

A Feature Preserving Smoother with Application
to the Coal-Mining Disaster Data of Britain

by

Shean-Tsong Chiu¹

Technical Report 87-1, January 1987

¹Mathematical Sciences Department, Rice University, Houston, Texas 77251.

A Feature Preserving Smoother with Application to the Coal-Mining Disaster Data of Britain

By Shean-Tsong Chiu

Rice University

Abstract

The problem of estimating a discontinuous mean function was studied. A feature preserving smoothing procedure was proposed. The procedure can preserve the discontinuities of the function and detect outliers in the observations. The procedure was applied to analyze the famous data set of the coal-mining disasters of Britain.

Key Words: Change-point, Coal-mining disasters of Britain, Discontinuity, Outlier.

1. Introduction

We consider that we have a sequence of independent random variables $X(1), \dots, X(T)$. The random variables are often represented in the form

$$X(t) = \mu(t) + \epsilon(t).$$

Here $\mu(t)$ is the mean function of $X(t)$ and $\epsilon(t)$ is a sequence of independent random errors with mean zero and variance $\sigma^2(t)$. In this paper we are interested in the problem of estimating the mean function $\mu(t)$ which might be discontinuous at some points. Since t is discrete, we cannot have a rigorous definition of discontinuity. When we say that a function is discontinuous at some points, we mean that the function changes rapidly at these points. It is of interest to estimate the locations of the discontinuities because the discontinuities are often caused by the occurrence of events, and the locations of the discontinuities provide important information about the impact of these events. We remark that sequences of observations which contain outliers can also be viewed as processes with discontinuous mean functions.

Most statistical research concerning estimating the discontinuous function takes the parametric approach. The mean function is assumed to be a function of some unknown parameters. The locations of the discontinuities are treated as unknown parameters. The change-point problem, which assumes that the mean function is a step function, is a particular case of this approach. Zacks (1983) surveyed the research concerning the change-point problem. Some recent studies on the change-point problem can be found in Siegmund (1986), Raftery and Akman (1986), Worsley (1986), Yao (1984), and Pollak and Siegmund (1985). Most of these only consider the case of a step function with a single jump. Siegmund (1986) discussed the case of two jumps. Hinkley (1970) considered using functions which is more general than step functions to fit the mean function.

For the nonparametric methods, the classical kernel estimator is used most often for estimating a smooth mean function. The estimate of $\mu(t_0)$ is a weighted average of the random variables in the window centered at t_0 . The weights are assigned according to the geometric distances without considering whether the neighboring points are in the same region of t_0 . Therefore, the kernel estimator will smooth (blur) the discontinuities and make it difficult to detect and locate the discontinuities accurately. Intuitively, if we knew the locations of the discontinuities, we would not include the observations from different regions to estimate the mean function. In practice we usually do not know the locations of discontinuities, however, we can classify the window into regions and estimate $\mu(t_0)$ from the observations in the region containing t_0 .

Based on this idea, we propose a procedure for estimating discontinuous mean functions (Section 2). The smoothing procedure has the ability of preserving the discontinuities. In Section 3, we apply this procedure to smooth the famous coal-mining disasters data. The result is compared with the results obtained from a classical kernel smoother and a robust smoothing procedure. The estimates of the locations of the discontinuities are compared with those obtained by Worsley (1986) and Raftery and Akman (1986). In Section 4, The significance of the discontinuities is studied. We remark that this procedure was originally developed in Chiu (1985) for smoothing noisy images.

2. The Feature Preserving Smoother

We wish to estimate $\mu(t_0)$ from the observations in a window with length l . The window is usually centered at t_0 . We first classify the window into regions and then estimate $\mu(t_0)$ by a weighted average over the points in the same region of t_0 . The basic idea of the classification procedure is to merge the neighboring regions with minimum "distance" until there are only a few regions. The merger process is terminated when the distances between neighboring regions are too large.

We now describe the procedure.

Step 1. At the beginning, each point is a region.

Step 2. The distances between neighboring regions are computed. The distance between two neighboring regions R_1 and R_2 is defined as

$$d = |\bar{x}_1 - \bar{x}_2| / \sqrt{(1/n_1) + (1/n_2)},$$

where \bar{x}_k is the sample mean of the region R_k and n_k is the number of points in the region R_k .

Step 3. Find the minimum distance d_{\min} among all neighboring regions.

