Neural Causality Detection for Multi-dimensional Point Processes

Tianyu Wang(✉), Christian Walder, and Tom Gedeon

Australia National University College of Engineering and Computer Science, Canberra, Australia
u6014854@anu.edu.au

Abstract. In the big data era, while correlation detection is relatively straightforward and successfully addressed by many techniques, causality detection does not have a generally-used solution. Causality provides valuable insights into data and guides further studies. With the overall assumption that causal influence can only be from prior history events, time plays an essential part in causality analysis, and this important feature means the data with strict temporal structure needs to be modelled. Traditionally, temporal point processes are employed to model data containing temporal structure information. The heuristic parameterization property of such models makes the task difficult. Domain related knowledge are needed to design proper parameterization. Recently, Recurrent Neural Networks (RNNs) have been used for time-related data modelling. RNN's trainable parameterization considerably reduces the dependency on domain-related knowledge. In this work, we show that combining neural network techniques with Granger causality framework has great potential by presenting an RNN model integrated with a Granger causality framework. The experimental results show that the same network structure can be applied to a variety of datasets and causalities are detected successfully.

Keywords: Granger causality · Recurrent neural network
Temporal point process

1 Introduction

In the big data era, extracting and understanding relations between data has become increasingly important. The goal is to analyse how influence flows between events from different sources or dimensions, i.e., whether one event's behaviour is affected by an earlier event and if so estimating the strength of that influence. Researchers have shown that many fields can benefit from causality detection.

Granger causality has been shown to be a reliable causality analysis method in fields including econometric [6] and neuroscience [8]. However, the number of the application scenarios of traditional Granger causality is significantly limited. The Granger causality utilizes the dependency between historical events and

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11304, pp. 509–521, 2018.
https://doi.org/10.1007/978-3-030-04212-7_45

present events to detect causal relations. For most sequence data, the dependency is complicated, but in most cases, a simple linear dependency is assumed, i.e. the value of the current event is a weighted sum of the past events. This oversimplified linear parameterization cannot capture complex dependencies, thereby limiting the performance of Granger causality applications and restricts application scenarios to regression problems on regular time grids. Efforts are made to overcome these limitation by employing radial basis function [1,10], kernel method [11] or locally linear neighborhoods [3,5]

Regarding sequence data modelling, point process models allow the training of sophisticated parameterized models with a maximum likelihood approach [2, 12]. Traditional point process models also have limitations. One limitation is that the sophisticated parameterization needs to be heuristically designed to suit the dependencies of the real world data. Thus, domain related knowledge plays an essential role in the model design. Heuristic parameterization also makes designed model not reusable through different domains. Recently, a recurrent neural network based model called Recurrent Marked Temporal Point Process (RMTPP) [4] was proposed targeting the heuristic parameterization problem for marked point process models, and achieved state-of-the-art results. This model solved the heuristic parameterization problem by approximating the dependency using a recurrent neural network. Experiments showed that the RMTPP model well approximated different one-dimensional point process models with minor hyperparameter tweaking and zero prior knowledge of the true parameterization.

In our work, we aimed to show that the advantage of recurrent neural network based models can be combined with Granger causality framework and provide a domain independent Granger causality detection model. The Granger causality is designed to explicitly detect causality between different dimensions, i.e, the model of each dimension should be independent given the input data. Here we can consider dimensions to be different event sources. To solve the dimension-wise modelling problem, we proposed the Recurrent Multi-dimensional Temporal Point Process (RMDTPP) model that can effectively model multi-dimensional data.

We tested our model on both data on regular time grids (the time differences are constant one and not explicitly modelled) and data with real-valued timestamps (the time differences are explicitly modelled as real value). Experimental results showed that RMDTPP can be integrated into the Granger causality framework seamlessly. Causality between dimensions are successfully detected and the RMDTPP model can be applied to different data without parameterization changing.

