Applied Network Science

RESEARCH Open Access

Interaction networks from discrete event data by Poisson multivariate mutual information estimation and information flow with applications from gene expression data

Jeremie Fish[1,2*], Jie Sun[1,2] and Erik Bollt[1,2]

*Correspondence:
fishja@clarkson.edu

[1] Department of Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, New York 13699, USA
[2] Clarkson Center for Complex Systems Science, Clarkson University, 8 Clarkson Avenue, Potsdam, New York 13699, USA

Abstract

In this work, we introduce a new methodology for inferring the interaction structure of discrete valued time series which are Poisson distributed. While most related methods are premised on continuous state stochastic processes, in fact, discrete and counting event oriented stochastic process are natural and common, so called time-point processes. An important application that we focus on here is gene expression, where it is often assumed that the data is generated from a multivariate Poisson distribution. Nonparameteric methods such as the popular k-nearest neighbors are slow converging for discrete processes, and thus data hungry. Now, with the new multi-variate Poisson estimator developed here as the core computational engine, the causation entropy (CSE) principle, together with the associated greedy search algorithm optimal CSE (oCSE) allows us to efficiently infer the true network structure for this class of stochastic processes that were previously not practical. We illustrate the power of our method, first in benchmarking with synthetic datum, and then by inferring the genetic factors network from a breast cancer micro-ribonucleic acid sequence count data set. We show the Poisson oCSE gives the best performance among the tested methods and discovers previously known interactions on the breast cancer data set.

Keywords: Network inference, Poisson distribution, Conditional mutual information, Information theory

Introduction

Understanding the behavior of a complex system requires knowledge of its underlying structure. However prior knowledge of the network of interactions is often unavailable, necessitating estimation from data. Perhaps no complex system is more important to our health and well being than that of the gene expression network. However, these data are generally time-point process (TPP), and discretely distributed, rather than continuous valued as most mutual information inference methods presume. Specifically, we assume a jointly distributed Poisson process. While TPP are relatively common, as far as we know, no efficient joint entropy estimator exists which does not make the assumption

that the data is continuously distributed. Generally for the purposes of entropy, the assumption is that we may simply use mutual information of the normal distribution to approximate that of the Poisson distribution. As we will show, this can lead to poor results, particularly when using these estimates for the purposes of network inference, leaving a gap in our ability to estimate the structure when the data is Poisson distributed. To this end, the main goal of this paper is to fill that gap.

Understanding which variables have a causal relationship with one another, the causal network, is an essential aspect of the ability to drive a system toward a desired outcome. In this work we focus on Granger causality in order to derive the causal structure of a given system. Granger causality (Granger 1969) has been used for network inference when interpreted as a causation inference concept. For linear stochastic processes, an example of a method which fits within the Granger causality framework is transfer entropy (TE) (Schreiber 2000) which is based on information theory for nonlinear processes. However when applied to a system with more than two factors, transfer entropy is unable to distinguish direct versus indirect effects or confounders, and therefore they will tend to yield false positive connections. These false positives may lead to interventions which do not achieve the desired result, for instance in the context of the gene expression network, if a gene is inferred to have far more outgoing edges than it truly has, one might conclude that its removal will lead to a desired outcome when in fact it will not.

To this end, we developed causation entropy (CSE) as a generalization of transfer entropy (Sun 2014; Sun et al. 2015), that explicitly defines the information flow between two factors, conditioned on tertiary intermediary factors. This, together with a greedy search algorithm to construct the network of interactions of the complex stochastic process, provably reveals network structure of certain stochastic processes, (Sun et al. 2015). Additionally numerical evidence suggests that optimal causation entropy (oCSE) produces very few spurious connections, even while finding the vast majority of true connections. In past studies, TE as well as CSE were computed nonparametrically, by the Kraskov–Stögbauer–Grassberger (KSG) (Kraskov and Stögbauer 2004) mutual information estimator which is a K-nearest neighbors (KNN) method. However, if specific knowledge of the joint distribution of the process allows considerable computation efficiencies particularly faster convergence, such as our previous work where jointly Gaussian variables (Sun et al. 2015) or jointly Laplace distributed variables in Ambegedara et al. (2016) were relevant then it becomes preferable to use these distributions.

Here, we focus on gene expression networks, which are an application of considerable scientific importance due to their foundational relevance as a building block tool to understanding details of life science. It is well understood that many diseases associate with variations of the expression of a single gene (Rogers 2008; Sebastiani et al. 2005; De Boulle 1993), e.g., famously such as sickle cell disease and cystic fibrosis. However it remains a difficult problem with considerable health implications to explain and to infer complex interactions and associations when many genes may be involved in common and even deadly disease. Such diseases are called polygenetic, and these include the breast cancer example that we study here. According to the Centers for Disease Control (CDC), in the USA, breast cancer is considered to be the second most common form of cancer amongst women, (https://www.cdc.gov/cancer/breast/statistics/index.htm), that

in 2019 was forecast to 268,600 cases and 42,260 deaths in 2019. In previous work, gene expression data has been assumed to be drawn from a multivariate Poisson distribution (Allen 2013; Gallopin et al. 2013). This highlights an application for the usefullness of the development of a Poisson version of oCSE.

We advance here a new methodology to probe variations in expression of a group (network) of genes that may lead to disease. Understanding the gene interaction network structure may be crucial to the development of future treatments. Network inference itself has many applications beyond cancer research, including functional magnetic resonance imaging (fMRI) network inference (Smith 2012; Bassett 2011; Stoltz and Harrington 2017; Fish et al. 2021), drug-target interaction networks (Yamanishi et al. 2008), and earthquake network inference (Zhang et al. 2016) and economy issues (Iori 2008) to name a few.

With this motivation, the main technical premise of this paper is to develop a computationally efficient approach to estimate joint entropy and related information theoretic measures for multivariate Poisson processes, which are necessary for utilizing oCSE for network estimation. Data derived from these are discrete-valued data, and typically consist of a significant fraction of zeros punctuated with nonzero values describing event counts in a given epoch. From the multi-variate Poisson model, we derive an analytical series representation of the joint entropy and the mutual information. Then, a practical finite partial-sum estimator allows estimation of mutual information, toward transfer entropy and causation entropy (Fig. 1).

This paper is structured as follows: first, we provide a brief introduction to mathematical background including a multivariate Poisson model and also relevant information theoretic quantities which are necessary to define information flow. Then, we derive our multivariate Poisson joint entropy estimator, which we relate to network inference. Finally, in the Results section we demonstrate our method and performance for benchmark synthetic data and then we study the breast cancer gene expression data sets.

Background

The goal of this work is to expand the optimal causation entropy (oCSE) (Sun et al. 2015) to handle Poisson distributed data. Below we highlight the necessary background of a multivariate Poisson distribution, Granger causal inference, and finally causation entropy for the development of a new Poisson oCSE algorithm.

