MULTISTAGE CURVE FITTING

CHRISTOPH HAEHLING VON LANZENAUER and DON WRIGHT

INTRODUCTION

One of the most important properties of a distribution function is that it fits the data well enough for the decision-makers' or analysts' purposes. The statisticians' problem is to select a specific form for the distribution function and to determine its parameters from the available data. Various methods (graphical method, method of moments, maximum likelihood method) are available for that purpose.

In many real world situations a single distribution function, however, may not be appropriate over the entire range of the available data. This suggests that the underlying process changes over the range of the respective variable. This fact should be considered in curve fitting. A typical example of such a situation is given in Figure 1 representing third party liability losses for trucks.

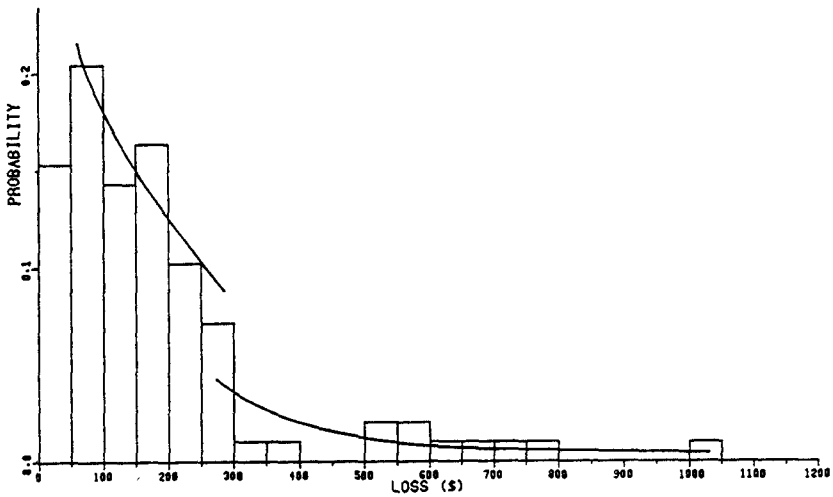

Fig. 1. Loss Distribution.

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

It is interesting to speculate about the different raisons d'être (Seal [5]) for the observed discontinuity. It may be the result of out-of-court or in-court settlements or could stem from differences between bodily injury and property damages.

To represent such data a combination or a mixture of distribution functions appears to be more appropriate. Various authors have considered this problem. While Almer [1] discusses the problem in general terms, Andreasson [2] represents the distribution of the claim size in the Swedish third party motor insurance by a sum of exponentials (exponential polynomial) and uses a graphical procedure to estimate the parameters. Coppini [3] derives the distribution of the length of sickness as the weighted sum of two gamma distributions, one referring to sick males and the other to sick females. The purpose of this paper is

(a) to present a different approach in mixing distribution functions to represent data as shown in Figure 1, and
(b) to use a computer based search procedure to determine the parameters.

Multistage Curve Fitting

Let $x(x \geq 0)$ be a random variable whose distribution function $F(X)$ has to be determined from a given set exhibiting such discontinuities. Since a single function for $F(X)$ appears to be inappropriate, one can think of $F(X)$ being composed of various expressions which are defined over specific intervals only. Let the index $k(k = 1, 2, \ldots, K)$ represent the kth interval of the random variable. We define as T_k the transition point between interval k and $k + 1$, postulate $T_k < T_{k+1}$ and set $T_0 = 0$ and $T_K = \infty$. The function representing the kth interval is defined as $g_k(x)$. Thus, the integral

$$\int_{T_{k-1}}^{T_k} g_k(x)dx \tag{1}$$

is contribution to $F(X)$. Adjustments however, must be made to (1) to insure that the sum of the integrals over all intervals equal to 1. Let α_k be the adjustment factor for interval k. Thus we can define

$$F(X) = \sum_{j=1}^{k-1} \alpha_j \int_{T_{j-1}}^{T_j} g_j(x)dx + \alpha_k \int_{T_{k-1}}^{X} g_k(x)dx \quad T_{k-1} \leq X \leq T_k \tag{2}$$

