear response surfaces (but the method can be extended to random nonlinear response surfaces). Besides a description of Sequential Bifurcation, we discuss its efficiency and a large scale application.

Keywords: Simulation, design of experiments: screening.

Tilburg University, School of Economics and Management,
P O Box 90153, 5000 LE Tilburg, Netherlands.
Phone: 013-662359

1. INTRODUCTION

Suppose we are faced with a problem in which a response depends on a great many explanatory variables, but we think that only a few of these variables are really important. There are several methods to find out which variables are important. The one-factor-at-a-time method uses one basic observation with all variables at one level and - in case of N variables - N extra observations; in each of these observations N-1 variables are at the basic level and one variable is at some other level. A more efficient method uses a Resolution-III design where the number of observations equals N+1, rounded up to the next multiple of four; see Box and Hunter (1961a,b). But an observation can be very expensive, in which case *obtaining N or more observations is prohibitive.*

A number of techniques have been developed to tackle this problem. In this paper we propose a modification of Sequential Bifurcation (SB). We restrict ourselves to a model that contains only the main effects of the input variables, and to observations without random errors. In most applications these assumptions are not realistic, but the present model can be viewed as a step towards more realistic models, containing random errors and/or interactions. Interactions as well as random errors will be treated in forthcoming papers.

In section 2 of this paper we describe SB and introduce a special notation. In section 3 we derive the number of observations needed to find k important variables out of N candidate variables. In section 4 we discuss a large scale example. In section 5 we draw conclusions and indicate further research.

2. DESCRIPTION AND NOTATION

Suppose the outcome y can be expressed as a first-order linear model of N variables $x_1,x_2,\ldots,x_N$:

$$y = y(x_1,x_2,\ldots,x_N) = \beta_0 + \beta_1 x_1 + \ldots + \beta_N x_N. \qquad (1)$$

A crucial assumption is that there is a priori knowledge of the *direction* of the influence of each x_i (i=1,...,N), if such an influence is present at all. Consequently, we can recode x_i such that all variables have non-negative influence. So we assume $\beta_i \geq 0$ (i=1,...,N). We want to find the important variables in an efficient way, where a variable is called important iff its regression parameter is large. We quantify "large" as "exceeding δ", where δ is assumed to be some given non-negative number. We shall return to this issue later. We do not know which of the β's are large, nor do we know the number of large β's, but we expect this number (say) k to be small ($k << N$).

Our SB is a modification of the Jacoby and Harrison (1962) approach. For sake of convenience we assume that $N=2^m$; if not, we can add dummy variables to the model (with β_i known to be zero). In a linear model like (1) we may investigate only two levels per independent variable, which we can denote as "low" and "high", or "off" and "on", or "0" and "1".

Let $y_{(i)}$ (i=0,1,...,N) denote the observation with $x_1 = x_2 = \ldots = x_i = 1$ and $x_{i+1} = x_{i+2} = \ldots = x_N = 0$. Then (1) gives

$$y_{(i)} = \beta_0 + \Sigma_{\ell=1}^{i} \beta_\ell \quad (i=0,1,\ldots,N), \tag{2}$$

and the sequence $(y_{(i)})_{i=0,1,\ldots,N}$ is non-decreasing (because $\beta_i \geq 0$). For $i<j$ we have

$$y_{(j)} - y_{(i)} = \Sigma_{\ell=i+1}^{j} \beta_\ell,$$

where we denote the latter sum by $\beta_{i+1,j}$; as a consequence $\beta_{j,j}=\beta_j$. We begin our search for the important variables by observing the two responses with all independent variables at their high and low levels, respectively: $y_{(N)}$ and $y_{(0)}$ (see eq. 2). The difference between their outcomes, $\beta_{1,N}$, equals the sum of all regression parameters, which are all non-negative. So if $\beta_{1,N} \leq \delta$, all parameters are small (i.e., do not exceed δ), and our problem is solved, using only two observations. However, if $\beta_{1,N} > \delta$, then one or more parameters may be large.

