Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

System Identification with Analog and Counting Process Observations II: Mutual Information

Victor Solo

Abstract— **We develop new formulae for the mutual information between jointly observed analog signals and point processes allowing also that there may be an underlying unobserved state. Our derivation method delivers existing results as special cases while throwing new light on them.**

I. Introduction

Demand from a number of areas including communication networks and (what stimulated the current work) neuroscience [1],[2] has led to renewed interest in fundamental information theory calculations relating to systems observed through point processes. In a companion paper [3] we have constructed new likelihood ratio formulae for jointly observed point process and analog signals. Here we similarly produce new formulae for mutual information.

In the counting process literature there is little to report. There is the seminal paper of [4] which calculates the entropy of a point process in terms of its stochastic intensity. There is also the important work of [5] who gives a formula for the mutual information between a point process and an underlying unobserved analog state.

Following the approach in [3] we make no attempt at a rigorous development. That would require more space and will be pursued elsewhere. Rather we use the conditional Bernoulli heuristic where one discretises time to tiny subintervals and treats the point process as a conditonal Bernoulli process - taking limits at the end to get the continous time result. This yields very simple derivations suited to an applied audience but also throws new light on existing results. The heuristic is well known having been mentioned briefly in [6] in connexion with a likelihood derivation and also used as a computational tool [7]. But here we push the method into completely new territory.

The remainder of the paper is organized as follows. In the next section we derive McFadden's result and extend it in section III to obtain entropy and then mutual information between multivariate point processes. In section IV we rederive Bremaud's state space formula as well as an extension of it. Then in section V we introduce the hybrid stochastic intensity defined in [3] and use it to develop a new formula for the mutual information between jointly observed analog and point process signals. We extend this in section VI to allow also an unobserved analog state. Conclusions are offered in section VII.

Point Process Notation. In the sequel δ denotes a tiny time interval; t denotes a continuous time and k a discrete time so that $t = k\delta$. $N_{(t)}$=# events up to time t and in discrete time $N_k = N_{(k\delta)}$. Next,$\delta N_{(t)}$ = incremental count = # events in $(t, t+\delta]$. Also $\delta N_k = \delta N_{(k\delta)}$. Continuing, the history of the counting process up to time k, will be denoted $N_0^k = (\delta N_0 = \delta n_0, \cdots, \delta N_k = \delta n_k)$. with also $N^0 = (\delta N_0 = \delta n_0)$. While the associated accumulating sequence of random variables will be denoted, $N_{0,k} = (\delta N_0, \delta N_1, \cdots, \delta N_k)$. We write $a(\delta) = o(\delta)$ to mean $a(\delta)/\delta \to 0$ as $\delta \to 0$. If $[0, T]$ is an observation interval we write $T = n\delta$.

Entropy. Recall that for a digital random vector X with probability mass function $p(x) = P(X = x)$, the entropy is defined as [8] $\quad H(X) = -\Sigma_{allx} p(x) lnp(x)$.

We also need to introduce the stochastic conditional entropy(**SCE**) which is a random variable

$$H(X|Y = y) = -\Sigma_{allx} p(x|Y = y) lnp(x|Y = y)$$

Then the (average) conditional entropy (which is a constant) [8] is given by $H(X|Y) = \Sigma_{ally} P(Y = y) H(X|Y = y)$. Note that **SCE** is not defined in [8], a standard information theory reference. The mutual information between two random vectors is [8] defined as

$$\mathcal{I}(X;Y) = H(X) + H(Y) - H(X, Y)$$

In the sequel we will subscript $H, \mathcal{I}$ by δ where appropriate.

II. Univariate Point Process Entropy

We introduce the following point process assumptions.

NS No Simultaneity: $P(\delta N_k > 1|N_0^{k-1}) = o(\delta)$

This means that in a small time interval δ only 1 or 0 events occur. This property is called orderliness in the point process literature [9].

SI Stochastic Intensity.

$$\begin{aligned} P(\delta N_k = 1|N_0^{k-1}) &= \lambda_{(k\delta)}\delta + o(\delta) \\ &= \lambda_k \delta + o(\delta) \end{aligned}$$

Here $\lambda_{(t)}$ is called the stochastic (conditional) intensity and is a non-negative functional of the past history. A more formal definition of the stochastic intensity can be found in [6],[9]. In view of assumptions **NS** and **SI** we have:

CBD Conditional Bernoulli Description.

$$\begin{aligned} P(\delta N_k = 0|N_0^{k-1}) &= 1 - \lambda_{(k\delta)}\delta + o(\delta) \\ &= 1 - \lambda_k \delta + o(\delta) \end{aligned}$$

We will also need the marginal rate function defined through

$$P(\delta N_k = 1) = \beta_{(k\delta)}\delta + o(\delta) = \beta_k \delta + o(\delta)$$

V Solo is with Dept of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. vsolo@umich.edu

0-7803-9568-9/05/$20.00 ©2005 IEEE

But since $P(\delta N_k = 1) = E(\delta N_k) = E(E(\delta N_k|N_0^{k-1}))$ we find that $\beta_k = E(\lambda_k)$.
Now we wish to calculate the discrete time entropy of the point process increment sequence $N_{0,n}$ and then let $\delta \to 0$ to obtain an analog result. We have

$$\begin{aligned} H_\delta(N_{0,n}) &= H_\delta(\delta N_0, \delta N_1, \cdots, \delta N_n) \\ &= -\Sigma P(N_0^n) lnP(N_0^n) \end{aligned}$$