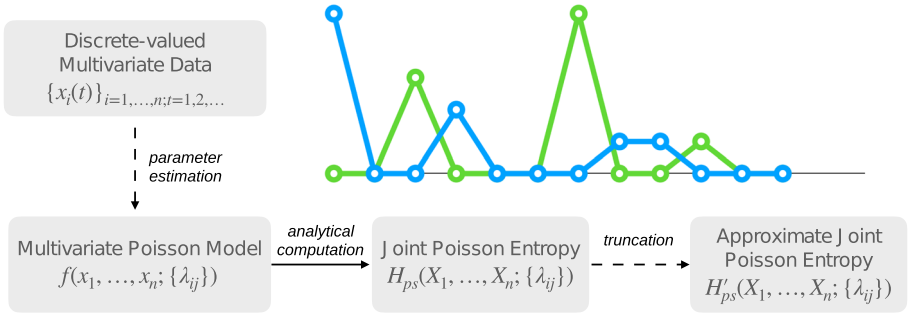

Fig. 1 Work flow of our computationally efficient approach to estimate the joint entropy of multi-variate Poisson distributed variables. From data, we proceed to distribution parameter estimation to approximate joint entropy

Multivariate Poisson model

First let us recall the single variate Poisson Model, (Reiss 2012; Bollt and Santitissadeekorn 2013):

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{1}$$

The Poisson model has a multivariate generalization as follows, (Karlis and Meligotsidou 2007):

$$P(X_1 = x_1, ..., X_n = x_n) = e^{-\sum_{i=1}^{n}\sum_{j\geq i}\lambda_{ij}} \sum_{C} \frac{\prod_{i=1}^{n} \lambda_{ii}^{(x_i - \sum_j a_{ij})} \prod_{i=1}^{n}\prod_{j>i} \lambda_{ij}^{a_{ij}}}{\prod_{i=1}^{n}(x_i - \sum_j a_{ij})! \prod_{i=1}^{n}\prod_{j>i} a_{ij}!}, \tag{2}$$

where the set

$$C = \{A = [a_{ij}]_{n\times n} : a_{ij} \in \mathbb{N}_0, a_{ii} = 0, a_{ij} = a_{ji} \geq 0, \sum_j a_{ij} \leq x_i\}, \tag{3}$$

and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. This model is based on assuming that the x_i are linearly transformed from a set of independently drawn Poisson variables. We begin with

$$\begin{aligned} X \in \mathbb{N}_0^{n\times t} &= (x_1, x_2, ...x_n)^T = BY, \\ Y \in \mathbb{N}_0^{m\times t} &= (y_{11}, y_{22}, ..., y_{nn}, y_{12}, y_{13}, ..., y_{(n-1)n})^T. \end{aligned} \tag{4}$$

Here each y_{ij} is independent Poisson, that is: $y_{ij} \in \mathbb{N}_0^t \sim$ Poisson (λ_{ij}), (for $i = 1, ..., n, j \geq i$), so $m = n + \frac{n(n-1)}{2}$, $B \in \mathbb{N}_0^{n\times m}$. Note that in this case $\lambda_{ij} = \lambda_{ji}$. The rows of X thus represent Poisson random variables which have t observations. Although the number of parameters needed to specify this model grows quickly, there are some nice properties. For instance, this model allows a simple estimate of each λ_{ij}, since the sum of independent Poisson variables yields the following covariance matrix structure:

$$\text{Cov}\,(X) = \begin{bmatrix} \lambda_{11} + \sum_{j\neq 1}^{n} \lambda_{1j} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{12} & \lambda_{22} + \sum_{j\neq 2}^{n} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{nn} + \sum_{j=1}^{n-1} \lambda_{nj} \end{bmatrix} \tag{5}$$

with $\lambda_{ij} = \lambda_{ji}$. The (i, j) entries of the covariance matrix represent cov (x_i, x_j), the covariance between the two random variables, x_i and x_j). Proof of Eq. 5 may be found in the appendix.

This model is a multivariate extension of the Poisson model that does not assume the random variables are necessarily independent. However, there are some limitations to this model. First, the rapid growth in the number of states and parameters with respect to the number of variables, making calculation of the joint distribution computationally unwieldy and expensive. Another limitation is that model cannot handle negative covariance (Karlis and Meligotsidou 2007). Some of these difficulties will be handled later on, however we do not address negative covariances, as a fix to this generally requires models which are not Poisson.

Transfer entropy and causation entropy

We briefly review certain Shannon entropies, building toward the concepts of transfer entropy and causation entropy. These are the fundamental concepts of information flow we use to consider network inference. The Shannon *entropy* of a (discrete) random variable X is given by Shannon (1948), Cover and Thomas (2012):

$$H(X) = -\sum_{x \in X} P(x) \log\left(P(x)\right), \tag{6}$$

where $P(x)$ is the probability that $X = x$, and $0 \log(0) = 0$ is the usual interpretation in this context. For the remainder of this paper we choose the natural log and thus all entropies will be measured in nats. Entropy can be thought of as a measure of how uncertain we are about a particular outcome. As an example we can imagine two scenarios, in one case we have a random variable $X_1 = (x_1^{(1)}, x_1^{(2)}, ..., x_1^{(n)})$ with $x_1^{(t)} = 0 (\forall t)$, that is $P(X_1 = 0) = 1$, in the other case the random variable $X_2 = (x_2^{(1)}, x_2^{(2)}, ..., x_2^{(n)})$ with $P(X_2 = 0) = 0.5$, $P(X_2 = 1) = 0.5$. Here $H(X_1) = 0$ nats, while $H(X_2) = \ln(0.5)$ nats which happens to be the maximum for this case (Cover and Thomas 2012). It is easy to see that Shannon entropy reaches its greatest value when we are the most uncertain about the outcome, and its minimal value (0) when we are completely certain about the outcome. We can now examine the case of two random variables X and Y. The *joint entropy* of a discrete random variable is defined (Cover and Thomas 2012):

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \ln\left(P(x, y)\right). \tag{7}$$

When the two random variables X and Y are independent $H(X, Y) = H(X) + H(Y)$ which is the maximum joint entropy. Thus $H(X, Y) \leq H(X) + H(Y)$, taken all together this means the joint entropy is largest when the variables are independent, and decreases as they become more and more dependent. This is an important feature, which we take advantage of for network inference.

There are comparable definitions of differential entropies for continuous random variables in terms of integration. The *conditional entropy* is defined:

$$H(X \mid Y) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \ln P(x \mid y). \tag{8}$$

The conditional entropy gives us another way to describe the relationship between variables, which is the key to network inference. If knowledge of the variable Y gives us

complete knowledge of the variable X then the conditional entropy will be $H(X \mid Y) = 0$ nats. Note that there is a relationship between the conditional entropy and the joint entropy namely:

$$H(Y \mid X) = H(X, Y) - H(X). \tag{9}$$

This is convenient in situations where it may be easier to compute one of the two entropies. Another important Shannon entropy is the *mutual information* which is defined as Cover and Thomas (2012):

$$\begin{aligned} I(X, Y) &= \sum_{x \in X} \sum_{y \in Y} P(x, y) \ln \left(\frac{P(x, y)}{P(x)P(y)} \right) \\ &= H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X, Y). \end{aligned} \tag{10}$$

The mutual information is exactly 0 when the variables are independent, and grows along with the mutual dependence. A feature which separates mutual information from correlation, is that the mutual information generally works well as a measure of dependence even when the relationships are not linear, which is not necessarily the case for correlation, for the interested reader an example of this is shown in Smith (2015). Finally, the *Kullback-Leibler (KL) divergence* (D_{KL}) (Cover and Thomas 2012) is stated:

$$D_{KL}(P \mid\mid Q) = -\sum_{x \in X} P(x) \ln \left(\frac{Q(x)}{P(x)} \right). \tag{11}$$

The KL divergence describes a distance-like quantity between two probability distributions, though it is not a metric as for one, it is not symmetric (that is in general $D_{KL}(P \mid\mid Q) \neq D_{KL}(Q \mid\mid P)$), and also it does not satisfy the triangle inequality. Mutual information Eq. 10 can be written in terms of KL divergence as Cover and Thomas (2012):

$$I(X, Y) = D_{KL}(P(x, y) \mid\mid P(x)P(y)), \tag{12}$$

describing a deviation from independence of a joint random variable (x, y). In other words, the mutual information can be recast as a distance like measure between two variables in the distributional sense. This is a key component for the use of mutual information, and specifically conditional mutual information, for inferring direct causal connections between variables.