Step 4. If the number of regions is bigger than L , go to Step 6, otherwise go to Step 5. The number L is chosen according to the complexity of the mean function. For most applications, $L=3$ or 4 is big enough.

Step 5. When there are few regions, we compute the statistic $D_K = d_{\min}/s$, s^2 is the mean squares within regions,

$$s^2 = \sum_{k=1}^K \left\{ \sum_{t=t_k}^{t_{k+1}-1} [X(t) - \bar{x}_k]^2 / (T-K) \right\},$$

where K is the number of regions and the k th region $R_k = \{t_k, \dots, t_{k+1}-1\}$. We note that D_K is scale and location invariant. The statistic D_K tends to be large when the means of

the regions are different. So, if D_K is bigger than c_K , we stop merging regions and go to Step 7, otherwise, go to Step 6. The critical values c_K are determined according to how much chance of finding a false discontinuity one is willing to tolerate.

Step 6. The neighboring regions with minimum distance are merged. If there is only one region now, go to Step 7, otherwise go to Step 2 and repeat the merger process.

Step 7. Estimate $\mu(t_0)$ from the observations which are in the same region of t_0 .

Though the value of c_K can be set by trial-and-error, we would prefer to set the value as the upper α point of the null distribution of D_K . Due to the complexity of the procedure, it is difficult to obtain the exact null distribution. However, we could find an approximate value by a Monte Carlo method.

3. Applied to the Coal-Mining Disasters Data

The feature preserving procedure described in the previous section is applied to the data set of the intervals between coal-mining disasters. This data set was originally given by Maguire *et al* (1952). Jarrett (1979) corrected and extended the data set to cover 191 disasters between 1851 and 1962. More detailed information can be found in Jarrett (1985). This data set has been extensively studied. The intervals have often been modeled as a sequence of independently and exponentially distributed random variables with a variable mean function $\mu(t)$. Barnard (1953), Cox and Lewis (1966), and Jarrett (1979) fitted the mean function by a smooth function with few parameters. It has been suspected that the data might have discontinuities (A.C. Atkinson, in the discussion of the paper of Leonard (1978)). Recently, Raftery and Akman (1986) and Worsley (1986) used step functions (with one jump) to model the mean function. We should point out that Barnard (1953), Cox and Lewis (1966) and Worsley (1986) used the data set given in Maguire (1952). In the following discussion, we will use the convention that the j th observation is the number of days between the $(j-1)$ th and the j th disasters.

The data and the curve obtained by the feature preserving smoother are plotted in Figure 1. The length of the window is 21 and $L=4$. The critical values used are $c_2=4.16$, $c_3=6.31$ and

$c_4=3.82$. These values are the approximate 95th percentiles of the statistics D_2 , D_3 and D_4 under the hypothesis that the observations in the window are identically independently exponentially distributed. These values are obtained from the empirical distribution of 4000 simulations. The estimate of $\mu(t_0)$ is the weighted average over the the region containing t_0 . The weights are proportional to a triangular kernel function $w(t) = 11-|t|$, $|t| \leq 10$. From Figure 1, we see that the mean function is quite constant before 120 and begins to increase gradually after 120. Except for the spike at 182 and the sharp increase at 187, the mean function is gradually decreasing after the peak at 153. The 182nd disaster occurred in December 1986. There are two big events around this date. The Second World War ended in 1945 and the coal industry was nationalized on January 1, 1947 (Saxena (1955)). There is another spike at 14; the 14th disaster occurred in May 1856.

Figure 2 shows the resultant curve of a classical kernel smoother. The smoother use the same window length and weighting function as the ones used in the feature preserving smoother. Figure 3 displays the result of applying the robust smoother of Cleveland (1979). This procedure is available in Becker and Chambers (1984). The fraction of data used for smoothing at each t point is 0.11, given the same window length as the one used above. From Figures 2 and 3, we see that the discontinuity at 182 is smoothed by these two procedures. The robust procedure also fails to show the two modes at 135 and 153.

Figure 4 illustrate the effects of using smaller critical values in the feature preserving smoother. The critical values are $c_2=2.95$, $c_3=4.74$ and $c_4=3.82$. c_2 and c_3 are the approximate 85th percentiles and c_4 is the approximate 95th percentile. Same window length and kernel function are used. Figure 4 shows a patch from 134 to 137 and another spike at 41.

Before going to the next section to discuss how to test the significance of the discontinuities found, we would like to indicate the locations of the discontinuities estimated by Worsley (1986) and Raftery and Akman (1986). The disasters which are close to the estimated locations of the discontinuities are also indicated.