By the chain rule [8]

$$H_\delta(N_{0,n}) = \Sigma_0^n H_\delta(\delta N_k | N_{0,k-1})$$

To calculate a typical term in the sum we first calculate the **SCE** . The CBD (and $ln(1-\theta\delta) = -\theta\delta + o(\delta)$) gives

$$\begin{aligned} & H_\delta(\delta N_k|N_0^{k-1}) \\ = \; & -\Sigma P(\delta N_k = \delta n_k|N_0^{k-1}) lnP(\delta N_k = \delta n_k|N_0^{k-1}) \\ = \; & -P(\delta N_k = 1|N_0^{k-1}) lnP(\delta N_k = 1|N_0^{k-1}) \\ - \; & P(\delta N_k = 0|N_0^{k-1}) lnP(\delta N_k = 0|N_0^{k-1}) \\ = \; & -\lambda_k \delta ln(\lambda_k \delta) - (1-\lambda_k\delta) ln(1-\lambda_k\delta) + o(\delta) \\ = \; & -\lambda_k \delta ln \lambda_k + \lambda_k \delta - \lambda_k \delta ln\delta + o(\delta) \end{aligned} \tag{2.1}$$

Continuing, for a typical term in the chain rule sum

$$\begin{aligned} & H_\delta(\delta N_k|N_{0,k-1}) = E(H_\delta(\delta N_k|N_0^{k-1})) \\ = \; & -\delta E(\lambda_k ln\lambda_k) + E(\lambda_k)\delta - E(\lambda_k)\delta ln\delta + o(\delta) \end{aligned}$$

Summing now gives

$$\begin{aligned} H_\delta(N_{0,n}) &= h_\delta(N_{0,n}) - \Sigma_0^n \delta\beta_{(k\delta)} ln\delta + T\frac{o(\delta)}{\delta} \\ h_\delta(N_{0,n}) &= -\Sigma_0^n \delta E(\lambda_{(k\delta)} ln\lambda_{(k\delta)}) + \Sigma_0^n \beta_{(k\delta)}\delta \end{aligned}$$

As $n \to \infty, \delta \to 0, n\delta = T$ we find:
$\boxed{\text{Result I}}$: Univariate Entropy; $h_\delta(N_{0,n}) \to h(N_{(0,T)})$

$$h(N_{(0,T)}) = -\int_0^T E(\lambda_{(t)} ln\lambda_{(t)})dt + \int_0^T \beta_{(t)}dt$$

The second term is of order $\int_0^T \beta_{(t)} dt ln\delta$ and explodes. The third term $\to 0$.

We call $h(N_{(0,T)})$ the analog (differential) entropy of the point process. The time derivative of this expression agrees with the entropy derivative given by [4] in his (3.10) once we substitute his (3.9) into (3.10) and carry out a few lines of elementary algebra. As usual with analog entropy we will be able to avoid the explosion problem since we will be interested in mutual information which deals in entropy differences [8]. We have then obtained McFadden's classic result by an entirely new argument. We note in passing that the notion of a general stochastic intensity usually attributed to [10] and also [11],[12] (see [9]) also appears in [4] albeit unnamed.

III. Multivariate Point Process Entropy and Mutual Information

We start with the bivariate case which turns out to be generic. We have two counting processes $N_{(t)}, M_{(t)}$ with corresponding discrete time counts $N_k = N_{(k\delta)}, M_k = M_{(k\delta)}$ and so on as before. We also denote the joint history as $\mathcal{H}_0^k = (N_0^k, M_0^k)$ as well as $\mathcal{H}_{0,n} = (N_{0,n}, M_{0,n})$. We now introduce the following assumptions.
NS No-simultaneity: $P(\delta N_k + \delta M_k > 1|\mathcal{H}_0^k) = o(\delta)$
Given any past trajectory only 0 or 1 events (of either type) can occur in the next small time interval.
This of course implies marginal no-simultaneity.

SI Joint Stochastic Intensities: For$C = N, M$

$$\begin{aligned} P(\delta C_k = 1|\mathcal{H}_0^{k-1}) &= \lambda_{(k\delta)}^{CJ}\delta + o(\delta) \\ &= \lambda_k^{CJ}\delta + o(\delta) \end{aligned}$$

These two stochastic intensities depend on the joint history and so will differ from the marginal stochastic intensity previously introduced. As before assumptions **NS** ,**SI** yield:
CBD Conditional Multi-Bernoulli Description.
Firstly we have the semi-marginal relations; for $C = N, M$

$$P(\delta C_k = 0|\mathcal{H}_0^{k-1}) = 1 - \lambda_k^{CJ}\delta + o(\delta)$$

But we need to consider bivariate conditional probabilities. There are four such probabilities but joint no-simultaneity ensures $P(\delta N_k = 1, \delta M_k = 1|\mathcal{H}_0^{k-1}) = o(\delta)$ and then the other three are determined from the two marginal conditional probabilities and the requirement that probabilities sum to 1 [3]. As explained in [3] we have the following remarkable property:
$\boxed{\text{Result II}}$:**CI** Conditional Independence.
$\delta M_k, \delta N_k$ are conditionally independent given $\mathcal{H}_0^{k-1}$ i.e.

$$\begin{aligned} & P(\delta N_k = \delta n_k, \delta M_k = \delta m_k|\mathcal{H}_0^{k-1}) \\ = \; & (P(\delta N_k = \delta n_k|\mathcal{H}_0^{k-1}) + o(\delta)) \\ \times \; & (P(\delta M_k = \delta m_k|\mathcal{H}_0^{k-1}) + o(\delta)) \end{aligned}$$

The full proof is in [3] but here we illustrate one of the four cases. On the one hand

$$\begin{aligned} & P(\delta N_k = 1, \delta M_k = 0|\mathcal{H}_0^{k-1}) \\ = \; & P(\delta N_k = 1|\mathcal{H}_0^{k-1}) - P(\delta N_k = 1, dMk = 1|\mathcal{H}_0^{k-1}) \\ = \; & \lambda_k^{NJ}\delta + o(\delta) \end{aligned}$$

$$\begin{aligned} \text{while} \quad & P(\delta N_k = 1|\mathcal{H}_0^{k-1}) P(\delta M_k = 0|\mathcal{H}_0^{k-1}) \\ = \; & (\lambda_k^{NJ}\delta + o(\delta))(1 - \lambda_k^{MJ}\delta + o(\delta)) \\ = \; & \lambda_k^{NJ}\delta + o(\delta) \end{aligned}$$

and the result follows.

