Decomposition of Event Sequences into Independent Components

Heikki Mannila [*] ***and Dmitry Rusakov*** [†]

1 Introduction

Many real-world processes result in an extensive logs of sequences of events, i.e., events coupled with time of occurrence. Examples of such process logs include alarms produced by a large telecommunication network, web-access data, biostatistics, etc. In many cases, it is useful to decompose the incoming stream of events into the number of independent streams. Such decomposition may reveal valuable information about the event generating process, e.g. dependencies among alarms in the telecommunication network, relationships between web-users and relevant symptoms of the decease. It may, as well, facilitate further analysis of the data by working with independent components separately.

In this paper we describe a theoretical framework and practical methods for finding event sequence decompositions. These methods use the probabilistic modeling of the event generating process. The probabilistic model predicts (with given confidence) the range of observed statistics for the independent subsequence event generation processes. The predicted values are used to determine the independence relations among the event types in the observed sequence of events, and these relations are used to decompose the sequences.

The presented techniques were validated by analyzing real data from a telecommunication network and on synthetic data that was generated under two different models. In the first dataset, the a priori event distribution was uniform, and in the second dataset events have followed a predefined burst-type a priori distribu-

[*] Nokia Research Center, e-mail: heikki.mannila@nokia.com
[†] Technion - Israel Institute of Technology, e-mail: rusakov@cs.technion.ac.il

tion. The algorithms were implemented under Matlab and a large number of event sequence analysis routines and visualization tools were written.

The area of decomposing sequences of events seems to be new to data mining. There are, of course, several topics in which related issues have been considered. The whole area of mixture modeling can even be viewed as finding independent components. For reasons of brevity, we omit discussions of these analogues in this version of the paper. We just mention the close similarity to work on independent component analysis (ICA), see, e.g., [2, 3, 6]. Our methods assume that the underlying event-generation process is stationary, i.e., does not change over time, and ergodic which means that we can draw general conclusions from the single, sufficiently long event log. We also assume a quasi-Markovian property of the observed process; i.e., the distribution of events in the particular time frame depends only on some finite neighborhood of this frame. Our approach resembles the marked point process modeling used in various fields, e.g., forestry [7].

We start by introducing the reader to the concept of event sequences and event sequence decomposition in Section 2 and we continue in Section 3 by presenting the underlying event-generating stochastic process and by explicitly mentioning the assumptions that are made about the properties of this process. Readers more interested in practical methods should skip Sections 3.1 and 3.2, and go straight to Section 3.3 and following subsections, which describe the proposed technique in details, and to Section 4, which presents experimental results on the telecommunications and synthetic data.

2 Event Sequences and Independent Events

The goal of this paper is to analyze event sequences and to partition the set of events into independent subsets. We first introduce the event sequence concept and then formulate the notion of independent event sequences. We follow definitions presented in [4].

2.1 Event sequences

We consider the input as a sequence of events, where each event has an associated time of occurrence. Given a set $E = \{e_1, \ldots, e_k\}$ of *event types*, an *event* is a pair (A, t), where $A \in E$ is an event type and $t \in \mathbb{N}$ is the occurrence time of the event. Note, that we often use the term event referring to the event type; the exact meaning should be clear from the context.

An *event sequence* s on E is an ordered sequence of events,

$$s = \langle (A_1, t_1), (A_2, t_2), \ldots, (A_n, t_n) \rangle \tag{1}$$

such that $A_i \in E$ for all $i = 1, \ldots, n$, and $t_i \in [T_s, T_e]$, $t_i \leq t_{i+1}$ for all $i = 1, \ldots, n-1$, where T_s, T_e are integers denoting the starting and ending time of the observation. Note that we can have $t_i = t_{i+1}$, i.e., several events can occur at the same time. However, we assume that for any $A \in E$ at most one event of type A occurs at any given time.

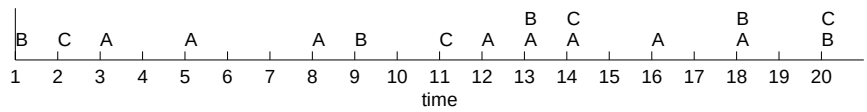


Figure 1. *Sequence of events A,B and C observed during* 20 *seconds.*

Given an event sequence s over a set of event types E, and a subset $E_1 \subseteq E$, the *projection* $s[E_1]$ of s to E_1 is the event sequence consisting of those events $\langle e, t\rangle$ from s such that $e \in E_1$. A *sub-sequence* of event e_i, denoted by s_{e_i}, is a subsequence of s consisting only of the events of type e_i from s, i.e., s_{e_i} is a projection of s onto $E_1 = \{e_i\}$.

Alternatively, we can view s as a function from the observed period, $[T_s, T_e]$, into $\{0,1\}^{|E|}$, and $\{s_{e_i}\}_{e_i \in E}$ as functions from $[T_s, T_e]$ into $\{0,1\}$, such that $s = s_{e_1} \times \ldots \times s_{e_k}$. In such formulation, $s(t)$ denotes the events that happened in the time unit t.

