STOCHASTIC METHODS AND THEIR APPLICATIONS

Papers in honour of CHRIS HEYDE
Edited by J. GANI and E. SENETA

Part 6. Stochastic processes

SCORING PROBABILITY FORECASTS FOR POINT PROCESSES: THE ENTROPY SCORE AND INFORMATION GAIN

DARYL J. DALEY, Australian National University
Centre for Mathematics and Its Applications, Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia. Email address: daryl.daley@anu.edu.au

DAVID VERE-JONES, Victoria University of Wellington and Statistics Research Associates
School of Mathematical and Computing Sciences, Victoria University of Wellington, PO Box 600, Wellington, New Zealand. Email address: david.vere-jones@mcs.vuw.ac.nz
SCORING PROBABILITY FORECASTS
FOR POINT PROCESSES: THE ENTROPY SCORE AND INFORMATION GAIN

DARYL J. DALEY
DAVID VERE-JONES

Abstract

The entropy score of an observed outcome that has been given a probability forecast \( p \) is defined to be \(-\log p\). If \( p \) is derived from a probability model and there is a background model for which the same outcome has probability \( \pi \), then the log ratio \( \log(p/\pi) \) is the probability gain, and its expected value the information gain, for that outcome. Such concepts are closely related to the likelihood of the model and its entropy rate. The relationships between these concepts are explored in the case that the outcomes in question are the occurrence or nonoccurrence of events in a stochastic point process. It is shown that, in such a context, the mean information gain per unit time, based on forecasts made at arbitrary discrete time intervals, is bounded above by the entropy rate of the point process. Two examples illustrate how the information gain may be related to realizations with a range of values of ‘predictability’.

Keywords: Entropy score; probability gain; information gain

2000 Mathematics Subject Classification: Primary 60G25
Secondary 62M20

1. Background

Both of us have had long associations with Chris Heyde, as consequences of location and editorial work. It is a pleasure to acknowledge his friendship and assistance, both personal and professional, during this long period. While Chris may be best known to the wider scholastic community for his editorial work, whether from the Applied Probability journals or occasional publications like Statisticians of the Century, we should like in this paper to recognize the significant contributions that Chris has made to the topic of inference in stochastic processes. His forays into analogues of results developed in a likelihood setting differ from what we investigate here, where the stimulus has come from recent developments in earthquake forecasting. These have brought into prominence the need for assessing and comparing the performance of probability forecasts for earthquakes. Since many models in this field are point process models, this raises the question of developing and assessing probability forecasts for point process models in general, topics which appear to have been relatively little considered. One aim of this paper is therefore to review briefly some issues in the assessment of probability forecasts for point processes. We illustrate our work with results for some simple renewal and queueing process models; these too have featured in Chris’s work through time—and the occasion here is likewise of time: happy birthday Chris.

The particular approach that we describe below to the assessment of probability forecasts is based on the use of the entropy score: if the observed outcome has forecast probability \( p \), it is given the score \(-\log p\); this score is a measure of the unpredictability of the particular
observed outcome. The expected value of this score, assuming that the true probabilities are used in the forecast, is then just the entropy of the distribution of possible outcomes. We shall interpret this expectation as a measure of the unpredictability of the model on which the forecasts are based. Both these aspects of the entropy score are important in applications. A convenient score is directly useful in empirically comparing the performance of different forecasting schemes or the quality of forecasts for different classes of events. At the model level, theoretical evaluations of the information gain (the expected difference of entropy scores—see below) can help to give insight into the relative predictive powers of different models, or the reasons why the actual performance of a given model might differ from its theoretical value.

Even in fields where probability forecasts are well established, such as weather forecasting, this approach appears to be relatively novel. The traditional methods of assessing such forecasts rely on dichotomizing both the outcomes and the forecasts, so reducing the problem to the assessment of a $2 \times 2$ (or similar) table. While other schemes for assessing probability forecasts certainly exist, the idea of scoring the forecasts in terms of their log probabilities appears to have been relatively little explored, despite its links to such familiar concepts as likelihood and entropy.

These links, however, allow us to develop some important properties of the entropy score. This requires a reshaping of earlier work on point process entropy such as McFadden (1965), Rudemo (1964), Fritz (1969), (1973) and Papangelou (1978). The important first step is simply to recognize that many basic results from information theory can be recast in forecasting terms by interpreting ‘uncertainty’ as ‘unpredictability’. From this starting point, the theory can be directed towards the quantification of ‘predictability’ and its assessment from observational data.

One important result is the existence of an upper bound for the expected difference in entropy scores—we call this the information gain—between forecasts based on the true model, and those based on a reference or null model. The forecasts have to be made in discrete time; the bound relates to the underlying process in continuous time.

Another by-product of this approach is to clarify the perception that the likelihood of a model contains within it an indication of the model’s forecasting performance. This has been part of the forecasting folklore for many years. In the earthquake context, for example, it appears in several early papers by Y. Y. Kagan—see for example Kagan and Knopoff (1977)—and more recent work by Imoto (2000). It is also closely linked to ideas used earlier by Akaike (1973), (1974) in developing the Akaike information criterion (AIC), although, in addition, the AIC takes into account the bias caused by using the same data both to estimate model parameters and then to evaluate the model. However, there is no indication in his papers that Akaike used the entropy score as such in developing his ideas; his motivation seems to have come rather from consideration of the expected mean square error of the forecast. Of course, this derives immediately from the log likelihood ratio in the Gaussian case, and to this extent can be considered a variation of the same general theme.

An initial exploration of these ideas was given by Vere-Jones (1998), and they are reviewed more broadly if somewhat informally in Daley and Vere-Jones (2003, Section 7.6). A second aim of this paper is thus to set these results on a firmer theoretical footing. Sections 2 and 3 largely review background ideas while Section 4 outlines some important approximation results. The final section give examples that provide a ‘visual predictability scale’, illustrating the associated changing information gains.
2. Entropy score, probability gain and information gain

2.1. Basic definitions

The simplest setting for the type of forecasting problems considered in this paper is a sequence of dependent trials ordered in time. The dependence relations can then be summarized by writing down the conditional distribution of the current trial given the results of previous trials (and, in general, of additional conditioning variables). Regardless of whether the background process is in continuous or discrete time, forecasts are made in practice at discrete times separated by longer or shorter time intervals. Consequently, such forecasts are for a sequence of discrete trials, and an underlying issue is to relate the performance of such forecasts to properties of the underlying point process.