In the latter case we observe $y_{(N/2)}$ and compute $\beta_{1,N/2} = y_{(N/2)} - y_{(0)}$ and $\beta_{N/2+1,N} = y_{(N)} - y_{(N/2)}$. This *bifurcation* is continued until we reach individual parameters: *sequential* bifurcation. The whole procedure can be described as follows:

(1) Observe $y_{(0)}$ and $y_{(N)}$; compute $\beta_{1,N}=y_{(N)}-y_{(0)}$.

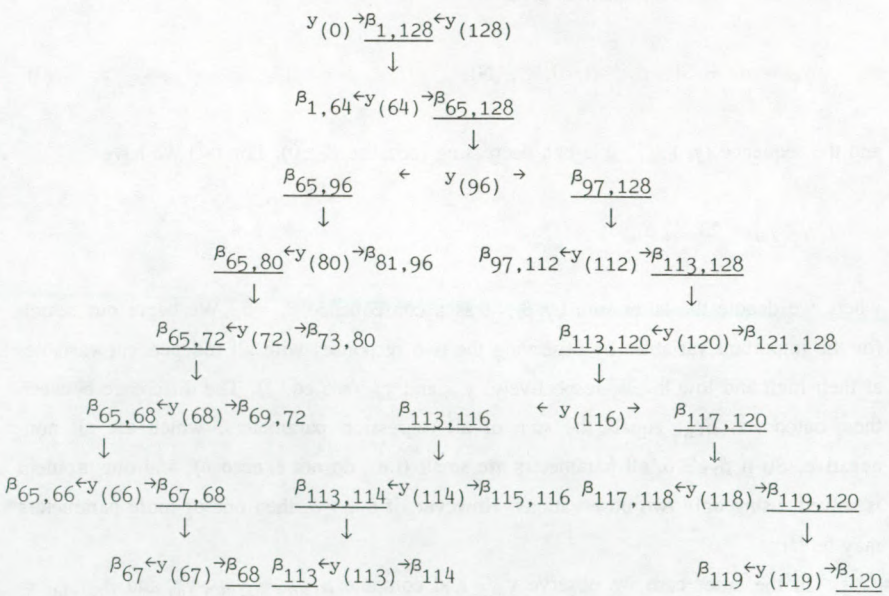
(2) If $\beta_{i,j}\leq\delta$, then all its β-components are small;
if $\beta_{i,j}>\delta$ and $i=j$, we have found a $\beta_i>\delta$;
if $\beta_{i,j}>\delta$ and $i<j$, proceed to step (3).

(3) Observe $y_{(k)}$, where $k=(j+i-1)/2$ and compute
$\beta_{i,k}=y_{(k)}-y_{(i-1)}$ and $\beta_{k+1,j}=y_{(j)}-y_{(k)}$;
proceed to steps (4a) and (4b).

(4a) The new β under consideration is $\beta_{i,k}$; proceed to step (2).

(4b) The new β under consideration is $\beta_{k+1,j}$; proceed to step (2).

FIGURE 1. Jacoby and Harrison (1962)'s Example of Sequential Bifurcation.



Example 1. Jacoby and Harrison (1962) give an example with $128=2^7$ variables, in which only the ones numbered 68, 113 and 120 have non-zero effects ($\delta=0$). Our SB analysis starts by observing $y_{(0)}$ and $y_{(128)}$, resulting in $\beta_{1,128}>0$. The next observation is $y_{(64)}$, which gives $\beta_{65,128}=y_{(128)}-y_{(64)}>0$ and $\beta_{1,64}=y_{(64)}-y_{(0)}=0$, so $\beta_1=\ldots=\beta_{64}=0$. Figure 1 represents the whole procedure. The positive parameters are underscored; the arrows indicate the order of observations and calculated parameters.

Jacoby and Harrison (1962) propose to take two observations for every group of β's to be split, that is, one observation with the first half of the variables under consideration "on", the other with the second half "on". Their procedure demands many more observations, actually almost twice as much: in example 1 we need 16 observations, whereas Jacoby and Harrison need 29 observations. The number of observations is further discussed in the next section.