`EXAMPLE:` Figure 1 presents the event sequence of three event types $E = \{A, B, C\}$ observed for 20 seconds, that is $T_s = 1$, $T_e = 20$ and

$$s = \langle (B,1), (C,2), (A,3), (A,5), (A,8), \ldots, (B,20), (C,20)\rangle.$$

Note that a number of events of different types can occur in the same second. The subsequences of sequence s are shown on Figure 2 and they are

$$\begin{aligned} s_A &= \langle (A,3), (A,5), (A,8), \ldots, (A,18)\rangle \\ s_B &= \langle (B,1), (B,9), (B,13), (B,18), (B,20)\rangle \\ s_C &= \langle (C,2), (C,11), (C,14), (C,20)\rangle \end{aligned}$$

It can be seen that event C always follows event B with one or two seconds lag. The C event that follows $(B, 20)$ was not observed due to finite observation time.

Treating s as a function from $[1, 20]$ into $\{0,1\}^3$ we have

$$s = \langle 010, 001, 100, 000, 100, \ldots, 000, 011\rangle$$

and s_A, s_B and s_C are just a binary vectors of length 20:

$$\begin{aligned} s_A &= 00101001000111010100 \\ s_B &= 10000000100010000101 \\ s_C &= 01000000001001000001. \quad \square \end{aligned}$$

2.2 Decomposition of event sequences

In order to discuss the independence properties we are interested in, we have to provide a way of probabilistic modeling of event sequences.

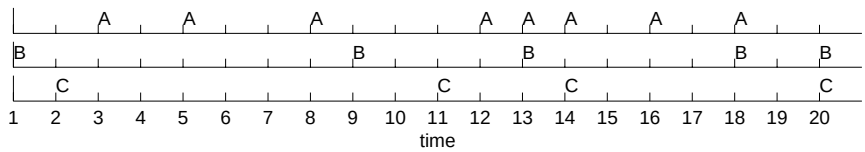


Figure 2. *Subsequences s_A,s_B and s_C of sequence s of events shown on Figure 1. Event C follows event B within two seconds lag.*

Given a set E of event types, the set of all event sequences over E can be viewed as the set $\mathcal{F}_E$ of all the functions $Z : [T_s, T_e] \to \{0,1\}^{|E|}$. That is, given a time t, the value $Z(t)$ indicates which events occur at that time.

A probabilistic model for event sequences is, in utmost generality, just a probability distribution μ_E on $\mathcal{F}_E$. For example, given some N, μ_E may depend only on the total number of the observed events and give a higher probability to the sequences that contain N events, e.g. $\mu_E(Z) = a \cdot e^{-(N-N_Z)^2/b^2}$ where N_Z denotes the total number of events in Z, $N_Z = \sum_{t=T_s}^{T_e} \|Z(t)\|_1$, and a, b are some appropriate constants. Note that in this example all event subsequences are dependent given N.

Next we define what it means that a distribution of event sequences is an independent composition of two distributions. We use the analogous concept from the distribution of discrete random variables: Let $\{X_1, \ldots, X_p\}$ be a discrete variables and denote by $P(X_1 = x_1, \ldots, X_p = x_p)$ the probability of observing the value combinations $(x_1, \ldots, x_p)$. Now P is an independent composition of distributions over variables $\{X_1, \ldots, X_j\}$ and $\{X_{j+1}, \ldots, X_p\}$ if for all combinations $(x_1, \ldots, x_p)$ we have

$$\begin{aligned} &P(X_1 = x_1, \ldots, X_p = x_p) = \\ &\qquad P_1(X_1 = x_1, \ldots, X_j = x_j) \cdot P_2(X_{j+1} = x_{j+1}, \ldots, X_p = x_p) \end{aligned} \tag{2}$$

where P_1 and P_2 are the marginal distributions defined by

$$\begin{aligned} &P_1(X_1 = x_1, \ldots, X_j = x_j) = \\ &\qquad \textstyle\sum_{(x_{j+1},\ldots,x_p)} P(X_1 = x_1, \ldots, X_j = x_j, X_{j+1} = x_{j+1}, \ldots, X_p = x_p) \\ &P_1(X_{j+1} = x_j, \ldots, X_p = x_p) = \\ &\qquad \textstyle\sum_{(x_1,\ldots,x_j)} P(X_1 = x_1, \ldots, X_j = x_j, X_{j+1} = x_{j+1}, \ldots, X_p = x_p). \end{aligned} \tag{3}$$

The above definition iseasily extended for the decomposition of $\{X_1, \ldots, X_p\}$ into more than two subsets.

Now, let E_1 be a subset of E. The distribution μ_E defines naturally the marginal distribution μ_{E_1} on $\mathcal{F}_{E_1}$

$$\mu_{E_1}(s_1) = \sum_{s \in \mathcal{F}_E, s[E_1] = s_1} \mu_E(s). \tag{4}$$

We can now provide a decomposition definition:

Definition 1 (Event set decomposition). *: The set of event types E* decomposes *into pairwise disjoint sets $E_1, \ldots, E_m$ with $E = \bigcup_{i=1}^{m} E_i$ and $\forall i \neq j, E_i \cap E_j = \emptyset$ if for all $s \in \mathcal{F}_E$:*

$$\mu_E(s) = \prod_{i=1}^{m} \mu_{E_i}(s[E_i]). \tag{5}$$

That is, the probability of observing a sequence s is the product of the marginal probabilities of observing the projected sequences $s[E_i]$.

If E decomposes into $E_1, E_2, \ldots, E_m$, we also say that μ_E decomposes into $\mu_{E_1}, \mu_{E_1}, \ldots, \mu_{E_m}$ and that E consists of independent components $E_1, E_2, \ldots, E_m$. As a special case, if E consists of two event types A and B, it decomposes into A and B provided

$$\mu_{\{A,B\}}(s) = \mu_A(s_A) \cdot \mu_B(s_B), \quad \forall s \in \mathcal{F}_{\{A,B\}}. \tag{6}$$

I.e., the occurrence probability of a sequence of A's and B's is the product of the probability of seeing the A's and probability of seeing the B's. Note that this definition is a standard definition of two independent processes ([5], page 296).

