Locating lines among scattered points

PETER HALL[1], NADER TAJVIDI[1,2] and P.E.MALIN[3]

[1]*Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*
[2]*Centre for Mathematical Sciences, Lund Institute of Technology, Box 118, SE-221 00 Lund, Sweden. E-mail: nader@maths.lth.se*
[3]*Division of Earth and Ocean Sciences, Old Chemistry Building, Box 90227, Duke University, Durham, North Carolina 27708, USA*

Consider a process of events on a line L, where, for the most part, the events occur randomly in both time and location. A scatterplot of the pair that represents position on the line, and occurrence time, will resemble a bivariate stochastic point process in a plane, P say. If, however, some of the points on L arise through a more regular phenomenon which travels along the line at an approximately constant speed, creating new points as it goes, then the corresponding points in P will occur roughly in a straight line. It is of interest to locate such lines, and thereby identify, as nearly as possible, the points on L which are associated with the (approximately) constant-velocity process. Such a problem arises in connection with the study of seismic data, where L represents a fault-line and the constant-velocity process there results from the steady diffusion of stress. We suggest methodology for solving this needle-in-a-haystack problem, and discuss its properties. The technique is applied to both simulated and real data. In the latter case it draws particular attention to events occurring along the San Andreas fault, in the vicinity of Parkville, California, on 5 April 1995.

Keywords: earthquake; hypothesis test; large-deviation probability; ley-line; point process; Poisson process; San Andreas fault; spatial process

1. Introduction

The problem and data which motivate this paper arise in geophysics, and are as follows. Consider an approximately linear segment of a geological fault-line, and suppose seismic events along the fault are recorded as pairs $x = (x_1, x_2)$, where x_1 denotes the distance along the line (from a fixed point) of the place where the event occurred, and x_2 is the occurrence time of the event. A small fraction of the pairs result from bursts of energy which propagate along the fault at approximately constant speeds, causing seismic events as they go. These bursts generate roughly linear clusters of points in the (space, time) plane, the gradient of the line being proportional to the speed at which energy travels along the fault. Identifying these clusters is of intrinsic geophysical interest. (If the triggering event were to start within the study region then a V-shaped pattern might be observed in the plane; it could be detected as two separate lines.)

  The fault-line in question is a section of the San Andreas fault near Parkfield, California. Data from this region are described by Malin and Alvarez (1992) and Malin *et al.* (2002); see Section 3.2 for further sources. A typical data set is graphed in Figure 1.1. The actual
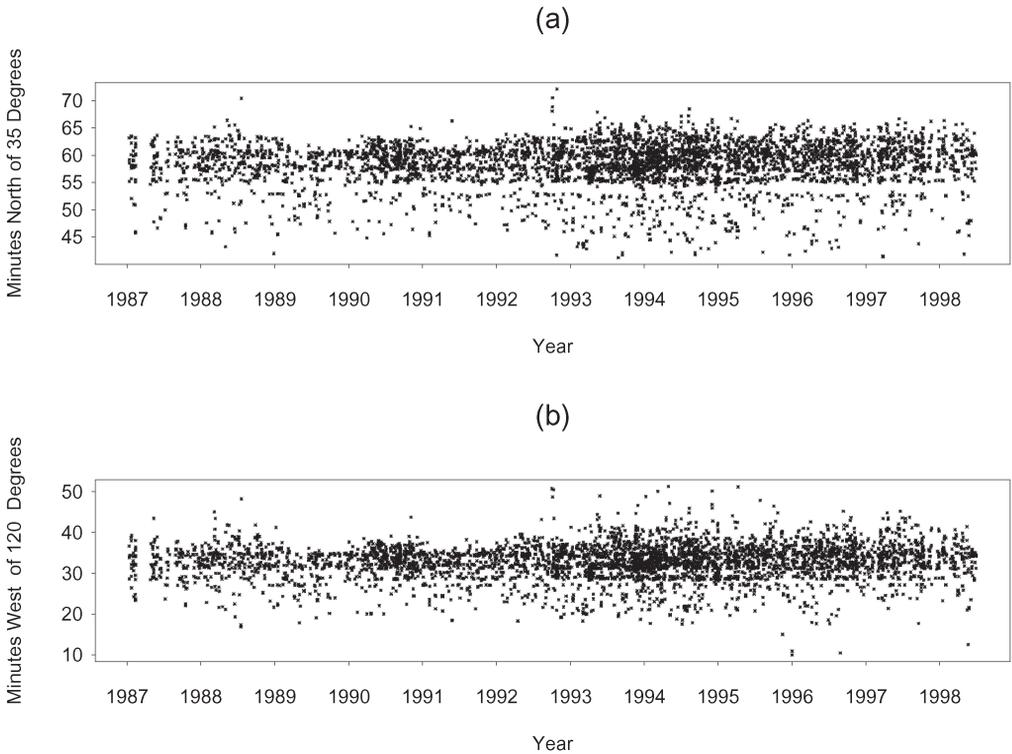
Figure 1.1. (a) Vertical and (b) longitudinal locations, in minutes of longitude and latitude respectively, and occurrence times of microearthquake events along the San Andreas fault near Parkfield, California. See Section 3.2 for further details.

data include spatial locations, and occurrence times with millisecond precision, of microearthquake events between 1987 and mid-1998.

Motivated by this problem, we suggest a method for detecting approximately linear clusters of points in the plane. Our algorithm searches for a surplus of points within a short but thin strip, $\mathcal{A}$, relative to the numbers of points in two other strips, $\mathcal{B}_1$ and $\mathcal{B}_2$, on either side of $\mathcal{A}$. We employ $\mathcal{B}_1$ and $\mathcal{B}_2$ to estimate the distribution of the number of points which would be expected to lie within the thinner, central strip $\mathcal{A}$, and thereby to provide an informal hypothesis test of significance. The null hypothesis is that the points are Poisson-distributed, although the intensity of this point process is not required to be uniform. Indeed, intensity is estimated locally each time the informal test is implemented; the method is used in the continuum. In particular, the maximum dimensions of $\mathcal{A}$, $\mathcal{B}_1$ and $\mathcal{B}_2$ are quite small, and so our approach is spatially and temporally adaptive. We show that our method is close to being optimal, in the sense that no other approach can correctly identify approximately linear clusters containing an order of magnitude fewer points.

In the context of our geophysical data, the approach described above has several advantages relative to some existing techniques. First, it is straightforward to apply in

inhomogeneous cases, where point-process intensity varies in the plane. Spatio-temporal fluctuation of intensity is a characteristic of seismic events, and requires a methodology which is intrinsically local in character. The need to accommodate this type of variation, even at the basic level of modelling the point process, means that methods for solving the problem are bound to be computer-intensive.

Secondly, the ease with which our informal test can be calibrated means that implementation adds relatively little computational complexity to the burden of calculation. In particular, the test does not require Monte Carlo simulation. Related techniques, based on probability calculations and stochastic-geometric formulae, can be difficult to implement without simulation, especially if the lines of points are not very clearly defined.

Thirdly, the informal test readily accommodates cases where prospective lines contain relatively many points, even though the points may not all be particularly close to lying on the same straight line.

In the discussion above we have stressed the informal nature of our test, to emphasize the fact that the test is part of a diagnostic procedure and not an end in itself. Indeed, the role of our test is quite similar to that of tests in many classification problems, where the diagnostic for classification can be viewed as an approximation to a likelihood ratio statistic that might be used to conduct a hypothesis test. This parallel becomes closer if we consider that we are trying to classify a linear cluster of points as being part of an approximate line, or as having arisen purely by chance, respectively. Implementing our approach depends on selecting a number of tuning parameters, which cannot be chosen adequately without referring to the data; and that reference cannot be made without, to some extent, impinging on the level and power of the informal tests, just as in the case of many tests that are parts of classification procedures. In Section 3.2, where we apply our method to a real data set, we shall argue that it is possible to make the reference conservatively, so that the errors are small and do not impact seriously on the performance of our method.