A. *Multivariate Entropy*

To calculate $H_\delta(\mathcal{H}_{0,n})$ we apply the chain rule to get

$$H_\delta(\mathcal{H}_{0,n}) = \Sigma_0^n H_\delta({}^{\delta M_k}_{\delta N_k}|\mathcal{H}_{0,k-1})$$

Again we first calculate the **SCE**

$$\begin{aligned} H_\delta({}^{\delta M_k}_{\delta N_k}|\mathcal{H}_0^{k-1}) &= -\Sigma P_k ln P_k \\ P_k &= P({}^{\delta M_k}_{\delta N_k} = {}^{\delta m_k}_{\delta n_k}|\mathcal{H}_0^{k-1}) \end{aligned}$$

Now applying the **CI** we find $P_k = P_k^M P_k^N$

$$P_k^N = P(\delta N_k = \delta n_k | \mathcal{H}_0^{k-1})$$
$$P_k^M = P(\delta M_k = \delta m_k | \mathcal{H}_0^{k-1})$$

Plugging this in gives

$$\begin{aligned} & H_\delta(\begin{smallmatrix}\delta M_k \\ \delta N_k\end{smallmatrix}|\mathcal{H}_0^{k-1}) = -\Sigma P_k^N P_k^M ln P_k^M P_k^N \\ = & -\Sigma_{\delta n_k}\Sigma_{\delta m_k} P_k^N P_k^M ln P_k^N \\ - & \Sigma_{\delta n_k}\Sigma_{\delta m_k} P_k^N P_k^M ln P_k^M \\ = & -\Sigma_{\delta n_k} P_k^N ln P_k^N - \Sigma_{\delta m_k} P_k^M ln P_k^M \\ = & H_\delta(\delta M_k|\mathcal{H}_0^{k-1}) + H_\delta(\delta N_k|\mathcal{H}_0^{k-1}) \end{aligned}$$

Now taking expectations we deduce:
$\boxed{\text{Result III}}$:**EA** Entropy Additivity

$$H_\delta(\mathcal{H}_{0,n}) = \Sigma_0^n H_\delta(\delta M_k|\mathcal{H}_{0,k-1}) + \Sigma_0^n H_\delta(\delta N_k|\mathcal{H}_{0,k-1})$$

Returning to the **SCE** we evaluate it in more detail. Indeed follwing the same steps as in (2.1) we get e.g.

$$H_\delta(\delta N_k|\mathcal{H}_0^{k-1}) = -\delta\lambda_k^{NJ} ln\lambda_k^{NJ} + \delta\lambda_k^{NJ} - \lambda_k^{NJ}\delta ln\delta + o(\delta)$$

Take expectations (note $\beta_k^N = E(\lambda_k^{NJ})$) to get

$$H_\delta(\delta N_k|\mathcal{H}_{0,k-1}) = \beta_k^N\delta - \delta E(\lambda_k^{NJ} ln\lambda_k^{NJ}) - \beta_k^N \delta ln\delta + o(\delta)$$

Summing and including the δM_k terms gives

$$\begin{aligned} H_\delta(\mathcal{H}_{0,n}) &= \\ h_\delta(\mathcal{H}_{0,n}) &- \Sigma_0^n(\beta_{(k\delta)}^M + \beta_{(k\delta)}^N + o(\delta))\delta ln\delta + T\frac{o(\delta)}{\delta} \\ h_\delta(\mathcal{H}_{0,n}) &= \delta\Sigma_0^n(\beta_{(k\delta)}^N - E(\lambda_{(k\delta)}^{NJ} ln\lambda_{(k\delta)}^{NJ})) \\ &+ \delta\Sigma_0^n(\beta_{(k\delta)}^M - E(\lambda_{(k\delta)}^{MJ} ln\lambda_{(k\delta)}^{MJ})) \end{aligned}$$

As $n \to \infty, \delta \to 0, n\delta = T$ we find the first term in $H_\delta(\mathcal{H}_{0,n})$ converges to an *additive* expression;
$\boxed{\text{Result IV}}$: Bivariate Entropy: $h_\delta(\mathcal{H}_{0,n}) \to h(\mathcal{H}_{(0,T)})$,

$$\begin{aligned} h(\mathcal{H}_{(0,T)}) &= \int_0^T (\beta_{(t)}^N - E(\lambda_{(t)}^{NJ} ln\lambda_{(t)}^{NJ}))dt \\ &+ \int_0^T (\beta_{(t)}^M - E(\lambda_{(t)}^{MJ} ln\lambda_{(t)}^{MJ}))dt \end{aligned}$$

The second term is of order $\int_0^T(\beta_{(t)}^N + \beta_{(t)}^M)dt ln\delta$ and explodes. The third term $\to 0$.

B. Mutual Information

The mutual information between the random histories $N_{0,n}, M_{0,n}$ will be, by definition [8],

$$\begin{aligned} & \mathcal{I}_\delta(N_{0,n}; M_{0,n}) \\ = & H_\delta(N_{0,n}) + H_\delta(M_{0,n}) - H_\delta(\mathcal{H}_{0,n}) \end{aligned}$$

Substituting the previous expressions yields

$$\begin{aligned} = & h_\delta(N_{0,n}) - \delta\Sigma_0^n\beta_k^N ln\delta \\ + & h_\delta(M_{0,n}) - \delta\Sigma_0^n\beta_k^M ln\delta \\ - & [h_\delta(\mathcal{H}_{0,n}) - \delta\Sigma_0^n(\beta_k^N + \beta_k^M) ln\delta] + o(\delta) \\ = & h_\delta(N_{0,n}) + h_\delta(M_{0,n}) - h_\delta(\mathcal{H}_{0,n}) + o(\delta) \end{aligned}$$