2.3 Finding independent components from observed sequences

Our goal is to start from observed sequence s over a set of event types E and to find sets $E_1, \ldots, E_m$ such that the probability distribution μ_E on $\mathcal{F}_E$ is decomposed into the marginal distributions $\mu_{E_1}, \ldots, \mu_{E_m}$. There are two obstacles to this approach: First, we only observe a single sequence, not μ_E. Second, the set of alternatives for $E_1, \ldots, E_m$ is exponential in size.

The first obstacle is considered in Section 3.1 where we show that certain quite natural conditions can be used to obtain information about μ_E from a single (long) sequence over E.

We next describe how to cope with the second obstacle. We overcome this problem by restricting our attention to pairwise interaction between event types. That is, given μ_E, two event types A and B are independent, if for all $s \in \mathcal{F}_E$ we have

$$\mu_{\{A,B\}}(s[\{A, B\}]) = \mu_A(s_A) \cdot \mu_B(s_B). \tag{7}$$

We show in the next section how we can effectively test this condition.

Given information about the pairwise dependencies between event types, we search for independent sets of event types. Let $G = (E, H)$ be a graph of E such that there is an edge between event types A and B if and only if A and B are dependent. Then our task is simply to find the connected components of G, which can be done in $O(|E|^2)$ by any standard algorithm (e.g., [1]).

Using the above procedure we separate E into the maximal number subsets $\tilde{E}_1, \ldots, \tilde{E}_l$, such that $\forall 1 \leq i \neq j \leq l, \forall e' \in \tilde{E}_i, \forall e'' \in \tilde{E}_j : e', e''$ are independent. Note, that pairwise independence generally does not imply the mutual independence ([5], page 184). In our case it means that $\tilde{E}_1, \ldots, \tilde{E}_l$ is not necessarily a decomposition of E. We use, however, $\tilde{E}_1, \ldots, \tilde{E}_l$ as a practical alternative to a

true decomposition of E. In the remainder of this paper we will concentrate on detecting pairwise dependencies among the events.

3 Detection of Pairwise Dependencies

The definition of decomposability given in the previous section is based on the use of the distribution μ_E on the set of all event sequences. This makes it impossible to study decomposability of a single sequence. If we have a large set of observed sequences, we can form an approximation of μ_E. Given a sufficiently long single sequence we can also obtain information about μ_E. In the following subsection we describe the conditions under which this is the case.

3.1 Basic assumptions

We expand our definitions a bit. Instead of considering event sequences over the finite interval $[T_s, T_e]$ of time, we (for a short while) consider infinitely long sequences. Such sequence $\tilde{s}$ is a function $\mathbb{Z} \to \{0,1\}^{|E|}$, and $\tilde{s}(t)$ gives the events that happened at time t.

We assume that the event sequence is generated by some underlying stochastic process $\{Z_t\}_{t\in\mathbb{Z}}$, where Z_t is a random variable that takes values from $\{0,1\}^{|E|}$. In this formulation $\mathcal{F}_E$ is a set of functions from $\mathbb{Z}$ into $\{0,1\}^{|E|}$, $\mathcal{F}_E = \{Z(t)|Z(t) : \mathbb{Z} \to \{0,1\}^{|E|}\}$, and μ_E is a probability measure on $\mathcal{F}_E$. Thus, the observed event sequence s is some specific realization $f(t) \in \mathcal{F}_E$ restricted to the interval $[T_s, T_e]$.

First two assumptions that we introduce will permit us to draw general conclusions from the single log, while the third assumption will allow us to restrict our attention to the local properties of the event generation process.

Assumption 1 (Stationary Process). *The observed process is a* stationary *process, i.e., it is shift-independent:*

$$\mu_E(S) = \mu_E(S_{+\tau}), \quad \forall \tau \in \mathbb{Z}, \forall S \subseteq \mathcal{F}_E \tag{8}$$

where $S_{+\tau} = \{f_{+\tau}(t)|\exists f \in S, \text{s.t. } \forall t \in \mathbb{Z} : f_{+\tau}(t) = f(t+\tau)\}$.

The assumption of stationary process means that process does not change over time. While this assumption by itself is somewhat unrealistic, in practice it can be easily justified by windowing, i.e., considering only a fixed sufficiently large time period. The question of stationary testing for a specific stochastic process is of great interest by itself, but it is beyond the scope of this paper.

Assumption 2 (Ergodic Process). *The observed process is an* ergodic *process, i.e., statistics that do not depend on the time are constant. That is, such statistics do not depend on the realization of the process.*

This is a very important assumption that means that any realization of the process is a representative of all possible runs. In particular it means that we can

average by time instead of averaging different runs of the process ([5], page 428). Let $X(f, u)$ denote the time average of the particular realization $f \in \mathcal{F}_E$ (event-log).

$$X(f, u) = \lim_{T \to \infty} (1/T) \int_{-T}^{T} f(u + t) dt. \tag{9}$$

This random variable is time invariant. If the process is ergodic, then X is the same for all f, i.e., $X(f, u) \equiv \bar{X}$, and for a stationary process we have

$$\bar{X} = E[X(f, u)] = \lim_{T \to \infty} (1/T) \int_{-T}^{T} E[f(u + t)] dt = \bar{f} \tag{10}$$

where $\bar{f} \equiv \bar{f}(t) = E[f(t)]$, so the expected value in every point, $\bar{f}$, is equal to the time average $\bar{X}$.