1) March 1890, the 124th disaster, the estimate obtained in Raftery and Akman (1986) by using a Bayesian approach. The reason cited in the paper is the emergence of the Miners' Federation at

the end of 1889. The feature preserving procedure did not find a sudden change around this date (though a small discontinuity at 122 was observed). From Figures 1 and 4, it seems that the emergence may have had a gradual, instead of abrupt, effect on the safety of coal-mining.

2). August 1891, the 126th disaster. The estimate obtained in Worsley (1986). The reason cited is that two Royal Commissions investigated coal-mining accidents and reported in 1886 and 1894. They concluded that the coal dust is the main cause of explosions and recommended wetting the sides and floors to keep down dust, and using blasting powder that produces less flame than gunpowder does. However, it is difficult to explain why the discontinuity occurs in the middle of these two reports.

3). April 1896, the 133rd disaster. From Figure 4, we find a discontinuity here. In this year, the recommendations mentioned above were implemented in the Explosive in Coal Mines Order of 1896.

Remark 1: From Figures 1 to 4, it can be seen that the model of a step mean function is not proper for the data set. We fitted a second order polynomial to the data in the interval from 125 to 181, and obtained the maximum likelihood estimates of the coefficients. We test the hypothesis of a constant mean function in this interval by the estimates. The approximate p-value is less than 0.0002.

Remark 2: This example illustrates that the testing procedures for the change-point problem are very likely to reject the null hypothesis when the mean function is smoothly increasing (or decreasing). It also shows that using an improper model could mislead us to wrong conclusions.

4. Test of Discontinuity

In order to confirm the discontinuities found by the feature preserving smoother, we have to test the significance of the discontinuities. In this section we consider a sequence of independently exponentially distributed random variables $X(t)$. If the mean function is discontinuous at t_0 , we will expect that, in the window $[t_0-l_1+1, t_0+l_2]$, the statistic

$$U(t_0) = \left\{ \sum_{i=t_0-l_1+1}^{t_0} X(i)/l_1 \right\} / \left\{ \sum_{i=t_0+1}^{t_0+l_2} X(i)/l_2 \right\}$$

will be very large or very small. Under the null hypothesis that the mean function is constant in the window, $U(t_0)$ has a F distribution with $2l_1$ and $2l_2$ degrees of freedom. Since we did not know t_0 and we are looking for discontinuity over the whole data set, we should compare the observed values of $U(t)$ at the possible discontinuous points with the quantiles of the distribution of the statistic

$$\bar{U} = \max_t \{ \max U(t), \max 1/U(t) \}$$

under the hypothesis that the mean function is constant.

$U(t_0)$ is not powerful for the alternatives that there is a small cluster (patch) of observations which are significantly larger than their neighbors. To test this kind of discontinuity, we consider the statistics

$$V(t_0) = \left\{ \sum_{i=t_0-k_1+1}^{t_0+k_2} X(i)/k \right\} / \left[\left\{ \sum_{i=t_0-l_1+1}^{t_0-k_1} X(i) + \sum_{i=t_0+k_2+1}^{t_0+l_2} X(i) \right\} / (l-k) \right]$$

where $k=k_2-k_1$ is the length of the patch and $l=l_2-l_1$ is the length of the window. Under the null hypothesis of constant mean, $V(t_0)$ has F distribution with $2k$ and $2(l-k)$ degrees of freedom. Similar to the discussion above, we should compare the observed values with the quantile of the null distribution of the statistic

$$\bar{V} = \max_t V(t).$$

The exact null distributions of \bar{U} and \bar{V} are not easy to find. We will show in the appendix that the distributions can be approximated by the distribution of the maximum of a sequence of independently identically distributed random variables with the corresponding F distribution. Therefore we have

$$P\{\bar{V} > u\} \approx 1 - [P\{V(t_0) > u\}]^{190}$$

and

$$P\{\max_t U(t) > u\} \approx 1 - [P\{U(t_0) > u\}]^{190}.$$

We now check how significant are the discontinuities found in the previous section. The results are shown in Table 1. In the table, we use $S_{p,q}$ to denote the sum of the variables $X(t)$ for t from p to q . The first and second columns describe the numerators and the denominators of the statistics, the third column is the observed values of these statistics, the fourth column is the asymptotical p-values and the fifth column is the approximate p-values obtained from 2000 simulations. From the table, we see that for small p-values, the asymptotic values agree well with the approximate values obtained from simulations. The discontinuity at 187 is tested in the fifth and sixth rows. The denominator in the sixth row does not include the 182nd observation. The p-value of the last row is obtained from the empirical distribution of 2000 simulations of the variable