For a stationary stochastic process, $\{X^t\}$, the *entropy rate* is defined as Cover and Thomas (2006), Bollt and Santitissadeekorn (2013):

$$H(\chi) = \lim_{t \to \infty} H(X^t \mid X^{t-1}, X^{t-2}, ..., X^1). \tag{13}$$

If the process is Markov (memoryless) then (Cover and Thomas 2012):

$$H(\chi) = \lim_{t \to \infty} H(X^t \mid X^{t-1}). \tag{14}$$

For this work we will assume that the networked system to be analyzed is a set of stationary stochastic processes, and therefore we can learn the Granger causal structure from observational data obtained from the system. The *transfer entropy* from X_2 to X_1 is defined as Schreiber (2000), Sun (2014):

$$T_{X_2 \to X_1} = H(X_1^{t+1} \mid X_1^t) - H(X_1^{t+1} \mid X_1^t, X_2^t). \tag{15}$$

The transfer entropy can be thought of as measuring the total amount of information flowing from one variable to another. Note however that some of that information may be flowing to one variable through another variable (for excellent figures illustrating this please see Sun 2014; Sun et al. 2015) and thus transfer entropy may infer a link between two variables that in reality do not have a direct connection to one another. This issue can be solved by conditioning as will be discussed below.

Causation entropy is a generalization of the transfer entropy, where (Sun 2014; Sun et al. 2015):

$$C_{\mathcal{Q} \to \mathcal{P} \mid \mathcal{S}} = H(\mathcal{P}^{t+1} \mid \mathcal{S}^t) - H(\mathcal{P}^{t+1} \mid \mathcal{S}^t, \mathcal{Q}^t). \tag{16}$$

$C_{\mathcal{Q} \to \mathcal{P} \mid \mathcal{S}}$ is designed to describe the remaining information flow from processes $\mathcal{Q}$ to processes $\mathcal{P}$ that may not be accounted for (conditioned on) processes $\mathcal{S}$. An example of causation entropy is shown in Fig. 2a. In theory if a process Z has no influence over another process Y, the causation entropy after conditioning out the remaining processes would be identically 0, allowing us to reject a connection from Z to Y. In practice however, when estimating these quantities by statistics from finite samples of noisy data, these will not compute to be identically 0, making it necessary to have a threshold, which is the purpose of using a shuffle test as discussed in Sun et al. (2015).

Network inference can be developed based on Eq. 16. However, considering the power-set of all possible subsets $\mathcal{P}, \mathcal{Q}, \mathcal{S}$ is clearly NP-hard and so not practical. This

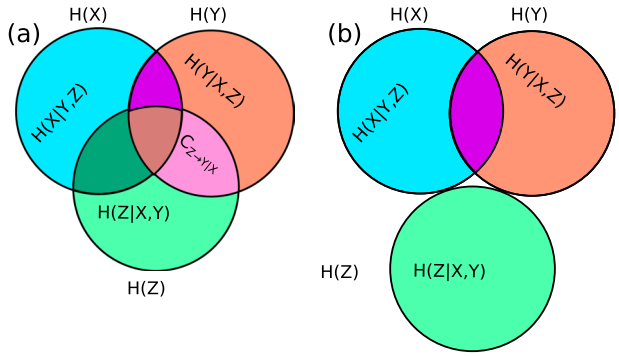

Fig. 2 **a** The causation entropy between two processes Z and Y is shown. In this case since we are only conditioning on a process X, $C_{Z \to Y \mid X} = T_{Z \to Y}$. Of course X may be replaced with a set of variables. **b** Here we show a special case where Z is independent of both X and Y (Z in this case may represent the history of X. In this case it becomes clear that $H(Z \mid X, Y) = H(Z)$, $H(X \mid Y, Z) = H(X \mid Y)$ and $H(Y \mid X, Z) = H(Y \mid X)$. As explained in the text, this special case helps us to discern what are the proper variables to use in the Poisson case

led to the development of a greedy search algorithm, we referred to as optimal causation entropy (oCSE) (Sun et al. 2015; Ambegedara et al. 2016) to a minimal network that explains the data, in terms of minimal causation entropy. This proceeds in two stages, aggregative discovery of statistically significant links, those that are maximally informative influencers in terms of the conditionally already significant links, with possible removal of statistically irrelevant links developed while growing the global network, and significance decided by a null hypothesis in terms of multiple random shuffles of the data. We were able to prove under mild hypothesis of the stochastic process that this procedure will discover the true network, also assuming a good statistical estimation of the entropies. It is precisely this problem of good data-driven statistical estimation of entropies specialized to the scenario of a multivariate Poisson process which is what we handle in this paper.

Entropy estimation from multivariate Poisson data

oCSE requires the calculation of conditional mutual information (CMI) (Sun et al. 2015), and there are numerous paths to obtain it. We reiterate that oCSE is the algorithm used for network inference, however the previously available versions of oCSE have proven to be insufficient in accurately reconstructing the network on synthetic Poisson data as will be shown in the results section.

In some cases, such as in the Gaussian case, the mutual information, and therefore the conditional mutual information, is easily estimated directly from data. Another possible path, which is the path we choose below, is to estimate the CMI using the joint entropy.

An estimator of the joint entropy of the Poisson distribution was derived in Guerrero-Cusumano (1995). However they derive their estimate by assuming the mutual information is the same as the Gaussian distribution. This would make the mutual information estimator the same as the one presented in Sun et al. (2015). However, particularly for smaller values of the rate parameter λ of the Poisson distribution, this estimator can be quite inaccurate, and as we will see in the results section, this can lead to highly inaccurate estimates for network structure.

In order to obtain an estimator which produces more accurate estimates of network structure, we take a different approach to estimating the mutual information of the Poisson distribution. Below an approximation of the joint entropy of the multivariate Poisson distribution is derived, and from this joint entropy the CMI needed for oCSE will be computed.

Estimating joint entropy of Poisson systems

Here we develop an estimator of entropies for the multivariate Poisson distribution, Eq. 2. To this end, we truncate partial sums from series representations.

Poisson entropy

We begin the Poisson Entropy:

$$H_{Poisson}(K) = -\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \ln\left(\frac{\lambda^k}{k!} e^{-\lambda}\right) =$$
$$-\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} [-\lambda + k \ln(\lambda) - \ln(k!)] = \tag{17}$$
$$\lambda - \lambda \ln(\lambda) + \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \ln(k!).$$

This expression for the entropy of a Poisson random variable is in terms of an infinite series, which is well approximated by a finite truncation partial sum.

Bivariate Poisson entropy

The Bivariate Poisson case is instructive to the n-variate Poisson case. Consider:

$$P(x_1, x_2) = e^{-\lambda_{11}-\lambda_{22}-\lambda_{12}} \frac{{\lambda_{11}}^{x_1}}{x_1!} \frac{{\lambda_{22}}^{x_2}}{x_2!} \left(\sum_{a_{12}=0}^{min(x_1,x_2)} \frac{x_1!}{(x_1 - a_{12})!} \frac{x_2!}{(x_2 - a_{12})! a_{12}!} \left(\frac{\lambda_{12}}{\lambda_{11}\lambda_{22}}\right)^{a_{12}} \right). \tag{18}$$

Let,

$$d_{12} = \frac{\lambda_{12}}{\lambda_{11}\lambda_{22}}, \tag{19}$$

and,

$$D(x_1, x_2) = \sum_{a_{12}=0}^{min(x_1,x_2)} \frac{x_1!}{(x_1 - a_{12})!} \frac{x_2!}{(x_2 - a_{12})!} \frac{d_{12}^{a_{12}}}{a_{12}!}. \tag{20}$$