Following the notation of Daley and Vere-Jones (2003), consider a set of discrete trials, each of which can take any one of a discrete set of ‘marks’ $\kappa$, labelled for convenience $1, \ldots, K$. Let $X_{\kappa,n}$ denote the indicator variable for the event that the $n$th trial results in outcome $\kappa$, with corresponding forecast probability $p^*_{\kappa,n}$. If the $n$th trial results in outcome $Y_n$, then the entropy score is $-\log p^*_{Y_n,n}$. If the $p^*_{\kappa,n}$ happen to be the true probabilities $p_{\kappa,n}^{(0)}$, the expected value of the score is then the entropy of the $n$th trial, that is,

$$E_n \equiv \text{entropy of } n\text{th trial} = E(-\log p^*_{Y_n,n}) = -\sum_{\kappa=1}^{K} p_{\kappa,n}^{(0)} \log p_{\kappa,n}^{(0)}.$$  

As mentioned in the introduction, we regard the entropy as a measure of the unpredictability of the outcome distribution, rather than its uncertainty, although the ideas are essentially equivalent. In any case, it is a characteristic of the process generating the $n$th trial, and is always positive, although possibly infinite if $K = \infty$.

Often, the requirement is to compare the performance of the forecast probabilities with the performance of some reference set of probabilities, for example, those derived from a model of independent and identically distributed (i.i.d.) trials. In the latter case, if the reference probabilities are denoted by $\pi_{\kappa}$ for $\kappa = 1, \ldots, K$, then the difference between the two scores can be written in the form

$$D_n = \sum_{\kappa=1}^{K} X_{\kappa,n} \log \frac{p^*_{\kappa,n}}{\pi_{\kappa}}.$$  

The ratio $p^*_{\kappa,n}/\pi_{\kappa}$ is the probability gain when the observed outcome $Y_n$ takes the value $\kappa$. The probability gain itself is a familiar concept in earthquake forecasting studies—see Aki (1989) for example. The expected value of the log probability gain, again assuming that the forecast probabilities are the true ones, is the information gain, or Kullback–Leibler distance, between the probability measures $P_n^{(0)}$ and $\Pi$ for the true and reference models respectively:

$$\mathcal{D}(P_n^{(0)}, \Pi) = \sum_{\kappa=1}^{K} p_{\kappa,n}^{(0)} \log \frac{p_{\kappa,n}^{(0)}}{\pi_{\kappa}}.  \tag{1}$$  

It is essential here that $p_{\kappa,n}^{(0)} > 0$ only if $\pi_{\kappa} > 0$, i.e. that $P_n^{(0)} \ll \Pi$, a condition we shall assume implicitly throughout our discussion.

In general, there is no requirement that the probability forecasts be based on the true, or indeed on any, probability model in the formal sense. In such a general case, the expected value
of the difference in the entropy scores can be written as the information gain for trial \( n \),

\[
J_n = \mathbb{E}[D_n] = \sum_{\kappa=1}^{K} p_{\kappa,n}^{(0)} \log \frac{p_{\kappa,n}^{*}}{\pi_{\kappa}}
\]

\[
= \sum_{\kappa=1}^{K} p_{\kappa,n}^{(0)} \log \frac{p_{\kappa,n}^{(0)}}{\pi_{\kappa}} - \sum_{\kappa=1}^{K} p_{\kappa,n}^{(0)} \log \frac{p_{\kappa,n}^{(0)}}{p_{\kappa,n}^{*}}
\]

\[= \mathcal{D}(P_{\kappa}^{(0)}, \Pi) - \mathcal{D}(P_{\kappa}^{(0)}, P^{*}). \tag{2}\]

The quantity \( J_n \) is a measure of the improvement in forecast performance in moving from the background probabilities \( \pi_{\kappa} \) to the forecast probabilities \( p_{\kappa,n}^{*} \), even though neither probability is based on the true model.

Since the Kullback–Leibler distance vanishes only when the two distributions coincide, the highest gain is achieved when the forecasts are based on the true probabilities. Then the second term on the right-hand side of (2) vanishes, and the information gain reduces to (1).

### 2.2. Stationary sequences of trials

When the trials are dependent, forming part of a stochastic process evolving in time, the ‘true’ probabilities become the conditional probabilities \( \Pr\{Y_n = \kappa \mid F_{n-1}\} = p_{\kappa,n|F_{n-1}} \) say, where the \( \sigma \)-algebra \( F_{n-1} \) is part of a history, or filtration, \( \mathcal{F} = \{F_n : n = 0, 1, \ldots\} \) to which the trials are adapted. In this case \( J_n \equiv J_{n|F_{n-1}} \) should be interpreted as the conditional information gain, given \( F_{n-1} \); the expected information gain itself is obtained by taking a further expectation over the past histories. Let us assume in addition that the process generating the trials is stationary. Then this further expectation reduces to a constant, independent of \( n \), but dependent in general on the history \( F \). We denote this constant by \( G_{\mathcal{F}} \), and call it the expected information gain per trial, under the history \( \mathcal{F} \); thus

\[
G_{\mathcal{F}} = \mathbb{E}[J_n|F_{n-1}] = \mathbb{E}[J_1|F_0].
\]

This quantity, which is related to the entropy rate per trial in the information theory context, is the key measure of the predictability of the sequence of trials. Its value depends crucially on the history \( \mathcal{F} \), summarizing the information to which we have access in developing the forecasts. The most straightforward situation is that in which \( \mathcal{F} \) is the internal history \( \mathcal{H} \) generated by the outcomes of the trials themselves. More generally, the outcome probabilities may be determined not just by the outcomes of previous trials, but by observations on additional explanatory variables, or precursors in the earthquake context. As the amount of precursory information increases, we might imagine the history to which we have access as being one in a sequence of histories of increasing richness, from complete randomness (unpredictability) to a limit situation in which the trials become perfectly predictable.

The increase in gain in moving from a basic history \( \mathcal{F} \) to a richer history \( \mathcal{G} \) is

\[
G_{\mathcal{G}} - G_{\mathcal{F}} = \mathbb{E}\left[ \sum_{\kappa=1}^{K} p_{\kappa,n|F_{n-1}} \log \frac{p_{\kappa,n|F_{n-1}}}{\pi_{\kappa}} \right] - \mathbb{E}\left[ \sum_{\kappa=1}^{K} p_{\kappa,n|F_{n-1}} \log \frac{p_{\kappa,n|F_{n-1}}}{p_{\kappa,n}^{*}} \right]
\]

\[= \mathbb{E}\left[ \sum_{\kappa=1}^{K} p_{\kappa,n|F_{n-1}} \log \frac{p_{\kappa,n|F_{n-1}}}{p_{\kappa,n|F_{n-1}}} \right]. \tag{3}\]

This is just the gain for the finer predictions, taking the coarser predictions as reference measure.
2.3. Relationship to the likelihood and rough goodness-of-fit