$$\bar{V} = \max_t \bar{V}(t),$$

where

$$\bar{V}(t_0) = \left\{ \sum_{i=t_0-k_1+1}^{t_0+k_2} X(i)/k \right\} / \left\{ \left(\sum_{i=t_0-l_1+1}^{t_0-k_1} X(i) + \sum_{i=t_0+k_2+1}^{t_0+l_2} X(i) \right) - X_{\max} \right\} / (l-k-1),$$

and X_{\max} is the maximum of $X(t)$ on $\{t_0-l_1+1 \leq t \leq t_0-k_1\} \cup \{t_0+k_2+1 \leq t \leq t_0+l_1\}$.

From the results, we conclude that if the mean function is smooth, it is very unlikely to have such sharp change as the one at 187, and the chance of having an observation as extreme as the 14th observation is less than 10 percent. The other discontinuities found in Figure 4 are not significant.

5. Summary

We consider the problem of estimating the mean function of a sequence of independent random variables. We are interested in the situation that the mean function might be discontinuous at some points. The classical estimates have difficulty in estimating a discontinuous function; the discontinuities are smoothed and the estimates at the points close to the discontinuities are biased. Therefore, the classical methods might not be able to help us find information concerning the discontinuities. To solve this difficulty, we propose a feature preserving smoothing procedure. The procedure preserves the discontinuities of the mean function and provides better estimates at

the points close to the discontinuous points. We applied this procedure and two classical methods to the data set of the coal-mining disasters of Britain. The results clearly demonstrate the advantage of the feature preserving smoother. The procedure can also detect outliers in the observations. We will generalize this procedure to estimate discontinuous regression functions in a further paper.

Acknowledgement

This research was supported in part by ARO Grant DAAG 29-85-K-0212 and ONR Grant N0001-485-K-0100.

Appendix

The Asymptotic Distributions of the Test Statistics

We apply Theorem 3.5.1 of Leadbetter *et al.* (1982) to find the asymptotic distribution of the maximum of the process $U(t)$. Similar result can be obtained for $V(t)$.

Suppose $Z(i)$, $i=1, \dots$, is a stationary sequence and F is the distribution of $Z(i)$. Define $M_n = \max\{Z(1), \dots, Z(n)\}$ and $\hat{M}_n = \max\{\hat{Z}(1), \dots, \hat{Z}(n)\}$, where $\hat{Z}(i)$ is a sequence of independent random variables with the same distribution function F . The result of Leadbetter *et al.* (1982) showed that M_n and \hat{M}_n have the same asymptotic distribution if, for any sequence u_n such that $\text{pr}(\hat{M}_n \leq u_n)$ converges to $\theta > 0$, the conditions $D(u_n)$ on page 53 and $D'(u_n)$ on page 58 are satisfied. Since $U(t_1)$ and $U(t_2)$ are independent when $|t_1 - t_2| \geq l$, it is clear that the condition $D(u_n)$ is satisfied. The condition $D'(u_n)$ holds if we can show that, for any $j \leq l$, $n \text{pr}\{U(t_1) > u_n, U(t_1 + j) > u_n\}$ converges to zero. We now proceed to prove this. In the following discussion, the notations M_n and \hat{M}_n denote the maxima of the sequences $U(t)$ and $\hat{U}(t)$.

Since $U(t)$ has a F distribution with $2l_1$ and $2l_2$ degrees of freedom, $\text{pr}(U(t) > u)$ is of order u to the power $-l_2$ as u goes to infinity. Also, for large u ,

$$\text{pr}(\hat{M}_n \leq u) = [1 - \text{pr}\{V(t) > u\}]^n \approx 1 - n \text{pr}(V(t) > u).$$