As expected the exploding terms have canceled out. Now letting $n \to \infty, \delta \to 0, n\delta = T$ gives

$$\begin{aligned} & \mathcal{I}_\delta(N_{0,n}; M_{0,n}) \to \mathcal{I}(N_{(0,T)}, M_{(0,T)}) \\ = & h(N_{(0,T)}) + h(M_{(0,T)}) - h(\mathcal{H}_{(0,T)}) \\ = & \Sigma_{C=N,M}\int_0^T (\beta_{(t)}^C - E(\lambda_{(t)}^C ln\lambda_{(t)}^C))dt \\ - & [\int_0^T (\beta_{(t)}^N + \beta_{(t)}^M)dt \\ - & \int_0^T E(\lambda_{(t)}^{NJ} ln\lambda_{(t)}^{NJ} + \lambda_{(t)}^{MJ} ln\lambda_{(t)}^{MJ}))dt] \end{aligned}$$

where $\lambda_{(t)}^N, \lambda_{(t)}^M$ are the marginal stochastic intensities. Collecting terms we get:
$\boxed{\text{Result V}}$: Bivariate Mutual Information

$$\begin{aligned} & \mathcal{I}(N_{(0,T)}; M_{(0,T)}) \\ = & \int_0^T [E(\lambda_{(t)}^{NJ} ln\lambda_{(t)}^{NJ}) - E(\lambda_{(t)}^N ln\lambda_{(t)}^N)]dt \\ + & \int_0^T [E(\lambda_{(t)}^{MJ} ln\lambda_{(t)}^{MJ}) - E(\lambda_{(t)}^M ln\lambda_{(t)}^M)]dt \end{aligned}$$

There are two important features;(i) the additive structure induced by no-simultanteity via conditional independence;(ii) as long as the joint stochastic intensities $\lambda_{(t)}^{NJ}$ and or $\lambda_{(t)}^{MJ}$ differ from the marginal stochastic intensities $\lambda_{(t)}^N$ and or $\lambda_{(t)}^M$ then the mutual information is $\neq 0$.

Given the additive structure induced by conditional independence itself induced by no-simulataneity, we see the extension to the full multivariate case is clear. For two groups of point processes represented by index sets A, B:
$\boxed{\text{Result VI}}$: Multivariate Mutual Information;

$$\begin{aligned} & \mathcal{I}(N_{(0,T)}^A; N_{(0,T)}^B) = \Sigma_{c\in A\cup B} E(\lambda_{(t)}^c ln\lambda_{(t)}^c)dt \\ - & \Sigma_{a\in A} E(\lambda_{(t)}^a ln\lambda_{(t)}^a)dt - \Sigma_{b\in B} E(\lambda_{(t)}^b ln\lambda_{(t)}^b)dt \end{aligned}$$

where $\lambda_{(t)}^c$ is the stochastic intensity of point process c given the past of all point processes in $A \cup B$; while $\lambda_{(t)}^a$ is the stochastic intensity of point process a given the past of all point processes in A and with a similar definition for $\lambda_{(t)}^b$.

IV. **State Space Models**

Here we suppose the stochastic intensity depends on an underlying unobserved state. We take the state to be an analog stochastic process $x_{(t)}$ for simplicity but the point process case can easily be treated. The sampled signal is $x_k = x_{(k\delta)}$. Since δN_k looks ahead we match the history $X_1^k = (X_1 = x_1, \cdots, X_k = x_k)$ with N_0^{k-1}. Also denote $X_{1,n} = (X_1, X_2, \cdots, X_n)$. *Additional Notation and Definitions*. We use the notation $X_k \sim x$ to mean $x \le X_k \le x + h$ with $0 < h << 1$. And then $\tilde{X}_1^k = (X_1 \sim x_1, \cdots, X_k \sim x_k)$. Finally by $P(A|X_1^k)$ we mean

$$\lim_{h\to 0} P(A|\tilde{X}_1^k) = \lim_{h\to 0}\frac{P(A, \tilde{X}_1^k)\frac{1}{h^{k+1}}}{P(\tilde{X}_1^k)\frac{1}{h^{k+1}}}$$

The assumptions now become:
NS No simultaneity: $P(\delta N_k > 1|N_0^{k-1}, X_1^k) = o(\delta)$
SI State dependent Stochastic Intensity.

$$\begin{aligned} P(\delta N_k = 1|N_0^{k-1}, X_1^k) &= P(\delta N_k = 1|X_k = x_k) \\ = \lambda_{(k\delta, x_{(k\delta)})}\delta + o(\delta) &= \lambda_{k,x_k}\delta + o(\delta) \end{aligned}$$

$\Rightarrow$ **CBD** Conditional Binomial Description.

$$\begin{aligned} P(\delta N_k = 0|N_0^{k-1}, X_1^k) &= P(\delta N_k = 0|X_k = x_k) \\ = 1 - \lambda_{(k\delta, x_{(k\delta)})}\delta + o(\delta) &= 1 - \lambda_{k,x_k}\delta + o(\delta) \end{aligned}$$

We now calculate the state/point process mutual information

$$\begin{aligned} &\mathcal{I}_\delta(X_{1,n}; N_{0,n-1}) \\ = \;& H_\delta(X_{1,n}) + H_\delta(N_{0,n-1}) - H_\delta(X_{1,n}, N_{0,n-1}) \end{aligned}$$

By the chain rule applied to each term we can write

$$\begin{aligned} & \mathcal{I}_\delta(X_{1,n}; N_{0,n-1}) = \Sigma_1^{n-1}\mathcal{I}_{\delta,k} \\ \mathcal{I}_{\delta,k} &= H_\delta(X_{k+1}|X_{1,k}) + H_\delta(\delta N_k|N_{0,k-1}) \\ &- H_\delta(\delta N_k, X_{k+1}|X_{1,k}, N_{0,k-1}) \end{aligned} \tag{4.1}$$

Now the Markov assumption on X_k gives $H_\delta(X_{k+1}|X_{1,k}) = H_\delta(X_{k+1}|X_k)$. And the chain rule+ state dependent stochastic intensity give

$$\begin{aligned} & H_\delta(\delta N_k, X_{k+1}|X_{1,k}, N_{0,k-1}) \\ = \;& H_\delta(\delta N_k|X_{1,k}, N_{0,k-1}) + H_\delta(X_{k+1}|X_{1,k}, N_{0,k}) \end{aligned} \tag{4.2}$$

$$\begin{aligned} = \;& H_\delta(\delta N_k|X_k) + H_\delta(X_{k+1}|X_k) \\ \Rightarrow \;& \mathcal{I}_{\delta,k} = H_\delta(X_{k+1}|X_k) + H_\delta(\delta N_k|N_{0,k-1}) \\ - \;& [H_\delta(X_{k+1}|X_k) + H_\delta(\delta N_k|X_k)] \\ = \;& H_\delta(\delta N_k|N_{0,k-1}) - H_\delta(\delta N_k|X_k) \end{aligned}$$