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

which satisfies

$$\sum_{k=1}^{K} \alpha_k \int_{T_{k-1}}^{T_k} g_k(x)dx = 1 \tag{3}$$

if α_k is defined as

$$\alpha_k = \begin{cases} 1 & k = 1 \\ \prod_{j=1}^{k} \dfrac{\int_{T_j}^{\infty} g_j(x)dx}{\int_{T_j}^{\infty} g_{j+1}(x)dx} & k = 2, 3, \ldots, K \end{cases}$$

or in its recursive equivalent

$$\alpha_k = \begin{cases} 1 & k = 1 \\ \alpha_{k-1} \dfrac{\int_{T_{k-1}}^{\infty} g_{k-1}(x)dx}{\int_{T_{k-1}}^{\infty} g_k(x)dx} & k = 2, 3, \ldots, K \end{cases}$$

For a given form of $g_k(x)$ the problem remaining is to determine

(a) the number of intervals K,
(b) the transition points T_k, and
(c) the parameters of the distribution to represent the kth interval.

The values selected depend of course on the criterion used in the curve fitting process. Various criteria are available with the squared sum of the error being used most frequently. The squared sum of the errors can be defined by

$$S = \sum_{i=1}^{N} [F(X = y_i) - (i/N)]^2 \tag{5}$$

with y_i $(i = 1, 2, \ldots, N)$ being the ith observation and $y_i \leq y_{i+1}$. Since accuracy in the tail areas appears to be of relevance in the evaluation of risk, heavier weights of the errors in the tails may be appropriate. It will be shown below that the suggested multistage process improves specifically the fit in the tails without using any arbitrarily assigned weights. Furthermore, for premium calculations it seems that the mean of the fitted distribution should be

as close as possible to the sample mean, $\bar{y}$. Thus we can augment the criterion of minimizing the squared

$$\sum_{k=1}^{K} \alpha_k \int_{\tau_{k-1}}^{\tau_k} x\, g_k(x)dx = \bar{y} + d_1 - d_2 \tag{6}$$

with

$$d_1, d_2 \leq c$$

d_1 and d_2 are tolerances which must be less than or equal to a managerially determined level c. Of course c can be zero.

Thus the problem is to determine optimally the above parameters using a given criterion. Although a number of methods are available for solving optimization problems, the success of any one method depends on the problem. Because of the existing discontinuities in the response surface a multidimensional search technique will be used for determining all parameters. An excellent discussion of search techniques can be found in Wilde [6].

Pattern Search

The search method to be used here has been developed by Hooke and Jeeves [4] and is known as pattern search. Their method takes advantage of the fact that most response surfaces have one or more ridges which lead to the optimum. Thus the purpose is to find a ridge and follow it to the optimum. In pattern search the search begins by exploring the response surface in the vicinity of a randomly or otherwise selected base point. With repeated success the explorations become longer taking advantage of an established pattern. Failure to improve the criterion, however, indicates that one must abandon the old pattern and try to find a new one which will be followed until the pattern is broken again and the process has to be repeated. The so determined pattern will coincide with the ridge. In the neighbourhood of the optimum, the steps become very small to avoid overlooking any promising directions. The optimum is reached and the search terminates when the predetermined final step size fails to improve the criterion. Repeated searches from different starting points reduce the likelihood of the optimum being a local extreme point. The ideas of pattern search are exemplified for a two dimensional search problem in the Appendix.

ILLUSTRATIONS

The multistage curve fitting is illustrated by two examples. Both examples come from the authors' experience in analysing insurance problems for a company operating a large fleet of vehicles. Various distributions can be used to present $g_k(x)$ and there is no restriction to use the same distributions for all intervals k. For the purpose of these examples, $g_k(x)$ was chosen to be exponential for all intervals with parameter λ_k, since it appeared appropriate and easy to integrate.

Example 2

This example consists of 75 data points representing collision claims for cars during 1969/70. The data are exhibited in Figure 2 by asterisks and have a mean of $\bar{y} = \$\ 363.13$. The optimal values of the parameters of the distribution function $F(X)$ with the squared sum of the errors and the mean of $F(X)$ resulting from the pattern search are given in Table 1. The initial step size for $\lambda_k = .0005$ and for $T_k = \$\ 50.00$ while the final step size is .00001 and $\$\ 1.00$ respectively.