Hence $\text{pr}(\hat{M}_n \leq u_n)$ converges for u_n of order n to the power $1/l_2$. To simplify the argument, we

introduce some notations. Letting $t_1 < t_2$ be two integers and $1 \leq t_2 - t_1 < l_1$, define

$$\begin{aligned} Y_1 &= X(t_1 - l_1 + 1) + \cdots + X(t_2 - l_1), \\ Y_2 &= X(t_2 - l_1 + 1) + \cdots + X(t_1), \\ Y_3 &= X(t_1 + 1) + \cdots + X(t_2), \\ Y_4 &= X(t_2 + 1) + \cdots + X(t_1 + l_2), \\ Y_5 &= X(t_1 + l_2 + 1) + \cdots + X(t_2 + l_2). \end{aligned}$$

Also define $W_1 = (Y_1 + Y_2) / (Y_3 + Y_4)$ and $W_2 = (Y_2 + Y_3) / (Y_4 + Y_5)$. Note that $U(t_1) = W_1(l_2/l_1)$ and $U(t_2) = W_2(l_2/l_1)$.

The conditional probability of $\{W_1 > u \text{ and } W_2 > u\}$ relative to $\{Y_1, Y_2, Y_3 \leq (\log n)^2/2\}$ is smaller than the conditional probability of $\{(Y_3 + Y_4) \leq (\log n)^2/u \text{ and } (Y_4 + Y_5) \leq (\log n)^2/u\}$ relative to $Y_3 \leq (\log n)^2/2$ which is equal to

$$\int_0^w \text{pr}\{Y_3 \leq w - Y_4 \mid Y_3 \leq (\log n)^2/2 \text{ and } Y_4 = y\} \text{pr}\{Y_5 \leq w - Y_4 \mid Y_4 = y\} f_4(y) dy \quad (1)$$

where $w = (\log n)^2/2$ and $f_4(y)$ is the density function of Y_4 . (1) is less than

$$\text{pr}(Y_3 < w \mid Y_3 \leq (\log n)^2/2) \text{pr}(Y_4 < w) \text{pr}(Y_5 < w)$$

which is of order w to the power $l_2 + t_2 - t_1$ as u goes to infinity. Hence (1) is of order $o(1/n)$ for u_n of order n to the power $1/l_2$. From this result and the fact that $\text{pr}\{Y_i > (\log n)^2/2\}$ is of order $n^{-(\log n)/2}$ for $i=1,2,3$, it is clear that, for $1 \leq t_2 - t_1 < l_1$, $n \text{pr}\{U(t_1) > u_n, U(t_2) > u_n\}$ converges to zero for u_n of order n^{1/l_2} . Following the same argument, we can obtain a similar result for $l_1 \leq t_2 - t_1 < l$ and the proof is finished.

References

- [1] Barnard, G. A. (1953). Time intervals between accidents-a note on Maguire, Pearson and Wynn's paper. *Biometrika* 40, 211-213.
- [2] Becker, R.A. and Chambers, J.M. (1984) *S: An interactive environment for data analysis and graphics*, Wadsworth, Belmont, California.
- [3] Chiu, S. T. (1985). Smoothing Noisy Images. *Proceedings of the 1985 Conference on Applied Analysis (Mathematics & Statistics) in Areospace, Industry and Medical Sciences*, 131-136.

- [4] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *JASA*, **74**, 829-836.
- [5] Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events*. Methuen, London.
- [6] Hinkley, D. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57**, 1-17.
- [7] Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika* **66**, 191-193.
- [8] Jarrett, R. G. (1985). Coal-mining disasters, in *Data: A collection of problems from many fields for the student and research worker*, ed. D.F. Andrews and A.M. Herzberg, Springer-Verlag, New York, 51-56.
- [9] Leadbetter, M.R., Lindgren, G. and Rootzen, H. (1982). *Extremes and related properties of random sequences and processes*. Springer-Verlag, New York.
- [10] Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. R. Statist. Soc. B*, **40**, 113-146.
- [11] Maguire. B. A., Pearson, E. S. and Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika* **39**, 168-180.
- [12] Pollak, M. and Siegmund, D. (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika* **72**, 267-280.
- [13] Raftery, A. E. and Akman V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73**, 85-89.
- [14] Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *Ann. Statis.* **14**, 361-404.
- [15] Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* **73**, 91-104.

- [16] Yao, Y. C. (1984), Estimation of a noisy discrete-time step function: Bayes and empirical Bayes. *Ann. Statis.* **12**, 1434-1447.
- [17] Zacks, S. (1983). Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation. 245-269, in *Recent Advances in Statistics*. ed. Rizvi, M. H., Rustagi, J. and Siegmund D., Academic Press, New York.