Note that not every stationary process is ergodic. For example, a process that is constant in time is stationary, but it is not ergodic, since different realizations may bring different constant values. A good introduction to the concept of ergodicity is given in [5], Section 13-1.

The assumption of ergodicity is very intuitive in many natural systems, e.g., in telecommunications alarms monitoring. In such systems, we feel that logs from different periods are independent and are a good representative of the overall behavior of the system. This observation is also the basis for the next assumption.

Assumption 3 (Quasi-Markovian Process). *The observed process is* quasi-Markovian *in the sense that local distributions are completely determined by the process values in some finite neighborhood, i.e.*

$$p(Z_t \in D | Z_{t'}, t' \neq t) = p(Z_t \in D | Z_{t'}, t' \neq t, |t - t'| \leq \mathcal{K}) \tag{11}$$

where $D \subseteq \{0, 1\}^{|E|}$ *and* $\mathcal{K}$ *is some predefined positive constant, which is called* maximal lag.

We call this assumption *Quasi-Markovian* in order to distinguish it from the classical definition of Markovian process where $\mathcal{K} = 1$. We specify that local probabilities depend not only on the past, but also on the future to account for cases with lagged alarms and alarms that originate from unobserved joint source but have variable delay times.

Note that Markovian property does not say that random variables that are too far apart (i.e., lagged by more than $\mathcal{K}$ second) are independent. It simply says that the information that governs the distribution of some particular random variable is contained in its neighborhood, i.e., in order for one variable to have an influence on another over the maximum lag period this variable should 'pass' the influence information in time steps smaller than $\mathcal{K}$ seconds.

3.2 First order dependencies

The straightforward way to detect pairwise dependencies among the events is by direct test of the pairwise independence condition. However, such approach is infeasible even for the simplest cases: Consider that two events are generated by

stationary, ergodic and quasi-Markovian process with $\mathcal{K} = 30$ seconds. In this case, we would like to approximate probabilities of the event distribution on some arbitrary 30 seconds interval (the start-time of the interval is unimportant since the process is stationary). This task will require approximation of probability of $2^{30} \cdot 2^{30} \approx 10^{12}$ joint event sequences. Supposing that the average of 100 observations of each sequence are needed to approximate its true frequency one should observe the event generation process for about 10^{14} seconds, which is approximately 31 million years.

The example given above demonstrates that there is no feasible way to detect all possible event dependencies for arbitrary event generation process. For many inter-event dependencies, however, there is no need to compute the full probabilities of event distribution functions on interval $\mathcal{K}$, since the dependencies among the events are much more straightforward and are detectable by simpler techniques. For example, one event may always follow another event after a few seconds (see example on Figures 1,2). Such dependency, called episode, is easily detectable [4].

This work deals with detection of event dependencies of first order. Such event dependencies can be described by specifying the expected density of events of one type in the neighborhood of events of second type. These *neighborhood densities* can usually be approximated with sufficient precision given the typical number of events (hundreds) in the data streams that we have encountered. Note also, that in the many applications the event streams are very sparse so it is reasonable to calculate densities in the neighborhood of events and not in the neighborhood of 'holes' (periods with no events occurring). Otherwise, the meaning of event and not-event may be switched.

3.3 Cross-correlation analysis

Consider two events e_1 and e_2. We observe a joint stochastic process that consists of two (possibly dependent) processes: one is generating events of type e_1 and second is generating events of type e_2. Consequently we have two streams of events s_1, s_2 of first and second event respectively. We can view s_1 and s_2 as a functions from the observed time period $[1; \mathcal{T}]$ (where $\mathcal{T}$ is the length of observation) into event frequencies, $\{0, 1\}$. An example of such process is given on Figure 3(a).

Supposing the quasi-Markovian property of the event generation process, the first order dependency should expose itself in the $2\mathcal{K}+1$ neighborhood of each event. We define the cross correlation with maximum lag $\mathcal{K}$ and with no normalization:

$$c_{12}(m) = \begin{cases} \sum_{n=1}^{\mathcal{T}-m} s_1(n)s_2(n+m) & m \geq 0 \\ c_{21}(-m) & m < 0 \end{cases}, \quad -\mathcal{K} \leq m \leq \mathcal{K}. \tag{12}$$

Note that the cross correlation vector c_{12} is the reverse of c_{21}. By dividing c_{12} by the observed frequencies of e_1 and e_2 we get the estimate of the neighborhood densities of e_2 in the neighborhood of e_1 and of e_1 in the neighborhood of e_2.

Ideally, if two events are unrelated and the length of observation ($\mathcal{T}$) is sufficiently large, the average density in the event neighborhood should be the same as everywhere on the observed period. It is the same as to require that lagged

cross-covariance is everywhere zero, i.e.

$$cov_{12}(m) = c_{12}(m)/(\mathcal{T} - m) - p_1 p_2 = 0, \quad \forall m \in [-\mathcal{K}, \mathcal{K}], \tag{13}$$

where p_1, p_2 are the a priori event probabilities, that does not depend on the time of observation since the process is supposed to be stationary. These probabilities can be estimated by averaging the observed frequencies of e_1 and e_2 over the length of the process (this is the direct usage of ergodicity assumption), i.e. let η_1, η_2 denote the observed number of events e_1 and e_2 respectively, thus

$$\begin{array}{l} p_1 \approx \eta_1/\mathcal{T} \\ p_2 \approx \eta_2/\mathcal{T} \end{array} \tag{14}$$

In practice, the neighborhood densities are deviating from the expected values even if events are independent; this is due to the random nature of the event generation process and due to finite number of observations. Thus, we should introduce some model that will account for these effects and give us a *threshold* values, that will allow detection of the event dependencies that are beyond random phenomenon.