2 A Review of Multi-dimensional Point Process

2.1 Multi-dimensional Point Process

Point process is commonly used to study sequence data. Normally, one sequence contains multiple event points, each event point has a timestamp describing the time that the event is observed of the time that the event happens. When will the next event happen is affected by all the previous events. If all the events are generated from one source, we use one-dimensional point process to model the data sequence. Combining multiple one-dimensional point processes with extra parameters capturing interactions between them leads to a multi-dimensional point process. In a multi-dimensional point process, more than one dimension can evolve at the same time and influence each other. One multi-dimensional point process can be uniquely defined by it conditional intensity function (CIF) $\lambda^*(t)$. Here * serves as a reminder that the function is conditional on the past, as first introduced by Daley and Vere-Jones [7]. The CIF describes how many events are expected to occur in an arbitrary time interval. Dimensions can self-intervene or cross intervene in a multi-dimension point process, which means that the intensity of one dimension receives the historical influence from all dimensions.

If we take a multi-dimensional Hawkes point process as an example, the CIF of the multi-dimensional Hawkes is defined as follows:

$$\lambda(t)^{m*} = \lambda_0^m + \sum_{n=1}^{M} \sum_{t_i^n < t} \alpha^{mn} e^{-\beta^{mn}(t - t_i^n)}, \tag{1}$$

where λ_0^m is the initial intensity of m dimension, α^{mn} and β^{mn} are employed to capture the influence flowing from n dimension to m dimension [9,13]. The size of matrix α and matrix β is $M \times M$ if there are in total M dimensions.

The heuristic parameterization problem refers to that the parameterization shown in (1) specifies how the point process evolves. Each event from one dimension stimulates the intensity at the same strength. The intensities of all dimensions gradually decay over time. Only the sequence data that roughly follows the specified evolving pattern can be well modelled. Thus, before we define any CIF for a dataset, we need to know how the data evolves, which is not always possible.

If we treat the length of time interval between two consecutive events as a random variable, the distribution of this random variable can be specified and calculated by a conditional density function $f^*(t)$. One CIF is corresponding to one conditional density function via

$$f^{m*}(t) = \lambda^{m*}(t) \exp\left(-\int_{t_j}^{t} \lambda^{m*}(s) \mathrm{d}s\right). \tag{2}$$

The likelihood of the point process of dimension m is the product of the conditional density values of each event. One thing to notice, in a multi-dimensional point process, the process of each dimension ends at the same time, i.e. the domain $[0, T]$ of all dimensions are the same. However, the last event of one dimension may not happen at the timestamp T, which means that the process of one single dimension does not end with the last event on that dimension. Normally, we add a pseudo event for each dimension at timestamp T with the intensity $\lambda^{m*}(T) = 1$. We denote the number of events in dimension m including the pseudo event as L^m and t_i^m is the timestamp of i^{th} event in dimension m, $t_{L_m}^m = T$ and $t_0^m = 0$. We have

$$\mathcal{L}(D_m) = \prod_{i=1}^{L^m} f^{m*}(t_i^m) \tag{3}$$

$$= \prod_{i=1}^{L^m} \lambda^{m*}(t_i^m) \exp\left(- \int_{t_{i-1}^m}^{t_i^m} \lambda^{m*}(s)\mathrm{d}s \right). \tag{4}$$

When $i = L^m$, we have

$$\lambda^{m*}(t_i^m) \exp\left(- \int_{t_{i-1}^m}^{t_i^m} \lambda^{m*}(s)\mathrm{d}s \right) = \exp\left(- \int_{t_{i-1}^m}^{t_i^m} \lambda^{m*}(s)\mathrm{d}s \right), \tag{5}$$

which is the likelihood of the process from the last real event of dimension m to the end.