Table 1. Test of Significance of Discontinuities

Z_1	Z_2	Z_1/Z_2	p-value	
			Asymptotic	Monte Carlo
$S_{14,14}/2$	$(S_{4,13}+S_{15,24})/40$	9.605	0.072	0.076
$S_{41,41}/2$	$(S_{31,40}+S_{42,51})/40$	4.439	0.969	0.982
$S_{134,137}/8$	$(S_{125,133}+S_{138,146})/34$	3.573	0.547	0.422
$S_{182,182}/2$	$(S_{166,181}+S_{183,186})/40$	8.624	0.136	0.133
$S_{187,190}/8$	$S_{170,186}/34$	5.150	0.055	0.045
$S_{187,190}/8$	$(S_{170,181}+S_{183,186})/32$	7.720		0.009

Figure 1. The intervals (in days) between coal-mining disasters. The line is the estimated mean function obtained by the feature preserving smoother, the critical values are $c_2=4.16$, $c_3=6.31$ and $c_4=3.82$.

Figure 2. The intervals (in days) between coal-mining disasters. The line is the estimated mean function obtained by the kernel smoother.

Figure 3. The intervals (in days) between coal-mining disasters. The line is the result obtained by the robust smoother of Cleveland (1979), the fraction of data used at each point is 0.11.

Figure 4. The intervals (in days) between coal-mining disasters. The line is the estimated mean function obtained by the feature preserving smoother, the critical values are $c_2=2.95$, $c_3=4.74$ and $c_4=3.82$.