Summing the log probability gains over a sequence of trials $1, \ldots, T$ and dividing by $T$ gives the mean log probability gain per trial. But the mean log probability gain is nothing other than the time average of the log likelihood ratio for the model $P^*$ against the model $\Pi$. All we are doing here is writing out the joint probability of the observed sequence of outcomes as a chain product of conditional probabilities, namely the probabilities of the observed outcomes of successive trials, each conditional on the outcomes of the preceding trials, and taking the average:

$$\frac{1}{T} \log L = \frac{1}{T} \sum_{n=1}^{T} \log \frac{p^*_Y}{\pi_Y}.$$ (4)

This provides a natural and simple link between the log likelihood and the log probability gains. The two are essentially equivalent. Selecting a model to maximize the likelihood ratio is roughly equivalent to selecting a model to maximize the average value of the log probability gain, or the entropy score. This is because, under typical ergodicity assumptions, the mean log likelihood ratio will be close to its expected value, which is just the information gain per trial. Moreover, roughly speaking, the selected model will be closer to the true model than its competitors, in the sense of the Kullback–Leibler distance.

In practice, the mean log likelihood ratio can be used not only to discriminate between models, but also to provide a rough internal check on a given model. If the observed information gain, calculated from the mean log likelihood ratio, falls well below that predicted (analytically or from simulation studies) by the model, it is an indication that the real data differ substantially from typical data produced from the model. Like any other numerical characteristic of the model, the expected information gain per trial, in the stationary case, can be used as the basis of a goodness-of-fit test, with null distribution found analytically or by simulation from realizations of the model.

Of course the link between the sample value of the log likelihood and its expected value is the key to many discussions both in inference and in information theory. Here we are simply concerned to point out the additional interpretation of such results in terms of probability gains and entropy scores.

3. Point process entropy

3.1. Entropy of a finite point process

Fundamental to any discussion of finite point processes is the recognition that a finite point process on a set $A$ can be thought of as a random variable taking its values on the portmanteau space $A^1$ that consists of a single point (corresponding to the empty realization) and the union of product sets $A(r)$ for $r = 1, 2, \ldots$. Thus, the dimension as well as the size of the components may vary from realization to realization. Strictly speaking, the components should be thought of as quotient spaces with respect to permutations of the axes (Macchi (1975), Daley and Vere-Jones (2003, Chapter 5)). This feature is important in selecting a reference measure to define the entropy of a point process, as distinct from the entropy of a process of distinguishable particles.

The likelihood of a realization of a finite point process on $A$ can written down as a product of two terms: a discrete probability distribution $\{p_r(A) : r = 0, 1, \ldots\}$, governing the number of points in $A$, $N(A)$ say, and a distribution giving the joint probability of finding $r$ points around $\{x_1, \ldots, x_r\}$ within $A$, given that $N(A) = r \geq 1$. Assuming absolute continuity with respect to Lebesgue measure, and bearing in mind that the points are indistinguishable, the latter
term can be written in the form
\[ \sum_{r=0}^{\infty} p_r \log(r! p_r) + \sum_{r=1}^{\infty} p_r \int_{A^{(r)}} s_r \log s_r \, dx_1 \ldots dx_r, \]
where the \( s_r \) are symmetric probability densities over \( A^{(r)} \). The entropy of the point process on the set \( A \) can then be found by taking expectations of the negative log likelihood, and takes the form
\[ H = - \left[ \sum_{r=0}^{\infty} p_r \log(r! p_r) + \sum_{r=1}^{\infty} p_r \int_{A^{(r)}} s_r \log s_r \, dx_1 \ldots dx_r \right], \tag{5} \]
where for brevity we have suppressed the arguments in the densities. The extra factor \( \log r! \), compared with the corresponding representation for sets of distinguishable points, represents the loss of information, for given \( r \) and a given set of locations, about which particle is located at which location. Under the assumption of indistinguishable points, there are \( r! \) permutations from which to choose, all of them equally likely, corresponding to a distribution with entropy \( \log r! \).

On the other hand, the information gain is obtained by taking expectations of the log likelihood ratio, and is the same whether the points are treated as distinguishable or indistinguishable. It is given by
\[ G = p_0 \log \frac{p_0}{p_0} + \sum_{r=1}^{\infty} p_r \left[ \log \frac{p_r}{p_0} + \int_{A^{(r)}} s_r \log s_r s_0 \, dx_1 \ldots dx_r \right]. \]
These two expressions illustrate the difference between an absolute or generalized entropy as at (5), giving the entropy relative to a measure reflecting the structure of the space on which it is defined, and the information gain, which compares the entropies of two probability measures defined within similar structural constraints.

Returning to (5), suppose that the expected number \( M = M(A) = E[N(A)] \) of points in \( A \) is fixed. The first sum in (5) is maximized by taking \( p_r = (M'/r)e^{-M} \), while for given \( r \) the second sum is maximized by taking \( s_r = 1/[(\ell(A))^r] \), where \( \ell(\cdot) \) denotes Lebesgue measure of dimension determined by the context. Thus, the maximizing distribution is just the uniform distribution over \( A \), and the homogeneous Poisson process on \( A \) appears as the finite point process with maximum entropy for a given expected number of points in \( A \). The corresponding probability distribution on \( A \cup \) then provides a natural reference measure for assessing the information gain for point processes on \( A \).

Rudemo (1964) and McFadden (1965) obtained (5). Fritz (1969), (1973) studied the entropy of a general finite point process with probability measure \( P \), using the generalized entropies of Csiszár (1969). Fritz also established a link between these generalized entropies and what may be called the Rényi dimension of a finite point process. Suppose that we are given a totally finite nonatomic measure \( \mu \) on \( (A, \mathcal{B}_A) \). Define a nonnormalized reference measure \( Q \) on \( (A \cup, \mathcal{B}_A \cup) \) by omitting the exponential factors \( e^{-\mu(B)} \) from the finite-dimensional probabilities for a Poisson process with parameter measure \( \mu \) on \( A \), so that if, for example, \( E = \{ \omega : N(B) = r \} \), then \( Q(E) = [\mu(B)]^r / r! \).