The reader will have noticed that our approach involves converting a three-dimensional problem, where data have two spatial coordinates x and y and a temporal coordinate t, into two bivariate problems, with data sets in the (x, t) and (y, t) planes respectively. The advantages of this approach are threefold. Firstly, if we were to treat the problem in its original trivariate form then the strips we would use for informal testing would become cylinders, with axial cylinders running down their cores. Since the orientation of a cylinder has two degrees of freedom, rather than one (a cylinder should be rotated in two angular dimensions), then the amount of calculation needed in three dimensions is greatly increased, relative to that required in two one-dimensional subproblems. Secondly, even in the setting of the Parkfield data, with some 5000 points, data sparsity becomes a problem if we work in three rather than two dimensions. Thirdly, if we attempt to solve two one-dimensional problems then each provides a check on the other, in the sense that a line which is clearly present in the three-dimensional problem should be manifest in both the two-dimensional subproblems. As we shall show in Section 3.2, this opportunity for 'cross-validating' the results between the two bivariate data sets provides helpful insight into whether a suggested line might be illusory. In particular, after analysis of the (x, t) and (y, t) point patterns, the findings in the two patterns can be combined. This will prove to be a powerful tool in our analysis of the real-data example.

The Poisson assumption, on which our analysis is partly based, is perhaps open to query. It tends to produce accurate or conservative results (i.e. informal tests which err on the side of relatively low level) in the case of processes where the variance of the number of points in a region is less than a constant multiple of the area of the region. However, it tends to be anticonservative otherwise. The Poisson assumption, or its conditioned version (i.e. a uniform distribution of points in a region), is commonly made in related work, such as that discussed above, and seems difficult to avoid unless one has a specific alternative model and is prepared to simulate extensively in order to compute critical points. One might expect the process producing the seismic data to be closer to a Poisson cluster process than a pure Poisson process. In this case the conservatism (or otherwise) of our procedure would likely depend on the distribution of spatial clump size.

There is a related literature in the area of archaeology, termed 'post-hole analysis', which is devoted to the problem of finding regular structures (usually of rectangular shape, representing the outlines of rooms or buildings) in scanned excavation plans. See, for example, Fletcher and Lock (1981, 1984, 1991) and material cited therein. The work of Fletcher and Lock includes discussion of alternative methods for assessing the significance of alignments, based on perturbing the data sufficiently to destroy alignments but not enough to alter the distribution of scattering. See also Small (1996), who addresses movable-strip methods in this context. In related work, Mack (1950) derives elegant formulae for the expected numbers of sets of k points, out of n points randomly distributed in the plane, that can be covered by a square or triangle. Still another approach to discovering geometric structure is given by Kent *et al.* (1983) in the context of palaeomagnetic data.

In related work, Broadbent (1980) studies the incidence of 'ley-lines', or approximate straight lines linking sites of archaeological interest, in Britain. Kendall and Kendall's (1980) study of theoretical properties of triads plays a role in the development of statistical shape theory (see Silverman and Brown 1978; Kendall 1984, 1985, 1986; Small 1984, 1988; Goodall 1991). Small (1982) extends Kendall and Kendall's (1980) methodology. Gates (1986) introduces measures of collinearity alternative to those proposed by Kendall and Kendall (1980).

Statistical work on testing significance by reference to movable strips is well known in statistics, for example through a very large literature on the scan statistic. See Cressie (1977) and Saunders (1978) for early accounts of its properties, Kulldorff *et al.* (1998) for a recent example of its application to space-time clustering, and Weinstock (1981) and Priebe *et al.* (1997) for related examples, in medical settings, of applications of the scan statistic to detecting clusters.

There is a very extensive, and more recent, literature in the area of computer vision and machine understanding, on the problem of locating lines among scattered points. It includes work by Stewart (1995a, 1995b), Danuser and Stricker (1998), Frigui and Krishnapuram (1999), Meer *et al.* (2000), Desolneux *et al.* (2003a, 2003b) and Susaki *et al.* (2004).

2. Methodology

We search for lines of points by using a sequence of 'test strips'. A test strip, $\mathcal{S}$, is an $a \times b$ rectangle divided into three disjoint, parallel substrips. One substrip is a $c \times b$

rectangle, where $0 < c < a$, and is axial to the main $a \times b$ rectangle. The other substrips are $\frac{1}{2}(a - c) \times b$ rectangles, and lie on either side of the axial substrip. We shall call b the length of the test strip and of the substrips, and shall refer to a as the width of the test strip.

Thus, the axial $c \times b$ substrip is sandwiched between two other substrips. Supposing that points in the plane come from a Poisson process, $\mathcal{P}$ say, and assuming the hypothesis H_0 that Poisson intensity does not vary significantly within $\mathcal{S}$, we shall use data that fall within the two $\frac{1}{2}(a - c) \times b$ substrips to estimate the average intensity of $\mathcal{P}$ in $\mathcal{S}$. Employing this estimator, we shall reject H_0 if the number of points of $\mathcal{P}$ that lie in the axial substrip is unduly large, in particular if it exceeds an empirically determined critical point. If the axial substrip is quite narrow relative to its length, and if our test leads to rejection, then we argue that the axial substrip contains an approximately linear array of points that has been adjoined to the Poisson process $\mathcal{P}$.

More generally, provided the test strips are not particularly large, we do not need to assume $\mathcal{P}$ is homogeneous, although we do ask that point-process intensity not change dramatically in the plane. Neither do we have to specify the test strip, or its substrips, particularly narrowly; wide latitude in their choice is possible.

The informal hypothesis test discussed above will be applied repeatedly, for many different locations, orientations and sizes of the test strip. Next we discuss critical points for these tests. Let $\log_* t = \log t$, the natural logarithm of t, if $t \geqslant e$, and let $\log_* t = 1$ otherwise. Put

$$x_u(t) = t + u(t \log_* t)^{1/2}, \tag{2.1}$$

where $u > 0$. Given a region $\mathcal{R}$ of area $\|\mathcal{R}\|$, and a constant $\lambda > 0$, we may interpret $x_u(\lambda \|\mathcal{R}\|)$ as a critical point of the distribution of the number of points of a homogeneous Poisson process, with intensity λ, within $\mathcal{R}$. Indeed, if, in (2.1), we were to replace $\log_* t$ by simply 1, and if λ and $\mathcal{R}$ were to vary in such a way that $\lambda \|\mathcal{R}\|$ diverged to infinity, then the probability that the number of points within $\mathcal{R}$ exceeded $x_u(\lambda \|\mathcal{R}\|)$ would converge to the probability that a standard normal random variable exceeded u. We shall use $x_u(\lambda \|\mathcal{R}\|)$, with λ replaced by an estimator, and incorporating a minor adjustment, as a critical point for repeated testing during our search for lines of points.

Suppose the points of $\mathcal{P}$ are distributed within a region containing $\mathcal{R}$, and let $N(\mathcal{R})$ denote the number of points which fall within $\mathcal{R}$. Given a test strip $\mathcal{S}$, let $\mathcal{A}$ denote its axial $c \times b$ substrip, and write $\mathcal{B} = \mathcal{S} \backslash \mathcal{A} = \mathcal{B}_1 \cup \mathcal{B}_2$ for the union of the other two substrips, $\mathcal{B}_1$ and $\mathcal{B}_2$ say. Put $\hat{\lambda}_j = N(\mathcal{B}_j)/\|\mathcal{B}_j\|$. As our estimator of average point-process intensity within $\mathcal{S}$ we could use either

$$\tilde{\lambda} = N(\mathcal{B})/\|\mathcal{B}\| = \tfrac{1}{2}(\hat{\lambda}_1 + \hat{\lambda}_2) \quad \text{or} \quad \hat{\lambda} = \max(\hat{\lambda}_1, \hat{\lambda}_2). \tag{2.2}$$

However, $\hat{\lambda}$ gives a procedure which is more robust against departures from homogeneity, and so is less likely to produce false positive results owing to fluctuations in point-process intensity. To appreciate why, note that the maximum is never less than the average, and so, if we use the maximum to estimate background point-process intensity, then in order for an

informal test to result in rejection there has to be a greater number of points in the axial substrip than would be required if we used the average.