TABLE 1

Results: *Example 1*

	Number of Stages (K)			
	$K = 1$	$K = 2$	$K = 3$	$K = 4$
λ_1	.003633	.002363	.002148	.002187
T_1	—	\$ 35.94	\$ 54.69	\$ 5.62
λ_2	—	.004039	.004969	.004996
T_2	—	—	\$ 243.75	\$ 199.99
λ_3	—	—	.001871	.002906
T_3	—	—	—	\$ 453.12
λ_4	—	—	—	.001402
S	.13684	.11688	.02890	.01948
$\bar{x}$	\$ 275.26	\$ 261.88	\$ 346.86	\$ 368.27

The number of transition points K is determined similar to the multiple regression model. The value of K will be increased as long as a "worthwhile" improvement in S justifies doing so. Figure 2 illustrates the distribution functions for $K \leq 4$ indicating

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

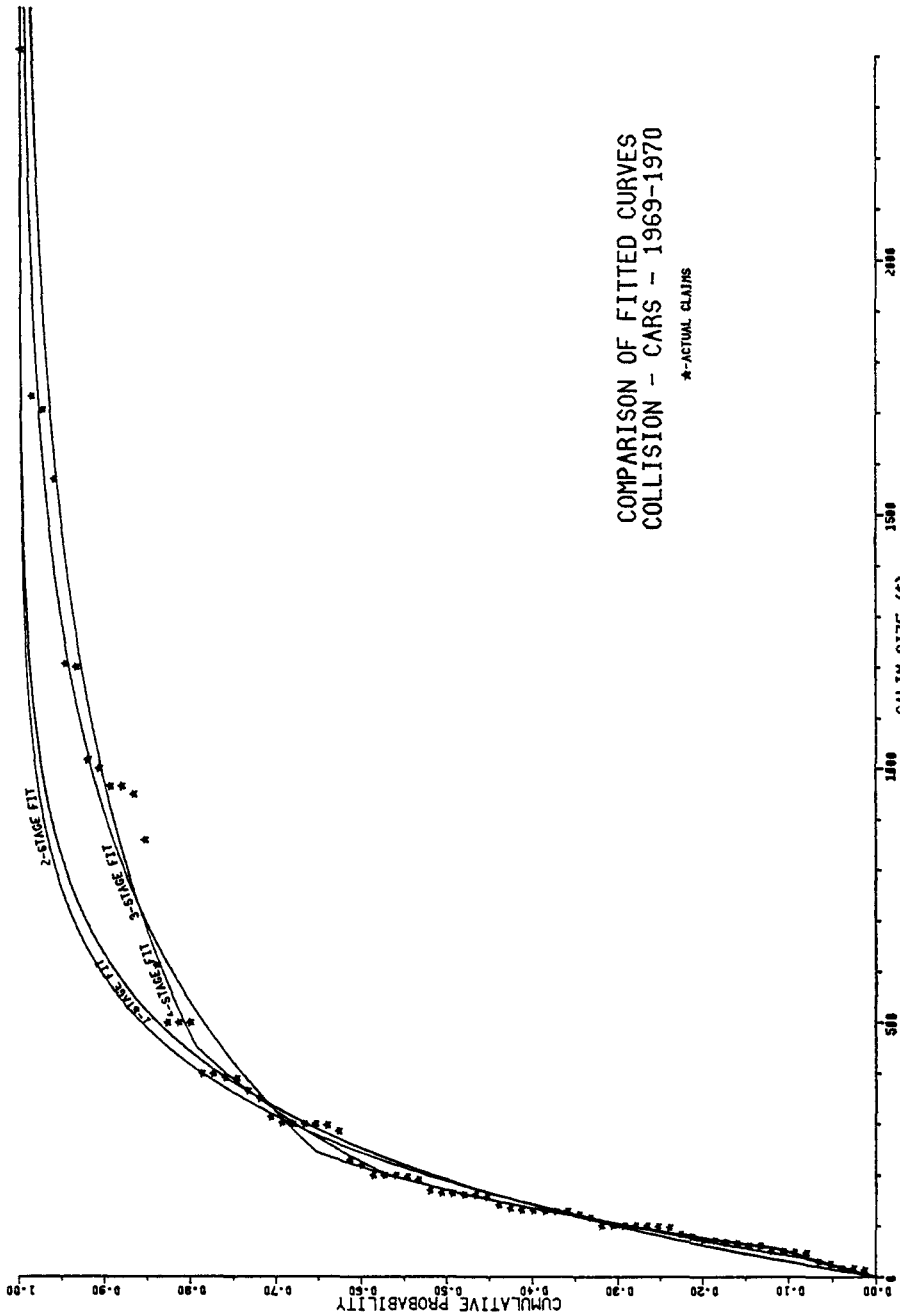

Fig. 2.