3. NUMBER OF OBSERVATIONS

We compute the number of observations that is necessary to find k non-zero coefficients ($\delta=0$) out of $N=2^m$ non-negative ones. First we consider worst-case (upper-limit) computations.

Suppose $k=0$. Trivially, we need only two observations: $y_{(0)}$ and $y_{(N)}$.

Suppose $k>0$. We can define an integer ℓ ($0\le\ell\le m$) such that $2^{\ell-1}<k\le 2^\ell$. If $m=\ell$, then in the worst case situation we need $1+2^\ell$ observations to find the important variables. If $m>\ell$, we first divide the 2^m variables into 2^ℓ groups of size $2^{m-\ell}$ each. In the worst case, the important variables are as dispersed as possible: they are in k different groups of size $2^{m-\ell}$, and we need $1+2^\ell$ observations to identify these groups. For each of these k groups we need $m-\ell$ more observations to identify the individual important variables. Hence, in the worst case, the number of observations is

$$1 + 2^\ell + k(m-\ell). \qquad (3)$$

If k is equal to some power of two, then (3) reduces to

$$1 + k + k(^2\log N - {}^2\log k) = 1 + k(1 - {}^2\log(k/N)), \qquad (4)$$

which can also serve as an approximation to (3) if k is not a power of two.

Example 2. In example 1 we have $128=2^7$ variables with 3 non-zero effects. So $m=7$, $k=3$, and as a consequence $\ell=2$. The maximum number of observations, according to (3), is 20; approximation (4) gives 20.2. The actual number of observations is 16 because the important factors are somewhat clustered (they all lie in the second half).

In appendix 1 we describe a number of rival screening techniques and compare them to Sequential Bifurcation. This yields table 1: the results are in favour of our method.

Example 3. Suppose we are dealing with $N=1024$ variables ($m=10$). For $k=0,1,\ldots,8$ the worst case number of observations is given in table 1, in which G2 stands for Two-stage group-screening, GM for Multi-stage group screening, JH for the Jacoby and Harrison (1962) Sequential Bifurcation, and SB for our version of Sequential Bifurcation.

Table 1.

Maximum number of observations for given k (number of non-zero variables).

k	0	1	2	3	4	5	6	7	8
G2	4	68	96	116	136	148	160	172	188
GM	2	20	35	49	62	74	85	96	107
JH	3	21	39	55	71	85	99	113	127
SB	2	12	21	29	37	44	51	58	65

For G2, JH and SB we compute the number of observations for the worst case, i.e.,, there is as little clustering as possible. For G2 we assume that for the first stage we guessed the number of non-zero coefficients correctly. For GM we apply the formula derived in appendix 1 and round up to the next integer.

An alternative for worst-case behaviour is the "expected" number of observations, using either a binomial model (where each variable has an a priori probability p of being important) or a random permutation of the N factors (where k is fixed). In a given physical system, however, the importance of a factor does not depend on a random mechanism; it is a property of the system. Moreover, the worst-case calculation can be justified by (i) its computational ease, (ii) worst-case results are only slightly higher than the expected values over all random permutations (for small k and large N), and (iii) the apparent superiority of SB using worst-case computations (see appendix 1). Nevertheless, for the sake of completeness, we derive the expected number of observations under the binomial model in appendix 2. This yields table 2.

Table 2.
Expected Number of Runs for N=1024 input variables for
Watson's (1961) Two-Stage Group Screening,
Morris's (1987) Multiple Grouping,
and Sequential Bifurcation (SB).

prior probability	Expected Number of Runs		
	Watson	Morris	SB
.0001	21.9	10.3	3.0
.0002	29.8	12.6	4.0
.0005	48.8	20.6	6.9
.001	64.3	26.4	11.4
.002	94.6	41.4	19.6
.005	139.0	81.7	40.8
.01	198.2	130.3	70.5
.02	262.7	253.1	120.1
.05	399.5	399.5	234.5
.1	521.1	521.1	374.2