Fig 1

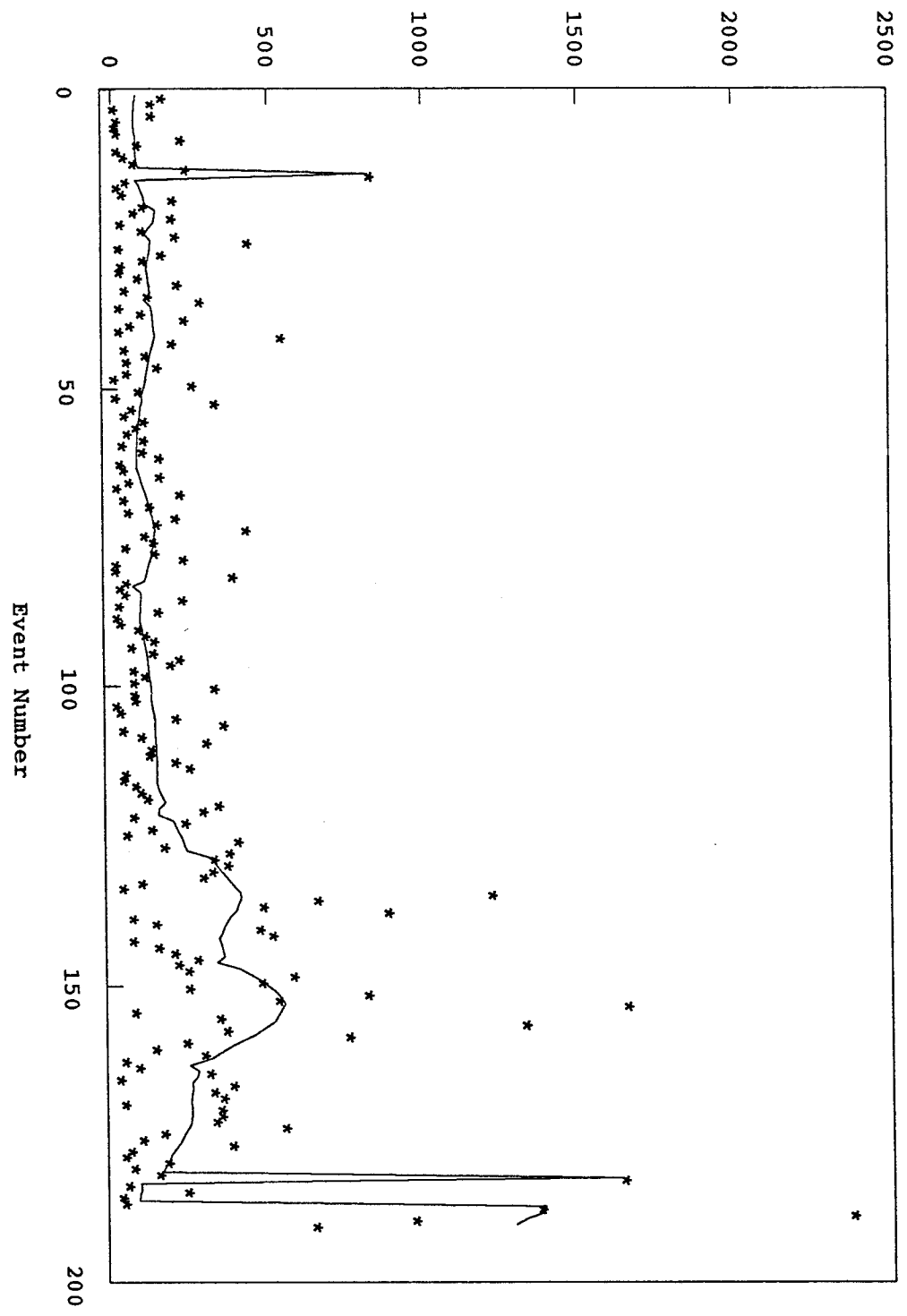


Fig 2

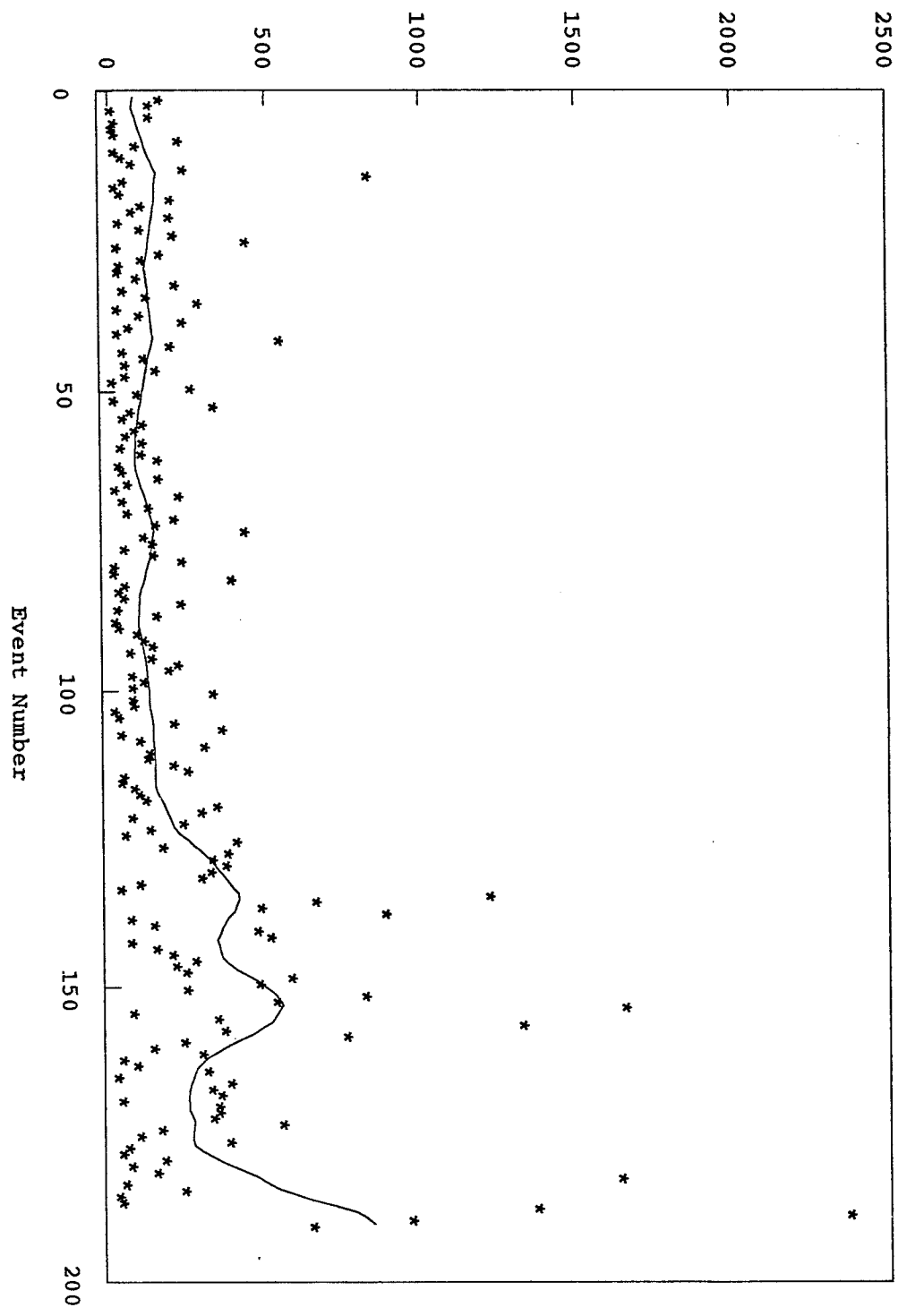