In a slight abuse of notation, we shall consider $\mathcal{S}$ to denote all dimensional and configurational information about the test strip $\mathcal{S} = \mathcal{S}(a, b, c, x, \theta)$ and its substrips. In particular, the notation $\mathcal{S}$ conveys the values of a, b and c, and the centre x of $\mathcal{S}$ and the orientation, θ, of $\mathcal{S}$ with respect to a fixed axis in the plane.

Assume $\mathcal{P}$ is a Poisson process, not necessarily homogeneous. Let $v > 0$, let x_u and $\hat{\lambda}$ be as at (2.1) and (2.2), and consider the hypothesis $H_0 = H_0(\mathcal{S})$ that the integral average of point-process intensity within the axial substrip $\mathcal{A}$ exceeds the maximum of the average intensities within $\mathcal{B}_1$ and $\mathcal{B}_2$. Our informal test of $H_0(\mathcal{S})$ will be based on the following rule:

$$\text{reject } H_0(\mathcal{S}) \text{ if and only if } N(\mathcal{A}) \geqslant \max\{x_u(\hat{\lambda} \, \|\mathcal{A}\|), v\}. \tag{2.3}$$

By inserting $v > 0$ here we are guarding against false positives arising from stochastic fluctuations, for example of the estimator $\hat{\lambda}$.

3. Simulation results and application to real data

3.1. Simulation

To assess performance of our method we simulated data sets of n points sampled independently and uniformly within the unit square. Such points can be considered to come from a homogeneous Poisson process, conditioned on the number of points in the square equalling n. The goal of the simulation study was to address performance of the test when (i) no line was present inside the unit square (call this H_0), (ii) a single straight line of points was added to the square (H_1 without noise), or (iii) points on the added line were observed with error (H_1 with noise). In case (iii) we 'jittered' the points on the line by adding random noise, using the S-Plus function jitter. In the present case this is virtually the same as perturbing a point x to $x + R$, where each component of R is uniformly distributed in $[-z, z]$ and $z = $ 'noise factor' $\times(1/50)$. The 'noise factor' was taken to lie between 0.1 and 3; see below for details. For all the results reported in this section we took $v = 2$.

To implement the test we searched for lines on a 10×10 grid as follows. First, the centre of a test strip was located at each point of the grid. Then we rotated the test strip 5 degrees at a time, so that 36 tests were carried out at each grid point (resulting in 3600 tests in each simulation run.) In practice a greater number of locations and rotations would be considered; see Section 3.2. The fact that every parameter configuration had to be simulated many times was a limiting factor in the present section.

As expected, the mean number of falsely detected lines was a decreasing function of u and an increasing function of c. For example, when $(n, a, b, c) = (100, 0.1, 0.6, 0.01)$, and simulation was under H_0, the average number of false positive tests was approximately 3 for $5.9 \leqslant u \leqslant 6$; 6 for $5.4 \leqslant u \leqslant 5.8$; 9 for $4.1 \leqslant u \leqslant 5.3$; and 12 for $u = 4$. Each small cluster of points in the plane that gave rise to at least one detection, produced about two

other false positive tests for different rotations or locations of the test strips. Hence, the above numbers have to be divided by 3 in order to give the number of small clusters of points which misleadingly suggested the presence of a line. These false clusters therefore numbered, on average, approximately 1, 2 and 3 for $5.9 \leqslant u \leqslant 6$, $5.4 \leqslant u \leqslant 5.8$ and $4 \leqslant u \leqslant 5.3$, respectively. When c was increased from 0.01 to 0.05 the average number of these misleading clusters rose to approximately 3 and 4 for $5.4 \leqslant u \leqslant 5.8$ and $4.1 \leqslant u \leqslant 5.3$, respectively. (The number stayed at about 1 when $5.9 \leqslant u \leqslant 6$.)

Of course, in practice the putative line of points is not perfectly straight. To provide information about the effects of perturbation we shall report false discovery rates, and detection probabilities, under the model 'H_1 with noise'. To generate the results in Figure 3.1 we took $(n, a, b, c) = (100, 0.1, 0.6, 0.01)$, as used above, and added a line of 10 points, except that the points in the line were jittered using noise factors 0.1, 0.25, 0.5 and 1.0 in the respective panels, from left to right, in each of the two parts of Figure 3.1. The added line was said to be 'partially detected' if at least two of its points were among those in the axial substrip which gave rise to a significant test result; and 'fully detected' if all its points were among those points. More generally, any collection of points that gave rise to a positive result was said to have produced a 'detected line'.

Figure 3.1(a) shows average numbers of line detections (i.e. statistically significant tests, either true positive or false positive), shown as circles, and average numbers of (at least) partial detections (true positives), shown as plus signs. Figure 3.1(b) gives the chance (expressed as a percentage) of detecting at least one cluster of points containing at least two points from the added line, shown as a circle, and also the probability of detecting a cluster which contains all 10 of the added points, shown as a plus sign.

As in the $n = 100$ example discussed earlier, on average each small cluster of points in the plane that gave rise to at least one detection in the experiments summarized in Figure 3.1, produced approximately two other detections. This is true for the cluster produced by the added line as well as for spurious linear-looking clusters. Therefore, each count in each panel is usually greater, by about two or three, than the respective counts reported earlier when $n = 100$; see Figure 3.1(b). (The last of the four panels there is an exception, and will be discussed shortly.) These two or three extra significant test results correspond to correct detections of the added line. Equivalently, the number of small point-clusters detected is roughly 2, 3 or 4 when $5.9 \leqslant u \leqslant 6$, $5.4 \leqslant u \leqslant 5.8$ or $4 \leqslant u \leqslant 5.3$, respectively, except that the number tends to decrease as the amount of noise (used to perturb points on the added line) increases.

This reflects the decreasing likelihood of detecting any of the added points as the noise factor increases. The probability is virtually 100% in the first panel (representing noise factor 0.1). It falls to between 75% and 100% in the third panel, and to between 20% and 70% in the fourth, depending on the value of u. Using a larger u increases the probability of detecting at least part of the added line, but, as shown in Figure 3.1(a), and of course as expected, it also increases the number of false positives. The chance of detecting all 10 points on the line never exceeds 30%, and is virtually zero in the case of the fourth panel. This is not necessarily a practical problem, however; having detected a cluster of points containing the added line, visual inspection can then usually determine most of the 10 points, generally with some 'false' points as well.
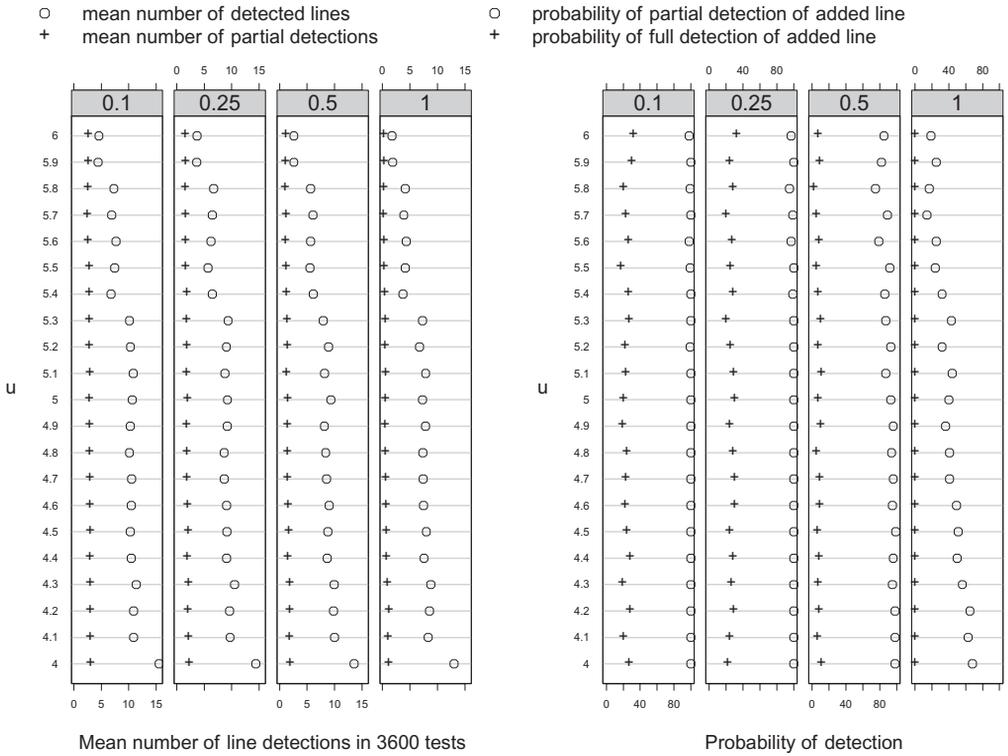
Figure 3.1. Numbers of detected lines and probabilities of detections obtained with $(n, a, b, c) =$ (100, 0.1, 0.6, 0.1) and with the noise factors as indicated at the top of each panel. (a) Monte Carlo estimates of expected numbers of positive test results (indicated by circles) and the numbers of those which contain at least two points of the added line (shown by plus signs). The number of detections should be divided by about 3 to get the number of small-point clusters, in the unit square, that produced statistically significant test results. In particular, the added line of 10 points was at least partially detected approximately three times, if it was detected at all. (b) The probability, expressed as a percentage, that the added line was at least partially detected (indicated by circles) or wholly detected (shown by plus signs).