4. AN APPLICATION: THE GREENHOUSE MODEL

We use Sequential Bifurcation to screen 281 explanatory variables in a deterministic simulation model of the CO_2 concentration in the year 2100. The model is developed at the Dutch National Institute of Public Health and Environmental Protection RIVM as a part of IMAGE: the Integrated Model for the Assessment of the Greenhouse Effect; see Rotmans (1990). IMAGE aims at giving quantitative insight into the greenhouse phenomenon. It consists of separate, autonomously functioning modules, which are concatenated and integrated with each other. We restrict ourselves to the carbon cycle module, which describes the global circulation of CO_2, Carbondioxide, in the atmosphere, oceans, and terrestrial biosphere.

To this simulation model we apply SB as if we had 512 input variables (512 being the smallest power of 2 exceeding 281) with a priori knowledge: $\beta_{282} = \beta_{283} = \ldots = \beta_{512} = 0$, so $y_{(281)} = y_{(282)} = \ldots = y_{(512)}$. Setting all input variables at their low and high levels yields a simulated CO_2 concentration in the year 2100 of $y_{(0)}=987.51$ and $y_{(281)}=1495.66$ parts per million. After 80 observations we stop with 32 regression parameters exceeding $\delta=5$, and all remaining (groups of) parameters smaller than 5.

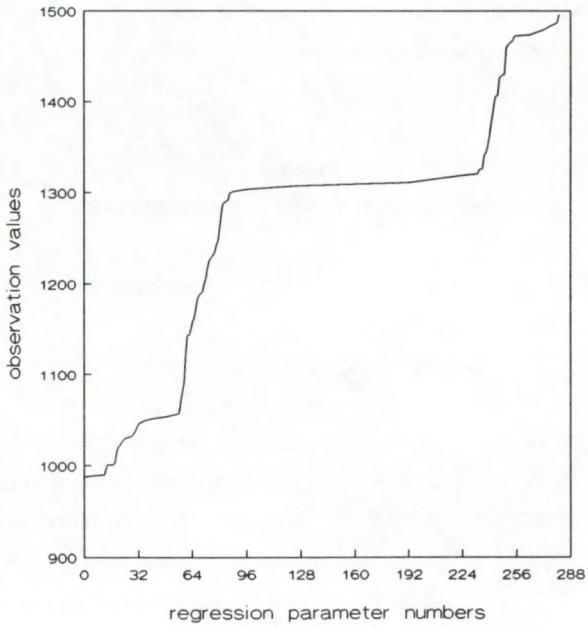
Remark. We do not chose δ beforehand. Instead, by considering the group of parameters with the largest sum, we can, at any time during the analysis, decide to either take this sum as δ and to stop, or to split this group. The decision to stop with $\delta=5$ is more or less arbitrary.

Figure 2 gives the outcomes $y_{(i)}$; intermediate values come from linear interpolation, e.g., we have no actual observations between $y_{(96)}=1303.77$ and $y_{(128)}=1307.57$, nor between $y_{(128)}$ and $y_{(192)}=1310.86$.

To verify our results, we select eight variables, three of which were found by SB as well as by previous (small-scale) analyses of the system, four were found by SB but neglected in previous analyses, and one was previously assumed to be important, but was not found by SB. We simulate 16 combinations of these 8 factors (resolution IV design; see Kleijnen 1987), and compute the regression parameters. The results are given in table 3, together with the SB results. Table 3 shows that the estimated main effects are quite close to the SB estimates.

The simulation observations are almost perfectly monotone; the only deviation from monotonicity is

Figure 2. Results of Greenhouse Study.



$$y_{(14)} - y_{(16)} = 0.54.$$

Whether we assume that this deviation is caused by wrong coding of some variables (deterministic model) or by noise (random model), is irrelevant: the conclusions of our investigation, namely which variables are important, hold in either case.

The outcomes of the SB analysis did not coincide with the expectations of the system experts. Especially the shifts from and to temperate forest had received too little attention in their prior studies of the system. The results of our analysis gave rise to further investigation, which takes place now.