Then Eq. 18 will become:

$$P(x_1, x_2) = e^{-\lambda_{11}-\lambda_{22}-\lambda_{12}} \frac{{\lambda_{11}}^{x_1}}{x_1!} \frac{{\lambda_{22}}^{x_2}}{x_2!} D(x_1, x_2). \tag{21}$$

Now to get the joint entropy of the Bivariate Poisson we have:

$$H(X_1, X_2) = -\sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} P(x_1, x_2) \ln(P(x_1, x_2)) = -\sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty}$$
$$e^{-\lambda_{11}-\lambda_{22}-\lambda_{12}} \frac{{\lambda_{11}}^{x_1}}{x_1!} \frac{{\lambda_{22}}^{x_2}}{x_2!} D(x_1, x_2) \times \tag{22}$$
$$[-\lambda_{11} - \lambda_{22} - \lambda_{12} + x_1 \ln(\lambda_{11}) + x_2 \ln(\lambda_{22}) - \ln(x_1!) - \ln(x_2!)$$
$$+ \ln(D(x_1, x_2))].$$

A scenario of interest arises when λ_{11}, λ_{22}, and λ_{12} are all small and $\lambda_{12} \ll \lambda_{11}\lambda_{22}$. In this case we have

$$D(x_1, x_2) \approx \sum_{a_{12}=0}^{min(x_1,x_2)} \frac{d_{12}^{a_{12}}}{a_{12}!}, \tag{23}$$

since the d_{12} term dominates. Small λ_{11} and λ_{22} ensures that the large x_1 and x_2 terms to become insignificant in Eq. 22. Thus, $D(x_1, x_2) \approx 1 + \frac{d_{12}^2}{2!} + ... \approx 1$. Grouping terms and remembering (the middle part of) Eq. 17, and estimating $D(x_1, x_2) = 1$, a finite partial sum of Eq. 22 can be written:

$$H(X_1, X_2) = e^{-\lambda_{12}}[H(X_1) + H(X_2) + \lambda_{12}]. \tag{24}$$

Remembering the assumption $\lambda_{12} << 1$, the expression Eq. 24 reduces further:

$$H(X_1, X_2) = [H(X_1) + H(X_2) + \lambda_{12}]. \tag{25}$$

As Fig. 3 shows, this approximation works well when $d_{12} << 1 \implies \lambda_{12} << \lambda_{11}\lambda_{22}$, and in this regime the error will be small. Similar analysis can be carried out for the larger multivariate cases which allows us to arrive at a general formula for our approximation given by:

$$H(X_1, X_2, ..., X_n) \approx [H(X_1) + ...H(X_n) + \sum_{j>i} \lambda_{ij}], \tag{26}$$

where we are assuming that λ_{ij} are small for all (i, j) pairs. Fortunately as we can see in Eq. 26, all of the quantities on the right hand side are computationally efficient to compute. This in fact greatly reduces the computational time necessary for estimation of the joint entropy. This formulation requires asymptotic assumptions that may not be valid in general in nature. However we find empirically in simulations that by scaling the rates λ_{ij} to be in [0, 1] the estimate performs well, as described by Fig. 3 and, verified in the network simulations, regardless of what the true underlying rates this scaling produces similar results.

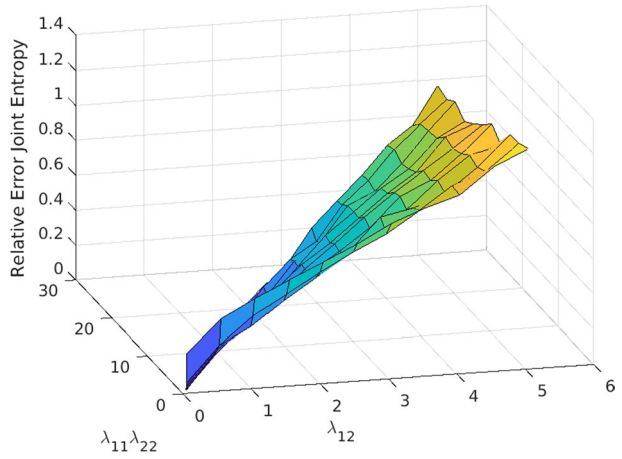

Fig. 3 The relative error in the joint entropy calculation between the joint entropy calculated through truncation and the joint entropy calculated by our approximation. It is clear that when both λ_{12} and $\lambda_{11}\lambda_{22}$ are small, the relative error is small. Thus we expect this approximation to work well when all of the estimated rates are small. In practice we find that when scaling the rates to be in [0, 1] we get good results, regardless of how high the true rates were

Now we have available to us an approximation of the joint entropy for Poisson variables, which can be used to infer the edges of a network (in a Granger causal sense) by estimating the conditional mutual informations necessary for the oCSE algorithm.

As a note of caution, consider that when calculating the mutual information in the Poisson model, care must be taken due to how the marginals of a joint Poisson process are drawn. For example from Eq. 10 it may be tempting to assert:

$$I_{Poisson}(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2), \tag{27}$$

with $X_1 \sim$ Poisson (λ_{11}) and $X_2 \sim$ Poisson (λ_{22}). However this is not exactly correct, though the error here is subtle. In fact we must make a small change to Eq. 27 to be:

$$I_{Poisson}(X_1, X_2) = H(\hat{X_1}) + H(\hat{X_2}) - H(X_1, X_2), \tag{28}$$

here $X_1 \sim$ Poisson (λ_{11}) and $X_2 \sim$ Poisson (λ_{22}), but $\hat{X_1} \sim$ Poisson $(\lambda_{11} + \lambda_{12})$ and $\hat{X_2} \sim$ Poisson $(\lambda_{22} + \lambda_{12})$. This subtle difference is important, because without recognizing this fact, the calculated mutual information becomes negative, which violates our well established condition that mutual information be positive. The need for $\hat{X_1}$ and $\hat{X_2}$ is apparent from Eq. 5, when two Poisson random variables are summed together their *marginals* then are drawn from the sum of the underlying rate (i.e. λ_{ii}) and the coupling rate (i.e. λ_{ij}). This also transfers to computing the conditional mutual information. To better illuminate this calculation it is helpful to refer to Fig. 2b.

$$I(X, Y \mid Z) = H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z). \tag{29}$$

In the special case presented in Fig. 2b Eq. 29 becomes

$$I(X, Y \mid Z) = I(X, Y) = H(X) + H(Y) - H(X, Y), \tag{30}$$

therefore,

$$H(X) + H(Y) - H(X, Y) = H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z). \tag{31}$$

In this special case we can note the following:

$$H(Y, Z) = H(Y) + H(Z). \tag{32}$$

Applying Eq. 32 to Eq. 31 we find that:

$$H(X) - H(X, Y) = H(X, Z) - H(X, Y, Z). \tag{33}$$

We know from Eq. 28 that in the Poisson case this becomes:

$$H(\hat{X}) - H(X, Y) = H(X, Z) - H(X, Y, Z). \tag{34}$$

Applying the following facts to Eq. 34

$$\begin{cases} H(X, Y, Z) = H(X, Y) + H(Z), \\ \text{and } H(X, Z) = H(X) + H(Z), \end{cases} \tag{35}$$

we find that:

$$H(\hat{X}) = H(X). \tag{36}$$

Similar analysis also shows that:

$$H(\hat{Y}) = H(Y), \tag{37}$$

this implies that we must use the Poisson marginals in the computation of the conditional mutual information. That is in the Poisson case we must have:

$$I(X, Y \mid Z) = H(\hat{X}, Z) + H(\hat{Y}, Z) - H(X, Y, Z) - H(Z). \tag{38}$$

Note the use of $\hat{X}$ and $\hat{Y}$ in this case. This distinction in the Poisson case is important because we note that without using the proper marginals the computation results in *negative* conditional mutual information which is clearly not correct since conditional mutual information must be positive (Cover and Thomas 2012).