3.4 Modeling the independent event generation processes

Consider two independent, stationary stochastic processes that are generating events of types e_1 and e_2. We assume that the each event generation process is not autocorrelated, i.e., in each process the probability of event(s) occurring at any given time is independent on the nearby events. Such assumption maybe justified in the case of sparse, quasi-Markovian processes where the average distance between the events of the same type is large comparing to the maximum lag distance.

We are interested in computing the probability of encountering $c_{12}(m) = k$ for some particular m over the observed stream of length $\mathcal{T}$. Since the event generation processes are assumed to be independent and stationary the above question is equivalent to calculating the probability of observing $c_{12}(0) = k$. We are also interested in computing the probability that $c_{12}(m)$ will not exceed some predefined values on the range $m \in [-\mathcal{K}, \mathcal{K}]$. We formulate the following problem:

Problem 1 (Pairs Distribution). *Assume we observe $2\mathcal{T}$ independent binary random variables $s_1(t)$, $s_2(t)$ for $t = 1, \ldots, \mathcal{T}$, with $P(s_1(t) = 1) = p_1$ and $P(s_2(t) = 1) = p_2$ for all $t = 1, \ldots, \mathcal{T}$. Let c_{12} be defined by Equation 12. The questions are:*

- *What is the distribution of $c_{12}(0)$?*
- *What is the exact from of $P(\forall m \in [A, B], a \leq c_{12}(m) \leq b)$ for some given $\{A, B, a, b\}$?*

We answer the first question exactly and give the approximation scheme for the second.

Under the assumptions of Problem 1 the generation of pair of events e_1 and e_2 is independent on its neighborhood and the probability of events e_1, e_2 occurring together is $p_{12} = p_1 p_2$, where p_1, p_2 are a priori event probabilities. Thus the

probability of observing exactly k pairs of events e_1, e_2 is described by binomial distribution:

$$P(c_{12}(0) = k) = \binom{\mathcal{T}}{k} p_{12}^k (1 - p_{12})^{\mathcal{T}-k} \tag{15}$$

To assess the probability of observing a random phenomenon we would like to estimate the probability of obtaining more or equally extreme number of observations as k, i.e.

$$\mathcal{R} = P(\text{random phenomenon}) = P(|c_{12}(0) - \mathcal{T} \cdot p_{12}| \geq |k - \mathcal{T} \cdot p_{12}|) \tag{16}$$

Direct calculation of $\mathcal{R}$ may be a hard computational task, which is unnecessary since we can use one of the known approximations to binomial distribution, namely to approximate binomial distribution by Normal or Poisson distributions. Since the typical values of p_{12} encountered in practice are very small (for the two most frequent events in telecommunications alarms data $p_{12} = 1.6 \cdot 10^{-6}$) the Poisson approximation is more appropriate:

$$P(c_{12}(0) = k) \approx \frac{\nu^k e^{-\nu}}{k!}, \quad \nu = \mathcal{T} \cdot p_{12}. \tag{17}$$

Thus the risk of accepting a random phenomenon with lag m as a true correlation is

$$\mathcal{R}(c_{12}(m) = k) \approx \sum_{i=k}^{\mathcal{T}} \frac{\nu^i e^{-\nu}}{i!} \approx 1 - \sum_{i=0}^{k-1} \frac{\nu^i e^{-\nu}}{i!} \tag{18}$$

The minimal extreme values are not considered by the above formula, since for the typical data we worked with the probability of observing zero pairs is quite large.[1] It is worth to mention, that we have only observed $\mathcal{T} - m$ trials of the events lagged by m seconds, but $\mathcal{T}$ is usually much bigger than m so the error of assuming $\mathcal{T}$ trials is insignificant.

We approximate the probability of $c_{12}(m)$ to achieve the certain values on the range of $[-\mathcal{K}, \mathcal{K}]$ by assuming that $c_{12}(m)$ and $c_{12}(m')$ are independent for $m \neq m'$. We have:

$$P(\forall m \in [-\mathcal{K}, \mathcal{K}], a \leq c_{12}(m) \leq b) \approx [P(a \leq c_{12}(0) \leq b)]^{2\mathcal{K}+1}. \tag{19}$$

Let δ denote the tolerable error probability of accepting two events as dependent, so threshold on the risk of each particular lagged correlation should be set as:

$$R_{th} = 1 - (1 - \delta)^{1/(2\mathcal{K}+1)}. \tag{20}$$

`EXAMPLE:` Consider two independent event streams shown on Figure 3(a). We would like to know what is the random probability of the observed event: 'The maximum number of 7 pairs was encountered while analyzing lagged pairs, for lags in $[-20, 20]$'. Under the model described above the probability of observing 7 or

[1] The usage of particular approximation (Poisson or Normal), as well as, the usage symmetric or asymmetric risk calculations are dictated by particular application.