Fig 3

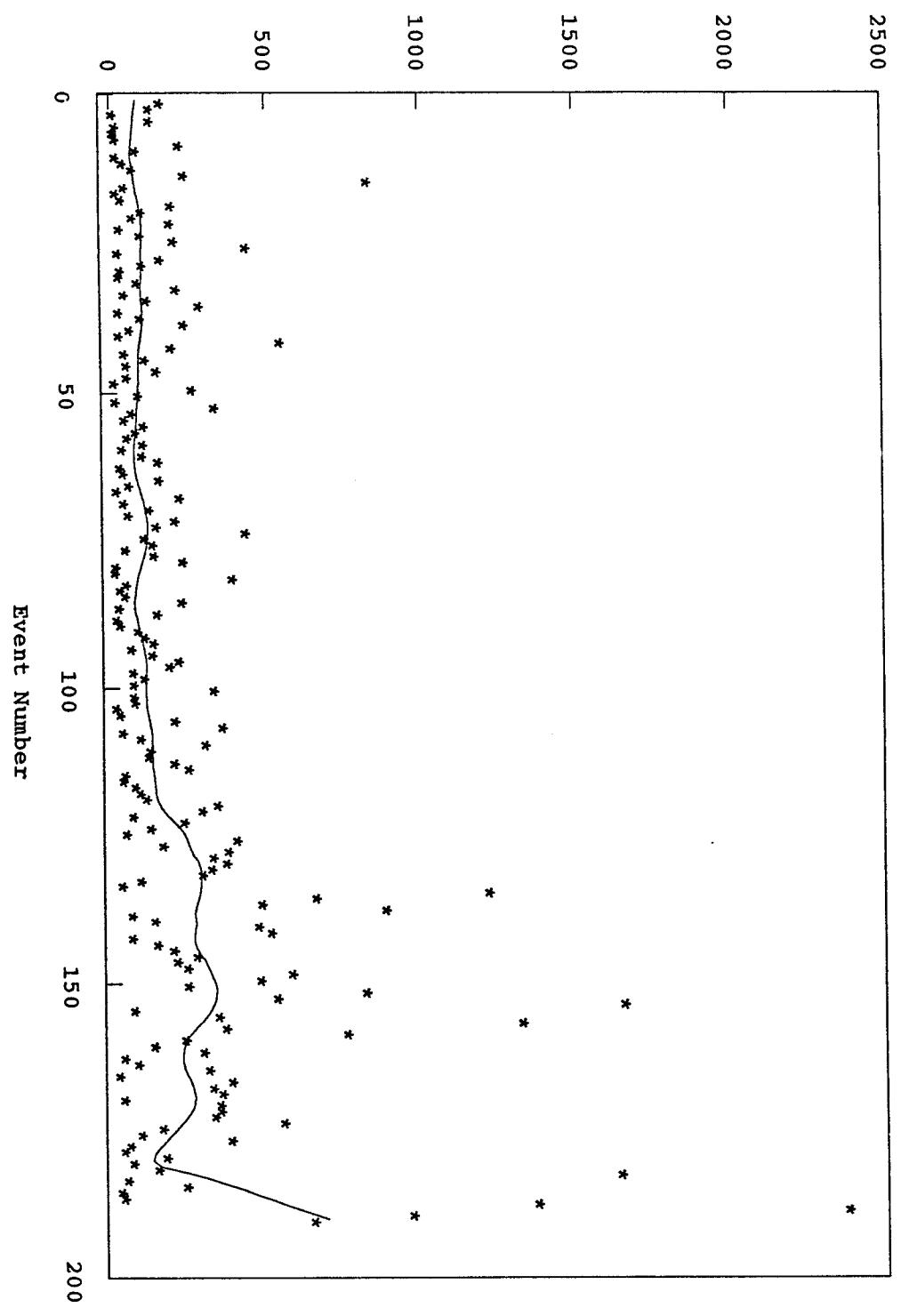


Fig 4