We calculate each term in turn via **SCE** . Firstly

$$\begin{aligned} & H_\delta(\delta N_k|X_k = x_k) \\ = \;& -P(\delta N_k = 1|X_k = x_k)lnP(\delta N_k = 1|X_k = x_k) \\ - \;& -P(\delta N_k = 0|X_k = x_k)lnP(\delta N_k = 0|X_k = x_k) \\ = \;& -\lambda_{k,x_k}\delta ln(\lambda_{k,x_k}\delta) - (1 - \lambda_{k,x_k}\delta)ln(1 - \lambda_{k,x_k}\delta) \\ = \;& -\lambda_{k,x_k}\delta ln\lambda_{k,x_k} + \lambda_{k,x_k}\delta - \lambda_{k,x_k}\delta ln\delta + o(\delta) \end{aligned}$$

Taking expectations gives

$$\begin{aligned} H_\delta(\delta N_k|X_k) &= -\delta E(\lambda_{k,x_k} ln\lambda_{k,x_k}) + E(\lambda_{k,x_k})\delta \\ &- E(\lambda_{k,x_k})\delta ln\delta + o(\delta) \\ E(\lambda_{k,x_k}) &= \int \lambda_{(k\delta,x)} p_{(k\delta,x)} dx = \beta_k \\ E(\lambda_{k,x_k} ln\lambda_{k,x_k}) &= \int \lambda_{(k\delta,x)} ln\lambda_{(k\delta,x)} p_{(k\delta,x)} dx \\ \Rightarrow H_\delta(\delta N_k|X_k) &= -\delta E(\lambda_{k,x_k} ln\lambda_{k,x_k}) + \beta_k\delta \\ &- \beta_k\delta ln\delta + o(\delta) \end{aligned}$$

with $p_{(t,x)}$ = marginal density function of $x_{(t)}$. Secondly

$$\begin{aligned} & P(\delta N_k = 1|N_0^{k-1}) \\ &= \int P(\delta N_k = 1|X_k, N_0^{k-1}) p(X_k|N_0^{k-1}) dX_k \\ &= \hat{\lambda}_k\delta + o(\delta) \\ \hat{\lambda}_k &= \int (\lambda_{(k\delta, x_{(k\delta)})} p(x_k|N_0^{k-1}) dx_k \\ &= E(\lambda_{k,x_k}|N_0^{k-1}) \end{aligned}$$

Similarly $P(\delta N_k = 0|N_0^{k-1}) = 1 - \hat{\lambda}_k\delta + o(\delta)$.
Thus we find for the **SCE**

$$\begin{aligned} H_\delta(\delta N_k|N_0^{k-1}) &= -\hat{\lambda}_k\delta ln(\hat{\lambda}_k\delta) \\ &- (1 - \hat{\lambda}_k\delta)ln(1 - \hat{\lambda}_k\delta) \\ = -\hat{\lambda}_k\delta ln\hat{\lambda}_k + \hat{\lambda}_k\delta &- \hat{\lambda}_k\delta ln\delta + o(\delta) \end{aligned}$$

as usual. Taking expectations and noting that $E(\hat{\lambda}_k) = E(E(\lambda_k|N_0^{k-1})) = E(\lambda_k) = \beta_k$ we get

$$\begin{aligned} & H_\delta(\delta N_k|N_{0,k-1}) \\ = \;& -\delta E(\hat{\lambda}_k ln\hat{\lambda}_k) + \delta\beta_k - \delta\beta_k\delta ln\delta + o(\delta) \end{aligned}$$

Putting these expressions together gives

$$\begin{aligned} \mathcal{I}_{\delta,k} &= H_\delta(\delta N_k|N_{0,k-1}) - H_\delta(\delta N_k|X_k) \\ &= -\delta E(\hat{\lambda}_k ln\hat{\lambda}_k) + \delta\beta_k - \delta\beta_k ln\delta \\ &- [-\delta E(\lambda_{k,x_k} ln\lambda_k) + \delta\beta_k - \delta\beta_k ln\delta] \\ &= \delta E(\lambda_{k,x_k} ln\lambda_{k,x_k}) - \delta E(\hat{\lambda}_k ln\hat{\lambda}_k) + o(\delta) \end{aligned}$$

Summing and letting $n \to \infty, \delta \to 0, n\delta = T$ gives:
$\boxed{\text{Result VII}}$: Mutual Information between observed point-process and unobserved analog state,

$$\begin{aligned} & \mathcal{I}_\delta(X_{1,n}; N_{0,n-1}) \to \mathcal{I}(X_{(0,T)}; N_{(0,T)}) \\ = \;& \int_0^T E(\lambda_{(t,x)} ln\lambda_{(t,x)})dt - \int_0^T E(\hat{\lambda}_{(t)} ln\hat{\lambda}_{(t)})dt \\ & \hat{\lambda}_{(t)} = E(x_{(t)}|N_{(0,t)}) \end{aligned}$$

This formula was originally obtained by [11] in a very different way. As before additivity makes the multivariate extension straightforward.