more pairs for one particular lag is (assuming lag is much smaller than observation time):

$$P(\#pairs \geq 7|\text{lag} = -17) \approx \sum_{i=7}^{100} \frac{3^i e^{-3}}{i!} = 0.0335. \tag{21}$$

Thus, assuming the pair generation trials were independent, we have:

$$P(\#pairs \geq 7| -20 \leq \text{lag} \leq 20) = 1 - (1 - 0.0335)^{41} = 0.7528. \tag{22}$$

So the probability of observing 7 (or more) pairs in the analyzed lag interval $[-20, 20]$ is 75% for data in Figure 3(a), thus these streams can not be considered dependent. □

3.5 In-burst event independence analysis

In many event sequences events tend to appear in bursts. Burst is a sudden increase of activity of the event generating process. For example in the telecommunication alarms data that we have analyzed, the mean inter-event distance is 68 seconds, the median is only 20 and maximal time between subsequent events is 3600 seconds(!). This facts indicate that alarms in the telecommunication network data tend to appear in bursts with long intervals between them.

In burst-type event sequences, most of the events are dependent just because they are almost always grouped together. However, we may still want to perform in-burst analysis of event independence. Such analysis can be seen as deciding on the event independence given that events are grouped in bursts. Note that this model describes the situation when bursts are ignited by some external event and knowledge of these external events may rend many of the in-burst events independent.

To assess the risk of assuming the random phenomenon as true in-burst event dependence we would like to address the following problem, which is the generalization of Problem 1:

Problem 2 (Pairs Distribution, Non-Uniform density). *Let η_1, η_2 be a positive constants and let $\mathcal{D}(t) : \{1, \ldots, \mathcal{T}\} \rightarrow [0, \frac{1}{\max(\eta_1,\eta_2)}]$ be a function with integral one, i.e., $\sum_{t=1}^{\mathcal{T}} \mathcal{D}(t) = 1$. We observe $2\mathcal{T}$ independent binary random variables $s_1(t)$, $s_2(t)$ for $t = 1, \ldots, \mathcal{T}$, such that $P(s_1(t) = 1) = \eta_1 \mathcal{D}(t)$ and $P(s_2(t) = 1) = \eta_2 \mathcal{D}(t)$ for all $t = 1, \ldots, \mathcal{T}$. Let c_{12} be defined by Equation 12. The questions are:*

- *What is the distribution of $c_{12}(m)$?*
- *How to estimate $P(\forall m \in [A, B], a \leq c_{12}(m) \leq b)$ for some $\{A, B, a, b\}$?*

This problem is illustrated on Figure 3(b), where the two independent event streams are generated according to some a priori density $\mathcal{D}$. Formally speaking, we assume two-stage event-generation process. First a subset $\mathcal{G}$ of $\mathcal{F}_{e_1,e_2}$ is chosen

such that expected event density is equal to $\mathcal{D}$, i.e. $E_{f\in\mathcal{G}}[f_{e_i}] = \eta_i\mathcal{D}$ (for $i = 1, 2$), and only then specific $f \in \mathcal{G}$ is chosen. In this way the observed event streams are 'forced' to comply with some a priori event density $\mathcal{D}$. We would like to find if e_1, e_2 are independent given $\mathcal{D}$.

Simple in-burst event dependence analysis scheme

The problem with the above approach lies in the estimation of a priori event density, which is too biased to the current realization of the random process. One way to overcome this difficulty, and introduce a robust density approximation scheme, is to assume that $\mathcal{D}$ is of some fixed form, e.g., mixture of Gaussians. The simplest assumption is a 'binary' form of a priori distribution, i.e., the assumption that $\mathcal{D}$ specifies only 'yes' or 'no' information about bursts and in-burst event density is uniform.

An approach described in this section is based on the fact that events stream is very sparse, and there are usually long intervals between subsequent bursts. Many of the intervals are greater than the maximal lag time, and thus the event stream can be safely separated into a number of independent events subsequences that correspond to bursts and inter-burst intervals that are free from any events.

The standard 'uniform' analysis of Section 3.4 is performed on the burst periods only, i.e., on the series of events that are separated by no-event intervals of length $\mathcal{K}$ at least. Such analysis allows detecting first-order independent events given the bursts (but assuming nothing about burst magnitude). Technically, estimating the event probabilities p_1 and p_2 from bursts areas only gives larger estimates for p_1 and p_2 (Equation 14) thus rendering more of the found dependencies random comparative to the 'uniform' analysis of Section 3.4.

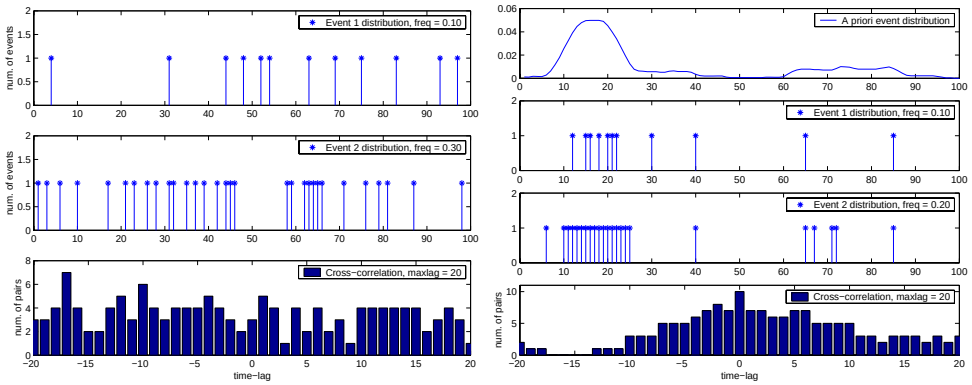
The algorithm for such simplified in-burst event independence analysis is outlined below (step one is the same for all event pairs):