3.2. Application to real data

The Parkfield data set, summarized in Figure 1.1, is comprised of 5102 points. It contains time and location information on microearthquake activities of the San Andreas fault near Parkfield, California, between 1987 and 1998.

The measurements that produced the data were obtained using a network of borehole seismographs placed in different locations in the area. From a geophysical viewpoint, both vertical and longitudinal 'migrations' of events are of interest. Among the studies that have been made of seismic events in the Parkfield area, and of mathematical models for those

events, we mention the contributions of McEvilly *et al.* (1967), Bakun and McEvilly (1979, 1984), Bakun and Lindh (1985), Stuart *et al.* (1985), Lindh and Malin (1987), Segall and Harris (1987), Malin and Alvarez (1992) and Malin *et al.* (2002).

Compared to the simulations in Section 3.1 we used a greater number of grid points, employing a 20×20 lattice and increasing the number of test-strip rotations from one every 5 degrees to one every degree. Therefore, in total we carried out $400 \times 180 = 72\,000$ tests. For simplicity, and for a degree of comparability with the results obtained in Section 3.1, we standardized both axes so that the data were analysed on the unit square; the discussion below should be interpreted in this context.

The main challenge facing an experimenter is choosing the tuning parameters a, b, c, u and v. Taking $(a, b, c) = (0.05, 0.1, 0.001)$ corresponds to using a test strip of which the length is exactly twice the width. This shape reflects the fact that, in the case of inhomogeneity, we do not want the strip to be too wide in a direction perpendicular to the line we are trying to detect, in case it reaches into regions where point-process intensity is significantly different from that near the line. On the other hand, we want the test strip to be wide enough to ensure that the intensity estimators $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are based on sufficiently many data.

When $(a, b, c) = (0.05, 0.1, 0.001)$ and the point process in the unit square is homogeneous with the same number of points as the real data set, the estimators $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are each based on approximately 12.5 data, whereas the axial substrip contains only 0.245 data, on average. (For the real data, and with $(a, b, c) = (0.05, 0.1, 0.001)$, the mean number of points in the two substrips on either side of the axial substrip was 13.5.) For a homogeneous Poisson process with the same number of points as the real data set, Monte Carlo simulations in the unit square show that when $u \geqslant 7$ and $v = 10$ the probability of one or more positive results among the $2 \times 72\,000$ tests (i.e. one test for each of $72\,000$ test-strip positions in the longitudinal-migration and lateral-migration data sets) is much less than 0.1.

While these results encourage the choices $(a, b, c) = (0.05, 0.1, 0.001)$, $u \geqslant 7$ and $v = 10$, they do not adequately treat the case of extreme counts in high-intensity parts of a Poisson process that resembles the real data set. To address this aspect of the problem we argued as follows. For each of the $72\,000$ locations of the test strip, there are two non-axial substrips in the case of longitudinal-migration data, and two non-axial substrips in the vertical-migration setting. The maximum number of counts in these substrips is 220, occurring in a non-axial substrip for one of the $72\,000$ locations in the case of longitudinal migrations. Dividing 220 by the area, $0.002\,45$, of the non-axial substrip, we obtain $89\,800$, being the largest of the $4 \times 72\,000 = 288\,000$ local intensity estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$. Reflecting this value, we consider properties of a Poisson processes, in the unit square, with intensity $90\,000$. In a simulation study of 200 such processes, none of the $200 \times 72\,000 = 1.44 \times 10^7$ tests with $(a, b, c) = (0.05, 0.1, 0.001)$, $u \geqslant 7$ and $v = 10$ gave a positive result.

Together these conclusions suggest that, for the real data set, spurious lines are unlikely to arise if we take $(a, b, c) = (0.05, 0.1, 0.001)$, $u \geqslant 7$ and $v = 10$. This will be our choice below, although we shall also report results for smaller values of u, in order to set the larger values into context.

For $u = 4$, and looking at vertical migrations (i.e. using the data shown in Figure 1.1(a)), our test found 19 lines in 7 different locations. Here 'location' is defined as the coordinates of the centre of the test strip, and a line is specified by its location and angle, which are measured counterclockwise from the horizontal. For $u = 5$, 7 and 8 the numbers of lines (again for vertical migrations) reduced to 16, 5 and 3, respectively. In the case of horizontal migrations we found 18, 12, 5 and 4 lines for $u = 4$, 5, 7 and 8, respectively. As with the simulation results reported in Section 3.1, the size (and indeed the discreteness) of these numbers reflects the number of points in the grids of locations and rotation angles. It nevertheless indicates which locations and angles are of particular interest.

Table 3.1 details the lines which are statistically significant when $u = 7$, 7.5 or 8. It can be seen that in each case there are essentially three timepoints which generate all the lines. Two of these are shared between the two data sets represented in Figures 1.1(a) and 1.1(b), respectively. Hence, for each set of deleted lines there is one timepoint which is not represented in the other. These timepoints may correspond to false positives; in the case of

Table 3.1. Date, time and (a) latitude or (b) longitude of the centres of test strips for different values of u. Angles are measured in degrees, counterclockwise from horizontal axis. The symbol • indicates that a statistically significant result, for the given value of u, was obtained for a test strip corresponding to the stated date, time, position and angle

(a)

Date and time	Minutes north of 35 degrees	Angle	u		
			7	7.5	8
7 Aug. 94, 08:41:22.753	58.02	86	•		
	58.02	87	•	•	•
	58.02	88	•	•	•
5 Apr. 95, 20:29:12.341	58.02	119	•	•	•
5 Oct. 90, 11:35:20.859	59.48	87	•		

(b)

Date and time	Minutes west of 120 degrees	Angle	u		
			7	7.5	8
7 Aug. 94, 08:41:22.753	32.13	87	•	•	•
	32.13	88	•	•	•
5 Dec. 94, 09:52:00.954	32.13	54	•	•	•
5 Apr. 95, 20:29:12.341	32.13	118	•	•	
	32.13	119	•	•	•

vertical migration, the corresponding line is not statistically significant when $u = 8$. The remaining two timepoints, repeated in each part of Table 3.1 and representing events which occurred in August 1994 and April 1995, respectively, are therefore singled out for special attention.

Figure 3.2 shows the test strips which produce statistically significant results when $u = 8$, for (a) vertical migration (and (b) longitudinal migration. As indicated in Table 3.1, there are essentially only two test strips in the vertical case, corresponding to events occurring on 7 August 1994 and on 5 April 1995. As suggested by Table 3.1(a), the two lines of points deriving from events of 7 August 1994 are so close together that the respective test strips appear as one. In Figure 3.2(b) there are, however, three visible test strips, corresponding to: two coalesced test strips representing the events of 7 August 1994; one test strip for the events of 5 April 1995 (both of which dates are 'shared' with Figure 3.2(a)); and a third test strip, produced by events of 5 December 1994, which has no counterpart in Figure 3.2(a) and which is possibly a false positive.