As described above, in a multi-dimensional point process, the intensity of one dimension is under the influence of the events from all other dimensions. During the interval from t_{i-1}^m to t_i^m, events from another dimension may happen and bring sudden changes to the intensity of dimension m. Thus, the integration has to be done piecewise as

$$\mathcal{L}(D_m) = \prod_{i=1}^{L^m} \lambda^{m*}(t_i^m) \exp\left(- \int_{t_{i-1}^m}^{t_i^m} \lambda^{m*}(s)\mathrm{d}s \right) \tag{6}$$

$$= \left(\prod_{i=1}^{L^m} \lambda^{m*}(t_i^m) \right) \exp\left(- \int_0^T \lambda^{m*}(s)\mathrm{d}s \right) \tag{7}$$

$$= \left(\prod_{i=1}^{L^m} \lambda^{m*}(t_i^m) \right) \exp\left(- \sum_{j=1}^{L} \int_{t_{j-1}}^{t_j} \lambda^{m*}(s)\mathrm{d}s \right), \tag{8}$$

where L is the total number of events produced by the multi-dimensional point process plus the pseudo-event at the end of the domain at timestamp T. The likelihood function of a multi-dimensional point process is the product of the likelihood of each single dimension point process as

$$\mathcal{L}(D) = \prod_{m=1}^{M} \mathcal{L}(D_m) \tag{9}$$

$$= \prod_{m=1}^{M} \left(\prod_{i=1}^{L^m} \lambda^{m*}(t_i^m) \right) \exp\left(- \sum_{j=1}^{L} \int_{t_{j-1}}^{t_j} \lambda^{m*}(s)\mathrm{d}s \right). \tag{10}$$

3 The Proposed Approach

3.1 Recurrent Multi-dimensional Temporal Point Process

While one hot is a group of bits where only one bit is allowed to be 1 and the rest are 0, many-hot allows as many bits to be 1 as needed. Simultaneous events can be easily represented by setting the corresponding bits to 1. In the RMDTPP model we use the many-hot representation to allow simultaneous events. The input of RMDTPP at step i contains the many-hot representation of the observation y_i and the time difference $d_i = t_i - t_{i-1}$. The hidden units of the RNN now should contain all the historical information needed to predict the conditional intensity of the selected dimension.

We adopted the CIF parameterization from RMTPP model [4] as

$$\lambda^*(t) = \exp(\mathbf{v}^\top \cdot \mathbf{h}_j + w(t - t_j) + b), \tag{11}$$

where $\mathbf{h}_j$ is the output of the RNN and $\mathbf{v}^\top$, w as well as b are learnable parameters. Then according to Eq. (2), the conditional density function is

$$\log f^*(t) = \mathbf{v}^\top \cdot \mathbf{h}_j + w(t-t_j) + b + \frac{1}{w}\exp(\mathbf{v}^\top \cdot \mathbf{h}_j + b) - \frac{1}{w}\exp(\mathbf{v}^\top \cdot \mathbf{h}_j + w(t-t_j) + b). \tag{12}$$

Equation (10) shows that the calculation of likelihood of dimension m only requires the conditional intensity value $\lambda^{m*}(t)$ and the integral of CIF between two events as $\Lambda^m_{t_{i+1}} = \exp\left(-\int_{t_{i-1}}^{t_i} \lambda^{m*}(s)\mathrm{d}s\right)$. Thus, at each step, we let the RMDTPP model outputs $\lambda^{m*}(t)$ and $\Lambda^m_{t_{i+1}}$.

In a formal way, given an M dimensional sequence data $D \triangleq \{(t_i, \tilde{y}_i)\}$, $i = 0, ..., L$, where $L-1$ is the total number of events in the sequence. $t_i \in [0, T]$ is the time stamp of event i, $t_0 = 0$, $t_L = T$ and $\tilde{y}_i \in \{0, 1\}^M$ is the many-hot representation of the dimensions of event i, $(\tilde{y}_i)_m$ is the m^{th} elements of $\tilde{y}_i$. Notice that there should be no $\tilde{y}$ that every element of $\tilde{y}$ is zero. That is for any t_i in D, there is at least one event is observed at that timestamp. Here, we make the start point $(t_0, \tilde{y}_0)$ and the end point $(t_L, \tilde{y}_L)$ of the point process as pseudo events to simplify our work and $\tilde{y}_0 = \tilde{y}_L = (0, 0, ..., 0)^T$. RMDTPP model maximizes the likelihood of point process of dimension m with duration T:

$$\mathcal{L}(D_m) = \prod_{i=1}^{L} [\lambda^{m*}(t_i)]^{(\tilde{y}_i)_m} \exp\left(-\int_{t_{i-1}}^{t_i} \lambda^{m*}(s)\mathrm{d}s\right). \tag{13}$$