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

that the multistage process clearly improves the fits. Furthermore it is interesting to note that the improvements take place primarily in the right tail. While the means of the fitted distribution functions for smaller values of K deviate substantially from $\bar{y}$, $\bar{x}$ approaches $\bar{y}$ reasonably closely for $K \geq 3$.

Example 2

The data in the second example are 98 third party liability losses for trucks during 1970/71. The data are exhibited in Figure 3 by asterisks and have a mean of \$ 399.49. A first run of the pattern search using the same step sizes as in Example 1 resulted in means of the fitted distribution functions $\bar{x}$ as given in Table 2 which are too far off from $\bar{y} =$ \$ 399.49. Thus the criterion of minimizing

TABLE 2

Means of the Fitted Distribution Functions

	Number of Stages (K)			
	$K = 1$	$K = 2$	$K = 3$	$K = 4$
$\bar{x}$	\$ 198.44	\$ 171.89	\$ 254.63	\$ 256.05

the squared sum of errors was augmented by (6) with $d_1 = d_2 = 0$. Of course this implies that the number of degrees of freedom is reduced by one. The parameter determined as a result of the others was selected to be λ_K. Table 3 summarizes the results.

TABLE 3

Results: Example 2

	Number of Stages (K)			
	$K = 1$	$K = 2$	$K = 3$	$K = 4$
λ_1	.002503	.005081	.004280	.004280
T_1	—	\$ 448.64	\$ 134.36	\$ 135.93
λ_2	—	.000492	.008851	.009163
T_2	—	—	\$ 307.80	\$ 254.68
λ_3	—	—	.0004899	.001839
T_3	—	—	—	\$ 1,148.30
λ_4	—	—	—	.0002099
S	3.5262	.22751	.07630	.06804