For any finite partition \( \mathcal{T} \) of \( A \), let \( \delta(\mathcal{T}) = \max_{A_i \in \mathcal{T}} \mu(A_i) \). Then as \( n \to \infty \) through a dissecting system \( \mathcal{T}_n \) on \( A, \delta(\mathcal{T}_n) \to 0 \) because \( \mu(\cdot) \) is nonatomic and, as shown in Fritz (1973),
\[
\sum_{r_1 \geq 0, \ldots, r_n \geq 0} \mathcal{P}(N(A_i) = r_i, i = 1, \ldots, k_n) \log \mathcal{P}(N(A_i) = r_i, i = 1, \ldots, k_n) + E[N(A)] \log \delta(\mathcal{T}_n)
\to -H(\mathcal{P}, \mathcal{Q}), \quad \text{say.} \tag{6}
\]
The role played by \( \mathbb{E}[N(A)] \) here is analogous to that played by the Rényi dimension (Rényi (1959), (1960)) for probability distributions on a finite-dimensional Euclidean space. Thus, \( \mathbb{E}[N(A)] \) can be interpreted as the Rényi dimension of the point process. This is not altogether surprising: if \( A \) is an interval in \( \mathbb{R}^d \) and we consider the event \( \{N(A) = k\} \), then the Rényi dimension of an absolutely continuous distribution for the \( k \) points in \( A \) is \( kd \). Taking expectations over \( N(A) \) gives the ‘average’ Rényi dimension of the point process as \( d \mathbb{E}[N(A)] \). This corresponds to the result in (6) if we take \( \mu \) to be Lebesgue measure in \( \mathbb{R}^d \) and take the sets \( A_i \) in the partition to be \( d \)-dimensional cubes with edges of length \( \epsilon \), so that \( \log \mu(A_i) = d \log \epsilon \).

### 3.2. Representations in terms of the conditional intensity

In an evolutionary context, where there is a time-like dimension ordering the points in the realization, the results for both entropies and their discrete approximations can be significantly extended and simplified. The key concept in these simplifications is the conditional intensity function, \( \lambda(t, \omega) \). If the simple point process \( N(t) \equiv N(0, t] \) is adapted to the history \( \mathcal{F} = \{ \mathcal{F}_t : t \geq 0 \} \), then, heuristically,

\[
\lambda(t) \, dt \approx \mathbb{E}[dN(t) \mid \mathcal{F}_t],
\]

where for brevity we omit the explicit dependence on the sample point \( \omega \). The use of conditional intensities is closely linked to martingale ideas which play an important role in the general theory of point processes (e.g. Brémaud (1981), Liptser and Shiryaev (2001)). Their application to entropy rates was introduced by Grigelionis (1974), (1975). Papangelou (1978) gives a thorough discussion of entropy rates for stationary point processes, based on ideas of conditional intensities.

In the case of a renewal process, the conditional intensity function at \( t \) reduces to the hazard function evaluated for the time interval back to the most recent event before \( t \). More generally, it can be considered as a hazard function conditioned by the information in a history \( \mathcal{F} \) up to the present time \( t \).

When the likelihood is re-expressed in terms of the hazard function, and hence of the conditional intensity, it takes the outward form of the likelihood for a Poisson process, much as successive conditioning allows the likelihood for a sequence of dependent trials to retain the product form it would have if the trials were independent (cf. (4)). In particular, the log likelihood ratio between a point process with \( \mathcal{F} \)-conditional intensity \( \lambda(\cdot) \) and a stationary Poisson process of rate \( \mu \) can be written as

\[
\log L = \sum_{i=1}^{N(T)} \log \frac{\lambda(t_i)}{\mu} - \int_0^T [\lambda(t) - \mu] \, dt
= \int_0^T \left\{ \log \frac{\lambda(t)}{\mu} \, dN(t) - [\lambda(t) - \mu] \, dt \right\},
\]

where the \( \{t_i : i = 1, \ldots, N(T)\} \) are the observed occurrence times of events over the observation window \((0, T]\).

Taking expectations gives an alternative expression for the mean log probability gain for the point process over the interval \((0, T]\) in the form

\[
G(\mathcal{P}, \mathcal{P}^0; \mathcal{F}) = \mathbb{E} \left\{ \int_0^T \lambda(t) \log \frac{\lambda(t)}{\mu} \, dt \right\} - \int_0^T [m(t) - \mu] \, dt,
\]

(7)
Taking expectations gives the total information gain over the interval 

\[ T \] at time \( t \).

### 3.3. Marked point processes

The representations of the log likelihood and entropy in terms of conditional intensity extend easily to marked point processes \( \{ (t_i, \kappa_i) \} \). The conditional intensity function here has the heuristic interpretation

\[
\lambda(t, \kappa) dt \ell(\mathrm{d}x) \approx E[N(\mathrm{d}t \times \mathrm{d}x) \mid \mathcal{F}_{t-}],
\]

(8)

where we have assumed that the marks \( \kappa_i \) have absolutely continuous distributions in some Euclidean space \( \mathcal{K} = \mathbb{R}^d \) with Lebesgue measure \( \ell(\cdot) \). The times \( t_i \) form the ground process for the point process, denoted by \( N_{\mathcal{G}}(\cdot) \). Assuming that the ground process is simple (i.e. there are no multiple marks), these times constitute the set \( \{ t : N([t \times \mathcal{K}] > 0 \}. \) The mark density \( f(\kappa \mid t) \) at time \( t \) should be construed as the density of a distribution conditional on the history \( \mathcal{F}_t \), so, in (8), \( \lambda(t, \kappa) = \lambda_g(t) f(\kappa \mid t) \) (Daley and Vere-Jones (2003, Equation (7.3.4))). The history may contain information from external variables as well as the times and marks of previous points of the process.

Extending our earlier arguments from the case of an unmarked simple point process, the log likelihood for a simple marked point process with observations \( \{ t_i, \kappa_i \} \) over an observation region \( (0, T] \times \mathcal{K} \) can be written as

\[
\log L = \sum_{i=1}^{N_{\mathcal{G}}(T)} \log \frac{\lambda_g(t_i)}{\mu} - \int_0^T [\lambda_g(t) - \mu] \mathrm{d}t + \sum_{i=1}^{N_{\mathcal{G}}(T)} \log \frac{f(\kappa_i \mid t_i)}{\phi(\kappa_i)},
\]

where the reference process is taken as a compound Poisson process on \( (0, T] \) with constant Poisson intensity \( \mu \) and i.i.d. marks with density \( \phi(\kappa) \) with respect to Lebesgue measure on \( \mathcal{K} \). Taking expectations gives the total information gain over the interval \( (0, T] \) as

\[
G(\mathcal{P}, \mathcal{P}^0; \mathcal{F}) = E \left[ \int_0^T \lambda_g(t) \log \frac{\lambda_g(t)}{\mu} \mathrm{d}t \right] - \int_0^T [m_g(t) - \mu] \mathrm{d}t + E \left[ \int_0^T \left[ \int_{\mathcal{K}} \frac{f(\kappa \mid t)}{\phi(\kappa)} \mathrm{d}k \right] \lambda_g(t) \mathrm{d}t \right].
\]

### 4. Discrete approximations

#### 4.1. A basic lemma

We turn next to the approximation of point process entropies and information gains by observations on an approximating system of discrete trials. We suppose that the observation interval \( (0, T] \) is partitioned into subintervals \( A_i = (u_{i-1}, u_i] \) for \( i = 1, \ldots, k_n \), with \( u_0 = 0 \), \( u_{k_n} = T \), for which forecasts are required, and that we observe not the individual points but only either the values \( X_i = N(A_i) \) or the more limited set of indicator variables \( Y_i = I_{\{N(A_i)>0\}} \).
Our aim is to show how the information gain and associated quantities for the point process can be approximated by the analogous quantities for the $X_t$ or the $Y_t$, both of which may be thought of as sequences of dependent trials.