With the output of the RMDTPP, we can compute the likelihood of the dimension $\mathcal{L}^m$ and use backpropagation to train the whole model in a maximum likelihood fashion. The structure of the RMDTPP model is shown in Fig. 1.

Another advantage of RMDTPP is that at each step after the output of RNN is computed, the likelihood calculation of each dimension can be run in parallel. We call the parallel training process as joint training. The hidden units in a joint training model should compress all the historical information instead of the information needed by a single dimension.

To achieve joint training, we initialize weights $\mathbf{v}$, w and b for each dimension. At each step, the output of the RNN is sent to each dimension, and from here the operations of each dimension are done independently and simultaneously. The likelihood is now the joint likelihood of all the dimensions, i.e. the product of the likelihood of each dimension. During the backpropagation phrase, the hidden layer of the RNN is updated by the gradients from all dimensions and the RNN is forced to learn a history representation that compresses the information needed by every dimension.

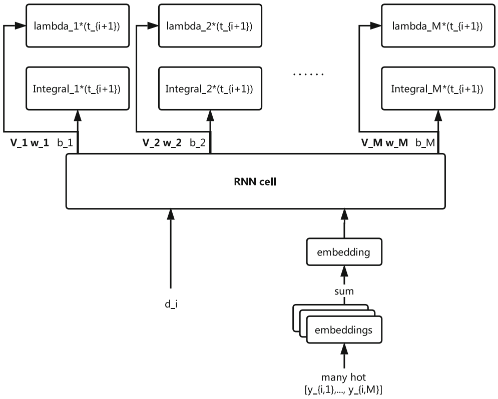


Fig. 1. Structure of RMDTPP model. M is the number of dimensions. As shown in the figure the conditional intensity and integral calculation of each dimension can run in parallel

3.2 Discrete RMDTPP

For data on a regular time grid, the length of each grid cell is fixed, which means that the time difference d between two time steps is a constant. In this case, the temporal structure of the data is simple and does not need to be modelled explicitly.

The input data is now a matrix D of M rows and T columns where M is the number of dimensions, and T is the length of the data. One row in D corresponds to the events record of one dimension. $D_{n,t} = 1$ if a event from dimension n is observed at time step t, else $D_{n,t} = 0$. Notice that in this setup, each column in D is a many-hot representation.

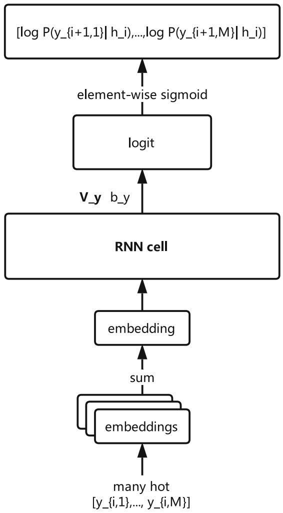


Fig. 2. Discrete RMDTPP structure

As shown in Fig. 2, the input of the discrete RMDTPP at time step t is simply the elements at the corresponding column $D_{:,t}$ and the output is a vector of predicted probabilities $P(D_{:,t+1} = 1|\mathbf{h}_t)$. The likelihood that the discrete RMDTPP tries to maximize is:

$$\log \mathcal{L}(D_n) = \sum_{t=1}^{T} \sum_{n=1}^{M} D_{n,t} P(D_{n,t} = 1|\mathbf{h}_t) + (1 - D_{n,t}) P(D_{n,t} = 0|\mathbf{h}_t). \quad (14)$$