Importantly the new definition given in Eq. 26 becomes more computationally efficient than computing the Poisson joint entropy directly from the joint probability. This requires calculation of only separate *single* variate entropies which is requires less computation. This naturally leads to the question of the accuracy of this new model. As can be seen in Fig. 4 the new definition of entropy still leads to accurate identification of network structure. This new definition also fits into the general framework of entropy which was developed above, allowing us to apply the optimal

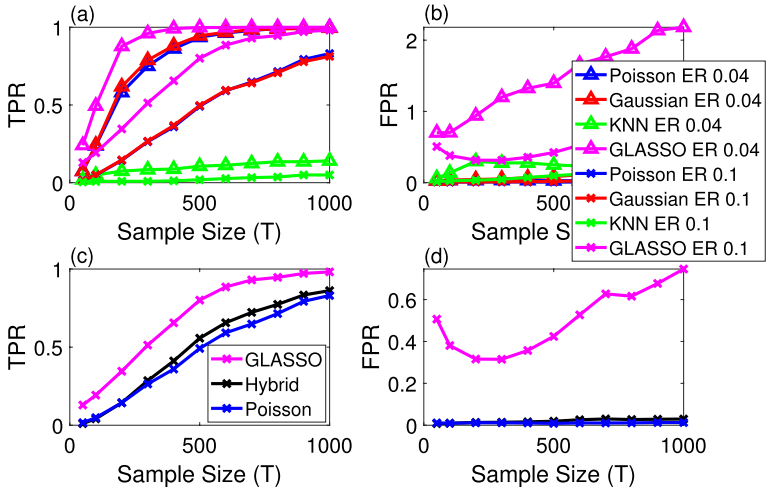

Fig. 4 True Positive and False Positive Rates for several test methods on ER graphs of two different levels of sparsity. Erdős-Rényi (ER) graphs with triangles for a 50 nodes graph with strong sparsity due to $p = 0.04$, and the x's for 50 nodes ER graphs with due to denser $p = 0.1$. The magenta lines represent GLASSO, the blue lines represent the Poisson oMII, the red lines represent the Gaussian oMII, and the green lines represent the KNN oMII. In **a** the true positive rate (TPR) is shown for different sample sizes, each point is averaged over 50 realizations of the network dynamics. In **b** the false positive rate (FPR) is shown. Clearly GLASSO finds more true edges, but at the expense of a significantly higher false positives. In fact, for the highly sparse ER network GLASSO finds 3 times as many edges as actually exist in the network with 1000 data points. The FPR increases with data set. As can be seen the Gaussian oMII performs as well as Poisson oMII in TPR with the KNN performing poorly, but the Poisson oMII significantly outperforms all other methods in terms of FPR. It appears that the Poisson oMII is the only method that converges to the true network structure with increasing sample size. **c** Comparing TPR between GLASSO, the hybrid method and Poisson oCSE. The hybrid method has an increased TPR relative to Poisson oCSE. **d** The FPR increases slightly for the hybrid method, but is substantially lower than GLASSO

mutual information interaction (oMII) (Ambegedara et al. 2016) algorithm, which is a version of oCSE without time-shift, to the data.

Network structure and inference

In a gene interaction network, understanding how future treatments could be developed, especially in the cases where more than a single gene may be implicated in a disease, may help in designing targeted for therapies. Genes interact with outcome such as disease reduces to a network inference problem. We do not assume apriori knowledge of the underlying network structure, but instead we have data describing time series of evolving stochastic processes at each of the states, related to each individual gene. The network is stated as a graph $\mathcal{G}$ defined as a set of vertices $\mathcal{V} \subset \mathbb{N}$ and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. Note that $|\mathcal{V}| = n$ denotes that there are n vertices (or nodes) in $\mathcal{G}$, by the cardinality, $|\cdot|$, of a set. The adjacency matrix $\mathcal{A} \in \mathbb{N}_0^{n \times n}$ is a convenient way to encode a graph,

$$\begin{cases} \mathcal{A}_{ij} = 1 \text{ if } (i,j) \in \mathcal{E}, \\ \mathcal{A}_{ij} = 0 \text{ otherwise}. \end{cases} \tag{39}$$

When a system has a graph structure it is often referred to as a network. The adjacency matrix then encodes the network structure of the system. Our goal is to estimate network structure $\hat{\mathcal{A}}$ closely as possible to the true network structure $\mathcal{A}$, that is we want, $\sum_{i,j} |\mathcal{A} - \hat{\mathcal{A}}|$, to be as small as possible (ideally 0). We would also like for this to be accomplished with as little data (t) as possible, since we are often limited in the amount of real world data we receive. Our estimation of the network structure relies on nodes sharing information with one another. Thus $\hat{\mathcal{A}}$ may be thought of as which nodes are directly communicating with one another, rather than strictly being the physical structure. In our previous work, (Sun 2014; Sun et al. 2015), we proved that under mild hypothesis, the multi-variate stochastic evolving by coupling on a complex network can be derived perfectly by optimal causation entropy (oCSE), errors arising from estimation issues such as model entropies of observations from various distributions, and finite data effects, but the information network structure align accurately in most situations.

In the first example demonstration of our methods, we benchmark with synthetic simulated by the multivariate Poisson model, Eqs. 2 and 4. To explicitly incorporate the adjacency matrix $\mathcal{A}$ and noise E as shown in Allen (2013), Gallopin et al. (2013), consider as:

$$X = BY + E, \tag{40}$$

$$B = [I_n; P \odot (1_n tri(\mathcal{A})^T)] \tag{41}$$

where I_n is the $(n \times n)$ identity matrix, $P \in \mathbb{N}_0^{n \times (m-n)}$ is a permutation matrix with exactly n ones per row $\odot$ represents the Hadamard product (componentwise multiplication of same sized arrays). $1_n \in \mathbb{N}^{n \times 1}$ is the vector of all ones, and $tri(\mathcal{A}) \in \mathbb{N}_0^{\frac{n(n-1)}{2} \times 1}$ denotes the vectorized upper triangular portion of the adjacency matrix, and $E \in \mathbb{N}_0^{n \times t}$. In the results section, the methods developed will be contrasted with GLASSO, a

popular method used for benchmark comparison as discussed in Allen (2013), Gallopin et al. (2013) and elsewhere for inference of gene expression networks.

We have established in previous discussion that there is no analytical solution for the entropy of the multivariate Poisson, instead an approximation has been made. Since the Poisson distribution resembles the Gaussian distribution often the latter is assumed for estimates, we thus compare the performance of oMII assuming both distribution types. Figure 4 shows that the oMII method, but even using the rough Gaussian best estimates of entropies, nonetheless does reasonably well finding the true edges with a high true positive rate (TPR). This is contrasted to network inference based on other entropy estimators, including the nonparametric kNN method, GLASSO, both of which are discussed below, and also the Poisson estimator developed here. However, the Gaussian oMII finds the edges at the expense of a much larger false positive rate (FPR). Specifically, define TPR and FPR as follows: let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be the true network structure and $\hat{\mathcal{G}} = \{\hat{\mathcal{V}}, \hat{\mathcal{E}}\}$ be the estimated network structure. Then:

$$\text{TPR} = \frac{|\mathcal{E} \cap \hat{\mathcal{E}}|}{|\mathcal{E}|}, \tag{42}$$

and

$$\text{FPR} = \frac{|\hat{\mathcal{E}} \setminus \mathcal{E}|}{|\mathcal{E}|}. \tag{43}$$

In this case $\setminus$ represents set subtraction. Note that from this definition $0 \leq \text{TPR} \leq 1$ while $\text{FPR} \geq 0$.