Table 3.

The important variables of the RIVM model.

name	effect		meaning
	SB	Verification (Resolution IV)	
CHREF(31)	30.16	26.14	shift from temperate forest to agricultural land
CHREF(24)	19.89	17.07	shift from temperate forest to grassland
TC2A	11.97	14.63	residence time in the (thick) cold mixed layer
STIM	8.80	10.32	biotic stimulation factor
PRECIP	7.72	8.85	rate of precipitation of carbon in the oceans
CHAREF(2)	7.42	7.08	fraction of charcoal formed upon burning of branches
MFLOW	5.93	6.20	circulating massflow (Gordon flow)
DIFF	<5	3.95	effective diffusivity in the oceans

5. DISCUSSION

In this paper a modification of the Jacoby and Harrison (1962) Sequential Bifurcation is described, which is easy to perform and turns out to be very efficient. Our method is a screening method: it is designed to get rid of the, presumably overwhelming, amount of unimportant input variables.

Our model (1) is very restrictive; the method presented can be used for models without random errors, but also for models with parameters which are very large, compared to the noise. Augmentation of the model with random errors, as well as with interactions, is feasible; details will be given in future papers.

ACKNOWLEDGMENT

I am indebted to Professors Jack Kleijnen (Tilburg University) and Peter Sander (Eindhoven University of Technology) for their support during my work on this subject.

APPENDIX 1. COMPARISON OF SB WITH OTHER SCREENING TECHNIQUES

We describe some alternative screening techniques, and compare them with our version of SB, theoretically as well as by means of an example. We shall discuss Two-Stage Group-Screening, Multi-Stage Group-Screening, and Jacoby and Harrison (1962)'s original version of Sequential Bifurcation.

(a) Two-Stage Group-Screening (Watson (1961), Mauro (1984), Mauro and Burns (1984)). The N variables are divided into G groups of size $g=N/G$ each (if N is not a multiple of G, the group sizes are taken as "evenly" as possible). As a first step, the G groups are studied by using a Plackett and Burman (or PB) design, see Kleijnen (1987, p. 302). The variables within each group are treated as a whole, i.e. they are varied simultaneously. In the second stage the variables in the groups that in the first stage turn out to be important (if any), are submitted to a next PB design.

If k variables are important and these variables are all in different groups (worst case), then the total number of observations is approximately $G+kg=G+kN/G$. The optimal G is about $\sqrt{(kN)}$, resulting in $\sqrt{(kN)}$ observations, as is easily verified. For small k and large N our procedure, which is of the order $k^2 \log N$, is superior.

(b) Multi-Stage Group-Screening. Both Patel (1962) and Li (1962) generalize two-stage group-screening to its multi-stage analogue. We shall briefly describe Patel's version; Li's approach differs only in detail.

In the first stage, the group of N variables is divided into g_1 groups of N/g_1 variables each, and analyzed in g_1+1 experiments (as in two-stage group-screening the variables within one group are varied simultaneously). Suppose the first stage turns out to give k_1 important groups. In the second stage, each of these groups is divided into g_2 groups of $N/(g_1g_2)$ variables and analyzed by using g_2 experiments. So, in the second stage a total number of k_1g_2 experiments are performed, resulting in (say) k_2 important groups. This is continued until we reach groups of size 1.

Patel neglects the difficulties arising from the fact that all divisions should result in integer numbers, so his results tend to be optimistic. He assumes that every variable has an a priori probability p of being important, and finds that in c stages approximately $1+cNp^{1-1/c}$ experiments are needed. This boiles down to an optimal c (straightforward differentiation) of -ln p, so the optimal number of experiments is $1-Npe \ln p$. The worst case number of observations in SB can be approximated by $1+Np-Np^2 \log p$ (see (4), with

k replaced by Np). It is easily verified that 1-Npe ln p < 1+Np-Np^{2}log p iff p > exp(1/(1/ln 2 - e)) = .4566. As we assume p=k/N small, and as we compare Patel's expectation with our worst case, the preceding inequality proves the superiority of our version of SB.