In order to imbed this situation into a limit process, we suppose that the given partition is a member of a dissecting family of partitions, say $\mathcal{T}_n$ for $n = 1, 2, \ldots$. A dissecting family is defined by the properties (see Kallenberg (1986, Section 1.1), Daley and Vere-Jones (2003, Section A1.6)) that the subsets $A_i$ of each partition are Borel measurable, that the partitions are nested and that the family separates points as $n \to \infty$, so that ultimately any two distinct time points $u', u''$ end up and remain in distinct subsets of the partitions. Similarly, in a marked point process, we consider a given discretization of the time interval and the product space as a member of a dissecting family of partitions of the space $(0, T] \times \mathcal{K}$, the partitions themselves being of product form $\mathcal{T}_n \times \mathcal{U}_n$.

The crucial partition, however, is not the partition of the state space (whether $(0, T]$ or $(0, T] \times \mathcal{K}$) but the partition this induces on the probability space $\Omega$. Consider a partition $\mathcal{T} \times \mathcal{U}$ of $(0, T] \times \mathcal{K}$, comprising $r$ divisions of the time interval and $s$ subdivisions of the mark space, yielding $r \times s$ sets overall. If we introduce the conventional mark 0 to denote the occurrence of no events within the subinterval under consideration, such a partition induces a partition of $\Omega$ into $(s + 1)^r$ events, each corresponding to a sequence of possible outcomes from the $r$ subintervals of the observation interval. In the case of the $Y_i$ alone, we have $s = 1$ and the partition of $\Omega$ corresponds to the set of all $2^r$ possible outcome sequences from a set of $r$ Bernoulli trials.

To consider the effects of refining the partition, take first a simple, unmarked point process and, for some fixed $n_0$, choose a fixed element $A \in \mathcal{T}_{n_0}$. For $n > n_0$, the associated processes

$$\eta(n)(A) = \sum_{i: A_i \leq A} Y_i = \sum_{i: A_i \leq A} 1_{\{N(A_i) > 0\}}$$

count the numbers of subintervals of $\mathcal{T}_n$ in $A$ which contain a point of the process. It is clear that, for increasing $n$, the $\eta(n)$ are nondecreasing. Indeed, because of the dissecting property and the assumed simplicity of the process, $\eta(n)(A) \uparrow N(A)$. Thus, any event $\{N(A) = k\}$, or any finite union or intersection of such events, can be approximated by the corresponding events $\{\eta(n)(A) = k\}$.

Similarly, for a marked point process with a simple ground process, take $A$ and $K$ from one of the product partitions $\mathcal{T}_{n_0} \times \mathcal{U}_{n_0}$ and consider the associated random variables

$$\zeta(n)(A \times K) = \sum_{i: A_i \leq A} \sum_{j: K_j \leq K} 1_{\{N(A_i) \times K_j > 0\}}.$$

Once again we find that $\zeta(n)(A \times K) \uparrow N(A \times K)$.

Now let $\mathcal{B}^{(n)} = \sigma(\mathcal{T}_n \times \mathcal{U}_n)$ denote the (finite) $\sigma$-algebra of sets in $(0, T] \times \mathcal{K}$ generated by the partitions up to and including the $n$th stage. Similarly, let $\mathcal{A}^{(n)}$ denote the corresponding finite $\sigma$-algebra of events (subsets of $(\Omega, \mathcal{E})$) generated by $\zeta^{(n)}(A \times K)$ for $A \times K \in \mathcal{B}^{(n)}$. The following lemma is basic.

**Lemma 1.** For a marked point process whose ground process is simple, the processes $\zeta^{(n)}(A \times K)$ suffice to generate the internal history $\mathcal{H}_{(0, T]}$, that is, $\mathcal{H}_{(0, T]} = \sigma(\bigvee \mathcal{A}^{(n)})$.

**Proof.** The properties of the dissecting sequences have two consequences. First, the collection of sets $\mathcal{A}^{(n)}_i \times \mathcal{K}_i^{(n)}$, over all $n$, is rich enough to generate the Borel $\sigma$-algebra in $(0, T] \times \mathcal{K}$. 

Scoring probability forecasts for point processes

305
Second, for any finite family of sets in this collection, events defined by the \( N(A \times K) \) can be approximated by the corresponding events defined by the \( \zeta^{(n)}(A \times K) \). The lemma follows from standard extension theorems.

4.2. Generalized entropies and information gains

We start by recalling the definition of generalized entropy in Csiszár (1969) and Fritz (1969).

Suppose that we are given a probability measure \( \mathcal{P} \) on \((\Omega, \mathcal{E})\) and a reference measure \( \mathcal{Q} \) (not necessarily a probability measure) with respect to which \( \mathcal{P} \) is absolutely continuous: \( \mathcal{P} \ll \mathcal{Q} \) on some \( \sigma \)-algebra \( \mathcal{A} \subseteq \mathcal{E} \). Following Csiszár (1969), if \( \frac{d\mathcal{P}}{d\mathcal{Q}} \) denotes the Radon–Nikodym derivative on \( \mathcal{A} \), we define the generalized entropy by

\[
H(\mathcal{P}, \mathcal{Q}; \mathcal{A}) = -\int_{\Omega} \frac{d\mathcal{P}}{d\mathcal{Q}} \log \frac{d\mathcal{P}}{d\mathcal{Q}} \mathcal{Q}(d\omega) = -\int_{\Omega} \log \frac{d\mathcal{P}}{d\mathcal{Q}} \mathcal{P}(d\omega). \tag{10}
\]

This corresponds to the standard definition for the entropy of an absolutely continuous distribution in \( \mathbb{R}^d \) when \( \mathcal{Q} \) is the corresponding Lebesgue measure (which we shall denote by \( \ell \), with the dimension \( d \) to be understood from the context).