Results

Synthetic data

We compare the performance of several methods on simulated data sets, including various types of oMII, as well as GLASSO (Friedman et al. 2008). Unlike oMII, which involves conditional mutual information as its engine, GLASSO involves maximizing the log-likelihood provided in Eq. 44 over values of a regularization parameter ρ,

$$\mathcal{L}(X, \rho) = \log(\det(\hat{A})) - \text{trace}(\text{Cov}(X)\hat{A}) - \rho||\hat{A}||_1. \tag{44}$$

A common method for the choice of ρ is maximazation of the Bayesian information criterion (BIC). We utilize 1000 log-spaced values of ρ in $[10^{-2}, 1]$ which varies $\hat{A}$ between a complete network to a completely disconnected network with zero edges. Following (Gallopin et al. 2013), first we use a box-cox transformation of the Poisson distributed data, to make the data more Gaussian like, prior to using GLASSO. The box-cox transformation of a random variable z is

$$bc(z \mid \gamma) = \begin{cases} \frac{z^\gamma - 1}{\gamma} \text{ if} \gamma \neq 0 \\ \log(z) \text{ if} \gamma = 0. \end{cases} \tag{45}$$

GLASSO results are shown in Fig. 4.

The Poisson oMII method is tested on data simulated as described in the section above. In Fig. 4 each data point is averaged over 50 realizations of the network dynamics. Two different Erdős-Rényi (ER) graph types are used, one with $p = 0.04$ and one with $p = 0.1$. The parameter p in an ER graph controls the sparsity of the graph, thus the graphs with $p = 0.1$ will have considerably more edges on average than graphs with $p = 0.04$. For these simulations $n = 50$ was chosen. The rates were chosen to be $\lambda_{ij} = 1$ $(\forall\, i, j)$ and $E_i \sim$ Poisson (0.5) $(\forall\, i)$ where $E_i \in \mathbb{N}_0^{t\times 1}$ are the columns of E. This is the high SNR scenario from Allen (2013). To estimate the rates, we simply use correlation between all pairs in the data. We note that this differs from above where we utilized the covariance matrix. Using correlation rather than covariance guarantees the calculated rates will be relatively small since the values of correlation do not exceed 1 in absolute value, this allows the estimated rates to stay in the small relative error regime shown in Fig. 3. The correlation matrix then gives us all of the off diagonal rates λ_{ij} $(i \neq j)$ and to obtain the rates λ_{ij} $(i = j)$ we can see from Eq. 5 that we simply need to subtract the sum of the non-diagonal elements from the diagonal elements. That is if we let

$$\text{Corr}\,(X) = \begin{bmatrix} e_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{12} & e_{22} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \cdots & e_{nn} \end{bmatrix}, \tag{46}$$

then $\lambda_{ii} = e_{ii} - \sum_{j\neq i} \lambda_{ij}$. In Fig. 4 it can be seen that in terms of TPR all of the methods perform quite well with the exception of the KNN version of oMII which exhibits poor performance across all examined sample sizes, likely due to slow convergence. In fact, for networks with few connections the poorest performing method in terms of TPR is the Poisson oMII method, with the best performing method being GLASSO. However GLASSO produces a very high FPR, in fact GLASSO finds *more* false positives than there are total edges in the true network, thus producing an FPR of greater than 1. By contrast both Gaussian and Poisson versions of oMII produce significantly lower FPR and the Poisson oMII produces the lowest rate of FPR across all sample sizes. It should be noted as well that the FPR of the Poisson version of oMII maintains an approximately constant level across all sample sizes, while the Gaussian version of oMII has an increasing FPR with sample size. For the denser networks, which had an expected average degree of 5, as expected all methods had a decreased TPR for low sample size. The FPR also fell for all methods due to the larger denominator (more edges). The conclusions remain the same for both network densities.

We offer a comparison of the complexity as well. For the oCSE/oMII algorithm it is difficult to pin down the theoretical complexity as has been noted previously (Runge 2018), however it appears from numerical experiments to scale polynomially in time (Sun et al. 2015). The computation time of oCSE is generally related to the number of edges in the network, more so than its size in the situation when the network is sparse. GLASSO on the other hand was found to be $\mathcal{O}(n^3)$ (Friedman et al. 2008). For this reason, it is difficult to compare the performance between oCSE and GLASSO directly though we compare the performance in Table 1. The listed performance

Table 1 Computational complexity

	Poisson	Gaussian	KNN	GLASSO
Full network ($n = 50$)	5824.1	109.2	3265.6	20.4
CMI $n = 3, t = 100$	0.0175	0.0002	0.0021	–
CMI $n = 3, t = 1000$	0.0123	0.0002	0.0170	–
CMI $n = 3, t = 5000$	0.0119	0.0002	0.2944	–
CMI $n = 502, t = 100$	0.0279	0.0082	0.0019	–
CMI $n = 502, t = 1000$	0.0270	0.0084	0.0708	–
CMI $n = 502, t = 5000$	0.0271	0.0070	2.1338	–
CMI $n = 1002, t = 100$	0.0486	0.0377	0.0022	–
CMI $n = 1002, t = 1000$	0.0446	0.0402	0.1187	–
CMI $n = 1002, t = 5000$	0.0438	0.0314	4.6982	–

assumes sequential processing, however we note that the performance of oCSE can be substantially improved in parallel as the permutation test is performed many times and need not be sequential. Additionally the edges for each node can be estimated in parallel. As can be seen, GLASSO has the best performance in terms of speed over the whole network in serial. Also for calculation of the conditional mutual information (CMI) the Gaussian version of oCSE tends to be the fastest, though for large n situations the speed of the Poisson version is roughly comparable. Finally for the large t scenario, the KNN version gets substantially slower, as would be expected.

In the first row we compare the performance of all of the algorithms used in the paper on the denser network scenario. All times listed are in seconds. For the full network time the times are averaged over 50 networks, and we can see that GLASSO has the best performance in terms of time. In the next rows we compare the performance of the 3 versions of oCSE in terms of the average time it takes to calculate the conditional mutual information, with each variation averaged over 1000 runs. We see that the performance of the Poisson version of oCSE is generally quite a bit slower for a small number of variables (n) and small length time series (t), but as n and t grow the performance of the Poisson version of oCSE approaches that of the Gaussian oCSE. The non-paramaetric (KNN) estimator is most sensitive to large t.

Breast cancer dataset

We now examine data derived from breast cancer patients who have been screened for different micro RNA's (miRNA's) occurrence counts of is analyzed by the Poisson oMII method featured in this paper. In previous work (Allen 2013; Gallopin et al. 2013) the gene expression data has been assumed to be drawn from a multivariate Poisson distribution and we follow that convention in this work, while noting that in some cases the marginals are not perfectly Poisson but rather have some overdispersion (i.e. the situation when the variance is larger than the mean). These data sets are publicly available at https://portal.gdc.cancer.gov website, described as TCGA-BRCA sequencing miRNA. In this case, $t = 1207$ and $n = 1881$ different miRNA samples are available. Of these, 1881 miRNA's ≈ 1000 pass the two sample Kolmogorov–Smirnov (KS) (Lilliefors 1967)

test comparing to the Poisson distribution, to confidence level $\alpha = 0.05$. The remaining $\approx$ 900 miRNA data were then scaled as follows:

$$x_i^* \in \mathbb{N}_0^{1207\times 1} = \lfloor \frac{x_i}{< x_i >} \rfloor. \tag{47}$$

The notation, $< \cdot >$ represents the mean and $\lfloor \cdot \rfloor$ componentwise, to integers. The scaled data is well fitting, again by KS-test, to a negative binomial distribution, with only $\approx$ 200 failing as both Poisson and negative binomial. Recall that the Poisson distribution is a special case of the negative binomial distribution, since:

$$P_{NegBin}(k) = \binom{k+r-1}{k} \lambda^k (1-\lambda)^r. \tag{48}$$

In the limit, $r \to \infty$ in Eq. 48 it is easy to see that the term $(1-\lambda)^r \to e^{-\lambda}$, and rewriting $\binom{k+r-1}{k} = \frac{(k+r-1)!}{k!(r-1)!} \to \frac{1}{k!}$. Combining these facts, as $r \to \infty$, the negative binomial distribution limits to a Poisson distribution.