3.3 Granger Causality

In this work we adopted the Granger causality defined by likelihood reduction [8]. The general idea is that for two dimensions X and Y, if the historical events of X contribute to the prediction of feature events of Y, then we say dimension X Granger-causes Y. The reason we use 'Granger-cause' or 'g-cause' in short instead of 'cause' is that there is a gap between the Granger causality and the real causality in terms of philosophy. The prediction contribution is measured by likelihood reduction. First, we train a model to use both historical events from dimension X and dimension Y to predict the feature event of Y. We use the model to calculate a likelihood $\mathcal{L}(Y)$ of the events of dimension Y. Then we train a new model to use only the historical events from dimension Y to predict the feature event of Y. We use the new model to calculate a new likelihood $\hat{\mathcal{L}}(Y)$ of the events of dimension Y. If $\hat{\mathcal{L}}(Y) - \mathcal{L}(Y) < 0$ then we say X contributes to the prediction of Y.

The detailed framework is introduced in Algorithm 1. Then the causality will be predicted according to the resulting likelihood reduction matrix.

Data: The training and testing dataset D_{train} and D_{test}
The number of dimensions M
Result: Likelihood reduction matrix Φ
$\Phi = M \times M$ zeros matrix;
$rmdtpp$ = RMDTPP.initialize(M);
$rmdtpp$.train(D_{train});
$\mathcal{L}(D_{test}) = [\mathcal{L}(D_{test\ 1}), ..., \mathcal{L}(D_{test\ M})] = rmtpp$.predict($D_{test}$);
for $m = 1; m \leq M; m = m + 1$ **do**
  $D'_{train} = D_{train} - D_{train\ m}$;
  $D'_{test} = D_{test} - D_{test\ m}$;
  $rmdtpp$ = RMDTPP.initialize(M-1);
  $rmdtpp$.train(D'_{train});
  $\mathcal{L}(D'_{test}) = [\mathcal{L}(D_{test\ 1}), ..., \mathcal{L}(D_{test\ m-1}), \mathcal{L}(D_{test\ m+1}), ..., \mathcal{L}(D_{test\ M})] =$ $rmdtpp$.predict(D'_{test});
  for $n = 1; n \leq M; n = n + 1$ **do**
    if $n < m$ **then**
      $\Phi_{n,m} = \mathcal{L}(D'_{test})_n - \mathcal{L}(D_{test})_n$
    end
    else
      $\Phi_{n,m} = \mathcal{L}(D'_{test})_n - \mathcal{L}(D_{test})_{n+1}$
    end
  end
end

Algorithm 1: Granger causality detection process

4 Experiment

4.1 Dataset and Evaluation Metric

To demonstrate the capacity of RMDTPP model, we first test our model on a three-dimensional piecewise homogeneous Poisson process dataset. The homogeneous Poisson point process is one of the most simple point processes. If we treat the time difference between two adjacent events in a homogeneous Poisson point process as a random variable r, then r obeys an exponential distribution. The probability density function of an exponential distribution is $\lambda e^{-\lambda x}$ when $x > 0$ and 0 when $x \leq 0$. The λ is a positive parameter. The intensity of a homogeneous Poisson process will be constant $\frac{1}{\lambda}$. We generate 1,000 sequences, each of them containing 300 events. 300 sequences are randomly chosen to serve as the test set. The rest 700 serve as the training set.

We also adopted the neural activity simulation (NAS) framework specified by [8]. A dataset containing 20 sequences is generated to test the discrete RMDTPP model. Each sequence contains 100,000 events from 5 different dimensions. Half of the dataset is used as training set and the other half is used as testing set.

Then the discrete NAS dataset is converted into a continuous NAS dataset by removing all the time step where there are no events observed and the time difference between events is calculated as integer.

Moreover, a four-dimensional Hawkes point process dataset is generated. Both the four-dimensional Hawkes point process dataset and the continuous NAS dataset are used to test the continuous RMDTPP model.