The \( \sigma \)-algebra \( \mathcal{A} \) plays an important role in this definition. It may be the full \( \sigma \)-algebra of events \( \mathcal{E} \), or some sub-\( \sigma \)-algebra, or the \( \sigma \)-algebra, say \( \mathcal{F}_x \), generated by a partition \( \delta \) of \( \Omega \) into \( \mathcal{F} \)-measurable sets.

As already noted below (5), when \( \mathcal{Q}(\cdot) \) is itself a probability measure, we prefer to use the same expression as the generalized entropy (10), but with the sign changed. We then write

\[
G(\mathcal{P}, \mathcal{P}^0; \mathcal{A}) = \int_{\Omega} \frac{d\mathcal{P}}{d\mathcal{P}^0} \log \frac{d\mathcal{P}}{d\mathcal{P}^0} \mathcal{P}^0(d\omega) = \int_{\Omega} \log \frac{d\mathcal{P}}{d\mathcal{P}^0} \mathcal{P}(d\omega) \geq 0,
\]

and interpret the last integral as an information gain, or as a Kullback–Leibler distance; it is set equal to \(+\infty\) if in fact \( \mathcal{P} \not\ll \mathcal{P}^0 \), but the integral can diverge even when the absolute continuity condition does hold.

The following two properties of the generalized entropy, taken from Csiszár (1969) as in Fritz (1969) but restated here in terms of information gains, are of crucial importance.

**Property 1.** The information gain is nondecreasing under refinement of the \( \sigma \)-algebra \( \mathcal{A} \).

This property is a corollary of the convexity of the function \( x \log x \) for \( x > 0 \). In particular, if \( \delta_1 \) and \( \delta_2 \) are two fine partitions of \( \Omega \), with \( \delta_1 \subseteq \delta_2 \) and which generate the \( \sigma \)-algebras \( \mathcal{F}_1 \), \( \mathcal{F}_2 \) respectively, then \( G(\mathcal{P}, \mathcal{P}^0; \mathcal{F}_1) \leq G(\mathcal{P}, \mathcal{P}^0; \mathcal{F}_2) \). Equation (3) is another example.

**Property 2.** Let \( \{\mathcal{A}_a\} \) be a family of \( \sigma \)-algebras generating \( \mathcal{A} \) and such that, for every \( \mathcal{A}_{a_1}, \mathcal{A}_{a_2} \), there exists an \( \mathcal{A}_{a_0} \in \{\mathcal{A}_a\} \) for which \( \mathcal{A}_{a_1} \cup \mathcal{A}_{a_2} \subseteq \mathcal{A}_{a_0} \). Then

\[
G(\mathcal{P}, \mathcal{P}^0; \mathcal{A}) = \sup_a G(\mathcal{P}, \mathcal{P}^0; \mathcal{A}_a). \]

Suppose, in particular, that \( \mathcal{A}^{(n)} \) is the set of events derived from the \( \zeta^{(n)}(A) \) associated with a dissecting family of partitions as in Lemma 1. Then Properties 1 and 2 taken in conjunction with Lemma 1 imply the following.

**Proposition 1.** Let \( N(\cdot) \) be a marked point process defined on \((0, T] \times \mathcal{K})\), with simple ground process, probability measure \( \mathcal{P} \), internal history \( \mathcal{F} \) and finite expected information gain \( G = G(\mathcal{P}, \mathcal{P}^0; \mathcal{H}_{(0, T]}) \) against some reference measure \( \mathcal{P}^0 \). Also let \( \mathcal{T}_n \times \mathcal{U}_n \) be a dissecting family of partitions of \((0, T] \times \mathcal{K})\). Then the expected information gain \( G_n \), associated with the partition \( \mathcal{T}_n \times \mathcal{U}_n \), satisfies both \( G_n \leq G \) (for all \( n \)) and \( G_n \uparrow G \) as \( n \to \infty \).
4.3. Extensions to predictable partitions and general histories

Up until now we have supposed that the end points in the partitions of \((0, T]\) were fixed times, but the bound given in Proposition 1 extends also to predictable partitions, that is, to situations where the successive subintervals \(\Delta_i = (u_{i-1}, u_i]\) are random but \(\Delta_i\) is \(\mathcal{H}_{u_{i-1}}\)-predictable. A condition must be imposed to ensure that the partition remains finite with probability 1 (e.g. the successive interval lengths remain bounded below with probability 1). Then, for each \(\Delta_i\), the probability distribution of \(N(\Delta_i)\) can be conditioned on \(\mathcal{H}_{u_{i-1}}\) in the same way as if the interval \(\Delta_i\) were fixed. Consequently, the joint distribution can be written down by the usual chain rule. Moreover, if the interval \(\Delta_i\) is further subdivided, say at \(\tau\) with \(u_{i-1} < \tau < u_i\), and \(\tau\) is likewise required to be \(\mathcal{H}_{u_{i-1}}\)-predictable, then the chain rule continues to apply, the partition in \(\Omega\) is refined, and Property 1 shows that the information gain is nondecreasing. It is simplest to suppose that, apart from their dependence on the initial random partition, the refinements form a dissecting family with fixed partition points. Conditional on the initial partition, therefore, the bounds in Proposition 1 continue to hold.

Such an extension shows that the bound given in Proposition 1 cannot be relaxed by selecting the length of the next partition interval on the basis of information already to hand.

An even more important extension is to situations where the forecasts may depend on external variables as well as the internal history, or equivalently where the process is \(\mathcal{F}\)-adapted to some history \(\mathcal{F}\) more general than its internal history. We have only partial results concerning this situation. Conditional intensities and the form of the likelihood ratio in (9) still hold (see e.g. Brémaud (1981, Chapter VI), Liptser and Shiryaev (2001), Jacod and Shiryaev (1987)), but in general we lose the simple behaviour of the discrete approximations embodied in Proposition 1. One underlying reason is that a great variety of situations are covered in the comparison of the forecasting performance of two point process models, each dependent in some way on external variables. For example, in some situations, the dependence on the external variables may be scale dependent; in such a case, relative to the true model, a reference model may perform better on small scales than on larger scales. Refining the partition, therefore, need not always result in an improvement in the information gain.

There is one important situation, however, where relatively straightforward statements can be made, namely, where the reference model \(\mathcal{P}^0\) is taken to be the projection of the full model (involving both the point process of interest and the external variables) onto the internal history \(\mathcal{H}\) of the point process. That is to say, the reference model is defined by the \(\mathcal{H}\)-conditional intensity of the point process, say \(\lambda(\mathcal{H})\).