Given that the majority of this miRNA data is distributed as scaled negative binomial (the Poisson data also can be fit as negative binomial) we must interpret the results of with caution especially in light of the results shown in Fig. 4. The results of the application of the Poisson oMII still are interesting, especially in light of the fact that the negative binomial distribution can be viewed as a compound Poisson distribution (Anscombe 1950; Bissell 1972). To obtain the networks shown in Fig. 5 we first restricted the data to having a minimum of $>$ 100 total counts,this was to avoid including data that had zero variation or near zero variation. This restriction left us with 1072 miRNA's, oMII was then used to analyze the remaining miRNA data without any further pre-processing, which resulted in the network shown in Fig. 5. The network has many miRNA's which are non-interacting, however there is a large weakly connected component. Focusing on the nodes which are members of the largest weakly connected component (LWCC) we found that many miRNA's that have been previously identified as up or down regulated in breast cancer end up in this component, this component included most of the miRNA's listed in Table 1 of Iorio et al. (2005). The miRNA's which land in the LWCC will be labeled as *interesting* miRNA's for brevity.

Focusing on this set of 656 miRNA's, the plot of Fig. 5 focuses in on this component by sizing the nodes relative to their out degree. The nodes with no out degree are so small that they are difficult to see in the figure, while the nodes with largest out degree are prominent. A feature of this network is that there are miRNA's that are "drivers" of the network, in that they have much larger out degree than the majority of other nodes. We list the top 20 miRNA's in order of their centrality based on out degree, betweenness centrality and eigenvector centrality in Table 2. For all three measures the top 4 miRNA's are identically ordered, all 4 of which have been noted for a prominent role in breast cancer (Iorio et al. 2005; Lim 2013; Antolín 2015; Thammaiah and Jayaram 2016; Medimegh et al. 2014; Tanic et al. 2015) and they seem to be the main drivers. This suggests that it may be possible to target a small number of miRNA's for some desired behavior of the system of miRNA's in drug development.

Table 2 The top 20 genes discovered from the hybrid method in terms of: out degree, betweenness centrality, and eigenvector centrality. All of these genes have been linked to breast cancer by previous studies

Out degree	Betweenness centrality	Eigenvector centrality
Mir-200c	Mir-200c	Mir-200c
Mir-141	Mir-141	Mir-141
Mir-143	Mir-143	Mir-143
Mir-200a	Mir-200a	Mir-200a
Mir-21	Mir-205	Mir-205
Mir-205	Mir-21	Mir-21
Mir-30a	Mir-26b	Mir-30a
Mir-26b	Mir-30a	Mir-183
Mir-183	Mir-183	Mir-26b
Mir-199b	Mir-199b	Mir-326
Mir-210	Mir-125b-2	Mir-200b
Mir-125b-2	Mir-134	Mir-210
Mir-134	Mir-326	Mir-125b-2
Mir-326	Mir-3607	Mir-199b
Mir-200b	Mir-379	Mir-429
Mir-379	Mir-210	Mir-32
Mir-3607	Mir-1976	Mir-3607
Mir 429	Mir-150	Mir-134
Mir-32	Mir-203a	Mir-766
Mir-337	Mir-100	Mir-100

Conclusion

In this paper we have given an approximation to the mutual information of a multivariate Poisson system, which is needed for applications such as inferring the gene expression network. We have shown through numerical experiments that this approximation works efficiently, and the results of network estimation indicate that the approximation is justified. We have also developed the oMII (and by extension the oCSE) algorithm for computation of the causation entropy of a Poisson system based on the joint entropy approximation discussed above. We have shown that this model is superior to simply assuming the data is Gaussian, which is likely related to the strange behavior of the marginals in a Poisson sytem, as we have outlined above. The Poisson oMII algorithm also significantly outperforms the nonparametric KNN version of oMII. Finally, we have applied the Poisson oMII algorithm to a breast cancer miRNA expression count dataset, which has produced potentially interesting insights into the network of miRNA's as it relates to breast cancer. Our network inference on the breast cancer miRNA network has shown that there is a relationship between the highest variance (in expression values) of miRNA's. There seems to be unidirectional connections between these miRNA's, with certain miRNA's taking on the role of drivers in the network. This may suggest a future course of action for future drug development.

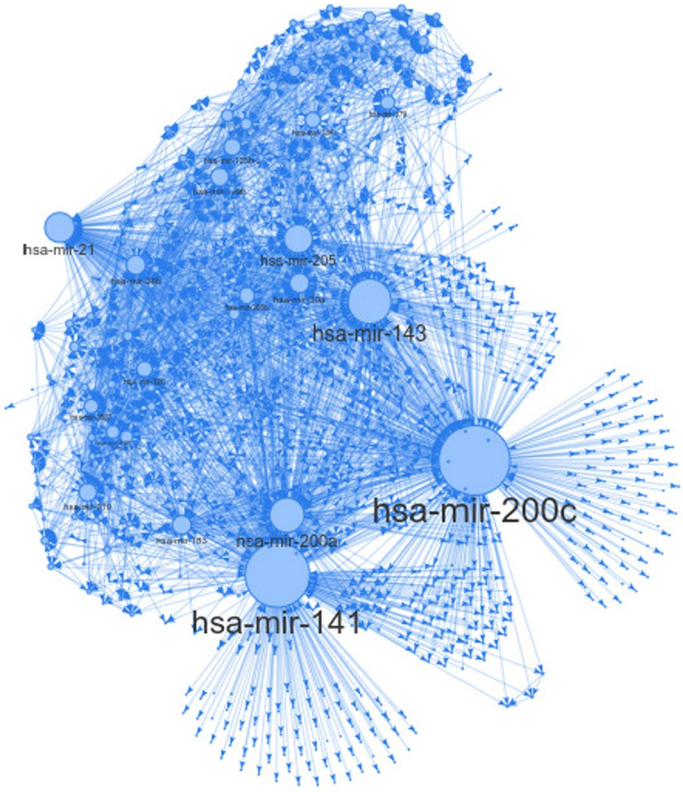

Fig. 5 Example network generated by the hybrid oMII algorithm. Nodes and text are sized relative to the out degree of the node. The nodes with largest out degree have previously been connected with breast cancer

Appendix 1 covariance of multivarite Poisson

Below we offer proof of Eq. 5.