The causality ground truth is obtained according to the parameters the parameters of the data generation frameworks. If the parameters of the data generation model indicate that events from dimension A increase or decrease the intensity of dimension B then we consider that there is a causal relation from A to B.

We treat the causality detection as a binary classification problem. The receiver operating characteristic curve (ROC) and area under curve (AUC) are used for performance measurement. Higher AUC indicates a better classification performance.

4.2 Experiment Results

To demonstrate the data fitting performance of the RMDTPP model, we first test our model on the three-dimensional piecewise Poisson process dataset.

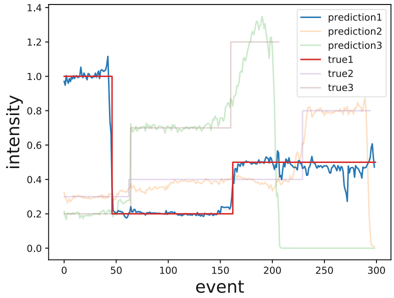
As shown in the Fig. 3, The RMDTPP model successfully predicts the conditional intensity value, captures the un-natural sudden change of the intensity and predicts intensity to be zero when the process of that dimension finishes. It is worth noticing that the model is trained using maximum likelihood approach. The model is not directly trained to minimize the gap between the predicted conditional intensity value and the true value. The result shows that our RMDTPP model can approximate the true CIF with no priori knowledge about it.

Then we combine the RMDTPP model with the Granger causality framework. The causality detection performance of the discrete RMDTPP model is tested on the discrete NAS dataset. The resulting likelihood reduction matrix and the ground truth causality matrix is shown in Fig. 4(a) and (b). We also test the causality detection performance of the continuous RMDTPP model on the continuous NAS dataset. The resulting likelihood reduction matrix is shown in Fig. 4(c). The ROC curve of both continuous and discrete NAS data is shown in Fig. 4(d).

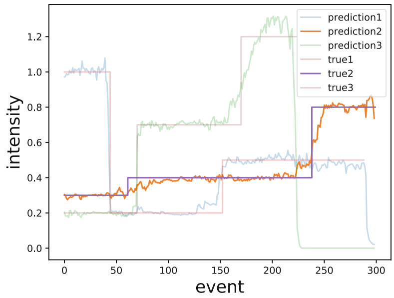
Considering that the NAS dataset is not a natural continuous time dataset, we then test the causality performance on the four-dimensional Hawkes point process dataset. The results are shown in Fig. 5.

The experiments results are reported in Table 1. The first row is the performance of the NAS model [8]. The NAS model itself is a discrete time point process. NAS dataset can be seen as the results of performing sampling operation on the corresponding NAS model.

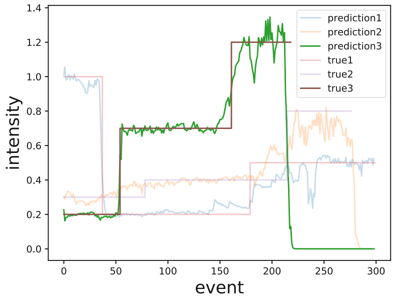
From the comparison, we can see that the performance of discrete RMDTPP matches the performance of the NAS model which has full awareness of the true parameterization. Continuous RMDTPP model is outperformed by both NAS



(a) Predicted intensity of dimension 1.



(b) Predicted intensity of dimension 2.



(c) Predicted intensity of dimension 3.

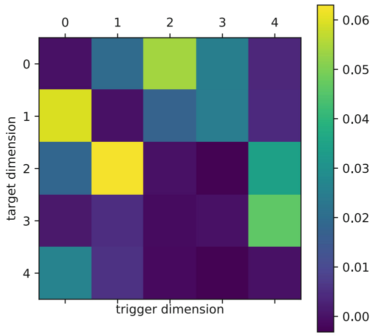
Fig. 3. Predicted intensity of the three-dimension piecewise Poisson process toy dataset.

model and discrete RMDTPP model on the NAS dataset. Part of the reason is that the likelihood reduction value is closely related to the true likelihood that is calculated with the full dataset. Likelihood reductions are calculated based on the true likelihood. The fluctuation of the true likelihood increases after we make the RMDTPP model to learn the real-valued time difference. As a result, it is harder to find a reasonable threshold for all dimension pairs.