Suppose that for the full model there exists a regular family of conditional probabilities for the point process, given the external variables; then, for every realization \(y\) of the external variables, there exists a probability distribution \(\mathcal{P}^y\) for the point process, given \(y\). Associated with this conditional distribution is a conditional information gain \(G(\mathcal{P}^y, \mathcal{P}^0; \mathcal{H}(0,T))\). It is a random variable since \(y\) is random, and we can show that

\[
E[G(\mathcal{P}^y, \mathcal{P}^0; \mathcal{H}(0,T))] \geq 0. \tag{11}
\]

Such a statement implies that the expected information gain, taking into account the explanatory variables, can never be less than the information gain calculated without them. It is the analogue, for information gains, of the basic inequality for entropies,

\[
H(X \mid Y) \leq H(X).
\]

The inequality (11) can also be applied to show that, in this situation, refinement of the partitions cannot decrease the information gain. Consider the effect of introducing a further
partition point \( \theta \) inside a subinterval \((u_{i-1}, u_i]\) as considered earlier. Suppose first that for the time interval \( \mathcal{H}_{[0,T]} \) we take just the interval \((u_{i-1}, u_i]\), and that both histories in (11) incorporate information on the point process and the explanatory variables up to time \( u_{i-1} \). Then, in the first new subinterval \((u_{i-1}, \theta]\), no additional information is available to the given process that is not also available to the reference process, and so the information gain over this part of the interval is equal to zero, as it is for the interval as a whole. In the second subinterval \((\theta, u_i]\), however, the given process is able to incorporate information about both the point process and the explanatory variables in \((u_{i-1}, \theta]\), whereas the reference process can incorporate only the additional information about the point process. Then (11), applied now to the second subinterval, implies that an additional potential gain occurs.

The inequality is only sharpened if account is taken also of the additional information available to the full process as distinct from the reference process in the earlier subintervals. In this special situation, therefore, the information gains increase with refinement of the partitions, and the continuous limit provides an upper bound as in Proposition 1.

5. Examples and a predictability scale

There are several families of point processes for which the information gain can be computed or even expressed in closed analytical form. Such families can be used to provide a visual illustration of the degree of predictability represented by different values of the information gain. In this section we use renewal processes with gamma-distributed lifetime distributions to illustrate the predictability of univariate point processes, and a simple type of bivariate Poisson process to illustrate some of the effects on the information gain of providing supplementary information.

5.1. A predictability scale for univariate point processes

In the univariate case, assuming that only the internal history is available for prediction purposes, the ‘more predictable’ the point process (the larger the value of \( G \)) the more closely it resembles a deterministic point process, for example, a process with constant interevent distances. Referring to (7), a simple stationary point process has information gain per unit time against a Poisson process with rate \( m \) expressible as

\[
G = \mathbb{E} \left( \lambda^*(t) \log \frac{\lambda^*(t)}{m} \right),
\]

where \( \lambda^*(\cdot) \) denotes the stationary conditional intensity function (cf. Daley and Vere-Jones (2003, Example 7.6(b))). For a renewal process with lifetime distribution \( F(\cdot) \), tail (or survivor function) \( \bar{F}(\cdot) \), density \( f(\cdot) \) and mean \( 1/m = \int_0^\infty \bar{F}(u) \, du \), \( \lambda^*(t) = f(B_t)/\bar{F}(B_t) \) where \( B_t \) denotes a (stationary) backward recurrence time random variable; \( B_t \) has density \( m\bar{F}(\cdot) \). Then

\[
G = \int_0^{\infty} \frac{f(y)}{\bar{F}(y)} \log \left( \frac{f(y)}{m\bar{F}(y)} \right) m\bar{F}(y) \, dy.
\]

Integrating by parts,

\[-\int_0^{\infty} f(y) \log \bar{F}(y) \, dy = -(-\bar{F}(y)) \log \bar{F}(y) \bigg|_{y=0}^{\infty} + \int_0^{\infty} (-\bar{F}(y)) \frac{-f(y)}{\bar{F}(y)} \, dy = 1,
\]

so, as in Equation (7.6.16) of Daley and Vere-Jones (2003),

\[
G = m \left[ 1 - \log m + \int_0^{\infty} f(y) \log f(y) \, dy \right],
\]
We are interested in computing the integral in (15) equals

Table 1 lists some values of the information gain $G$ truncated to terms of the same order.

Suppose the lifetime distribution has the gamma density function

We are interested in computing

where $\psi(z) = (d/dz) \log \Gamma(z) = \Gamma'(z)/\Gamma(z)$. Using integral representations for $\log \Gamma(\kappa)$ and $\psi(z)$ (e.g. 6.1.50 and 6.3.21 in Abramowitz and Stegun (1992)) or else exploiting known expansions for them for large $\kappa$ (e.g. 6.1.41 and 6.3.18 in Abramowitz and Stegun (1992)),

The integral in (15) equals $(1/\kappa)[1+o(1)]$ for large $\kappa$, as is consistent with the series expansion truncated to terms of the same order.

For such renewal processes with variable shape parameter $\kappa$ but a common mean, equal to 1 say, we should want $\mu = \kappa$, and then the first term on the right-hand side of (15) equals $\log 2\pi/\kappa$. The case $\kappa = 1$ gives a Poisson process, the term in $\psi$ vanishes and $H_1(1, 1) = -1$. Table 1 lists some values of the information gain $G = 1 + H_1(\kappa, \kappa)$ in the column with the delay $a$ set to 0.

Another family of distributions for which the integral in (13) is computable is the family of delayed gamma density functions, meaning that the density equals $f_\kappa(x-a; \mu)$ for $x > a$.
Table 1: Information gains for delayed and gamma-distributed intervals (mean = 1).