(d) Jacoby and Harrison (1962)'s Sequential Bifurcation . The original version of SB starts by considering $\beta_{1,N/2}$, computed as $y_{(N/2)}$-$y_{(0)}$, and $\beta_{N/2+1,N}$, computed as the observation with $x_1=\ldots=x_{N/2}=0$ and $x_{N/2+1}=\ldots=x_N=1$ minus $y_{(0)}$. A $\beta_{i,j}$ (i<j) that is found to be positive, gives rise to two observations: one with $x_i=\ldots=x_{(i+j-1)/2}=1$, and one with $x_{(i+j+1)/2}=\ldots=x_j=1$, and both with all other independent variables equal to 0. Subtraction of $y_{(0)}$ gives $\beta_{i,(i+j-1)/2}$ and $\beta_{(i+j+1)/2,jj}$. This is repeated until the individual variables are reached. The number of observations in the worst case can be computed analogously to the derivation of (3), and is equal to $2^{\ell+1}$-1+2k(m-ℓ) if k out of 2^m parameters are positive and $2^{\ell-1}<k\leq 2\ell$ (for k=0, ℓ=1). The number of observations is about twice as much as our modification of SB needs.

APPENDIX 2. EXPECTED NUMBER OF OBSERVATIONS

Suppose each variable has an a priori probability p of being important (binomial model). Then, if δ=0, any group of n variables contains at least one important variable with probability 1-(1-p)^n. The expected number of groups of size 2^{m-j} containing at least one important variable is

$$2^j(1-(1-p)^{2^{m-j}}) \ ,$$

and this is also the expected number of observations that split groups of size 2^{m-j} into groups of 2^{m-j-1} (j=0,1,...,m-1). So the expected total number of observations is

$$2 + \sum_{j=0}^{m-1} 2^j(1-(1-p)^{2^{m-j}}) \ .$$

Because $\sum_{j=0}^{m-1} 2^j = 2^m-1$, the expected total number of observations can be rewritten as

$$1 + 2^m - \sum_{j=0}^{m-1} 2^j (1-p)^{2^{m-j}} .$$

Morris (1987, table VI) compares his Multiple Grouping technique to Watson's (1961) Two-Stage Group-Screening with respect to the expected number of runs, for some values of p and N=1024 variables. We add Sequential Bifurcation to this comparison; the results, given in table 3, are self-explanatory.

REFERENCES

BOX, G.E.P., AND J.S. HUNTER (1961a), The 2^{k-p} fractional factorial designs, Part I, *Technometrics 3*, 311-351.

BOX, G.E.P., AND J.S. HUNTER (1961b), The 2^{k-p} fractional factorial designs, Part II, *Technometrics 3*, 449-458.

JACOBY, J.E., AND S. HARRISON (1962), Multi-variable experimentation and simulation models, *Naval Research Logistic Quarterly 9*, 121-136.

KLEIJNEN, J.P.C. (1987), *Statistical tools for simulation practitioners*, Marcel Dekker, New York.

MORRIS, M.D. (1987), Two-stage factor screening procedures using multiple grouping assignments, *Communications in Statistics, Theory and Methods 16*, 3051-3067.

LI, C.H. (1962), A sequential method for screening experimental variables, *American Statistical Association Journal 57*, 455-477.

MAURO, C.A. (1984), On the performance of two-stage group screening experiments, *Technometrics 26*, 255-264.

MAURO, C.A., AND K.C. BURNS (1984), A comparison of random balance and two-stage group screening designs: a case study, *Communications in statistics, theory and methods 13*, 2625-2647.

PATEL, M.S. (1962), Group-screening with more than two stages, *Technometrics 4*, 209-217.

ROTMANS, J. (1990), *IMAGE: An Integrated Model to Assess the Greenhouse Effect*, Kluwer, Dordrecht.

WATSON, G.S. (1961), A study of the group screening method, *Technometrics 3*, 371-388.

ontvangen 6-5-1992
geaccepteerd 11-6-1992