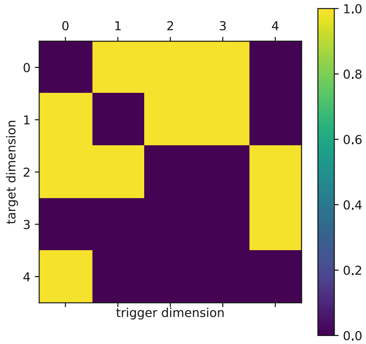
No previous application of the Granger causality framework with traditional continuous time point process models can be found. Since that the causality can be easily identified via the learned parameters model.

Table 1. Model AUC value comparison

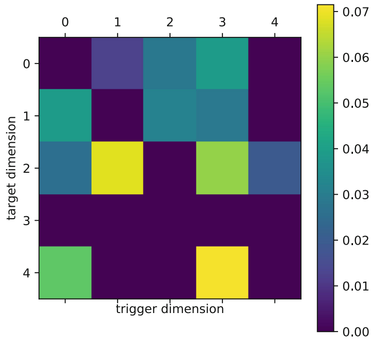
	NAS	4 dimensional hawkes point process
NAS model	**1.0**	N/A
Discrete RMDTPP	**1.0**	N/A
Continuous RMDTPP	0.83	**0.72**



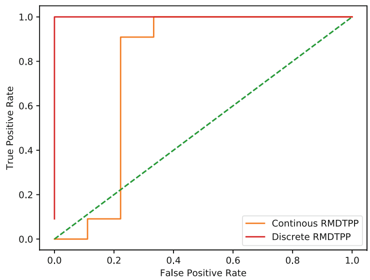
(a) Likelihood reductions of discrete NAS dataset.



(b) The ground truth causality matrix.



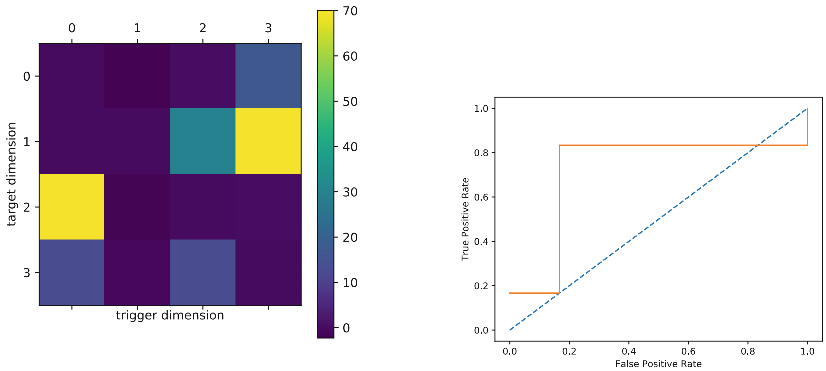
(c) RMDTPP likelihood reduction of NAS.



(d) ROC curve of the causality detection of NAS.

Fig. 4. The likelihood reduction matrices and the ROC curves.

However, only continuous RMDTPP model can work with both the NAS dataset and the multi-dimensional Hawkes point process dataset. The only hyperparameter that has to be manually adjusted is the size of the RNN units, which can be decided by observing the likelihood results and picking the one with the highest likelihood. Thus, we can say that combining RMDTPP model with Granger causality has a high potential. In discrete cases, the discrete RMDTPP model is practically usable. The heuristic parameterization problem of traditional point process model is overcome allowing Granger causality to be applied to various scenarios.



(a) Likelihood reduction of 4-dimensional Hawkes point process. In order to show heat map clearly, the max likelihood reduction is clamped to 70

(b) ROC curve of causality detection on 4-dimensional Hawkes point process.