Although the lines in August 1994 and April 1995 are both statistically significant, there are intrinsic differences between them. Perhaps most importantly, the line in August 1994 is nearly vertical, whereas that in April 1995 is at an angle which corresponds to that of the San Andreas fault. This line should therefore be considered more significant from a seismological point of view, independently of any measure of statistical significance it might have.
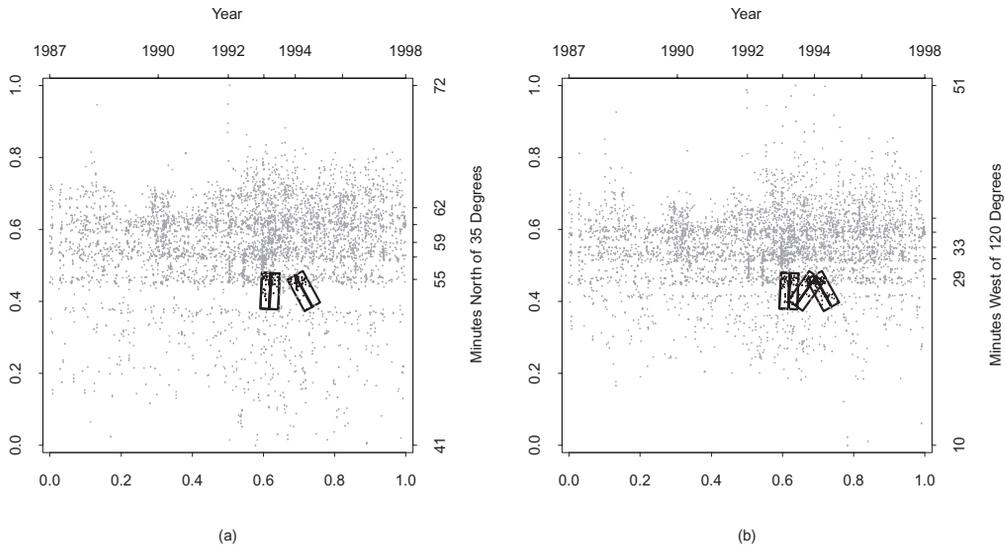

Figure 3.2. Lines detected in the data set of events in Parkfield, California. (a) Vertical and (b) longitudinal migration lines, respectively, and the test strips for which they are statistically significant. Both time and location axes were transformed linearly to the interval [0,1].

4. Theoretical properties

First, we construct an asymptotic model which we shall use to assess properties of our tests. We shall allow the intensity of points in $\mathcal{P}$ to increase, in proportion to ℓ, say. This will permit us to search in increasingly small regions for approximately linear arrays of points. Therefore, at the same time as we increase ℓ we shall shrink the test strips $\mathcal{S}$. Indeed, we shall let

$a = \alpha h$ and $b = \beta h$, where $h = h(\ell)$ converges to zero as $\ell \to \infty$, but α

and β are constrained to lie within fixed intervals, specifically $\alpha \in [\alpha_1, \alpha_2]$ (4.1)

and $\beta \in [\beta_1, \beta_2]$, with $0 < \alpha_1 < \alpha_2 < \infty$ and $0 < \beta_1 < \beta_2 < \infty$.

Construct the $a \times b$ test strip $\mathcal{S} = \mathcal{S}(a, b, c, x, \theta)$, centred at $x \in \mathcal{D}$, with its sides of length b making an angle $\theta \in [0, \pi)$ to some fixed line in the plane, and with axial substrip $\mathcal{A}$ of width c.

We shall suppose too that $c = c(\ell)$ satisfies

$c \in [0, \gamma h]$,   where $\gamma \in (0, \alpha_1)$ is fixed.   (4.2)

In particular, we allow the width of the axial substrip $\mathcal{A}$ to be arbitrarily small; no lower bound is placed on c. Of course, this reflects the fact that we are seeking an array of points arranged approximately in a straight line. Responding to this objective, we could ask that $0 \leqslant c \leqslant \delta$, where $\delta = \delta(\ell)$ satisfies $\delta/h \to 0$ as $\delta \to 0$, implying that we search only within thin axial substrips. However, the nature of our informal test makes this unnecessary.

The following theorem shows that for a wide range of choices of the tuning parameters, there is only a small probability that our algorithm 'detects' spurious lines. This result and Theorems 4.2 and 4.3 below show that if the test strip is broadly of the type used in Section 3 – for example, with its sides in moderate proportion and its axial substrip not so narrow that it could not cover most of the points on a 'jittered' line – then our sequence of informal tests is able to fill the functions intended of it, and reliably find added lines without suffering too many false positives. The theorems are not intended to provide an algorithm for choosing the tuning parameters. Derivations of Theorems 4.1 and 4.2 are given in Section 5, and a proof of Theorem 4.3 is provided in a longer version of this paper (Hall *et al.*, 2005).

Theorem 4.1. *Assume the point process $\mathcal{P}$ is Poisson with intensity $\ell f(x)$, at $x \in \mathbb{R}^2$, where the fixed function f has two bounded derivatives and satisfies $f > 0$ in $\mathbb{R}^2$. Suppose too that for some $\varepsilon > 0$, $h = O(\ell^{-1/6-\varepsilon})$ and $\ell^{1-\varepsilon} h^2 \to \infty$ as $\ell \to \infty$. Let $\mathcal{D}$ denote a bounded, measurable subset of the plane, and consider all configurations of test strips $\mathcal{S} = \mathcal{S}(a, b, c, x, \theta)$, where $x \in \mathcal{D}$, $\theta \in [0, \pi)$, and substrip dimensions a, b, c are constrained only by (4.1) and (4.2). Taking u fixed and using $v = w \log \ell$, for fixed w, conduct the test described at (2.3) for all such configurations of $\mathcal{S}$. Then, provided $u, w > C$, where $C > 0$ is sufficiently large, the probability that the hypothesis test rejects $H_0(\mathcal{S})$, for some $\mathcal{S}$, converges to 0 as $\ell \to \infty$.*

Provided the area, $\|\mathcal{D}\|$, of $\mathcal{D}$ is non-zero, Theorem 4.1 continues to hold if the identity '$v = w \log \ell$' is replaced by '$v = w \log \hat{\mu}$', where $\hat{\mu} = N(\mathcal{D})/\|\mathcal{D}\|$ is an estimator of the average intensity of $\mathcal{P}$ in $\mathcal{D}$. In this sense the critical points for the test can be chosen empirically.

The condition $h = O(\ell^{-1/6-\varepsilon})$ as $\ell \to \infty$, imposed in the theorem, is needed to ensure that the estimate of intensity within the axial substrip $\mathcal{A}$, being based on the substrips $\mathcal{B}_1$ and $\mathcal{B}_2$ on either side of $\mathcal{A}$, is not unduly biased upwards through the intensity of $\mathcal{P}$ being too large on either side of $\mathcal{A}$. It is readily seen that significant bias can occur if h is too large. Note also that, in view of the assumption that the Poisson intensity equals ℓf, each test strip includes approximately ℓh^2 points. Therefore, the assumption that $\ell^{1-\varepsilon} h^2 \to \infty$ as $\ell \to \infty$ ensures that the estimator $\hat{\lambda}$ is, with high probability, based on at least const. $\times \ell^{\varepsilon}$ points, that is, on a polynomially large number of points of $\mathcal{P}$.

Next, we address the power of our procedure when an approximately linear, randomly perturbed array of points is added to $\mathcal{P}$. Specifically, suppose $n = n(\ell)$ points are adjoined to $\mathcal{P}$ within a $\delta \times \beta h$ strip centred in $\mathcal{D}$, where $\beta \in [\beta_1, \beta_2]$. (The distribution of the adjoined points within the $\delta \times \beta h$ strip can be arbitrary.)