V. Hybrid Mutual Information

We begin for simplicity with the bivariate case of a jointly observed scalar analog signal $y_{(t)}$ and a point process $N_{(t)}$. We extend previous notation in the natural way to cover $y_{(t)}$. In particular we introduce the joint history $\mathcal{H}_{N,Y}^{k-1} = (N_0^{k-1}, Y_1^k)$. It is not immediately clear how to define a stochastic intensity to cover this case and the utility of our definition will become clear below. We assume:
NSNo simultaneity : $P(\delta N_k > 1|\mathcal{H}_{N,Y}^{k-1}, Y_{k+1} = y) = o(\delta)$.
HSI Hybrid Stochastic Intensity

$$\begin{aligned} & P(\delta N_k = 1|\mathcal{H}_{N,Y}^{k-1}, Y_{k+1} = y) \\ = \;& \lambda_{(k\delta,y)}\delta + o(\delta) = \lambda_{k,y}\delta + o(\delta) \end{aligned}$$

As usual **NS,HSI** deliver:
CBD Conditional Bernoulli Description.

$$P(\delta N_k = 0|\mathcal{H}_{N,Y}^{k-1}, Y_{k+1} = y) = 1 - \lambda_{k,y}\delta + o(\delta)$$

There are two associated quantities of importance.
Conditional Density

$$q(k\delta, y) = \lim_{h\to 0} \frac{1}{h} P(Y_{k+1} \sim y|\mathcal{H}_{N,Y}^{k-1})$$

Induced Stochastic Intensity

$$\begin{aligned} & \lambda_{(k\delta)}\delta + o(\delta) = P(\delta N_k = 1|\mathcal{H}_{N,Y}^{k-1}) \\ = & \int P(\delta N_k = 1|\mathcal{H}_{N,Y}^{k-1}, Y_{k+1} = y) q(k\delta, y)dy \\ & \Rightarrow \lambda_{(t)} = \int \lambda_{(t,y)} q_{(t,y)} dy \end{aligned}$$

Now we can develop the new hybrid mutual information. Applying the chain rule exactly as we did in the state space case, but not assuming any state space relation, we get firstly (4.1) (with X replaced by Y) and then substituting the chain rule (4.2) (with reversed chaining order) delivers

$$\begin{aligned} & \mathcal{I}_\delta(Y_{1,n}; N_{0,n-1}) = \Sigma_1^{n-1}\mathcal{I}_{\delta,k} \\ \mathcal{I}_{\delta,k} &= \mathcal{I}_{\delta,k}^d + \mathcal{I}_{\delta,k}^a \\ \mathcal{I}_{\delta,k}^a &= H_\delta(Y_{k+1}|Y_{1,k}) - H_\delta(Y_{k+1}|Y_{1,k}, N_{0,k-1}) \\ &= \text{analog mutual information} \\ \mathcal{I}_{\delta,k}^d &= H_\delta(\delta N_k|N_{0,k-1}) - H_\delta(\delta N_k|Y_{k+1}, Y_{1,k}, N_{0,k-1}) \\ &= \text{digital mutual information} \end{aligned}$$

Now we calculate **SCE** in each case and introduce: $p_{(k\delta,y)} = p(Y_k|Y_1^{k-1})$. We get

$$\begin{aligned} \mathcal{I}_{\delta,k}^a &= -E\int p_{(k\delta,y)} ln p_{(k\delta,y)} dy \\ &+ E\int q_{(k\delta,y)} ln q_{(k\delta,y)} dy \end{aligned}$$

For the digital component we find much as before, that the first **SCE** is $\quad H_\delta(\delta N_k|N_0^{k-1})$

$$\begin{aligned} &= -\lambda_k\delta ln(\lambda_k\delta) - (1-\lambda_k\delta) ln(1-\lambda_k\delta) + o(\delta) \\ &= -\lambda_k\delta ln\lambda_k + \lambda_k\delta - (\lambda_k\delta) + o(\delta)) ln\delta + o(\delta) \end{aligned}$$

While the second **SCE** is (dropping $o(\delta)$ terms)

$$\begin{aligned} & H_\delta(\delta N_k|Y_{k+1} = y_{k+1}, Y_1^k, N_0^{k-1}) \\ = & -\lambda_{k,y_k}\delta ln(\lambda_{k,y_k}\delta) - (1-\lambda_{k,y_k}\delta) ln(1-\lambda_{k,y_k}\delta) \\ = & -\lambda_{k,y_k}\delta ln\lambda_{k,y_k} + \lambda_{k,y_k}\delta - (\lambda_{k,y_k}\delta) ln\delta \end{aligned}$$

Taking expectations and noting that $E(\lambda_{k,y_k}) = E(\lambda_k) = \beta_k$ we find upon subtraction that

$$\mathcal{I}_{\delta,k}^d = \delta(E(\lambda_{k,y_k} ln\lambda_{k,y_k}) - E(\lambda_k ln\lambda_k))$$

Summing up gives $\mathcal{I}_\delta(Y_{1,n}; N_{0,n-1}) = \mathcal{I}_\delta^d + \mathcal{I}_\delta^a$
where, as $n \to \infty, \delta \to 0, n\delta = T$

$$\begin{aligned} & \mathcal{I}_\delta^d \to \mathcal{I}^d(Y_{(0,T)}; N_{(0,T)}) \\ = & \int_0^T E(\lambda_{(t,y_{(t)})} ln\lambda_{(t,y_{(t)})})dt - \int_0^T E(\lambda_{(t)} ln\lambda_{(t)})dt \\ & \mathcal{I}_\delta^a \to \mathcal{I}^a(Y_{(0,T)}; N_{(0,T)}) \\ = & \int_0^T E(q_{(t,y)} ln q_{(t,y)})dydt - \int_0^T E(p_{(t,y)} ln p_{(t,y)})dydt \end{aligned}$$

So we get: $\boxed{\text{Result VIII}}$: Mutual Information between observed point process and observed analog process;
$\mathcal{I}_\delta(Y_{1,n}; N_{0,n-1}) \to \mathcal{I}(Y_{(0,T)}; N_{(0,T)})$

$$\mathcal{I}(Y_{(0,T)}; N_{(0,T)}) = \mathcal{I}^d(Y_{(0,T)}; N_{(0,T)}) + \mathcal{I}^a(Y_{(0,T)}; N_{(0,T)})$$