Fig. 5. Resulting Granger causality matrices.

5 Conclusion

In this work, we propose the recurrent multi-dimension temporal point process model. RMDTPP model is inspired by the mathematical definition of the multi-dimensional point process. A separate likelihood calculation of each dimension allows seamless integration with Granger causality framework. Granger causality detection experiments show that the same RMDTPP model can be applied to various sequence data. The RMDTPP model also has limitations. In this conditional intensity function, w is a learned constant scalar value, which means that the time difference between the last event and current time $(t - t_j)$ can only linearly influence the current conditional intensity in log space. This feature is not true for all temporal point process models. It is possible to give the conditional intensity function a more flexible parameterization. Despite of the limitations, we proved that the combining neural network work techniques with Granger causality leads to a powerful model. The heuristic parameterization problem is overcome and causal relation between dimensions can be detected. Our work enables the application of Granger causality on sequence data from any domain.

References

1. Ancona, N., Marinazzo, D., Stramaglia, S.: Radial basis function approach to nonlinear granger causality of time series. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **70**(5 Pt 2), 056221 (2004)

2. Bacry, E., Delattre, S., Hoffmann, M., Muzy, J.: Some limit theorems for Hawkes processes and application to financial statistics. Stoch. Process. Their Appl. **123**(7), 2475–2499 (2013). https://doi.org/10.1016/j.spa.2013.04.007. http://www.sciencedirect.com/science/article/pii/S0304414913001026. A Special Issue on the Occasion of the 2013 International Year of Statistics
3. Chen, Y., Rangarajan, G., Feng, J., Ding, M.: Analyzing multiple nonlinear time series with extended Granger causality. Phys. Lett. A **324**, 26–35 (2004). https://doi.org/10.1016/j.physleta.2004.02.032
4. Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: Recurrent marked temporal point processes: embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 1555–1564. ACM, New York (2016). https://doi.org/10.1145/2939672.2939875. http://doi.acm.org/10.1145/2939672.2939875
5. Freiwald, W.A., et al.: Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. J. Neurosci. Methods **94**(1), 105–119 (1999). https://doi.org/10.1016/S0165-0270(99)00129-6. http://www.sciencedirect.com/science/article/pii/S0165027099001296
6. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969). http://www.jstor.org/stable/1912791
7. Guttorp, P.: An introduction to the theory of point processes (D. J. Daley and D. Vere-Jones). SIAM Rev. **32**(1), 175-2 (1990). https://search.proquest.com/docview/926135796?accountid=8330. Copyright © 1990 Society for Industrial and Applied Mathematics. Last updated 05 Mar 2012. CODEN - SIREAD
8. Kim, S., Putrino, D., Ghosh, S., Brown, E.N.: A granger causality measure for point process models of ensemble neural spiking activity. PLOS Comput. Biol. **7**(3), 1–13 (2011). https://doi.org/10.1371/journal.pcbi.1001110
9. Liniger, T.J.: Multivariate Hawkes processes. Ph.D. thesis, ETH Zurich. Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403 (2009). https://doi.org/10.3929/ethz-a-006037599
10. Marinazzo, D., Pellicoro, M., Stramaglia, S.: Nonlinear parametric model for Granger causality of time series **73**(6), 066216 (2006). https://doi.org/10.1103/PhysRevE.73.066216
11. Marinazzo, D., Pellicoro, M., Stramaglia, S.: Kernel method for nonlinear granger causality. Phys. Rev. Lett. **100**(14), 144103 (2008). https://doi.org/10.1103/PhysRevLett.100.144103
12. Ogata, Y.: Statistical models for earthquake occurrences and residual analysis for point processes. J. Am. Stat. Assoc. **83**(401), 9–27 (1988). https://doi.org/10.1080/01621459.1988.10478560. https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1988.10478560
13. Toke, I.M.: An introduction to Hawkes processes with applications to finance, p. 6, February 2011