Theorem 4.2. *Assume the conditions of Theorem 4.1, and that $n = n(\ell)$ and $\delta = \delta(\ell)$ satisfy $\ell \delta h \to \infty$, $\delta/h \to 0$,*

$$\lim_{\ell \to \infty} n^{-1}(\ell \delta h)^{1/2} \to 0, \qquad \limsup_{\ell \to \infty} n^{-1} \log \ell \leqslant \eta. \tag{4.3}$$

Let $\mathcal{S}$ denote any test strip $\mathcal{S} = \mathcal{S}(a, b, c, x, \theta)$ which satisfies (4.1) and has as its axial substrip $\mathcal{A}$ the $\delta \times \beta h$ region within which the n new points, adjoined to $\mathcal{P}$, are distributed. Write $\pi(\ell)$ for the probability that our algorithm rejects $H_0(\mathcal{S})$. Then, provided $\eta > 0$ in (4.3) is sufficiently small, relative to the fixed values of u and w in the critical point $\max\{x_u(\hat{\lambda}\|\mathcal{A}\|), w \log \ell\}$, $\pi(\ell) \to 1$ as $\ell \to \infty$.

Of course, it is likely the algorithm defined by (2.3) will reject $H_0(\mathcal{S})$ for a number of test strips $\mathcal{S}$, not just those specified in Theorem 4.2. However, it follows from Theorem 4.1 that the test is not likely to reject $H_0(\mathcal{S})$ for any $\mathcal{S}$ whose axial substrip $\mathcal{A}$ does not include any of the n adjoined points. In practice, if an approximately linear cluster of n points exists then it is straightforward to find it using the information gained from the hypothesis tests described by (2.3). As an aid in this process one can choose $\mathcal{S}$ to maximize the probability that, under $H_0(\mathcal{S})$, the inequality at (2.3) is satisfied. This allows us to estimate the orientation of the line on which the points approximately lie, and hence to estimate the speed at which the associated seismic events are propagated down the fault. See Section 3 for discussion.

The first part of (4.3) asks that n be of larger order than the standard deviation of our estimator of the mean number of points that would lie in the axial substrip, $\mathcal{A}$, if that number were Poisson-distributed and the strip were of width δ. Indeed, under the Poisson assumption, and the condition in Theorem 4.1 that the intensity of points per unit area is approximately ℓ, the expected number of points in a strip with dimensions $\delta \times \beta h$, centred near x, is approximately $\ell \delta h \beta f(x)$, and the variance is equal to the square root of this

quantity. The second part of (4.3) asks that n not be less than a small constant multiple of the lower level, $v = w \log \ell$, of the critical point $\max\{x_u(\hat{\lambda} \| \mathcal{A}\|), v\}$ at (2.3).

Next we show that the level of performance described by Theorem 4.2 is optimal, in the sense that no method can detect 'contamination' of $\mathcal{P}$ by n points for an order of magnitude smaller value of n than allowed by (4.3). For simplicity, take $\mathcal{D}$ to be a unit square, and divide it into m columns and m rows, of sizes $a = b = h = m^{-1}$. As this notation suggests, we are viewing each square subdivision of $\mathcal{D}$ as a test strip $\mathcal{S}$. The axial substrips of the test strips are parallel to the rows of $\mathcal{D}$. Denote their common widths by δ. Let $h = h(\ell)$ and $\delta = \delta(\ell)$ satisfy the conditions imposed in Theorems 4.1 and 4.2:

$$h = O\left(\ell^{-1/6-\varepsilon}\right), \qquad \ell^{1-\varepsilon} h^2 \to \infty, \qquad \ell \delta h \to \infty, \qquad \delta/h \to 0, \tag{4.4}$$

for some $\varepsilon > 0$. Assume too that

$$\text{the Poisson process in the plane is homogeneous with intensity } \ell. \tag{4.5}$$

That is, in the notation of Theorem 4.1, we take $f \equiv 1$. Thus, the number of points within each axial substrip is Poisson-distributed with mean $\ell \delta h$.

To each of the m^2 $\delta \times h$ axial substrips we decide with probability $\frac{1}{2}$ to add a further n points, and decide with probability $\frac{1}{2}$ not to augment the Poisson-distributed points already in the substrip, making the decision independently for each substrip. We know from Theorem 4.1 that the probability that our test can identify the fact that no new points have been added to any of the axial substrips, given that none have, converges to 1 as $\ell \to \infty$. Likewise, we know from Theorem 4.2 that if $n = n(\ell)$ satisfies (4.3) then the probability that our test can correctly identify the fact that n points have been distributed in one of the axial substrips, given that they have, also converges to 1 as $\ell \to \infty$. Theorem 4.3 below provides a converse to Theorem 4.2. The second part of Theorem 4.3 holds for a general technique, not just the one discussed in Theorems 4.1 and 4.2. In the theorem the addition of n points to one of the $h \times h$ regions is referred to as a 'contamination'.

Theorem 4.3. *Assume conditions* (4.4) *and* (4.5)*; in the definition of the critical point at* (2.3), *take $u > 0$ and $v = w \log \ell$ for $w > 0$; and take f, in the notation of Theorem 4.1, to be identically equal to 1. If $n = n(\ell)$ satisfies* (4.3), *and if u, w are sufficiently large and η (at* (4.3)) *sufficiently small, then, applying the test at* (2.3) *to each of the m^2 test strips, for each $k \geqslant 1$,*

$$\sup_{1 \leqslant j \leqslant k} \{1 - P(\text{algorithm correctly detects each contamination and each}$$
$$\text{non-contamination} \mid \text{there are exactly } j \text{ contaminations})\} \to 0 \tag{4.6}$$

as $\ell \to \infty$. Conversely, suppose that for some algorithm, (4.6) *holds. Then n satisfies the first part of* (4.3). *If, in addition, $(\ell \delta h)^{-1} \log \ell$ is bounded, then the second part of* (4.3) *is also true, for some $\eta > 0$.*

5. Technical arguments

5.1. Proof of Theorem 4.1

Without loss of generality, $\mathcal{D}$ is a square; in the contrary case, replace $\mathcal{D}$ by a square region containing the original $\mathcal{D}$. Let $d > 0$ be a constant, let $v = v(\ell)$ denote the integer part of ℓ^d, and place v regularly spaced points along each side of $\mathcal{D}$, with one point at either end of each side. The set of point pairs that results is a $v \times v$ lattice within $\mathcal{D}$; call it $\mathcal{D}_d = \mathcal{D}_d(\ell)$. Likewise, divide $\Theta = [0, \pi]$, $\mathcal{L} = [\alpha_1 h, \alpha_2 h]$, $\mathcal{M} = [\beta_1 h, \beta_2 h]$ and $\mathcal{N} = [0, \gamma h]$ into respective lattices Θ_d, $\mathcal{L}_d$, $\mathcal{M}_d$ and $\mathcal{N}_d$ of v regularly spaced points. Let S (S_d) denote the set of all test strips that are centred at a point in $\mathcal{D}$ (in $\mathcal{D}_d$), are inclined at an angle in Θ (in Θ_d), whose widths and lengths equal elements of $\mathcal{L}$ and $\mathcal{M}$ (of $\mathcal{L}_d$ and $\mathcal{M}_d$), and whose axial substrip widths are elements of $\mathcal{N}$ (of $\mathcal{N}_d$).