Proof *Covariance of the multivariate Poisson*

In the model presented in Eqs. 2, 3, 4, we can see that:

$$\begin{aligned} x_1 &= y_{11} + y_{12} + ... + y_{1n} \\ x_2 &= y_{12} + y_{22} + ... + y_{2n} \\ &\vdots \\ x_n &= y_{1n} + y_{1n} + ... + y_{nn} \end{aligned} \tag{49}$$

Without loss of generality we will look at the pair $(i = 1, j = 2)$. In this case we see that the covariance between this pair of random variables is defined:

$$\text{cov}\,(x_1, x_2) = \mathbb{E}[x_1 x_2] - \mathbb{E}[x_1]\mathbb{E}[x_2], \tag{50}$$

Considering Eqs. 49, 50 and noting $y_{12} = y_{21}$, we have:

$$\begin{aligned}\operatorname{cov}(x_1, x_2) &= \mathbb{E}\left[y_{12}^2 + \sum_{\substack{i=1\\ i\neq 2}}^{n}\sum_{j=2}^{n} y_{1i}y_{2j}\right] \\ &- \mathbb{E}[y_{11} + y_{12} + \ldots + y_{1n}]\mathbb{E}[y_{12} + y_{22} + \ldots + y_{2n}]\end{aligned} \tag{51}$$

Because the expectation is a linear operator,Eq. 51 can be expressed as:

$$\begin{aligned}\operatorname{cov}(x_1, x_2) &= E[y_{12}^2] + \mathbb{E}\left[\sum_{\substack{i=1\\ i\neq 2}}^{n}\sum_{j=2}^{n} y_{1i}y_{2j}\right] - \\ &\left(\mathbb{E}[y_{12}] + \sum_{\substack{i=1\\ i\neq 2}}^{n}\mathbb{E}[y_{1i}]\right)\left(\mathbb{E}[y_{12}] + \sum_{j=2}^{n}\mathbb{E}\left[y_{2j}\right]\right).\end{aligned} \tag{52}$$

From the independence of each y_{ij} the covariance can thus be expressed:

$$\begin{aligned}\operatorname{cov}(x_1, x_2) =& \mathbb{E}[y_{12}^2] + \sum_{\substack{i=1\\ i\neq 2}}^{n}\sum_{j=2}^{n}\mathbb{E}\left[y_{1i}y_{2j}\right] - \\ &\mathbb{E}^2[y_{12}] - \sum_{\substack{i=1\\ i\neq 2}}^{n}\sum_{j=2}^{n}\mathbb{E}\left[y_{1i}y_{2j}\right] = \\ &\mathbb{E}[y_{12}^2] - \mathbb{E}^2[y_{12}] = \\ &\operatorname{Var}(y_{12}).\end{aligned} \tag{53}$$

Since y_{12} is independent Poisson and from the variance of an independent Poisson random variable $\operatorname{Var}(y_{12}) = \lambda_{12}$. Applying this to each $i, j(i \neq j)$ pair gives the desired covariance structure. □

Abbreviations

TPP	Temporal point process
KNN	K-nearest neighbors
CSE	Causation entropy
oCSE	Optimal causation entropy
miRNA	Micro-ribonucleic acid
TE	Transfer entropy
KSG	Kraskov–Stobauer–Grassberger
CDC	Centers for disease control
fMRI	Functional magnetic resonance imaaging
oMII	Optimal mutual information intereaction
KS	Kolmogorov–Smirnov
LWCC	Largest weakly connected component

Acknowledgements
Not Applicable

Author Contributions
J.F. is the primary author, J.S. and E.B. gave insights into the theory and were an integral part of the editing process. All authors read and approved the final manuscript.

Funding
E.B. was supported by the Army Research Offics (N68164-EG) and DARPA, J.F and J.S were supported by DARPA.

Availability of data and materials
All data and materials will be made available upon reasonable request to the corresponding author.

Declarations

Conflict of interest
All authors declare that they have no competing interests.

Received: 22 September 2021 Accepted: 30 September 2022
Published online: 11 October 2022

References

Allen GI, Liu Z (2013) A local Poisson graphical model for inferring networks from sequencing data. IEEE Trans Nano-Biosci 12:189–198
Ambegedara AS, Sun J, Janoyan K (2016) Bollt EM information theoretical noninvasive damage detection in bridge structures. Chaos 26:116312
Anscombe FJ (1950) Sampling theory of the negative binomial and logarithmic series distributions. Biometrika 37:358–382
Antolín S et al (2015) Circulating miR-200c and miR-141 and outcomes in patients with breast cancer. BMC Cancer 15:1–15
Bassett D et al (2011) Dynamic reconfiguration of human brain networks during learning. PNAS 108:7641–7646
Bissell AF (1972) A negative binomial model with varying element sizes. Biometrika 59:435–441
Bollt EM, Santitissadeekorn N (2013) Applied and computational measurable dynamics. SIAM
Cover T, Thomas J (2006) Elements of information theory, 2nd edn. Wiley, Hoboken
Cover TM, Thomas JA (2012) Elements of information theory. Wiley, Hoboken
De Boulle K et al (1993) A point mutation in the FMR-1 gene associated with fragile-X mental retardation. Nat Genet 3:31
Fish J, DeWitt A, Almomani AAR, Laurienti PJ, Bollt E (2021) Entropic regression for neurological motivated applications arxiv
Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441
Gallopin M, Rau A, Jaffrézic F (2013) A hierarchical Poisson log-normal model for network inference from RNA sequencing data. PloS One 8:e77503
Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econom J Econom Soc 37:424–438
Gregory PA, Bert AG, Paterson EL et al (2008) The MiR-200 family and MiR-205 regulate epithelial to mesechymal transition by targeting ZEB1 and SIP1. Nat Cell Biol 10:593
Guerrero-Cusumano JL (1995) The entropy of the multivariate Poisson: an approximation. Inf Sci 86:1–17
Inouye DI, Yang E, Allen GI, Ravikumar P (2017) A review of multivariate distributions for count data derived from the Poisson distribution. Wiley Interdiscip Rev Comp Stat 9:e1398
Iori G et al (2008) A network analysis of the Italian overnight money market. J Econ Dyn Control 32:259–278
Iorio MV, Ferracin M, Liu C, Veronese A et al (2005) MicroRNA gene expression deregulation in human breast cancer. Cancer Res 65:7065–7070
Karlis D, Meligotsidou L (2007) Finite mixtures of multivariate Poisson distributions with application. J Stat Plan Inference 137:1942–1960
Kraskov A, Stögbauer H (2004) Grassberger P estimating mutual information. Phys Rev E 69:066138
Lilliefors HW (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 62:399–402
Lim Y et al (2013) Epigenetic modulation of the miR-200 family is associated with transition to a breast cancer stem-cell-like state. J Cell Sci 126:2256–2266
Medimegh I, Troudi W, Stambouli N et al (2014) Wild-type genotypes of BRCA1 gene SNPs combined with micro-RNA over-expression in mammary tissue leading to familial breast cancer with an increased risk of distant metastases' occurrence. Med Oncol 31:255
Reiss RD (2012) A course on point processes. Springer Science, Berlin
Rogers CS et al (2008) Disruption of the CFTR gene produces a model of cystic fibrosis in newborn pigs. Science 321:1837–1841
Runge J (2018) Causal network reconstruction from time series: from theoretical assumptions to practical estimation. Chaos Interdiscip J Nonlinear Sci 28:075310

Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85:461
Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH (2005) Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet 37:435
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
Smith SM (2012) The future of FMRI connectivity. Neuroimage 62:1257–1266
Smith R (2015) A mutual information approach to calculating nonlinearity. Stat 4:291–303
Stoltz BJ, Harrington HA, Porter MA (2017) Persistent homology of time-dependent functional networks constructed from coupled time series. Chaos Interdiscip J Nonlinear Sci 27:047410
Sun J (2014) Bollt EM causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. Phys D 267:49
Sun J, Taylor D, Bollt EM (2015) Causal network inference by optimal causation entropy. SIAM J Appl Dyn Sys 14:73
Tanic M, Yanowski K, Gómez-López G et al (2015) MicroRNA expression signatures for the prediction of BRCA1/2 mutation-associated hereditary breast cancer in parrafin-embedded formalin-fixed breast tumors. Int J Cancer 136:593–602
Thammaiah CK, Jayaram S (2016) Role of let-7 family MicroRNA in breast cancer. Non Coding RNA Res 1:77–82
Yamanishi Y, Vert JP, Kanehisa M (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 21:i468–i477
Zhang Y, Zhao H, He X, Pei FD, Li GG (2016) Bayesian prediction of earthquake network based on space-time influence domain. Phys A 445:138–149

Publisher's Note