DIMACS Working Group on Adverse Event/Disease Reporting, Surveillance, and Analysis II

Group discussion on methods for monitoring multiple data streams to detect incidence of disease.

The detection problem

Multiplicity is present in this problem in many forms. We are interested in events, such as release of a toxic agent, which affect the population's health. The variety of possible agents and mode of release mean that an event may be of several types, and it can happen at any time and location. There are also multiple data streams forming the source of information from which we are to deduce whether an event has taken place.

When an event occurs, its subsequent effects will depend on various factors which affect dispersal, including population movement and weather conditions (e.g., wind direction). While deterministic models may be available for daily population movement, it is natural to treat weather as a random factor (a more sophisticated approach would be to add observed weather information to the input data). Thus, the question of what one should look for in the data to detect an event is complex.

One approach to combining data streams is to monitor individual streams separately, then, when unusual phenomena are seen in one type of data, to look for supporting evidence in others. This raises questions about the nature of the joint effect each possible event type will have on the set of data streams. In assessing the "false alarm rate", the degree of association between different data streams must also be considered.

In this report we summarise discussion of an alternative approach in which joint modelling of the full set of data streams is undertaken in order to create statistically powerful detectors of event incidence. There are pre-cursors to this in the treatment of multiple endpoints in statistical process control and in the sequential monitoring of clinical trials. In these areas, models are established for the joint behaviour of a number of observed endpoints, including the effect of an "intervention", i.e., the process becoming out of control in the first example or a successful treatment effect in the latter. It is then a very natural step, and in statistical terms the most efficient option, to define monitoring procedures using the whole set of endpoints together. The challenge in our detection problem is to do something similar while keeping the whole procedure manageable.

Problem formulation and detection algorithm

Let X denote the complete collection of data from all the different streams, recorded at the finest level of detail. In principle, this comprises a set of time series for each data-type, recorded at numerous locations throughout the country. If we wish, X can be extended to include underlying events arising in models of the spread of disease but which are not directly observable.

For several reasons we shall monitor a compressed form, Y, of these data. As noted above, some elements of X may not really be observable. Certain summaries of the complete data may be more rapidly available than others. Finally, a monitoring process based on key elements of information will be more manageable, and computationally tractable.

The basis of our approach is a model for how Y behaves over time

(a) when no incident has occurred, and

(b) when there has been an incident.

Case (a) is still quite diverse as levels of, say, coughs, colds and influenza vary through the year, often with random outbreaks in localised areas. This is an important qualitative difference from the common assumption in industrial process control that measurements continue to follow a fixed "in control" distribution until an event occurs. Case (b) is certainly diverse: there are possible types of event, time of occurrence, and location. In the analogy to process control, we shall be testing a large group of hypotheses, each stating that there has been a recent incident of a certain type, at a given place and a given time.

Let H_0 denote the null hypothesis that no incident has occurred and let $f_0(y;t)$ be the probability density of the (compressed) data y available up to time t. Let the hypothesis H_{ijk} be that an event of type i occurred at location j and time k, and let λ_{ijk} be a parameter defining the scale of this event, with $\lambda_{ijk} = 0$ if no such event has occurred and $\lambda_{ijk} > 0$ otherwise. Considering just one triple (i, j, k) for the moment, we wish to test the null hypothesis H_0, under which $\lambda_{ijk} = 0$, against the alternative H_{ijk}: $\lambda_{ijk} > 0$.

Suppose we have a model for the consequences of an event (i, j, k) which gives a density $f_{ijk}(y; t, \lambda_{ijk})$ for observations y, agreeing with $f_0(y;t)$ up for $t < k$ and then diverging as time passes beyond the incident time $t = k$. In the analogous situation in statistical process control, a CUSUM chart uses a sequential probability ratio test, rejecting H_0 in favour of H_{ijk} as soon as

$$\frac{f_0(y;t)}{f_{ijk}(y;t,\tilde{\lambda}_{ijk})} > C, \quad t > k, \tag{1}$$

where $\tilde{\lambda}_{ijk}$ is a pre-chosen size for the shift from control conditions and the critical value C controls the rate of false alarms.

Rule (1) remains an option but it is possible to generalise the stopping criterion, drawing on methods for group sequential testing developed, in particular, for clinical trials. A more general rule would be to reject H_0 in favour of H_{ijk} as soon as

$$g_{ijk}(y;t) > C_{ijk}(t), \quad t > k, \tag{2}$$

where the test statistic $g_{ijk}(y;t)$ is based on the cumulative information over the period k to t on departures from H_0 in the direction to be expected under H_{ijk}. It is desirable to optimise the speed of detection for the most relevant values of λ_{ijk}. This needs the multivariate densities $f_0(y;t)$ and $f_{ijk}(y;t,\lambda_{ijk})$ to be clearly defined in order to choose a suitable test statistic $g_{ijk}(y;t)$ and critical values $C_{ijk}(t)$ giving an acceptable false alarm rate. Importantly, these densities also determine the rate of accrual of information (in the precisely defined statistical sense) about whether or not an incident has occurred at a given time and place.

The full procedure is then the super-position of tests of H_0 against all alternatives H_{ijk}. Thus, at time t, one considers the possibility of incidents of each type, at all times $k < t$, and at all locations. This is similar to monitoring a scan statistic but now the dimensions are space, time and event type. The overall false alarm rate is the rate of reaching the first alarm when, in fact, no incident occurs. Given the inter-connections between tests for detecting all the different H_{ijk}s, simulation is liable to be needed to assess the overall false alarm rate. The choice of critical values $C_{ijk}(t)$ in rules (2) governs the play between the false alarm rate and speed of detection of actual events (but bear in mind the interpretation of "false alarm" discussed in Remark 1 below).

Choosing and modelling Y

The above approach relies on having a suitable choice of the response vector Y and models for how Y will behave both under normal conditions and following a specified type of incident. Where a detailed model is available for the larger vector X, one is faced with the task of deducing a tractable model for the chosen Y, most likely on the basis of simulation data. In other situations, a distribution may have to be deduced directly from observational data and specialist knowledge of the inter-relations between certain variables.

It is clear that choosing and modelling Y is a major task. It would be appealing to find simple, approximate models in which Y is analytically tractable. An ideal case would be a low-dimensional multivariate normal distribution with a shift in mean under H_{ijk} of λ_{ijk} times a known unit vector $\boldsymbol{\delta}_{ijk}$ under H_{ijk}. In particular,

this would make it easier to derive effective group sequential tests of the form (2). The shift in mean vector has an appealing interpretation here, namely, a template of how the occurrence of an event will affect the recorded data in a spatial pattern, evolving as time elapses after the event.

Data-reduction is a key part of this task. If starting from a model for the very large response variable X, it is important to identify those parts of X which are most clearly affected by the types of event to be detected. Simplicity is also of value as there is a danger that over-precise predictions of the consequences of one type of event might mask a major incident following a slightly different pattern. Basing methods on a small set of key variables would appear to be a good path towards transparent decision rules which do not depend heavily on fine details of model assumptions.

Remarks

1. *Event types*. The notion of an event type should be interpreted broadly. Given that one is trying to guard against the unexpected, it may well be advisable to include one or more types of event defined to produce rather general effects across data streams. The lack of fine tuning of the models for such events will make rapid detection difficult but adding these will at least ensure power over a wider range of possible incidents.

2. *False alarms*. The term "false alarm" should be taken to mean an initial warning which will lead to more detailed examination of data currently in hand plus other investigations. A public announcement would only follow if further evidence backed up the initial findings. Thus, a suitable rate for false alarms may be higher than one might initially expect.