Given a test strip $\mathcal{S} \in S$, and $D \geqslant 2$, let $\mathcal{A}$ denote the axial substrip of $\mathcal{S}$, and let $\mathcal{T}(\mathcal{A})$ be the set of all points in the plane that are no further distant than ℓ^{-D} from the perimeter of $\mathcal{A}$. We may choose $d = d(D)$ so large that for each $\mathcal{S} \in S$ with axial substrip $\mathcal{A}$, there exists $\mathcal{S}' \in S_d$, with axial substrip $\mathcal{A}'$, such that $\mathcal{A} \Delta \mathcal{A} \subseteq \mathcal{T}(\mathcal{A}')$, where $\mathcal{A} \Delta \mathcal{A}'$ denotes the symmetric difference of $\mathcal{A}$ and $\mathcal{A}'$. In this case,

$$N(\mathcal{A}) \leqslant N(\mathcal{A}') + N(\mathcal{A} \Delta \mathcal{A}') \leqslant N(\mathcal{A}') + N\{\mathcal{T}(\mathcal{A}')\}. \tag{5.1}$$

Let $K \geqslant 1$ be an integer. Since $D \geqslant 2$, the probability that $\mathcal{T}(\mathcal{A}')$ contains at least $K + 1$ points of $\mathcal{P}$ equals $O(\ell^{-K-1})$, uniformly in $\mathcal{S}' \in S_d$, as $\ell \to \infty$. Since the number, $\#S_d$ say, of elements of S_d increases with ℓ only polynomially fast as ℓ increases, then if $K = K(d)$ is sufficiently large, $\#S_d = O(\ell^K)$. In this case, the probability that, for some $\mathcal{S}'$ in S_d, $\mathcal{T}(\mathcal{A}')$ contains at least $K + 1$ points equals $O(\ell^{-1})$ as $\ell \to \infty$. Therefore, by (5.1), it suffices to show that for any fixed integer $k \geqslant 1$, the probability that $N(\mathcal{A}') + k \geqslant \max\{x_u(\hat{\lambda} \|\mathcal{A}'\|), v\}$ for some $\mathcal{S}' \in S_d$ with axial substrip $\mathcal{A}'$, converges to 0. For this it is sufficient to prove that if $k \geqslant 1$ and $K > 0$ are given, and $u, w > 0$ are sufficiently large,

$$\sup_{\mathcal{S} \in S_d} P\left[N(\mathcal{A}') + k \geqslant \max\left\{x_u(\hat{\lambda} \|\mathcal{A}'\|), w \log \ell\right\}\right] = O(\ell^{-K}). \tag{5.2}$$

Of course, in this formula $\hat{\lambda}$ is computed from data in the two non-axial substrips, $\mathcal{B}'_1$ and $\mathcal{B}'_2$ say, within $\mathcal{S}'$.

Let the random variable $M(\lambda)$ have a Poisson distribution with parameter $\lambda > 0$. Given $x > 0$, let m denote the least integer not less than $\lambda + \lambda^{1/2}x$. In particular, $m \geqslant \lambda + \lambda^{1/2}x$. Using Stirling's formula, and taking C_1 and C_2 to be positive absolute constants, we may prove that

$$P\left\{M(\lambda) \geqslant \lambda + \lambda^{1/2}x\right\} = P\{M(\lambda) \geqslant m\}$$

$$\leqslant \frac{\lambda^m}{m!}e^{-\lambda}\sum_{j=0}^{\infty}\left(1 + \lambda^{-1/2}x\right)^{-j} = \left(1 + \lambda^{-1/2}x\right)\frac{\lambda^{m+(1/2)}}{m!\,x}e^{-\lambda}$$

$$\leqslant C_1\left(1 + \lambda^{-1/2}x\right)\frac{\lambda^{m+1/2}}{m^{m+1/2}\,x}e^{m-\lambda}$$

$$\leqslant C_2\,x^{-1}\left(1 + \lambda^{-1/2}x\right)^{1/2-(\lambda+\lambda^{1/2}x)}\exp\left(\lambda^{1/2}\,x\right)$$

$$= C_2\,x^{-1}\left(1 + \lambda^{-1/2}x\right)^{1/2}\exp\left\{-\lambda\,\xi\left(x/\lambda^{1/2}\right)\right\}, \qquad (5.3)$$

where $\xi(t) \equiv (1 + t)\log(1 + t) - t = \frac{1}{2}t^2 + O(|t|^3)$ as $t \to 0$.

More simply, if m is an integer then

$$\sup_{0<\lambda\leqslant e} P\{M(\lambda) \geqslant m\} \leqslant C_3/m! \leqslant C_4\exp\{-m(\log m - C_4)\},$$

for absolute constants $C_3, C_4 > 0$. The latter formula implies that if $m \geqslant (1 + a)\log \ell$, where $a > 0$, then

$$\sup_{0<\lambda\leqslant e} P\left[M(\lambda) > \max\left\{\lambda + a(\lambda\log_* \lambda)^{1/2}, m\right\}\right]$$

$$\leqslant \sup_{0<\lambda\leqslant e} P\{M(\lambda) > m\} \leqslant C_4\exp\{-m(\log m - C_4)\} = O(\ell^{-K})$$

for all $K > 0$. From this result and (5.3) we deduce that if $K > 0$ is given and $a = a(K)$ is sufficiently large,

$$\sup_{0<\lambda<\infty} P\left[M(\lambda) > \max\left\{\lambda + a(\lambda\log_* \lambda)^{1/2}, (1 + a)\log \ell\right\}\right] = O(\ell^{-K}). \qquad (5.4)$$

Given a subset C of the plane, put $\lambda(C) = \ell \int_C f(x)\,dx$, denoting the expected number of points of $\mathcal{P}$ that lie within C. Recall that $\mathcal{A}$ denotes the axial substrip of the test strip $\mathcal{S}$, and that $\mathcal{B}_1$ and $\mathcal{B}_2$ are the substrips that sandwich $\mathcal{A}$. Put $\hat{\lambda}_j = N(\mathcal{B}_j)/\|\mathcal{B}_j\|$ and $\hat{\lambda}(\mathcal{S}) = \max(\hat{\lambda}_1, \hat{\lambda}_2)$. Since f has two bounded derivatives then

$$\max\left\{\frac{\lambda(\mathcal{B}_1)}{\|\mathcal{B}_1\|}, \frac{\lambda(\mathcal{B}_2)}{\|\mathcal{B}_2\|}\right\} \geqslant \frac{\lambda(\mathcal{A})}{\|\mathcal{A}\|} + O(\ell h^2),$$

uniformly in test strips $\mathcal{S}$ satisfying the conditions of Theorem 4.1. Moreover, if $K > 0$ is given, and if $C_5 = C_5(K) > 0$ is sufficiently large, then by (5.3),

$$\sup_{\mathcal{S}\in S_d}\max_{j=1,2} P\left[|\hat{\lambda}_j - \lambda(\mathcal{B}_j)| > C_5\{\lambda(\mathcal{B}_j)\log \ell\}^{1/2}\right] = O(\ell^{-2K}).$$

From these results, and the fact that $\lambda(\mathcal{B}_j)$ is bounded below by a constant multiple of ℓh^2, and $\ell^{1-\varepsilon}h^2 \to \infty$ for some $\varepsilon > 0$, we deduce that for some $C_6 > 0$,

$$\sup_{\mathcal{S} \in S_d} P\left[\|\mathcal{A}\|^{-1}\left\{\hat{\lambda}(\mathcal{S})\,\|\mathcal{A}\| - \lambda(\mathcal{A})\right\} \leqslant -C_6\,\ell h^2\right] = O(\ell^{-K}).$$

Since $h = O(\ell^{-1/6-\varepsilon})$ then $\ell h^2\,\|\mathcal{A}\| = o\{\lambda(\mathcal{A})^{1/2}\}$, uniformly in $\mathcal{S} \in S_d$. Hence, for each $\varepsilon > 0$,

$$\sup_{\mathcal{S} \in S_d} P\left[\lambda(\mathcal{A})^{-1/2}\left\{\hat{\lambda}(\mathcal{S})\,\|\mathcal{A}\| - \lambda(\mathcal{A})\right\} \leqslant -\varepsilon\right] = O(\ell^{-K}). \tag{5.5}$$

Result (5.2) follows from (5.4) and (5.5).