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

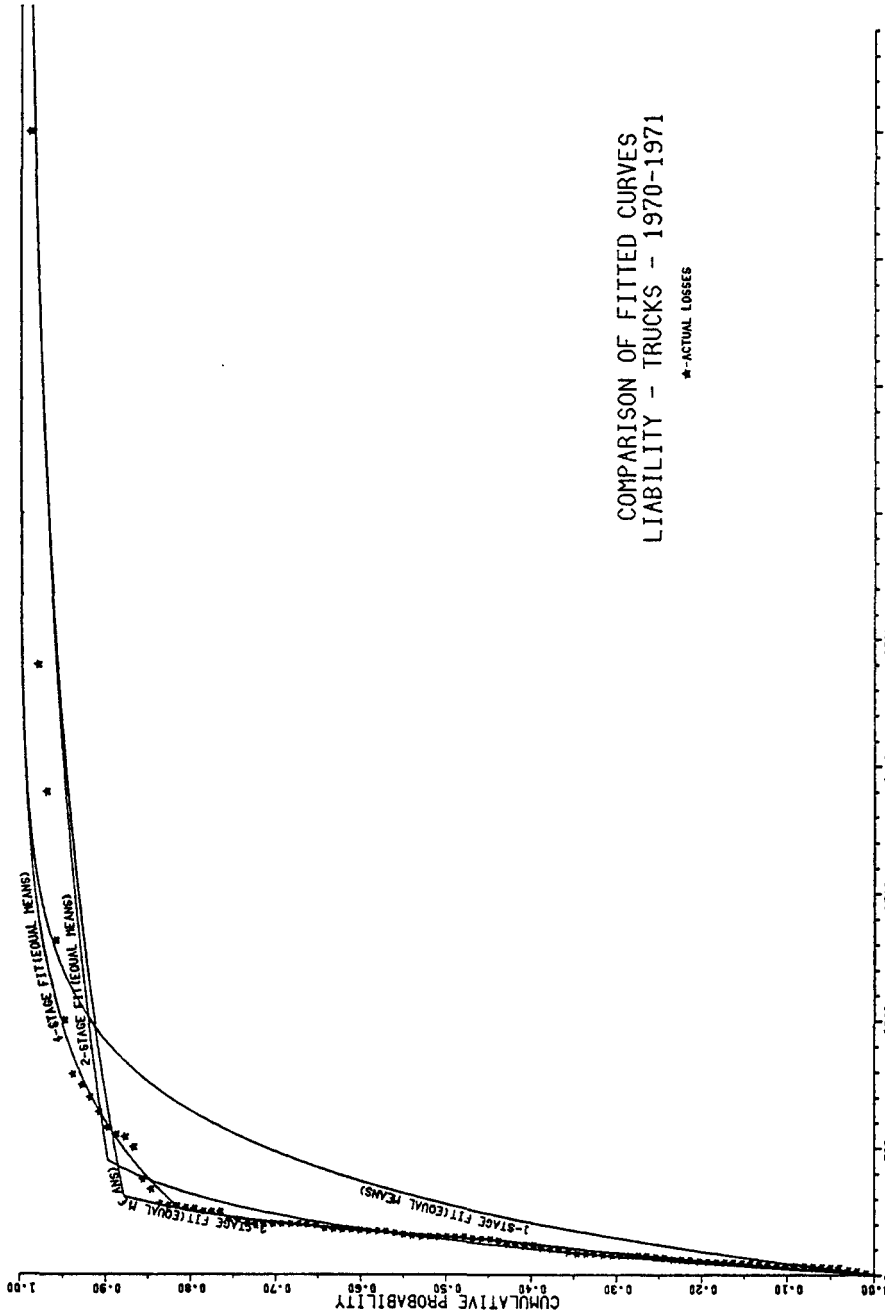

Fig. 3.

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

Figure 3 again illustrates the distribution functions for $K \leq 4$. Again considerable improvements resulted over a one stage fit.

Appendix

The concept of pattern search is explained and illustrated for the two stage fit with equal means of example 2. The example has two independent parameters, the exponential parameter λ_1, and the transition point T_1. The exponential parameter λ_2 is determined by λ_1, T_1 and the restriction of equal means. The contour lines of the response surface expressed by the squared sum of errors for values of the independent variables λ and T are plotted in Figures 4 and 5.

The search is illustrated in Figures 4 and 5 with solid lines representing successful perturbation and pattern moves while broken lines indicate perturbations and pattern moves which fail to improve the objective function. The search begins by exploring the response surface at base point $B_1 = H_1$ through changes in the transition point in T (Figure 4). An improvement in the criterion leads to a temporary head $h_1(T)$. From here local explorations through changes in λ lead to $h_1(T, \lambda)$ and the second base point B_2 since only two independent variables exist. Reasoning that another perturbation about B_2 would produce similar results, one creates a new temporary head H_2 by adding the vector $B_1 B_2$ to Point B_2. This represents a pattern move. Local explorations about H_2 produce B_3. As above local explorations about B_3 are omitted and a new temporary head H_3 is determined by adding the vector $B_2 B_3$ to point B_3. As can be observed from Figure 4, H_3 fails to improve the criterion. The pattern is broken and local explorations must take place at B_3 which lead to B_4 and via a new pattern eventually to the temporary head H_8. At H_8 the pattern is broken again and local explorations about B_8 must resume which lead via pattern moves to H_{11} (Figure 4). This process is continued with reduced step sizes and illustrated in Figure 5. The optimum, B_{16}, is reached when perturbations with the predetermined minimum step size fail to improve the criterion. Repeated searches from different initial base points should be performed to insure the optimum is a global optimum.