<table>
<thead>
<tr>
<th>κ</th>
<th>a = 0</th>
<th>a = 0.25</th>
<th>a = 0.5</th>
<th>a = 0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.15.72608</td>
<td>6.01376</td>
<td>6.41923</td>
<td>7.11237</td>
</tr>
<tr>
<td>0.2</td>
<td>1.89773</td>
<td>2.18541</td>
<td>2.59088</td>
<td>3.28402</td>
</tr>
<tr>
<td>0.4</td>
<td>0.42386</td>
<td>0.71154</td>
<td>1.11701</td>
<td>1.81015</td>
</tr>
<tr>
<td>0.6</td>
<td>0.10719</td>
<td>0.39487</td>
<td>0.80034</td>
<td>1.49348</td>
</tr>
<tr>
<td>0.8</td>
<td>0.01780</td>
<td>0.30548</td>
<td>0.71095</td>
<td>1.40409</td>
</tr>
<tr>
<td>1</td>
<td>0.00000</td>
<td>0.28768</td>
<td>0.69315</td>
<td>1.38629</td>
</tr>
<tr>
<td>1.5</td>
<td>0.04449</td>
<td>0.33217</td>
<td>0.73764</td>
<td>1.43078</td>
</tr>
<tr>
<td>2</td>
<td>0.11593</td>
<td>0.40361</td>
<td>0.80908</td>
<td>1.50222</td>
</tr>
<tr>
<td>3</td>
<td>0.25103</td>
<td>0.53871</td>
<td>0.94418</td>
<td>1.63732</td>
</tr>
<tr>
<td>4</td>
<td>0.36289</td>
<td>0.65057</td>
<td>1.05604</td>
<td>1.74918</td>
</tr>
<tr>
<td>6</td>
<td>0.53486</td>
<td>0.82254</td>
<td>1.22801</td>
<td>1.92115</td>
</tr>
<tr>
<td>8</td>
<td>0.66377</td>
<td>0.95145</td>
<td>1.35692</td>
<td>2.05006</td>
</tr>
<tr>
<td>10</td>
<td>0.76653</td>
<td>1.05421</td>
<td>1.45968</td>
<td>2.15282</td>
</tr>
<tr>
<td>15</td>
<td>0.95768</td>
<td>1.24556</td>
<td>1.65083</td>
<td>2.34397</td>
</tr>
<tr>
<td>20</td>
<td>1.09580</td>
<td>1.38348</td>
<td>1.78895</td>
<td>2.48209</td>
</tr>
<tr>
<td>30</td>
<td>1.29286</td>
<td>1.58054</td>
<td>1.96001</td>
<td>2.67915</td>
</tr>
<tr>
<td>40</td>
<td>1.45079</td>
<td>1.72157</td>
<td>2.12704</td>
<td>2.82018</td>
</tr>
<tr>
<td>50</td>
<td>1.54377</td>
<td>1.83145</td>
<td>2.23692</td>
<td>2.93006</td>
</tr>
</tbody>
</table>

and 0 otherwise, where $a$ is a finite positive constant which, if the lifetime has unit mean, must be constrained to $(0, 1)$ and $(κ, µ)$ must satisfy

$$a + \frac{κ}{µ} = 1.$$ 

Write $H_1(κ, µ, a)$ for the analogue of the integral in (14) with such $f_κ(x − a; µ)$. From inspection, $H_1(κ, µ, a) = H_1(κ, µ)$, so the shift $a$ enters only through the constraint on the mean. For example, a delayed exponential lifetime distribution with unit mean has $G = −\log(1 − a)$, providing an even simpler parametric family than the gamma density functions considered earlier. More generally, a delayed gamma density function as above with delay parameter $a$, shape parameter $κ$ and unit mean, has

$$H_1(κ, κ/(1 − a)) = H_1(κ, κ) − log(1 − a).$$

Some values of the corresponding information gains $G = 1 + H_1(κ, κ) − log(1 − a)$ are shown in Table 1.

Observe from Table 1 that the Poisson process has zero information gain (i.e. for $(κ, a) = (1, 0), G = 0$), and that ‘predictability’, as indicated by larger values of $G$, increases as $κ \to \infty$ or $a \uparrow 1$ (this is associated with the interevent density function becoming more like a spike (delta function) at 1), and also for $κ \downarrow 0$, being associated there with the interevent distribution having a spike near $0+$ and otherwise a relatively long flat tail with weak exponential damping (because a mean 1 is still required). In the former case, increased predictability arises from approximate regularity of the interevent intervals, while in the latter case it is associated with clustering (these comments can be seen from inspection of Figure 1).
Figure 1: Parts of realizations of four different gamma-distributed renewal processes.

Figure 1 shows realizations of four renewal processes, each of 40 lifetimes that are gamma distributed with the same mean and respective shape parameters $\kappa = 0.2, 1, 5$ and $25$, and gains $G$ as shown (cf. Table 1; Poisson has $\kappa = 1$ and $G = 0$). The 41 vertical lines denoting ends of lifetimes are discernible for the three cases with $\kappa \geq 1$; for the case $\kappa = 0.2$ the 20 visibly distinct lines represent, respectively, actual ends of lifetimes.

5.2. A bivariate Poisson process

We consider finally the simple bivariate Poisson process $(N^I, N^H)$ used in Example 7.3(a) of Daley and Vere-Jones (2003) to illustrate the dependence of the conditional intensity on the history. Here the process $N^I$ is a unit rate Poisson process, as also is $N^H$ in the stationary regime, being constructed by delaying $N^I$ by exponentially distributed lags $\Delta$ with mean $1/\mu$, independently for each point. An equivalent description is that $N^H$ is the output process of a stationary $M/M/\infty$ queue.

Suppose that we wish to predict points in $N^H$ on the basis of the internal history (a) of $N^H$ alone, or else (b) of both $N^I$ and $N^H$. Under the former condition, $N^H$ is Poisson at unit rate and no positive information gain ensues. Under the latter condition, the situation changes because, if we let $Q(t)$ denote the number of points of $N^I$ for which the corresponding points at lags distributed like $\Delta$ have not yet appeared, then the stationary version of the conditional intensity equals $\mu Q(t)$. It is known that the stationary process $Q(t)$ is a birth-and-death process whose stationary distribution is Poisson with mean $E(Q(t)) = 1/\mu$. Appealing to (12), the information gain per unit time against a unit rate Poisson process (i.e. the process without the additional information) is thus given by

$$G = E(\mu Q(t) \log [\mu Q(t)])$$
$$= \log \mu + E(Q(t) \log Q(t))$$
$$= \log \mu + \mu e^{-1/\mu} \sum_{k=2}^{\infty} \frac{(1/\mu)^k k \log k}{k!}.$$  (16)

For $\mu$ large, the sum here is dominated by its first term, and $G = \log \mu + (\log 2)/\mu + O(\mu^{-2})$. For $\mu$ small, the last expectation in (16) is

$$\mu E(Q(t) \log (Q(t))) \approx \mu (E(Q(t))(\log E(Q(t))) = -\log \mu.$$
Table 2: Information gain in output of M/M/∞ queue with known queue size.

<table>
<thead>
<tr>
<th>μ</th>
<th>0.1</th>
<th>0.2</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.05094</td>
<td>0.10467</td>
<td>0.28485</td>
<td>0.57340</td>
<td>1.00702</td>
<td>1.74253</td>
<td>2.37049</td>
<td>3.03003</td>
</tr>
</tbody>
</table>

and therefore $G \approx 0$. Thus, $N^{II}$ is much more predictable given the full history of the bivariate process only for larger $\mu$, and hence for ‘small’ delays $\Delta$ following the occurrence of a point in $N^I$. Table 2 is based on (16).

References