5.2. Proof of Theorem 4.2

Let δ be as in the statement of the theorem, and let $\mathcal{S}$ denote the test strip whose axial substrip, $\mathcal{A}$, is the $\delta \times b$ strip within which the additional n points are located. The presence of these points increases $N(\mathcal{A})$ by n, but has no impact on the value of $\hat{\lambda}(\mathcal{S})\,\|\mathcal{A}\|$. When computed solely from points in $\mathcal{P}$, $N(\mathcal{A})$ is asymptotically normal with mean $\lambda(\mathcal{A})$ and standard deviation $\lambda(\mathcal{A})^{1/2}$, and moreover, the value of $\hat{\lambda}(\mathcal{S})\,\|\mathcal{A}\|$ is not less than $\lambda(\mathcal{A}) + \lambda(\mathcal{A})^{1/2} + o_p\{\lambda(\mathcal{A})^{1/2}\}$. (The proof of the latter property is similar to that leading to (5.5).) Note too that $\lambda(\mathcal{A})/\ell\delta h$ is bounded away from zero and infinity as $\ell \to \infty$, and so by choice of δ, $n/\lambda(\mathcal{A})^{1/2} \to \infty$. Therefore,

$$P\left[N(\mathcal{A}) \geqslant x_u\{\hat{\lambda}(\mathcal{S})\,\|\mathcal{A}\|\}\right] \to 1, \tag{5.6}$$

where the probability is calculated under the model in which the n points are added to $\mathcal{P}$ as prescribed by Theorem 4.2. More simply, if $w > 0$ and if η in (4.3) satisfies $\eta < w^{-1}$ then $P\{N(\mathcal{A}) \geqslant w \log \ell\} = 1$. This property and (5.6) imply that the probability that the hypothesis test described at (2.3), when applied to our particular choice of $\mathcal{S}$, leads to rejection, converges to 1 as $\ell \to \infty$.

Acknowledgements

References

Bakun, W.H. and Lindh, A.G. (1985) The Parkfield, California, earthquake prediction experiment. *Science*, **229**, 619–624.

Bakun, W.H. and McEvilly, T.V. (1979) Earthquakes near Parkfield, California: Comparing the 1934 and 1966 sequences. *Science*, **205**, 1375–1377.

Bakun, W.H. and McEvilly T.V. (1984) Recurrence models and Parkfield, California earthquakes. *J. Geophys. Res.*, **89**, 3051–3058.

Broadbent, S. (1980) Simulating the ley hunter. *J. Roy. Statist. Soc. Ser. A*, **143**, 109–140.

Cressie, N. (1977) On some properties of the scan statistic on the circle and the line. *J. Appl. Probab.*, **14**, 272–283.

Danuser, G. and Stricker, M. (1998) Parametric model fitting: From inlier characterization to outlier detection. *IEEE Trans. Pattern Anal. Machine Intelligence*, **20**, 263–280.

Desolneux, A., Moisan, L. and Morel, J.M. (2003a) Maximal meaningful events and applications to image analysis. *Ann. Statist.*, **31**, 1822–1851.

Desolneux, A., Moisan, L. and Morel, J.M. (2003b) Computational gestalts and perception thresholds. *J. Physiol. Paris*, **97**, 311–324.

Fletcher, M. and Lock, G. (1981) Computerised pattern perception within posthole distributions. *Sci. Archaeol.*, **22**, 1520.

Fletcher, M. and Lock, G. (1984) Post built structures at Danebury Hilfort: An analytical search method with statistical discussion. *Oxford J. Archaeol.*, **3**, 175–196.

Fletcher, M. and Lock, G. (1991) *Digging Numbers: Elementary Statistics for Archaelogists.* Oxford: Oxbow.

Frigui, H. and Krishnapuram, R. (1999) A robust competitive clustering algorithm with applications to computer vision. *IEEE Trans. Pattern Anal. Machine Intelligence*, **21**, 450–465.

Gates, J. (1986) Measures and tests of alignment. *Biometrika*, **73**, 731–734.

Goodall, C. (1991) Procrustes methods in the statistical-analysis of shape. *J. Roy. Statist. Soc. Ser. B*, **53**, 285–339.

Hall, P., Tajvidi, N. and Malin, P. (2005) Locating lines among scattered points. http://www.maths.lth.se/matstat/staff/nader/papers/htm.pdf.

Kendall, D.G. (1984) Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.*, **16**, 81–121.

Kendall, D.G. (1985) Exact distributions for shapes of random triangles in convex sets. *Adv. in Appl. Probab.*, **17**, 308–329.

Kendall, D.G. (1986) Further developments and applications of the statistical theory of shape. *Teor. Veroyatnost. i Primenen.*, **31**, 467–473.

Kendall, D.G. and Kendall, W.S. (1980) Alignments in two-dimensional random sets of points. *Adv. in Appl. Probab.*, **12**, 380–424.

Kent, J.T., Briden, J.C. and Mardia, K.V. (1983) Linear and planar structure in ordered multivariate data, as applied to progressive demagnetization remanence. *Geophys. J. Roy. Astronom. Soc.*, **75**, 593–621.

Kulldorff, M., Athas, W.F., Feurer, E.J., Miller, B.A. and Key, C.R. (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *Amer. J. Public Health*, **88**, 1377–1380.

Lindh, A.G. and Malin, P.E. (1987) *Introduction to Parkfield Earthquake Studies: Instrumentation and Geology.* El Cerrito, CA: Seismological Society of America.

Mack, C. (1950) The expected number of aggregates in a random distribution of n points. *Proc. Cambridge Philos. Soc.*, **46**, 285–292.

Malin, P.E. and Alvarez, M.G. (1992) Stress diffusion along the San Andreas fault at Parkfield, CA. *Science*, **256**, 1005–1007.

Malin, P.E., Oancea, V.G., Shalev, E. and Tang, C. (2002) A friction-feedback model for microearthquake sequences at Parkfield, California. Manuscript.

McEvilly, T.V., Bakun W.H. and Casady, K.B. (1967) The Parkfield, California earthquakes of 1966. *Bull. Seismol. Soc. Amer.*, **57**, 1221–1244.

Meer, P., Stewart, C.V. and Tyler, D.E. (2000) Robust computer vision: An interdisciplinary challenge.

Comput. Vision Image Understanding, **78**, 1–7.

Priebe, C.E., Olson, T.H. and Dennis, M. Jr. (1997) A spatial scan statistic for stochastic scan partitions. *J. Amer. Statist. Assoc.*, **92**, 1476–1484.

Saunders, I.W. (1978) Locating bright spots in a point process. *Adv. in Appl. Probab.*, **10**, 587–612.

Segall, P. and Harris, R. (1987) Earthquake deformation cycle on the San Andreas fault near Parkfield, California. *J. Geophys. Res.*, **92**, 10 511–10 525.

Silverman, B. and Brown, T. (1978) Short distances, flat triangles and Poisson limits. *J. Appl. Probab.*, **15**, 815–825.

Small, C.G. (1982) Random uniform triangles and the alignment problem. *Math. Proc. Cambridge Philos. Soc.*, **91**, 315–322.

Small, C.G. (1984) A classification theorem for planar distributions based on the shape statistics of independent tetrads. *Math. Proc. Cambridge Philos. Soc.*, **96**, 543–547.

Small, C.G. (1988) Techniques of shape analysis on sets of points. *Internat. Statist. Rev.*, **56**, 243–257.

Small, C.G. (1996) *The Statistical Theory of Shape.* New York: Springer-Verlag.

Stewart, C.V. (1995a) MINIPRAN – new robust estimator for computer vision. *IEEE Trans. Pattern Anal. Machine Intelligence*, **17**, 925–938.

Stewart, C.V. (1995b) Robust parameter estimation in computer vision. *Comput. Vision Image Understanding*, **76**, 54–69.

Susaki, J., Hara, K., Kajiwara, K. and Honda, Y. (2004) Robust estimation of BRDF model parameters. *Remote Sensing of Environment*, **89**, 63–71.

Stuart, W.D., Archuleta, R.J. and Lindh, A.G. (1985) Forecast model for moderate earthquakes near Parkfield, California. *J. Geophys. Res.*, **90**, 592–604.

Weinstock, M.A. (1981) A generalised scan statistic for the detection of clusters. *Internat. J. Epidemiol.*, **10**, 289–293.