VI. State Space Hybrid Mutual Information with Analog and Point Process Observations

We expand the set of definitions and assumptions.
NS No simultaneity

$$P(\delta N_k > 1|\mathcal{H}_{N,Y}^{k-1}, X_1^k, Y_k = y) = o(\delta)$$

SDHSI State Dependent Hybrid Stochastic Intensity

$$\begin{aligned} & P(\delta N_k = 1|\mathcal{H}_{N,Y}^{k-1}, X_1^k, Y_k = y) \\ = & P(\delta N_k = 1|\mathcal{H}_{N,Y}^{k-1}, X_k = x_k, Y_k = y) \\ = & \lambda_{(k\delta, x_{(k\delta)}, y)}\delta + o(\delta) = \lambda_{k,x_k,y}\delta + o(\delta) \end{aligned}$$

As usual **NS,SDHSI** deliver:
CBD Conditional Bernoulli Description.

$$\begin{aligned} & P(\delta N_k = 0|\mathcal{H}_{N,Y}^{k-1}, X_1^k, Y_k = y) \\ & P(\delta N_k = 0|\mathcal{H}_{N,Y}^{k-1}, X_k = x_k, Y_k = y) \\ = & 1 - \lambda_{k,x_k,y}\delta + o(\delta) \end{aligned}$$

There are two associated quantities of importance.
Conditional Density

$$q(k\delta, y) = \lim_{h\to 0} \frac{1}{h} P(Y_k \sim y|\mathcal{H}_{N,Y}^{k-1})$$

SDCD State dependent conditional density

$$\begin{aligned} & \lim_{h\to 0} P(Y_k \sim y|\mathcal{H}_{N,Y}^{k-1}, X_k = x_k) \\ = & P(Y_k \sim y|X_k = x_k) = p(y|x_k) = q_{k\delta, x_{k\delta}, y} \end{aligned}$$

Now the mutual information is $\mathcal{I}_\delta(X_{1,n}; \mathcal{H}_{1,n-1}^{N,Y})$ and application of the chain rule as in (4.1) and in section V gives

$$\begin{aligned} & \mathcal{I}_\delta(X_{1,n}; \mathcal{H}_{1,n-1}^{N,Y}) = \Sigma_0^{n-1}\mathcal{I}_{\delta,k} \\ \mathcal{I}_{\delta,k} &= H_\delta(X_{k+1}|X_{1,k}) + H_\delta(\delta N_k, Y_{k+1}|\mathcal{H}_{1,k-1}^{N,Y}) \\ &- H_\delta((\delta N_k, Y_{k+1}), X_{k+1}|X_{1,k}, \mathcal{H}_{1,k-1}^{N,Y}) \end{aligned}$$

Applying the chain rule to the third term gives

$$\begin{aligned} & H_\delta((\delta N_k, Y_{k+1}), X_{k+1}|X_{1,k}, \mathcal{H}_{1,k-1}^{N,Y}) \\ = & H_\delta(\delta N_k, Y_{k+1}|X_{1,k}, \mathcal{H}_{1,k-1}^{N,Y}) \\ + & H_\delta(X_{k+1}|X_{1,k}, \mathcal{H}_{1,k}^{N,Y}) \end{aligned}$$

And then applying the Markov property leaves

$$\begin{aligned}\mathcal{I}_{\delta,k} &= H_\delta(\delta N_k, Y_{k+1}|\mathcal{H}^{N,Y}_{1,k-1}) \\ &- H_\delta(\delta N_k, Y_{k+1}|X_{1,k}, \mathcal{H}^{N,Y}_{1,k-1})\end{aligned}$$

Now applying the chain rule to each of these terms gives $\mathcal{I}_{\delta,k} = \mathcal{I}^a_{\delta,k} + \mathcal{I}^d_{\delta,k}$

$$\begin{aligned}\mathcal{I}^a_{\delta,k} &= H_\delta(Y_{k+1}|\mathcal{H}^{N,Y}_{1,k-1}) - H_\delta(Y_{k+1}|X_{1,k}, \mathcal{H}^{N,Y}_{1,k-1}) \\ \mathcal{I}^d_{\delta,k} &= H_\delta(\delta N_k|Y_{k+1}, \mathcal{H}^{N,Y}_{1,k-1}) \\ &- H_\delta(\delta N_k|Y_{k+1}, X_{1,k}, \mathcal{H}^{N,Y}_{1,k-1}) \\ &= H_\delta(\delta N_k|Y_{k+1}, \mathcal{H}^{N,Y}_{1,k-1}) \\ &- H_\delta(\delta N_k|Y_{k+1}, X_k, \mathcal{H}^{N,Y}_{1,k-1})\end{aligned}$$

To continue we calculate the **SCE** in each case to find

$$\begin{aligned}\mathcal{I}^a_{\delta,k} &= -E\int q_{(k\delta,y)} ln q_{(k\delta,y)} dy \\ &+ E\int q_{(k\delta,x_{k\delta},y)} ln q_{(k\delta,x_{k\delta},y)} dy\end{aligned}$$

For $\mathcal{I}^d_{\delta,k}$ we proceed similarly to before. Firstly

$$\begin{aligned}& H_\delta(\delta N_k|Y_{k+1}=y, X_k = x_k, \mathcal{H}^{k-1}_{N,Y}) \\ =\ & -\Sigma_{\delta n_k=0,1} P(\delta N_k = \delta n_k|Y_{k+1}=y, X_k=x_k, \mathcal{H}^{k-1}_{N,Y}) \\ \times\ & lnP(\delta N_k = \delta n_k|Y_{k+1}=y, X_k = x_k, \mathcal{H}^{k-1}_{N,Y}) \\ =\ & -\delta\lambda_{k,x_k,y} ln(\delta\lambda_{k,x_k,y}) \\ +\ & (1-\lambda_{k,x_k,y}\delta) ln(1-\lambda_{k,x_k,y}\delta) \\ =\ & -\delta\lambda_{k,x_k,y} ln\lambda_{k,x_k,y} - \lambda_{k,x_k,y}\delta ln\delta - \lambda_{k,x_k,y}\delta\end{aligned}$$