1. Remove all no-event intervals that are greater that $\mathcal{K}$ seconds.
2. Calculate the lower limit on the number of observed lagged pairs for a given δ (Equation 18 and Equation 20) using the a priori event probabilities estimated from remaining data (Equation 14 with $\mathcal{T}_{new} = \mathcal{T} - time(\text{no_event_intervals})$).
3. Calculate the observed number of pairs for each lag in $[-\mathcal{K}, \mathcal{K}]$ and announce the events as dependent if the observed number of pairs for some lag exceeds lower limit calculated on Step 2.

An example of in-burst independence analysis

As example, consider two event streams on Figure 3(b). Performing the uniform random risk calculations (Section 3.4) without taking into account the distribution of events, we get the probability of 0.19% to observe correlation 10 or higher. On the over hand, removing the no-event areas, and working only with about 50 seconds of observed bursts, we get 18% probability that the observed phenomenon is random.

This analysis shows that two events are clearly dependent in general, without considering an information about prior distribution. The in-burst analysis, however, demonstrates that these events are independent given the burst information.



(a) (b)

Figure 3. *(a)Two independent streams of events and their cross correlation. The probability of observing* 7 *or more pairs in the analyzed lag interval is* 75%*. (b)Two independent streams of events following the same joint density and their cross correlation. The probability of observing* 10 *or more pairs in the analyzed lag interval is only* 0.19% *for a uniform model, but it is* 18% *considering the burst regions only.*

Note, that this result is achieved under the very simple model, without even taking into account the actual form of event density. A natural extension may be to threshold the estimated a-priori density function at some over label (and not at zero, like in the above approach). This method will allow gradual adjustment of the event dependencies, from the events independent regardless to bursts to the events that are dependent even given the information that they occur together in some dense bursts (applying threshold on the estimated event density function at some high level).

4 Experimental Results

The empirical analysis of the proposed dependency detection methods was performed on the telecommunications alarm log data and on two synthetic datasets that were especially created to test the dependency detection algorithm. The data was analyzed using two dependency detection methods, as summarized below:

- Event dependency analysis under uniform event generation model, as described in Section 3.4.
- Event dependency analysis using only the yes/no burst information, as described in Section 3.5.

All algorithms were applied with maximum lag $\mathcal{K} = 300$, and error probability threshold $\delta = 0.01$.

4.1 Telecommunication alarms data

The telecommunications alarm log consists of 46662 alarms in telecommunication network logged over the period of one month (2626146 seconds). The alarms are of 180 different types and 27 alarms are occurring with relative frequency of more than one percent. The original data contains a number multiple events of the same time occurring in the same second. We suppress these multiple entries to allow only one event of each type in any particular second. This operation leaves 38340 events in the log, which correspond to 82 percent of the original data. The mean inter-event distance is about 68 seconds, while the median is only 20 indicating that events tend to group together in bursts.

We restrict our analysis to the ten most frequent alarms that are responsible for more than 51 percent of the alarm data. These alarms, occurring more than a thousand times each, are frequent enough to allow various meaningful statistical approximations. On the other hand, such restriction enables to follow algorithm behavior in detail and not to be overwhelmed by large amount of data and inter event relationships.

We first analyze the telecommunications alarm data using the 'uniform' event generation model, i.e., without respect to the burst information. We demand random probability of the observed lagged correlation to be less than one percent and we are analyzing lags smaller than five minutes, i.e., $\delta = 0.01$, $\mathcal{K} = 300$ and $\mathcal{R}_{th} = 1 - (1 - 0.01)^{1/601} \approx 1.67 \cdot 10^{-5}$. Pairwise event dependencies that were detected in this analysis are shown on Figure 4a.

Removing the no-events intervals that are longer than $\mathcal{K} = 300$ seconds and applying the dependencies detection technique with the same parameters we get fewer dependencies, as illustrated on Figure 4b. The dependencies that are dropped are dependencies in pairs $(2, 6)$, $(6, 8)$ and, most noticeable, $(1, 2)$ and $(4, 10)$. Note that every in-burst dependency is also a dependency in the general sense.

Note that the set of inter-event dependencies consistently decreases as we restrict the definition of dependent events, Figures 4(a,b).

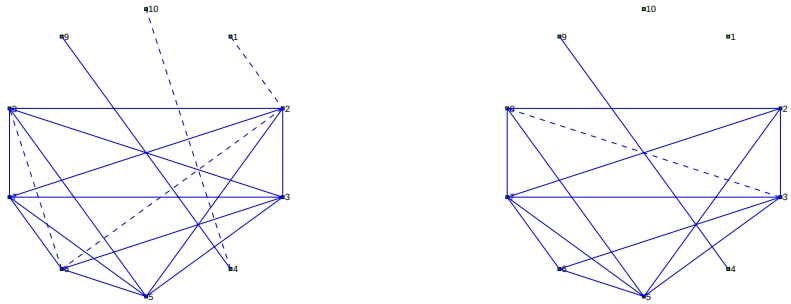
4.2 Experiments with synthetic data

We conduct additional experiments with two sets of synthetic data. The synthetic data streams contain events of five different types occurring over the period of 600000 seconds. The events frequencies are about 5000,3000,1000,1000 and 100 in the observed streams. Two rules were applied during the generation of event streams:

- Event #2 follows event #3 with 80% probability in time frame $[95, 105]$ seconds.
- Event #2 follows event #4 with 90% probability in time frame $[8, 12]$ seconds.