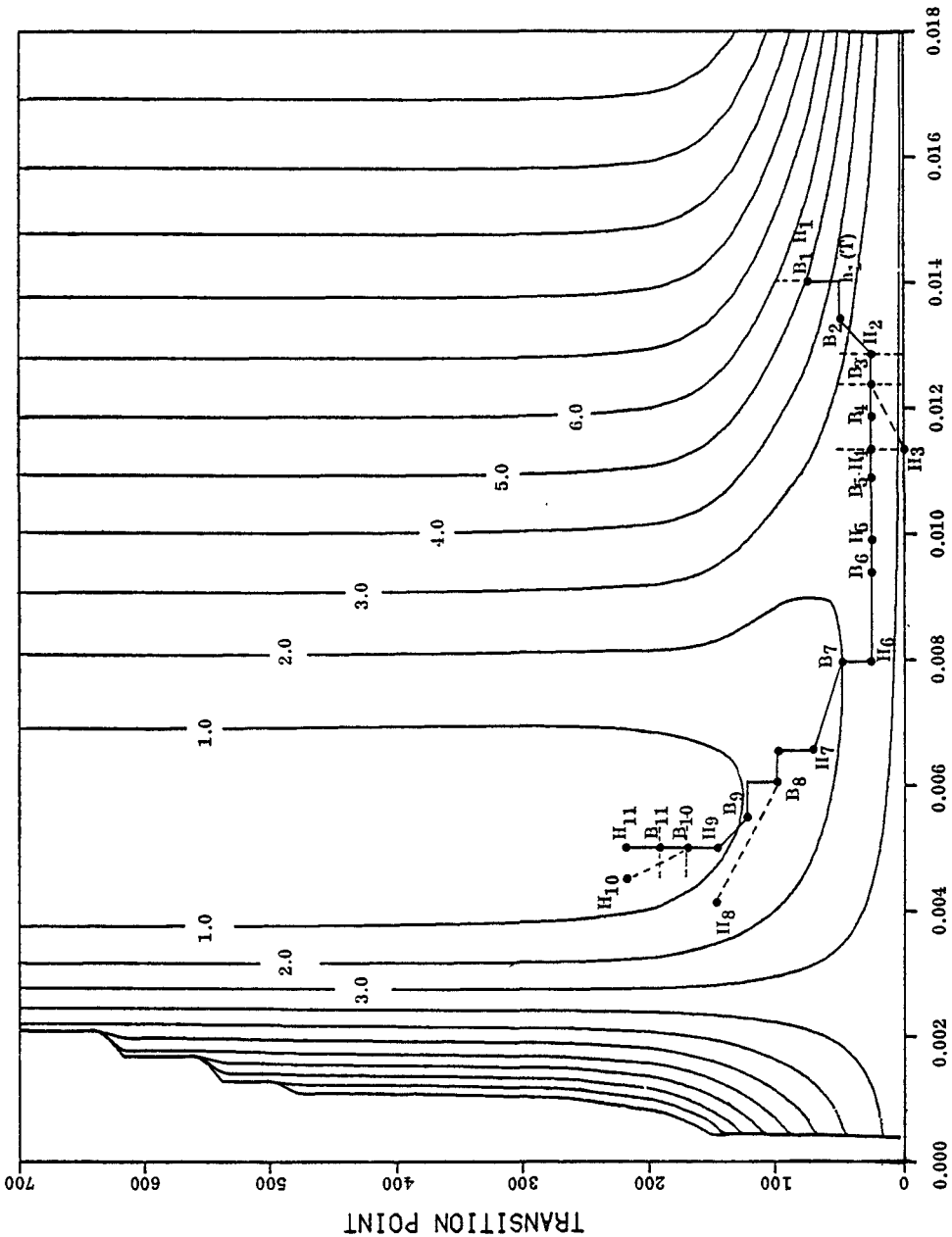

Fig. 1 Pattern Search

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

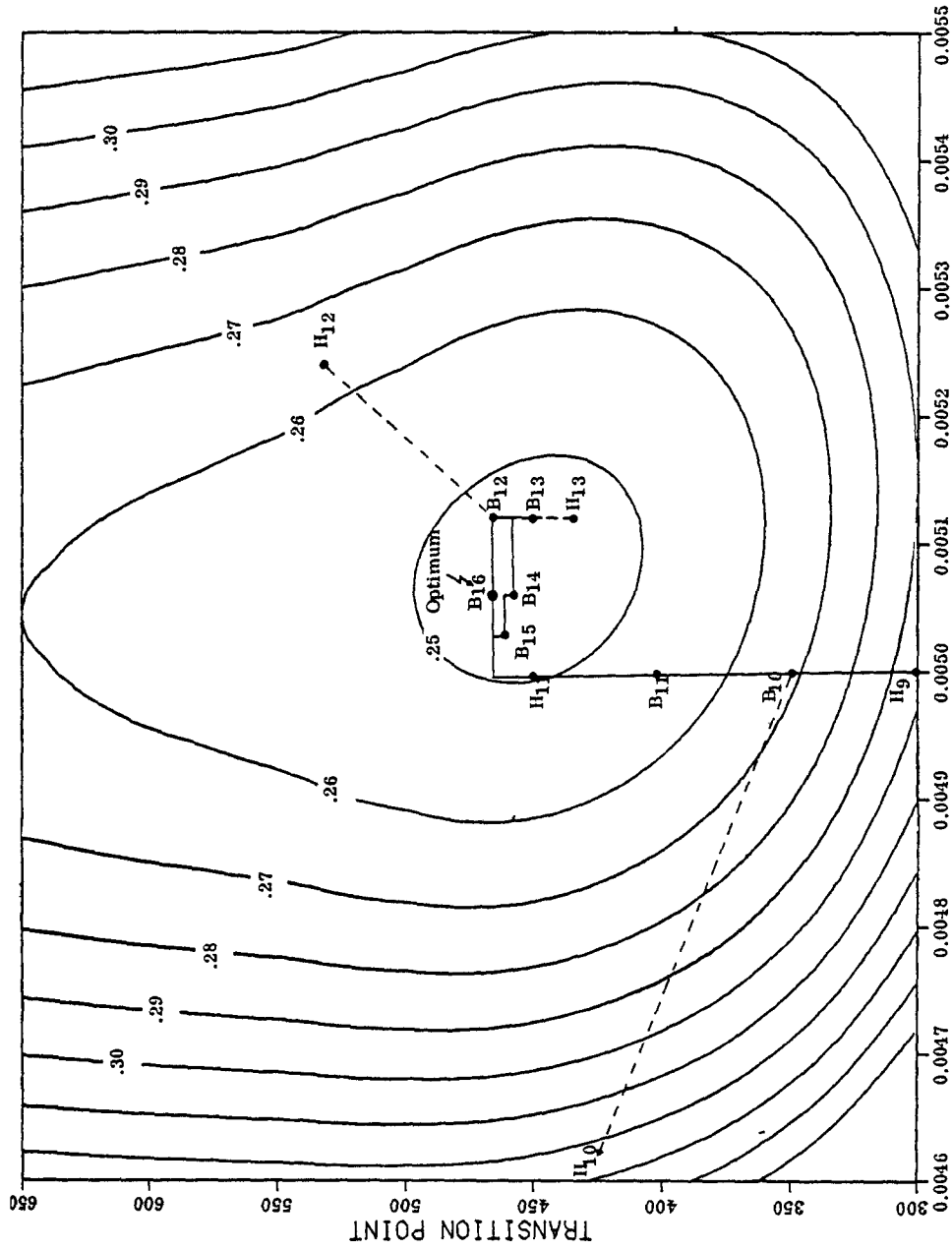

Fig. 5. Pattern Search.

https://doi.org/10.1017/S0515036100011491 Published online by Cambridge University Press

The authors gratefully acknowledge the support of this research from the Associates Fund, School of Business Administration, The University of Western Ontario.

References

[1] Almer, B., Risk Analysis in Theory and Practical Statistics, *Trans. XV International Congress of Actuaries, New York*, 1957, 2, pp. 314-349.
[2] Andreasson, G., Distribution for Approximations in Applied Risk Theory, *The ASTIN Bulletin*, 1966, 4, pp. 11-18.
[3] Coppini, M., A Propos de la Distribution des cas de Maladie Entre les Assurés et par Rapport à la duree, *The ASTIN Bulletin*, 1963, 2, pp. 45-61.
[4] Hooke, R., and T. A. Jeeves, Direct Search Solution of Numerical and Statistical Problems, *J. Assoc. Comp. Mach.*, 1961, 8, pp. 212-229.
[5] Seal, H., *Stochastic Theory of a Risk Business*, Wiley, 1969.
[6] Wilde, D., *Optimum Seeking Methods*, Prentice Hall, 1964.