So that, since $E(\lambda_{k,x_k,y}) = \beta_k$ we get

$$\begin{aligned}& H_\delta(Y_{k+1}, X_k, \mathcal{H}^{N,Y}_{1,k-1}) \\ =\ & -\delta E(\lambda_{k,x_k,y} ln\lambda_{k,x_k,y}) - \beta_k\delta ln\delta - \beta_k\delta\end{aligned}$$

Secondly $H_\delta(\delta N_k|Y_{k+1}=y, \mathcal{H}^{k-1}_{N,Y})$

$$\begin{aligned}=\ & -\Sigma_{\delta n_k=0,1} P(\delta N_k = \delta n_k|Y_{k+1}=y, \mathcal{H}^{k-1}_{N,Y}) \\ \times\ & lnP(\delta N_k = \delta n_k|Y_{k+1}=y, \mathcal{H}^{k-1}_{N,Y}) \\ & \text{But } P(\delta N_k = 1|Y_{k+1}=y, \mathcal{H}^{k-1}_{N,Y}) \\ =\ & \int P(\delta N_k = 1|X_k = x_k, Y_{k+1}=y, \mathcal{H}^{k-1}_{N,Y}) \\ \times\ & p(x_k|Y_{k+1}=y, \mathcal{H}^{k-1}_{N,Y}) dx_k \\ =\ & \int \delta\lambda_{k,x_k,y} \frac{p(Y_{k+1}|X_k=x_k, \mathcal{H}^{k-1}_{N,Y}) p(x_k|\mathcal{H}^{k-1}_{N,Y})}{p(Y_{k+1}|\mathcal{H}^{k-1}_{N,Y})} dx_k \\ =\ & \delta\int \lambda_{k,x_k,y} \frac{p(y_k|x_k)p(x_k|\mathcal{H}^{k-1}_{N,Y})}{q_{(k\delta,y_{(k\delta)})}} dx_k = \delta\hat{\lambda}_{k,y_k}\end{aligned}$$

Similarly, noting

$$\begin{aligned}& q_{(k\delta,y_{(k\delta)})} = \int p(y_k|x_k)p(x_k|\mathcal{H}^{k-1}_{N,Y}) dx_k \\ & \Rightarrow P(\delta N_k = 0|Y_{k+1}=y, \mathcal{H}^{k-1}_{N,Y}) = 1 - \hat{\lambda}_{k,y_k}\delta\end{aligned}$$

Putting these together we find (dropping $o(\delta)$ terms)

$$\begin{aligned}& H_\delta(\delta N_k|y_{k+1}, \mathcal{H}^{k-1}_{N,Y}) \\ =\ & -\delta E(\hat{\lambda}_{k,y_k} ln\hat{\lambda}_{k,y_k}) - \delta ln\delta\beta_k - \beta_k\delta\end{aligned}$$

Collecting terms together delivers

$$\mathcal{I}^d_{\delta,k} = -\delta E(\hat{\lambda}_{k,y_k} ln\hat{\lambda}_{k,y_k}) + \delta E(\lambda_{k,x_k,y} ln\lambda_{k,x_k,y})$$

Summing and taking the usual limits gives:

$\boxed{\text{Result IX}}$: Analog and Point Process Mutual Information with unobserved state.

$$\begin{aligned}& \mathcal{I}_\delta(X_{1,n}; \mathcal{H}^{N,Y}_{1,n-1}) = \mathcal{I}^d_\delta + \mathcal{I}^a_\delta \to \mathcal{I}^a + \mathcal{I}^d \\ \mathcal{I}^a &= \int_0^T E(\int q_{(t,x_{(t)},y)} ln q_{(t,x_{(t)},y)} dy) dt \\ &- \int_0^T E(\int q_{(t,y)} ln q_{(t,y)} dy) dt \\ \mathcal{I}^d &= \int_0^T E(\lambda_{(t,x_{(t)},y_{(t)})} ln\lambda_{(t,x_{(t)},y_{(t)})}) dt \\ &- \int_0^T E(\hat{\lambda}_{(t,y_{(t)})} ln\hat{\lambda}_{(t,y_{(t)})}) dt\end{aligned}$$

VII. **Conclusions**

In this paper we have used the conditional Bernoulli heuristic to provide elementary rederivations of known point process mutual information results (I,VII). We have also developed new results for mutual information between multivariate point processes, and between observed point process and observed analog process (V,VI). And involving observed analog and point processes (VIII) together with an unobserved state (IX).

References

[1] F. Rieke, D. Warland, R. de Ruyter van Stvenink, and W. Bialek, *Spikes: Exploring the neural code*, MIT Press, Boston, 1997.

[2] Dayan P and Abbott LF, *Theoretical Neuroscience*, MIT Press, Cambridge MA, 2001.

[3] V Solo, "System identification with analog and counting process observations I: Hybrid stochastic intensity and likelihood ratios.", Tech. Rep. submitted to IEEE Conf on Decision and Control 2005, Dept EECS, Univ. Michigan, Ann Arbor, 2005.

[4] J A McFadden, "The entropy of a point process", *J Soc Indust Appl Math*, vol. 13, pp. 988–994, 1965.

[5] P Bremaud, "On the information carried by a stochastic point process", *Revue du Cethedecdc*, vol. 43, pp. 45–70, 1975.

[6] Karr A, *Point Processes and their Statistical Inference, second edition*, Marcel Dekker, New York, 1991.

[7] M Berman, "Approximate point process likelihood with GLIM", *Applied Statistics*, vol. 41, pp. 31–38, 1992.

[8] T Cover and J Thomas, *Elements of Information Theory*, J Wiley, New York, 1991.

[9] D J Daley and D Vere-Jones, *An introduction to the Theory of Point Processes, Volume I (2nd. ed.)*, Springer-Verlag, New York, 2003.

[10] I Rubin, "Regular point processes and their detection", *IEEE Trans. Inf. Thy.*, vol. 18, pp. 547–557, 1972.

[11] P Bremaud, "A martingale approach to point processes", unpub. PhD Thesis. Univ California, Berkeley, 1972.

[12] D P Gaver, "Random hazard in reliability problems", *Tecnometrics*, vol. 5, pp. 211–216, 1963.