All other events were generated independently.

The first data stream was generated with events appear uniformly over the whole time period, while the second data stream was generated according to a priori distribution, which consisted of uniform distribution plus 200 randomly located



(a) (b)

Figure 4. *Pairwise dependencies for the first ten most frequent events of telecommunication alarms data. Error probability threshold is one percent ($\delta = 0.01$). (a) Without the use of burst information. (b) Using only the yes/no burst information. Dashed lines show weaker dependencies. Demanding greater confidence (smaller δ) renders these events independent.*

Table 1. *Experimental results with synthetic data. Found dependencies shown for each pairwise dependency detection method.*

Event dependence analysis	*Uniform density*	*Burst-like density*
Uniform event generation model	$(2,3)$, $(2,4)$	Every pair.
Yes/No burst information	$(2,3)$, $(2,4)$	Almost every pair.
Burst density approximation	$(2,3)$, $(2,4)$	$(2,3)$, $(2,4)$, $(1,2)$.
Episodes detection	$(3 \to 2)$, $(4 \to 2)$	$(3 \to 2)$, $(4 \to 2)$.

Gaussians with variances varying from 300 to 3600 seconds.

The same pairwise dependency analysis techniques were applied on these synthetic datasets to test the performance and stability of the proposed dependency analysis methods. To support the consistency of the results the techniques were applied with exactly the same parameters as for the telecommunication alarms data, namely $\delta = 0.01$, $\mathcal{K} = 300$. In addition two other pairwise dependency detection methods were applied on this dataset: dependency detection using the estimated event density (that is a more precise solution to Problem 1, which is not described here due to space limitations) and episode detection method similar to the one described in [4].

The experimental results are shown in Table 1. In the first dataset, the only dependencies in pairs $(2,3)$ and $(2,4)$ are correctly detected by all methods, and all other pairs were announced independent.

In the second dataset, ignoring the burst information renders all of events to be dependent, and it is expected, since all the events are inherently dependent because they follow the same a priori distribution. The simple in-burst independence

analysis, which takes only burst existence into account, also announces almost all (except $(5,1)$, $(5,2)$ and $(5,3)$) of the events to be dependent. Possible source to that behavior can be even higher (comparing to telecommunications data) in-burst event density, since the median inter-event distance is only 2 seconds, comparing with 20 for telecommunications data, while the mean inter-event distances are about the same (60 and 68 seconds respectively).

5 Conclusions

This paper has presented two methods for detection of pairwise event dependencies of first order for quasi-Markovian event sequences. The methods are distinguished by the amount of a priori information that is used to decide on event dependence. The methods vary from the basic one, that does not consider the burst information, up to the more advanced methods that consider burst information to various degrees.

To summarize, the proposed methods have the following properties:

- All independent events are reported as independent, with error probability δ.
- All first order pairwise event dependencies are found, with probability $1-\delta$.
- Some of the independent pairs reported may have dependency of higher order (undetectable by first order methods).

As a negative example of the dependencies that can not be found by analysis of first order moments imagine that first event has a priori density of 2 events per maximal lag ($\mathcal{K}$) and is distributed uniformly everywhere except the $\mathcal{K}$ seconds after occurrences of second event (which is, suppose, much sparsely distributed). Suppose also, that after each event of second type there are always two events of first type in time frame of $\mathcal{K}$ seconds and they are always separated by p or $\mathcal{K}-p$ seconds. While the distributions of these two events are clearly dependent this fact can not be detected by analyzing the neighborhood event densities of first event around second event and vice versa.

The presented methods can be extended to treat second-order dependencies, i.e., second-order moments of the distribution of one event in the neighborhood of another. One should be careful, however, to ensure that he has enough data to make a sufficient estimation of the measured quantities. This may be possible in the independent component analysis of the usual, non-sequential data, e.g., market basket data. In a market basket data all 'events' (purchases) in the dataset happen in the same time and we have multiple observations of the same variables (the set of all available goods). Removing the additional time variable may permit investigation of the higher order dependencies by approximating higher order moments with sufficient precision. It also may allow analysis of more complex, non-pairwise dependencies.

Bibliography

[1] T. H. Corman, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press, 1990.

[2] A. Hyvarinen and P. Pajunen, *Nonlinear independent component analysis: Existence and uniqueness results*, Neural Networks, 12 (1999), pp. 429–439.

[3] T. W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, *A unifying information-theoretic framework for independent component analysis*, International journal of computers and mathematics with applications, (In press, 1999).

[4] H. Mannila, H. Toivonen, and A. I. Verkamo, *Discovery of frequent episodes in event sequences*, Data Mining and Knowledge Discovery, 1 (1997), pp. 259–89.

[5] A. Papoulis, *Probability, Random Variables, and Stohastic Processes*, McGraw-Hill, 3rd ed., 1991.

[6] J. Storck and G. Deco, *Nonlinear independent component analysis and multivariate time series analysis*, Physica D, 108 (1997), pp. 335–349.

[7] D. Stoyan and A. Penttinen, *Recent applications of point process methods in forestry statistics*, Statistical Science, 15 (2000), pp. 61–78.